Journal of Basrah Researches (Sciences) **51**(2), 270 (2025)
DOI: https://doi.org/10.56714/bjrs.51.2.19

Robust and Interpretable Chest X-ray Classification via Diffusion Purification and Concept-Based Adversarial Detection

Amna Kadhim Ali*

College of Veterinary medicine, University of Basrah, Basrah, Iraq.

ARTICLE INFO

Received 09 November 2025
Revised 21 December 2025
Accepted 24 December 2025
Published 31 December 2025

Keywords:

Concept Activation Vectors (TCAV), Medical Images, Random Forest, Resnet18.

Citation: A. K. Ali., J. Basrah Res. (Sci.) **50**(2), 270 (2025). DOI:https://doi.org/10.56714/bjrs.51.2.19

ABSTRACT

Adversarial attack is an approach that primarily compromise the integrity of deep learning system in medical imaging by adding subtle changes to the model inputs that humans don't notice, it leads the model to make an incorrect decision, so the attacker can corrupt both the input data and the prediction system. In this paper A hybrid detection method based on diffusion purification, deep feature extraction, and ensemble learning has been proposed to address this issue using chest x-ray images.
Two phases make it up. First, cleansing the adversarial samples. This is accomplished by a diffusion model. The samples are regenerated to remove the small perturbations fully. In the second phase, the images are categorized into two groups. Clean and purified images are collected in a safe category, while harmful images are unsafe category. In the second stage, we utilize a pre-trained ResNet18 feature extractor to retrieve salient from chest X-rays. Furthermore, a Random Forest model classifies the features obtained from the ResNet18 model into harmful and non-harmful image. The experimental results show that the proposed framework can detect more than 99% of adversarial samples. The satisfactory detection rate of the proposed purification and ensemble-based feature detection for making AI-assisted disease detection more reliable and safer.

1. Introduction

Computer-aided medical image analysis through deep learning and machine learning has proven to be effective. Nonetheless, adversarial perturbations pose a severe threat to medical utilization [1]. The high sensitivity of AI-assisted diagnostic systems to even minimal adversarial attacks with modest effects may cause severe false positives in medical imaging [2].
Studies emphasize the necessity of employing modern technologies to protect medical images due to their sensitivity and importance, which must be differ from traditional identification methods [3]. which are no longer effective in detecting adversarial processes such as malicious training and pre-processing [4].
Despite the wide body of recent research on common images and adversarial attacks, limited work has been reported in medical imagery yet due to its special considerations e.g., maintaining anatomy consistencies during defense [5]. One such one is medical images –e.g., chest X-rays, where models have been underperforming in achieving strong and generalizable routine. This challenge is due to

***Corresponding author email:** amna.kadhim@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq

variations in information content, imaging protocol differences, and diverse clinical readings by specialists [6]. These challenges make the development and evaluation of defense methods more difficult; subtle imperceptible perturbations in medical images could disrupt important clinical topographies.

Mixed architectures presented as Deep Learning models with classical classifiers have demonstrated great potential for adversarial detection, illustrated by astonishing performance in medical imaging tasks [7]. Another example of works towards this direction is the diffusion model that has found to be effective for denoising and clean up against adversarial examples [8, 9].

Lately, deep feature representations by leading ML models such as ResNet can be fed to random forest classifiers and neural networks to yield robust, interpretable systems that are high performing across different domains, specifically in adversarial image perturbations within the realm of medical imaging. They have demonstrated their capability to capture discriminative features, and are therefore ideal for adversarial detection in medical domain [10, 11]

Despite the existence of specialized detection systems capable of adapting to multiple strategies designed to counter attacks such as FGSM and PGD, the lack of standardized evaluation criteria continues to hinders a fair comparison between methods of filtering and detecting attacks in medical images [12,13]. Recent research has focused on developing real-time, interpretable, and effective defense strategies that are suitable for clinical medical images [14].

There are significant research gaps in the field that remain unexplored despite the published results. One of them is the lack of consideration for adversarial detection within the framework of clinical healthcare models. Namely, the existing methodologies for detection in clinical settings are typically founded on the application of traditional CNN-based classifiers. Therefore, they are often void of specialized detection approaches and reliable preprocessing to counterattack adversarial perturbations. The diffusion-based purification could mitigate adversarial noise, but its application to medical imaging is highly constrained. Moreover, to my best knowledge, no study has thoroughly examined several adversarial attacks, i.e., FGSM, PGD, BIM, DeepFool, at once in the medical imaging context in a systematic and comparable way. As a result, the lack of benchmarks prohibits the integration of existent defense strategies into a clinical setting. This is the research gap that I am going to fill with the target study. Develop an effective mechanism for detecting adversarial images in chest X-ray datasets.

1. Enhance detection accuracy by integrating the features generated by training ResNet18 with a Random Forest classifier.
2. Exploring the effectiveness of diffusion-based purification on reversing adversarial degenerations on actual medical images.
3. Examining the identity of the framework against different types of adversarial attack.
4. Evaluability of class correlation between clean, adversarial, and purified datasets for determining defense efficacy.
5. TCAV to help with interpretable: to ensure semantic transparency and directional robustness.

