Saihood *et al.*
BMC Medical Informatics and Decision Making (2024) 24:403
https://doi.org/10.1186/s12911-024-02820-9

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Fusion-driven semi-supervised learning-based lung nodules classification with dual-discriminator and dual-generator generative adversarial network

Ahmed Saihood[1], Wijdan Rashid Abdulhussien[1], Laith Alzubaid[2,3,4*], Mohamed Manoufali[5] and Yuantong Gu[2]

Abstract

Background The detection and classification of lung nodules are crucial in medical imaging, as they significantly impact patient outcomes related to lung cancer diagnosis and treatment. However, existing models often suffer from mode collapse and poor generalizability, as they fail to capture the complete diversity of the data distribution. This study addresses these challenges by proposing a novel generative adversarial network (GAN) architecture tailored for semi-supervised lung nodule classification.

Methods The proposed DDDG-GAN model consists of dual generators and discriminators. Each generator specializes in benign or malignant nodules, generating diverse, high-fidelity synthetic images for each class. This dual-generator setup prevents mode collapse. The dual-discriminator framework enhances the model's generalization capability, ensuring better performance on unseen data. Feature fusion techniques are incorporated to refine the model's discriminatory power between benign and malignant nodules. The model is evaluated in two scenarios: (1) training and testing on the LIDC-IDRI dataset and (2) training on LIDC-IDRI, testing on the unseen LUNA16 dataset and the unseen LUNGx dataset.

Results In Scenario 1, the DDDG-GAN achieved an accuracy of 92.56%, a precision of 90.12%, a recall of 95.87%, and an F1 score of 92.77%. In Scenario 2, the model demonstrated robust performance with an accuracy of 72.6%, a precision of 72.3%, a recall of 73.82%, and an F1 score of 73.39% when testing using Luna16 and an accuracy of 71.23%, a precision of 67.56%, a recall of 73.52%, and an F1 score of 70.42% when testing using LungX. The results indicate that the proposed model outperforms state-of-the-art semi-supervised learning approaches.

Conclusions The DDDG-GAN model mitigates mode collapse and improves generalizability in lung nodule classification. It demonstrates superior performance on both the LIDC-IDRI and the unseen LUNA16 and LungX datasets, offering significant potential for improving diagnostic accuracy in clinical practice.

Keywords Generative Adversarial Network, Lung Nodules Classification, Semi-supervised learning, Deep learning, Medical imaging

*Correspondence:
Laith Alzubaid
l.alzubaidi@qut.edu.au
Full list of author information is available at the end of the article

Introduction

Generative Adversarial Networks (GANs) are a breakthrough and landmark at the forefront of modern advancements in the neural network, wherein data mirroring specific distributions are synthesized precisely through the exercise [1]. From photorealistic image synthesis to excellent and subtle domains, such as medical image analysis, where GAN is characterized by its core architecture and competitive synergy between a generator and a discriminator, GANs have been applied broadly. The medical industry terms GANs as a disruptive technology that is out to the problem of annotation scarcity coupled with medical imagery's complex, high-dimensional nature.

The fusion of deep learning with semi-supervised learning paradigms harnesses the potent capabilities of deep neural networks to extract intricate features and learn representations directly from raw data [2, 3]. Deep learning models are data-hungry, and despite being trained with unlabeled data, they still require substantial computational resources to train and the presence of large-scale labelled datasets. However, labels are challenging to obtain.

Recent advancements in semi-supervised learning for medical image classification have significantly improved medical image analysis, particularly in detecting and classifying lung nodules. Applying semi-supervised learning methods in this field leverages the limited availability of labelled data alongside a larger pool of unlabeled data to enhance model performance.

Self-training is one of the semi-supervised models used to train on the labelled data first, then used to predict labels for the unlabeled data [4, 5]. The predictions deemed most confident are added to the training set, and the process is repeated. However, Incorrect labels can be reinforced, leading to degradation in model performance. Moreover, the accuracy of adequately labelled data depends on the initial model trained on labelled data.

Co-training [6, 7] involves training two separate models on different data views (i.e., different feature sets). Each model labels the unlabeled data used to re-train the other model. It is not easy to get two independent views to merge the decision of the final prorate labels. Hence, graph-based methods are used for semi-supervised learning [8–11], which uses graphs to represent the data, where nodes represent samples and edges represent similarities between samples. Labels are propagated from labelled to unlabeled nodes based on the graph structure. However, the performance in this method is highly dependent on the graph's quality and structure; constructing and processing large graphs can be computationally expensive.

Also, pseudo-labelling semi-supervised learning involves assigning pseudo-labels to the unlabeled data using the current model and then re-training the model using these pseudo-labels [12–14]. However, the accuracy of this model can be sensitive to the threshold used for selecting pseudo-labeled data. Ensemble semi-supervised learning is also used to classify medical images [15, 16]. It combines predictions from multiple models to generate more robust labels for the unlabeled data. However, training multiple models is resource-intensive.

One notable method, FocalMix [17], represents a pioneering approach to leveraging semi-supervised learning for 3D medical image detection. The study showed that semi-supervised learning methods could achieve substantial improvements. Loyman et al. [18] employ a two-step approach that includes automatic annotation of partially labelled datasets and learning a semantic similarity metric space based on the predicted annotations. These methods bolster models' capability to differentiate between nodules, which could enhance the precision and dependability of lung nodule classification systems. However, their performance may be hindered by a class imbalance or a scarcity of labelled data.

To address the class imbalance in semi-supervised learning, Chen et al. [19] aim to enhance the utilization of unlabeled data by adjusting the consistency loss to better match the class distribution of augmented unlabeled data with the original unlabeled data. Another approach explored using a semi-supervised Generative Adversarial Network (SSGAN) [20] for medical image classification, such as lung X-ray classification, with minimal labelled samples. The study extended this model with pseudo-labelling, proposing a novel GAN model (PLAB-GAN) that utilizes unlabeled data for sample distribution estimation and direct classifier training. Li et al. [21] proposed a semi-supervised learning method through graph-embedded random forests. By embedding labelled and unlabeled data into a graph and assuming all data form a manifold, this method aims to mine label information from unlabeled data to compute more accurate information gain. However, if the unlabeled data set is not sufficiently representative of the full spectrum of the problem space or contains biases, these methods might perform differently than expected.

A semi-supervised multi-task learning framework for lung cancer diagnosis was explored [22], integrating segmentation and classification tasks. This method begins with training on a labelled dataset and iteratively refines the model by predicting and incorporating labels for the unlabeled data. This model simultaneously addresses the challenge of false positive reduction and nodule segmentation, leveraging morphology information like size, volume, and shape. This methodology exhibits the potential

for enhanced representational capacity but encounters challenges in equitably distributing the learning emphasis across concurrent tasks.

Yan et al. [23] address the scarce labelled datasets in medical imaging by combining a multi-discriminator GAN with an encoder. Nevertheless, employing multiple discriminators can confer advantages in discerning subtle distinctions; concurrently, generating distinct classes is imperative to facilitate this discrimination effectively.

Therefore, to mitigate these challenges, this paper proposes a novel architecture called dual-discriminator and dual-generator GAN (DDDG-GAN), designed particularly for the considered semi-supervised classification problem of lung nodules. This novel design incorporates dual Generators and Discriminators, each carefully trained to sample and assess its results over the labelled data containing benign and malignant lung nodules. Such a bifurcated strategy would nuance each class of nodules against the different attributes of the other, overcoming the limitation of previously presented models in differentiating very small but diagnostically crucial features. The DDDG-GAN model is built around three critical components: the DDDG-GAN architecture, a feature fusion mechanism, and a classification level. This explains the architecture of a design around training the dual generators and discriminators; hence, the overall arch doubles the number of generators, making it more effective in synthesizing class-specific, hyper-realistic images. Feature fusion details the process of the fusion of discriminative feature maps being generated from the DDDG-GAN into an attentive feature map that encapsulates important features indicative of the potential malignancy of the nodule. Finally, the classification level explains the fused features through convolutional and fully connected layers to precisely classify the nodules into benign and malignant categories.

The issues addressed in this article are: 1) By employing dual generators, DDDG-GAN mitigates the issue of mode collapse, where traditional GANs tend to generate a limited variety of outputs. This ensures the production of a more diverse set of synthetic images, enhancing model robustness. 2) The architecture's dual generators allow for generating high-fidelity, class-specific images. This is crucial in medical image analysis for accurately representing distinct categories, such as benign or malignant nodules. 3) Including dual discriminators, each specializing in a particular class, improves the model's ability to generalize from training data to unseen data, enhancing diagnostic accuracy. 4) The model enhances the capability to distinguish subtle differences between classes (e.g., benign vs. malignant), a critical requirement in medical diagnosis and many other applications where fine-grained classification is essential.

The main contributions of this article are:

- It proposes dual generators to accurately capture the inherent differences between benign and malignant pulmonary nodules. The dual generators, each trained on different datasets, learn and replicate the distinguishing features and characteristics of each type of nodule. This architecture confirms the generation of diverse, high-fidelity, class-specific synthetic lung nodule images, an essential requirement for enhancing model robustness and diagnostic accuracy.
- Proposed a new dual discriminator that inputs volumetric data and generates a two-dimensional feature map. The approach encapsulates the most critical discriminative features to distinguish between benign and malignant nodules. The proposed model improves the capability to differentiate fine-grained morphological contrasts between benign and malignant nodules.
- It incorporates dual generators and discriminators through feature map fusion, wherein each is carefully trained to sample and appraise their outputs using labeled data of benign and malignant lung nodules. This will ensure that diagnostically significant attributes are effectively employed to enhance the classification of lung nodules.
- The proposed model, in effect, is subjected to extensive experiences on two datasets alternately, both in terms of effectiveness and robustness. As for generalization capacity, the model is too competent in different datasets, which leads to reliable segmentation and classification of pulmonary nodules.

Related works

Semi-supervised learning (SSL) has been increasingly applied in medical image analysis to overcome the challenge of limited labelled data. Various methods, including self-training, co-training, graph-based approaches, and generative models, have been explored to enhance segmentation and classification accuracy.

Cai et al. [4] and Dzieniszewska et al. [5] introduced the self-training on CT scans and dermoscopy images, respectively. The teacher model is used to generate pseudo-labels for unlabeled data. Improvement was shown by Cai et al. in DSC, IoU, Precision, and Recall for CT segmentation. Dzieniszewska et al. presented very high skin lesion segmentation mIoU. Nevertheless, all of such models strongly depend on the initial good quality of the teacher model. Otherwise, it may lead to error induction and propagation in learning. However, these models may not be generalized well in other datasets.

Xie et al. [24] introduced a small-paced self-training method, which assumes that the distributions of labelled and unlabeled data can be aligned. This method improved Accuracy, Precision, Recall, and F1 Score for X-ray image classification. Despite its potential, the assumption of distribution alignment may only hold in some practical scenarios, limiting its effectiveness.

Yang et al. [7] and Tang et al. [6] applied co-training methods to MRI and CT scans, respectively. These approaches leverage multiple views of the data to improve learning. Yang et al. demonstrated DSC and Hausdorff Distance (HD) enhancements for MRI segmentation. At the same time, Tang et al. improved Accuracy, Precision, Recall, and F1 Score for multi-label protein–protein interaction prediction. However, co-training methods face computational complexity, reliance on initial model quality, and scalability issues for larger datasets.

Graph-based methods, explored by Sun et al. [11] and Miller et al. [10], utilize relationships between data points to enhance learning. Sun et al. applied these techniques to RGB images, demonstrating improvements in Accuracy, Precision, Recall, F1 Score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). Miller et al. applied graph-based learning to Synthetic Aperture Radar (SAR) images, achieving similar improvements. However, these methods are computationally intensive and sensitive to constructing similarity graphs, which can limit their scalability. Zha et al. [9] extended graph-based methods to multi-label CT and MRI data, highlighting challenges with label imbalance and data diversity.

Su et al. [12] and Li et al. [13] explored pseudo-labeling methods for 3D gadolinium-enhanced MR imaging scans and pancreas CT images, respectively. These methods showed significant improvements in the Dice Similarity Coefficient (DSC). However, the quality of initial pseudo-labels is crucial, and generating and refining these labels can be computationally expensive.

Ensemble methods, discussed by Li et al. [16] and Kallipolitis et al. [25], combine multiple models to improve performance. Li et al. applied this approach to colon and laryngoscopy CT images, achieving high accuracy, precision, recall, and F1 scores. Kallipolitis et al. used ensemble methods for histopathology images, demonstrating similar improvements. However, these methods can overfit the training data if not adequately regularized and require high-quality annotated datasets, which are costly and time-consuming.

Generative methods, such as GANs, generate new data samples to enhance learning. Hardy et al. [26] and Toizumi et al. [27] applied these techniques to RGB and satellite images, respectively. These powerful methods face node synchronization and scalability challenges in distributed settings. Generative methods improved accuracy, precision, recall, F1 Score, and Specificity for RGB images and improvements in Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) for satellite images.

Li et al. [28] integrated GANs with a pyramid attention mechanism and transfer learning to improve segmentation accuracy for CT scans. This approach leverages pre-trained models to enhance performance even with limited labelled data, showing improvements in Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). However, the success of this approach depends heavily on the quality and relevance of the pre-trained models used in transfer learning.

Several techniques have been suggested to handle challenges in medical image analysis. For example, SimTrip [29] includes a self-supervised learning framework leveraging triplet manifestation to extract expressive features using extended views of data, effectively handling computational limitations and small batch sizes. However, it focuses primarily on feature representation without managing data generation, a crucial requirement in medical imaging for augmenting datasets. Another method named LCGANT [30] incorporates a GAN-based generator and a transfer learning classifier (VGG-DF) to manage overfitting and perform high classification accuracy. While practical, its dependence on a separate classifier raises additional complexity.

In contrast, our DDDG-GAN unifies the generative and discriminative strategies within a single architecture, simplifying training and improving generalization. Furthermore, weakly supervised learning methods, as reviewed in [31] emphasize practical solutions for managing restricted labeled data via insufficient or noisy labels. Despite their utility, these methods can suffer from decreased generalization capability. Our proposed method surpasses these approaches by incorporating dual generators for class-specific data augmentation, dual discriminators for precise feature extraction, and a feature fusion mechanism to integrate critical attributes, achieving robust classification with minimal labeled data.

Table 1 summarizes semi-supervised learning methods in medical image analysis. These methods offer significant promise in addressing the challenge of limited labelled data. Each method has strengths and weaknesses, with common challenges including computational complexity, dependency on initial model quality, and scalability. Future research should address these challenges to enhance the robustness and applicability of SSL methods, broadening their impact on various medical imaging tasks.

