Journal of Computer Science

Research Article

Robust Adversarial Attack Detection via Generative Adversarial Network With Residual Multi-Layer Aggregation Based Contrastive Loss Function

Amudha Gopalakrishnan and Nalini Joseph

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, India

Article history
Received: 26-04-2025
Revised: 06-08-2025
Accepted: 16-08-2025

Corresponding Author:
Amudha Gopalakrishnan
Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, India
Email: amudhag.cse@gmail.com

Abstract: Adversarial attacks in medical imaging refer to subtle modifications to images that mislead diagnostic systems, resulting in inaccurate diagnoses and assessments. These attacks exploit vulnerabilities in image processing, leading to misclassification or altered visual features that often go unnoticed. This raises serious concerns about the security and reliability of medical diagnosis, directly impacting clinical decision-making and patient safety. This research proposes a Generative Adversarial Network with Residual Multi-Layer Aggregation-based Contrastive Loss Function (GRMLA-CLF) to effectively identify adversarial attacks using medical images. In the generator, Residual Multi-Layer Aggregation (RMLA) is incorporated to capture fine-grained information and structural patterns of adversarial attacks, improving the model's adaptability. The Contrastive Loss Function (CLF) enhances adversarial attack detection by increasing the distance between genuine and adversarial samples, ensuring a clear distinction in latent space, and ensuring distinct representation. This enhances model robustness by reducing sensitivity to small perturbations while preserving significant features necessary for accurate decision-making. The proposed GRMLA-CLF achieves high accuracy rates of 99.81, 99.64, and 98.65% on the ISIC2019, Chest X-ray, and APTOS2019 datasets, respectively, outperforming existing methods like Global Attention Noise (GATN).

Keywords: Adversarial Attacks, Contrastive Loss Function, Generative Adversarial Network With Residual Multi-Layer Aggregation, Medical Images, Vulnerabilities

Introduction

Medical image classification is crucial in modern healthcare, enabling efficient and accurate diagnosis across numerous conditions. This process involves the analysis and classification of medical images based on various factors like imaging modalities and clinical information (Hussain et al., 2025). Deep Learning (DL), a subset of Artificial Intelligence (AI), has gained significant popularity in medical image analysis due to its strong performance in classifying and interpreting complex patterns. It enhances the ability to extract essential features and offers flexibility in addressing intricate diagnostic challenges. Nevertheless, DL methods possess inherent vulnerabilities that make them susceptible to adversarial perturbations, which can lead to misclassification by exploiting model vulnerabilities (Alzubaidi et al., 2024; Haque and Zafar, 2024). Adversarial Deep Learning (ADL) aims to compromise DL models by generating deceptive data with subtle modifications. These adversarial examples exploit DL vulnerabilities, raising concerns about model integrity and reliability (Ng and Hargreaves, 2023; Sheikh and Zafar, 2024). Therefore, addressing such attacks is essential to ensure the security and robustness of DL-based medical image analysis systems (Jiang et al., 2024).

Security threats are typically categorized into two groups: Causal and probing attacks. During training, causal attacks degrade model performance by introducing adversarial samples that disrupt the learning process (Kanca Gulsoy et al., 2024; Kanca et al., 2025). Several defense strategies have been developed,

including adversarial training, FreeLB, and robust optimization. Among these, Adversarial Propagation (AdvProp) is an effective training method that enhances robustness by learning from both clean and adversarial examples. Unlike traditional training, AdvProp (Xu et al., 2023) introduces separate Batch Normalization (BN) layers to minimize conflicts between learning tasks. However, AdvProp increases computational overhead due to the generation of adversarial samples and the management of additional BN layers, making it less suitable for resource-constrained environments.

Some of the primary causal attacks include backdoor and poisoning attacks, which are designed to produce specific effects. A probing attack, categorized as an evasive attack, alters test data after training, thereby outperforming detection mechanisms (Vaddadi et al., 2024; Pervin et al., 2023). Recent advancements in generative methods have revolutionized image manipulation and generation, enabling the creation of highly realistic images that are nearly indistinguishable from authentic counterparts (Pasqualino et al., 2024; Anand et al., 2024). Adversarial attacks not only raise critical concerns but also have potentially life-threatening consequences, particularly in the medical field (Chanakya et al., 2024; Gbashi et al., 2023; Priya and Dinesh Peter, 2025). By exploiting vulnerabilities in image processing, these attacks introduce visual alterations that often remain undetected. To address this issue, a Generative Adversarial Network with Residual Multi-Layer Aggregation-based Contrastive Loss Function is proposed to enhance adversarial attack detection in medical images. Unlike traditional GANs, the proposed GRMLA-CLF employs Residual Multi-Layer Aggregation (RMLA) in the generator to fuse deep and shallow features, effectively capturing both global patterns and fine-grained information in perturbed inputs. The Contrastive Loss Function (CLF) ensures better separation between genuine and adversarial features, thereby enhancing model robustness. This process not only improves adversarial resistance but also strengthens feature representation across diverse medical image datasets. Consequently, the proposed method increases the reliability of medical diagnoses and supports accurate clinical decisions, ultimately improving patient safety.