2. Related Work

Several recent studies have introduced different adversarial detection and defense techniques like semantic feature analysis and diffusion-based purification. While these methods have potential implementation for real-life settings, many are confined by either poor generalizability, interpretability, or a densification of how devices operate in a realistic demographic.

survey [15] classifies adversarial attacks into gradient-based, optimization-based, and generative approaches, and reviews defensive methods such as adversarial training, input preprocessing, and robustness enhancement. The work also underscores the trade-off between robustness and accuracy, and exposes the lack of generalizability in a variety of defense methods. While it combines several methods, it does not provide quantitative criteria and experimental comparisons especially in medical domain.

This paper [16] leverages the diffusion-based adversarial attacks: AdvDiffuser and DiffAttack to craft transferable realistic alterations. Although the stealth and success rate of these kind of attacks are higher than GAN-based ones, they found that previous defense mechanisms including adversarial training and input bottleneck fail to defend against them. Although these new attack types would

pose a huge threat, the work does not propose or evaluate potential countermeasures and light-weight detection methods specifically designed for propagation-based adversarial examples are missing.
The review [17] introduces the TPre-ADL framework and focuses on integrating robustness, interpretability, and trustworthiness. It introduces ensemble learning, defensive distillation, and robust training strategies to state that robustness is limited by accuracy on clean data. As the presented framework is mostly theoretical and not validated in practice, this concept may need additional evaluation to be effective. The authors also presume a lack of research on defending multiple attack types with interpretable outputs, which is a critical gap in the development of defense systems.
This work [18] introduces a detection method named Concept Activation Vectors for measuring the degree to which adversarial perturbations perturb high-level concepts when injected into deep models and achieves < 95% detection accuracy with strong transferability for GoogLeNet and ResNet34. The model is interpretable and requires little computation, but it has not been tested on newer attack types such as diffusion- or patch-based attacks and has not been tested on medical image perturbations.
in paper [19] the researchers proposed a text-based detection method. This measure converts CNN feature maps into text embeddings to identify adversarial changes through sentiment analysis. Text-based detection utilizes the measure in less than 5 ms per sampling and is capable of obtaining the same precision as the original CNN detector. However, this has just been used to straightforward datasets like CIFAR-10 and SVHN, its ability to high-resolution healing photos or large models remains unmeasured.
Study [20] introduces UnMask,a method for Detecting adversarial examples with UnMask by measuring the semantic features expected and heating from the input. For example, an image labeled “bird” has to fire “feathers” and “eyes”. However, adversarial perturbations stop this. It reaches 93% accuracy with PGD combat and is 31% better in classification accuracy under attack than adversarial training. Nevertheless, it has not yet been tested against recent attack types thanks to diffusion or patch-based methods and real-time application uncertainties.
the paper [21] evaluates statistical deviations in the intermediate layer outputs of numerous DNN architectures and modalities (images, audio, video). It demonstrates the following wide usability and higher detection accuracy and computational efficiency conducive to almost real-time application. In contrast, the technique exhibits insufficient detailed complexity analysis and superior adversarial attacks of high-risk domains such as media imaging.
The method [22] compares the semantic features of clean and adversarial inputs and quantifies the lack of equality in comparison to the original input. It is not sensitive to different perturbation quantities and preprocessing techniques and demonstrates stable detection behavior in the variety of attack machinations. Nevertheless, the model was tested only on CIFAR-10, and there are no comparisons to the best existing models; thus, it is unclear whether the mechanism would suit complex datasets or clinical use.
Research [23] explores conditional diffusion models that employ a denoising technique applied iteratively to adversarially perturbed images to object detections. This approach’s attack success rate was significantly reduced, while detection accuracy was maintained, indicating robustness with reasonable computational overhead. Unfortunately, this approach is limited to object detection and may be impractical in real-time or resource-constrained conditions.

3. Methodology

In this paper, we proposed a multi-stage adversarial detection framework for chest X-ray images based on purification by diffusion, deep feature extraction by ResNet18, classification by Random Forest. Our implementation is in Python using PyTorch and scikit-learn and contains four stages including data preprocessing, feature extraction, model training and evaluation.

3.1. Dataset Description and Preprocessing

This study uses the Chest X-Ray Images (Pneumonia) dataset, available online from Kaggle. This dataset contains a total of 5,863 grayscale chest X-ray images classified into two groups: Normal and Pneumonia, which include both bacterial and viral cases. They are taken in the standard anterior-posterior (AP) projection and saved in compressed JPEG format.

The Chest X-ray images downloaded were divided into three sets: training, test, and validation. Only the training and test sets were used, as the validation set contains a very small number of images (approximately 16), which does not represent a complete sample. The training set includes 5,216 images, while the test set contains 647 images. These sets are entirely independent, with no overlap between the images, thereby ensuring the absence of data leakage.
Preprocessing steps include resizing all images into 224×224 pixels for ResNet18 input, converting images into 1-channel grayscale tensors, and normalization into pixel values with a mean of 0.485 and standard deviation of 0.229. The dataset consists of three distinct types of images: clean images, i.e., original chest X-rays before any perturbations, adversarial images, i.e., clean images after perturbations from the FGSM, PGD, BIM, and DeepFool methods, the key parameters used for the four attacks are presented in Table 1.

Table 1. the main parameters employed in FGSM, PGD, BIM, and DeepFool attacks.