The integration of GANs and diffusion models into medical imaging research denotes a paradigm shift with the possibility of addressing longstanding challenges in

Table 1 Summary of semi-supervised learning methods in medical image analysis

Method	Ref	Image modality	Metrics	Issues
Self-Training semi-supervised learning	[4], 2023	CT scan	Dice Similarity Coefficient (DSC), Intersection over Union(IoU), Precision, Recall	The model 's effectiveness heavily relies on the initial teacher model 's quality
	[5], 2024	skin lesion images were obtained using dermoscopy	mIoU	The generalizability of the model to other datasets is lacking
	[24], 2023	X-ray	Accuracy, Precision, Recall, F1 Score	The method assumes that the distribution of labelled and unlabeled data can be aligned using small-paced self-training
Co-Training semi-supervised learning	[7], 2024	MRI	DSC, HD	Potential computational complexity due to the co-training process, reliance on the initial quality of the models, and scalability issues for larger datasets
	[6], 2024	CT scans	Accuracy, Precision, Recall, F1 Score	The quality and diversity of training data significantly impact the model's performance. Sparse or noisy data can affect prediction accuracy
Graph-Based Semi-supervised learning	[11], 2023	RGB	Accuracy, Precision, Recall, F1 Score, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)	The spectral decomposition process can be computationally intensive, which might limit the method 's scalability for enormous datasets
	[10], 2024	Synthetic Aperture Radar (SAR) images	accuracy, precision, recall, F1 score, and Specificity	The method 's accuracy is sensitive to how well the similarity graph represents the underlying data structure
	[9], 2009	CT and MRI	accuracy, precision, recall, F1 score, and Specificity	Multi-label datasets often suffer from imbalanced label distributions, which can affect the performance of the graph-based semi-supervised learning methods
Pseudo-Labeling semi-supervised learning	[12], 2024	3D gadolinium-enhanced MR imaging scans (GE-MRIs	Dice Similarity Coefficient (Dice)	The effectiveness of the method relies on the accuracy of the initial pseudo-labels
	[13], 2023	Pancreas-CT	Dice Similarity Coefficient (Dice)	The process of generating and refining soft pseudo-labels can be computationally expensive
Ensemble semi-supervised learning	[16], 2023	colonoscopy and laryngoscopy CT	accuracy, precision, recall, F1 score, and Specificity	Ensembles with many models can still overfit the training data if not adequately regularized or if the base models are too complex
	[25], 2021	Colon-CT	accuracy, precision, recall, F1 score, and Specificity	The model 's effectiveness relies on the availability of high-quality annotated datasets, which can be time-consuming and costly to produce

Table 1 (continued)

Method	Ref	Image modality	Metrics	Issues
Generative semi-supervised learning	[26], 2019	RGB	accuracy, precision, recall, F1 score, and Specificity	Ensuring synchronization between the central server and distributed nodes can be challenging, especially as the number of nodes increases
	[27], 2019	satellite images	Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR)	The effectiveness of the model on different types of cloud formations and satellite data
	[28], 2024	CT scans	Dice Similarity Coefficient (DSC), Intersection over Union (IoU)	The effectiveness of transfer learning depends on the quality and relevance of the pre-trained models used

the domain. GANs have already shown remarkable power in generating synthetic medical images that near copyist accurate data distributions, delivering valuable resources for training machine learning models in techniques where labeled datasets are lacking. Based on these hits, diffusion models are appearing as a groundbreaking process for generating high-quality synthetic medical images with improved commitment and diversity. For instance, studies like [32] and [33] underscore their ability to create realistic images even with limited data availability.

These improvements are incredibly transformative in medical imaging, where the investment and annotation of large, various datasets are resource-intensive and often impracticable. Diffusion models leverage their ingrained noise-denoising tools to create synthetic images that not only increase existing datasets but also enhance model training robustness. By generating high-quality, manifold, and representative datasets, these models pave the way for more accurate diagnostics, improved disease detection, and enhanced generalizability of machine learning systems in clinical applications. Integrating these techniques into research frameworks holds significant promise for overcoming data scarcity, minimizing annotation burdens, and accelerating progress in developing reliable and scalable medical imaging solutions.

Proposed method: DDDG-GAN

In this paper, we incorporate two generators and two discriminators trained to generate samples of two lung nodule classes (i.e., benign and malignant) from labelled actual data samples through a semi-supervised learning model. The proposed model consists of three parts explained in three subsections: the first is DDDG-GAN, in which the two generators and two discriminators are explained in detail and how they will be trained using their different classes. The second is feature fusion, in which the fusion process of the maps produced by the DDDG-GAN is explained, and how the obtained attentive vector is fed to the classification level. The third is the classification level, in which we define the mapping of the fused features at the feature fusion level through the convolutional layer and fully connected layer to be classified into benign and malignant.

Figure 1 presents the proposed model, which features two generators (G1 and G2) and two discriminators (D1 and D2), mainly devised to tackle mode defeat and enhance class-specific generalization. G1 and G2 are independently trained to generate high-fidelity synthetic images of benign and malignant lung nodules. This separation allows each generator to concentrate on the unique attributes of its assigned class, such as smooth edges for benign nodules or irregular, spiculated patterns for malignant nodules. D1 and D2 act as evaluators, distinguishing between real and synthetic images while emphasizing critical features unique to their respective classes. The outputs of the discriminators are incorporated through a feature fusion mechanism, incorporating discriminatory information into an attentive feature map. This map improves the classification procedure by encapsulating subtle elements. The architecture is conceived to balance synthetic image quality, class separability, and robust classification performance, addressing critical challenges in medical image analysis.

The proposed method comprises three components: 1) the dual generators to generate samples of two lung nodule classes, 2) discriminators to extract discriminative feature maps being generated, 3) Feature maps fusion aims to create a composite representation that magnifies the differential characteristics recognized by each discriminator and 4) classification level fuse features through convolutional and fully connected layers to classify the nodules into benign and malignant categories precisely.

Generator network architecture

Given nodules $X \in \mathcal{R}^{H \times W \times D}$, the first generator network (G_1) was trained using labelled benign nodules in the dataset to map from a latent space (a predefined noise distribution) to the distribution of nodule $\widehat{X}_B$. At the same time, the second generator network (G_2) was trained using the sample of the labelled malignant nodules to map from a latent space to the distribution of nodules $\widehat{X}_M$. The latent space (Z) is a multivariate Gaussian distribution where each dimension is independent and identically distributed (i.i.d). For both generators, the process starts from a latent vector, progressively upscales it through deconvolutional layers and finally outputs a 3D tensor that represents $\widehat{X} \in \mathcal{R}^{H \times W \times D}$ for both malignant and begin.

G1 and G2 share an identical network architecture, as shown in Fig. 2, to ensure that any differences in the generated images are solely due to the nature of the data they were trained rather than model complexity or capability variations. The input to the generator is a latent vector of dimension L, which is densely connected to a higher-dimensional space. This initial transformation maps the latent space to a format conducive to 3D volume generation. This layer is followed by batch normalization and a LeakyReLU activation to introduce non-linearity and stabilize the training process.

Conv3DTranspose layers are used to up-sample the volume to the desired dimensions. G1 is trained only on a dataset of benign pulmonary nodules. This dataset covers a wide range of benign nodule images. This is done so that G1 learns to generate varied representations covering both benign nodules' standard and rare features. Conversely, G2

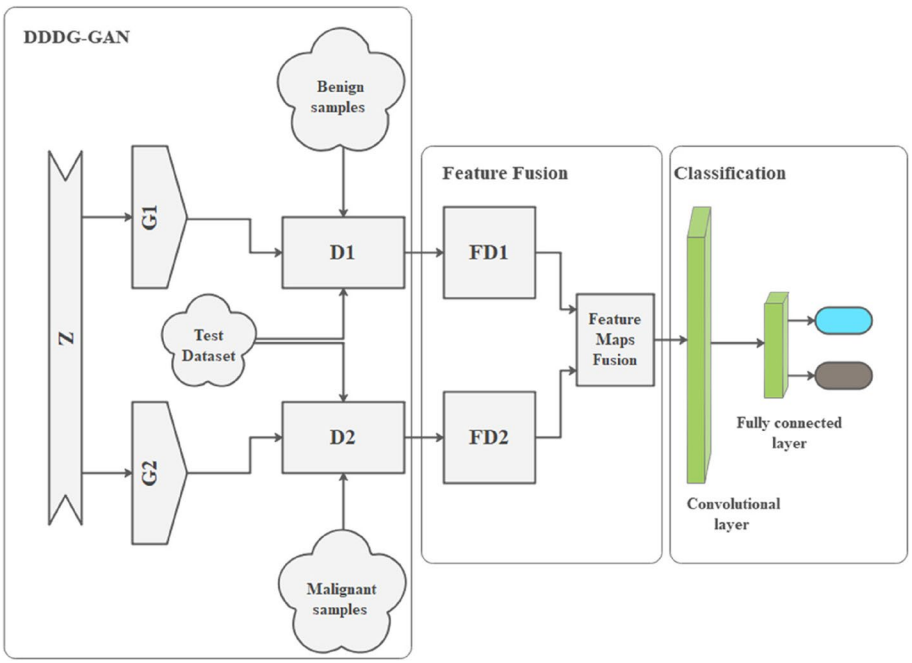

Fig. 1 Dual-Discriminator and Dual-Generator Generative Adversarial Network (DDDG-GAN) architecture. The model incorporates two generators (G1 and G2) and two discriminators (D1 and D2)

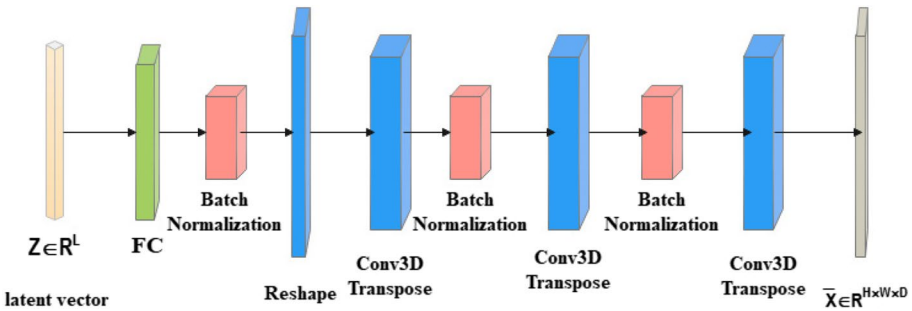

Fig. 2 Generator Network Architecture. The generator network is designed to map a latent vector sampled from a multivariate Gaussian distribution to a 3D representation of lung nodules

is trained on a dataset of malignant pulmonary nodules, capturing the heterogeneity and, therefore, the complexity of malignant formations. Thus, G2 can generate images that accurately reflect all the features of malignancies in the pulmonary nodules. The idea behind using two generators is to develop synthetic 3D nodule images that are not distinguishable from accurate scans and, therefore, constitute a valuable resource for enhanced diagnostic accuracy. This approach with two generators allows for comparing the features between benign and malignant nodules and informs about their differences in morphology and texture.

$$\widehat{X}_B = G_1(Z, \theta_1^i)$$

where $\widehat{X}_B$ indicate the generated 3D benign nodules, G_1 indicates the generator network layers that map Z to $\widehat{X}_B$., and θ_1^i refers to the parameters of each layer (i) learned through the training process in G_1.

In the same context, the second generator maps the input latent space to $\widehat{X}_M$, which indicates the generation of 3D malignant nodules.

$$\widehat{X}_M = G_2(Z, \theta_2^i)$$

The primary motivation behind training two separate generator networks (G1 for benign nodules and G2 for malignant nodules) with the same architecture but on different datasets lies in the intrinsic differences between benign and malignant pulmonary nodules.

These differences, though subtle in some instances, are crucial for accurate diagnosis, treatment planning, and prognosis. By training G1 and G2 on distinct datasets, each generator learns to encapsulate and reproduce the unique characteristics and features inherent to each type of nodule.

Discriminator architecture

The discriminators in GANs focus on distinguishing between natural and synthetic medical images, particularly for extracting discriminative features of benign and malignant nodules, which are critical in refining the generative models' capabilities. The discriminators, D1 and D2, are tasked with analyzing the outputs of G1 and G2, respectively, along with actual images of benign and malignant nodules. Their goal is to not only discriminate between real and generated images but also to highlight the distinctive features characteristic of benign and malignant nodules in a format conducive to further analysis or decision-making processes.

A typical binary classification (real vs. fake) discriminator must be adapted to highlight discriminative features. This adaptation focuses on feature extraction layers that retain spatial information, ultimately leading to an output layer redesigned to produce a 2D feature map rather than a single scalar indicating real or fake.

To elucidate the discriminative capabilities between benign and malignant pulmonary nodules, we designed a discriminator network employing a sequential model architecture, leveraging three-dimensional convolutional layers to process input volumes of size $H \times W \times D$. This architecture facilitates the extraction and analysis of spatial features inherent to the nodules, which are crucial for their classification and understanding.

The discriminator begins with a three-dimensional convolutional layer, as shown in Fig. 3, configured with 64 filters of kernel size 5×5x3 and a stride of 2×2x1, applied with 'same' padding to maintain the spatial dimensions of the input volume. This layer is designed to perform an initial feature extraction, capturing the nodules' local and global spatial features. Activation of the convolutional layer outputs is mediated by a LeakyReLU function with an alpha value of 0.2, introducing non-linearity while allowing for a slight gradient when the unit is inactive to mitigate the vanishing gradient problem. A dropout layer follows, with a dropout rate of 0.25, to prevent overfitting by randomly omitting a subset of features during training.

After the initial feature extraction, a second three-dimensional convolutional layer, equipped with 128 filters and a kernel size of 5×5x3, further refines the feature map. This layer employs strides of 2×2x2 and 'same' padding, augmented with zero padding to adjust the spatial dimensions as needed, ensuring consistent feature representation across the volume. Batch normalization is applied post-convolution to stabilize learning by normalizing the layer's inputs with a momentum of 0.8, facilitating faster convergence and improved generalization. This is followed by another application of the LeakyReLU activation function and a dropout layer, reinforcing the network's ability to learn complex, robust features resistant to overfitting.