The key contributions of this research are explained below:

- RMLA is incorporated into the generator to enhance its ability to produce more realistic and structurally consistent adversarial examples by fusing deep and shallow residual features. This multi-layer aggregation enables the generator to better capture the complex patterns present in medical images
- CLF enforces greater separation between clean and adversarial features in the embedding space, enhancing class discriminability. This helps the model learn robust feature representations, thereby improving resistance to adversarial attacks
- Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to enhance low-contrast images by preventing over-amplification of noise, making features more distinguishable for the model

Literature Survey

Dai et al. (2023) introduced a Global Attention Noise (GATN) injection by containing global and attention noise into the feature layers. Global noise-enhanced lesion features in medical images by preserving sharp areas where the model was vulnerable. Attention noise locally smoothed the model, mitigating the effect of small perturbations. Based on medical images, GATN-Related noise (GATN-R) was introduced with clearer lesion boundaries. Trainable attention noise was subsequently included in the feature layers to further smooth the model locally and highlight salient regions, thereby enhancing model resistance to small perturbations. Tsai et al. (2023) proposed one- and multi-pixel level attacks using Deep Neural Networks (DNNs) to classify medical images. The primary multi-class and multi-label datasets were utilized to conduct one-pixel attacks. Multiple experiments were conducted by varying the number of altered pixels, which enhanced both the model's performance and the robustness of the DNN-based method.

Annamalai et al. (2023) developed a Convolutional Neural Network (CNN) by integrating an Auction-Based Optimization Approach (ABOA) and Dice Similarity Coefficient (DSC) to predict pulmonary disease. The CNN eliminated irrelevant features during feature extraction. This limitation was effectively addressed by incorporating ABOA and DSC for improved classification of pulmonary disease types. An autoencoder block was employed to transfer image features across multiple convolutional layers. However, the non-convex nature of DSC resulted in unstable gradients, requiring additional training iterations to achieve optimal performance.

Nazir et al. (2024) established a weighted average ensemble method by integrating Inception-V3, VGG16, and CNN to classify severity levels in diabetic cases. An automated detection system was designed to assist in early disease diagnosis and reduce the incidence of vision loss across diverse patient groups. The established weighted average ensemble model proved to be efficient, robust, and accurate. Ben Graham's method was adopted to address various lighting resolution issues, and OpenCV Gaussian blur

was applied to smooth image corners. Harini et al. (2024) presented a Self-Attention-based Cycle-Consistent GAN–Archerfish Hunting Optimization Approach for Melanoma Classification on Dermoscopic Images (SACCGAN-AHOA-MC-DI). Dermoscopic images were pre-processed by utilizing Adjusted Quick Shift Phase with Dynamic Range Compression (AQSP-DRC) to remove noise and enhance image quality. These pre-processed images were then segmented using Piecewise Fuzzy C-Means Clustering (PF-CMC) to isolate the Region of Interest (ROI). The segmented ROI was further processed using the Hexadecimal Local Adaptive Binary Pattern (HLABP) to extract radiomic features. Finally, SACCGAN was employed to classify skin cancer effectively and accurately.

Zhao et al. (2025) developed a Scale Enriching Method (SEM) to improve the transferability of adversarial examples by applying an input scale-enriching model. SEM enhanced significant regions and increased tolerance to variations across different target models, thereby improving the transferability of adversarial examples. During perturbation, SEM prevented the introduction of noise, preserving textural features across varying scales. Rahman et al. (2025) proposed a Deep Neural Network (DNN) for analyzing adversarial attack methods such as Projected Gradient Descent (PGD), Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and others for image classification. Additionally, two adversarial ensemble strategies, Mean and Weighted ensemble, were employed to generate adversarial examples using different attack techniques. The DNN demonstrated improved performance after applying defensive measures. Table 1 provides a summary of existing methods, highlighting their key advantages and limitations.

From the overall analysis, existing methods exhibit limitations such as reduced model interpretability, poor generalization, instability, overfitting, and mode collapse, which hinder the generation of diverse outputs. Additionally, adversarial attacks exploit vulnerabilities in image processing, leading to misclassification. To address these issues, the GRMLA-CLF is proposed to identify adversarial attacks by incorporating Residual Multi-Layer Aggregation (RMLA) and Contrastive Loss Function (CLF). These components enhance feature learning, ensure better stability and generalization, and prevent model collapse. CLF maximizes the separation between adversarial and genuine samples in the feature space. As a result, the proposed approach improves model robustness, enhances reliability, and reduces misclassification in the presence of adversarial attacks.

Table 1: Summary of existing methods by representing the advantages and limitations

Author, Year	Methods	Advantages	Limitations
Dai et al. (2023)	GATN	It enhances model resistance to small perturbations	GATN led to excessive feature distortion, which minimized interpretability, generalization, and instability
Tsai et al. (2023)	DNN	It automatically learn complex patterns due to depth and non-linearity	Exploiting less perturbation leads to suboptimal performance
Annamalai et al. (2023)	CNN-ABOA-DSC	This method optimally selects discriminative features and minimizes redundancy	The non-convex nature of DSC leads to unstable gradients and model performance
Nazir et al. (2024)	weighted average ensemble method	Minimize the number of vision losses for diverse patients.	Suffers from overfitting due to combining multiple DL methods
Harini et al. (2024)	SACCGAN-AHOA-MC-DI	This method provides high-quality and realistic dermoscopic images.	CGAN suffers from mode collapse, where it fails to generate diverse output because of inconsistencies in attention mapping

Materials and Methods

This research proposes GRMLA-CLF to effectively identify adversarial attacks. Initially, the ISIC2019 (link: https://challenge.isic-archive.com/landing/2019/ (Accessed on 10 April 2025)), Chest X-ray, and APTOS2019 datasets are used to evaluate the model's performance. The obtained images are pre-processed using resizing and CLAHE to standardize input dimensions and enhance low-contrast images. Finally, GRMLA-CLF is applied to detect adversarial attacks in medical images. Figure 1 illustrates the workflow of the proposed methodology.

