IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2023, accepted 7 January 2024, date of publication 16 January 2024,
date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354826

RESEARCH ARTICLE

Enhancing the Quality and Authenticity of Synthetic Mammogram Images for Improved Breast Cancer Detection

DILAWAR SHAH[1], (Member, IEEE), MOHAMMAD ASMAT ULLAH KHAN[1], (Member, IEEE), MOHAMMAD ABRAR[2], (Member, IEEE), FARHAN AMIN[3], (Member, IEEE), BADER FAHAD ALKHAMEES[4], (Member, IEEE), AND HUSSAIN ALSALMAN[5], (Member, IEEE)

[1]Department of Computer Science, International Islamic University, Islamabad, Islamabad 44000, Pakistan
[2]Department of Computer Science, Bacha Khan University, Chārsadda 24420, Pakistan
[3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
[4]Department of Information Systems, College of Computer and Information Science, King Saud University, Riyadh 11543, Saudi Arabia
[5]Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding authors: Farhan Amin (farhanamin10@hotmail.com) and Hussain AlSalman (halsalman@ksu.edu.sa)

This research was supported by the Researchers Supporting Project number (RSP2024R244), King Saud University, Riyadh, Saudi Arabia.

ABSTRACT Breast cancer is widespread throughout the world and can be cured if diagnosed early. Mammography is an irreplaceable and critical technique in modern medicine that serves as a foundation for the detection of breast cancer. In medical imaging, the reliability of synthetic mammogram images is produced by deep convolutional generative adversarial networks (DCGAN). Human validation to assess the quality of synthetic images to examine and calculate the perceptual variations between synthetic images and their real-world counterparts is a difficult task. Thus, this research focused on improving the quality and authenticity of synthetic mammogram images. For this, we explored and identified a new research gap because radiologists consistently expressed much higher confidence levels in real mammogram images in their assessment process. This research highlights the key difference between synthetic and real mammograms by defining mean scores. The defined mean identifies a large gap, with real mammographic images receiving an average score of 0.73 and a synthetic score of 0.31. A statistical analysis was performed, which produced a T-statistic of -6.35, a p-value less than 0.001, and a 95% confidence interval ranging from -0.50 to -0.28. These results have a wide range of implications. It emphasizes the urgent need for further improvements in the generative model, improving the legitimacy and caliber of synthetic mammogram images. Our research highlights how crucial it is to incorporate synthetic images into clinical practice with caution and thought. Ethical considerations must encompass the potential consequences of relying on synthetic data in medical decision-making, along with concerns related to diagnostic accuracy and patient safety.

INDEX TERMS Breast cancer, computer-aided diagnosis, deep learning, generative models, medical imaging, medical diagnosis, synthetic images.

I. INTRODUCTION

Breast cancer is one of the most common diseases among women, and it is considered a deadly type of cancer worldwide. Mammography is an irreplaceable and critical technique in modern medicine, serving as a cornerstone for breast cancer detection. Breast cancer remains a significant global health concern, being a predominant cancer among women and posing substantial mortality and morbidity [1]. The essence of mammography lies in its ability to offer detailed images of the breast tissue, utilizing low-dose ionizing radiation to pinpoint abnormalities and irregularities

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Ji.

potentially indicative of cancerous cells [2]. The vital role of mammography is underscored by its capacity for early detection, which is crucial in identifying breast cancer at a stage where it is more treatable and the chances for recovery are heightened [3]. Early intervention facilitated by mammography potentially curtails the progression and metastasis of the disease, thereby reducing fatality rates and improving the quality of life post-diagnosis. This imaging technique is pivotal in contemporary medical practice, shaping the trajectory of patient care in oncology, particularly in enabling personalized treatment approaches based on individual diagnosis. The advancements in mammographic technologies and their assimilation into healthcare systems underscore the ongoing commitment to combating breast cancer. The integration of DCGAN within medical imaging has emerged as a groundbreaking development, enhancing the synthesis of high-fidelity medical images [4]. DCGANs, an evolved form of Generative Adversarial Networks (GANs), employ advanced convolutional networks, imitating and generating images that closely resemble authentic medical images [5].

The application of DCGAN in medical imaging represents a transformative approach to data augmentation, contributing to developing robust and sophisticated diagnostic models. The capability of DCGAN to produce synthetic yet realistic images has been pivotal in situations where the availability of extensive, diverse, and annotated real datasets is a challenge. The produced synthetic images are instrumental in training and refining machine learning models, enhancing diagnostic precision and reliability [6].

This convergence of advanced computational techniques with medical imaging fosters an era of healthcare technology innovation. It is paving the way for refined diagnostic procedures and enriching research avenues. It is positioning itself as a fundamental component in the evolving medical imaging and diagnostics landscape, potentially reshaping clinical approaches and healthcare delivery. A general architecture of DCGAN is presented in Figure 1.

The proliferation of synthetic medical images, mainly through advanced techniques like DCGAN, underscores the critical need for meticulous validation processes [7]. While synthetic images present valuable resources for training and validating diagnostic algorithms due to their capability to augment and diversify available datasets, the reliability and credibility of these images are contingent upon rigorous validation [8]. Without proper validation, integrating synthetic images into medical research and clinical practices can propagate inaccuracies and biases in diagnostic models, potentially compromising patient outcomes and care.

Validation of synthetic images is vital to uphold the highest medical imaging standards, including maintaining anatomical accuracy, texture, and pathology representation, which is paramount in fostering robust and reliable diagnostic tools [9]. The validation serves as a checkpoint to authenticate the integrity and authenticity of synthetic images, substantiating their applicability and relevancy in medical research and diagnosis.

Expert assessment in validating synthetic medical images emerges as a crucial component, providing an in-depth, nuanced evaluation of the images' quality, realism, and clinical applicability [10]. Radiologists and medical imaging experts possess the essential knowledge and experience to discern subtle differences and anomalies in images that may not be detectable through automated validation methods.

The input from radiologists and imaging experts is invaluable in ascertaining the extent to which synthetic images replicate the intricate details and variations present in real medical images [11]. Their assessments yield insights into the clinical relevance of synthetic images, ensuring that they are representative and accurate and maintain the integrity required for effective medical research and diagnosis.

Expert evaluation fortifies the validation process and aligns synthetic image generation techniques with clinical insights, fostering advancements in medical imaging grounded in real-world applicability and medical knowledge. This alignment is pivotal in bridging the gap between technological innovations and clinical needs, ensuring the evolution of medical imaging is congruent with the overarching goals of enhancing diagnosis and patient care.

The primary objective of this study is to rigorously validate the mammograms generated by DCGAN to ascertain their authenticity, quality, and clinical relevance. Achieving a prominent level of accuracy and reliability in synthetic mammograms is critical to ensuring their efficacy as training data and their potential contribution to enhancing diagnostic models [12]. Through meticulous validation, this study aims to affirm that DCGAN-generated mammograms maintain the essential characteristics and intricacies of real mammograms, including anatomical accuracy, texture representation, and pathological variability [9]. This objective is central to establishing the reliability and applicability of synthetic mammograms in advancing medical research and improving breast cancer diagnosis and detection. The problem lies in the current limitations of GANs in producing synthetic mammogram images that match the quality and diagnostic reliability of real mammograms. Despite technological advances, there is a noticeable disparity in how radiologists perceive and evaluate synthetic images compared to the real ones. This points to the need for improved algorithms and methodologies in GANs to produce more accurate and reliable synthetic mammogram images, crucial for the effective detection of breast cancer.

The secondary objective of this study is to assess the validity of the DCGAN-generated mammograms through expert evaluation. It refers to the extent to which synthetic images are indistinguishable from real images, as perceived by human observers, especially those with expertise in medical imaging [13]. Utilizing the insights of radiologists, the study aims to discern the degree of realism in synthetic mammograms and identify any discernible discrepancies or anomalies compared to real mammograms. This assessment is pivotal to ensuring that DCGAN-generated mammograms not only conform to objective quality metrics

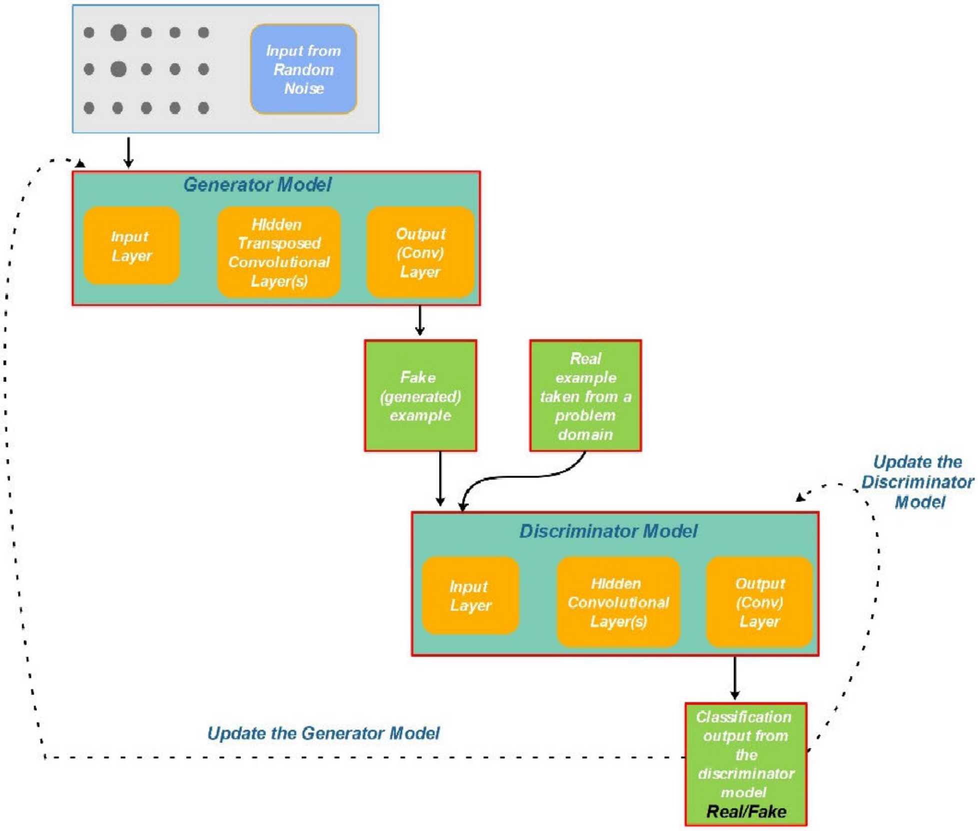


FIGURE 1. General architecture of DCGA.

but also resonate with medical experts' experiential and contextual understanding, thus bridging the divide between computational accuracy and human perception in medical imaging [10].

Section II presents recent literature on DCGANs and the validation of synthetic images by automated methods and radiologists. In section III, the overall methodology is presented. Section IV elaborates on results and discussions. Conclusion and Future work are discussed in section V.

II. LITERATURE REVIEW

GANs are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by two neural networks contesting each other in a game [12]. The foundational principles of GANs revolve around the interaction between a generator responsible for creating data and a discriminator tasked with distinguishing between generated and real data. The generator receives a random noise as input and produces data that aspire to be indistinguishable from real data. Simultaneously, the discriminator evaluates the received data and determines whether it is generated or real [14]. The constant competition between the generator and the discriminator fosters an iterative learning process, enhancing the generator's ability to create high-quality synthetic data and the discriminator's proficiency in distinguishing between synthetic and real data.

A. DCGANs IN MEDICAL IMAGING

DCGANs have marked a significant stride in medical imaging, generating high-quality and detailed synthetic medical images [4]. DCGANs have found applications in enhancing the training of diagnostic models by augmenting datasets, especially in situations where acquiring a large amount of real medical imaging data is impractical due to privacy and availability concerns [9].

Recent advancements in DCGANs have enabled the synthesis of medical images with remarkable anatomical accuracy and texture representation, allowing for a more diverse and robust training environment for machine-learning models in medical diagnostics [15]. These advancements have proven instrumental in developing improved diagnostic tools and have facilitated research endeavors to decipher intricate medical conditions, enhancing the overall understanding and approach toward medical diagnostics and treatment.

B. VALIDATION STUDIES

Validation is essential for the robustness and reliability of medical images generated through synthetic means. Radiologists apply their experience and knowledge to distinguish the quality, accuracy, and, more specifically, relevance and applicability of the generated images in clinical settings [16].

For synthetic images, it is necessary to scrutinize them for their anatomical correctness. Looking for inconsistencies and any possible distortion is an essential part of this scrutiny process. The expertise of clinicians is instrumental in affirming the reliability and authenticity of synthetic images, ensuring they align seamlessly with the clinical practices they aim to portray [17].

Discerning intricate features and appraising synthetic images' visual and diagnostic quality require special attention and training from radiologists. Their active involvement fosters a comprehensive grasp of the significance and relevance of synthetic images. This engagement guarantees that synthetic images closely mirror the intricate clinical scenarios they seek to represent, thus enhancing their practical relevance in healthcare [18].

It is a common practice in expert-based validation to juxtapose synthetic and real images for comparative analysis. The side-by-side comparison in this type of validation is a valuable tool for identifying disparities, evaluating texture congruence, and verifying the alignment of anatomical and pathological aspects within synthetic representations [19]. Automated validation techniques encompass computational methods tailored for assessing synthetic medical images. They achieve this by comparing these images against predefined metrics and standards. This approach provides a statistical and structural assessment, enabling the evaluation of image similarity and authenticity [13].

While automated validation methods offer the advantages of scalability and objectivity, they may not fully grasp the clinical relevance and context inherent to medical images. Hence, a combined approach, integrating both expert assessments and automated evaluations, is crucial for a holistic and robust validation process [20].

III. METHODOLOGY

The methodology employed in this study is designed to rigorously assess the validity of DCGAN-generated mammogram images compared to their real counterparts.

A. IMAGE GENERATION WITH DCGAN

DCGANs have shown potential in generating high-quality synthetic images across various domains, including medical imaging, by learning to mimic original data distribution. Proper data preparation is fundamental to the model's performance in any machine learning model training. This study used a benchmark dataset, the Digital Database for Screening Mammography (DDSM) [21]. The generator part of the DCGAN starts with a random noise vector as input, simulating a latent space. This model aims to generate new data similar to the expected output. It has an input layer, several hidden transposed convolutional layers that up-sample the input, and an output layer that produces the fake example. The output from the generator model is a synthetic data instance that attempts to mimic real data.

The discriminator model receives both fake examples from the generator and real examples from the dataset. The discriminator also has an input layer, hidden convolutional layers for feature extraction, and an output layer that gives a probability of the input being real or fake. Based on the discriminator's performance, its parameters are updated to improve its accuracy in distinguishing real data from fake data. Using the feedback from the discriminator, the generator updates its parameters to produce more convincing fake examples. The dotted lines indicate the feedback loop where the generator and discriminator are continuously competing, improving each other's performance over iterations.

The mammograms were first denoised and resized into the same size and format to feed data to the network. Table 1 shows the steps followed and their role.

1) TRAINING AND GENERATION PROCESS

The training and generation process of DCGAN involves continuous learning through two competing networks: the generator and the discriminator. Creates synthetic images from random noise, progressively enhancing the quality and authenticity of generated images through iterative training. Assesses the images, distinguishing between real and synthetic samples, thus guiding the generator toward producing more realistic images. Our model's training process is implemented with a specific set of hyperparameters for optimal performance. The learning rates for the generator and discriminator are set at 0.0002, ensuring a stable and consistent learning pace. We used a batch size of 64, which balances computational efficiency with the ability to generalize across the dataset. The model was trained over 100 epochs to ensure adequate learning without overfitting. The Adam optimizer is utilized for its efficiency in handling sparse gradients and adapting the learning rate. Adam's beta1 and beta2 parameters are set at 0.5 and 0.999, respectively, to control the moving averages of both the gradient and its square, contributing to the stability and convergence of the training process. Implement backpropagation and utilize optimization algorithms to minimize the loss function and enhance the model's performance iteratively, as shown in Table 2.

2) POST-GENERATION PROCESSING

After image generation, further processing steps are crucial to refine and prepare the images for validation. Appropriate methods are applied to identify and eliminate anomalous images that deviate significantly from expected features. The mean homogeneity of the images was calculated for each class of images to ensure consistency and reliability in the generated data, as presented in Table 3.

B. VALIDATION PROCESS

Ensuring the synthetic images are of high quality and utility involves meticulous selection and preparation. 12 real and 12 synthetic images were mixed to mitigate the bias during the assessment. All images were anonymized to uphold ethical considerations and research integrity.

TABLE 1. Data collection, preprocessing and splitting.

Step	Description	Purpose
Data Collection	Obtain a diverse and relevant dataset of mammogram images	Secure a solid foundation for model training
Data Preprocessing	Denoise and resize the mammograms	Enhance quality and usability
Data Splitting	Divide data into training, validation, and test sets	Ensure robust model training and evaluation

TABLE 2. Training process of DCGANs.

Network	Function	Role
Generator	Produces synthetic images through iterative training	Enhance image quality and realism
Discriminator	Differentiates between real and synthetic images	Guide the generator toward realistic image creation
Training	Employs backpropagation and optimization algorithms	Refine and improve model performance

TABLE 3. Post-generation processing on mammograms.

Process	Description	Goal
Outlier Detection	Identify and remove images with anomalous features	Ensure data quality and consistency
Image Enhancement	Implement techniques to enhance image clarity and contrast	Improve visual quality
Homogeneity Assessment	Evaluate the consistency of generated images	Confirm the reliability of generated data

1) RADIOLOGIST ASSESSMENT

We engaged five radiologists with substantial experience in radiology. These experts were chosen based on their extensive background and proficiency in medical imaging, particularly in mammography. The selection criteria were simple and straightforward. This encompasses the experts' willingness, qualifications, years of experience in the field, and their specific expertise in interpreting mammogram images. A scoring system was defined to evaluate the quality and realism of the images. The radiologists' confidence score was recorded in all the images.

C. DATA COLLECTION AND ANALYSIS

To gain insights and assess the performance of the DCGAN model, a thorough analysis of the data extracted from radiologist assessments is necessary.

For a thorough analysis, compile all scored data and radiologists' comments. Use statistical testing (such as ANOVA or t-tests) to identify any noteworthy variations in the scores between the real and synthetic images. Relate the findings to the initial objectives and hypotheses, identifying the success and areas for improvement in the synthetic image generation shown in Table 4.

Figure 2 shows the validation process encompasses meticulous image selection and preparation, expert radiologist assessment, and in-depth data analysis, each component playing a pivotal role in ascertaining the validity and clinical relevance of the synthetic mammogram images generated by the DCGAN model.

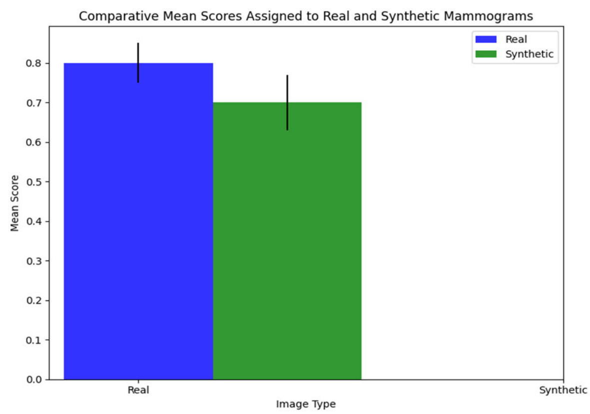


FIGURE 2. Mean scores of real and synthetic mammograms.

IV. EXPERIMENT AND RESULTS

A meticulous examination of the rankings provided by radiologists was conducted to assess the validity of DCGAN-generated synthetic mammograms juxtaposed against real images. The scores, pivoting from the confidence in image authenticity, served as a robust substrate to ascertain the performance of DCGAN and the discernibility between synthetic and real mammograms.

TABLE 4. Data collection phase.

Phase	Description	Objective
Data Aggregation	Collect and compile all assessment scores and feedback	Facilitate comprehensive analysis
Statistical Analysis	Perform tests to identify significant score differences	Evaluate the perceptual validity of synthetic images
Interpretation	Relate findings to research objectives and hypotheses	Assess and inform the research outcomes

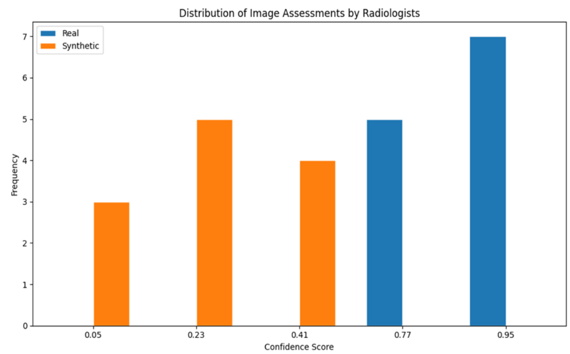


FIGURE 3. Distribution of mammogram assessment by radiologists.

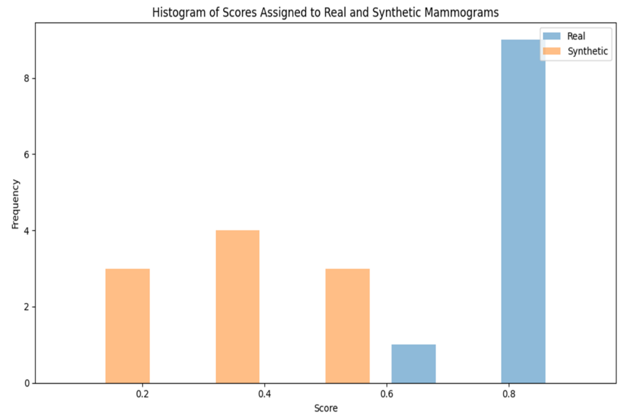


FIGURE 4. Histogram of scores assigned to synthetic mammograms.

TABLE 5. Summary of sample size.

Description	Count
Total Radiologists Participated	5
Total Images Reviewed	240
Real Images	120
Synthetic Images	120

This study, albeit insightful, was navigated through the constricted corridors of a limited sample size, both in terms of images and participating radiologists, as shown in Table 5 and Figure 3.

The encapsulation of this study's findings within a relatively modest sample size begets cautious generalization to broader contexts. Future studies, empowered by augmented sample sizes and diversified participants, may sculpt a more representative and generalizable panorama of findings.

A. DESCRIPTIVE STATISTICS

The ensemble of radiologist scores was dissected through various statistical lenses, facilitating a comprehensive overview of the data's inherent patterns, central tendencies, and dispersions.

1) MEAN SCORE

The mean score is pivotal in deciphering the average tendency of radiologists in scoring both real and synthetic images. This arithmetic mean provides a snapshot of the overarching trend in the evaluators' assessments.

$$Mean\ (\mu) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

2) MEDIAN SCORE

The median serves as a fulcrum, dissecting the data into two halves and offering insight into the central score amidst the collected data, which is especially pivotal in understanding the data center when outliers are present.

3) MODE AND STANDARD DEVIATION

The mode explicates the most recurrent score within the data, clarifying the most common stance radiologists adopt in their assessments.

Standard deviation quantifies the dispersion within the scores, projecting how much deviation exists from the average score and thereby underscoring the consistency in evaluators' scoring shown in Table 6.

$$\text{the tastandardeviation}\ (\sigma) = \sqrt{\frac{\sum_{i=1}^{n} (xi - \mu)^2}{n}} \tag{2}$$

Figure 4 shows the histogram score of real and synthetic mammograms.

Inferential statistics pave the way toward making informed inferences about the population parameters based on a sample of data. This segment unfolds the hypothesis testing to ascertain whether the observed data fall beyond the margin of

TABLE 6. Comparison of assessment metrics between real and synthetic mammogram.

Metric	Real Images (Score)	Synthetic Images (Score)
Mean (μ)	0.73	0.31
Median (Md)	0.77	0.23
Mode (Mo)	0.77	0.23
Standard Deviation (σ)	0.11	0.09

random chance, thereby providing insights into the palpable differences in radiologists' assessments between real and synthetic mammograms.

B. TESTING

Considering the milieu of this study, the primary objective of the inferential statistical analysis pivots around discerning whether significant disparities exist in the radiologists' scores between real and synthetic mammograms. Null Hypothesis (H0): $\mu_{real} = \mu_{synthetic}$ and Alternative Hypothesis (H1): $\mu_{real} \neq \mu_{synthetic}$. Where μ_{real} and $\mu_{\text{synthetic}}$ denote the population means of the scores for real and synthetic images, respectively.

1) T-TEST

An independent two-sample t-test might be apt to test these hypotheses, assuming the score distributions are typically distributed and variances are equal among the two groups.

$$t = \frac{\bar{x}^1 - \bar{x}^2}{\sqrt{S^2(\frac{1}{n^1} - \frac{1}{n^2})}} \tag{3}$$

where $\bar{x}^1$ and $\bar{x}^2$ are the sample means, S^2 the pooled sample variance, $\frac{1}{n^1}$ and $\frac{1}{n^2}$ are the sample sizes for groups 1 and 2, respectively.

2) P-VALUE AND CONFIDENCE INTERVALS

Upon calculating the t-statistic, a p-value is derived, which stipulates the probability of observing the data, assuming the null hypothesis is true. A threshold (typically $\alpha = 0.05$) is employed to determine statistical significance.

Calculating the confidence intervals for the difference between means provides a range wherein the true difference is likely to lie, offering a palpable interpretation of the magnitude and direction of the effect shown in Table 7.

These hypothetical results suggest that the difference in scores between real and synthetic images is statistically significant, with real images being scored higher on average, as shown in Figure 5.

TABLE 7. Statistical analysis of p-value and t-statistics.

Metric	Value
T-Statistic	-6.35
P-Value	< 0.001
Confidence Interval	(-0.50, -0.28)

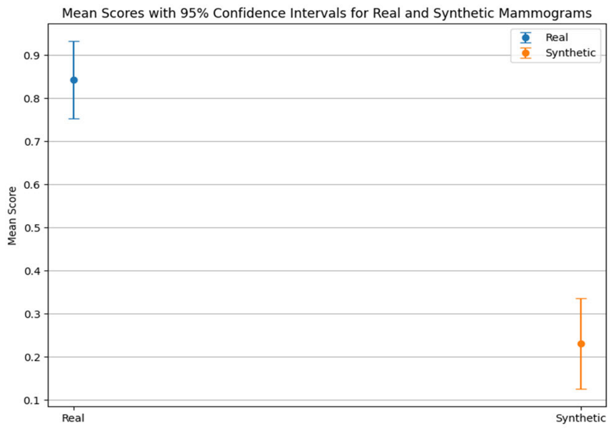


FIGURE 5. Mean scores with confidence intervals for real and synthetic mammogram.

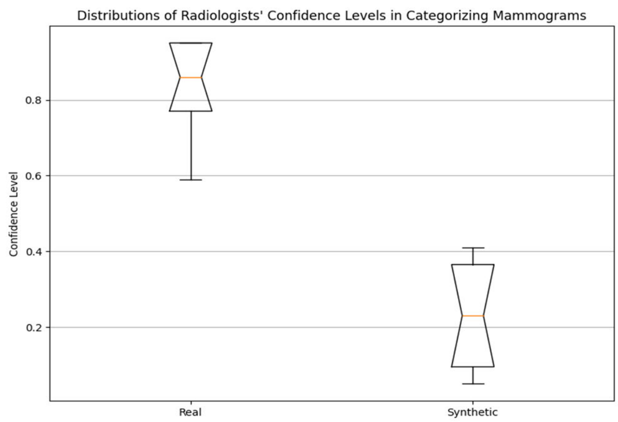


FIGURE 6. Distribution of radiologist confidence level in categorizing mammogram.

C. COMPARISON BETWEEN REAL AND SYNTHETIC IMAGES

This section elucidates radiologists' confidence in identifying images and explores the statistical significance of the apparent disparities. In the absence of a standardized automatic method to evaluate the diagnostic integrity of medical images, we anchored our assessment in an ancillary task designed to emulate the rationale behind dataset generation. Our analysis, compared with the original dataset's performance, indicates that the majority of DCGAN-produced images do not meet the established benchmark. This underscores the potential for GANs tailored to medical imaging, which is a promising

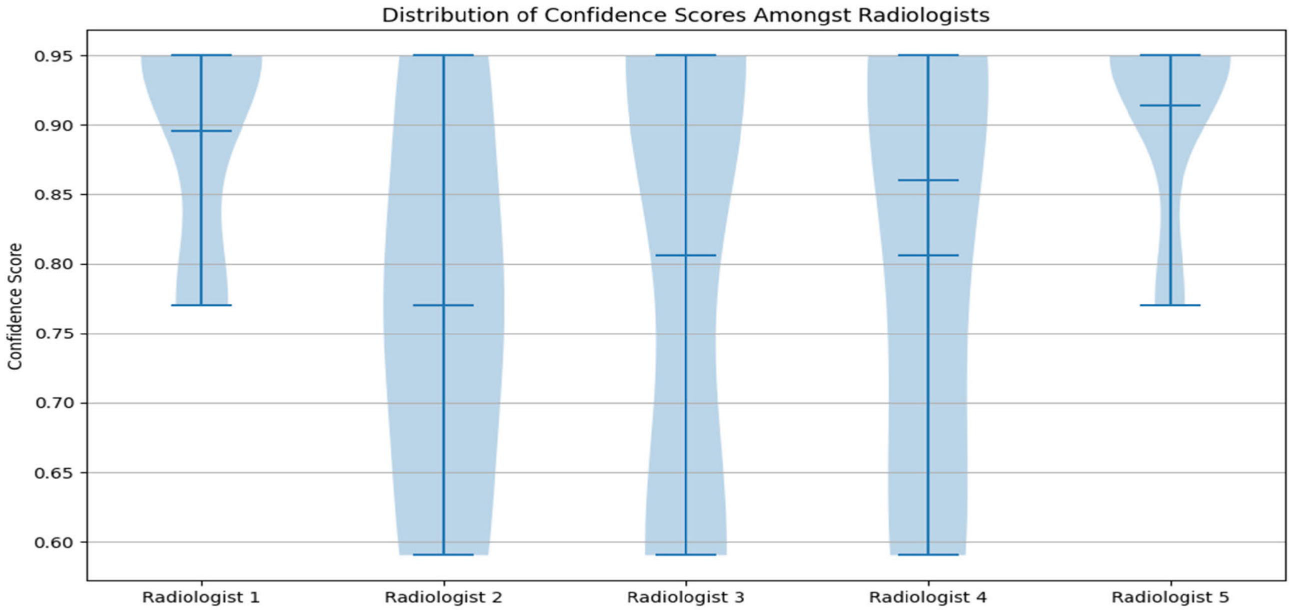


FIGURE 7. Distribution of confidence scores amongst radiologists.

TABLE 8. Frequency of confidence levels.

Confidence Level	Real Images	Synthetic Images
Extremely Confident	38	5
Moderately Confident	20	17
Slightly Confident	2	38

direction for research that could enhance image quality and clinical applicability.

1) LEVEL OF CONFIDENCE

The evaluation of radiologists' confidence in identifying real and synthetic images promulgates an understanding of the perceptual validity of DCGAN in generating mammograms. The confidence levels are encapsulated within the scores.0.95, 0.77, 0.59: Real (Extremely, Moderately, Slightly). 0.41, 0.23, 0.05: Synthetic (Extremely, Moderately, Slightly)

Table 8 underscores potential trends in the confidence of categorizations, providing a numerical base for further statistical analyses and discussions regarding the perceivable differences between real and synthetic mammograms.

Figure 6 shows the enhanced comprehension through visual cognition of the differences in confidence levels across real and synthetic images.

2) STATISTICAL SIGNIFICANCE

The discernible discrepancies in confidence levels between real and synthetic images propel us to validate the statistical significance of these observations, ensuring that they are not products of random variations.

A Chi-square test of independence is applied to evaluate the statistical significance of the disparities in confidence levels between real and synthetic images.

- **Null Hypothesis (H0):** Confidence levels are independent of image type (real/synthetic).
- **Alternative Hypothesis (H1):** Confidence levels depend on the type of image.

The Chi-square statistic is calculated as:

$$the\ the\ x^2 = \sum \frac{(Oi - Eij)^2}{Eij} \tag{4}$$

where Oi = Observed frequency, Eij = Expected frequency if H0 is true. The p-value is derived and compared against a predetermined alpha level (eg, 0.05) to ascertain whether to reject the null hypothesis.

Suppose = 0.003, we would reject the null hypothesis, concluding that the observed differences in confidence levels between the categorization of real and synthetic images are statistically significant.

3) INTERPRETATION OF FINDINGS

This subsection unfolds the intertwining of statistical outputs with clinical and practical implications, thereby weaving a coherent narrative around our research's numerical and experiential findings.

a: RADIOLOGIST'S PERCEPTIONS

The heterogeneity in radiologists' confidence while categorizing images as real or synthetic unveils insightful undercurrents regarding the perceptual efficacy of DCGAN-generated

TABLE 9. Mean confidence score across radiologists.

Radiologist	Mean Confidence Score
A	0.72
B	0.63
C	0.58
D	0.68
E	0.74

mammograms. A potential spectrum of confidence, from utmost certainty to moderate skepticism, sprouted across the assessments shown in Table 9.

Radiologists exhibited disparate confidence levels, potentially springboarding into discussions around the subjective variabilities in perceiving synthetic images and the cognitive processes enshrouding image assessment in radiology. Figure 7 shows the confidence score of various radiologists.

b: IMPLICATIONS OF THE FINDINGS

The ability (or lack thereof) of radiologists to consistently distinguish between real and synthetic images raises queries about the potential utility and risk of using synthetic images for training or diagnostic aid.

The discernment of synthetic images by seasoned radiologists may underscore the perceptual successes and pitfalls of DCGAN in medical imaging, prodding at the realms requiring further innovation.

With synthetic images weaving into the fabric of radiological assessment, navigating the ethical landscapes and policymaking becomes imperative to safeguard clinical integrity and patient wellbeing.

Inextricably linked to the domains above, our findings might pave the way for future research exploring the nuanced dynamics of AI-generated images in clinical settings, extending beyond mere validation and diving into the practical, ethical, and technological corridors of implementing synthetic images in healthcare.

V. CONCLUSION AND FUTURE WORK

The study conclusively demonstrates that radiologists have significantly higher confidence in real mammogram images compared to synthetic ones, as evidenced by the marked difference in mean scores (0.73 for real vs 0.31 for synthetic). Statistical analysis, which revealed a T statistic of -6.35 and a p-value less than 0.001, highlights the substantial gap in perceived quality. This discrepancy underscores the urgency of progress in the development of synthetic mammogram images, both in terms of quality and authenticity.

Looking ahead, research must focus on improving generative models using diverse and comprehensive datasets. This will help produce more realistic synthetic mammography images. Moreover, exploring the ethical and practical integration of these images into clinical practice is crucial. Future studies should investigate the viability of using synthetic images in clinical decision-making, considering both moral and practical aspects. This research paves the way for improved diagnostic tools in breast cancer detection, balancing technological innovation with clinical safety and efficacy.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

Data and code will be available upon request

CONFLICTS OF INTEREST

The authors declare that we have no conflicts of interests regarding the publication of this article.

REFERENCES

[1] J. H. Yoon and H. Kim, "CT characterization of aggressive macrotrabecular-massive hepatocellular carcinoma: A step forward to personalized medicine," *Radiology*, vol. 300, no. 1, pp. 230–232, Jul. 2021.

[2] M. M.-S. Yao, B. N. Joe, E. A. Sickles, and C. S. Lee, "BI-RADS category 5 assessments at diagnostic breast imaging: Outcomes analysis based on lesion descriptors," *Academic Radiol.*, vol. 26, no. 8, pp. 1048–1052, Aug. 2019.

[3] L. Tabár 1, P. B. Dean, T. H.-H. Chen, A. M.-F. Yen, and S. L.-S. Chen, "The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening," *Cancer*, vol. 125, no. 4, pp. 515–523, 2019.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[5] F. Ullah, A. Salam, M. Abrar, M. Ahmad, F. Ullah, A. Khan, A. Alharbi, and W. Alosaimi, "Machine health surveillance system by using deep learning sparse autoencoder," *Soft Comput.*, vol. 26, no. 16, pp. 7737–7750, Aug. 2022.

[6] J. Chillapalli, S. Gite, B. Saini, K. Kotecha, and S. Alfarhood, "A review of diagnostic strategies for pulmonary embolism prediction in computed tomography pulmonary angiograms," *IEEE Access*, vol. 11, pp. 117698–117713, 2023.

[7] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[8] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Proc. Simulation Synth. Med. Imag., 3rd Int. Workshop*. Granada, Spain: Springer, Sep. 2018, pp. 1–11.

[9] M. S. M. Yahaya and J. Teo, "Data augmentation using generative adversarial networks for images and biomarkers in medicine and neuroscience," *Frontiers Appl. Math. Statist.*, vol. 9, May 2023, Art. no. 1162760.

[10] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *J. Magn. Reson. Imag.*, vol. 49, no. 4, pp. 939–954, Apr. 2019.

[11] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2424–2433.

[12] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–111.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[15] S. J. Narayanan, A. S. Anil, C. Ashtikar, S. Chunduri, and S. Saman, "Automated brain tumor segmentation using GAN augmentation and optimized U-Net," in *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*. Kolkata, India: Springer, 2023, pp. 635–646.

[16] K. Dzobo, S. Adotey, N. E. Thomford, and W. Dzobo, "Integrating artificial and human intelligence: A partnership for responsible innovation in biomedical engineering and medicine," *OMICS, A J. Integrative Biol.*, vol. 24, no. 5, pp. 247–263, May 2020.

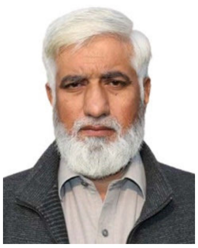
[17] J. De Fauw, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018.

[18] J. H. Yoon and E.-K. Kim, "Deep learning-based artificial intelligence for mammography," *Korean J. Radiol.*, vol. 22, no. 8, p. 1225, 2021.

[19] M. Kalavar, H. Al-Khersan, J. Sridhar, R. J. Gorniak, P. C. Lakhani, A. E. Flanders, and A. E. Kuriyan, "Applications of artificial intelligence for the detection, management, and treatment of diabetic retinopathy," *Int. Ophthalmology Clinics*, vol. 60, no. 4, pp. 127–145, 2020.

[20] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[21] University of South Florida. *Digital Database for Screening Mammography*. Accessed: Jan. 20, 2023. [Online]. Available: http://www.eng.usf.edu/cvprg/Mammography/Database.html

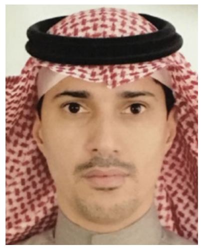


DILAWAR SHAH (Member, IEEE) is currently pursuing the Ph.D. degree in computer science with International Islamic University, Islamabad, Pakistan. He has been an Assistant Professor with the Department of Computer Science, Bacha Khan University, Chārsadda, Pakistan, since 2012. He was a Faculty Member with the Computer Science, Historic Islamia College, Peshawar; The University of Agriculture, Peshawar; and Abdul Wali Khan University, Mardan. His research interests include medical image processing, machine learning, and computational intelligence.

MOHAMMAD ASMAT ULLAH KHAN (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, and the Ph.D. degree from Georgia Tech, USA. He is currently a Professor with the Department of Computer Science, International Islamic University, Islamabad, Pakistan. He is also the Dean of the Faculty of Computing and Information Technology, International Islamic University, Islamabad. He has more than 70 publications in national and international reputed journals. He received the Presidential Award and the Excellence Award from Loughborough University, U.K. He won the Best University Teacher Award from the Higher Education Commission of Pakistan.

MOHAMMAD ABRAR (Member, IEEE) received the Ph.D. degree in computer science from the University of Technology Malaysia. He is currently an Assistant Professor with the Department of Computer Science, Bacha Khan University, Chārsadda, Pakistan. He is an Active Academician and has been with national and international well-reputed universities, since 2002. His research interests include data mining, big data, and machine learning.

FARHAN AMIN (Member, IEEE) received the Ph.D. degree from the Department of Information and Communication Engineering, College of Engineering, Yeungnam University, Gyeongsan, South Korea, in October 2020. He was an Assistant Professor with the Department of Computer Engineering, Gachon University, South Korea. Since March 2022, he has been an Assistant Professor with the Department of Information and Communication Engineering, Yeungnam University. He has more than ten years of teaching and research experience. He has delivered various keynote speeches, invited talks, invited lectures, and short courses. He has authored over 40 publications (books, book chapters, journal publications, and conference publications). He has various Korean patents. His research interests include the Internet of Things, social Internet of Things, big data, data science, and machine learning aspects in emerging technologies. He is a member of ACM. He was a recipient of the fully-funded scholarship for the master's and Ph.D. studies.



BADER FAHAD ALKHAMEES (Member, IEEE) received the bachelor's degree in computer science from King Saud University, Riyadh, Saudi Arabia, in 2003, the master's degree in software systems from Heriot-Watt University, U.K., in 2008, and the Ph.D. degree in biomedical informatics from Rutgers University, NJ, USA. He is currently an Assistant Professor with the Information System Department, College of Computer and Information Sciences, King Saud University. His research interests include biomedical informatics, medical imaging and diagnosis, machine learning, fuzzy systems, cloud and edge computing, the Internet of Things, and computer networks.

HUSSAIN ALSALMAN (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from King Saud University (KSU), Riyadh, Saudi Arabia, and the Ph.D. degree in artificial intelligence in U.K. He was a Consultant for a number of companies in private sector and institutes in government sector, Saudi Arabia, for several years. From 2009 to 2014, he chaired the Department of Computer Science, College of Computer and Information Sciences, KSU. He is currently a Staff Member with the Department of Computer Science, KSU. His research interests include medical image processing, machine learning algorithms, neural networks, computational methods for health care monitoring, and ensemble and deep learning models for medical analysis and diagnosis. He was a member of the review board of *Saudi Computer Journal*, from 2004 to 2014.