Crucially, the network transitions from analyzing volumetric data to producing a two-dimensional feature map that encapsulates discriminative features critical for distinguishing between benign and malignant nodules. This is achieved by flattening the three-dimensional feature maps into a vector and passing them through a densely connected layer configured to reshape the output into

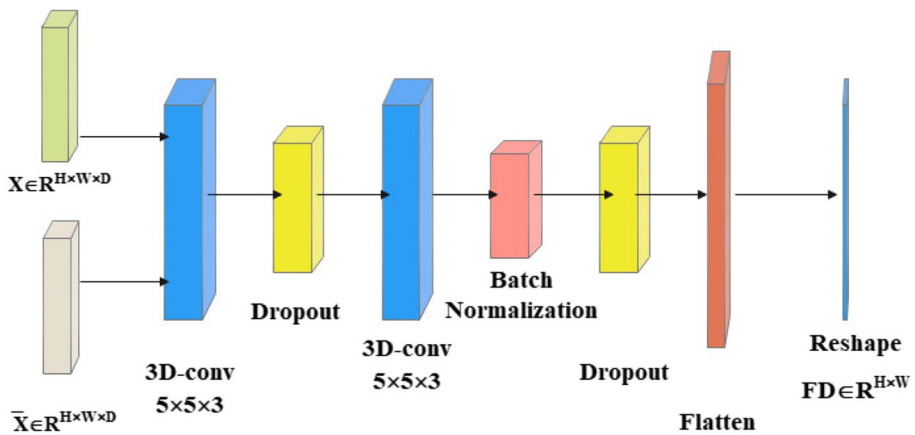

Fig. 3 Discriminator Network Architecture. The discriminator network is designed to distinguish between real and synthetic images while extracting class-specific features critical for classification

FD1 and FD2, both in the shape of a HxW (i.e., 2D feature map). This allows the discriminator to classify inputs as real or synthetic and highlight spatial features indicative of the nodule's pathology, providing a visual representation of the discriminative characteristics learned through the network.

Feature maps fusion

The fusion of feature maps from FD1 and FD2 for the same input aims to create a composite representation that magnifies the differential characteristics recognized by each discriminator. This composite map improves nodule classification by highlighting the distinct features indicative of benignity or malignancy.

Before the fusion process, we normalize the feature maps obtained from FD1 and FD2 to ensure they are on a similar scale using Z-score standardization. The absolute difference between the normalized feature maps of FD1 and FD2. This difference map will highlight regions where the discriminators disagree, indicating areas of interest that might differentiate between benign and malignant nodules more effectively. After the fusion, the model is trained to be classified based on the fused feature map. This convolutional layer is followed by a fully connected layer, culminating in a softmax output for binary classification. D1 is trained to focus on benign nodules and D2 on malignant nodules, and both output feature maps highlight characteristics of the input nodules. An innovative approach is needed for the fusion and analysis of these feature maps.

We use a cross-entropy loss function for the fused feature map and the subsequent classification layer, which is standard for binary classification tasks. We incorporate contrastive loss with cross-entropy loss function to encourage the model to generate fused feature maps that are distinctly far apart for benign and malignant nodules, enhancing separability. The total loss L_{total} is formulated as a combination of cross-entropy loss L_{CE} and contrastive loss $L_{Contrastive}$:

$$L_{total} = \alpha L_{CE}(\mathrm{y}, \widehat{\mathrm{y}}) + \beta L_{Contrastive}(F_{D1}, F_{D2})$$

where y is the true label (benign or malignant), $\widehat{\mathrm{y}}$ is the predicted label based on the fused feature map, $F_{D1} and F_{D2}$ are the feature maps from FD1 and FD2, respectively, and α and β are weights that balance the contributions of each loss component. This allows for effectively using both discriminators' outputs, leveraging their specialized focus to improve nodule classification. Through this approach, the GAN model can generate more informative and distinctive representations of pulmonary nodules, facilitating better understanding and identification of their benign or malignant nature. Semi-supervised learning seeks to leverage a small amount of labelled data and a large pool of unlabeled data to improve learning efficacy. The dual loss function approach fits well into such contexts for enhancing feature discrimination.

Loss functions

The loss function plays a crucial role in guiding the training process of both the generator and discriminator models. When dealing with two separate GANs, G1 and G2, with their respective discriminators D1 and D2, each trained on different data distributions (benign and malignant nodules, respectively) to achieve the desired objectives for both generation and discrimination tasks. Formulating a loss function that measures the discrepancy between the outputs of D1 and D2 can provide insights into the distinguishability of benign and malignant nodules as learned by the GANs.

For generators, the loss is formulated to measure how well it deceives the discriminator into believing that the generated images are real. If we consider G1 generating benign nodules and G2 generating malignant nodules, the generator loss for each GAN can be defined using the binary cross-entropy (BCE) as follows:

$$L_G = -\frac{1}{N}\sum_{i=1}^{N} log(D(G(z_i)))$$

where L_G is the generator loss, N is the number of samples, G represents the generator G1 or G2, D represents the discriminator (D1 for G1's output and D2 for G2's output), and z_i is the input latent vector to the generator. The goal of G is to minimize this loss to generate images that D will classify as real.

The discriminator loss combines the error on real and generated (fake) images, aiming to classify both types correctly. The loss for a discriminator D can be defined as:

$$L_D = \frac{1}{N}\sum_{i=1}^{N}[log(D(x_i)) + log(1 - D(G(z_i)))]$$

where L_D is the discriminator loss and x_i represents real images from the dataset (benign for D1 and malignant for D2). The discriminator seeks to minimize this loss to accurately distinguish between real and generated images.

Experimental results

In this study, we introduced two scenarios to evaluate the proposed method, considering two types of datasets for assessment. The LIDC-IDRI and LUNA16 datasets are used here. As following:

- Scenario 1: The labeled nodules in the LIDC-IDRI dataset are evaluated using tenfold cross-validation for training and validation, with an additional independent test set comprising 20% of the dataset.
- Scenario 2: The entire labelled nodules in the LIDC-IDRI dataset for training and 500 nodules from the LUNA16 dataset for testing as an unseen dataset. Moreover, 73 nodules in total, with 37 benign and 36 malignant nodules from the LUNGx dataset, are used as an unseen dataset to confirm the assessment of the model's generalization ability.

Datasets

The LUNA16

The LUNA16 dataset contains 549,714 benign candidate nodules and 1,351 malignant candidate nodules. It is highly imbalanced towards benign nodules. Thus, this article used 500 nodules to test the proposed model. We randomly selected 250 benign and 250 malignant nodules. We trained our model using the LIDC dataset and then tested it using LUNA16.

The LIDC-IDRI

The LIDC-IDRI dataset comprises 1,018 instances, each consisting of a CT scan and an XML file detailing the nodule outlines identified by four radiology specialists. The LIDC-IDRI dataset does not directly label nodules as benign or malignant; instead, it includes a malignancy rating on a scale from 1 to 5 based on the assessments of up to four experienced thoracic radiologists. The malignancy score is as follows: improbable to be malignant, unlikely to be malignant, indeterminate/uncertain, likely to be malignant, and highly likely to be malignant. We consider nodules with average scores below a certain threshold (e.g., less than 3) benign and those above it (e.g., greater than 3) malignant. For this study, we selected 3246 nodules randomly divided as follows: 1054 malignant nodules, 987 benign nodules, and 1205 uncertain nodules.

Results

The evaluation of the proposed model across two distinct scenarios reveals insightful differences in performance metrics. As illustrated in Table 2 and Table 3, the model indicated impressive results in the first scenario, where tenfold cross-validation was conducted on the LIDC-IDRI dataset, and further independent testing was performed using 20% of the dataset as a held-out test set. The accuracy was 92.56%, with a precision of 90.12%, a recall of 95.87%, and an F1 score of 92.77%. These metrics indicate that the model effectively distinguishes between benign and malignant nodules within this dataset. High recall especially expresses a strong capability of the model to identify malignant nodules where accurate and reliable diagnostics should be guaranteed.

The other, on the other hand, tested generalization capability by training the model on the entire LIDC-IDRI data set and testing it on the unseen dataset, LUNA16; in this case, the metrics for model testing dropped significantly: accuracy- 72.6%, precision- 72.3%, recall- 73.82%, and F1 score- 73.39%. This reduction in performance indicates that while the model performs well on familiar data, it struggles with unseen data, suggesting potential overfitting during training.

In Scenario 2, the model trained on the full LIDC-IDRI dataset was tested on the independent LungX dataset, reaching an accuracy of 71.23%**,** a precision of 67.56%, a recall of 73.52%**,** and an F1-score of 70.42%**.** These outcomes emphasize the model's capability to generalize to unseen data, with balanced performance across precision and recall. The LungX dataset's diversity and real-world variability posed additional challenges. However, the model presented strong classification capabilities, emphasizing its potential for clinical applicability in recognizing and classifying lung nodules across varied patient populations and imaging conditions.

Figure 4 shows the proposed model's confusion matrix in two cases. Case 1 It can be seen from the confusion matrix that the proposed model performs exceedingly well on the test set formed from the LIDC-IDRI dataset. It has 201 true positives and 185 true negatives, indicating that the model performs very well in accurately identifying malignant and benign nodules. The 23 false positives and nine false negatives show that

Table 2 The results obtained over two scenarios

Dataset	Testing	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LIDC-IDRI (10-Fold Cross-Validation)	Scenario 1 (Validation Average)	92.56	90.12	95.87	92.77
LIDC-IDRI (20% Testing Set)	Scenario 1 (Independent Testing)	93.3	91.36	95.71	93.48
LIDC-IDRI (100% Training)	Scenario 2 (Test on LUNA16)	72.6	72.3	73.82	73.39
LIDC-IDRI (100% Training)	Scenario 2 (Test on LUNGx)	71.23	67.56	73.52	70.42

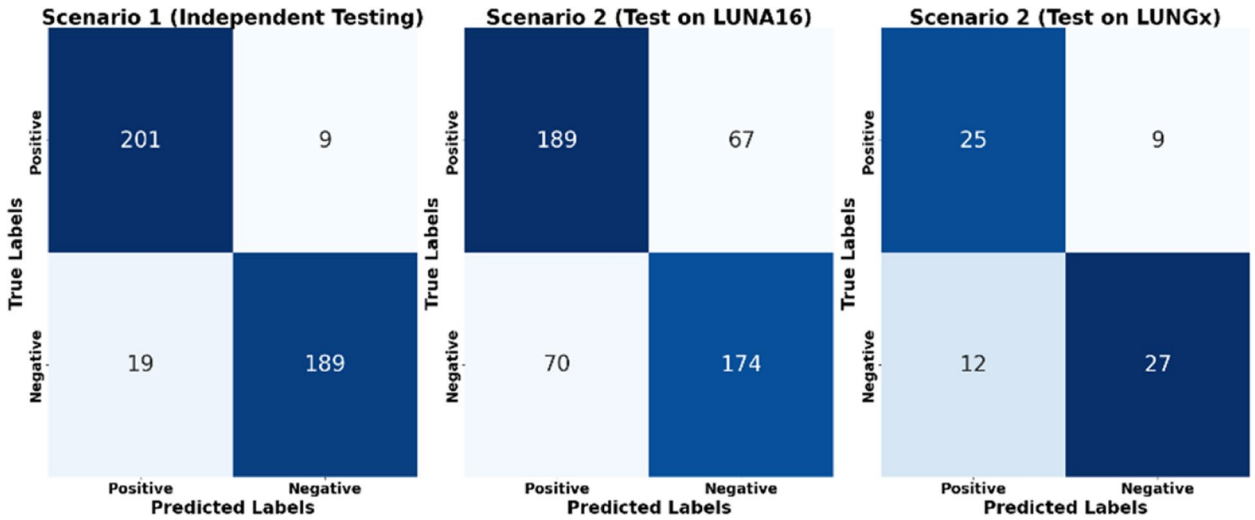

Fig. 4 Confusion matrix for the proposed model in two scenarios

this model has a balanced and safe performance with the least risk of misclassification. This concludes that this model is well-trained on LIDC-IDRI datasets and efficiently distinguishes between benign and malignant nodules.

On the other hand, in the second case, when the LIDC-IDRI dataset is considered for the model's training purpose, LUNA16 lies in the testing category. The confusion matrix generates somewhat agile but less performing results. It creates many false positive results, with 81, and false negative ones, with 68. This means this unseen data makes it harder for the user model to handle or process it. Although this model has rightly classified the majority of malignant nodules as true positives—188 and the majority of benign nodules as true negatives—163, these high misclassification rates raise concerns in terms of generalization. More false positives might lead to more unnecessary treatments of benign cases, and a higher false negative proportion could lead to missing cases of malignancies, both of which are significant causes for clinical concern regarding diagnostics.

Inception score

The Inception Score (IS) [34] is a metric used to evaluate the quality of images generated by the proposed model. It provides a quantitative measure of how the generated images are reasonable based on two key aspects: their clarity and diversity. The Inception Score uses a pre-trained Inception model trained using RGB lung nodules. The Good-generated images should contain easily recognizable lung nodules. For each image, the Inception model should output a high probability for one class and a low probability for others. The set of all generated images should cover various nodule types. The Inception Score is calculated using these probabilities to compute the KL divergence between the conditional label distribution for each image and the marginal label distribution over the entire set of generated images. The formula for IS is:

$$IS(G) = exp(\mathrm{E}_{x-p_g}[D_{KL}(p(y\backslash x) \parallel p(y))])$$

G is the proposed model, p_g is the model's distribution over generated images, D_{KL} is the KL divergence, and $(p(y\backslash x) and p(y)$ are the conditional and marginal label distributions, respectively.

We have generated 15 images using the proposed GAN, where the images are meant to represent either benign or malignant nodules. The high confidence levels (90% for benign and 95% for malignant) suggest that the images are clear and distinct, which is suitable for the first component of the IS (Table 3). Producing both benign and malignant nodules with the distribution of 60% malignant and 40% benign) means there is diversity in the generated images, aligning with the second component of the IS.

Table 3 Inception Score-based DDDG-GAN evaluation

Image No	Predicted Class	Confidence	Notes
1–6	Benign	90%	High confidence
7–15	Malignant	95%	High confidence

Interpretability visualization

SHAP (SHapley Additive exPlanations) [31] values and Grad-CAM (Gradient-weighted Class Activation Mapping) [30] visualizations offer profound insights into the decision-making process of the DDDG-GAN model for

lung nodule classification. SHAP values provide a quantitative breakdown of how each feature influences the model's predictions, highlighting which attributes push the classification towards being benign or malignant. This detailed feature contribution analysis enhances the interpretability of the model, revealing key features that significantly impact its decisions. On the other hand, Grad-CAM visualizations generate heatmaps overlaid on the original images, illustrating the regions of the input that the model focuses on when making its predictions. These visual cues help validate whether the model is attending to diagnostically relevant areas, ensuring that its decisions are based on appropriate visual information.