Attack	ε value	Iterations	Step size (α)	Norm type	Targeted/Untargeted
FGSM	0.03	1 (single-step)	—	L∞	Untargeted
PGD	0.03	40	0.01	L∞	Untargeted
BIM	0.03	20	0.005	L∞	Untargeted
DeepFool	—	up to 50	0.02 (default)	L2	Untargeted

and finally purified images, i.e., perturbed chest X-rays that were reprocessed through a diffusion model to remove the perturbations. Diffusion models are able to reverse the mathematical noise to output original images while maintaining the anatomical details [24]. For diffusion-based image purification, the Stable Diffusion v2.1 checkpoint was employed. The sampling method utilized was DDIM with 20 inference steps. The classifier-free guidance scale was set to 5.0, and the image-to-image noise strength was fixed at 0.5.
The three types of images are used for feature extraction, classification, and interpretability processes. Fig. 1 compiles several sample comparisons between these three groups, demonstrating the distortive power of attacks and the restoring processing of purified models.

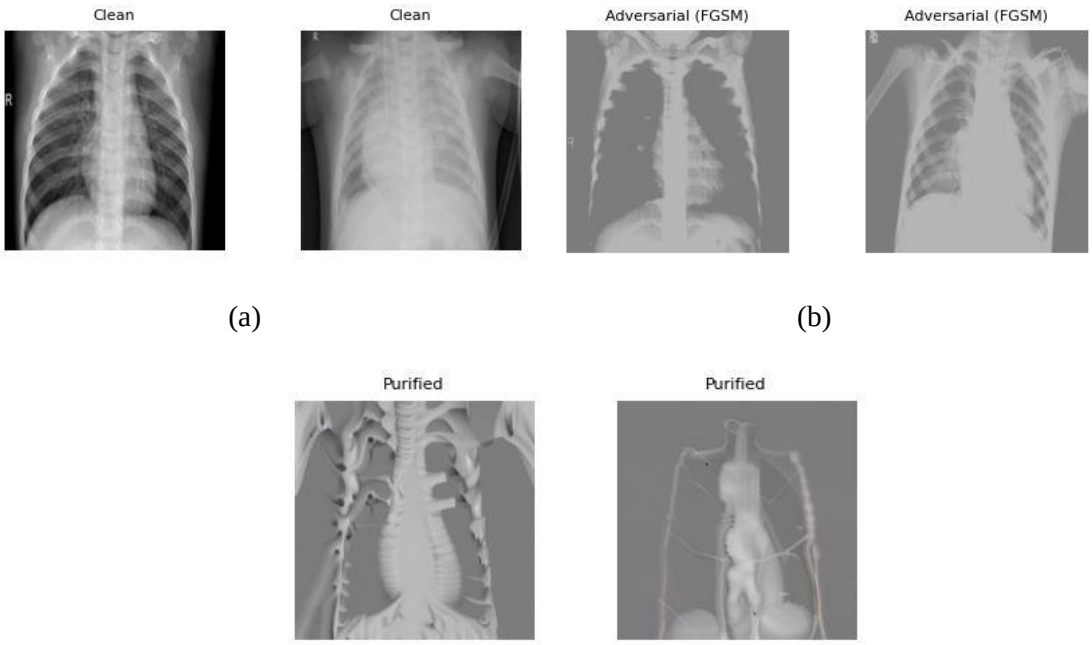

Fig. 1. A visual comparison of sample images from (a) clean, (b) adversarial and (c) purified images

3.2. Deep Feature Extraction Using ResNet18

We employed a deep convolutional neural network named ResNet18 to extract meaningful representations from chest X-ray images. The concept behind ResNet, short for Residual Network, is specifically designed to address the degradation problem in deep networks. While the network becomes deeper, the accuracy of deep networks deteriorates quickly, finally reaching zero accuracy as a resultant of vanishing gradients. By incorporating residual connections, ResNet enables the network to learn identity mappings and effectively train very deep architectures.

ResNet18 consists of 18 layers with batch normalization, convolutional blocks, ReLU activations, and shortcut connections, which enable it to learn high-level abstractions while preserving low-level properties [26].

This is particularly of interest in medical imaging as it permits to maintain subtle anatomical features at different scales. In this research, ResNet18 has been used for feature extraction instead of decision marking.

In particular, the first convolutional layer was adapted to accept inputs of 1-channel grayscale, making it suitable for use with radiological images. The last fully connected layer was replaced by an identity mapping so that the model produced 512D feature vectors.

The model was pre-trained and fine-tuned on the chest X-ray data to increase sensitivity to disease specific patterns. And finally, clean/adversarial/purified images were fed as inputs to generate deep feature embeddings and the with no-gradient inference mode.

Due to the residual structure of the ResNet, it can resist small changes and preserve fundamental characteristics of clinical data, which seems to have met the requirements for adversarial detection. Fig. 2 demonstrates the two-step use of ResNet18. During training, this model differentially classifies clean chest X-rays as Normal or Pneumonia. And then at the feature extraction stage, these 512 dimensional image features are extracted using the trained ResNet-18 on different imaging modalities without its final classification layer.

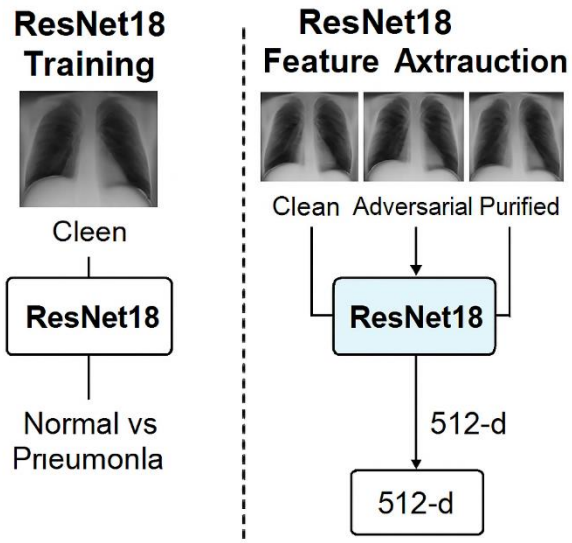

Fig. 2. illustrates the two-stage utilization of the ResNet18 model.

