

# Generative adversarial deep learning in images using Nash equilibrium game theory

Syeda Imrana Fatima, Yugandhar Garapati

Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad, India

## Article Info

### Article history:

Received Apr 12, 2023

Revised Jul 8, 2023

Accepted Jul 9, 2023

### Keywords:

Canadian Institute for Advance Research

Convolutional neural network

Deep learning

Generative adversarial learning algorithm

Nash equilibrium game theory

## ABSTRACT

A generative adversarial learning (GAL) algorithm is presented to overcome the manipulations that take place in adversarial data and to result in a secured convolutional neural network (CNN). The main objective of the generative algorithm is to make some changes to initial data with positive and negative class labels in testing, hence the CNN results in misclassified data. An adversarial algorithm is used to manipulate the input data that represents the boundaries of learner's decision-making process. The algorithm generates adversarial modifications to the test dataset using a multiplayer stochastic game approach, without learning how to manipulate the data during training. Then the manipulated data is passed through a CNN for evaluation. The multi-player game consists of an interaction between adversaries which generates manipulations and retrains the model by the learner. The Nash equilibrium game theory (NEGT) is applied to Canadian Institute for Advance Research (CIFAR) dataset. This was done to produce a secure CNN output that is more robust to adversarial data manipulations. The experimental results show that proposed NEGT-GAL achieved a greater mean value of 7.92 and takes less wall clock time of 25,243 sec. Therefore, the proposed NEGT-GAL outperforms the compared existing methods and achieves greater performance.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Syeda Imrana Fatima

Department of Computer Science and Engineering, GITAM (Deemed to be University)

Hyderabad, India

Email: isyeda@gitam.in

## 1. INTRODUCTION

The perturbations are created to deep learning (DL) models to make the original image perturbed and resulted in an adversarial image that deceives and confuses the DL model. The DL model got more secure and alerted to adversarial attacks by training the model with creating perturbations on the original image. A generative adversarial network (GAN)-based on the DL model was presented to overcome the problem of the low-quality image which causes poor performance. The low-quality defect images are reconstructed by using the GAN model, and to recognize the reconstructed images a VGG-16 network is developed [1]. The GAN model is mainly used to differentiate between fake and real samples to improve the performance of the model [2]. The process of capturing high-dimensional images was complex in the steganography task hence the AdvSGAN which evaluates the restricted neural coder represents the steganography of images by performing an adversarial game between the adversary model and neural code. The adversarial GAN networks provide high performance in steganography tasks [3]. The GAN based DL methods is also used for tumor classification in magnetic resonance (MR) images to extract the related features and structure of MR images known by the convolutional layers to improve the performance in data augmentation of medical MR images [4]. The quality of the image and the accuracy of computed tomography

(CT) of cone-beam computed tomography (CBCT) is improved using DL methods with GAN networks [5]. The automated medical diagnostics of brain images is performed by a novel method to know high-resolution generative models of brain magnetic resonance imaging (MRI) images using the texture transformation and deformation field together with deep neural networks [6]. The abnormal-to-normal translation GAN (ANT-GAN) model is used in medical images which consist of medical imaging information like classification and lesion segmentation to produce the improved lesion image [7]. Medical images are time-consuming and expensive hence to reduce them and to exploit the required information, a novel semi-supervised DL method that trains the adversarial images itself with specific regularization is presented for medical images in large-scale classification [8].

The DL models mainly depend on the number of training samples which affects the performance of the model hence a multitask generative adversarial network (MTGAN) to extract the highly required information from unlabeled data. The MTGAN performs the classification and reconstruction tasks [9]. The DL models do not give suitable solutions for sophisticated adversaries in image transformation due to their non-differential nature, and filtering of orientation hence non-deep learning approaches are presented. The non-deep learning approach performs well in image transformation like discrete sine transform (DST) and supports vector-based classifiers [10]. Due to the limited availability of labeled samples, traditional deep learning (DL) algorithms were not appropriate. As a solution, a feature-oriented adversarial active learning (FAAL) approach was developed to extract high-level features from an intermediate layer of a DL classifier. These features were then used to design a heuristic-based GAN algorithm. The fake features are evaluated and differentiated between the features of real and fake [11]. DL algorithms have shown impressive performance in various machine learning tasks but they are vulnerable to adversarial attacks, particularly in the form of adversarial images. To address this issue, experts have suggested a new approach called the spatial-frequency ensemble relation network based on GANs. This approach aims to improve the performance of DL models against adversarial attacks on images. The ensemble relation network extracts the features of training images, extracts the relation between images, and transforms the relationship into related categories by using GAN [12]. The representation of features for targeted images was performed by  $A^3GN$  by using channel-wise attention and geo-metric attention to result in improved performance [13]. The effectiveness of traditional image segmentation is performed by a novel two-stage image augmentation architecture which results in the synthetic mask and image pairs. The effectiveness of image segmentation is improved by increasing the size of the training dataset in synthesized image mask pairs [14]. The Nash equilibrium game theory has been applied to Canadian Institute for Advance Research (CIFAR) dataset in the research, to implement generative adversarial learning algorithm (NEGT-GAL) on images to overcome adversarial data manipulations [15].

Yang *et al.* [16] presented a network security task method depending on adversarial DL. The deep auto-encoder-deep neural network (AEDNN) method was proposed by using deep auto-encoder (DAE) for feature extraction and deep neural network (DNN) for binary classification of attacks. AEDNN can manage huge amount of network data and the efficiency, robustness, and generalization of the network was improved. The limitations of AEDNN consists of imbalanced data of various categories in dataset and accuracy need to be improved in finding the minority classes. Further, the calculation method of security needs to be optimized and need to validate the model in detail. Jeong *et al.* [17] presented the accuracy for classification model using DL and decreased the malicious attackers. The Modified National Institute of Standards and Technology (MNIST) dataset with image samples and NSL-KDD dataset consists the data of network was taken as the input datasets. The accuracy was calculated by giving the adversarial samples into convolutional neural network (CNN) and auto-encoder classification models which were designed using the libraries of PyTorch and TensorFlow. The exploited adversarial samples consist of insufficient data can cause to a great damage to the performance of model. Further, the accuracy was improved by applying the learning method of recurrent neural network (RNN) and deep fool method. Ma *et al.* [18] presented the analysis of issues on adversarial attacks on DL depending on medical image analysis. The experiment was conducted by four various detection methods and attacks on three medical image datasets. The DNN models of medical images more vulnerable for adversarial attacks compared to natural images. The medical adversarial attacks detection is simple and can evaluate 98% detection accuracy. The wrong decision in medical images leads to difficulties in finding the adversarial attacks on medical images compared to natural images. Further, defense approaches need to improve the robustness of adversarial medical images. Huang *et al.* [19] presented the synthetic aperture radar (SAR) image recognition methods depended on DL models. The adversarial examples of SAR images are generated which was used to attack the classical DL models. The performance of the adversarial attacks was tested using SAR images and the advantages evaluated are automatic train classifiers and learning features which improves the performance of image recognition methods. The drawbacks of SAR image recognition were overfitting and it was not suitable for other attack models except iterative least likely class method (ILCM) algorithm. Rahman *et al.* [20] presented a novel defense

Mockingbird in machine learning domains related to adversarial attacks of website fingerprinting (WF) which was a traffic analysis attack, improves the accuracy up to 98%. Initially the straight forward techniques were used to save the traces against adversarial attacks but it does not give robust classification. Hence the Mockingbird technique evaluated for traces and gives adversarial training by moving in the space of traces where the predictable gradients are not followed. The main advantage of it is having high security of the data against adversarial attacks. There are limitations associated with high computational time as well as the rapid transmission of large amounts of information. Further, leverage Mockingbird for server-side and make it more robust against other models.

According to Pasini *et al.* [21], deep convolutional (DC-GANs) model training could be distributed. The task distribution relies on categorizing the data, and it is decided by the qualities that differ between the classes of data, which take precedence over the features that differ between data points of identical classes. The pattern of the wall-clock time shows that the training of distributed DC-CGANs on all four datasets resulted in less scaling as long as the computational workload for every message passing interface (MPI) procedure stays unchanged. Because, the computational time required to complete the training was not affected by the increase in the number of data classes. Meanwhile, the distributed DC-CGANs produced results that were almost relevant and already included in the CIFAR10 dataset. A distributed method to train DC-CGANs models has been shown by Pasini *et al.* [21]. By dividing the training data into groups based on data labels, this strategy lessens the disparity among the generator and discriminator and improves scalability by doing parallel training on many generators, each of which was trained with a single data label in mind. The variation across classes is eliminated during data splitting based on labels, which also corrects the imbalance in normal GAN training. Since everyone is independent of the others, the generators were trained simultaneously, which improves scalability. Quicker data processing and steady training may not constantly translate into faster convergence. Karim *et al.* [22] presented an adversarial transform network (ATN) model to attack different time series classification models and a distilled model was used to observe the behavior of classification models of time series. The student-teacher framework was used in proxy attacks on a specific model. The capability of generalizing adversarial models performed well on the samples which do not see before by the adversarial models. The disadvantage of ATN was not suitable for time series classification models which influences model robustness and performance evaluation. Further, time series classification models need to be developed for the targeted adversaries.

The main contributions involved in this research are given as follows: i) a Nash equilibrium game theory (NEGT) is used to attain the equilibrium condition as a result, of more similar manipulation images; ii) a generative adversarial learning (GAL) algorithm is used to generate additional data which helps to improve the training data; iii) also, the wall clock time is reduced by generating the additional data using the GAL algorithm. Moreover, the structure of this paper is given as follows: the overall concept of proposed methodology and the mathematical equations for NEGT are explained in section 2 and section 3. Whereas the experimental results of the proposed methodology are presented section 4. Finally, the conclusion of the overall work is given in section 5.

## 2. PROPOSED METHOD

Nash equilibrium game theory (NEGT) [23] has been developed on the CIFAR dataset to identify a manipulative change in the data that can influence the decision boundaries of the learner. Through this approach, it is possible to adjust many favorable labels to negative values. The Nash equilibrium assists a player in determining the optimum incentive in a circumstance based not just on their actions but also on the choice of the other people involved. NEGT identifies a set of tactics that are mutually optimal for both the learner and the opponent [24]. Here, neither the learner nor the opponent has any incentive to deviate from these tactics. The learner would then be retrained across all adversarial data manipulations made by many players to suggest a secure CNN that is resistant to future adverse data manipulations. Figure 1 (see in appendix) shows the flow chart demonstrating a two-player games.

### 2.1. Dataset validation

A labelled input training data  $X_{train}$  is used to train CNN original and then is tested using testing data  $X_{test}$ . To obtain manipulated CNN, the adversarial data manipulation function ( $\alpha^*$ ) is added to  $X_{test}$  data. The data is considered from the CIFAR dataset [25], [26].

## 3. NASH EQUILIBRIUM GAME THEORY

The adversarial learning is simulated by the training algorithm as a fixed sum Stackelberg game among two players, an adversary who plays leader (L) role and learner who plays the follower (F) role. The game begins with the leader taking the first action/move/play. The adversary's gain is thought to represent

the learner's loss in a constant sum game, and vice versa. During each encounter, each of the opponents and the learner executes a move [27]. A target CNN is tested on  $X_{test}$  after being trained on  $X_{train}$ . The objective of the game is to identify  $\alpha^*$  such that  $X_{test} + \alpha^*$  reduces this CNN's performance found on  $X_{test}$ .

### 3.1. Adversary (leader)

The adversary is supposed to be looking for genuine positives with only an understanding of the learner's class label errors. The adversary seeks data changes that increase classification error,  $error(\omega)$  using an evolutionary algorithm. In the evolutionary algorithm,  $J_L(\alpha, \omega)$  is defined as the fitness function which improves the game's progresses. The game has converged when the opponent does not notice an improvement in the payout function or when the highest number of repetitions is achieved. The game convergence criteria are determined by the evolutionary algorithm's search and optimization criteria in each game round. The game eventually devolves into an aggressive data manipulation game using weights  $\omega$  on the learner. Each player is assigned to the L and F strategy areas A and W, respectively [28]. The strategy space is a set of possible moves for each participant. The reward function  $J_L$  and  $J_F$  of the player determines the result of a strategy. For a specific statement of  $\omega \in W$ , the best strategy  $\alpha^* \in A$  for the leader is expressed in (1),

$$\alpha^* = \arg \max_{\alpha \in A} J_L(\alpha, \omega) \quad (1)$$

For labelled input training data  $X_{train}$ ,  $X_{test}$  available during the game, the adversary seeks a move that maximizes the fitness function  $J_L()$ , where,  $error(\omega)$  is the categorization error evaluated by recall for the current adversary data. The term cost is the  $\ell_2$  norm for the current  $\alpha$ . Hence, the (2) to (4) is written as:

$$J_L(\alpha, \omega) = 1 + \lambda * error(\omega) - cost(\alpha) \quad (2)$$

$$error(\omega) = 1 - recall(\omega) \quad (3)$$

$$cost(\alpha) = \|\alpha\|_2 \quad (4)$$

The negative cost ( $\alpha$ ) term ensures that the adversary makes as little alteration to the current  $\alpha$  while maximizing the positive error ( $\omega$ ) term. By necessity, the fitness function maximizes error ( $\omega$ ) by minimizing the associated recall ( $\omega$ ). For each iteration of the game, recall ( $\omega$ ) is computed on the manipulated training data  $X_{train} + \alpha$  and the iteration that produces the highest value for  $J_L(\alpha, \omega)$  is chosen for subsequent iterations.  $cost(\alpha)$  is improved by an empirically determined weighting term for each dataset. To provide a positive fitness value in the evolutionary process, a constant 1 is added to  $J_L(\alpha)$ .

### 3.2. Learner (follower)

CNN acts as the learner. The input and output layers of the CNN architecture are available in TensorFlow as the CIFAR10 model. Convolution layers, max pooling layers, regularization layers, and activation units comprise CNN's input layers. The CNN has a softmax probability distribution function output layer. The CNN's input and output layers define the learner's overall loss function.

Following the adversary's attack, the learner retrains the model. At equilibrium, the adversary can uncover examination data that is notably distinct from the dataset, while the learner can modify its model with antagonistic data to account for new threats. For a given observation of  $\omega \in W$ ,

For  $L$ 's move  $\alpha$ ,  $F$ 's best strategy is formulated in (5).

$$\omega^* = \arg \max_{\omega \in W} J_F(\alpha, \omega) \quad (5)$$

The empirical difference among the given input training data  $X_{train}$  and the adversary testing distribution of data is characterized by  $X_{train} + \alpha^*$  terms of the attacker's cost  $cost(\alpha)$  as well as the learner's error  $error(\alpha)$ . Throughout the game's versions, we can find  $\alpha$  that maximizes the adversary's payoff  $J_L(\alpha, \omega)$  by manipulating the training data distribution  $X_{train}$  into  $X_{train} + \alpha$ . After game convergence, we can find the  $\alpha^*$  that minimizes the learner's payoff  $J_F(\alpha, \omega)$  by manipulating the dispersion of testing data  $X_{test}$  into  $X_{test} + \alpha^*$ . The CNN is re-trained on the new bridge sample to modify DL processors for hostile data [29].

## 4. RESULTS AND DISCUSSION

This segment provides the results and analysis of the proposed NEG-T-GAL [30] model where is implemented and simulated using Python 3.7 software whereas the computer is powered by the INTEL i5 processor running at 2.4 GHz with 16 GB RAM on Windows 10 OS. Where, Figure 2 depicts data

integration on strong labels that seem to be unfavorable in the harmful procedure. Adding and removing pixels, as well as modifying the shape and scale of the picture, are examples of changes that prevent detection.

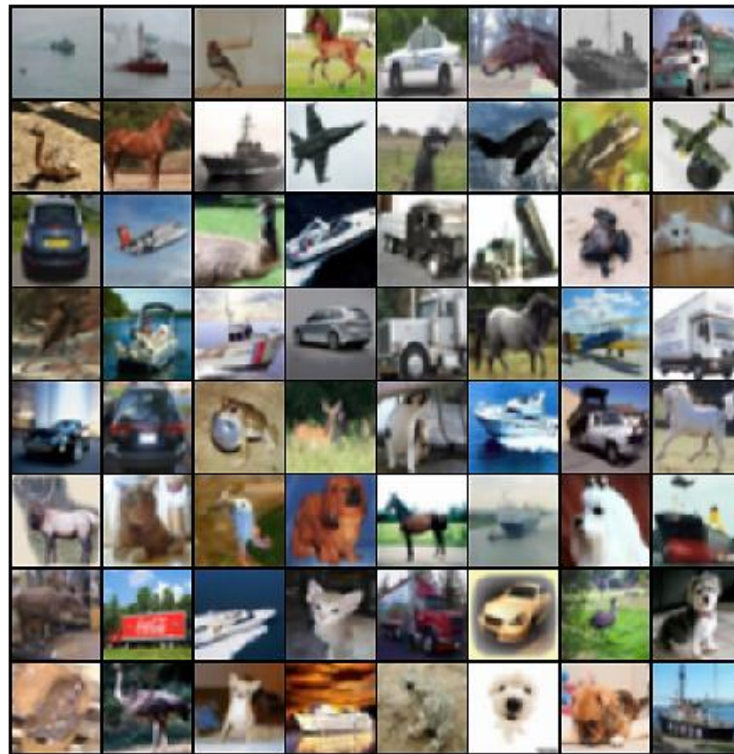


Figure 2. Original images

In Figure 3, the images from the CIFAR dataset have undergone a form of alteration where additional pixels have been added to make them appear similar. This manipulation can impact the accuracy of image recognition algorithms trained on the dataset, as they may struggle to distinguish between artificially modified images. In Figure 4, the animal face has been altered to seem the same by altering the thickness and shape. In Figure 5, overall images from the CIFAR dataset images generated by adding and removing pixels. Figure 5 shows that the CIFAR landscape images were modified to appear as if 9 pixels had simply been added without altering the geometry of the images.

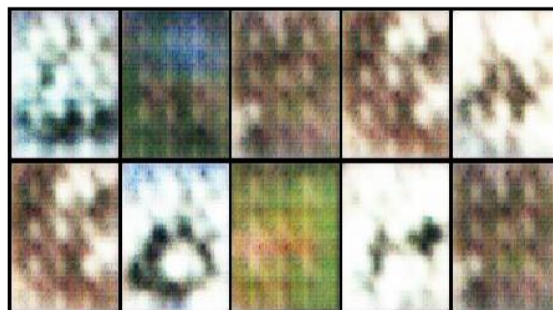


Figure 3. Manipulated images

#### 4.1. Comparative analysis

The proposed NEGT-GAL algorithm is used to reduce the adversarial data manipulations which result in a secured CNN as an output. The existing algorithms like synthetic aperture radar (SAR), adaptive

iteration fast gradient method (AI-FGM), and adversarial transform network (ATN) models do not suitable to provide high security in adversary manipulations. The adversarial manipulations were analyzed and the features were recognized by using the SAR image on DL algorithms. The SAR images was well suitable to report the manipulations that occurred but the overfitting and not suitable for algorithms except ILCM algorithm [19]. The AI-FGM method uses a gradient searching process in an iterative method related to adversarial attacks on DL to result in high success in adversary attacks by the modifications done in pixels of the image but it does not give suitable performance [21]. An ATN model was utilized to launch attacks against various time series classification models. However, this approach may not yield precise outcomes for time series classification [22]. Hence to result in a secured clear image that can give high performance in any manipulations, a NEGT is applied to CIFAR dataset in GAL algorithm. The loss function of generator and discriminator performance on CIFAR data manipulations with Nash equilibrium has been plotted against each other for its monotonically generated images and classification in Figure 6.

Here, Tables 1 to 4 represents the comparison analysis between various existing methods. The following are the methods used for the comparison (i.e.) DC-GAN, deep convolutional conditional GANs (DC-CGAN's), distributed DC-CGANs [21] and the proposed NEGT-GAL method. By using those methods, various performance parameters such as Wall-clock time, mean, standard deviation (SD), Fréchet inception distance (FID) are evaluated.



Figure 4. Manipulated images from CIFAR



Figure 5. Algorithm generated images

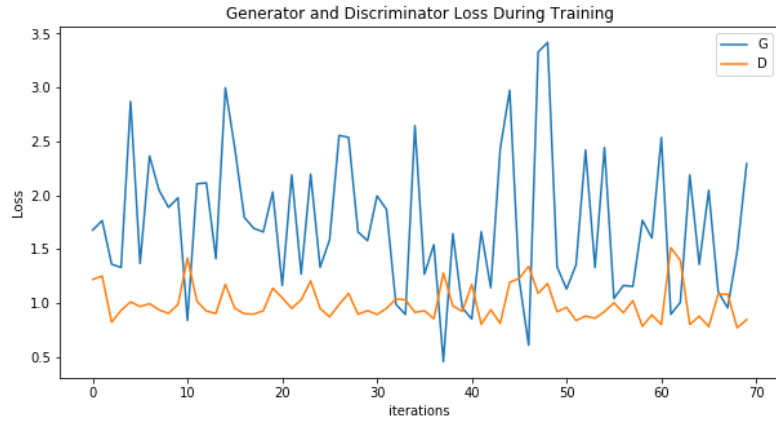


Figure 6. Loss variation of generator and discriminator

Table 1. Comparison of wall-clock time

Methods	Wall-clock time (seconds)
DC-GANs	3,75,520
DC-CGAN's	3,93,100
Distributed DC-CGANs [21]	39,011
Proposed NEGT-GAL	25,243

Table 2. Comparison of mean value

Methods	Mean value
DC-GANs	4.39
DC-CGAN's	5.69
Distributed DC-CGANs [21]	6.43
Proposed NEGT-GAL	7.92

Table 3. Comparison of SD

Methods	SD
DC-GANs	0.28
DC-CGAN's	0.31
Distributed DC-CGANs [21]	0.25
Proposed NEGT-GAL	0.18

Table 4. Comparison of FID

Methods	FID
DC-GANs	14.13
DC-CGAN's	11.12
Distributed DC-CGANs [21]	9.41
Proposed NEGT-GAL	6.9

Table 1 shows that the wall clock time consumed by the DC-GAN is 375,520 sec, DC-CGAN's is 393,100 sec and distributed DC-CGANs [21] is 39,011 sec whereas the proposed NEGT-GAL method consumes wall clock time of 25,243 sec. Therefore, the proposed NEGT-GAL takes less time while compared to the existing methods and outperforms the existing methods. Figure 7 illustrates the graphical comparison of the proposed NEGT-GAL with existing methods in terms of wall clock time.

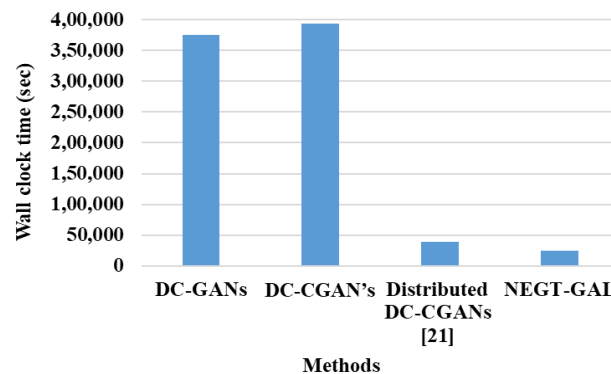


Figure 7. Graphical comparison in terms of wall clock time

Similarly, Table 2 shows mean value of the DC-GAN, DC-CGAN's, and distributed DC-CGANs [21] as 4.39, 5.69, and 6.43 respectively. Whereas, the proposed NEGT-GAL method achieved a maximum mean value of 7.92 which is greater than the compared existing methods. Here also the proposed NEGT-GAL method outperforms the compared existing methods. The graphical comparison of the proposed NEGT-GAL with existing methods in terms of mean value is illustrated in Figure 8.

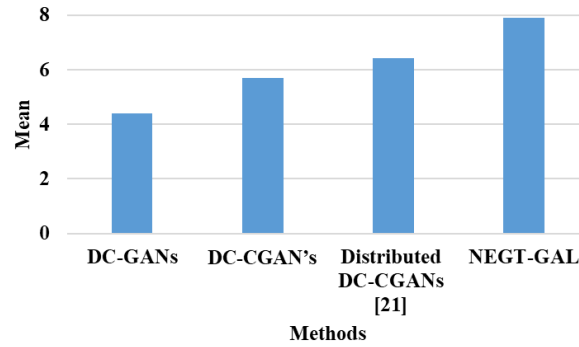


Figure 8. Graphical comparison in terms of mean

Here, Table 3 shows the achieved SD value of the DC-GAN, DC-CGAN's, and distributed DC-CGANs [21] as 0.28, 0.31, and 0.25 respectively whereas, the proposed NEGT-GAL method achieved a SD of 0.18. So, the proposed NEGT-GAL achieved a less SD value and outperforms the compared existing methods. Figure 9 illustrates the graphical comparison of the proposed NEGT-GAL with existing methods in terms of SD.

Finally, the DC-GAN, DC-CGAN's, and distributed DC-CGANs [21] achieved FID of 14.13, 11.12 and 9.41 respectively which is given in Table 4 whereas, the proposed NEGT-GAL method achieved a minimum FID of 6.90. Therefore, the results demonstrate that the proposed NEGT-GAL outperforms the DC-GAN, DC-CGAN's, and distributed DC-CGANs [21] in terms of FID. The graphical comparison of the proposed NEGT-GAL with existing DC-GANs, DC-CGAN's and distributed DC-CGANs [21] using various parameters is illustrated in Figure 10.

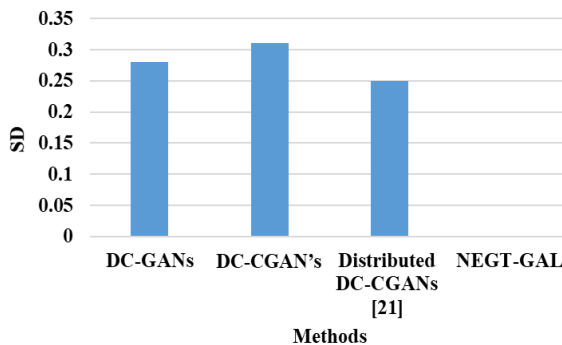


Figure 9. Graphical comparison in terms of SD

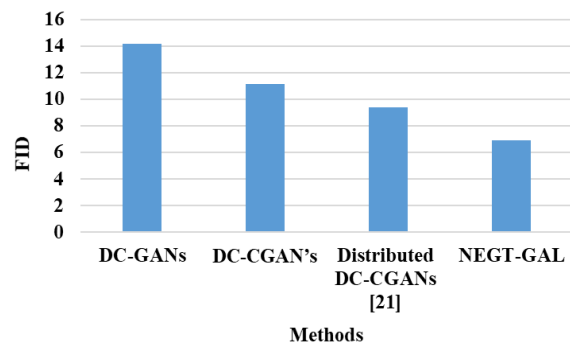


Figure 10. Graphical comparison in terms of FID

## 5. CONCLUSION

The complex task of adversarial data manipulations was reduced by presenting a GAL algorithm using NEGT. NEGT is applied to the CIFAR dataset as the input dataset whereas a fitness function is employed in a sequential game. During the sequential game, an adversary manipulates the input CIFAR dataset multiple times, which affects the learner's assessment results. In the game theory the adversary generates the manipulations on data and the learner retains all the manipulations held by the adversary and resulting in the secured CNN as output. The generative adversarial algorithm converges the affecting performance of testing on adversarial manipulations in DL networks which improves the security of adversarial manipulations. The Generative adversarial algorithm including sequential games with both players and stochastic games in deep neural networks resulted in an improved performance in secured CNN. Moreover, to evaluate the results of the proposed NEGT-GAL algorithm, it is compared with the conventional approaches such as DC-GAN, DC-CGAN's, and distributed DC-CGANs. The proposed NEGT-GAL achieved a greater mean value of 7.92, minimal SD of 0.18, minimal FID of 6.9 and less wall clock time of 25,243, which are superior when compared to the existing methods. In the future, some modifications will be included in the GAL algorithm to improve the classification performance.



## APPENDIX

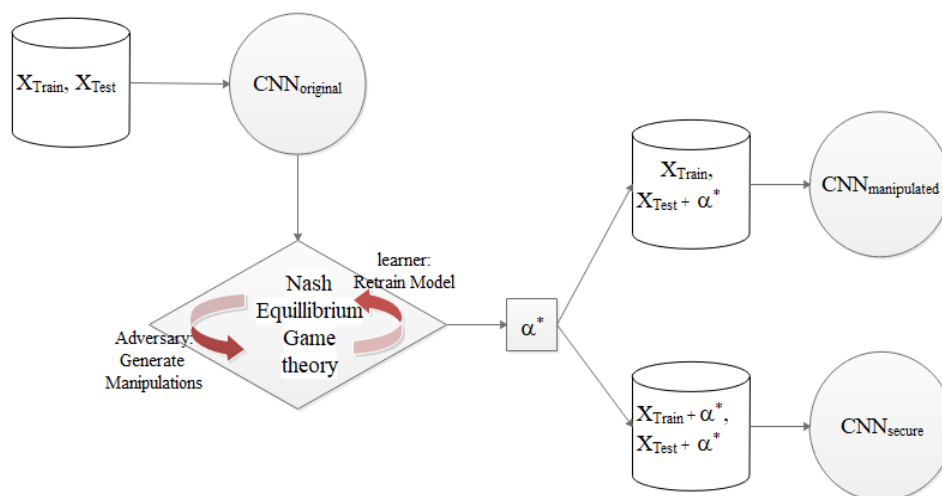


Figure 1. A flow chart demonstrating a two-player game




## REFERENCES

- [1] Y. Gao, L. Gao, and X. Li, "A generative adversarial network based deep learning method for low-quality defect image reconstruction and recognition," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3231–3240, May 2021, doi: 10.1109/TH.2020.3008703.
- [2] T. Alipour-Fard and H. Arefi, "Structure aware generative adversarial networks for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5424–5438, 2020, doi: 10.1109/JSTARS.2020.3022781.
- [3] L. Li, M. Fan, and D. Liu, "AdvSGAN: Adversarial image steganography with adversarial networks," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25539–25555, Jul. 2021, doi: 10.1007/s11042-021-10904-1.
- [4] N. Ghassemi, A. Shoeibi, and M. Rouhani, "Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images," *Biomedical Signal Processing and Control*, vol. 57, Mar. 2020, doi: 10.1016/j.bspc.2019.101678.
- [5] Y. Zhang *et al.*, "Improving CBCT quality to CT level using deep learning with generative adversarial network," *Medical Physics*, vol. 48, no. 6, pp. 2816–2826, Jun. 2021, doi: 10.1002/mp.14624.
- [6] C. K. Chong and E. T. W. Ho, "Synthesis of 3D MRI brain images with shape and texture generative adversarial deep neural networks," *IEEE Access*, vol. 9, pp. 64747–64760, 2021, doi: 10.1109/ACCESS.2021.3075608.
- [7] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley, "An adversarial learning approach to medical image synthesis for lesion detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2303–2314, Aug. 2020, doi: 10.1109/JBHI.2020.2964016.
- [8] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical Image Analysis*, vol. 70, May 2021, doi: 10.1016/j.media.2021.102010.
- [9] R. Hang, F. Zhou, Q. Liu, and P. Ghamisi, "Classification of hyperspectral images via multitask generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1424–1436, Feb. 2021, doi: 10.1109/TGRS.2020.3003341.
- [10] A. Agarwal, R. Singh, M. Vatsa, and N. K. Ratha, "Image transformation based defense against adversarial perturbation on deep learning models," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2020, doi: 10.1109/TDSC.2020.3027183.
- [11] G. Wang and P. Ren, "Hyperspectral image classification with feature-oriented adversarial active learning," *Remote Sensing*, vol. 12, no. 23, Nov. 2020, doi: 10.3390/rs12233879.
- [12] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Fighting fire with fire: A spatial-frequency ensemble relation network with generative adversarial learning for adversarial image classification," *International Journal of Intelligent Systems*, vol. 36, no. 5, pp. 2081–2121, May 2021, doi: 10.1002/int.22372.
- [13] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 855–875, Jan. 2021, doi: 10.1007/s11042-020-09604-z.
- [14] S. Pandey, P. R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," *Biomedical Signal Processing and Control*, vol. 57, Mar. 2020, doi: 10.1016/j.bspc.2019.101782.
- [15] A. S. Chivukula and W. Liu, "Adversarial deep learning models with multiple adversaries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1066–1079, Jun. 2019, doi: 10.1109/TKDE.2018.2851247.
- [16] H. Yang, R. Zeng, G. Xu, and L. Zhang, "A network security situation assessment method based on adversarial deep learning," *Applied Soft Computing*, vol. 102, Apr. 2021, doi: 10.1016/j.asoc.2021.107096.
- [17] J. Jeong, S. Kwon, M.-P. Hong, J. Kwak, and T. Shon, "Adversarial attack-based security vulnerability verification using deep learning library for multimedia video surveillance," *Multimedia Tools and Applications*, vol. 79, no. 23–24, pp. 16077–16091, Jun. 2020, doi: 10.1007/s11042-019-7262-8.
- [18] X. Ma *et al.*, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, Feb. 2021, doi: 10.1016/j.patcog.2020.107332.
- [19] T. Huang, Q. Zhang, J. Liu, R. Hou, X. Wang, and Y. Li, "Adversarial attacks on deep-learning-based SAR image target recognition," *Journal of Network and Computer Applications*, vol. 162, Jul. 2020, doi: 10.1016/j.jnca.2020.102632.




- [20] M. S. Rahman, M. Imani, N. Mathews, and M. Wright, "Mockingbird: Defending against deep-learning-based website fingerprinting attacks with adversarial traces," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1594–1609, 2021, doi: 10.1109/TIFS.2020.3039691.
- [21] M. L. Pasini, V. Gabbi, J. Yin, S. Perotto, and N. Laanait, "Scalable balanced training of conditional generative adversarial neural networks on image data," *The Journal of Supercomputing*, vol. 77, no. 11, pp. 13358–13384, Nov. 2021, doi: 10.1007/s11227-021-03808-2.
- [22] F. Karim, S. Majumdar, and H. Darabi, "Adversarial attacks on time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3309–3320, Oct. 2021, doi: 10.1109/TPAMI.2020.2986319.
- [23] P. Adesso, M. Barni, M. Di Mauro, and V. Matta, "Adversarial kendall's model towards containment of distributed cyber-threats," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3604–3619, 2021, doi: 10.1109/TIFS.2021.3082327.
- [24] Y. Kartal, A. T. Koru, F. L. Lewis, Y. Wan, and A. Dogan, "Adversarial multiagent output containment graphical game with local and global objectives for UAVs," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 2, pp. 875–886, Jun. 2023, doi: 10.1109/TCNS.2022.3210861.
- [25] N. Mani, M. Moh, and T.-S. Moh, "Defending deep learning models against adversarial attacks," *International Journal of Software Science and Computational Intelligence*, vol. 13, no. 1, pp. 72–89, Jan. 2021, doi: 10.4018/IJSSCI.2021010105.
- [26] S. Sarkar, S. Agrawal, T. Baker, P. K. R. Maddikunta, and T. R. Gadekallu, "Catalysis of neural activation functions: Adaptive feed-forward training for big data applications," *Applied Intelligence*, vol. 52, no. 12, pp. 13364–13383, Sep. 2022, doi: 10.1007/s10489-021-03082-y.
- [27] Y. Zhou, M. Kantarcioglu, and B. Xi, "A game theoretic perspective on adversarial machine learning and related cybersecurity applications," in *Game Theory and Machine Learning for Cyber Security*, Wiley, 2021, pp. 231–269.
- [28] A. S. Chivukula, X. Yang, B. Liu, W. Liu, and W. Zhou, "Game theoretical adversarial deep learning," in *Adversarial Machine Learning*, Cham: Springer International Publishing, 2023, pp. 73–149.
- [29] L. Yin *et al.*, "Haze grading using the convolutional neural networks," *Atmosphere*, vol. 13, no. 4, Mar. 2022, doi: 10.3390/atmos13040522.
- [30] F. Ni, Z. He, S. Jiang, W. Wang, and J. Zhang, "A generative adversarial learning strategy for enhanced lightweight crack delineation networks," *Advanced Engineering Informatics*, vol. 52, Apr. 2022, doi: 10.1016/j.aei.2022.101575.

## BIOGRAPHIES OF AUTHORS



**Syeda Imrana Fatima**    is research scholar in Computer Science and Engineering department at GITAM Deemed University Hyderabad, Telangana, India. Her area of interest is artificial intelligence, machine learning and deep learning. she has also worked on cloud computing. She has completed her bachelors in information technology and masters in computer science and engineering. She can be contacted at email: isyeda@gitam.in.



**Yugandhar Garapati**    is Assistant Professor at Computer Science and Engineering Department GITAM Deemed University Hyderabad, Telangana, India. He completed his diploma in computer science and engineering in (2004), and bachelors in computer science and engineering in (2007), he has done his masters in computer science and technology in (2010), he completed his Ph.D. in CSE in (2018). He is recognized and awarded as Associate Fellow from Andhra Pradesh Akademi of Sciences in 2018, and Elevated as Senior Member of IEEE and ACM in 2021. He can be contacted at email: ygarapat@gitam.edu.