In the context of Grad-CAM visualizations, as shown in Fig. 5, these images demonstrate the model's ability to focus on diagnostically significant features. Heat maps for benign nodules correctly classified as benign highlight smooth and well-defined areas associated with benign characteristics. This indicates that the model captures and utilizes benign-specific features, validating the dual generator's role in creating high-fidelity benign nodule images.

For malignant nodules correctly classified as such, Grad-CAM heatmaps highlight irregular, spiculated regions and heterogeneous textures characteristic of malignancies. This demonstrates that the model has learned to accurately represent and recognize the subtle features of malignant nodules, an essential aspect of medical diagnosis. The dual-discriminator setup further enhances the model's ability to distinguish fine-grained differences in features, as reflected through the accurate highlighting of malignant features in the heatmaps. In this design, individual discriminators specialize in a single class.

In cases where the model classifies an unknown nodule as benign, the Grad-CAM visualizations (Fig. 5) provide insights into what drove the model's decision. Highlighted regions in such cases exemplify areas that the model identifies as benign features. If the classification is correct, it demonstrates the model's generalization and robustness.

In the DDDG-GAN model, addressing mode collapse is crucial for generating diverse and accurate representations of benign and malignant lung nodules. The model incorporates dual generators to mitigate this issue. Each generator generates high-fidelity images that capture the distinct features of benign or malignant nodules, thereby ensuring a broad data distribution coverage. This approach helps in producing diverse synthesized images that reflect the various characteristics seen in real-world medical images. Additionally, dual discriminators, each focusing on a specific class, further enhance the model's ability to generalize from the training data to unseen data, ensuring robust performance. The effectiveness of this architecture is supported by Grad-CAM visualizations, which demonstrate that the model accurately identifies and focuses on diagnostically relevant features across different categories of lung nodules.

The SHAP result breaks down the contributions of individual features to the model's prediction. Positive SHAP values indicate features pushing the classification towards malignancy, while negative values suggest benignity (Fig. 5). The magnitude of these values shows each feature's importance, providing a quantitative measure of their influence. SHAP analysis reveals that certain features were underemphasized, leading to incorrect classifications and offering a deeper understanding of the model's decision-making process. It highlights which features need adjustment or reweighting, ensuring the decision is based on logical and medically relevant factors.

The combined use of Grad-CAM and SHAP provides a comprehensive understanding of the model's decision-making process. Grad-CAM visualizes the spatial focus, confirming that the model's attention aligns with clinically significant regions. SHAP quantifies the importance of each feature, enhancing interpretability by explaining the model's decision regarding feature contributions. Together, they validate the robustness and interpretability of the DDDG-GAN model in classifying lung nodules. Correct classifications are supported by Grad-CAM, which focuses on relevant regions and SHAP and demonstrates reliance on critical features. Misclassifications reveal areas where the model needs improvement, guided by Grad-CAM insights and SHAP feature importance.

However, despite the strengths of the DDDG-GAN model, misclassifications do occur, as indicated by the combined SHAP and Grad-CAM analyses in Fig. 6. When a benign nodule is misclassified as malignant, the Grad-CAM heatmap may show that the model erroneously focused on irregular textures or shapes associated with malignancy, leading to a false positive. Conversely, when a malignant nodule is misclassified as benign, the heatmap might reveal insufficient attention to spiculated patterns or other malignant features, resulting in a false negative. These misclassifications underscore the importance of continuous model refinement by incorporating more diverse training samples and adjusting the feature selection process to minimize errors. The combined use of SHAP values and Grad-CAM visualizations is instrumental in identifying these areas for improvement, ensuring that the model's performance is robust, accurate, and reliable in medical diagnostics.

The combined strength of Grad-CAM and SHAP proves the model's power. Grad-CAM shows the model's focus on relevant regions, providing spatial context to the decision-making process, while SHAP offers a detailed

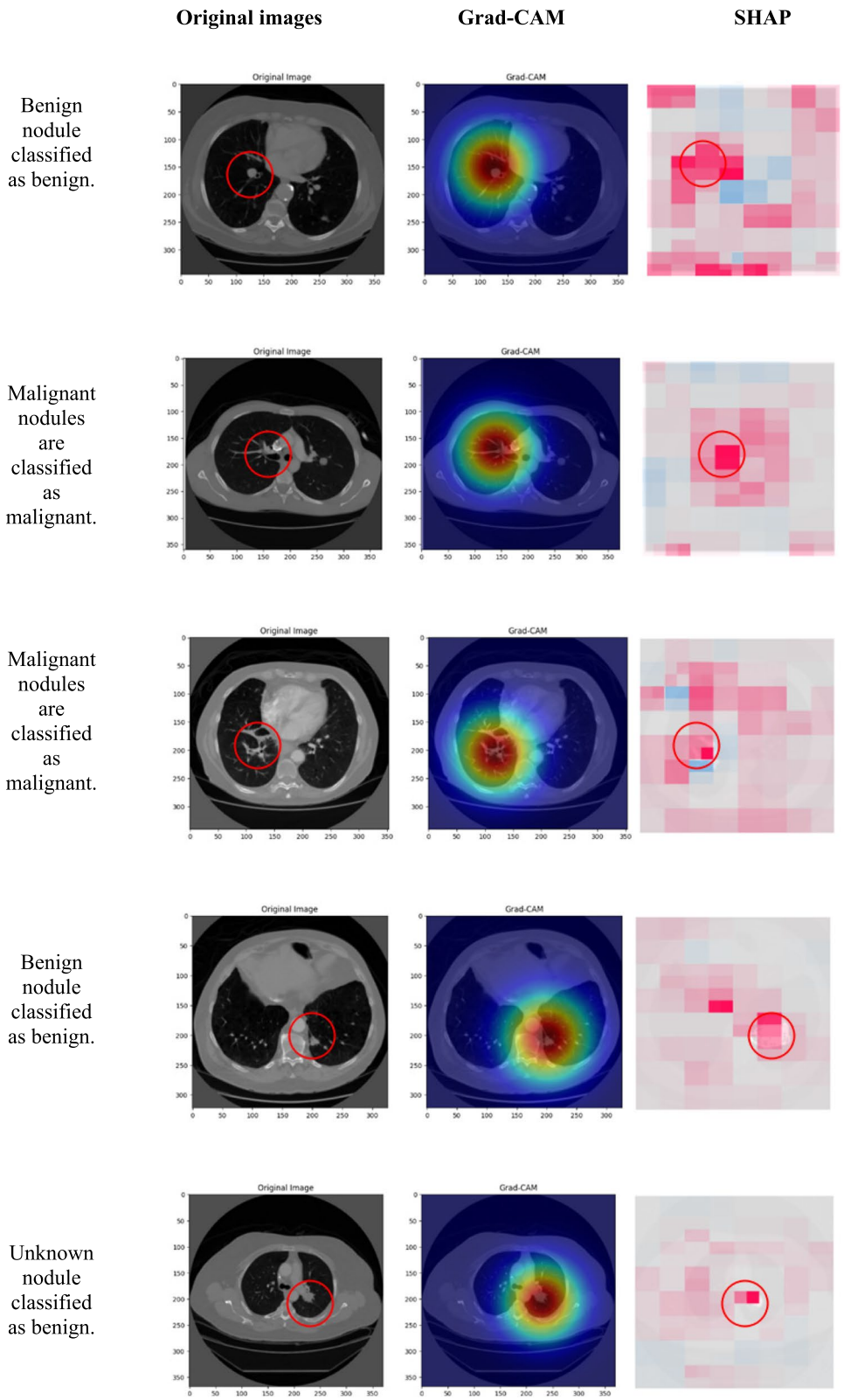

Fig. 5 Grad-CAM and SHAP visualizations of true classified nodules

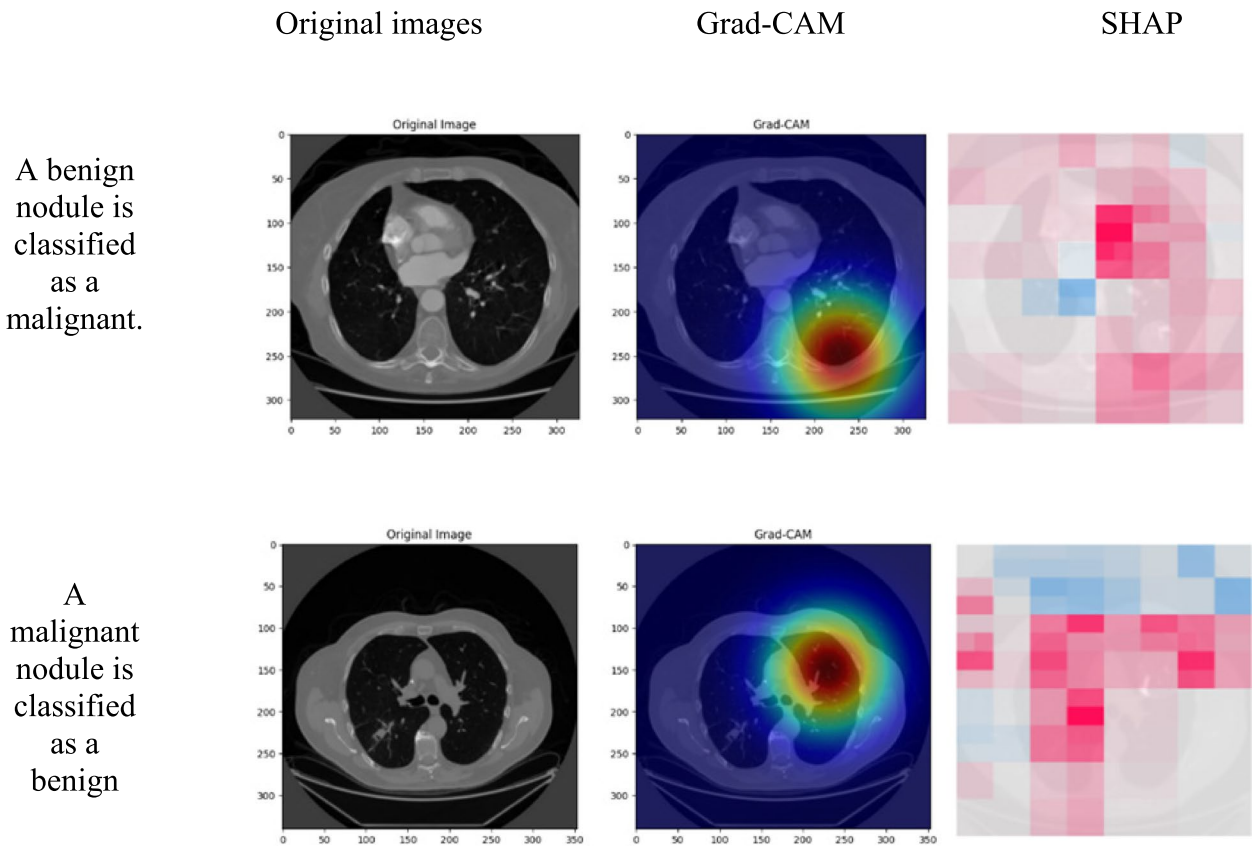

Fig. 6 Grad-CAM and SHAP visualizations of misclassified nodules

numerical breakdown of feature importance. This dual approach ensures accurate classification, builds trust in the model's predictions, and identifies specific areas for refinement. The DDDG-GAN model's effectiveness is evident in its ability to correctly classify nodules, with visual validation from Grad-CAM and feature importance from SHAP. Misclassifications highlight the need for feature adjustment and additional training data to capture critical characteristics better. This comprehensive analysis demonstrates the model's strengths and areas for improvement, ensuring its robustness and accuracy in medical diagnostics.

The strong interpretability of Grad-CAM and SHAP proves the model's power. Grad-CAM shows the model's focus on relevant regions, providing spatial context to the decision-making process, while SHAP offers a detailed numerical breakdown of feature importance. This dual approach ensures accurate classification, builds trust in the model's predictions, and identifies specific areas for refinement. The DDDG-GAN model's effectiveness is evident in its ability to correctly classify nodules, with visual validation from Grad-CAM and feature importance from SHAP. Misclassifications highlight the need for feature adjustment and additional training data to capture critical characteristics better. This comprehensive analysis demonstrates the model's strengths and areas for improvement, ensuring its robustness and accuracy in medical diagnostics.

To validate the rate and faithfulness of the generated lung nodule images, we show an illustrated comparison between real and synthetic examples in Fig. 7. The generated images nearly resemble their real counterparts, capturing vital morphological features such as texture, edge smoothness, and structural patterns. This underscores the capability of the proposed DDDG-GAN model to synthesize high-resolution, class-specific nodule images that are indistinguishable from real data. These outcomes highlight the model's significance in generating diverse, realistic examples, which are required to enhance diagnostic accuracy and the robustness of downstream classification tasks.