3.3. Feature Aggregation and Classification Using Random Forest

This step is to convert the 512-dimensional features extracted from chest X-ray images with ResNet18 into one single dataset for classification.

Every image, whether clean, corrupted or restored is represented by a single feature vector using its structural and textural information. These vectors are stored in the NumPy array X and the corresponding binary labels (0/1) in a vector y. All clean and released images are labeled as 0 while

the distorted FGSM generated, BIM, PGD, or DeepFool are labeled as 1. Such binary label type of partition makes the classification task much simpler, and allows the model to concentrate on learning boundary between distorted and undistorted samples.
Pre-processing such as resizing scale, grayscale format conversion and normalization were conducted on the data prior to feeding into the ResNet18 for feature extraction. Finally, a Random Forest classifier was used to classify the feature vectors for final predictions.
Random Forest is a type of ensemble learning algorithm which can train multiple decision trees and combine their outputs. In order to avoid overfitting and improve model diversity, each tree is constructed based on a random sample of the data points and features. The last prediction is decided by the majority among all trees [26].
For all analyses, the full feature set was employed to train a Random Forest model with 200 decision trees (n_estimators = 200). For model training we used the implementation in scikit-learn and serialized trained model for future inference with Joblib. After aggregating the feature maps and classifying them, we used a multi-tier evaluation process to test how robust they are by adversarial perturbation. The main parameters applied in all the proposed pipeline steps are provided Table 2. This pipeline can be broken down into four key stages: supervised training, adversarial attacks generation, purification through diffusion models and final classification.

Table 2. The most critical parameters used throughout the pipeline

Component	Parameter	Value / Description
Model Architecture	ResNet18 (pretrained)	Modified for 1-channel input, 512-d output
Training Setup	Optimizer	Adam
	Learning rate	0.001
	Epochs	10
	Batch size	32
	Scheduler	ReduceLROnPlateau (factor=0.5, patience=2)
Adversarial Attacks	FGSM / PGD ε values	[0.01, 0.05, 0.1, 0.2]
Purification	Model	Stable Diffusion v2.1
	Prompt	“a realistic chest X-ray image”
Classifier	Type	Random Forest (200 trees)

3.4. Evaluation and Performance

The Random Forest classifier was evaluated on the full dataset using standard performance metrics:
•Accuracy: Measures the proportion of correctly classified instances over the total number of samples. It provides a general sense of model correctness but may be misleading in imbalanced datasets.
•Confusion Matrix: A tabular representation of true versus predicted labels, showing the counts of true positives, true negatives, false positives, and false negatives.
•Classification Report: Includes precision, recall, and F1-score for each class:
oPrecision: The ratio of true positives to all predicted positives. It reflects the model’s ability to avoid false alarms.
oRecall: The ratio of true positives to all actual positives. It indicates the model’s ability to detect relevant instances.
o F1-score: The harmonic mean of precision and recall, providing a balanced measure of both.
These definitions were used to interpret the classifier’s behavior across different image types and to assess the impact of purification on detection reliability.

3.5. Concept-Based Interpretability Using TCAV

To illustrate the impact of inconsistent concepts on model’s internal representations, we test via Testing with Concept Activation Vectors (TCAV). TCAV measures the extent to which a model’s outputs are sensitive in specific directions human-comprehensible concepts by computing gradients on learned feature representations [27].

We considered every attack type (FGSM, PGD, BIM and DeepFool) as a binary concept, clean and purified samples represented the reference. We computed the activations from Layer 4 of the ResNet18 network and pooled them to be 512-dimensional feature vectors. The logistic were computed for both target set, safe and attack. For each concept, a logistic regression classifier that identifies which features of the image is most influenced by that concept (versus from base) was obtained and it achieves unifies for all concepts (CAV).

TCAV scores were then calculated as the ratio of angles in which gradients align in Positive direction to each CAV. Statistical significance was evaluated using a permutation-based switching test of 500 iterations.

The scores were as follows, high TCAV score for attack targets (≈1), which indicates greater influence of adversarial concepts. And moderate-to-low TCAV scores for safe targets (≈0) indicate that a partial semantic recovery occurred after purification. This work shows that not only are adversarial patterns identifiable, but consistently and directionally encoded in the latent space of the model.

Figure 3 presents an overview of the proposed methodology for detecting and interpreting adversarial bias in medical image classification. The flowchart illustrates the sequential stages of the framework, beginning with dataset preparation and classification into clean and adversarial categories using the ResNet18 classifier, followed by feature extraction. It also includes the simulation of adversarial attacks FGSM, PGD, BIM, and DeepFool to distort clean images, and subsequent refinement using a diffusion model.

The extracted features from all images are then fed into a Random Forest classifier to distinguish between safe (clean and purified) and adversarial samples. Finally, verification is performed using Testing with Concept Activation Vectors (TCAV), which quantifies the influence of adversarial bias on the model's internal representations.

