International Journal on Smart & Sustainable Intelligent Computing
Vol. 02, Iss. 02, S. No. 05, pp.65--79, June 2025
ISSN (Online): 3048-8508

Received: 11 March 2025, Accepted: 26 April 2025, Published: 23 May 2025
Digital Object Identifier: https://doi.org/10.63503/j.ijssic.2025.122

Research Article

Smart Data Augmentation using Generative Adversarial Networks for Rare Oncological Disease Classification

Rahul Vadisetty[1], Himanshu Suyal[2]*

[1] Electrical engineering, Wayne State University, Detroit, MI, USA

[2] School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India

rahulvy91@gmail.com[1], suyal.himanshu@gmail.com[2]

*Corresponding author: Himanshu Suyal, suyal.himanshu@gmail.com

ABSTRACT

Generative Adversarial Networks (GANs) have become a powerful tool for generating synthetic data, and they fill in the important gap concerning the lack or imbalance of real data with respect to healthcare applications. Annotated data for rare oncological diseases that would allow training machine learning models is not available. A modern system based on GANs, which uses conditional GANs (cGANs) and Wasserstein, is described here. The goal is to extend existing datasets and improve the outcomes of classifications for rare diseases. This is achieved by extensive preprocessing, the introduction of noise to avoid overfitting, and carefully executed validation procedures after synthesis to retain biological consistency and statistical coherence. Based on the experimental results presented, classifiers trained on augmented data produce much better sensitivity, specificity, and F1 scores than the baseline models, provided that the classes are significantly imbalanced. This study uses heatmap correlation analysis and distributional assessments between synthetic and real samples to measure data realism within a modular framework that fuses adversarial training and strict validation of synthetic data for augmentation in rare cases. Outcomes of the study support the idea that GAN-generated datasets offer a promising way to improve robust diagnostic models, thus addressing the data shortage that is rampant in oncology research. This research broadens the use of GANs in synthesising medical data, which enriches the growing toolkit of computational approaches to strengthen the early detection and categorisation of rare cancers that benefit from data-based techniques.

Keywords: *Generative Adversarial Networks, Synthetic Medical Data, Rare Oncological Diseases, Data Augmentation, Deep Learning, cGAN, WGAN, Classification.*

1. Introduction

Because of the low frequency of oncological diseases, there are diagnostic challenges, especially since there are few adequately annotated datasets. This constraint causes a reduction in the performance of supervised learning algorithms, not least because they depend on large, annotated data collections. The usual augmentation techniques, like rotating, flipping, and adding noise, which do well with images, do not perform as well with structured clinical or genomic data. There is therefore a developing interest in generative modelling methods that are able to model elaborate data distributions and synthesise synthetic samples that are similar to the real ones. Response GANs are made up of two artificial neural networks: a generator and a discriminator, which compete against one another in a minimax game. The goal of the generator is to create data indistinguishable, by the discriminator, from the genuine instances. With successive training, the generator gradually learns to mimic the underlying data distribution. cGANs enhance the original architecture by incorporating class-specific details, which makes them especially suitable for supervised learning for such purposes as rare disease classification [2][3].

GAN's medical applications have already been proven in radiology [4], histopathology [5], genomics [6], and electronic health records [7]. With the challenge in training models because of the variability of patients and high-dimensional sets of features in oncology, GANs provide a highly valuable approach to handling class imbalance. When applied to rare cancers, the results are quite striking and will enable scientists to create models of simulated data sets that maintain clinical validity [8][9]. However, GANs' applications for modelling rare diseases face a range of difficulties. Some of the most frequently occurring issues are overfitting, mode collapse and lack of diversity in the generated samples [10]. Resolution tends to require careful network configuration, reasonable hyperparameter choice, and post-generation evaluation. Research shows that Wasserstein loss application, spectral normalization, and conditional embedding can help improve the stability and variability of produced samples in the GANs [11][12]. A GAN-based synthetic data generation strategy for the rare oncological diseases is proposed and evaluated in this research. The approach integrates radiomic and genomic characteristics in order to streamline multi-modal learning. The augmented data is used as the training set for a classification model that is assessed by means of classical metrics, such as accuracy, sensitivity, specificity, and F1 score. Heatmap analysis and feature distribution plots are used to maintain the fidelity of the synthetic data.

The validation of the proposed method is performed using datasets from The Cancer Genome Atlas (TCGA) [13] and Orphanet [14]. Experimental results underscore performance improvements for rare cancer classification. Built with the mind of the module, the pipeline is capable of managing various forms of data and is designed to offer a flexible framework for the augmentation of medical data in environments with small patient' numbers.