CNNs, which comprise the backbone of the proposed DDDG-GAN, intrinsically capture spatial features with isotropic filters that tend to highlight central regions precisely when the input data is pre-processed to center the areas of diagnostic interest, such as lung nodules. This

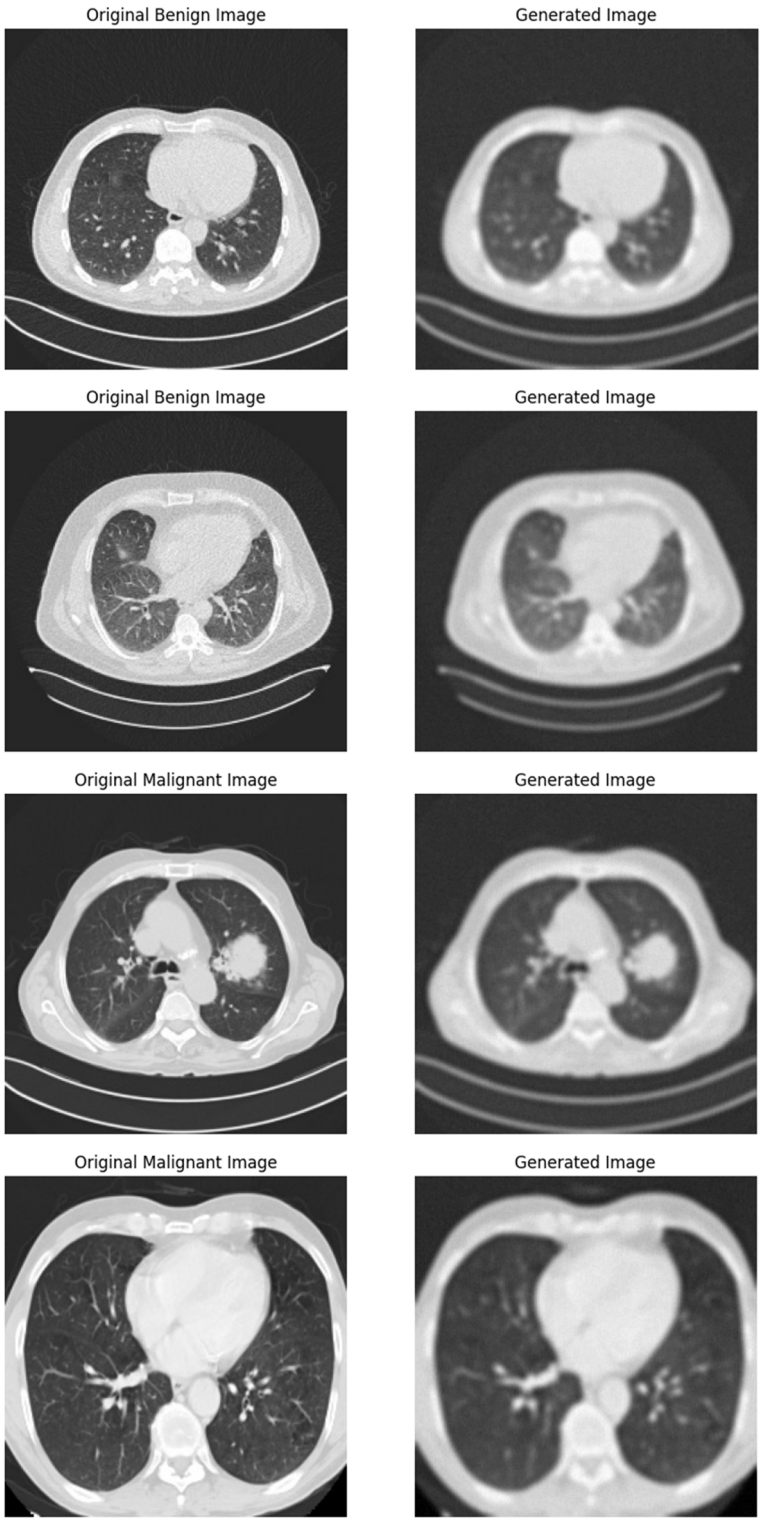

Fig. 7 Explanation of the reconstruction operation. The left column displays the input images, which are encoded into a latent vector via the encoder. The decoder then reconstructs these images from the latent vector, indicating the model's ability to keep and recreate key features of the original input

uniformity results in heatmaps with circular patterns, mirroring the model's invariant attention to the central features, where most nodules are located in the dataset. Moreover, the circular focus aligns with the imaging structure and annotations, as nodules are typically separated and centered during pre-processing to assure standardization.

Comparison with the state-of-the-art

The table compares various semi-supervised learning methods for binary classification of interstitial lung disease patterns from the LIDC dataset.

Among self-training methods, the highest performance is observed with a recall of 87.1, precision of 86.5, and accuracy of 85.8, indicating adequate selective re-training. Co-training methods show varied effectiveness, with the best results demonstrating recall of 87, precision of 88, and accuracy of 86, underscoring robust performance when leveraging different data views. This variation highlights how different data feature sets can impact the success of co-training.

Graph-based training methods stand out for their strong performance, particularly with one method achieving recall of 90.2, precision of 90.5, and accuracy of 89.5. This highlights the advantage of utilizing the intrinsic structure of the data to improve segmentation results. The ability of graph-based methods to model complex relationships within the data contributes significantly to their high performance.

Pseudo-labeling and ensemble learning methods show moderate to good results. Ensemble learning, with a recall of 86.2, precision of 87, and accuracy of 85.5, outperforms pseudo-labelling due to the combination of multiple models, which enhances robustness and overall performance. Ensemble methods benefit from reducing individual model biases, leading to more accurate predictions.

Other methods, such as rethinking semi-supervised learning, exhibit the weakest performance with a recall of 78.3, precision of 78, and accuracy of 77. This suggests significant limitations in their approach and potential areas for improvement. The method's lower performance indicates challenges in effectively utilizing labelled and unlabeled data.

Simple GANs exhibit high accuracy (89.7) but lower precision (78.1), indicating a trade-off between overall performance and prediction precision. The discrepancy suggests that while GANs are effective in certain aspects, they may struggle with precise boundary delineation.

Multi-task learning achieves high recall (89) and accuracy (90.2) but lower precision (84), highlighting its strength in capturing diverse aspects of the task. The method's ability to learn multiple tasks simultaneously can improve recall and accuracy, though precision may be compromised.

The DDDG-GAN method had the best overall performance, with a recall of 95.87%, a precision of 90.12%, and an accuracy of 92.56%. Therefore, the excellent results produced by DDDG-GAN demonstrate its effectiveness in leveraging the strengths of GANs while surmounting their limitations, hence allowing for significant advances in semi-supervised segmentation tasks related to binary classification on LIDC.

Baseline methods comparison

We validate the performance of different GAN-based methods for lung nodule detection in two different scenarios. In scenario 1, we train the proposed model on 80% of the LIDC dataset and evaluate it on the remaining 20%. In scenario 2, we use all LIDC datasets to train and test the model on the unseen LUNA16 dataset.

The results shown in Tables 4 and 5 represent the evaluation of the performance of various GAN-based methods. Scenario 1 demonstrates that the Pyramid Attention-Based GAN [28] achieved an accuracy of 82%, precision of 81%, recall of 87%, and an F-score of 84%. This method showed high recall, indicating its effectiveness in identifying positive cases with relatively high accuracy and precision, making it a robust model for Scenario 1(see Fig. 8). In comparison, the Reinforcement Learning-Based GAN [29] exhibited balanced metrics with an accuracy of 77%, precision of 80%, recall of 77%, and an F-score of 78%, showing moderate performance across all metrics.

MD-GAN [26] showed good performance with an accuracy of 76%, precision of 74%, recall of 82%, and an F-score of 78%. The high recall and balanced F-score indicate its reliability in Scenario 1. However, the proposed DDDG-GAN outperformed all other methods in Scenario 1, achieving the highest accuracy of 92.56%, precision of 90.12%, recall of 95.87%, and an F-score of 92.77%. This indicates superior performance in accurately identifying positive and negative cases, making it the best-performing model in this scenario.

In Scenario 2 (see Fig. 9), where models are tested on the LUNA16 dataset as an unseen dataset, the Pyramid Attention-Based GAN [28] saw a significant drop in performance, with an accuracy of 65%, precision of 67%, recall of 65%, and an F-score of 66%. This decline suggests potential overfitting to the training data. Similarly, the Reinforcement Learning-Based GAN [36] also showed a decrease in performance, with an accuracy of 62%, precision of 63%, recall of 61%, and an F-score of 62%, indicating balanced but lower metrics across the board.

MD-GAN [26] maintained moderate performance in Scenario 2, with an accuracy of 60%, precision of 59%,

Table 4 Comparison with the state-of-the-art

Ref	Method	Accuracy (%)	Precision (%)	Recall (%)
Cai et al. [4],2023	Self-training	87.1	86.5	85.8
Dzien et al. [5], 2024	Self-training	84.8	84.2	83.5
Xie et al. [24], 2023	Self-training	79.5	79	78.3
Yang et al. [7], 2024	co-training	87	88	86
Tang et al. [6], 2024	co-training	81.2	82	80.5
Bai et al. [35], 2024	co-training	82	83	81
Sun et al. [11], 2023	Graph-based training	80.87	81.5	80
Miller et al. [10], 2024	Graph-based training	89	89.5	88
Miller et al. [9], 2009	Graph-based training	90.2	90.5	89.5
Li et al. [12], 2024	pseudo label	85	85	84
Li et al. [16], 2023	Ensemble Learning	86.2	87	85.5
You et al. [15], 2023	Rethinking Semi-Supervised learning	78.3	78	77
Salimans et al. [34], 2016	Simple GAN	89.7	78.1	81.5
Khosravan et al. [22], 2018	Multi-task learning	90.2	84	89
DDDG-**GAN (proposed)**		**92.56**	**90.12**	**95.87**

Table 5 Evaluation of the performance of various GAN-based methods

Method	Scenarios	accuracy	precision	recall	f_score
[28],2024 pyramid attention-based GAN	Scenario1	82%	81%	87%	84%
	Scenario2	65%	67%	65%	66%
[36], 2024 Reinforcement Learning-based GAN	Scenario1	77%	80%	77%	78%
	Scenario2	62%	63%	61%	62%
MD-GAN [26], 2019	Scenario1	76%	74%	82%	78%
	Scenario2	60%	59%	58%	59%
DDDG-GAN (proposed)	Scenario1	92.56%	90.12%	95.87%	92.77%
	Scenario2 (Luna16)	72.6%	72.3%	73.82%	73.39%
	Scenario2 (LUNGx)	71.23%	67.56%	73.52%	70.42%

recall of 58%, and an F-score of 59%, suggesting some level of generalization to unseen data. On the other hand, the proposed DDDG-GAN showed relatively good performance in Scenario 2, maintaining the highest accuracy of 72.6%, precision of 72.3%, recall of 73.82%, and F-score of 73.39% among the compared methods. This suggests a better generalization capability to unseen data.

The Receiver Operating Characteristic (ROC) curves for the proposed DDDG-GAN under two scenarios are shown in Fig. 10. Scenario 1 illustrates the evaluation of the primary benchmark dataset, where the model performed an Area Under the Curve (AUC) of 96%, outperforming Reinforcement Learning-Based GAN (AUC = 0.88), Pyramid Attention-Based GAN (AUC = 0.91), and MD-GAN (AUC = 0.87), implying perfect classification performance with a vertical initial rise and close-to-perfect true positive rates at lower false positive rates. This mirrors the model's high sensitivity and particularity on the training-like dataset. Scenario 2, which assesses the model's generalizability on an unseen dataset simulating real-world situations, reached an AUC of 76%, followed by Pyramid Attention-Based GAN (AUC = 70%), Reinforcement Learning-Based GAN (AUC = 67%), and MD-GAN (AUC = 60%). Although lower than Scenario 1, the curve performs well, mirroring the challenges posed by data variability in clinical environments. These results emphasize the robustness and adaptability of the proposed model across various testing conditions, with compatible sensitivity and generalization in both controlled and real-world scenarios.

The confusion matrices in Fig. 11 reveal the strengths and weaknesses of the proposed model compared with the baseline models. The Pyramid Attention-Based GAN showed a high number of true positives (191) and true negatives (153), with relatively low false positives (45) and false negatives (29). This indicates its effectiveness in