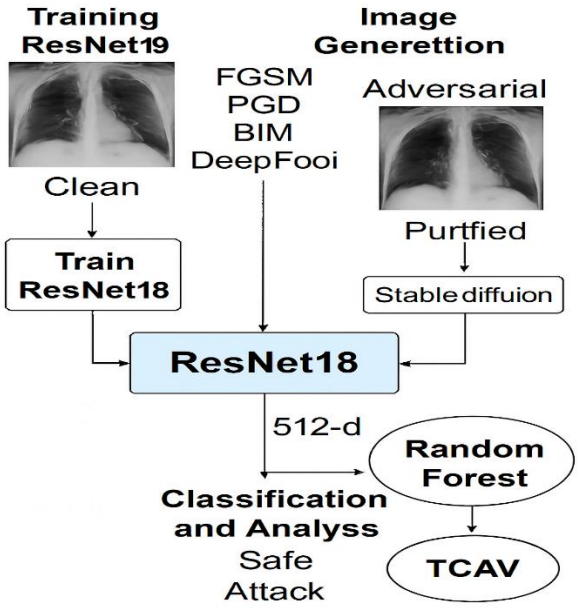

Fig. 3. The Proposed Methodology for Adversarial Detection.

4. Experimental results

4.1. Baseline Classification Performance

The training process of the fine-tuned model of ResNet18 on the clean chest X-ray data tested an accuracy rate of 91.19%, which further demonstrated that the modified ResNet18 has a good

performance in distinguishing normal and pneumonia cases in a clean environment. This performance serves as evidence that the proposed model can learn important clinical features including lung opacity, texture variation, and anatomical boundaries.
Nevertheless, the performance of the model decreased severely against adversarial perturbations. Two attack methods FGSM and PGD were performed on the test set under different ε. As in Table 3, small perturbations led to drastic drop of accuracy especially for iterative attacks as PGD.

Table 3. ResNet18 Accuracy Under Adversarial Attacks.

Attack Type	ε = 0.01	ε = 0.05	ε = 0.1	ε = 0.2
FGSM	~64.2%	~44.2%	~28.0%	~23.7%
PGD	~64.1%	~7.05%	~1.12%	~0.64%

These results highlight the vulnerability of deep models to imperceptible adversarial noise. PGD, in particular, reduced accuracy to nearly zero at ε = 0.2, despite the perturbations being visually imperceptible. This underscores the need for robust detection mechanisms beyond direct classification.

4.2. Robustness Recovery via Feature-Based Classification

To increase the performance of the Random Forest classifier, it was trained on deep feature representations learned from ResNet18 instead of classifying clinical signs directly. The model was trained to differentiate benign samples (i.e., clean and purified images) from adversarial samples produced by FGSM, PGD, BIM, and DeepFool attacks.
The feature vectors were crops of 512 dimensional representations pooled from the last convolutional block of ResNet. In total, 5616 samples were used for training and evaluation. We trained the classifier with 200 decision trees, in scikit-learn, and the performance of random forest classifier is shown in Table 4 and confusion matrix is given as Table 5 and Fig. 4 shows minimal misclassification.

Table 4. Performance Metrics of Random Forest Classifier.

Class	Precision	Recall	F1-Score	Support
Safe	0.993	0.996	0.994	3120
Attack	0.995	0.992	0.993	2496
Overall Accuracy	—	—	**99.02%**	5616

Table 5. Confusion Matrix of Random Forest Classifier.

	Predicted Safe	Predicted Attack
Actual Safe	3095	25
Actual Attack	30	2466

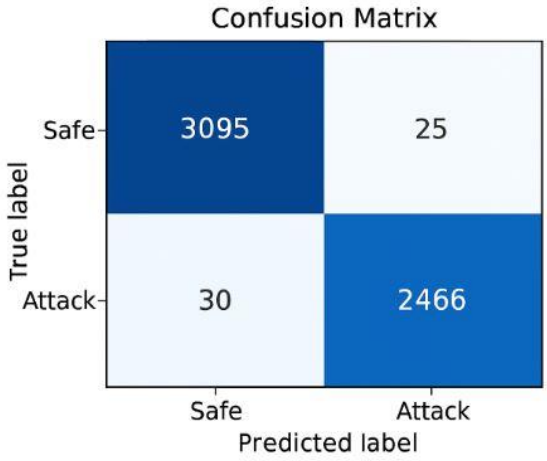

Fig. 4. Confusion Matrix.

These results confirm that purified images were successfully grouped with clean samples, indicating that the diffusion-based purification restored semantic consistency. The feature-based classification approach proved highly effective in detecting adversarial perturbations, even when direct classification failed.

4.3. Concept Sensitivity via TCAV Analysis

To interpret how adversarial concepts influence the model's internal representations, Testing with Concept Activation Vectors (TCAV) was applied. Each attack type FGSM, PGD, BIM, and DeepFool was treated as a distinct concept. TCAV scores were computed for both Safe and Attack targets across two logits, 0 for safe and 1 toward Attack.

The analysis was performed using pooled feature vectors from ResNet18's layer4, with 200 samples per concept and baseline. Gradients of the logits were computed with respect to these features, and logistic regression was used to generate Concept Activation Vectors (CAVs). Permutation testing with 500 iterations assessed statistical significance. Table 6 and Fig. 5 show the TCAV scores for Adversarial Concepts.

Table 6. TCAV Scores for Adversarial Concepts.

