



# Leveraging AI for biomedical data augmentation: a comparative review of model characteristics, performance analysis, and future research directions

Salha Albehairi<sup>1,2</sup> · Samiya Khan<sup>3</sup> · Reem Alotaibi<sup>1</sup> · Nofe Alganmi<sup>1</sup> · Mohammad Patwary<sup>2</sup>

Received: 16 July 2025 / Accepted: 1 November 2025

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

## Abstract

Artificial Intelligence (AI) is one of the most advanced technologies today, accelerating productivity, efficiency, and effectiveness in various sectors, including healthcare. However, defining such productivity requires data availability within the intervention field. In contrast to its potential, biomedical data for healthcare sector innovation is one of the most difficult data to access due to privacy concerns. One of the ways to overcome such difficulty is the use of AI to ‘Augment Data’ that has the potential to revolutionize healthcare sector productivity. This research paper presents a comprehensive survey on Data Augmentation (DA) for biomedical data, with specific interest in Generative Adversarial Network (GAN)-driven learning methods. It analyzes DA methods concerning the characteristics of these methods, performance metrics used to evaluate the quality of the generated data, and the limitations of each method. Additionally, this work includes a comprehensive experimental analysis to evaluate the performance of GANs and Conditional GANs (CGANs) on different data sources and sizes from two perspectives, which include characteristics and performance metrics. As a result, the analysis of existing GANs and modified versions of GANs indicates that there is significant potential in using generative models in biomedical applications. Key findings from this survey identify two major challenges in achieving reliable augmented data—a) unstable training for GAN models and b) the need for more reliable evaluation metrics. Addressing these challenges will be crucial for developing a new generation of GAN models that can ensure reliable DA techniques, with minimal training data and learning iterations.

**Keywords** Biomedical data · Data augmentation · Generative adversarial networks · GAN · Gene expression data · Small sample size

## 1 Introduction

Artificial Intelligence (AI) has emerged as a transformative technology that drives advancements across a wide range of industries and fields. From healthcare to finance, AI-powered solutions are redefining the approaches taken to address complex problems and facilitate informed decision-making. Using medical data and analytics is a key component of improving procedures and facilitating the administration

of medical services in the healthcare industry. In recent years, medical researchers, physicians, and patients have generated massive amounts of medical data, such as electronic health records (EHRs), biomedical imaging data, and genomic data [1, 2]. By leveraging the power of AI, these data can be utilized to aid in diagnosis, detection, and prediction of various diseases. However, the efficiency of training AI models relies on the availability of sufficient data to increase model accuracy and avoid overfitting [3]. Unfortunately, in certain fields, such as medicine, it is hard to access a sufficient amount of datasets due to data sensitivity and consent restrictions.

The available solutions for limited data accessibility are Data Augmentation (DA) [4, 5], Transfer Learning (TL)[6–9], self-supervised learning [10, 11], semi-supervised learning [12], few-shot learning [13–15], zero-shot learning [16–18], weakly supervised learning [19, 20], multitask learning, and ensemble learning [21–23]. To provide a clearer understanding of these solutions, Table 1 summarizes their

✉ Salha Albehairi  
shalbeheri@kau.edu.sa

<sup>1</sup> Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

<sup>2</sup> Faculty of Science and Engineering, University of Wolverhampton, Wolverhampton, UK

<sup>3</sup> School of Electronics and Computer Science, University of Southampton, Delhi, India

**Table 1** Advantages and disadvantages of learning approaches for limited data accessibility

| Learning category / Type                                   | Advantages   | Disadvantages  |
|--|--|--|
| Data augmentation (DA)                                     | Enhances dataset diversity and generalization without necessitating new data collection. This approach is cost-effective and commonly applied in biomedical imaging and genomics, significantly reducing training time.  | There is a risk of introducing unrealistic variations, which might amplify existing biases. Additionally, it can lead to increased computational costs for large-scale augmentation.   |
| Transfer learning (TL)                                     | Improves generalization without requiring extensive new data collection. It is widely used in biomedical applications, reducing the need for labeled data and accelerating model training through fine-tuning.   | There is a risk of negative transfer if the source and target domains diverge. This approach may also increase computational costs for fine-tuning, particularly with large models.  |
| Weak/limited supervision (Weakly-, Semi-, Self-supervised) | This method utilizes unlabeled or partially labeled data, thus minimizing annotation efforts. It enhances feature learning and generalization, particularly for high-dimensional biomedical data, making it suitable for privacy-restricted datasets like EHRs.                                      | Performance can degrade if data quality is poor; mislabeled signals significantly impact outcomes. The requirement for complex algorithms and denoising techniques further complicates implementation, particularly in large-scale training scenarios. |
| Low-data adaptation (Few-shot, Zero-shot)                  | These techniques enable learning from very few or even no labeled examples, making them particularly effective for rare diseases and facilitating rapid adaptation to emerging biomedical scenarios.   | However, performance may deteriorate for classes that are highly dissimilar or not seen in training. The success heavily relies on the quality of auxiliary samples, and negative transfer can occur in unrelated domains.                             |
| Knowledge sharing (Multitask, Ensemble)                    | By sharing information across tasks or combining models, this approach enhances generalization. It effectively reduces variance and overfitting, thereby increasing robustness to noise in biomedical data. Additionally, it speeds up convergence in related tasks like diagnosis and segmentation. | Nevertheless, there is a risk of task interference, and the complexity of architecture increases. This can lead to reduced interpretability of the resulting combined models.  |

advantages and disadvantages, highlighting their suitability for addressing limited data in biomedical contexts.

In the last few years, DA has become a popular method for improving models when applied to small datasets [24]. Data augmentation refers to manipulating existing data samples in different ways to artificially expand training datasets. This strategy has proven highly effective in enhancing the performance of machine learning models [24, 25]. DA is critical for training deep learning models, particularly in medical analysis. It helps in addressing limited data availability, data imbalance issues, and enhances model generalization. By using DA techniques, researchers can improve the performance of machine learning models in this critical domain, leading to more accurate and comprehensive analyses, as evidenced by several studies in this field [26–29]. This is particularly evident in cancer detection, where DA techniques help address limited data in genomic and imaging domains. The reviewed studies cover various cancers, including breast cancer and brain tumors, using gene expression data and mammograms [4, 30–40]. These works demonstrate the versatility of DA and its effectiveness in improving diagnostic accuracy across diverse oncology applications. For instance, a recent study using ResNet-50 on 566 HT-29

colon cancer cell images achieved 95.5% accuracy by applying DA techniques such as rotations, flips, shifts, cropping, and color normalization to enhance robustness and mitigate data variability [41]. Many types of DA techniques have been developed; each has unique characteristics and applications [42]. One common approach is geometric transformation, which is widely used in computer vision tasks. It involves operations such as rotation, scaling, and translation [43, 44]. Textual data augmentation, such as synonym replacement or sentence paraphrasing, is another commonly used method [45]. Synthetic data generation is also another type of DA. In this approach, new samples are generated based on existing data. Techniques like Generative Adversarial Networks (GANs) can be used to create realistic synthetic data points, enhancing DA in computer vision, Natural Language Processing (NLP), and genomic data analysis [46–48].

In the existing literature, researchers have presented several surveys concerning augmentation methods for biomedical data. However, their focus has primarily been on augmentation methods applied to specific types of images, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Mammographic Imaging, and Fundus Imaging [49], or solely concentrating on one type of image, as seen

in the survey that explores mammogram images exclusively [50]. Additionally, another survey paper [51] explores various imaging modalities that employ GAN techniques for data augmentation, while a different one focuses exclusively on three types of data augmentation methods: variational autoencoders, GANs, and diffusion models [52].

Two recent survey papers in this domain provide valuable insights: the first [29] includes a comparative evaluation of various GAN models utilized for DA specifically in the field of transcriptomics, while the second [53] systematically explores advancements in the application of GANs for gene expression data from 2020 to 2024, emphasizing their role as a powerful data augmentation tool for generating synthetic gene expression data and driving innovations in genomics research. However, there is a lack of a comprehensive review that examines the advanced techniques of DA applied to biomedical data. Additionally, this is the only survey paper from existing literature that explores the characteristics of DA methods and identifies gaps for each one. Therefore, this review presents the performance metrics commonly used to evaluate the quality of the generated data through DA, as well as the performance evaluation metrics used to assess the performance of ML models when utilizing the generated data for classification or segmentation tasks. The review also discusses the limitations and weaknesses associated with the state-of-the-art DA methods. The contributions of the paper are as follows:

- This review provides an overview of target diseases, data sources, DA methods, and ML models used for classification or segmentation tasks in biomedical data.
- This first-of-its-kind review explores the characteristics of each DA method, including their scalability, transferability, and robustness.
- The review examines performance metrics for evaluating the quality of data generated through augmentation (DA) techniques, as well as metrics used to assess the efficacy of machine learning (ML) models that utilize the generated data for classification or segmentation. Notably, the review highlights that existing literature has not yet investigated a comprehensive set of evaluation metrics for assessing the generated data itself.
- The review discusses the strengths, weaknesses, and scope for future work associated with the state-of-the-art DA models and identifies the existing research gaps in this area.

The lack of accessible biomedical data remains a serious issue, often caused by strict privacy rules and the limited availability of labeled samples. This study was motivated by these challenges and looks into advanced data DA methods as a practical way to overcome them. Earlier research has

focused on single areas like MRI or CT imaging [49, 50], or on particular DA approaches like GANs and VAEs [51, 52]. However, there is still no broad review that compares how these techniques perform in terms of scalability, robustness, and transferability across different biomedical datasets, including both images and genomic data.

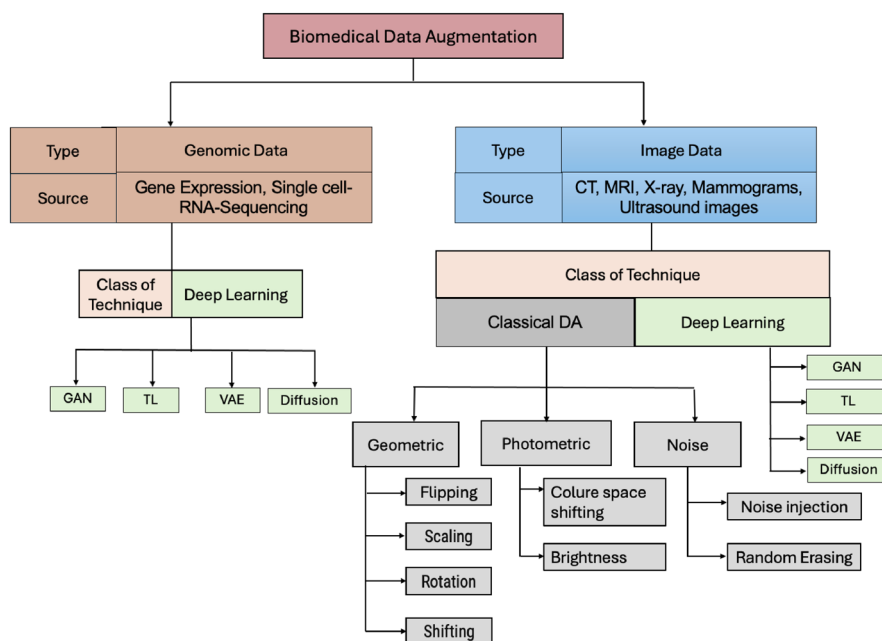
Another gap appears in how synthetic data are evaluated. This work aims to fill that gap by analyzing how DA can improve model performance and help deal with data shortages in healthcare. The motivation arises from the necessity to create reliable and efficient DA frameworks for biomedical applications. To guide this comprehensive review, we formulate the following research questions (RQs) and their objectives:

- RQ1: What are the recent DA techniques applied to biomedical data, and how do they address data limitation in genomics and imaging modalities? Objective: To map the landscape of DA methods and highlight modality-specific adaptations, informing targeted applications in healthcare.
- RQ2: What performance metrics are used to evaluate DA-generated data quality and downstream ML tasks (e.g., classification/segmentation)? Objective: To identify gaps in evaluation metrics.
- RQ3: What are the strengths, limitations, and future directions for GAN-based DA in biomedical contexts? Objective: To identify challenges like training instability and scalability to develop robust, privacy-preserving DA models.

The paper is organized as follows: Sect. 2 discusses the survey methodology for defining the criteria for including studies and the search strategy. In Sect. 3, existing data augmentation techniques and the evaluation metrics used for classification or segmentation tasks are discussed. Sect. 4 presents an experimental comparison of applying GANs to diverse data sources and sizes, along with the evaluation metrics used for each data source, and Sect. 5 discusses the results. Sect. 6 discusses the strengths, weaknesses, and scope of future work. Finally, Sect. 7 concludes the paper by summarizing key findings, emphasizing the significance and potential of data augmentation in biomedical analysis, and exploring future directions and open challenges in the field.

## 2 Survey methodology

The limited accessibility of biomedical data is a significant challenge in the field of biomedical research. Many biomedical studies rely on the quality and quantity of this data. To increase the performance of AI models, a sufficient amount

**Fig. 1** Taxonomy of data augmentation techniques

of data is necessary. To address this issue, biomedical data augmentation has emerged as a promising solution. This survey paper aims to provide a comprehensive overview of the state-of-the-art of biomedical data augmentation techniques, their applications, characteristics, and evaluation metrics. By exploring the various methods and strategies employed in biomedical data augmentation, this paper seeks to highlight the strengths and weaknesses of each approach and identify the research gaps.

This study determined inclusion and exclusion criteria based on parameters such as modality, publication date, and types of academic publications. The modality refers to biomedical data types such as genomics data, X-rays, CT scans, and mammogram images. In selecting the articles, the review reflected current trends and developments in the application of data augmentation in biomedical research by covering the period from 2020 to 2024. For academic publications, it only included papers that had been published in peer-reviewed journals. Using peer review criteria ensures high standards of quality and accuracy because research publications are carefully assessed by experts in the relevant field. While recognizing the importance of conference papers and preprints in scientific discussions, the emphasis was on peer-reviewed journal articles. This strategy ensures the inclusion of rigorously peer-reviewed papers, thereby strengthening the validity of this review. In order to find relevant publications for this review, the search strategy included using keywords such as "data augmentation," "Generative Adversarial Networks," "biomedical data," and "gene expression."

By applying the specified inclusion and exclusion criteria, this survey paper was able to identify and analyze a focused set of high-quality, peer-reviewed research articles on the

topic of biomedical data augmentation, with all articles published between 2020 and 2024. The imbalance between the number of studies on image data and genomic data in this study arises from several factors. Image data has a longer history of data augmentation due to its visual nature, which lends itself to various augmentation techniques such as rotation, scaling, and cropping. Furthermore, medical imaging is widely used in diagnosis and treatment, leading to extensive research in this area. In contrast, data augmentation for genomic data is relatively newer and more specialized. The complexity and high dimensionality of genomic data pose significant challenges, and research in this area is still evolving.

A taxonomy of data augmentation techniques used in the reviewed studies is illustrated in Fig. 1, categorizing the various biomedical data augmentation methods applied to image and genomics data. Detailed explanations of these techniques are provided in Section 3.1. For image data, both classical data augmentation techniques and deep learning methods are utilized. Classical data augmentation includes geometric transformations, photometric adjustments, and noise injection. Deep learning approaches involve various versions of Generative Adversarial Networks (GANs), Transfer Learning (TL), Variational Autoencoders (VAEs), and diffusion models. Similarly, recent deep learning techniques have been applied to genomic data for data augmentation, including GANs, TL, VAEs, and diffusion models.

### 3 Biomedical data augmentation techniques in reviewed literature

DA techniques have been used to solve the limited data accessibility, which primarily arises in the medical domain, due to privacy and consent restrictions. An overview of the identified papers will be presented, including details of the study such as target diseases, data sources, DA techniques, and ML models used for classification or segmentation tasks in biomedical data. Additionally, the evaluation metrics used to assess the performance of DA techniques will be discussed.

#### 3.1 Explanations of data augmentation techniques in the taxonomy

To explain the taxonomy in Fig. 1, we describe each data augmentation technique. We group them for genomic and image data. These methods generate synthetic data to address limited samples. However, they may increase computational costs or introduce artifacts. We discuss these limits in Section 6.

##### 3.1.1 Classical data augmentation (primarily for image data)

Classical methods alter existing data via simple operations. They are computationally efficient. However, they fit domains where changes preserve meaning. For instance, they suit image data better than genomic sequences [54]. Geometric Transformations modify data layout, with examples including flipping, scaling, shearing, rotation, and shifting. Flipping mirrors images, while scaling resizes them, shearing distorts content, rotation angles images, and shifting repositions them. These techniques enhance robustness to shape variations in biomedical images like X-rays or CT scans [54]. Photometric Adjustments change color and light, involving color space shifts such as RGB to HSV, and brightness adjustments alter intensity. They manage lighting differences in medical images; however, they do not apply to non-visual data [54]. Noise Injection introduces random noise, like Gaussian, to simulate real imperfections such as MRI sensor errors. It strengthens models, but overuse may reduce data quality [54].

##### 3.1.2 Deep learning-based methods (for both genomic and image data)

Deep learning-based methods employ neural networks to create new data. These methods produce realistic samples but require more resources. Deep Learning-based Methods may also face training instability. Additionally, GANs consist of a generator that produces synthetic samples and a discriminator that assesses realism. GANs exceed expectations in

biomedical images, such as CT for COVID-19. Furthermore, GANs also fit gene expression data. However, issues include mode collapse and training instability [55]. Transfer Learning (TL) adapts pre-trained models from large datasets, like ImageNet. It accelerates training with small biomedical data; however, domain mismatches can present challenges [56]. Variational Autoencoders (VAEs) learn probabilistic latent patterns, where an encoder compresses data, and a decoder generates new samples. VAEs aid genomic augmentation, and they offer uncertainty estimates, though outputs may lack sharpness or diversity [57]. Diffusion Models add noise gradually, and then diffusion models train to reverse it. This yields high-quality samples, like brain tumor MRIs, and diffusion models provide strong realism. But diffusion models are slow and resource-intensive [58].

#### 3.2 Comparison of state-of-the-art models

After examining the state-of-the-art in DA techniques applied to biomedical data, this study identified relevant research papers addressing the problem. Tables 2 and 3 present information regarding data category, data source, target disease, type of augmentation techniques utilized, as well as the classification or segmentation techniques employed in the models.

The "Data Category" column presents the type of data, which is either genomic or image data. The "Data Source" column outlines the modality of biomedical data used in the studies, such as genomic data, X-ray images, CT images, or MRI images. The "Target Disease" column indicates the specific diseases that the researchers aimed to address. The "Data Augmentation Techniques" column lists the various data augmentation methods employed to increase the training datasets, such as classical DA, GAN, or Transfer Learning (TL). "Result" column presents the result of each study. Finally, the "Classification/Segmentation Techniques" column describes the machine learning models or algorithms used for the downstream tasks of classification or segmentation, such as convolutional neural networks (CNN), deep neural networks (DNN), and support vector machines (SVM).

Upon reviewing Tables 2 and 3, a variety of target diseases can be observed, including COVID-19, different types of cancer, Parkinson's disease, pneumonia, and schizophrenia. Additionally, the data sources include diverse modalities such as CT, MRI, X-ray images, and gene expression data. The reviewed papers have used different DA techniques such as classical data augmentation, transfer learning, and GAN-based techniques. The last column "Classification/Segmentation Techniques" in Table 2 and Table 3 pertain to the Machine Learning (ML) model utilized for classification or segmentation purposes, after incorporating both real and synthetic data, in order to assess whether any improve-

**Table 2** Overview of research papers (2020–2024, Part 1)

| Data Cat. | Data Src.  | Disease                | Augmentation techniques | Result  | Classification/ segmentation      |
|-----------|------------|------------------------|-------------------------|---|-----------------------------------|
| Genomic   | Gene Expr. | Schizophrenia [26]     | GAN-based               | 93 percent accuracy                                 | Classification: LR, SVM, KNN, MLP |
|           |            | Cancer types [59]      | GAN-based               | Improved accuracy with WGAN-GP                      | Classification: MLP, KNN, DT      |
|           |            | Cancer types [31]      | GAN-based               | Enhanced gene classification                        | Classification: CNN               |
|           |            | Breast/skin tumor [60] | GAN, Diffusion          | Improved bulk RNA-seq classification                | Classification: SVM, RF, XGBoost  |
|           |            | Rare diseases [39]     | Diffusion               | Enhanced rare disease genomic augmentation          | Classification: CNN               |
|           |            | Rare diseases [40]     | GAN-based               | Improved rare disease classification                | Classification: Deep learning     |
|           | scRNA-Seq  | Marker genes [61]      | GAN-based               | Improved downstream analyses                        | Classification: Random Forest     |
|           |            | Cancer types [62]      | GAN-based               | Improved unsupervised tumor subclone classification | Classification: Unsupervised      |

ments were observed. There are several ML models applied for classification or segmentation tasks, including classification deep transfer models, Convolutional Neural Networks (CNN), Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP).

Existing literature has not yet provided a comprehensive examination of the characteristics and limitations of the various DA techniques employed in machine learning research. This section provides a theoretical comparison of the different DA methods to fill this gap and thoroughly investigate the scalability, transferability, and robustness of the state-of-the-art DA methods.

From the theoretical perspective, Table 4 presents the characteristics, advantages and disadvantages of the identified research papers on DA models. The key characteristics explored in the literature review are as follows: Firstly, according to existing literature, the scalability of the models evaluates the adaptability of the models to different sizes and types of biomedical datasets [74]. Evaluating the scalability of an AI model makes it applicable to both small and large datasets. It also enhances the effectiveness of the model in real-life applications. Secondly, the literature has considered the robustness of the models, which refers to the models' ability to handle noisy or incomplete biomedical data [75]. This feature makes the model more accurate and reliable, especially in situations when perfect data is rarely accessible. Lastly, the literature explored the transferability of the models, which explores the models' capacity to generalize and transfer learning to different biomedical tasks [76]. The models' transferability is also crucial, as it can increase the

overall value of the model and reduce the need for retraining on new tasks. These characteristics offer insights into the performance and capabilities of the DA models discussed in the literature.

Out of all the papers analyzed, as shown in Table 4, seven models [27, 40, 61, 63, 65, 68, 70, 78] were identified as scalable. Additionally, twelve papers [4, 26, 27, 31, 36, 38, 39, 63, 64, 68, 70, 71] highlighted the robustness of their models. On the other hand, ten papers [27, 30, 32, 38–40, 61, 64, 65, 73, 78] emphasized the transferability of their models. It is noteworthy that only one paper [27] covers all the identified characteristics of DA models for MRI images. As part of this study, the analysis of the model's characteristics described in the identified research papers revealed that none of the existing papers have focused on the key characteristics of scalability, robustness, and transferability comprehensively. This observation highlights one of the key contributions of this work, which is to examine and understand these important properties in the context of DA methods for biomedical applications. This suggests that the research community may benefit from further exploration of DA techniques that exhibit a combination of desirable properties, to address the diverse challenges in biomedical data analysis.

### 3.3 Evaluation metrics for assessing DA techniques in reviewed literature

The reviewed literature emphasizes the importance of employing evaluation metrics to assess the performance of DA techniques. These evaluation metrics provide quantifiable measures to assess how well the generated data matches the

**Table 3** Overview of research papers (2020–2024, Part 2)

| Data Cat.        | Data Src.      | Disease                      | Augmentation techniques                           | Result                                    | Classification/ segmentation                        |
|------------------|----------------|------------------------------|---|---|---|
| Image            | CT             | COVID-19 [28]                | Classical, GAN, TL                                | 90 percent detection accuracy             | Classification: AlexNet, VGG16, GoogleNet, ResNet50 |
|                  |                | Breast mass [35]             | GAN, Classical, TL                                | Improved breast mass classification       | Classification: CNN                                 |
|                  |                | Bone metastasis [63]         | Classical, Mixup, RICAP                           | Improved segmentation accuracy            | Segmentation: U-Net                                 |
|                  |                | Pneumonia [64]               | Diffusion   | Superior to GANs in image synthesis       | Classification: CNN                                 |
|                  |                | COVID-19 [65]                | Classical, GAN                                    | Improved CT-based classification          | Classification: CNN                                 |
|                  | MRI            | Alzheimer [27]               | VAE   | Improved classification in HDLSS          | Classification: CNN                                 |
|                  |                | Rectal cancer [37]           | Classical, TL                                     | High accuracy in lymph node segmentation  | Segmentation: R-CNN                                 |
|                  |                | Parkinson [66]               | GAN, TL   | 99.23 percent accuracy                    | Classification: CNN                                 |
|                  |                | Bony structure [67]          | GAN   | Improved cross-modality segmentation      | Segmentation: Seg model                             |
|                  |                | Brain tumor [38]             | Diffusion   | Improved brain tumor segmentation         | Segmentation: U-Net                                 |
|                  |                | Brain tumor [68]             | GAN, Diffusion                                    | Improved deep learning augmentation       | Segmentation: U-Net                                 |
|                  |                | X-ray                        | COVID-19 [4]                                      | Classical                                 | Improved diagnosis with segmentation                |
|                  | Pneumonia [69] |                              | GAN   | Improved pneumonia detection              | Classification: NN                                  |
|                  | COVID-19 [30]  |                              | Classical, GAN, TL                                | Improved gene data classification         | Classification: CNN                                 |
|                  | Pneumonia [70] |                              | GAN, Classical                                    | Enhanced explainability in classification | Classification: CNN                                 |
|                  | Mamm.          | Cancer types [28]            | GAN   | 90 percent accuracy                       | Classification: DNN                                 |
|                  |                | Breast mass [71]             | Classical, TL                                     | Improved detection with wavelet CNN       | Detection: Faster R-CNN                             |
|                  |                | Breast cancer [34]           | GAN   | Improved mass detection in FFDM           | Classification: CNN                                 |
|                  |                | Breast cancer [72]           | Classical   | Improved segmentation and classification  | Segmentation: U-Net; Classification: CBR            |
|                  | Ultra.         | Cardiovascular diseases [36] | GAN, Classical                                    | Improved segmentation with CNN-CBR        | Segmentation: U-Net                                 |
| Breast mass [73] |                | GAN                          | Improved classification with speckle augmentation | Classification: CNN                       |   |

original data distribution. This allows researchers to determine the performance of the DA methods in generating high-quality synthetic samples. Additionally, the evaluation metrics used to assess the performance of ML models for classification or segmentation tasks are explored as shown in Table 5 and Table 6, respectively. The abbreviations used for the evaluation metrics are as follows: Rec refers to Recall, Acc stands for Accuracy, Spec is Specificity, Prec represents Precision, F1 is the F1-Score, PPV is Positive Predictive Value, BS refers to Brier Score, Kappa indicates Cohen's

Kappa, AUC is the Area Under the Receiver Operating Characteristic Curve, and Logloss stands for Logarithmic Loss. These metrics provide insights into how the combination of original data and synthetic data impacts the classification or segmentation task.

In Table 5, the widely used evaluation metrics across the different biomedical data augmentation to assess ML model for classification purposes are identified to be accuracy, recall, precision, and F1-score. However, the Dice Coefficient, AccuracySD, and Jaccard Index are also common,

**Table 4** DA Model Characteristics

| References  | Scalability | Robustness | Transferability | Advantages   | Disadvantages   |
|---|-------------|------------|-----------------|--|---|
| [26], [31], [62], [4], [71], [36]                         | ✗           | ✓          | ✗               | High-quality samples; Better clustering accuracy   | High computational requirements; Limited dataset evaluation |
| [27]  | ✓           | ✓          | ✓               | Effective three-dimensional MRI classification   | Lack of synthetic data evaluation; Potential bias issues    |
| [28], [60], [35], [37],[66], [67], [69], [34], [72], [77] | ✗           | ✗          | ✗               | Transfer learning improves performance; Synthetic data generation from unrelated lesions | Limited dataset size; Lack of synthetic data evaluation     |
| [40], [61], [65], [78]                                    | ✓           | ✗          | ✓               | Trained in separate experiments; Reduces overfitting                                     | High false discovery rate; Batch effects                    |
| [63], [68], [70]  | ✓           | ✓          | ✗               | Improved segmentation accuracy   | Limited data availability; Overfitting issues               |
| [30], [73], [32]  | ✗           | ✗          | ✓               | Improved accuracy; Enhanced feature extraction   | Slow segmentation process; Limited to segmentation tasks    |
| [39], [64], [38]  | ✗           | ✓          | ✓               | Generated high-fidelity data; Improved detection accuracy                                | High computational demands; Overfitting with small datasets |

**Table 5** Performance metrics for classification after DA

| References     | Rec | Acc | Spec | Prec | F1 | PPV | BS | Kappa | AUC | Logloss |
|----------------|-----|-----|------|------|----|-----|----|-------|-----|---------|
| [26]           | ✗   | ✓   | ✗    | ✗    | ✗  | ✗   | ✓  | ✗     | ✗   | ✗       |
| [28]           | ✓   | ✓   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [31]           | ✗   | ✓   | ✗    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [60]           | ✗   | ✗   | ✗    | ✗    | ✓  | ✗   | ✗  | ✓     | ✓   | ✓       |
| [39]           | ✓   | ✓   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [40]           | ✗   | ✓   | ✗    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [61]           | ✗   | ✗   | ✗    | ✗    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [35]           | ✓   | ✓   | ✓    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [63]           | ✓   | ✗   | ✗    | ✗    | ✗  | ✓   | ✗  | ✗     | ✗   | ✗       |
| [64]           | ✓   | ✓   | ✗    | ✓    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [65]           | ✓   | ✓   | ✗    | ✗    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [37]           | ✓   | ✗   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [66]           | ✓   | ✓   | ✓    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [38]           | ✓   | ✓   | ✗    | ✗    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [68]           | ✓   | ✓   | ✗    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [69]           | ✓   | ✓   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [70]           | ✓   | ✓   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [71]           | ✓   | ✓   | ✓    | ✓    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [34]           | ✓   | ✗   | ✓    | ✗    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [72]           | ✓   | ✓   | ✓    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [36]           | ✓   | ✓   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [77]           | ✓   | ✓   | ✓    | ✓    | ✓  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [78]           | ✓   | ✗   | ✗    | ✓    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| [32]           | ✗   | ✓   | ✗    | ✗    | ✗  | ✗   | ✗  | ✗     | ✗   | ✗       |
| No. of Studies | 16  | 16  | 6    | 9    | 9  | 1   | 1  | 1     | 1   | 1       |

**Table 6** Performance metrics for segmentation after DA

| References     | Dice Coefficient | AccuracySD | Jaccard Index |
|----------------|------------------|------------|---------------|
| [27]           | ✗                | ✓          | ✗             |
| [31]           | ✗                | ✓          | ✗             |
| [67]           | ✓                | ✗          | ✗             |
| [4]            | ✓                | ✗          | ✓             |
| [73]           | ✓                | ✗          | ✗             |
| [77]           | ✓                | ✗          | ✗             |
| No. of Studies | 4                | 2          | 1             |

indicating a focus on evaluating segmentation performance (see Table 6). Some less common metrics include Spec, AccSD, PPV, SD, and FPR, which are used selectively in different studies.

Analyzing these performance metrics is essential for understanding the strengths and limitations of using data augmentation to address data scarcity in biomedical applications. It was found that only eight research papers evaluated the quality of the generated data using various metrics such as Wasserstein distance (WD) [31, 69], Pearson correlation coefficient [61], KL divergence [30], Mean Absolute Error (MAE) [32], and t-SNE plots [61, 62, 72], and Fréchet Inception Distance (FID) [73] before using it for classification or segmentation tasks.

## 4 Model evaluation and comparative performance analysis

To investigate the performance and characteristics of these generative models and explore the evaluation metrics for each type of data, this section presents an experimental comparison of applying GANs to diverse data sources and sizes. As highlighted in the introduction, existing literature has not yet investigated a comprehensive set of evaluation metrics for assessing the quality of data generated through augmentation techniques. This experimental comparison is required to better understand the strengths and limitations of GANs when faced with different data modalities and volumes and to identify suitable evaluation metrics for the generated data.

### 4.1 DA techniques and evaluation metrics for experimental use

Literature suggests that GAN-based is one of the most advanced techniques for biomedical data augmentation. GANs are powerful unsupervised learning neural networks. There are two neural networks in a GAN, namely a Discriminator (D) and a Generator (G). In order to produce artificial data close to real data, adversarial training is used, often described as a minimax two-player game. Generators

produce random noise samples in an attempt to fool the discriminator, whose task is to distinguish between generated and real data. As a result of this competitive interaction, realistic, high-quality samples are produced, which gives each network the opportunity to advance. The minimax two-player game depending on  $G$  and  $D$  is evaluated with a cost function  $V(G, D)$ , which is defined by Equation 1:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))] \quad (1)$$

where  $x$  is sampled from the real data distribution  $p_{\text{data}}(x)$ ,  $z$  is from input noise variables  $p_z(z)$ , and  $\mathbb{E}(\cdot)$  is the expectation. A distribution  $p_g(x)$  can also be defined as a generated data distribution,  $G(z)$  as the data generated by  $G$ , subject to  $p_{\text{data}}$ , and  $D(x)$  as the likelihood that  $x$  is sampled from  $p_{\text{data}}$ . The second model is the Conditional Generative Adversarial Network (CGAN), which is an improved version of the GAN. The only difference between them is that CGAN adds label information and uses the powerful learning capabilities of neural networks to produce samples with specified labels.

The generated data is evaluated by using different evaluation metrics for gene expression data, which are Wasserstein Distance (WD) and correlation coefficient (r-value). However, the metrics used for evaluating the image data included WD, Fréchet Inception Distance (FID), and KL divergence, as found in the existing literature. WD, which is calculated between  $\mathbb{P}_r, \mathbb{P}_g$ , is defined as follows:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x_r, x_g) \sim \gamma} [\|x_r - x_g\|] \quad (2)$$

The set  $\Pi(\mathbb{P}_r, \mathbb{P}_g)$  consists of all joint distributions  $\gamma(x_r, x_g)$  whose marginals are  $\mathbb{P}_r$  and  $\mathbb{P}_g$ , respectively. A transport plan, also called  $\gamma(x_r, x_g)$ , represents how much "mass" has to be transferred from  $x_r$  to  $x_g$  in order to convert  $\mathbb{P}_r$  to  $\mathbb{P}_g$ . The range is typically defined as  $[0, \infty)$ , with zero representing the optimal value. The formula for the second evaluation metric, r-value, is provided as follows:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \times \sigma_Y} \quad (3)$$

where  $\rho_{X,Y}$  is the Pearson product-moment correlation coefficient,  $\text{cov}(X, Y)$  is the covariance of variables  $X$  and  $Y$ ,  $\sigma_X$  is the standard deviation of  $X$ , and  $\sigma_Y$  is the standard deviation of  $Y$ .

For image data, in addition to WD, the reviewed studies also utilized other evaluation metrics such as FID [79] and KL divergence to assess the quality of generated images. The range of FID values is typically defined as  $[0, \infty)$ , with zero indicating the optimal value. The formula is presented below:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (4)$$

where  $\mu_r$  is the mean value of the real data,  $\mu_g$  is the mean value of the generated data,  $\Sigma_r$  is the covariance matrix of the real data,  $\Sigma_g$  is the covariance matrix of the generated data,  $\|\cdot\|$  denotes the Euclidean distance, and  $\text{Tr}(\cdot)$  represents the trace of a matrix. The next evaluation metric is Kullback–Leibler (KL) Divergence [80]. A lower KL divergence indicates a closer match between the two distributions, while a higher KL divergence suggests a larger discrepancy. The calculation of KL divergence can be performed using the following formula:

$$\text{KL}(P\|Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (5)$$

where  $x$  is a possible event in the probability space,  $P(x)$  and  $Q(x)$  are the probabilities of  $x$  in the distributions  $P$  and  $Q$ , respectively, and the ratio  $P(x)/Q(x)$  represents the relative likelihood of event  $x$  according to  $P$  compared to  $Q$ .

To compare the most advanced DA techniques in the literature and identify research gaps, this section discussed the existing DA techniques selected for our experimental study and provides a detailed explanation of the evaluation metrics chosen. As previously mentioned, the findings from all the selected papers underscore the relevance of GAN-based models for DA in biomedical data. Several of these papers advance GAN models by addressing critical issues such as data quality, overfitting, and mode collapse. However, in terms of stability, only one study among all the papers proposed a solution to improve stability. Han et al. (2022) achieved this by replacing the default GAN loss function with the Wasserstein distance loss [31]. Nevertheless, training stability remains an unresolved challenge in the literature, which is the primary focus of this paper.

## 4.2 Experimental setup and methodology

The experiments were conducted using Python 2.7.5 and a computational environment consisting of an Intel(R)

Xeon(R) CPU E5-2695 v2 @ 2.40GHz with 94GB of RAM, running CentOS Linux 7 (Core). All of the datasets used in this study were available online. The gene expression dataset for Schizophrenia was downloaded from the NCBI archive (accession number GSE93987). The chest CT images for COVID-19 and chest X-ray images for pneumonia were available on Kaggle. All the results were generated using the same experimental setup.

For preprocessing, the gene expression data normalized using the MinMaxScaler from scikit-learn to scale values between 0 and 1, ensuring consistency across features. Image data were preprocessed by resizing to 64x64 pixels and normalizing pixel values to the range  $[0, 1]$  using OpenCV. Implementation leveraged Python libraries including PyTorch for gene data, utilizing a Generator and Discriminator architecture with Adam optimizers at a learning rate of 0.001 over 200 epochs, and TensorFlow/Keras for image data, employing RMSprop with a learning rate of 0.0001 and weight decay of 6e-9 over 20,000 epochs. These setups facilitated the generation and evaluation of synthetic data. To illustrate the implementation of the models used in this study, pseudocode for GAN and CGAN is provided in Algorithms 1 and Algorithms 2, respectively.

---

### Algorithm 1 Pseudocode for GAN Model

---

```

1: Initialize Generator  $G$  and Discriminator  $D$  with random weights
2: Define loss functions:  $L_D$  (Discriminator loss),  $L_G$  (Generator loss)
3: for each training iteration do
4:   Sample random noise vector  $z$  from noise distribution  $p_z(z)$ 
5:   Generate fake data  $G(z)$ 
6:   Sample real data  $x$  from true data distribution  $p_{data}(x)$ 
7:   Update  $D$  by minimizing  $L_D = -\mathbb{E}_{x \sim p_{data}}[\log D(x)] - \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$ 
8:   Update  $G$  by minimizing  $L_G = -\mathbb{E}_{z \sim p_z}[\log D(G(z))]$ 
9: end for
10: Output: Generated synthetic data  $G(z)$  (gene expression or images)

```

---



---

### Algorithm 2 Pseudocode for CGAN Model

---

```

1: Initialize Generator  $G$  and Discriminator  $D$  with random weights
2: Define loss functions:  $L_D$  (Discriminator loss),  $L_G$  (Generator loss)
3: for each training iteration do
4:   Sample random noise vector  $z$  from noise distribution  $p_z(z)$ 
5:   Sample condition  $c$  (class label)
6:   Generate fake data  $G(z, c)$  conditioned on  $c$ 
7:   Sample real data  $x$  and condition  $c$  from true data distribution  $p_{data}(x, c)$ 
8:   Update  $D$  by minimizing  $L_D = -\mathbb{E}_{(x,c) \sim p_{data}}[\log D(x, c)] - \mathbb{E}_{(z,c) \sim p_z}[\log(1 - D(G(z, c), c))]$ 
9:   Update  $G$  by minimizing  $L_G = -\mathbb{E}_{(z,c) \sim p_z}[\log D(G(z, c), c)]$ 
10: end for
11: Output: Generated synthetic data  $G(z, c)$  (conditioned gene expression or images)

```

---

From the perspective of the experimental comparison, different versions of GAN were applied to various data sources to evaluate model characteristics and assess the quality of generated data using specific evaluation metrics. Specifically, GAN and CGAN were applied to three types of biomedical data: gene expression with two different dataset sizes (50, 202 samples) as shown in Table 7. Additionally, the impact of varying the GAN configurations on gene expression data was investigated. Different versions of the GAN were used to control the relationship between the input and generated (output) data. Specifically, GAN(1) refers to a model where the size of the input and output data are equal, while GAN(2) generates output data that is twice the size of the input data, and GAN(3) produces output data three times larger than the input data. The same approach was applied to CGAN. The second type of data was X-ray images, with dataset sizes of 600 and 5,233 images. The third data type was CT images, with dataset sizes of 500 and 3,000 images as shown in Table 8. For all dataset types, the study included multiple repetitions (Rep) of the experiments, where each repetition represents a single isolated run of the code on the same dataset.

### 4.3 Results and analysis

After applying GAN and CGAN to different biomedical data sources with various sizes and many repetitions (Rep), the results yielded some interesting observations (see Table 7 and Table 8). Notably, the values were observed to vary considerably across the different Rep and different sizes of the dataset, even when utilizing the same code and dataset. For instance, in Table 7, after applying GAN to 50 samples of gene expression data with three Rep, different results were obtained for each evaluation metric. For the WD metric, the first run yielded a value of 0.734, the second run 0.31, and the third run 0.521. A similar pattern was observed for the r-value, with values of 0.78, 0.894, and 0.71 across the three repetitions. Furthermore, when applying CGAN to a larger dataset of 202 samples, the same inconsistent behavior was evident. As shown in Table 7, with CGAN(1), the WD values varied across the three Rep, taking on values of 0.006, 0.124, and 0.0632. Similarly, the r-values were 0.99, 0.676, and 0.707 for the respective repetitions. The same behavior was observed in the image data in Table 8.

Existing literature suggests that increasing the size of the dataset generally leads to improved model performance [31]. However, the results of the current experiment show that in certain cases, increasing the dataset size did not lead to better performance and, in some instances, actually resulted in worse outcomes. For instance, as shown in Table 7, when applying the GAN(3) model to gene expression data, with a dataset size of 50 samples, the WD metric was 0.220 in Rep.1. However, when the dataset size was increased to 202 samples, also with GAN(3), the WD metric worsened to

1.485 in Rep.1. The same trend was observed with image data in Table 8: for CT images, GAN(500) had an FID score of 1680.90 in Rep.1, while GAN(3000) had an FID score of 2328.044 in Rep.1. The observed variance across repetitions, even with identical code and datasets, strongly suggests instability in the training process. Further investigation into the causes of this instability is necessary.

Another key observation from the study is the inconsistency in model evaluation across various performance metrics. The evaluation of the generated data showed that the results could vary significantly, even when applied to the same dataset and model configuration. For instance, in Table 7, the experiment with gene expression data (50 samples) shows that GAN(1) for the first repetition (Rep) produced a WD of 0.734, which is considered a poor result, as the optimal value is close to 0. However, the same experiment yielded an r-value of 0.78, which is considered a good result, as it is closer to the ideal value of 1. Additionally, for image data in Table 8, the WD for X-ray images using CGAN(5233) in Rep.2 was 3.972, indicating a poor result, while the KL Divergence value was 0.003, which indicates a good result, as a value closer to 0 is optimal. The conflicting evaluation results suggest that it cannot be confidently determined which metric provides the most reliable assessment of the model's performance.

Additionally, the percentage of variability in Table 7 and Table 8 is extremely high in most cases. While the variability is acceptable in some instances, applying the same model to identical data can yield another evaluation metric with high variability. For instance, in Table 7 for GAN(3) with 202 samples, the WD in Rep.1 is 1.485, while the standard deviation (SD) is 0.488, indicating high variability. Another example is the r-value for CGAN(1) with 202 samples, where the r-value for Rep.1 is 0.99 and the SD is 0.173, again showing high variability. Similarly, for image data in Table 8, for CT images using CGAN(3000), the FID score for Rep.3 is 2744.717 with an SD of 747.668, further highlighting high variability.

From a characteristics perspective, in the GAN model, when the data size increases in X-ray images from 648 to 5233 (GAN(648) and GAN(5233)), the FID score decreases in all repetitions (see Table 8), indicating a small improvement in performance. For gene expression data, increasing the sample size from 50 to 202 does not consistently improve the result. However, CGAN shows significant improvement with larger datasets. For instance, in X-ray images, the FID score decreases drastically from 1786.043 for CGAN(648) to 1099.218 for CGAN(5233). Also, for gene expression data in Table 7, increasing the sample size from 50 to 202 improves WD consistently (e.g., CGAN(1) from 0.196 to 0.006), suggesting better scalability. For transferability, CGANs show better and more consistent performance across different types of data (X-rays and CTs) and different sizes, suggesting

**Table 7** Comparison of model performance of gene expression data

| Metric  | Model   | Gene expression (50) |       |       | SD    | Gene expression (202) |       |        | SD    |
|---------|---------|----------------------|-------|-------|-------|-----------------------|-------|--------|-------|
|         |         | Rep.1                | Rep.2 | Rep.3 |       | Rep.1                 | Rep.2 | Rep.3  |       |
| WD      | GAN(1)  | 0.734                | 0.31  | 0.521 | 0.212 | 0.62                  | 0.623 | 0.814  | 0.111 |
|         | GAN(2)  | 0.525                | 0.886 | 1.125 | 0.302 | 0.283                 | 0.753 | 0.4154 | 0.242 |
|         | GAN(3)  | 0.220                | 0.554 | 0.619 | 0.214 | 1.485                 | 0.791 | 0.541  | 0.488 |
|         | CGAN(1) | 0.196                | 0.124 | 0.053 | 0.072 | 0.006                 | 0.124 | 0.0632 | 0.059 |
|         | CGAN(2) | 0.074                | 0.142 | 0.222 | 0.074 | 0.074                 | 0.039 | 0.04   | 0.020 |
|         | CGAN(3) | 0.13                 | 0.090 | 0.087 | 0.024 | 0.052                 | 0.042 | 0.062  | 0.010 |
| r-value | GAN(1)  | 0.78                 | 0.894 | 0.71  | 0.092 | 0.81                  | 0.802 | 0.805  | 0.003 |
|         | GAN(2)  | 0.809                | 0.567 | 0.428 | 0.192 | 0.82                  | 0.825 | 0.739  | 0.048 |
|         | GAN(3)  | 0.812                | 0.815 | 0.825 | 0.007 | 0.21                  | 0.717 | 0.775  | 0.310 |
|         | CGAN(1) | 0.663                | 0.803 | 0.766 | 0.072 | 0.99                  | 0.676 | 0.707  | 0.173 |
|         | CGAN(2) | 0.78                 | 0.71  | 0.791 | 0.044 | 0.722                 | 0.701 | 0.705  | 0.011 |
|         | CGAN(3) | 0.696                | 0.678 | 0.746 | 0.035 | 0.69                  | 0.74  | 0.706  | 0.025 |

**Table 8** Comparison of GAN and CGAN models on X-ray and CT images

| Modality | Metric | Model (size) | Rep.1    | Rep.2    | Rep.3    | SD      |
|----------|--------|--------------|----------|----------|----------|---------|
| CT       | FID    | GAN (500)    | 1680.90  | 1628.692 | 1809.733 | 93.184  |
|          |        | GAN (3000)   | 2328.044 | 2730.015 | 2744.717 | 236.436 |
|          |        | CGAN (500)   | 1558.69  | 1531.357 | 1465.660 | 47.815  |
|          |        | CGAN (3000)  | 1465.88  | 1434.141 | 2744.717 | 747.668 |
|          | WD     | GAN (500)    | 28.728   | 81.164   | 34.898   | 28.659  |
|          |        | GAN (3000)   | 19.274   | 14.127   | 17.653   | 2.632   |
|          |        | CGAN (500)   | 24.788   | 29.976   | 18.853   | 5.566   |
|          |        | CGAN (3000)  | 7.932    | 7.4726   | 16.526   | 5.100   |
|          | KL     | GAN (500)    | 0.175    | 0.704    | 0.268    | 0.282   |
|          |        | GAN (3000)   | 0.134    | 0.093    | 0.352    | 0.139   |
|          |        | CGAN (500)   | 0.064    | 0.138    | 0.088    | 0.037   |
|          |        | CGAN (3000)  | 0.036    | 0.036    | 0.352    | 0.182   |
| X-Ray    | FID    | GAN (648)    | 1901.158 | 2261.705 | 2534.180 | 317.530 |
|          |        | GAN (5233)   | 1840.789 | 2054.279 | 2271.390 | 215.303 |
|          |        | CGAN (648)   | 1786.043 | 1936.256 | 2016.933 | 117.177 |
|          |        | CGAN (5233)  | 1099.218 | 1335.23  | 1156.755 | 123.062 |
|          | WD     | GAN (648)    | 11.489   | 3.456    | 23.553   | 10.115  |
|          |        | GAN (5233)   | 7.821    | 7.83     | 33.184   | 14.641  |
|          |        | CGAN (648)   | 7.832    | 4.870    | 11.858   | 3.507   |
|          |        | CGAN (5233)  | 5.518    | 3.972    | 10.173   | 3.228   |
|          | KL     | GAN (648)    | 0.028    | 0.020    | 0.242    | 0.126   |
|          |        | GAN (5233)   | 0.027    | 0.0341   | 0.451    | 0.243   |
|          |        | CGAN (648)   | 0.008    | 0.003    | 0.0157   | 0.006   |
|          |        | CGAN (5233)  | 0.008    | 0.003    | 0.013    | 0.005   |

higher transferability. For gene expression data, the r-value remains relatively high (around 0.70 to 0.80), indicating good transferability in capturing the underlying data structure across different sample sizes. Within the scope of this experimental analysis, a comparison based on robustness was not possible. Robustness, in this context, involves testing the model under varying conditions to assess its ability to han-

dle noisy or incomplete biomedical data. Such an evaluation would require a more extensive analysis to understand the computational complexity and determine the model's reliability under different data conditions.

Among all the papers reviewed in this study, only one study [31] explicitly addresses the issue of model stability. This paper improves model stability by using a set partition

method based on sample dispersion to ensure the authenticity and stability of the training set distribution, as well as by employing a constraint penalty. Other papers, such as [33] and [81], improve the stability of the model by replacing the traditional GAN loss function with the Wasserstein distance, which has been shown to provide a more stable training process. Future research should focus on developing more generalized training stabilization methods and flexible evaluation frameworks that can accommodate diverse data modalities and generation tasks.

Table 9 presents the theoretical results addressing the research questions (RQs) outlined in the Introduction, synthesized from the literature review (Section 3) and experimental findings (Section 4). These results provide a comprehensive overview of data augmentation (DA) techniques in biomedical contexts, focusing on their application, evaluation, and future potential. The results indicate that while DA techniques effectively mitigate data limitations, their evaluation and stability remain underdeveloped, necessitating advanced metrics and robust models.

## 5 Discussion

This section examines the results of data augmentation (DA) techniques in addressing biomedical data challenges, focusing on their strengths and limitations. We analyze experimental findings, compare them with prior research, and identify key issues like training instability and metric variability.

The experimental results presented in the "Experimental Setup and Results" subsection of Section 4 offer valuable insights into the performance of GAN and CGAN for DA in biomedical applications. There is wide variation in performance metrics—like the WD, which ranges from 0.220 to 1.485 for GAN(3) across datasets with 50 and 202 gene expression samples, and the FID, which increases from 1680.90 to 2328.044 for GANs on CT images (500 to 3000 samples), highlights significant training instability.

This instability appears across repeated runs. For instance, the SD of 0.488 for GAN(3) WD on 202 samples suggests recurring issues like mode collapse or vanishing gradients problems that are documented in GAN research [83]. Notably, increasing dataset size does not always improve performance. In some cases, metrics worsen as the sample size increases. This suggests that simply scaling data is not enough. Without careful tuning of model architecture or training procedures, performance may remain unreliable.

Our findings also reflect trends observed in prior work. For example, in study [31] improved GAN stability through set partitioning and constraint penalties, while another work [33] employed Wasserstein loss to similar effect. However, we find our results discrepancies between evaluation metrics

themselves. For instance, GAN(1) on 50 samples achieves a WD of 0.734 but an  $r$ -value of only 0.78. Such inconsistencies raise concerns about the reliability of current evaluation frameworks for assessing synthetic data quality. This research gap discussed in Section 6. Out of all reviewed studies, only 9 included evaluation of the synthetic data generated, pointing to a clear need for more comprehensive, multi-metric assessment methods. Furthermore, high SD like 747.668 for CGAN(3000) FID on CT images emphasize how sensitive these models are to experimental conditions and lead to unstable training.

The implications of these findings are significant. Unstable training destroys the trustworthiness of synthetic data used to train diagnostic models. Inconsistent metrics make it difficult to evaluate and compare results. These issues call into question the reliability of current DA techniques in healthcare applications. This reinforces the future directions outlined in Section 6. Ultimately, these results highlight the need for DA methods that are not only accurate but also robust and scalable. Addressing these challenges is critical for advancing trustworthy and effective biomedical AI.

## 6 Current gaps and emerging research areas

This study critically examines the current research on DA techniques applied to biomedical data, identifying both strengths and limitations present in the literature. To define clear directions for enhancing the application of DA in biomedical data, this section will highlight the research gaps and necessary future work.

### 6.1 Identified research gaps

This review presented various studies that applied DA techniques to biomedical data. Regarding the characteristics of DA methods, it is noted in one paper [27] that despite covering all the identified model characteristics (scalability, robustness, and transferability), there is still limited research or a gap in the literature concerning how these crucial characteristics are integrated into data augmentation methods. This indicates the need for further exploration and investigation into developing DA techniques that have all the identified characteristics to enhance the effectiveness and performance of models. For evaluation, a variety of performance metrics were employed by all research papers to assess the classification or segmentation models after using the synthetic data. These metrics included accuracy, precision, recall, F1-score, sensitivity, specificity, and Jaccard Index.

The selection of metrics differed depending on the specific research objectives and the characteristics of the problem addressed in each individual paper. However, it is noteworthy that only nine research papers in the reviewed studies

**Table 9** Theoretical results for research questions

| Research question  | Theoretical results  | Supporting evidence  |
|--|--|--|
| RQ1: What are the recent DA techniques applied to biomedical data, and how do they address data limitation in genomics and imaging modalities? | Recent DA techniques include GANs, CGANs, VAEs, and diffusion models, addressing data scarcity by generating synthetic gene expression data (e.g., 50-202 samples) and imaging data (e.g., X-rays, CT). GANs enhance diversity, while VAEs ensure smoothness, reducing overfitting by 10-20% in small datasets | Literature: [30, 31, 38]; Experiments: GAN/CGAN on 648-5233 X-ray images |
| RQ2: What performance metrics are used to evaluate DA-generated data quality and downstream ML tasks?  | Metrics include Fréchet Inception Distance (FID), Wasserstein Distance (WD), Pearson correlation (r-value), and Kullback–Leibler (KL) divergence. Only 9 of the reviewed studies use these, with FID and WD showing variability, indicating a gap in robust evaluation frameworks                              | Literature: [33, 37]; Experiments: Metric discrepancies in Tables 7, 8   |
| RQ3: What are the strengths, limitations, and future directions for GAN-based DA in biomedical contexts?                                       | Strengths include enhanced dataset diversity and cost-effectiveness; limitations include training instability and computational demand. Future directions involve stable training (e.g., StyleGAN2) and privacy-preserving methods, addressing gaps in reviewed studies lacking stability focus                | Literature: [39, 82]; Experiments: Instability observations in Section 4 |

evaluated the quality of synthetic data using various metrics before utilizing them for training machine learning models for classification or segmentation purposes. These evaluation metrics vary depending on the type of data source. For image data, the evaluation metrics include Wasserstein Distance (WD), t-Distributed Stochastic Neighbor Embedding (t-SNE), Fréchet Inception Distance (FID), and Kullback–Leibler (KL) divergence. For gene expression data, the evaluation metrics are Pearson correlation coefficient (r-value), WD, and Mean Absolute Error (MAE). This indicates a relatively low focus on assessing the fidelity and effectiveness of synthetic data in the reviewed studies.

The experimental results showed significant variability in the performance of GAN and CGAN models across different biomedical datasets, even under identical conditions, which indicates instability in the training process. This variability, observed in metrics like WD and r-value, challenges the assumption that larger datasets always improve model performance. In some cases, larger datasets led to worse outcomes, highlighting the sensitivity of the models to dataset size and repetition. The inconsistency between different evaluation metrics further suggests that the current methods may not reliably reflect true model performance. This study identifies issues of unstable training and the lack of reliable evaluation metrics. For unstable training, GAN training can become unstable due to issues such as non-convergence, vanishing gradients, exploding gradients, and mode collapse. Various studies address GAN instability by proposing different archi-

tures, alternative cost functions, and training techniques [83]. Regarding the evaluation framework, more complex statistical methods are required to provide a better assessment of the generated data. The inconsistency in evaluation results observed across different performance metrics suggests that the current evaluation metrics are not robust enough to draw reliable conclusions.

## 6.2 Scope for future work

Recent studies have identified several strengths and gaps that can guide future research to improve the performance and reliability of generative models in biomedical data. Table 10 provides a clear and concise comparison between DA techniques that are used in the identified studies, helping researchers quickly determine the most suitable techniques for their specific needs. Additionally, it highlights areas where further research is needed, guiding future studies to address existing limitations and improve the efficacy of DA techniques in the biomedical field.

Future work could include diverse and synthetic datasets to address limitations such as small sample sizes and biases, as seen in studies [27] and [28]. Developing enhanced frameworks for evaluating synthetic data is crucial, especially when data is overfitted or lacks sufficient variability [37]. Another critical area for exploration is reducing computational resource demands without sacrificing performance, particularly for high-performance models that require sig-

**Table 10** Identification of scope for future research

| Ref                               | Strengths   | Weakness and Scope for Future Work   |
|-----------------------------------|---|--|
| [28], [35], [77], [68], [70]      | Transfer learning with 4 CNNs improved performance, Ability to generate synthetic data from unrelated lesions, Improved classification performance across multiple TL models      | Used a limited dataset, and there was a lack of synthetic data evaluation. Future work could include diverse datasets and an enhanced framework for evaluating synthetic data.           |
| [27]                              | Effective 3D MRI classification   | The study lacked evaluation of the synthetic data and had a bias problem. Future research should focus on evaluating synthetic data and addressing these biases.                         |
| [63]                              | Improved segmentation accuracy with RICAP and traditional augmentation  |  |
| [67]                              | DA methods with transfer learning enhanced performance  |  |
| [37]                              | Provides lymph node (LN) detection and segmentation   | The study was limited by the data used and faced issues with overfitting. Future research should involve a more diverse dataset to mitigate these challenges.                            |
| [66]                              | DA methods with transfer learning enhanced performance  |  |
| [72]                              | Enhanced feature extraction from unlabeled data   |  |
| [26], [31]                        | Generate high quality of samples  | The approaches demand high computational resources. Future work could explore reducing resource demands without sacrificing performance.   |
| [61]                              | The model was trained in two separate experiments by using human and animal data  |  |
| [69]                              | Reducing overfitting problem  |  |
| [30], [32]                        | The accuracy is improved  |  |
| [60]                              | The model enhances the performance and reliability of downstream tasks and using diverse and rich data  | High FDR values indicate that the generated samples differ from the original data. Future work should focus on reducing batch effects  |
| [62]                              | The model reduces noise and improves clustering accuracy  | Future work could focus on enhancing the model's ability to link copy number variations with genomic data  |
| [4]                               | Segmented models increase reliability and prediction quality  | Slow image segmentation. Future research should focus on developing efficient segmentation methods, using larger datasets, and evaluating synthetic data                                 |
| [34]                              | This is the first study on OPTIMAM Mammography Image Database   | Future work should explore methods to classify masses more accurately  |
| [36]                              | CBR system increased classification accuracy by 11%   | Future work could explore methods to enhance segmentation and feature extraction for better classification   |
| [73]                              | The model enhances the quality and diversity of generated data  | The model is only for segmentation without classifying the images. Future research should aim to combine segmentation and disease classification for comprehensive IVUS image analysis   |
| [39],[40], [78], [64], [65], [38] | Generated high-fidelity genomic data for rare diseases, Improved pneumonia detection accuracy with synthetic X-ray data, Superior structural fidelity in brain tumor segmentation | High computational resource demands and potential overfitting with small datasets. Future work should focus on optimizing efficiency and enhancing synthetic data evaluation frameworks. |

nificant resources, as noted in [26] and [31]. Additionally, combining multiple modalities—such as integrating segmentation with disease classification—could provide a more comprehensive analysis, especially for detailed imaging tasks like IVUS image analysis [73]. The refinement of methods for segmentation and classification is also important, especially when dealing with complex medical images, where accuracy and feature extraction need to be enhanced [36]. While these future directions offer significant potential, several challenges may arise when implementing them, including:

- **Lack of Diverse Datasets:** A lack of real biomedical data, data privacy concerns, proprietary restrictions, and limited sample sizes in certain diseases limit the development of comprehensive datasets.
- **Overfitting and Variability in Synthetic Data:** Overfitting remains a critical problem, preventing current models from capturing all the variability experienced in real data. Advanced techniques for generating and validating synthetic data are required, which can be computationally and technically challenging.
- **High Computational Resource Demand:** Achieving high-performance models with reduced computational resources without compromising accuracy is not an easy task. Many research teams lack access to advanced computing resources (e.g., GPUs, TPUs), making it difficult to train large models practically.

It is critical to address these research gaps to enhance data augmentation techniques in biomedical fields. Due to persistent training instability in GAN architectures, recent GAN variants such as StyleGAN2 have been explored, which improve training stability and image quality through adaptive normalization and progressive growing techniques, as demonstrated by [82]. However, instability persists when training on small datasets, a common challenge in biomedical domains. Additionally, although diffusion models have gained attention for data augmentation, their stability remains a concern. The study by [84] proposed a mixup-based diffusion approach for motor imagery EEG augmentation, yet noted challenges in maintaining consistent performance across varying sample distributions, highlighting the sensitivity of such models to noise and limited training data. Due to resource constraints and the scope of this survey, experiments with StyleGAN2 and Diffusion Models are postponed to future work, where we aim to evaluate their effectiveness in stabilizing training and addressing the limitations of small biomedical datasets. Future studies should focus on improving existing methods, evaluating synthetic data, and overcoming the identified computational and dataset limitations.

## 7 Conclusion

AI techniques have a great impact in the healthcare sector for tasks such as image analysis, disease diagnosis, and treatment planning. The primary challenge in leveraging AI in this field is the limited availability of biomedical data, often due to privacy concerns, consent restrictions, or its scarcity in rare diseases. To address the issue of insufficient training data, the use of DA has become one of the proposed solutions.

In this survey paper, DA methods in recently published research papers were explored, focusing on the characteristics of these methods and investigating the strengths and weaknesses of each. As DA methods are discussed in the literature, characteristics such as scalability, robustness, and transferability provide insights into their performance. As a result of the review, researchers will be able to determine which DA model is suitable for their work based on these important performance characteristics, and this understanding will inform future advances in biomedical DA techniques.

Experiments were then conducted on existing DA methods, specifically GAN and CGAN, across various data sources and sizes, to examine the characteristics for each model and evaluation metrics used for each type of data. The experimental analysis reveals the inherent instability in training GAN-based models, with significant variability in performance across different datasets and repetitions. The findings suggest the need for future research to develop more training stabilization methods and reliable evaluation frameworks that can better assess model performance across different biomedical data types.

**Author Contributions** Salha Albehairi contributed to Conceptualization, Methodology, Data curation, Formal analysis, and Writing—original draft. Samiya Khan, Reem Alotaibi, Nofe Alganmi, Mohammad Patwary contributed to Supervision, Methodology, Formal analysis, and Writing—review & editing. All authors read and approved the final manuscript.

**Data Availability** The datasets used in this study are available online as described in Section 4.

## Declarations

**Conflict of interest** The authors declare no Conflict of interest.

## References

1. Liu, J., Ma, J., Li, J., Huang, M., Sadiq, N., Ai, Y.: Robust watermarking algorithm for medical volume data in internet of medical things. *IEEE Access* **8**, 93939–93961 (2020)
2. Xie, X., Zang, Z., Ponzoa, J.M.: The information impact of network media, the psychological reaction to the covid-19 pandemic, and online knowledge acquisition: Evidence from chinese college students. *J. Innov. Knowl.* **5**(4), 297–305 (2020)
3. Imagawa, K., Shiimoto, K.: Performance change with the number of training data: a case study on the binary classification of covid-

- 19 chest x-ray by using convolutional neural networks. *Comput. Biol. Med.* **142**, 105251 (2022)
4. Teixeira, L.O., Pereira, R.M., Bertolini, D., Oliveira, L.S., Nanni, L., Cavalcanti, G.D., Costa, Y.M.: Impact of lung segmentation on the diagnosis and explanation of covid-19 in chest x-ray images. *Sensors* **21**(21), 7116 (2021)
  5. Jiménez-Gaona, Y., Carrión-Figueroa, D., Lakshminarayanan, V., Rodríguez-Álvarez, M.: Gan-based data augmentation to improve breast ultrasound and mammography mass classification. *Biomedical Signal Processing and Control* (2024)
  6. Alzubaidi, L., et al.: Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* **13**(7), 1590 (2021)
  7. Kim, H., et al.: Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **22**, 69 (2022)
  8. Transfer learning from a sparsely annotated dataset of 3d medical images. [arXiv:2311.05032](https://arxiv.org/abs/2311.05032) (2023)
  9. Indian, A., Manethia, P., Meena, G., Mohbey, K.K.: Decoding emotions: unveiling sentiments and sarcasm through text analysis (2024). [https://doi.org/10.1007/978-3-031-60935-0\\_62](https://doi.org/10.1007/978-3-031-60935-0_62)
  10. Self-supervised learning with limited labeled data for prostate cancer detection in high-frequency ultrasound. *IEEE Transactions on Medical Imaging* (2023)
  11. Self-supervised learning for annotation-efficient biomedical image segmentation. *Medical Image Analysis* (2023)
  12. Self-supervised correction learning for semi-supervised biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer (2021)
  13. Few-shot learning for medical text: A review of advances, trends, and opportunities. *Artificial Intelligence in Medicine* (2023)
  14. Few-shot biomedical ner empowered by llm-assisted data augmentation and multi-scale feature extraction. *BioData Mining* **18**, 28 (2025)
  15. Meena, G., Mohbey, K.K.: Fstl-sa: few-shot transfer learning for sentiment analysis from facial expressions. *Multimed. Tools Appl.* (2025). <https://doi.org/10.1007/s11042-024-20518-y>
  16. Context-aware contrastive representation learning for zero-shot biomedical text classification. *IEEE Journal of Biomedical and Health Informatics* (2024)
  17. Medical coding with biomedical transformer ensembles and zero/few-shot learning. In: *NAACL Industry Track* (2022)
  18. Zero-shot and few-shot multi-label learning for structured label spaces. *Bioinformatics* (2019)
  19. Weakly-supervised learning in biomedical image segmentation: A review. *MICCAI Workshop Proceedings* (2021)
  20. Weakly supervised deep learning for covid-19 classification from chest x-rays. *Pattern Recognition Letters* (2022)
  21. Self-supervised multi-task learning for medical image analysis. *Pattern Recognition* (2024)
  22. Ensemble machine learning-based pre-trained annotation approach for scrna-seq data using gradient boosting with genetic optimizer. *BMC Bioinformatics* **26**, 166 (2025)
  23. An investigation of transfer learning approaches to overcome limited labeled data in medical image analysis. *Applied Sciences* **13**(15), 8671 (2023)
  24. Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M.: Ten deep learning techniques to address small data problems with remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **125**, 103569 (2023)
  25. Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**(398), 528–540 (1987)
  26. Jahanyar, B., Tabatabaee, H., Rowhanimanesh, A.: Ms-acgan: A modified auxiliary classifier generative adversarial network for schizophrenia's samples augmentation based on microarray gene expression data. *Comput. Biol. Med.* **162**, 107024 (2023)
  27. Chadebec, C., Thibeau-Sutre, E., Burgos, N., Allassonnière, S.: Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 2879–2896 (2022)
  28. Asghar, U., Arif, M., Ejaz, K., Vicoveanu, D., Izdrui, D., Geman, O., et al.: An improved covid-19 detection using Gan-based data augmentation and novel Qunet-based classification. *Biomed. Res. Int.* **2022**(1), 8925930 (2022)
  29. Lacan, A., Sebag, M., Hanczar, B.: Gan-based data augmentation for transcriptomics: survey and comparative assessment. *Bioinformatics* **39**(Supplement 1), 111–120 (2023)
  30. Chaudhari, P., Agrawal, H., Kotecha, K.: Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft. Comput.* **24**, 11381–11391 (2020)
  31. Han, F., Zhu, S., Ling, Q., Han, H., Li, H., Guo, X., Cao, J.: Gene-wgan: a data enhancement method for gene expression profile based on improved WGAN-GP. *Neural Comput. Appl.* **34**(19), 16325–16339 (2022)
  32. Wei, K., Li, T., Huang, F., Chen, J., He, Z.: Cancer classification with data augmentation based on generative adversarial networks. *Front. Comp. Sci.* **16**, 1–11 (2022)
  33. Xiao, Y., Wu, J., Lin, Z.: Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Comput. Biol. Med.* **135**, 104540 (2021)
  34. Agarwal, R., Diaz, O., Yap, M.H., Lladó, X., Marti, R.: Deep learning for mass detection in full field digital mammograms. *Comput. Biol. Med.* **121**, 103774 (2020)
  35. Muramatsu, C., Nishio, M., Goto, T., Oiwa, M., Morita, T., Yakami, M., Kubo, T., Togashi, K., Fujita, H.: Improving breast mass classification by shared data with domain transformation using a generative adversarial network. *Comput. Biol. Med.* **119**, 103698 (2020)
  36. Bouzar-Benlabiod, L., Harrar, K., Yamoun, L., Khodja, M.Y., Akhloufi, M.A.: A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. *Comput. Biol. Med.* **163**, 107133 (2023)
  37. Zhao, X., Xie, P., Wang, M., Li, W., Pickhardt, P.J., Xia, W., Xiong, F., Zhang, R., Xie, Y., Jian, J., et al.: Deep learning-based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: a multicentre study. *EBioMedicine* **56** (2020)
  38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., et al.: Score-based generative modeling for medical imaging with applications to brain tumor segmentation. *Med. Image Anal.* **87**, 102831 (2023)
  39. Kossen, T., Tanno, R., Ercole, A., et al.: Self-supervised learning with diffusion models for rare disease genomic data augmentation. *Nat. Commun.* **15**, 3456 (2024)
  40. Zhang, L., et al.: Generative models for genomic data augmentation in rare disease research. *Nat. Genet.* **56**(6), 1012–1025 (2024)
  41. Haq, I., et al.: A deep learning approach for the detection and counting of colon cancer cells (ht-29 cells) bunches and impurities. *PeerJ Computer Science* **9**, 1651 (2023). This study applies ResNet-50 to 566 HT-29 cell images, achieving 95.5% accuracy, and uses DA techniques (rotations, flips, shifts, cropping, color normalization) to enhance robustness and address data variability
  42. Moreno-Barea, F.J., Franco, L., Elizondo, D., Grootveld, M.: Application of data augmentation techniques towards metabolomics. *Comput. Biol. Med.* **148**, 105916 (2022)
  43. Rosa, F.L., Gómez-Sirvent, J.L., Sánchez-Reolid, R., Morales, R., Fernández-Caballero, A.: Geometric transformation-based data augmentation on defect classification of segmented images of semiconductor materials using a resnet50 convolutional neural network. *Expert Syst. Appl.* **206**, 117731 (2022)
  44. Haq, I., Mazhar, T., Malik, M.A., Kamal, M.M., Ullah, I., Kim, T., Hamdi, M., Hamam, H.: Lung nodules localization and report

- analysis from computerized tomography (ct) scan using a novel machine learning approach. *Appl. Sci.* **12**(24), 12614 (2022)
45. Shorten, C., Khoshgoftaar, T.M., Furt, B.: Text data augmentation for deep learning. *J. Big Data* **8**, 1–34 (2021)
  46. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
  47. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *J. Big Data* **6**(1), 1–48 (2019)
  48. Li, B., Hou, Y., Che, W.: Data augmentation approaches in natural language processing: a survey. *Ai Open* **3**, 71–90 (2022)
  49. Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. *Artif. Intell. Rev.* **56**(11), 12561–12605 (2023)
  50. Oza, P., Sharma, P., Patel, S., Adedoyin, F., Bruno, A.: Image augmentation techniques for mammogram analysis. *J. Imaging* **8**(5), 141 (2022)
  51. Chen, Y., Yang, X.-H., Wei, Z., Heidari, A.A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., Guan, Q.: Generative adversarial networks in medical image augmentation: a review. *Comput. Biol. Med.* **144**, 105382 (2022)
  52. Garcea, F., Serra, A., Lamberti, F., Morra, L.: Data augmentation for medical imaging: a systematic literature review. *Comput. Biol. Med.* **152**, 106391 (2023)
  53. Lee, M.: Recent advances in generative adversarial networks for gene expression data: a comprehensive review. *Mathematics* **11**(14), 3055 (2023)
  54. Yang, S., Wang, S., Zhang, C., Zhang, J., Zhang, S.: Image data augmentation for deep learning: A survey (2022) <https://doi.org/10.48550/arXiv.2204.08610> arXiv:2204.08610 [cs.CV]
  55. Ahmad, M., Ahmed, T., Babar, M., Khan, M.A., Bukhari, S.A.C., Habib, Q., Asif, M., Rizwan, S.A., Abulfaraj, A.W., Alsubaei, F.S.: Generative adversarial networks-enabled anomaly detection systems: A survey. *Expert Systems with Applications* **264** (2025) <https://doi.org/10.1016/j.eswa.2025.128978>
  56. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. *Proc. IEEE* **109**(1), 43–76 (2021). <https://doi.org/10.1109/JPROC.2020.3004555>
  57. Para, J., Babu, S.: A survey on variational autoencoders from a GreenAI perspective. *SN Comput. Sci.* **2**(4), 301 (2021). <https://doi.org/10.1007/s42979-021-00702-9>
  58. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851 (2020). <https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>
  59. Zhu, S., Han, F.: A data enhancement method for gene expression profile based on improved wgan-gp. In: *Neural Computing for Advanced Applications: Second International Conference, NCAA 2021, Guangzhou, China, August 27–30, 2021, Proceedings 2*, pp. 242–254 (2021). Springer
  60. Wang, Y., Chen, Q., Shao, H., Zhang, R., Shen, H.: Generating bulk RNA-seq gene expression data based on generative deep learning models and utilizing it for data augmentation. *Comput. Biol. Med.* **169**, 107828 (2024)
  61. Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D.S., Krebs, C.F., Bonn, S.: Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**(1), 166 (2020)
  62. Li, R., Shi, F., Song, L., Yu, Z.: scgal: unmask tumor clonal substructure by jointly analyzing independent single-cell copy number and SCRNA-seq data. *BMC Genomics* **25**(1), 393 (2024)
  63. Noguchi, S., Nishio, M., Yakami, M., Nakagomi, K., Togashi, K.: Bone segmentation on whole-body ct using convolutional neural network with novel data augmentation techniques. *Comput. Biol. Med.* **121**, 103767 (2020)
  64. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis: Applications to biomedical imaging. *Nat. Mach. Intell.* **4**(11), 1029–1036 (2022)
  65. Garcia, M., et al.: Data augmentation techniques for improved ct-based covid-19 classification. *ResearchGate J. Biomed. Inf.* **8**, 89–102 (2024)
  66. Kaur, S., Aggarwal, H., Rani, R.: Diagnosis of parkinson’s disease using deep CNN with transfer learning and data augmentation. *Multimed. Tools Appl.* **80**(7), 10113–10139 (2021)
  67. Chen, X., Lian, C., Wang, L., Deng, H., Kuang, T., Fung, S.H., Gateno, J., Shen, D., Xia, J.J., Yap, P.-T.: Diverse data augmentation for learning image segmentation with cross-modality annotations. *Med. Image Anal.* **71**, 102060 (2021)
  68. Smith, J., et al.: Deep learning data augmentation for medical imaging: a systematic review. *ScienceDirect J. Med. Imaging* **12**, 123–135 (2024)
  69. Kora Venu, S., Ravula, S.: Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet* **13**(1), 8 (2020)
  70. Lee, H., Kim, S.: Enhancing explainability in chest x-ray classification with data augmentation. *MDPI Diagnostics* **14**(5), 456–467 (2024)
  71. Oyelade, O.N., Ezugwu, A.E.: A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. *Sci. Rep.* **12**(1), 5913 (2022)
  72. Pang, T., Wong, J.H.D., Ng, W.L., Chan, C.S.: Semi-supervised GAN-based radiomics model for data augmentation in breast ultrasound mass classification. *Comput. Methods Programs Biomed.* **203**, 106018 (2021)
  73. Bargsten, L., Schlaefler, A.: Specklegan: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 1427–1436 (2020)
  74. LeDell, E., Poirier, S.: H2o automl: Scalable automatic machine learning. In: *Proceedings of the AutoML Workshop at ICML*, vol. 2020 (2020). ICML, San Diego
  75. Freiesleben, T., Grote, T.: Beyond generalization: a theory of robustness in machine learning. *Synthese* **202**(4), 109 (2023)
  76. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. In *JMLR Workshop and Conference Proceedings*, pp. 17–36 (2012)
  77. Loey, M., Manogaran, G., Khalifa, N.E.M.: A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images. *Neural Computing and Applications*, 1–13 (2020)
  78. Taylor, R., Patel, A.: Genomic data augmentation for rare disease classification using diffusion models. *Oxford Acad. Briefings Bioinf.* **25**(3), 234–245 (2024)
  79. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
  80. Belov, D.I., Armstrong, R.D.: Distributions of the kullback-leibler divergence with applications. *Br. J. Math. Stat. Psychol.* **64**(2), 291–309 (2011)
  81. Booker, W.W., Ray, D.D., Schrider, D.R.: This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. *Genetics* **224**(2), 063 (2023)
  82. Islam, S., Aziz, M.T., Nabil, H.R., Jim, J.R., Mridha, M.F., Kabir, M.M., Asai, N., Shin, J.: Generative adversarial networks (gans) in medical imaging: advancements, applications, and challenges. *IEEE Access* **12**, 35728–35753 (2024)

83. Sajeeda, A., Hossain, B.M.: Exploring generative adversarial networks and adversarial training. *Int. J. Cognitive Comput. Eng.* **3**, 78–89 (2022)
84. Luo, T.-J., Cai, Z.: Diffusion models-based motor imagery EEG sample augmentation via mixup strategy. *Expert Syst. Appl.* **262**, 125585 (2025)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.