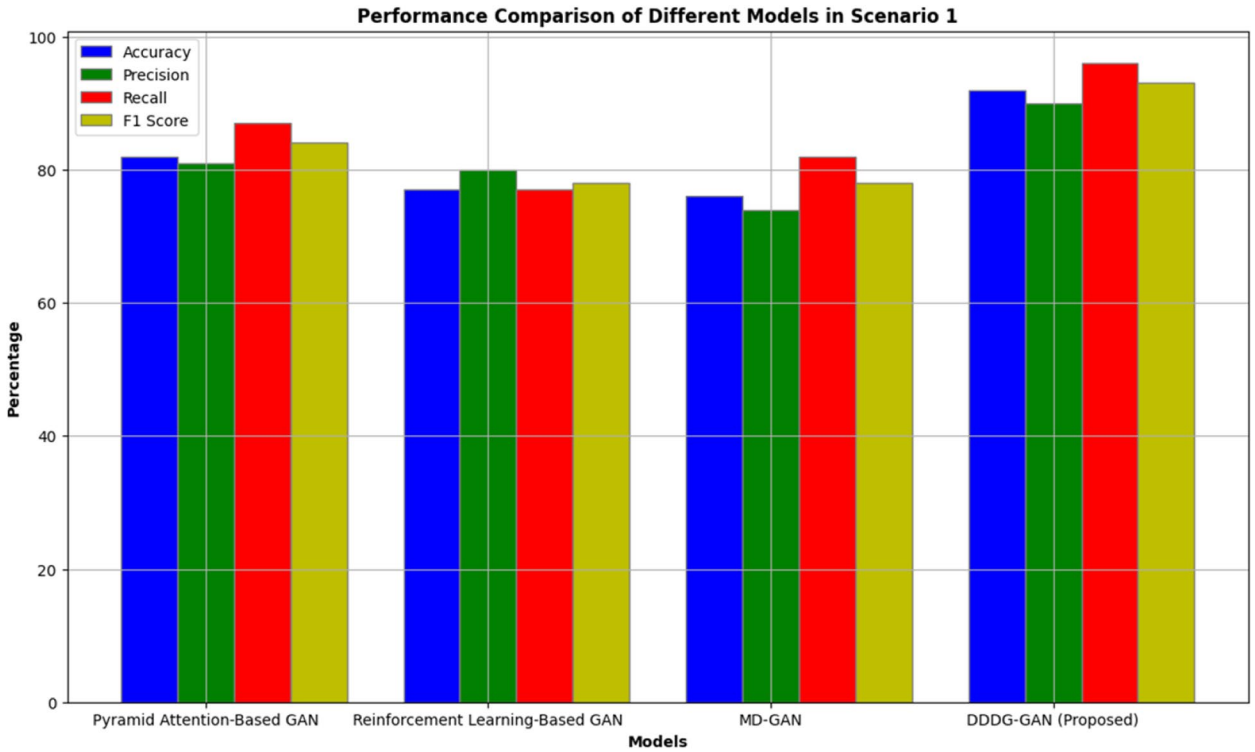

Fig. 8 Performance comparison of different models in scenario 1

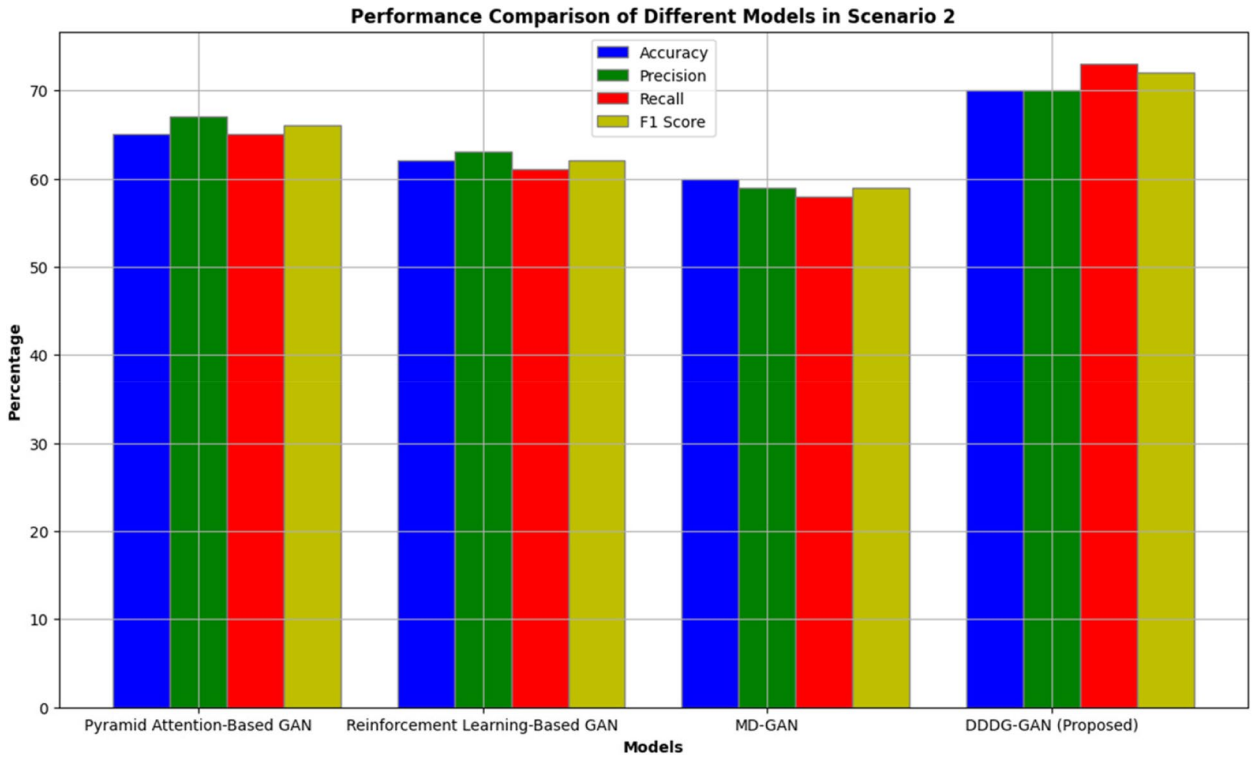

Fig. 9 Performance comparison of different models in scenario 2

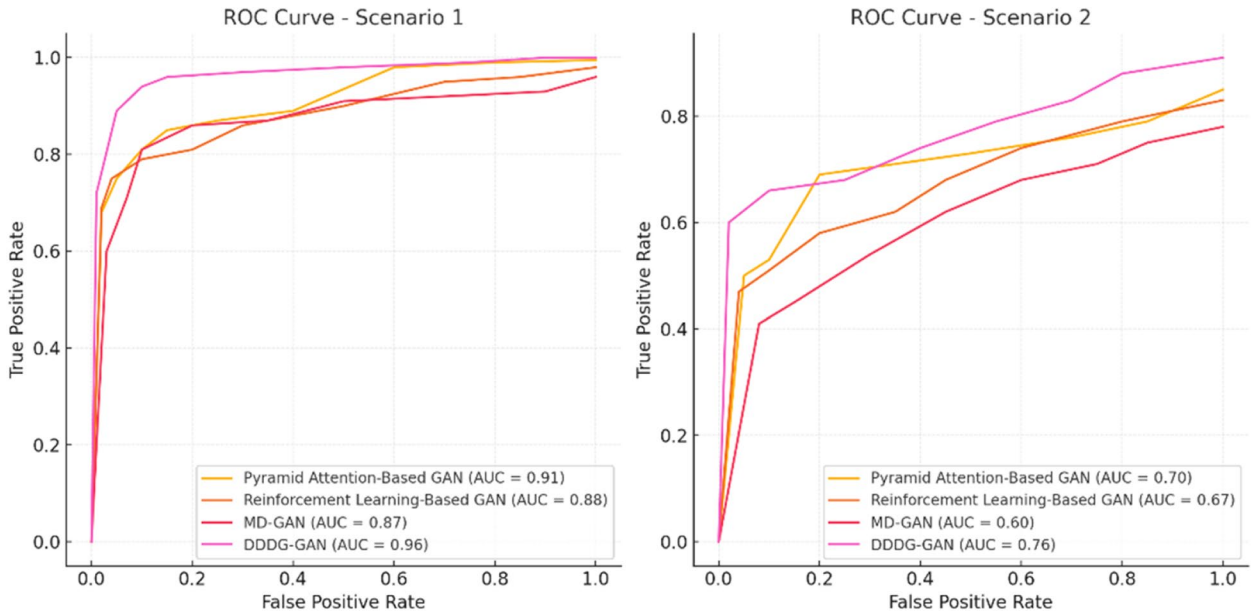

Fig. 10 ROC Curves for DDDG-GAN: Evaluation under two scenarios

identifying malignant and benign cases, with high recall suggesting it misses very few malignant cases. The Reinforcement Learning-Based GAN had a substantial number of true positives (174) and true negatives (148), with moderate false positives (43) and false negatives (53), demonstrating a balanced performance. MD-GAN [30] had good recall with true positives (176) and true negatives (143), though it faced moderate false positives (61) and false negatives (38). The proposed DDDG-GAN outperformed all, with true positives (201) and true negatives (185), minimal false positives (23), and false negatives (9), indicating robust performance.

In Scenario 2, the performance of each model declined when tested on the unseen LUNA16 dataset. The Pyramid Attention-Based GAN confusion matrix shows increased false positives (82) and false negatives (91), reflecting its struggle with new data. Similarly, the Reinforcement Learning-Based GAN had higher false positives (90) and false negatives (99), indicating difficulties in generalization. MD-GAN also saw a performance drop with increased false positives (98) and false negatives (101), highlighting the challenge of adapting to unseen data. The proposed DDDG-GAN, while facing difficulties, showed better generalization than other models with true positives (142), true negatives (159), false positives (98), and false negatives (101).

Computational complexity

The time complexity comparison over five epochs shown in Fig. 12 provides significant insights into the training efficiency of different models, including the Pyramid Attention-Based GAN, Reinforcement Learning-Based GAN, MD-GAN, and the proposed DDDG-GAN. The Pyramid Attention-Based GAN shows a starting time of 399 s in the first epoch, peaking at 425 s in the second epoch and fluctuating between 380 and 401 s over the remaining epochs. This variability indicates occasional optimization but a lack of consistent training efficiency. The Reinforcement Learning-Based GAN starts with a high training time of 477 s, increasing to 485 s in the second epoch, then decreasing to 439 s by the fourth epoch, before rising again to 462 s in the fifth epoch. These fluctuations and higher training times suggest less efficiency and more significant computational resource requirements.

The MD-GAN model was scientifically the highest in the training time, starting from 500 s in the first epoch, decreasing marginally to 480 s by the fourth epoch, and ending at 485 s in the fifth epoch. This means that efficiency is lower compared to other models. In contrast, the proposed DDDG-GAN starts with 468 s of training time, reduces considerably to 438 s in the second epoch, and stabilizes around 430 to 440 s in later epochs. This reduction and stabilization of training time show efficient learning and convergence by the model, making it more amenable for field applications with limited computational resources.

In terms of performance, DDDG-GAN outperforms every model in terms of efficiency and effectiveness. Lower training time across different epochs reflects

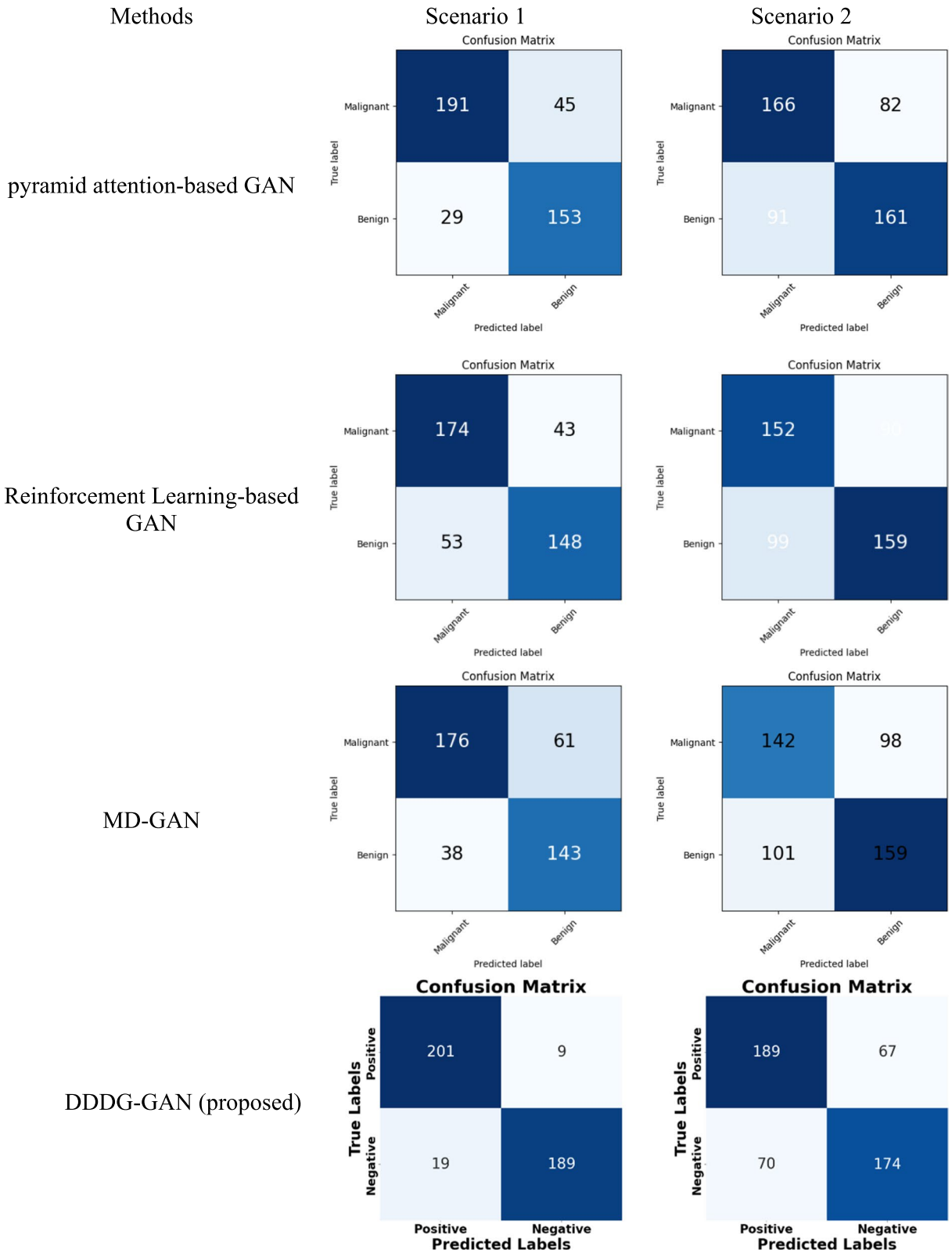

Fig. 11 The confusion matrices for baseline methods comparison

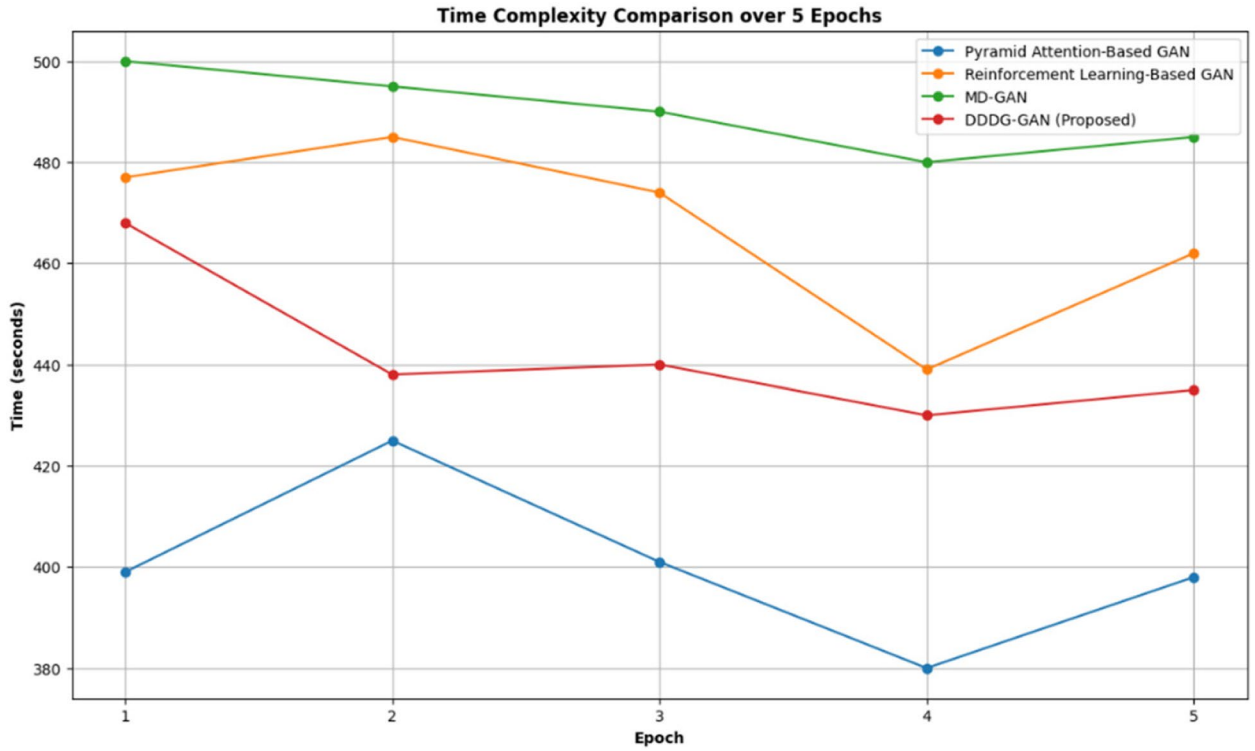

Fig. 12 Time complexity comparison over 5 epochs training

faster convergence that matters tremendously in scenarios where the environment is large-scale or resource constrained. The best accuracy, precision, recall, and F1 scores are reported for the DDDG-GAN in Scenario 1, each depicting superior classification performances. Also, even if, in the initial period, the training time for the DDDG-GAN is long, good generalizability can be noted in Scenario 2. It exhibits superior accuracy and more recall when compared to other models during the testing of the unseen LUNA16 dataset, meaning that the data variability can easily be handled as seen in this scenario; hence, it forms a foundation for robust performance regarding diverse clinical setups.