Concept	Safe→logit 0	Safe→logit 1	Attack→logit 0	Attack→logit 1
FGSM	0.633 (p=1.000)	0.700 (p=1.000)	0.667 (p=1.000)	0.833 (p=1.000)
PGD	0.100 (p=0.002)	0.000 (p=1.000)	0.033 (p=0.002)	0.000 (p=1.000)
BIM	0.733 (p=1.000)	0.500 (p=0.002)	0.767 (p=1.000)	0.300 (p=0.002)
DeepFool	1.000 (p=1.000)	0.767 (p=1.000)	1.000 (p=1.000)	0.800 (p=1.000)

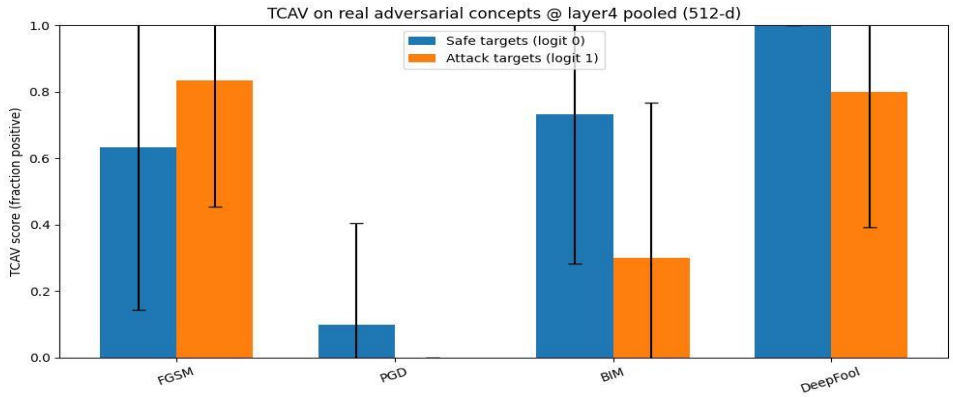

Fig. 5. TCAV Scores for Adversarial Concepts.

These results reveal that FGSM and DeepFool concepts are strongly encoded in the model's latent space, especially for Attack targets. PGD, despite its effectiveness in degrading accuracy, shows weak semantic alignment, suggesting it disrupts rather than reinforces concept direction, and BIM shows moderate influence and partial recovery after purification.

The presence of moderate TCAV scores for Safe targets especially in FGSM, BIM, and DeepFool suggests that purified images retain some semantic traces of the original attack, but are sufficiently restored to align with the Safe class. This validates the interpretability of the model and the effectiveness of the purification strategy. Importantly, these semantic patterns were consistent with TCAV scores computed across internal layers, where PGD showed low alignment ($p \approx 0.002$) and BIM exhibited stronger directional influence further reinforcing the conceptual interpretation and highlighting the layer-dependent nature of TCAV analysis.

To contextualize the methodology proposed in this study, Table 7 presents a comparative analysis of recent adversarial defense approaches in medical imaging. The comparison focuses on core aspects such as the defense strategy employed, interpretability support, dataset type, and reported performance. By contrasting our framework with prior works, we highlight key distinctions in terms of semantic recovery, clinical relevance, and conceptual transparency. This structured comparison underscores the novelty of integrating diffusion purification with feature-level classification and directional interpretability, setting our approach apart from existing methods.

Table 7. Comparative Analysis of Adversarial Defense Methods in Medical Imaging.

Study / Method	Defense Strategy	Interpretability	Medical Dataset Used	Reported Accuracy
[28]	Adversarial Autoencoder with Conditional Normalizing Flows	✗	Unlabeled medical images (unspecified)	~94.6%
[29]	Zero-shot Image Purification (ZIP) using Diffusion Models	✗	Chest CT scans, retinal images	~96.2%
[30]	Comparative evaluation of multiple adversarial defenses on medical DL systems	✗	CT, MRI, X-ray (varied modalities)	~95.1%
This Study	Diffusion Purification + ResNet Feature Extraction + Random Forest	✓ (via TCAV)	Chest X-ray (Normal vs Pneumonia)	99.02%

In this table, we provide a side-by-side comparison of our proposed method with the most recent adversarial defense approaches in medical imaging. Ji et al. (2023) used an autoencoder with conditional flows that did not relate to interpretability or clinical background. Nguyen & Luong (2024) presented feature-level analysis free diffusion-based purification. Puttagunta et al. (2023) compared a variety of adversarial defenses on both CT, MRI and X-ray data. The sense in their work covers wide range of medical modalities; however, feature-level categorizations and concept-based interpretability are also somewhat reduced, which makes it semantical. In contrast, we unify purification with feature extraction and concept-based interpretability (through TCAV), resulting in a state-of-the-art performance on a real chest X-ray dataset.

5. Discussion and future directions

This experiment with the modular pipeline of supervised network adversarial simulation, purification and feature based classification can be achieved very robust performance on chest X-rays. The high first-order vulnerability to adversarial attacks (especially PGD) of ResNet18 underlines a potential challenge in practical clinical settings with the requirement for high reliability. By using ResNet18 as a feature extractor and training a Random Forest classifier on learned representations, the system was able to recover error at 99.02% (minimal misclassification). This suggests that deep features may remain discriminative even when the original classifier is broken and ensemble methods

can effectively exploit these features to make decisions more robust. Furthermore, the accommodation of cleaned images into the Safe class was successful. The adoption of Stable Diffusion not only recovered the anatomy but also maintained semantic alignment, as shown by the ability of classifier to cluster purified samples with clean ones. This indicates that generative purification is an effective strategy to defend adversarial attacks for medical imaging. TCAV study confirmed system interpretability. Adversarial concepts held constant sway over the model's internal representations, since FGSM and DeepFool had very solid TCAV scores. Although effective, PGD led to only mediocre semantic alignment between the languages, indicating that certain attacks may damage feature space coherence rather than encode specific concepts. These observations highlight concept-level interpretability as an important part of evaluating model robustness.