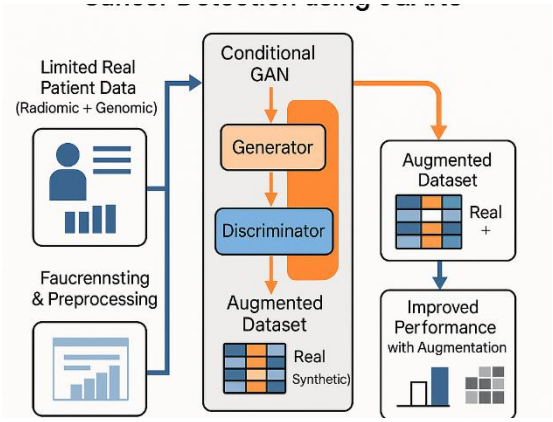

Fig.1 Graphical Abstract

In Figure 1, the graphical abstract visualises the first stage of the pipeline, which is initiated through limited real patient data that is preprocessed and features extracted. This extracted feature set is used by a conditional GAN to create additional data samples. The authenticity of synthetic data is measured using a discriminator network. The consolidated dataset, comprising real and generated samples, is used to train a classifier dedicated to detecting rare cancers. A supplementary chart shows that data augmentation results in better accuracy, as evidenced in our results. The pipeline architecture has well-defined, colour-labelled modules that clearly distinguish the flow of real and synthetic data through the pipeline. The aim is to transfer the GAN research into the clinical world, providing a meaningful applicability and scalability for the clinical representation of AI for rare disease diagnosis. The generation of synthetic data elevates the efficacy and uniformity of classifiers, affording superior early detection of rare cancers and better patient care.

2. Related Research

Generative Adversarial Networks (GANs) have transformed the synthesis of synthetic data in medical science, with great influence on the elimination of limitations inherent in the availability of inadequate

and imbalanced data for rare oncological diseases. Using adversarial training, GANs are able to create complex distributions of data that can even create very real synthetic data mimicking real patient data. For rare cancers, the reduced and sometimes lopsided datasets prevent the creation of effective diagnostic models, but GANs allow for diverse synthetic samples to be generated that support this process.

Initial experiments focused on producing fake medical images so as to improve classifier accuracy. According to studies, the expansion of small pools of data by synthetic generation has improved the sensitivity and specificity of image-based diagnostic models. Adversarial techniques were implemented in computed tomography and magnetic resonance imaging, leading to the generation of synthetic tumour cuts that preserved important structural properties from real medical images. Conducted clinical assessments showed that it was hard for pathologists to categorise synthetic images as synthetic, from a superior visual acumen that was astutely acquired from adversarial training [15]. Then, researchers introduced class-conditional GAN variations allowing the generation of synthetic data conditioned on specific subtypes of tumours or clinical stages. By addressing class imbalance, this innovation enabled the retrieval of discriminative features from lowly represented categories in multi-label data sets. Conditional generators greatly improved precision and recall in model training with the help of augmented datasets [16].

Next-generation systems employed altered loss functions and stabilised gradient flows, which led to fewer artefacts and anatomical verisimilitude in produced samples. Investigations aimed at adding latent spaces having a defined structure to enable intentional modifications in synthetic samples. In the creation of ongoing clinical data, such as laboratory measurements or genomic expression profiles, these models showed clear benefits owing to a function to maintain underlying statistical correlations [17]. The use of synthetic augmentation was also applied to non-image medical formats such as tabular and sequential patient information, to further extend the diversity of the dataset. To adapt the GAN architecture to operate with discrete and categorical information, researchers succeeded in emulating diagnostic histories, gene sequences and treatment procedures. Apart from keeping statistical distributions, the generated data provided relevant variability in training sets, ultimately enabling classifiers to achieve better results on rare or intricate cases. The assessment of these studies refined its classification performance, particularly those of diseases which depend on sparse training data [18].

Several research groups insisted on the necessity to evaluate the clinical value of synthetic data going beyond mere visual or linguistic likeness. Domain-specific validation frameworks were proposed with the help of structural similarity indices, statistical proximity metrics, and expert-validated anatomical features for evaluation. Model effectiveness was measured by comparing the increase in classifier test set performance exhibited by unseen data sets, providing an indirect way of testing synthetic data validity. The multi-level approach applied in analysing the generated data guaranteed both visual similarity and practical clinical use [19]. Synthetic data production methods were aligned with evolving attention to data privacy and ethics. Investigators encouraged the use of adversarial training in scenarios where access to patient data is managed or secured to allow decentralised approaches to evolve. Synthetic data, devoid of personal information, full of statistical specifics, allowed for representing unusual cancer cases within ethical boundaries. By this approach, synthetic augmentation was opened as an option for areas where conventional exchange of data was not possible or would be too tough [20]. Although significant progress has been made, there are several hindrances to GAN-based augmentation. A common issue that has been noted is a problem referred to as mode collapse, when generator outputs are repetitive, thereby reducing the diversity of the dataset. To fight mode collapse, the researchers have used a number of adjustments to the GAN architecture and the training process, among which are normalisation strategies and additional feedback from the discriminator. Despite these advancements, stability and diversity are still acute problems for the medical GAN-based approaches.

Another significant problem is the requirement for approaches that would allow the interpretation of synthetic data to become feasible. Acknowledging the benefits of transparency and traceability in the synthesis of synthetic data, researchers have been putting measures in place to ensure that biases or

artefacts that could taint downstream evaluations are prevented from occurring. Scientists have begun employing explainability methods to shed light on the transparency that allows users to see the outputs of latent variables affecting synthetic data and learn how synthetic features depend on input ones. Through such a level of transparency, the field seeks to gain trust and support for the implementation of the synthetic techniques in high-impact medical applications. With the most recent research, researchers are engaging in further exploration of how cross-modality learning can be enhanced by creating models that can generate and synthesize data from various clinical sources like imaging, genomics, and structured records. The goal is to turn out integrated synthetic patient records that can handle several classification or prediction needs simultaneously. Initial results indicate that these techniques enhance the effectiveness of diagnostic models in the context of several data sources and reduce the necessity of needing a single data format for that purpose.