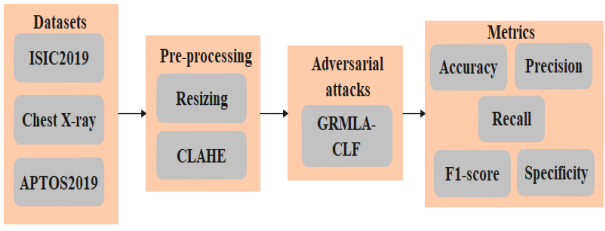

Fig. 1: Workflow process of the proposed methodology

Datasets

This research employs ISIC-2019, NIH Chest X-ray, and APTOS 2019 (Nazir et al., 2024) datasets to determine the model performance. These datasets include diverse medical imaging challenges for ensuring robustness across different modalities. By leveraging the available clinical data, the research enhances the model's reliability. Additionally, the Derma dataset (Tsai et al., 2023) is used to analyze the generalization capability of the proposed method. A detailed description of these datasets is provided below.

Derma: This is a multi-class dermatoscopic colored dataset containing pigmented skin lesions. It includes 10,015 images across 7 classes, each sized $3\times600\times450$, and is used for generalization analysis.

ISIC2019: This dataset is used to classify primary pigmented skin diseases and contains a total of 25,331 dermoscopic images categorized into 9 classes. It supports comprehensive analysis by providing diverse cases that facilitate the development of more accurate models.

Chest X-ray: This is a binary-class, multi-label frontal view X-ray dataset consisting of grayscale images with 14 classes and a total of 112,120 images. Each image has an original size of 1×1024×1024 pixels, featuring 247 diverse label combinations. However, most classes contain fewer than 100 images, with some having fewer than 10.

APTOS2019: The images were analyzed by an expert team and categorized into five stages: No Diabetic Retinopathy (DR), mild, moderate, severe, and proliferative DR. The dataset contains 3,662 training and 1,992 testing images. The images are split into 80% training and 20% testing sets and then fed into the pre-processing stage.

Pre-Processing

The obtained images are pre-processed through resizing and CLAHE (Mohammadi and Nguyen, 2024) to ensure uniform input dimensions and enhanced contrast, which improves overall quality for accurate analysis. The images are then resized to 224×224 for standardizing input dimensions, achieving compatibility with DL. This resolution balances computational effectiveness and preserves essential visual features for accurate classification.

After resizing, CLAHE is utilized to enhance the DR image's intricate details, low contrast, and textures by adjusting the input image's lightness value. Unlike conventional histogram equalization, which stretches intensity levels over the entire dynamic range, CLAHE addresses artifacts in low-texture regions and prevents over-amplification of noise by splitting the image into small, overlapping tiles. Each tile undergoes histogram equalization with a clip limit through five phases: Excess calculation, computation, mapping, scaling, and redistribution using the Cumulative Distribution Function (CDF). To enhance contrast, bilinear interpolation stitches the tiles together, enhancing local contrast and making edges and borders more distinct. Finally, the pre-processed images are passed to the model for adversarial attack identification.

Adversarial Attack

Projected Gradient Descent (PGD) is an iterative adversarial attack that perturbs an input sample to increase the model's loss while ensuring the perturbation remains within a defined bound. It enhances upon the Fast Gradient Sign Method (FGSM) by using multiple small steps instead of a single large step, making it more effective against DL models. PGD is selected over other attacks like FGSM and DeepFool because it performs iterative perturbations within a bounded region, making it more effective for analyzing model robustness. Unlike FGSM, which uses a single-step update, PGD explores the loss surface more thoroughly. Compared to Deep Fool, PGD is simpler to implement and more suitable for adversarial training. Its widespread adoption as a benchmark attack makes it ideal for assessing the performance of defense mechanisms.

At each iteration, the perturbation is projected into the *e*-ball around the original input to ensure validity. PGD is considered a robust first-order adversarial attack and is widely employed as a benchmark for evaluating model robustness. It is particularly important in adversarial training, where models are trained to defend against attacks by being exposed to adversarial examples. By incorporating PGD attacks, vulnerabilities in neural networks can be effectively assessed, and corresponding defenses can be developed to improve security in applications such as medical imaging. The adversarial training process exposes the model to perturbed inputs generated using white-box PGD attacks, which allow full access to gradients for crafting effective adversarial examples. During training, learning from these examples enhances the model's ability to distinguish between manipulated and clean inputs. This significantly improves overall robustness against adversarial threats.

Generative Adversarial Network With Residual Multi-Layer Aggregation Based Contrastive Loss Function (GRMLA-CLF)

After pre-processing, GRMLA-CLF is employed to identify adversarial attacks by learning the distribution of clean images and detecting deviations caused by perturbations. The proposed GRMLA-CLF is chosen

for its strengths in feature representation, class separability, and sample diversity. GAN (Devarajan and Khader, 2023) is used to reconstruct clean images from perturbed inputs, thereby minimizing the impact of attacks and continuously improving the model's ability to distinguish between real and adversarial samples. GAN consists of two neural networks: A generator and a discriminator, which compete against each other. The generator captures the training data distribution, while the discriminator differentiates between real and generated images. In the generator, Residual Multi-Layer Aggregation (RMLA) is incorporated to strengthen feature representation by aggregating deep and shallow features across layers. This enables the model to capture both fine-grained details and global context, which is essential in adversarial scenarios. Furthermore, the integration of the Contrastive Loss Function (CLF) enforces clear separation between clean and adversarial representations in the embedding space, ensuring the network learns to distinguish subtle differences effectively. Together, these components form an effective model that not only defends against adversarial attacks but also enhances feature robustness and class separability. A detailed explanation of GRMLA-CLF is provided below.

Generator: The generator is a neural network responsible for producing synthetic data samples that resemble real data. It takes a random noise vector as input and transforms it into realistic data through a series of learned transformations. An equilibrium state is achieved when the generator produces samples that the discriminator cannot distinguish from real data, assigning nearly equal probabilities to both real and generated samples. In an adversarial manner, both networks are trained together via error backpropagation of the loss function. The generator input is a pre-processed image that acts as noise from a prior distribution with variables $p_z(z)$. In this way, the generator produces varying samples from the data distribution x via mapping $G(z)$. The generator architecture consists of 7 down-sampling modules, 1 up-sampling convolutional layer, and 6 up-sampling modules. The up-sampling and down-sampling modules perform a “concatenation process” at corresponding levels to better synthesize extracted features from the bottom to the top layers.

Residual Multi-Layer Aggregation (RMLA): In the generator, ResNet50 is incorporated as the backbone due to its proven effectiveness in deep learning. Its robust feature extraction capabilities and residual connections enable stable gradient flow in deeper networks. Within ResNet50, the RMLA module is used to enhance the model’s sensitivity to subtle perturbations commonly found in medical images. This adaptation allows the network to capture fine-grained features and local distortions introduced by adversarial attacks more effectively. As a result, RMLA provides a novel contribution toward improving adversarial robustness in clinical applications, enabling the generator to produce more realistic and high-quality samples for applications like adversarial attacks in medical imaging.

A pre-trained ResNet50 model on ImageNet is employed to leverage both low- and high-level feature extraction capabilities. During training, the initial layers of ResNet50 are frozen to retain essential visual features such as textures and edges, which reduces training time and helps prevent overfitting. The deeper layers are fine-tuned using domain-specific datasets such as Chest X-ray, ISIC2019, and APTOS2019 to adapt to complex and subtle patterns in medical images. This approach enhances feature transferability and improves the model's generalization across different medical imaging modalities. Furthermore, ResNet50 addresses the problem of model degradation in deep networks by employing residual learning, which allows deeper networks to learn effectively. It adopts skip connections by changing the learning objective using Eq. (1):

$$F(x) = H(x) - x \tag{1}$$

Where x indicates input, $H(x)$ represents output after processing, and $F(x)$ denotes final output. Through circulating the original input with the processed output via skip connections, ResNet50 learns residual mapping effectively to acquire the final representation. This enables efficient training of deep networks without encountering vanishing gradients, ensuring stable learning and improved performance. The multi-layer aggregation residual network offers several benefits: It establishes residual learning by integrating the original input with the processed output, which simplifies the training of deep networks and prevents vanishing gradient issues. It facilitates multi-scale information aggregation during feature extraction, allowing for more comprehensive capture of image features. Additionally, it reduces the number of network parameters, simplifying the training process and enhancing computational efficiency. To further minimize complexity and increase efficiency, dimensionality reduction is applied before large-block convolutions, balancing the network’s depth and width to improve accuracy. Aggregating visual information at various scales also improves multi-scale feature extraction through two consecutive 3×3 convolution operations. Compared to traditional ResNet50, RMLA achieves the same receptive field using two sets of 5×5 convolutions, which reduces parameters and simplifies

training. These enhancements enable more effective and accurate feature extraction using RMLA. Fig. 2 shows the structure of the RMLA module.

Discriminator: A discriminator is used to differentiate between clean and adversarially perturbed inputs by identifying subtle distortions introduced through adversarial attacks. The discriminator network outputs a 1×16×16 matrix, where each pixel represents the discriminant value for a small region of the input image. This structure enables the discriminator to perform localized discrimination, which improves its ability to detect adversarial manipulations effectively. By evaluating smaller regions independently, the model enhances robustness and ensures stronger adversarial defense. Moreover, this fine-grained discrimination contributes to refining adversarial training strategies and strengthens overall model security against adversarial threats.

Contrastive Loss Function (CLF): During training, CLF plays a significant role in enhancing the adversarial robustness of the GAN by guiding both the generator and discriminator.

CLF minimizes the distance among feature representations of genuine images while maximizing the distance between genuine and adversarial pairs in the latent space. For the generator, this process facilitates the creation of high-fidelity and realistic outputs that are structurally similar to clean images. Moreover, CLF sharpens the discriminator's ability to distinguish adversarial distortions by separating the features of fake and real samples. This dual influence enables the generator to focus on robustness, whereas the discriminator becomes sensitive to subtle perturbations. Hence, CLF enhances the feedback loop in GAN training, ensuring an effective adversarial detection process across medical image datasets. The implementation steps for CLF are discussed below in detail.

Step 1: Compute the Euclidean distance d between the latent feature representations of two input samples in the embedding space.

Step 2: Use a predefined margin to control the minimum distance between dissimilar sample pairs.

Step 3: Calculate the Contrastive Loss Function (CLF) using Eq. (2):

$$L_{cont} = \sum_{ij} w(1 - y_{ij}) d_{ij}^2 + (1 - w) y_{ij} [\max(m - d_{ij}, 0)]^2 \quad (2)$$

Where d_{ij} denotes the distance between the feature vectors of F_o and F_l at position (i,j), m represents the margin for the revised feature pairs, w refers to the balance of weights of the two terms in Equation (2), and $\underline{y_{ij}}$ indicates the label at position (i,j). This enhances the model's ability to distinguish between adversarial and clean inputs by learning discriminative representations, making adversarial perturbations more detectable.

Step 4: Integrate the Contrastive Loss Function (CLF) with the adversarial loss of the GAN to jointly optimize both the generator and discriminator.

Step 5: Finally, perform gradient descent to minimize the total loss, ensuring the model effectively clusters genuine samples in the latent space.

Thus, the GRMLA-CLF strengthens adversarial defense and ensures security in DL methods, and its structure is shown in Fig. 3.

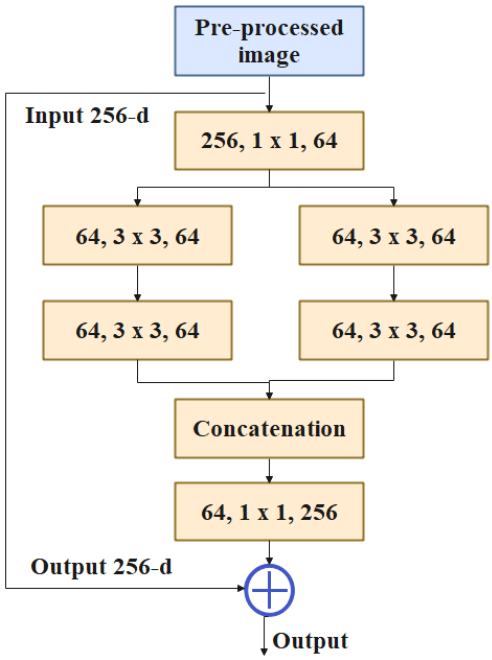

Fig. 2: Structure of RMLA module

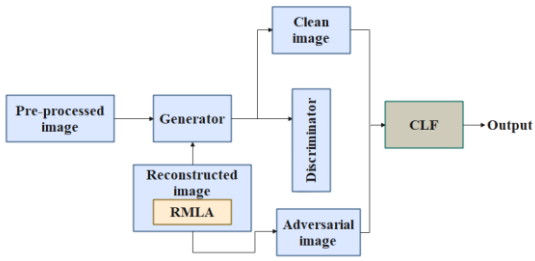

Fig. 3: Structure of GRMLA-CLF

Experimental Results

The proposed GRMLA-CLF is simulated in a Python 3 environment with a system configuration of 64 GB RAM, a Windows 10 operating system, and an Intel i5 processor. The metrics of recall, accuracy, precision, specificity, and F1-score are used to evaluate the model's performance, as mathematically equated in

Eqs. (3) to (7):

$$Recall\ \frac{TP}{TP+FN} \times 100 \quad (3)$$

$$Accuracy\ \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (5)$$

$$Specificity\ \frac{TN}{TN+FP} \times 100 \quad (6)$$

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \times 100 \quad (7)$$

Where *FN* denotes False Negative, *FP* denotes False Positive, *TP* denotes True Positive, and *TN* denotes True Negative.

Performance Analysis

Figure 4 represents the graphical representation of different DL methods. When compared to existing methods like Swin transformer-CLF, ViT-CLF, and GAN-CLF, the proposed GRMLA-CLF obtains high accuracies of 99.81, 99.64, and 98.65% on the ISIC2019, Chest X-ray, and APTOS2019 datasets, respectively. This is due to its superior ability to enhance adversarial robustness and feature representation. The RMLA assists in capturing multi-scale hierarchical features that enhance feature continuity and learn intricate attack patterns effectively. The CLF ensures better distinction among adversarial and genuine samples by enhancing inter-class variance and minimizing intra-class variance. Additionally, the proposed method enables the model to effectively learn robust discriminative features against PGD attacks during training. Therefore, this combination minimizes misclassification and increases resilience to adversarial perturbations, resulting in high model performance.

Figure 5 shows a graphical comparison of different loss functions used for training and evaluating performance. Compared to existing methods like Hinge Loss Function (HLF), Cross Entropy Loss Function (CELF), and KL-Divergence Loss Function (KL-DLF), the proposed Contrastive Loss Function (CLF) achieves higher accuracy of 99.81, 99.64, and 98.65% on the ISIC2019, Chest X-ray, and APTOS2019 datasets, respectively, by effectively distinguishing between similar and dissimilar samples.

CLF improves feature space separation, ensuring that samples from different classes are clearly identified. Moreover, it enhances the model's ability to learn discriminative representations, making it more robust to adversarial variations and noise. By minimizing the distance among genuine pairs and maximizing it for adversarial pairs, the function boosts both robustness and generalization, leading to higher accuracy.

Figure 6 shows a graphical representation of k-fold validation used for model performance evaluation. When k = 5, the model achieves higher accuracy compared to k values of 3, 7, and 9 because it provides an optimal balance between variance and bias. The case with k = 3 is more prone to overfitting due to higher sensitivity to noise, whereas k values of 7 and 9 increase bias, which reduces sensitivity to local patterns. With k = 5, the model effectively captures the underlying distribution while minimizing the impact of noise, ensuring a better decision boundary. Therefore, this balance improves the model's generalization and stability, resulting in enhanced performance.

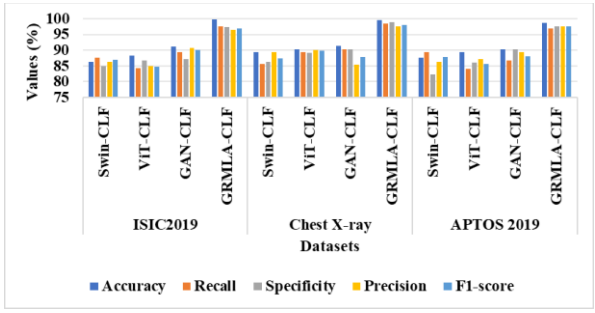

Fig. 4: Graphical representation of different DL methods

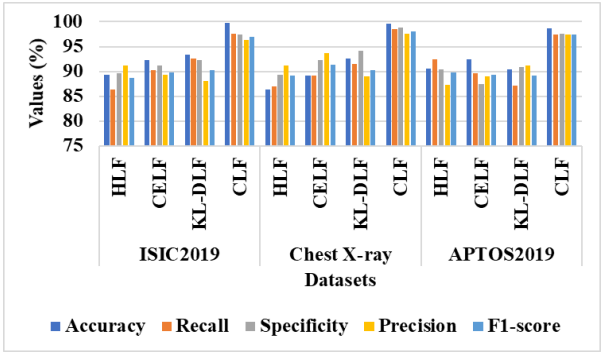

Fig. 5: Graphical representation of different loss functions used for training and evaluation performance

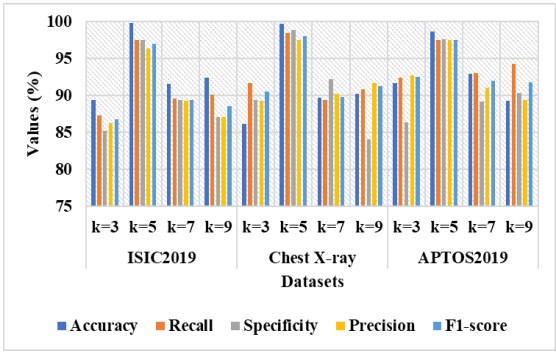

Fig. 6: Graphical representation of the k-fold validation process used for model performance evaluation

Table 2 presents a performance analysis of computational time complexity and memory consumption. The proposed GRMLA-CLF achieves lower time complexity of 59.25s, 54.26s, and 50.36s compared to existing methods such as Swin-CLF, ViT-CLF, and GAN-CLF due to its effective feature reuse process.

The RMLA module enables gradient flow through skip connections, minimizing redundant computations and improving convergence. Additionally, CLF optimizes the feature space with fewer parameters, reducing unnecessary calculations. This combination of CLF and residual learning increases efficiency by focusing on significant feature representations, which results in lower memory usage compared to traditional methods. Furthermore, the proposed method demonstrates improved statistical performance, showing lower p-values in t-tests and tighter confidence intervals compared to existing approaches.

Generalization Analysis

Table 3 presents a performance analysis of generalizability across different deep learning methods using the Derma dataset. Generalization refers to the model's ability to maintain strong performance on unseen data. Compared to existing methods, the proposed GRMLA-CLF achieves superior generalization by using RMLA to capture both high- and low-level features, while CLF enhances class separability. These components ensure the model effectively learns feature representations, leading to improved performance.

Cross-Dataset Validation

Table 4 presents the performance analysis of cross-dataset validation, where the model is trained on the Chest X-ray dataset and tested on the APTOS2019 dataset. The Chest X-ray dataset was chosen for training because it contains a large number of images, while APTOS2019 has fewer images for testing. The proposed method achieves a high accuracy of 95.36%, demonstrating its ability to learn scale-invariant and transferable representations compared to existing methods.

Table 2: Performance analysis of time complexity and memory consumption

Datasets	Methods	Computational time (s)	Memory consumption (MB)	p-value from t-test	95% of CI
ISIC2019	Swin-CLF	78.26	156	0.025	86.2
	ViT-CLF	75.16	167	0.022	87.5
	GAN-CLF	69.14	159	0.019	88.1
	GRMLA-CLF	59.25	145	0.015	91.3
Chest X-ray	Swin-CLF	71.26	156	0.028	84.7
	ViT-CLF	69.26	147	0.025	85.4
	GAN-CLF	65.17	139	0.021	86.2
	GRMLA-CLF	54.26	126	0.019	89.5
APTOS 2019	Swin-CLF	59.44	165	0.028	82.9
	ViT-CLF	62.31	147	0.025	84.0
	GAN-CLF	55.78	143	0.020	85.1
	GRMLA-CLF	50.36	132	0.017	88.6

Table 3: Performance analysis of generalizability with different DL methods using the Derma dataset

Methods	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-score (%)
Swin-CLF	89.25	85.64	82.67	83.69	84.65
ViT-CLF	90.58	87.29	86.94	84.29	85.76
GAN-CLF	92.68	89.38	88.59	86.39	87.85
GRMLA-CLF	97.25	96.17	95.18	95.06	95.61

Table 4: Performance analysis of cross-dataset validation with training on the Chest X-ray dataset and testing on APTOS2019

Methods	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-score (%)
Swin-CLF	78.25	76.39	80.34	82.17	79.17
ViT-CLF	80.29	83.45	82.49	86.98	85.17
GAN-CLF	82.69	86.39	86.39	88.29	87.32
GRMLA-CLF	95.36	94.20	93.12	95.39	94.79

Comparative Analysis

Tables 5 to 7 present a comparative analysis of existing methods on the ISIC2019, Chest X-ray, and APTOS2019 datasets. As shown in Table 5, the proposed GRMLA-CLF achieves a higher accuracy of 99.81% on the ISIC2019 dataset compared to Dai et al. (2023); Harini et al. (2024). Similarly, Table 6 shows that the proposed method attains superior accuracy, recall, and specificity of 99.64, 98.4, and 98.8%, respectively, on the Chest X-ray dataset compared to Annamalai et al. (2023). Table 7 indicates that the

method achieves accuracy, precision, F1-score, and recall of 98.65, 97.48, 97.02, and 96.57%, respectively, on the APTOS2019 dataset compared to Nazir et al. (2024). These improvements are attributed to residual aggregation, which enhances deep feature learning by preserving significant contextual and spatial information across layers, thereby increasing the model's robustness. Additionally, the contrastive loss function maximizes the distance between genuine and adversarial samples, ensuring better separation in feature space and reducing the model's vulnerabilities.

Table 5: Comparative analysis of existing methods on ISIC2019

Methods	Accuracy (%)	Recall (%)	Specificity (%)
GATN (Dai et al., 2023)	72.49	N/A	N/A
SACCGAN-AHOA-MC-DI (Harini et al., 2024)	99.5	96	93
Proposed GRMLA-CLF	99.81	97.54	97.45

Table 6: Comparative analysis of existing methods on Chest X-ray

Methods	Accuracy (%)	Recall (%)	Specificity (%)
CNN-ABOA-DSC (Annamalai et al., 2023)	96.5	97.3	96.6
Proposed GRMLA-CLF	99.6	98.4	98.8

Table 7: Comparative analysis of existing methods on APTOS 2019

Methods	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)
Weighted average ensemble (Nazir et al., 2024)	95.06	87.88	85.69	83.78
Proposed GRMLA-CLF	98.65	97.48	97.02	96.57

Discussion

This section describes the limitations of the existing methods, along with the advantages of the proposed GRMLA-CLF, based on their adversarial attack identification performance. The existing methods' limitations are noted as follows: The GATN (Dai et al., 2023) suffers from excessive feature distortion, which minimizes model interpretability, generalization, and creates instability in adversarial training. DNN (Tsai et al., 2023) compromises medical image classification by exploiting fewer perturbations, which results in misdiagnosis. The convex nature of DSC (Annamalai et al., 2023 results in unstable gradients, demanding more training iterations for optimal performance. The weighted average ensemble (Nazir et al., 2024) suffers from overfitting due to the combination of multiple DL approaches, while improper weight assignments amplify adversarial vulnerabilities. The CGAN (Harini et al., 2024) suffers from mode collapse, failing to generate diverse outputs due to inconsistencies in attention mapping. The proposed GRMLA-CLF overcomes these limitations by incorporating RMLA and CLF. The RMLA enhances feature extraction, rendering the model effective in capturing both high and low-level attack patterns. This enhances robustness against perturbations and minimizes sensitivity to minor adversarial noise. Furthermore, CLF enables better separation between genuine and adversarial samples, thereby reducing FP. Additionally, the GAN's generative ability effectively learns the real data distribution. Therefore, the proposed GRMLA-CLF improves generalization by learning richer representations over PGD attack types. Furthermore, the proposed method is designed for deployment in clinical environments, supporting practical applications. The use of CLAHE enhances image visibility, increasing the model's applicability. The model's strong performance across multiple datasets demonstrates its relevance in improving accuracy and minimizing misdiagnosis. Its robustness is evident not only under white-box PGD attacks but also through strong generalization across the Derma dataset and in cross-dataset validation.

Conclusion

This research proposes GRMLA-CLF for accurately identifying adversarial attacks in medical images. In the generator, RMLA captures information during feature extraction, reducing the number of network parameters, simplifying the training process, and enhancing computational efficiency. CLF improves the model's ability to differentiate between adversarial and genuine images, minimizing FP. This assists in preserving significant medical image details, enabling clinical decision-making and accurate diagnosis. In pre-processing, CLAHE enhances image contrast while avoiding over-amplification of noise, which is crucial for medical images. The CLAHE processes small regions adaptively to preserve local information and improve visibility. Compared to existing methods like GATN, the proposed GRMLA-CLF achieves high accuracies of 99.81, 99.64, and 98.65% on the ISIC2019, Chest X-ray, and APTOS2019 datasets, respectively. The GRMLA-CLF was evaluated under white-box adversarial attacks, where attackers have full access to perform gradient-based attacks. However, adversarial attacks in real-world scenarios often occur under black-box conditions, where the adversary has limited access, which poses challenges to the model's practical robustness. In future work, this research will be extended to defend against black-box

adversarial attacks using efficient methods across different datasets to further enhance the model's robustness and reliability.

Acknowledgment

The authors, Amudha Gopalakrishnan and Nalini Joseph, would like to express their sincere gratitude to the Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, India, for their constant support and encouragement throughout the course of this research work.

Funding Information

This research received no external funding.

Authors Contributions

Both the authors have equally contributed to this manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

Conflicts of Interest

The authors declare no conflict of interest.

References

Alzubaidi, L., AL–Dulaimi, K., Obeed, H. A.-H., Saihood, A., Fadhel, M. A., Jebur, S. A., Chen, Y., Albahri, A. S., Santamaría, J., Gupta, A., & Gu, Y. (2024). MEFF – A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging. *Intelligent Systems with Applications*, *22*, 200355. https://doi.org/10.1016/j.iswa.2024.200355

Anand, A., Krithivasan, S., & Roy, K. (2024). RoMIA: a framework for creating Robust Medical Imaging AI models for chest radiographs. *Frontiers in Radiology*, *3*(2023). https://doi.org/10.3389/fradi.2023.1274273

Annamalai, B., Saravanan, P., & Varadharajan, I. (2023). ABOA-CNN: auction-based optimization algorithm with convolutional neural network for pulmonary disease prediction. *Neural Computing and Applications*, *35*(10), 7463–7474. https://doi.org/10.1007/s00521-022-08033-3

Chanakya, P., Harsha, P., & Pratap Singh, K. (2024). Robustness of Generative Adversarial CLIPs Against Single-Character Adversarial Attacks in Text-to-Image Generation. *IEEE Access*, *12*, 162551–162563. https://doi.org/10.1109/access.2024.3491017

Dai, Y., Qian, Y., Lu, F., Wang, B., Gu, Z., Wang, W., Wan, J., & Zhang, Y. (2023). Improving adversarial robustness of medical imaging systems via adding global attention noise. *Computers in Biology and Medicine*, *164*, 107251. https://doi.org/10.1016/j.compbiomed.2023.107251

Devarajan, R. K., & Khader, S. S. (2023). Pose Sequence-Aware Generative Adversarial Network for Augmenting Skeleton Sequences to Improve Cerebral Palsy Detection by Deep Learner. *International Journal of Intelligent Engineering and Systems*, *16*(5), 512–522. https://doi.org/10.22266/ijies2023.1031.44

Gbashi, S., Maselesele, T. L., Njobeh, P. B., Molelekoa, T. B. J., Oyeyinka, S. A., Makhuvele, R., & Adebo, O. A. (2023). Application of a generative adversarial network for multi-featured fermentation data synthesis and artificial neural network (ANN) modeling of bitter gourd–grape beverage production. *Scientific Reports*, *13*(1), 11755. https://doi.org/10.1038/s41598-023-38322-3

Haque, S. B. ul, & Zafar, A. (2024). Robust Medical Diagnosis: A Novel Two-Phase Deep Learning Framework for Adversarial Proof Disease Detection in Radiology Images. *Journal of Imaging Informatics in Medicine*, *37*(1), 308–338. https://doi.org/10.1007/s10278-023-00916-8

Harini, P., Madhavi, B. N., Latha, B. S., & Sasikumar, A. N. (2024). Optimized self-attention based cycle-consistent generative adversarial network adopted melanoma classification from dermoscopic images. *Microscopy Research and Technique*, *87*(6), 1271–1285. https://doi.org/10.1002/jemt.24506

Hussain, T., Shouno, H., Hussain, A., Hussain, D., Ismail, M., Hussain Mir, T., Rong Hsu, F., Alam, T., & Anonna Akhy, S. (2025). EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification. *IEEE Access*, *13*, 54040–54068. https://doi.org/10.1109/access.2025.3554184

Jiang, S., Wu, Z., Yang, H., Xiang, K., Ding, W., & Chen, Z.-S. (2024). A prior knowledge-guided distributionally robust optimization-based adversarial training strategy for medical image classification. *Information Sciences*, *673*, 120705. https://doi.org/10.1016/j.ins.2024.120705

Kanca Gulsoy, E., Ayas, S., Kablan, Elif. B., & Ekinci, M. (2024). Enhancing the adversarial robustness in medical image classification: exploring adversarial machine learning with vision transformers-based models. *Neural Computing and Applications*, *35*, 1–19. https://doi.org/http://dx.doi.org/10.2139/ssrn.4605358

Kanca, E., Ayas, S., Baykal Kablan, E., & Ekinci, M. (2025). Evaluating and enhancing the robustness of vision transformers against adversarial attacks in medical imaging. *Medical & Biological Engineering & Computing*, *63*(3), 673–690. https://doi.org/10.1007/s11517-024-03226-5

Mohammadi, S. S., & Nguyen, Q. D. (2024). A User-friendly Approach for the Diagnosis of Diabetic Retinopathy Using ChatGPT and Automated Machine Learning. *Ophthalmology Science*, *4*(4), 100495. https://doi.org/10.1016/j.xops.2024.100495

Nazir, K., Kim, J., & Byun, Y.-C. (2024). Enhancing Early-Stage Diabetic Retinopathy Detection Using a Weighted Ensemble of Deep Neural Networks. *IEEE Access*, *12*, 113565–113579. https://doi.org/10.1109/access.2024.3432867

Ng, M. F., & Hargreaves, C. A. (2023). Generative Adversarial Networks for the Synthesis of Chest X-ray Images. *Proceeding of the 3rd International Electronic Conference on Applied Sciences*, 84. https://doi.org/10.3390/asec2022-13954

Pasqualino, G., Guarnera, L., Ortis, A., & Battiato, S. (2024). MITS-GAN: Safeguarding medical imaging from tampering with generative adversarial networks. *Computers in Biology and Medicine*, *183*, 109248. https://doi.org/10.1016/j.compbiomed.2024.109248

Pervin, Mst. T., Tao, L., & Huq, A. (2023). Adversarial attack driven data augmentation for medical images. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(6), 6285. https://doi.org/10.11591/ijece.v13i6.pp6285-6292

Priya, K. V., & Dinesh Peter, J. (2025). Enhanced Defensive Model Using CNN against Adversarial Attacks for Medical Education through Human Computer Interaction. *International Journal of Human–Computer Interaction*, *41*(3), 1729–1741. https://doi.org/10.1080/10447318.2023.2204697

Rahman, M., Roy, P., Frizell, S. S., & Qian, L. (2025). Evaluating Pretrained Deep Learning Models for Image Classification Against Individual and Ensemble Adversarial Attacks. *IEEE Access*, *13*, 35230–35242. https://doi.org/10.1109/access.2025.3544107

Sheikh, B. U. H., & Zafar, A. (2024). Removing Adversarial Noise in X-ray Images via Total Variation Minimization and Patch-Based Regularization for Robust Deep Learning-based Diagnosis. *Journal of Imaging Informatics in Medicine*, *37*(6), 3282–3303. https://doi.org/10.1007/s10278-023-00919-5

Tsai, M.-J., Lin, P.-Y., & Lee, M.-E. (2023). Adversarial Attacks on Medical Image Classification. *Cancers*, *15*(17), 4228. https://doi.org/10.3390/cancers15174228

Vaddadi, S. A., Somanathan Pillai, S. E. V., Addula, S. R., Vallabhaneni, R., & Ananthan, B. (2024). An efficient convolutional neural network for adversarial training against adversarial attack. *Indonesian Journal of Electrical Engineering and Computer Science*, *36*(3), 1769. https://doi.org/10.11591/ijeecs.v36.i3.pp1769-1777

Xu, W., Nie, L., Chen, B., & Ding, W. (2023). Dual-stream EfficientNet with adversarial sample augmentation for COVID-19 computer aided diagnosis. *Computers in Biology and Medicine*, *165*, 107451. https://doi.org/10.1016/j.compbiomed.2023.107451

Zhao, Y., Zheng, J., Gao, X., Liu, L., Zhang, Y., & Zhang, Q. (2025). Enhancing the transferability of adversarial attacks via Scale Enriching. *Neural Networks*, *189*, 107549. https://doi.org/10.1016/j.neunet.2025.107549