5.1. Future Work

Based on these results, some lines of future research are suggested:

1- Multiclass Extension: Moving from a binary classification (Normal vs. Pneumonia) to cover more entities, including COVID-19, TB and pleural effusion in order to test the scalability and generalization of our model, we aim for it to be able to perform classification tasks where the valid medical entities extend beyond two classes.

2- Cross-Dataset Validation: Investigating domain transferability and robustness across imaging sources by comparing results on the CheXpert versus MIMIC-CXR datasets for testing.

3- Purification Benchmarking: To assess the trade-offs between anatomy fidelity and semantic recovery, compare Stable Diffusion with denoising autoencoders or GAN-based restoration.

4- Multilingual Interpretability: Extending TCAV analysis to Arabic medical text corpora and its application to characterizing concept sensitivity in multilingual NLP systems trained on radiology reports.

References

[1] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *arXiv*, arXiv:1804.05296, 2018, doi: 10.48550/arXiv.1804.05296.

[2] K. D. Apostolidis and G. A. Papakostas, "A survey on adversarial deep learning robustness in medical image analysis," *Electronics*, vol. 10, no. 17, Art. no. 2132, 2021, doi: 10.3390/electronics10172132.

[3] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Survey on adversarial attack and defense for medical image analysis: Methods and challenges," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–38, 2024, doi: 10.1145/3702638.

[4] M. Surekha, A. K. Sagar, and V. Khemchandani, "Adversarial attack and defense mechanisms in medical imaging: A comprehensive review," in *Proc. IEEE Int. Conf. Computing, Power and Communication Technologies (IC2PCT)*, 2024, pp. 1657–1661, doi: 10.1109/IC2PCT60090.2024.10486235.

[5] G. Bortsova *et al*., "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, Art. no. 102141, 2021, doi: 10.1016/j.media.2021.102141.

[6] D. Rodriguez, T. Nayak, Y. Chen, R. Krishnan, and Y. Huang, "On the role of deep learning model complexity in adversarial robustness for medical images," *BMC Medical Informatics and Decision Making*, vol. 22, no. Suppl. 2, Art. no. 160, 2022, doi: 10.1186/s12911-022-01891-w.

[7] A. S. Musthafa, K. Sankar, T. Benil, and Y. N. Rao, "A hybrid machine learning technique for early prediction of lung nodules from medical images using a learning-based neural network classifier," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 3, Art. no. e7488, 2023, doi: 10.1002/cpe.7488.
[8] G. Webber and A. J. Reader, "Diffusion models for medical image reconstruction," *BJR | Artificial Intelligence*, vol. 1, no. 1, Art. no. ubae013, 2024, doi: 10.1093/bjrai/ubae013.
[9] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–44, 2025, doi: 10.1145/3721479.
[10] D. Qiu, L. Zheng, J. Zhu, and D. Huang, "Multiple improved residual networks for medical image super-resolution," *Future Generation Computer Systems*, vol. 116, pp. 200–208, 2021, doi: 10.1016/j.future.2020.11.001.
[11] B. P. Reddy, K. Rangaswamy, D. Bharadwaja, M. M. Dupaty, P. Sarkar, and M. S. Al Ansari, "Using generative adversarial networks and ensemble learning for multimodal medical image fusion to improve the diagnosis of rare neurological disorders," *Int. J. Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, doi: 10.14569/IJACSA.2023.01411108.
[12] G. W. Muoka *et al*., "A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense," *Mathematics*, vol. 11, no. 20, Art. no. 4272, 2023, doi: 10.3390/math11204272.
[13] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," *Expert Systems with Applications*, vol. 198, Art. no. 116815, 2022, doi: 10.1016/j.eswa.2022.116815.
[14] Z. Teng *et al*., "A literature review of artificial intelligence for medical image segmentation: From explainable AI to trustworthy AI," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 12, Art. no. 9620, 2024, doi: 10.21037/qims-24-723.
[15] A. Abomakhelb, K. A. Jalil, A. G. Buja, A. Alhammadi, and A. M. Alenezi, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," *Technologies*, vol. 13, no. 5, Art. no. 202, 2025, doi: 10.3390/technologies13050202.
[16] J. Liu, Y. Li, Y. Guo, Y. Liu, J. Tang, and Y. Nie, "Generation and countermeasures of adversarial examples on vision: A survey," *Artificial Intelligence Review*, vol. 57, no. 8, Art. no. 199, 2024, doi: 10.1007/s10462-024-10841-z.
[17] D. A. M. Akhtom, M. M. Singh, and C. Xinying, "Enhancing trustworthy deep learning for image classification against evasion attacks: A systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 7, Art. no. 174, 2024, doi: 10.1007/s10462-024-10777-4.
[18] J. Li, Y.-A. Tan, X. Liu, W. Meng, and Y. Li, "Interpretable adversarial example detection via high-level concept activation vector," *Computers & Security*, vol. 150, Art. no. 104218, 2025, doi: 10.1016/j.cose.2024.104218.
[19] Y. Wang, T. Li, S. Li, X. Yuan, and W. Ni, "New adversarial image detection based on sentiment analysis," *IEEE Trans. Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 14060–14074, 2023, doi: 10.1109/TNNLS.2023.3274538.
[20] S. Freitas, S.-T. Chen, Z. J. Wang, and D. H. Chau, "Unmask: Adversarial detection and defense through robust feature alignment," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 1081–1088, doi: 10.1109/BigData50022.2020.9378303.
[21] F. Mumcu and Y. Yilmaz, "Detecting adversarial examples," *arXiv*, arXiv:2410.17442, 2024, doi: 10.48550/arXiv.2410.17442.