Figure 10 compares the test time complexity by different models. This figure provides crucial information related to the efficiency of the other models regarding the testing phase. The proposed model, DDDG-GAN, has the lowest time for the testing phase, which is 70 s. This shows that the proposed method is much more efficient in this phase. Moreover, the Pyramid Attention-Based GAN model has the second highest time, with a testing time complexity of 80 s, which is slightly high when compared with the proposed model. The Reinforcement Learning-Based GAN portrays a higher test time of 90 s, thus increasing the requirement for computation. The MD-GAN still has the highest test time of 100 s, which translates to the lowest efficiency of the models. Apart from training efficiency, the DDDG-GAN is more remarkable in the testing phase, making it a practicable model in real-life applications, considering the demand for high training efficiency and testing.

Ablation study

To assess the effectiveness of the proposed feature fusion mechanism, we conducted an ablation study under two scenarios; we approximated various feature fusion techniques, including no fusion (independent outputs), simple averaging, concatenation with a dense layer, weighted averaging, and the proposed attentive feature fusion mechanism. The performance metrics, including accuracy, precision, recall, and F1-score, are summarized in Tables 6 and 7.

In Scenario 1, the proposed attentive fusion mechanism surpassed all other methods with an accuracy of 92%, a precision of 90%, a recall of 96%, and an F1-score of 93%. More straightforward techniques, such as No Fusion, achieved lower accuracy (85%) due to the absence of integration between the outputs of the discriminators, guiding to suboptimal performance (Table 7). Simple Averaging enhanced the metrics slightly (accuracy: 88%, F1-score: 88%) but was incapable of catching subtle dependencies between features. Concatenation with

Table 6 Ablation Study of Feature Fusion Strategies (tested based on scenario 1)

Feature Fusion Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
No Fusion (Separate Outputs)	85	83	86	84
Simple Averaging	88	86	90	88
Concatenation + Dense Layer	91	89	94	91
Weighted Feature Averaging	90	88	93	90
Proposed Fusion (Attentive Map)	92	90	96	93

Table 7 Ablation Study on Unseen Dataset (Scenario 2)

Feature Fusion Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
No Fusion (Separate Outputs)	62	60	65	62
Simple Averaging	65	64	67	65
Concatenation + Dense Layer	68	67	70	68
Weighted Feature Averaging	67	66	69	67
Proposed Fusion (Attentive Map)	70	70	73	72

a Dense Layer achieved better (accuracy: 91%, F1-score: 91%) by leveraging more decadent feature interchanges, while Weighted Feature Averaging performed moderate advancement (accuracy: 90%, F1-score: 90%) by balancing the contribution of each feature. The proposed attentive fusion mechanism presented a superior performance by effectively integrating discriminative features, capturing subtle attributes that improved classification performance.

In Scenario 2, which describes a real-world test scenario, the proposed attentive fusion mechanism also performed the most elevated metrics, with an accuracy of 70%, a precision of 70%, a recall of 73%, and an F1-score of 72%. In this scenario (Table 7), No Fusion demonstrated significant performance degradation (accuracy: 62%, F1-score: 62%), mirroring the incapacity to generalize without feature integration. Simple Averaging provided a slight improvement (accuracy: 65%, F1-score: 65%), while Concatenation with a Dense Layer and Weighted Feature Averaging achieved better (accuracy: 68% and 67%, respectively). However, these methods still needed to work on handling the variability and complexity of unseen cases—the attentive fusion mechanism excelled by integrating essential features and maintaining strong generalization in a challenging real-world setting.

Discussion

The proposed DDDG-GAN incorporates several innovations that make it superior to the state-of-the-art semi-supervised learning approaches to classify lung nodules. In this regard, the DDDG-GAN exploits, in an efficient way, semi-supervised learning paradigms to overcome the inherent problems of the limited labelled medical imaging data. This is motivated by the knowledge that a large amount of unlabeled data and a small quantity of labelled data improves both the aspects, the learning efficacy and generalizability of the model. The strategy makes the process independent of large labelled datasets that are usually complicated and expensive to obtain in medical contexts.

The t-SNE visualizations in Fig. 13 demonstrate the effectiveness of the feature extraction and fusion process in distinguishing between benign and malignant nodules. The distinct clustering in individual and fused feature maps highlights the model's capability to maintain class-specific information while allowing for proximity near the decision boundary. This characteristic is essential for enhancing the diagnostic accuracy of the model, ensuring that benign and malignant nodules are accurately identified based on their unique features. The balance achieved in the fused feature map indicates that the model is well-tuned to leverage the class's commonalities and differences for improved classification performance.

The clear separation of the clusters in the fused feature map indicates that the fusion process effectively combines the discriminative features from FD1 and FD2 without significant overlap. The proximity of the clusters suggests that while they are close enough to the decision boundary, they retain their class-specific characteristics. This balance is crucial for improving the model's accurate classification of nodules. The semi-supervised DDDG-GAN model benefits significantly from the apparent separation and balanced proximity of the clusters in the fused feature map. These characteristics enhance the

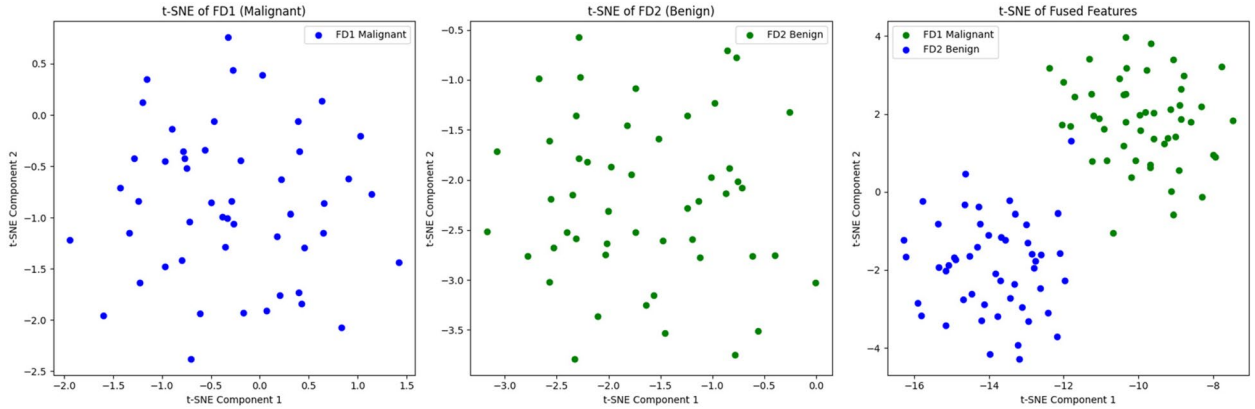

Fig. 13 The t-SNE visualizations, (left): the features extracted from a malignant nodule, (middle): the features extracted from a benign nodule, and (right): the fused features from both classes

model's discriminative power, generalization capability, and efficient use of labelled and unlabeled data, improving classification accuracy and enhancing decision-making in lung nodule classification.

Grad-CAM and SHAP outputs in Figs. 5 and 6 provide a robust framework for understanding and validating the DDDG-GAN model's decision-making process in lung nodule classification. Grad-CAM visualizations offer spatial context by highlighting the regions of interest, ensuring the model focuses on diagnostically relevant areas. At the same time, SHAP values provide a detailed numerical breakdown of feature importance, explaining the influence of individual features on the model's predictions. Together, these tools validate the model's strengths in correctly identifying benign and malignant nodules and identifying areas for improvement by revealing misclassifications and guiding targeted refinements. This dual approach enhances the model's accuracy, interpretability, and trustworthiness, which is crucial for reliable medical diagnostics.

What truly sets the DDDG-GAN apart and is a reason for its success is a two-generator, two-discriminator-based architecture. Each generator undergoes training on a different class of nodules, benign or malignant, ensuring their synthesized image strongly represents their respective courses Fig. 4. The architecture of the dual generator ensures mode collapse- a phenomenon that traditional GANs often suffer from, and it offers a rich set of high-fidelity images. Thus, The model can generate a much better generalized capability from the training data, leading to significant improvements in diagnostic accuracy with the dual discriminators specialized for each class respectively Fig. 6.

As this work has incorporated dual discriminators, giving them individual attention facilitates the differentiation of those features that can demarcate between the benign and malignant nodules. This advancement allows better classification and increases the model's reliability when applied to real-world medical applications with higher data variability. The advanced feature fusion mechanism employed by the DDDG-GAN places discriminative feature maps from the dual generators and dual discriminators into the attentive feature map. In this way, significant features representing nodule malignancy are captured by fusion and fed into convolutional and fully connected layers for accurate classification. Therefore, the capability increases the ability of a model to capture the fine-grained differences between benign and malignant nodules, precisely what is needed in medical diagnosis and other fine-grained classification applications.

The DDDG-GAN arranges the loss functions so that there is a balance in the contributions by the cross-entropy loss and contrastive loss; the combination leaves little doubt that this will generate fused feature maps which will be firmly and separately set apart in cases of both benign and malignant nodules for better feature discrimination and accurate classification. The attentive feature mapping afterwards will enhance the model's capability for diagnosing nodules, making it clinically viable with more accuracy.

Regarding a direct comparison, DDDG-GAN achieves the highest accuracy, 92.34%, and recall, 95.71%, in Scenario 1, as illustrated in Table 1. It is superior to the other semi-supervised methods, such as self-training, co-training, and graph-based training. It is highly desirable to have a high recall in medical diagnosis to avoid missing any cases of malignancy so that the patients can be treated early enough for a better outcome. The proposed model also had the highest precision, totalling 89.93%, with the highest F1 score, 92.62%, distorting both the Specificity and sensitivity for well-balanced performance

in discrimination between benign and malignant modules. It is crucial to avoid false positives and negatives, both of which are a source of serious concern in clinical diagnosis.

Although a decrease in performance was found in scenario 2 when DDDG-GAN was tested on the unseen LUNA16 dataset, it still presents a considerably high accuracy of 70.2 per cent and a recall of 73.43%. It demonstrates better generalization capabilities in comparison to current state-of-the-art methods. High generalizability is crucial for the model's applicability in real clinical scenarios where data variability is typically high. The Grad-CAM visualizations shown in Fig. 5 further aid in understanding the decisions made by the model. The visualizations below provide clear evidence of the specific parts of the nodule that the model focuses on when making the classification. If a nodule is classified correctly, the heatmaps offer insight into diagnostically relevant features that support the model's classification decisions. In misclassification cases, the heatmaps provide context on areas for improvement and guiding refinements.

The proposed DDDG-GAN demonstrates superior efficiency in the training and testing phases (see Figs. 12 and 14). During training, it shows a consistent reduction in time across five epochs, starting at 468 s and decreasing to 430 s, outperforming the Pyramid Attention-Based GAN, Reinforcement Learning-Based GAN, and MD-GAN, which all exhibit higher and more variable training times. In the testing phase, the DDDG-GAN maintains its efficiency with the lowest time of 70 s, compared to 80 s for the Pyramid Attention-Based GAN, 90 s for the Reinforcement Learning-Based GAN, and 100 s for the MD-GAN. This consistent performance in training and testing phases underscores DDDG-GAN's practicality and robustness for real-world applications, particularly in resource-constrained environments.

The possibility of progressive generative models, such as GANs and diffusion models, to revolutionize clinical workflows cannot be magnified. These models deliver a powerful solution to expand datasets, specifically for underrepresented patient groups, confirming that AI systems are trained on various and representative data. This ability handles a critical bottleneck in clinical AI development, where imbalanced or deficient datasets often determine the generalizability and fairness of models. Moreover, the power of these generative models to synthesize high-fidelity, class-specific medical images allows the creation of tailored datasets that mirror rare or minority cases, which are otherwise difficult to collect in sufficient quantities. By seamlessly combining synthetic data into training pipelines, these technologies can significantly improve the robustness and diagnostic accuracy of AI systems, paving the way for more fair and efficient clinical workflows. This advancement highlights their transformative function in making accurate medicine more accessible and trustworthy.

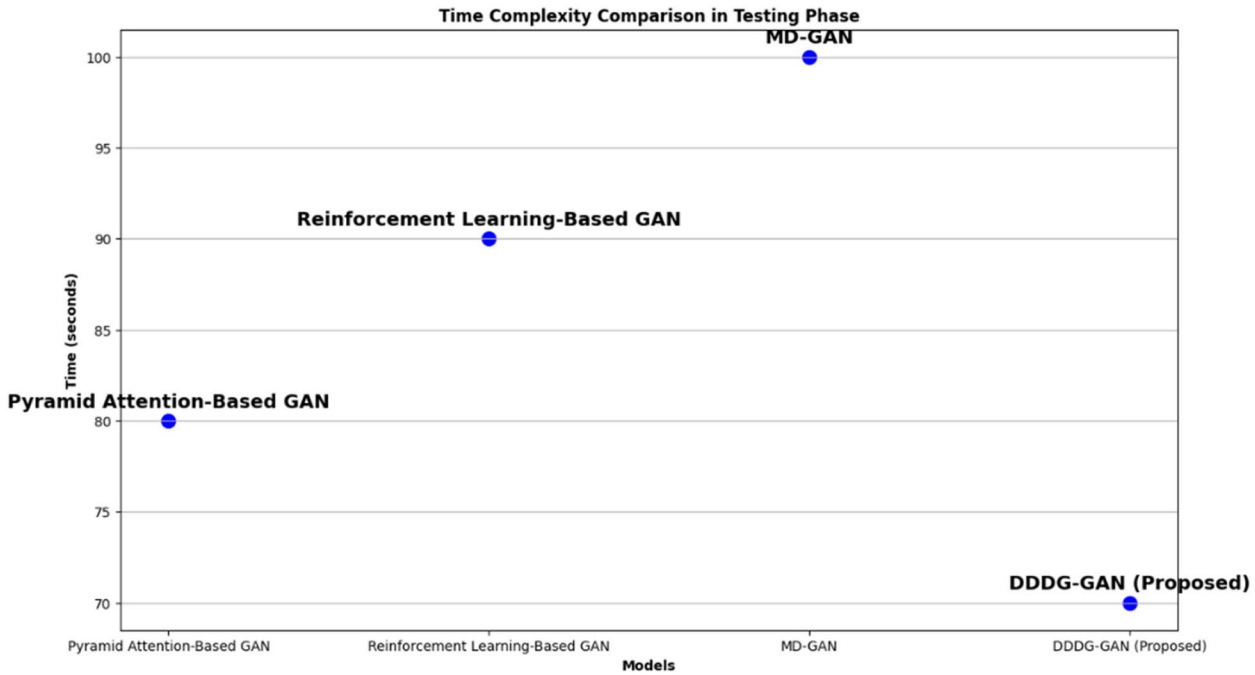

Fig. 14 Time complexity comparison in the testing phase

Conclusions

This study introduces a novel DDDG-GAN for the semi-supervised classification of lung nodules. The innovative architecture of DDDG-GAN addresses several critical challenges in medical image analysis, including data scarcity, mode collapse, and the need for generalizability.

Most importantly, the dual-generator concept helps the DDDG-GAN overcome the mode collapse problem commonly found in a traditional GAN. Since its two generators train individually on benign and malignant nodule data, the model ensures the generation of diverse and high-fidelity synthesized images with representative characteristics of their respective classes. Such diversity in the synthesized data makes the model more robust and better at representing those minor, diagnostically relevant differences between benign and malignant nodules.

Correlated with the dual-generator structure, the dual-discriminator framework enhances generalization from training to unseen data. Each discriminator specializes in a specific class and possesses in-depth knowledge of the unique features of that class. Consequently, the classification result becomes more accurate, and the model is more reliable under clinical scenarios where data variability can be very high. Due to the discriminators, they can focus on class-specific attributes. Therefore, the model learns to generate realistic images while most effectively distinguishing between different kinds of nodules.

The DDDG-GAN features a sophisticated mechanism of feature fusion. It fuses the discriminative feature maps output by the dual generators and dual discriminators into an attentive feature map. Critical features indicating nodule malignancy are contained in the attentive feature map. The convolutional and fully connected layers further process such captured features for accurate classification. In this way, the model becomes adept at classifying the minor differentiation between benign and malign nodules, which is extremely important in a medical diagnosis.

The experimental evaluations confirm the model's superior performance. In Scenario 1, training on 80% and testing on the remaining 20% of the LIDC-IDRI dataset, the model has shown outstanding performance in terms of accuracy, 92.34%, precision, 89.93%, recall, 95.71%, and the F1 score, 92.62%. These metrics indicate the model's ability to learn robustly and its effectiveness in classifying lung nodules. Our results showed that the DDDG-GAN, when tested on the unseen LUNA16, had a relatively good or high accuracy of 70.2% and perfect recall of 73.43%, far better than the results of some semi-supervised learning methods, which shows its effectiveness in generalizing in working with new unseen data. In other words, our method reached an accuracy of 92%, a precision of 90%, a recall of 96%, and an F1-score of 93%, significantly surpassing Pyramid Attention-based GAN (accuracy: 82%, F1-score: 84%), Reinforcement Learning-based GAN (accuracy: 77%, F1-score: 78%), and MD-GAN (accuracy: 76%, F1-score: 78%). Likewise, in Scenario 2, DDDG-GAN reached an accuracy of 70%, a precision of 70%, a recall of 73%, and an F1-score of 72%, exceeding the corresponding metrics of the approximated methods. These results emphasize the robustness and significance of the proposed dual-discriminator and dual-generator architecture in improving classification performance and handling challenges such as mode collapse and data lack.

Also, the corresponding time complexity analysis confirms the applicability of the DDDG-GAN since its training and testing times were always shorter than identified by the other state-of-the-art methods. Efficiency in computation time and high performance make the DDD-GAN an efficient and helpful tool in such accurate medical image analysis situations where this is applied in resource-constrained situations.

While the proposed DDDG-GAN demonstrates vital performance metrics on benchmark datasets such as LIDC-IDRI and LUNA16, its clinical validation still needs to be improved. The current study does not evaluate the model on real-world clinical datasets, which often include patients with comorbidities or imaging conditions that could alter the appearance of lung nodules. Such diverse cases could affect the model's generalizability. To address this limit, future work will focus on validating the model in a clinical environment by testing it on data collected from hospitals and clinics, enclosing a broader spectrum of patient demographics and medical conditions.

Despite the recent strides DDDG-GAN has taken towards pulmonary nodule semi-supervised classification, there is still much scope for further research to improve and enhance its applicability. It could be further improved by developing better generalization techniques than currently exist in the model, such as advanced data augmentation methods for domain adaptation to deal better with diverse and unseen datasets. Also, multimodal data, as in the case of MRI or PET scans and CT scans, could be incorporated, providing more insight into diagnosis information and thus making this model all the more accurate and robust. Automated hyperparameter optimization techniques, such as Bayesian optimization and genetic algorithms, would make the process automatic and result in better model performance without much manual intervention.

Acknowledgements

The authors would like to thank the Queensland University of Technology for supporting the project.

Authors' contributions
Authors' contributions: Ahmed Saihood: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. Wijdan Rashid Abdulhussien: Conceptualization, Data curation, Methodology, Writing – original draft. Laith Alzubaid: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Supervision, Validation, Writing – review & editing. Mohamed Manoufali: Validation, Writing – review & editing. Yuantong Gu: Funding acquisition, Supervision, Writing – review & editing.

Authors' information
Not applicable.

Funding
The authors would like to acknowledge the support received through the following funding schemes of the Australian Government: Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for Joint Biomechanics under Grant IC190100020.

Data availability
We have used public datasets which are: i) LUNA 2016: https://luna16.grandchallenge.org/Download/ ii) LIDC-IDRI: https://paperswithcode.com/dataset/lidc-idri.

Declarations

Ethics approval and consent to participate
We have used public datasets, which are: i) LUNA 2016: https://luna16.grandchallenge.org/Download/. ii) LIDC-IDRI: https://paperswithcode.com/dataset/lidc-idri.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]College of Computer Science and Mathematics, University of Thi-Qar, Thi Qar, Iraq. [2]School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia. [3]Centre for Data Science, Queensland University of Technology, Brisbane, QLD 4000, Australia. [4]Aptium Company, Brisbane, QLD 4059, Australia. [5]Space & Astronomy, Commonwealth Scientific Industrial Research Organisation (CSIRO), Kensington, WA 6151, Australia.

Received: 9 September 2024 Accepted: 16 December 2024
Published online: 24 December 2024

References

1. Goodfellow I, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. https://doi.org/10.1145/3422622.
2. van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn. 2020;109(2):373–440. https://doi.org/10.1007/s10994-019-05855-6.
3. Oliver A, Odena A, Raffel C, Cubuk ED, and Goodfellow IJ. Realistic evaluation of semi-supervised learning algorithms, 6th Int. Conf. Learn. Represent. ICLR 2018 - Work. Track Proc., no. NeurIPS; 2018.
4. Cai G-W. et al. Semi-Supervised Segmentation of Interstitial Lung Disease Patterns from CT Images via Self-Training with Selective Re-Training. Bioengineering. 2023;10(7).https://doi.org/10.3390/bioengineering10070830.
5. Alzubaidi L, Fadhel MA, Hollman F, Salhi A, Santamaria J, Duan Y, Gupta A, Cutbush K, Abbosh A, Gu Y. SSP: self-supervised pertaining technique for classification of shoulder implants in x-ray medical images: a broad experimental study. Artif Intell Rev. 2024;57(10):261. https://doi.org/10.1007/s10462-024-10878-0.
6. Tang T, Zhang X, Li W, Wang Q, Liu Y, Cao X. Co-training based prediction of multi-label protein–protein interactions. Comput Biol Med. 2024;177:108623. https://doi.org/10.1016/j.compbiomed.2024.108623.
7. Yang J, Li H, Wang H, Han M. 3D medical image segmentation based on semi-supervised learning using deep co-training. Appl Soft Comput. 2024;159:111641. https://doi.org/10.1016/j.asoc.2024.111641.
8. Chong Y, Ding Y, Yan Q, Pan S. Graph-based semi-supervised learning: A review. Neurocomputing. 2020;408:216–30. https://doi.org/10.1016/j.neucom.2019.12.130.
9. Zha Z-J, Mei T, Wang J, Wang Z, Hua X-S. Graph-based semi-supervised learning with multiple labels. J Vis Commun Image Represent. 2009;20(2):97–103. https://doi.org/10.1016/j.jvcir.2008.11.009.
10. Miller KS, Bertozzi AL. Model Change Active Learning in Graph-Based Semi-supervised Learning. Commun Appl Math Comput. 2024;6(2):1270–98. https://doi.org/10.1007/s42967-023-00328-z.
11. Sun Y, Shi Z, and Li Y. A Graph-Theoretic Framework for Understanding Open-World Semi-Supervised Learning. Adv Neural Inf Process Syst. 2023, vol. 36, pp. 23934–23967, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/4b6898c70d5b328deaf2216aefd8f77a-Paper-Conference.pdf.
12. Su J, Luo Z, Lian S, Lin D, Li S. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Med Image Anal. 2024;94:103111. https://doi.org/10.1016/j.media.2024.103111.
13. Li X, Wu Y, Dai S. Semi-supervised medical imaging segmentation with soft pseudo-label fusion. Appl Intell. 2023;53(18):20753–65. https://doi.org/10.1007/s10489-023-04569-6.
14. Qiu L, Cheng J, Gao H, Xiong W, Ren H. Federated Semi-Supervised Learning for Medical Image Segmentation via Pseudo-Label Denoising. IEEE J Biomed Heal Informatics. 2023;27(10):4672–83. https://doi.org/10.1109/JBHI.2023.3274498.
15. You C. et al. Rethinking Semi-Supervised Medical Image Segmentation: A Variance-Reduction Perspective. Adv Neural Inf Process Syst. 2023;3(CL).
16. Li J, Shi H, Chen W, Liu N, and Hwang K-S. Semi-Supervised Detection Model Based on Adaptive Ensemble Learning for Medical Images. IEEE Trans Neural Networks Learn Syst.2023:1–12.https://doi.org/10.1109/TNNLS.2023.3282809.
17. Wang D, Zhang Y, Zhang K, and Wang L. FocalMix: Semi-supervised learning for 3D medical image detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2020:3950–3959.https://doi.org/10.1109/CVPR42600.2020.00401.
18. Loyman M, Greenspan H. Semi-supervised lung nodule retrieval. 2020. [Online]. Available: http://arxiv.org/abs/2005.01805.
19. Chen W, Li K. Self-supervised Learning for Medical Image Classification Using Imbalanced Training Data. Commun Comput Inf Sci. 2022;1590:242–252.https://doi.org/10.1007/978-981-19-4109-2_23.
20. Liu K, Ning X, Liu S. Medical image classification based on semi-supervised generative adversarial network and pseudo-labelling. Sensors. 2022;22(24):1–12. https://doi.org/10.3390/s22249967.
21. Li R, Zhou L, Wang Y, Shan F, Chen X, Liu L. A graph neural network model for the diagnosis of lung adenocarcinoma based on multimodal features and an edge-generation network. Quant Imaging Med Surg. 2023;13(8):5333–48. https://doi.org/10.21037/qims-23-2.
22. Khosravan N, Bagci U. Semi-Supervised Multi-Task Learning for Lung Cancer Diagnosis. Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS.2018;2018:710–713.https://doi.org/10.1109/EMBC.2018.8512294.
23. Kuang YAN, Lan T, Peng X. Unsupervised Multi-Discriminator Generative Adversarial Network for Lung Nodule Malignancy Classification. IEEE Access. 2020;8:77725–34. https://doi.org/10.1109/ACCESS.2020.2987961.
24. Xie Z, Sun H, Li M. Semi-supervised Learning with Support Isolation by Small-Paced Self-Training. Proc 37th AAAI Conf Artif Intell AAAI 2023. 2023;37:10510–10518, 2023, https://doi.org/10.1609/aaai.v37i9.26249.
25. Kallipolitis A, Revelos K, Maglogiannis I. Ensembling EfficientNets for the Classification and Interpretation of Histopathology Images. 2021.
26. Hardy C, Le Merrer E, Sericola B. "MD-GAN: Multi-Discriminator Generative Adversarial Networks for Distributed Datasets", in. IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019;2019:866–77. https://doi.org/10.1109/IPDPS.2019.00095.
27. Toizumi T, Zini S, Sagi K, Kaneko E, Tsukada M, Schettini R. ARTIFACT-FREE THIN CLOUD REMOVAL USING GANS NEC corporation (Japan), 2 University of Milano-Bicocca (Italy). 2019;1:3596–3600.

28. Li G, Wang J, Tan Y, Shen L, Jiao D, Zhang Q. Semi-supervised medical image segmentation based on GAN with the pyramid attention mechanism and transfer learning. Multimed Tools Appl. 2024;83(6):17811–32. https://doi.org/10.1007/s11042-023-16213-z.
29. Ren Z, Lan Q, Zhang Y, Wang S. Exploring simple triplet representation learning. Comput Struct Biotechnol J. 2024;23:1510–21. https://doi.org/10.1016/j.csbj.2024.04.004.
30. Ren Z, Zhang Y, Wang S. A Hybrid Framework for Lung Cancer Classification. Electronics. 2022;11(10):1614.https://doi.org/10.3390/electronics1010000.
31. Ren Z, Wang S, Zhang Y. Weakly supervised machine learning. CAAI Trans Intell Technol. 2023;8(3):549–80. https://doi.org/10.1049/cit2.12216.
32. Huijben EMC, Pluim JPW, van Eijnatten MAJM. Denoising diffusion probabilistic models for addressing data limitations in chest X-ray classification. Informatics Med Unlocked. 2024;50:101575. https://doi.org/10.1016/j.imu.2024.101575.
33. Kim HK, Ryu IH, Choi JY, Yoo TK. A feasibility study on the adoption of a generative denoising diffusion model for the synthesis of fundus photographs using a small dataset. Discov Appl Sci. 2024;6(4).https://doi.org/10.1007/s42452-024-05871-9.
34. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. Adv Neural Inf Process Syst. 2016:2234–2242.
35. Bai Y, Mi J, Li L. Information granule optimization and co-training based on kernel method. Appl Soft Comput. 2024;158:111584. https://doi.org/10.1016/j.asoc.2024.111584.
36. Xu C, Zhang T, Zhang D, Zhang D, Han J. Deep Generative Adversarial Reinforcement Learning for Semi-Supervised Segmentation of Low-Contrast and Small Objects in Medical Images. IEEE Trans Med Imaging. 2024:1.https://doi.org/10.1109/TMI.2024.3383716.

Publisher's Note