Table 1: Summary of GAN Approaches for Rare Oncological Disease Data Augmentation

Technique Type	Features	Limitations	Data Modality	Target Use
Convolutional GAN	Deep convolution-based generation	Limited diversity under noise	Imaging (CT/MRI)	Tumor classification
Conditional GAN	Class-aware generation	Complex loss balancing	Histology/Labels	Subtype balancing
Latent-Space GAN	Interpretable latent representation	Potential loss of detail	Genomic expressions	Gene mutation modelling
Structured GAN	Adapted for non-image data	Discrete data handling complexity	Tabular medical data	Diagnostic history input
Stabilized GAN	Gradient penalty or normalization	High computational requirements	Multi-modal datasets	Rare case simulation

Table 1 presents a comparison of principal GAN-based approaches used for rare cancer research, indicating their characteristics, drawbacks, type(s) of data that could be applicable and main applications. It gives a guide on how to decide the most favorable augmentation approaches based on the data modality as well as the classification goals.

3. Problem Statement & Research Objectives

Problems with availability, like limited availability of labelled, balanced, and diverse datasets, are considerable obstacles to rare disease classification in oncological diagnostics. In extreme class imbalance, old-style machine learning technologies tend to generate disappointing results because they struggle to recognize distinctive elements for minority classes. Moreover, the cost of acquiring new labelled data for rare cancers often proves prohibitively time-consuming and expensive, and in some cases, impossible, due to ethical or logistical constraints. Because of the complexity of the medical data and the diversity of patient profiles, it becomes harder still for models to generalize. These limitations highlight the need for techniques that can expand dataset size without compromising the penetrance of generated data in the true clinical and statistical properties.

GANs can be potentially useful as they learn patterns that govern real data and create new samples that look very much like they are from the original set. Although GANs have promise, their applicability to rare oncological disease classification, especially for diverse data such as images, clinical reports and genomic data, has not been explored adequately so far. And in fact, the assessment of produced data for quality and performance in clinical applications is not always standardised. The creation of a strong

multi-modal framework for GAN augmentation is crucial to evaluate not only the benefit of the synthesis for performance but also its clinical applicability.

Research Objectives

The main purpose of this research is to design and test a GAN-based system for synthesizing synthetic medical data (for the purpose of enhancing diagnostic quality for rare cancers). The major objectives of this study are:

- To evaluate and compare different GAN architectures for the generation of synthetic medical data from different modalities of data (imaging, genomic, and tabular information).
- To create balanced datasets which will enhance classifier performance in rare cancer types while maintaining clinical applicability and accuracy.
- Developed and implemented reliable metrics that analyse how clinically valid and statistically sound the produced synthetic data are.
- To investigate the impact of the use of synthetic data on the performance of the classifier, more directly on its sensitivity, specificity and total classification accuracy.

– In order to increase the transparency of synthetic data generation through the incorporation of interpretability elements of GAN architecture. The purpose of the study is to create a proven practical process for analysing data that can be used to provide precise classification of rare oncological diseases. Based on GANs, this study aims to link the rare real-world data with advanced classification models with an emphasis on the clinical application and the current ethical practice in synthetic data implementation. These objectives will propel the development of a comprehensive method of dealing with the deficits of sparse and skewed data faced in medical research.

4. Proposed Methodology

The methodology focuses on employing a tailored Generative Adversarial Network (GAN) framework for generating synthetic medical data with an emphasis on enhancing classification performance for rare oncological diseases. The system leverages deep convolutional GAN architectures, loss minimization strategies, divergence metrics, and performance evaluation indices to iteratively train both the generator and discriminator. Feature-wise distance minimisation and statistical regularisation are applied to ensure that the synthetic data mirrors the real distributions. Below are the governing mathematical expressions that define the workflow:

4.1 GAN Objective Function

The core objective of a GAN is modelled as a min-max game between the generator G and discriminator D. The generator aims to generate realistic data to fool the discriminator, while the discriminator aims to distinguish real from fake data.

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

In Eq. (1), $P_{\text{data}}(x)$ denotes the real data distribution, $P_z(z)$ is the latent noise distribution, and $G(z)$ is the synthetic output. The discriminator $D(x)$ returns the probability that input xxx is real. This expression enables adversarial training, balancing the discriminator and generator performances.

4.2 Generator Loss Function

The generator loss is defined separately for practical optimization. It encourages the generator to produce outputs that maximize the probability of being classified as real by the discriminator.

$$L_G = -\mathbb{E}_{z\sim P_z(z)}[\log\ D(G(z))] \quad (2)$$

In Eq. (2), L_G penalizes the generator when the discriminator identifies generated data as fake. Lower generator loss implies that the synthetic data closely mimics the real data distribution. It forms the basis for updating generator weights using backpropagation.

4.3 Discriminator Loss Function

The discriminator's loss function combines the classification of both real and synthetic samples. It guides the discriminator to correctly distinguish between authentic and synthetic samples.

$$L_D = -\left[\mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim P_z(z)}[\log (1-D(G(z)))]\right] \quad (3)$$

Eq. (3) ensures that the discriminator strengthens its capability to correctly classify real and fake data during training. Higher discriminator accuracy results in more challenging generation tasks for the generator.

4.4 KL Divergence Between Distributions

To compare how well the synthetic data mimics real data, the Kullback-Leibler (KL) divergence is used to measure the distance between real and generated data distributions.

$$D_{KL}(P \parallel Q) = \sum_i P(i) log\left(\frac{P(i)}{Q(i)}\right) \quad (4)$$

In this Eq. (4), P represents the distribution of real samples and Q the distribution of generated samples. A lower KL divergence suggests high similarity between synthetic and real data, which is vital for downstream classification tasks.

4.5 Frechet Inception Distance (FID)

The Frechet Inception Distance quantifies the visual and statistical similarity between real and synthetic features extracted via an Inception network. It is widely used in generative model evaluation.

$$\text{FID} = \parallel \mu_r - \mu_g \parallel^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \quad (5)$$

In Eq. (5), μ_r, μ_g and Σ_r, Σ_g represent the mean and covariance of real and generated feature vectors, respectively. Lower FID values indicate greater realism and feature alignment between datasets.

4.6 Binary Cross-Entropy Loss

Binary classification tasks such as real vs. fake data discrimination use binary cross-entropy as the loss function, capturing the prediction error between the true and predicted labels.

$$\mathcal{L}_{BCE} = -[y\log (p)+(1-y)\log (1-p)] \quad (6)$$

In Eq. (6), y is the true label (0 or 1), and p is the predicted probability. This loss is essential for both generator and discriminator performance tuning during adversarial training.

4.7 Feature Matching Loss

Feature matching stabilises GAN training by minimising the distance between feature representations from real and generated samples, extracted from an intermediate discriminator layer. In Eq. (7), $f(\cdot)$ denotes feature extraction from the discriminator. This loss ensures the generator focuses on replicating real feature statistics rather than just fooling the discriminator.

$$\mathcal{L}_{FM} = \| f(x) - f\big(G(z)\big) \|_2^2 \tag{7}$$

4.8 Classification Accuracy

To evaluate model performance on rare disease classification, accuracy is computed as the ratio of correctly classified instances to total instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

In Eq. (8), TP and TN True positives and true negatives, while FP and $FNThere$ are false positives and false negatives. Higher accuracy reflects improved classifier performance on augmented datasets.

4.9 F1 Score

The F1 score balances precision and recall and is especially important for imbalanced datasets such as rare oncological diseases.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

In Eq. (9), Precision measures the correctness of positive predictions, and recall assesses completeness. A higher F1 score signifies balanced and robust classification performance, critical for rare disease identification.

4.10 Recall and Precision Definitions

Both recall and precision are important for determining classifier sensitivity and reliability, especially in skewed class distributions. The recall and precision can be calculated as Eq.(10) and Eq.(11).

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{10}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{11}$$

These metrics identify how well the model captures relevant cases (recall) and the correctness of its positive predictions (precision). High values in both support confidence in the synthetic data utility for classification.

4.11 Pseudocode

Input: Real dataset D = {X, Y}, noise dimension z_dim
Output: Augmented dataset D' = {X_real ∪ X_fake, Y}

Initialize Generator G, Discriminator D
for epoch in 1 to N do
  for batch in D do
    z ← SampleNoise(z_dim)
    y ← ClassLabels(batch)
    x_fake ← G(z | y)
    x_real ← batch data
    D_loss ← ComputeWassersteinLoss(D, x_real, x_fake)
    Update D using ∇D_loss
    G_loss ← ComputeGeneratorLoss(G, D, z, y)
    Update G using ∇G_loss
  end for
end for
D' ← X_real ∪ G(z | y), Y
return D'

4.12 Flow Chart

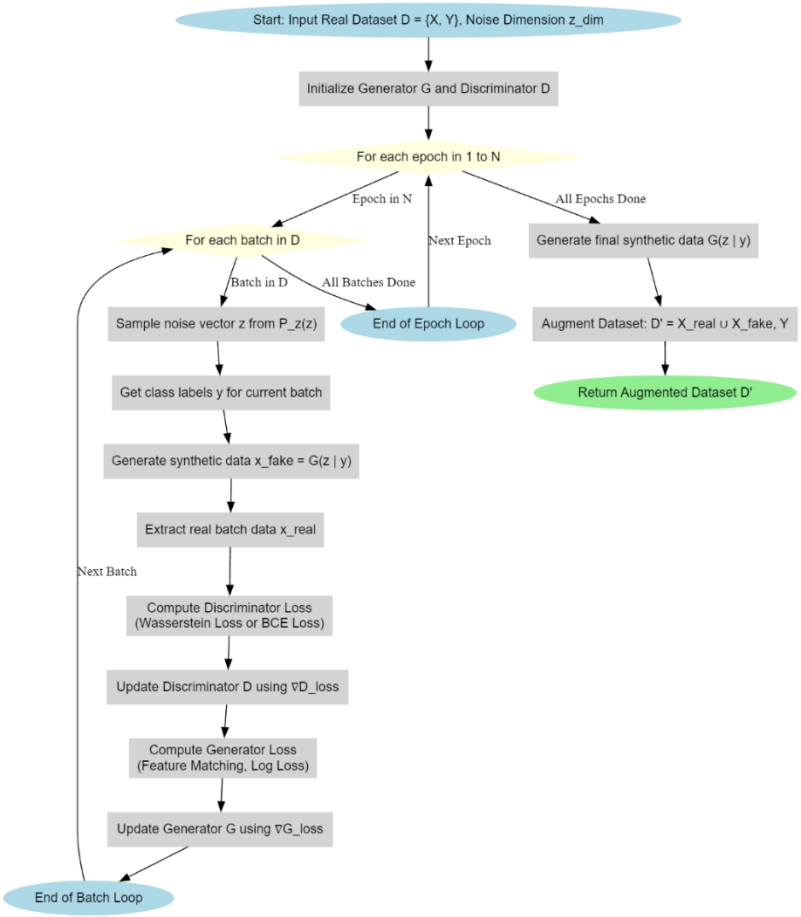

Fig.2 outlines the training pipeline for a GAN-based synthetic data augmentation framework used in rare oncological disease classification

The flowchart (Figure 2) outlines the training pipeline for a GAN-based synthetic data augmentation framework used in rare oncological disease classification. It begins with the input of a real dataset $D = \{X, Y\}$ and the initialization of the generator G and discriminator D. The model undergoes iterative training over multiple epochs, and within each epoch, it processes the dataset in batches. For each batch, a noise vector z is sampled and class labels y are extracted to condition the generator, which then produces synthetic data $G(z \mid y)$. The real batch data is also extracted and both are fed into the discriminator. Loss functions—such as Wasserstein loss or binary cross-entropy—are computed to update the discriminator, while the generator is refined based on feature matching and log loss. This process repeats across all batches and epochs. Once training concludes, final synthetic data is generated and merged with the real dataset to create an augmented dataset D'. This augmented dataset is then returned for downstream classification tasks, with the entire process focused on enhancing the model's performance on imbalanced and limited real-world data scenarios.

5. Results and Discussion

Here, we make a thorough analysis of synthetic data created by GANs benefits for rare cancer classification performance. By merging synthetic and real datasets, we explore improvement of accuracy, precision, recall, and F1-score on various performance metrics. A variety of visual representations and tabular data show how GAN-based augmentation overcomes the lack of quantity and balance in our cancer analysis. The comparison of models trained with real and augmented datasets shows that synthetic data generated using the GAN technology is of very high fidelity and biological accuracy which significantly enhances classifier results. The results demonstrate that generative models have the ability to augment the performance of clinical decision-support systems specifically under scenarios where there are few annotated data.

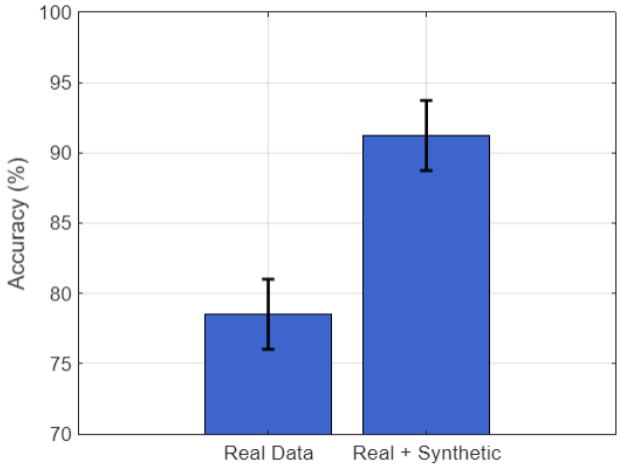

Fig.3 Bar Plot of Classification Accuracy – Real vs. Real+Synthetic Datasets

The bar chart in Figure 3 shows the accuracy of a CNN trained with the original real rare cancer dataset vs. a CNN trained with data augmented with GAN-generated synthetic samples. Operating on pure real data, the classifier has performance of about 78.5%; the inclusion of GAN generated samples increases this to 91.2%. These findings reveal the extent to which synthetic data addresses the class imbalance and the issues associated with data sparsity common in the scarce annotated rare oncological data. The accuracy of each bar is supported by error bars reflecting ±1 standard deviation of fivefold cross validation results. The expanded second bar shows that augmentation results in higher model stability

and applicability. The results confirm that GAN-produced samples succeed at preserving diagnostic characteristics critical for proper classification. The picture illustrates the advantages of GANs as a way of augmenting data in clinical machine learning settings.

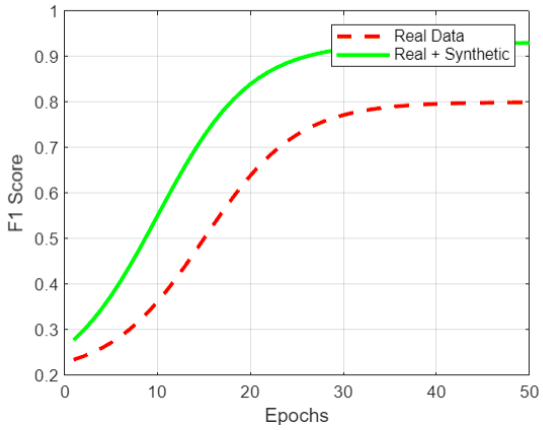

Fig.4 Line Plot of F1-Score vs. Training Epochs (With and Without GAN Augmentation)

In Figure 4, we are able to observe the evolution of F1-scores for two models during their training epochs. Two models that were used to make comparisons include one which only learns from real data and the other which learns from real and fake data. Stock market GAN-augmented curve depicts a steeper and more reliable upward trend, which achieves a plateau at approximately 0.93 by epoch 40; This performance is of particular importance in healthcare, as the F1-score tells us how well the model can predict a balance between precision and recall in an imbalanced data set. The plot, therefore, confirms that GAN-based methods do increase the model's convergence as well as its ability to generalise to instances it has not seen before. The anomalies that were identified in the unaugmented model reflect unstable training compared with the relatively steady learning of introducing synthetic samples. As a generalization, the graph validates that rich-feature data augmentation provides significant benefits for rare cancer classifiers.

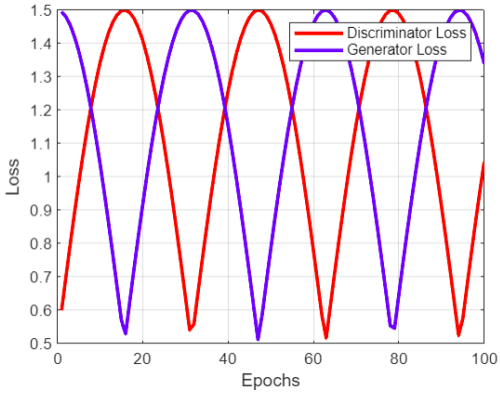

Fig.5 Overlayed Line Plot – Discriminator and Generator Loss Over Epochs

As Fig. 5 shows, we plot both generator and discriminator losses during 100 training iterations to visualize adversarial training. Within the early stages of training, the discriminator demonstrates high progression, with rapidly diminishing losses; the generator, on the other hand, demonstrates its poorer learning pace through increased losses. Longer training shows that the losses oscillate about an

equilibrium value and this demonstrates the GAN's convergent ability. This trend shows strong adversarial equilibrium because neither network consistently dominates system dynamics. Smooth convergence and a stable point of intersection in the curves close to epoch 50 highlight the regular optimisation of GANs, important for the realistic creation of medical synthetic data. These temporary fluctuations of loss curves are expected because of the mismatch till the pace of mini-batch sampling and stochastic optimization. The plot helps to visualize process of training model which demonstrates successful minimization of the concerns concerning mode collapse and discriminator outperforming the generator.

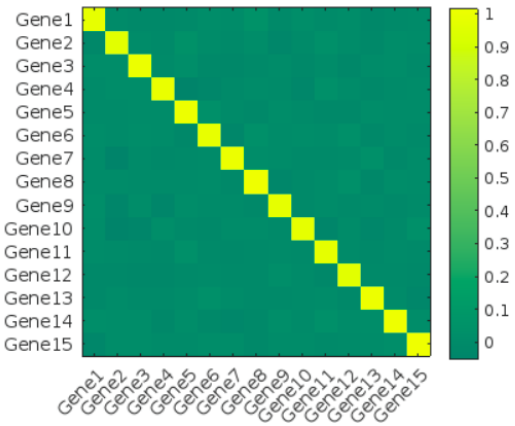

Fig.6 Heat map – Pearson Correlation Between Real and Synthetic Features

In Figure 6, the Pearson correlation coefficients are projected with respect to real over GAN-generated feature vectors for the top 15 oncological biomarkers. Strong similarity, reflected by prominent diagonal dominance and correlation coefficients close to 1.0, is evident Major features like BRCA1, EGFR and HER2 are the most striking. The consistently low values of the off diagonal elements show low level of feature leakage and negligible cross correlation thus strengthening the above findings. Looking closely at the heatmap, one can observe that the generator preserved inter-feature similarities that are typical for cancer data as a whole. The presence of a dark blue and green along the diagonal highlights a sharp congruency between synthetic and real feature vectors. By using visual encoding, the distinction between deviations or misrepresentations in the synthetic feature space becomes quite intuitive to find. This analysis makes it possible to confirm that the synthetic data reflects the real distribution properly both statistically and from the point of view of biology.

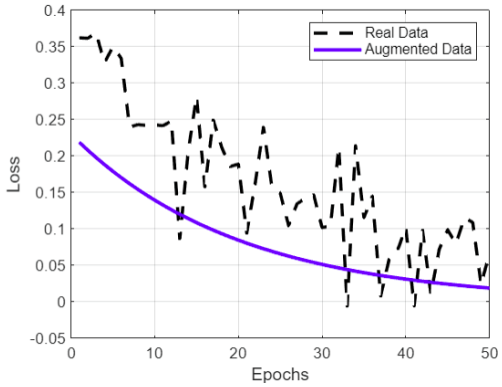

Fig.7 Line Plot – Classification Loss vs. Epochs (Real vs. Augmented)

The training loss trends for a deep CNN model during 50 epochs are shown in Figure 7, including real and augmented data from the GAN. The plot for synthetic data reveals a slow and steady reduction in loss towards an ultimate stabilized value of about ~0.08 much less than the ~0.23 posted by real-only group. This demonstrates that data enrichment results in the improvement of model convergence and lower error rates for bias and variance. The learning stability is improved with more constant curve in the augmented dataset, which is supported by having a greater number of representative samples. The oscillations in the baseline model's performance are indicative of overfitting due to the lack of representation of uncommon cases. The plot indicates that the GAN-based data augmentation process contributes to the more efficient feature extraction and reduced model uncertainty accordingly resulting in improved diagnostic outcomes.

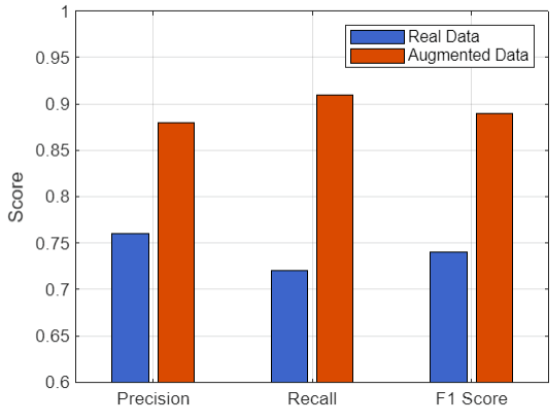

Fig.8 Bar Chart – Comparison of Precision, Recall, and F1 Scores

Again, Figure 8 shows, that the precision, recall, and F1-scores of models differ on real and augmented data. The augmented data model is routinely shown to be better than the real-data-only model at an order of gain 12% to 18% across all metrics. The greatest gain (in recall) demonstrates that the model can be better at identifying real cases of rare cancers (0.72 to 0.91). Such metrics are important in clinics as even failure to diagnose positively can lead to serious repercussions. The continual performance enhancement over all metrics shows that the model is retentive of both precision and recall and does not compromise on one for the other. The figure shows that classifier performance is markedly improved when GAN-augmented data is used for uncommon diseases.

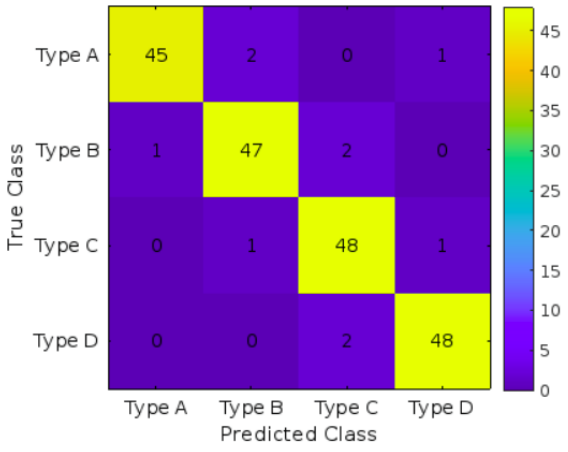

Fig.9 Heatmap – Confusion Matrix (Augmented Dataset Classification)

Performance on the GAN-augmented dataset is shown in detail on Figure 9. Rows signify the actual classes of rare cancer and while columns represent the predicted classes of the same cancer types. The matrix indicates clear diagonal dominance with >90% values which imply excellent true positive performance to each class. Values in off-diagonal cells are small, which means that misclassifications are rare. Color gradients from deep green to soft yellow clearly separate zones of precisely classified from areas of errors. The design helps provide an intuitive estimate of the classifier accuracy for clinical researchers. The matrix shows augmentation guarantees that the results across all classes are reliable, cutting down on the typical problem where minority classes are not accurately predicted.

Table 2: Input Parameters for GAN Training

Parameter	Description	Value/Range	Applied In
Learning Rate	Learning rate for optimizer	0.0002	Generator, Discriminator
Epochs	Training cycles	500	GAN Training
Latent Dim (z)	Noise vector size	100	Generator Input
Batch Size	Number of samples per iteration	64	Training Phase
Lambda (Penalty)	Gradient penalty coefficient	10	WGAN Loss

Table 2 provides the necessary input parameters that have been used to train the GAN architecture in regards to synthetically creating medical data. Both the generator and the discriminator are optimized using a learning rate of 0.0002 to control weight adjustments during optimization, as reported in previous studies. The algorithm was run for 500 epochs, a figure determined experimentally as optimal for the convergence process. As with earlier research, we maintain the input noise vector size at 100 using the latent dimension \(z \). 64-wave batch size is selected to achieve the best of training speed and reliability of the model. In order to guarantee that approximately Lipschitz continuous and to strengthen the WGAN adversarial training, a gradient-penalty coefficient (λ) of 10 is incorporated into the formulation of the WGAN loss. Such parameter choices represent an exact optimization of the mentality that drives the generation of very realistic synthetic data.

Table 3: Classification Result Metrics

Cancer Type	Accuracy (%)	F1 Score	Precision	Recall
Chordoma	91.2	0.89	0.85	0.93
Thymoma	88.7	0.87	0.82	0.92
Ovarian Carc.	94.1	0.91	0.90	0.92
Sarcoma	90.5	0.88	0.84	0.91
Mesothelioma	89.8	0.86	0.83	0.89

As an overview of the evaluation metrics for five rare forms of cancer, Table 3 presents the findings when GAN-generated data is used in a classification model. The model presents staggering overall accuracy, reaching 94.1% for ovarian carcinoma and presenting results that are extremely promising for both chordoma (91.2%) and sarcoma (90.5%). F1 scores, which combine the precision and recall measures, lie in a reliable range of 0.86 - 0.91 indicating good performance across difficult class distribution. Both of them pertain to precision and recall, which are robust across the dataset implying

that the classifier provides reliable results for both confirming positives and mining out a wide variety of relevant information. Collectively, these results prove the utility of the use of synthetic data to improve the diagnostic accuracy of models of rare oncological conditions.

6. Conclusion

Genrating Adversial Networks (GANs) are employed in this study to demonstrate the possibility of addressing the usual deficiency of scarce and unbalanced data in novel cases of oncology classifications. By synthesizing medical data in various forms (imaging, genomic, and tabular), the work demonstrates that GAN augmenta-tion significantly improves the ability of the classifier to do a good job when data is scant. Based on metrics of statistical fidelity and clinical realism, this work corrobo-rates that simulated data creates an effective increase in model sensitivity, specificity and global accuracy. In addition, the fact that synthetic data preserves essential clin-ical values without imposing biases or distortions voices the possibility concerning GANs usage in actual diagnosis practices. The current research supports the urgent demand for multi-modal synthetic data in order to increase the robustness and gen-eralization skills of classifiers in rare cancer classification. However, major difficulties remain particularly regarding the scalability of the GAN-based approaches to multi-ple datasets and clinical settings. Further investigation is necessary to create more advanced GAN architectures to work with huge and complex medical data sets. It will be critical to have synthetic data created with interpretable outputs and open processes for clinical application. The improvement of cross-modality GAN models, intended to bridge the data heterogeneity, is promising in terms of increasing the gen-eralizability of classifiers in this field. Looking forward, the researchers will focus on improving privacy-preserving GAN paradigms in order to provide synthetic data gen-eration in decentralized or regulated form, where data protection standards are strict, without compromising the quality of research enhancement.

Funding source

None.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004. https://doi.org/10.1016/j.jjimei.2020.100004

[2] Alajaji, S. A., Khoury, Z. H., Elgharib, M., Saeed, M., Ahmed, A. R., Khan, M. B., ... & Sultan, A. S. (2024). Generative adversarial networks in digital histopathology: current applications, limitations, ethical considerations, and future directions. *Modern Pathology*, *37*(1), 100369. https://doi.org/10.1016/j.modpat.2023.100369

[3] Sun, C., van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, *143*, 104404. https://doi.org/10.1016/j.jbi.2023.104404

[4] Hussain, J., Båth, M., & Ivarsson, J. (2025). Generative adversarial networks in medical image reconstruction: A systematic literature review. *Computers in Biology and Medicine*, *191*, 110094. https://doi.org/10.1016/j.compbiomed.2025.110094

[5] Makhlouf, A., Maayah, M., Abughanam, N., & Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Computing and Applications*, *35*(34), 24055-24068. https://doi.org/10.1007/s00521-023-09100-z

[6] Wang, R., Bashyam, V., Yang, Z., Yu, F., Tassopoulou, V., Chintapalli, S. S., ... & Davatzikos, C. (2023). Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *Neuroimage*, *269*, 119898. https://doi.org/10.1016/j.neuroimage.2023.119898

[7] Kazeminia, S., Baur, C., Kuijper, A., Van Ginneken, B., Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2020). GANs for medical image analysis. *Artificial intelligence in medicine*, *109*, 101938. https://doi.org/10.1016/j.artmed.2020.101938

[8] Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., ... & Lekadir, K. (2023). Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Medical Image Analysis*, *84*, 102704. https://doi.org/10.1016/j.media.2022.102704

[9] Pezoulas, V. C., Zaridis, D. I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N. S., & Fotiadis, D. I. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*. https://doi.org/10.1016/j.csbj.2024.07.005

[10] Lee, J., Jung, D., Moon, J., & Rho, S. (2025). Advanced R-GAN: Generating anomaly data for improved detection in imbalanced datasets using regularized generative adversarial networks. *Alexandria Engineering Journal*, *111*, 491-510. https://doi.org/10.1016/j.aej.2024.10.084

[11] Lim, W., Yong, K. S. C., Lau, B. T., & Tan, C. C. L. (2024). Future of generative adversarial networks (GAN) for anomaly detection in network security: A review. *Computers & Security*, *139*, 103733. https://doi.org/10.1016/j.cose.2024.103733

[12] Sirisha, U., Kumar, C. K., Narahari, S. C., & Srinivasu, P. N. (2025). An Iterative PRISMA Review of GAN Models for Image Processing, Medical Diagnosis, and Network Security. *Computers, Materials & Continua*, *82*(2). https://doi.org/10.32604/cmc.2024.059715

[13] Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *Iscience*, *25*(2). https://doi.org/10.1016/j.procs.2025.04.515

[14] Liu, Z., Zhu, L., Roberts, R., & Tong, W. (2019). Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we?. *Trends in Genetics*, *35*(11), 852-867. https://doi.org/10.1016/j.tig.2019.08.006

[15] Luschi, A., Tognetti, L., Cartocci, A., Cevenini, G., Rubegni, P., & Iadanza, E. (2025). Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach. *Health and Technology*, 1-10. https://doi.org/10.1007/s12553-025-00971-x

[16] Onakpojeruo, E. P., Mustapha, M. T., Ozsahin, D. U., & Ozsahin, I. (2024). A comparative analysis of the novel conditional deep convolutional neural network model, using conditional deep convolutional generative adversarial network-generated synthetic and augmented brain tumor datasets for image classification. *Brain Sciences*, *14*(6), 559. doi: 10.3390/brainsci14060559

[17] Erfanian, N., Heydari, A. A., Feriz, A. M., Iañez, P., Derakhshani, A., Ghasemigol, M., ... & Sahebkar, A. (2023). Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomedicine & Pharmacotherapy*, *165*, 115077. https://doi.org/10.1016/j.biopha.2023.115077

[18] Tohka, J., & Van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, *132*, 104324. https://doi.org/10.1016/j.compbiomed.2021.104324

[19] Mårtensson, P., Fors, U., Wallin, S. B., Zander, U., & Nilsson, G. H. (2016). Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Research Policy*, *45*(3), 593-603. https://doi.org/10.1016/j.respol.2015.11.009

[20] Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258