[22] H. Mu, C. Li, A. Peng, Y. Wang, and Z. Liang, "Robust adversarial example detection algorithm based on high-level feature differences," *Sensors*, vol. 25, no. 6, Art. no. 1770, 2025, doi: 10.3390/s25061770.
[23] X. Ye, Q. Zhang, S. Cui, Z. Ying, J. Sun, and X. Du, "Mitigating adversarial attacks in object detection through conditional diffusion models," *Mathematics*, vol. 12, no. 19, Art. no. 3093, 2024, doi: 10.3390/math12193093.
[24] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, 2023, doi: 10.1109/TPAMI.2023.3261988.
[25] P. Nasra and S. Gupta, "ResNet18-based deep learning approach for efficient and accurate nail disease detection," in *Proc. 3rd Int. Conf. Advancement in Computation & Computer Technologies (InCACCT)*, 2025, pp. 145–150, doi: 10.1109/InCACCT65424.2025.11011397.
[26] A. Yaqoob *et al*., "SGA-driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study," *Scientific Reports*, vol. 15, no. 1, Art. no. 10944, 2025, doi: 10.1038/s41598-025-95786-1.
[27] L. Schmalwasser, N. Penzel, J. Denzler, and J. Niebling, "FastCAV: Efficient computation of concept activation vectors for explaining deep neural networks," *arXiv*, arXiv:2505.17883, 2025, doi: 10.48550/arXiv.2505.17883.
[28] Z. Rguibi, A. Hajami, D. Zitouni, Y. Maleh, and A. Elqaraoui, "Medical variational autoencoder and generative adversarial network for medical imaging," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, pp. 494–505, 2023, doi: 10.11591/ijeecs.v32.i1.
[29] M. C. Nguyen and N. H. Luong, "Diffusion-based purification for adversarial defense in medical image classification," in *Proc. Int. Symp. Information and Communication Technology*, 2024, pp. 80–91, doi: 10.1007/978-981-96-4288-5_7.
[30] M. K. Puttagunta, S. Ravi, and C. N. K. Babu, "Adversarial examples: Attacks and defenses on medical deep learning systems," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 33773–33809, 2023, doi: 10.1007/s11042-023-14702-9.

Journal of Basrah Researches (Sciences) **51**(2), 270 (2025)
DOI: https://doi.org/10.56714/bjrs.51.2.19

تصنيف متين وقابل للتفسير لصور الأشعة السينية للصدر باستخدام التنقية بخوارزمية الانتشار والكشف العدائي المعتمد على المفاهيم

امنة كاظم علي

كلية الطب البيطري, جامعة البصرة, البصرة , العراق

معلومات البحث

الاستلام 09 تشرين ثاني 2025
المراجعة 21 كانون أول 2025
القبول 24 كانون أول 2025
النشر 31 كانون أول 2025

الكلمات المفتاحية

. الصور الطبية
Concept Activation Vectors (TCAV),
Random Forest,
. ResNet18

الملخص

الهجوم العدائي هو اسلوب يُعرض سلامة التعليم العميق المستخدم في التصوير الطبي للخطر عن طريق اضافة تغييرات طفيفة الى مدخلات النموذج لا يلاحظها البشر, مما يؤدي الى اتخاذ النموذج قرارات غير صحيحة, وبالتالي يمكن للمهاجم اتلاف كلاً من بيانات الادخال ونظام التنبؤ. في هذه الورقة تم اقتراح طريقة كشف هجينة تعتمد على تنقية الصور واستخراج الميزات العميقة والتعلم التجميعي لمعالجة هذه المشكلة باستخدام صور الاشعة السينية للصدر.

يتكون النهج من مرحلتين اولا, تنقية العينات المعادية من خلال نموذج الانتشار لازاله الاضطرات الصغيرة تماماً. ثم تصنيف الصور الي مجموعتين . يتم جمع الصور النظيفة والمنقاة في فئة امنة, بينما تجمع الصور الضارة في فئة غير امنة , لاحقاً يتم استخدام Resnet18 مدربة مسبقاً لاستخراج الميزات العميقة لصور الاشعة السينية ويصنف نموذج Random forest الميزات الى صور ضارة وغير ضارة. تظهر النتائج التجريبية ان الطريقة المقترحة يمكنها اكتشاف 99% من العينات المعادية وبالتالي فأن معدل الكشف المرضي بالاعتماد على تنقية الصور واستخراج الميزات القائم على المجموعة يجعل قرارات النموذج اكثر موثوقية وامان.

Citation: A. K. Ali., J. Basrah Res. (Sci.) **50**(2), 270 (2025). DOI:https://doi.org/10.56714/bjrs.51.2.19

***Corresponding author email:** amna.kadhim@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq