Features Normalisation and Standardisation (FNS): An Unsupervised Approach for Detecting Adversarial Attacks for Medical Images

Sreenivasan Mohandas and Naresh Manwani
Machine Learning Lab, International Institute of Information Technology, Hyderabad, India

Keywords: Adversarial Attacks, Adversarial Defenses, Multivariate Gaussian Models, Medical Applications, Features Normalization, and Standardization.

Abstract: Deep learning systems have shown state-of-the-art performance in clinical prediction tasks. However, current research suggests that cleverly produced hostile images can trick these systems. Deep learning-based medical image classification algorithms have been questioned regarding their practical deployment. To address this problem, we provide an unsupervised learning technique for detecting adversarial attacks on medical images. Without identifying the attackers or reducing classification performance, our suggested strategy FNS (Features Normalization and Standardization), can detect adversarial attacks more effectively than earlier methods.

1 INTRODUCTION

Deep learning-based medical imaging systems have significantly improved the accuracy and efficiency of clinical prediction tasks, thanks to the development of deep learning algorithms and the availability of high-quality labeled medical imaging datasets. For example, (Daniels and Metaxas, 2019) extracts features from X-rays for lung disease categorization, (Shaffie et al., 2019) uses computed tomography (CT) scans to detect lung cancer, and (Reda et al., 2018) uses magnetic resonance imaging (MRI) scans to establish an early diagnosis of prostate cancer. Several healthcare start-ups, including Zebra Medical Vision and Aidoc have recently secured FDA certifications for their AI medical imaging systems. According to these FDA clearances, deep learning-based medical imaging systems could shortly be used for clinical diagnosis.

Parallel to advancements in deep learning-based medical imaging systems, so-called adversarial images have shown flaws in these systems in various clinical areas (Finlayson et al., 2019). Adversarial images are purposely generated inputs to deep learning models to deceive image categorization. The method falsely labels "Pleural Thickening" as "Pneumothorax" when only minor perturbations are added to a clean X-ray image. As a result, users of such systems may be exposed to unforeseen harmful scenarios, such as diagnostic errors, medical reimbursement fraud, and so on, if sufficient safeguards are not in place. As a result, an adequate defense strategy must be devised to deploy these devices securely.

Several defensive strategies have been offered in response to the threat. Adversarial training, which enlarges the training dataset with adversarial images to improve the resilience of the trained Convolutional Neural Network (CNN) model, is a standard method in the natural imaging domain. Many diverse adversarial images are included in the training dataset, which can dramatically reduce classification accuracy. (Ma et al., 2021) develops a logistic regression classifier based on characteristics extracted from a trained CNN model to distinguish adversarial images from clean images. However, the usefulness of this approach is limited to a set of predefined attack methods. These issues are addressed in the following way.

(Taghanaki et al., 2018) adds a radial basis mapping kernel to CNN models, which translates data onto a linearly well-separated manifold to improve class separation and lessen the impact of perturbations. Global dependencies and contextual information can be leveraged to strengthen resilience, according to (He et al., 2019). To guard against adversarial attacks, they suggest a non-local context encoder in medical picture segmentation systems. Although both strategies improve robustness by changing the network design, the trade-off between accuracy and robustness (Zhang et al., 2019) may degrade system performance in practice.

This paper proposes a robust adversarial image detection technique that effectively counters adver-

Mohandas, S. and Manwani, N.
Features Normalisation and Standardisation (FNS): An Unsupervised Approach for Detecting Adversarial Attacks for Medical Images.
DOI: 10.5220/0011624000003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 140-145
ISBN: 978-989-758-623-1; ISSN: 2184-433X

sarial attacks on deep learning-based medical image classification systems. We focus on unsupervised anomalous detection utilizing features retrieved from a trained CNN classifier, as inspired by (Zheng and Hong, 2018) and (Li and Zhu, 2020). Our method successfully detects adversarial images and can effectively defend against unseen attacks, whether white-box or black-box, because it makes no assumptions about prior attack method knowledge. To demonstrate the success of our suggested defense strategy, we conduct extensive experiments using a publicly available X-ray dataset. We have considered the X-ray dataset to validate our algorithm. As per (Shi et al., 2022), X-ray and color fundus photographs are common diagnostic and prognostic imaging modalities in patient care.

2 BACKGROUND

In this section, we provide background information and a review of related studies on adversarial attack and defense mechanisms. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the instance space and $\mathcal{Y} = \{+1, -1\}$ be the label space. Let $\mathbf{x} \in \mathcal{X}$ be an example and y be its actual label. Let classifier $f : \mathcal{X} \to \mathcal{Y}$ is such that $f(\mathbf{x}) = y$. The purpose of adversarial attacks is to find an example $\mathbf{x}^* \in \mathcal{X}$ in the neighborhood of $\mathbf{x}$ that is misclassified by the classifier f (i.e., $f(\mathbf{x}^*) \neq y$). There are two types of hostile examples: non-targeted and targeted examples. A non-targeted adversarial example $\mathbf{x}^*$ is produced by adding minor noise to $\mathbf{x}$ without changing the label, yet misleads the classifier as $f(\mathbf{x}^*) \neq y$. A targeted adversarial example tries to trick the classifier by producing a specific label as $f(\mathbf{x}^*) = y^*$, where $y^* \neq y$ is the adversary's target label. In most circumstances, the adversarial noise's L_p norm must be less than a $\|\mathbf{x}^* - \mathbf{x}\|_p \leq \varepsilon$ for some $\varepsilon > 0$, and $p \in \{0, 1, 2, ...\}$. We will now discuss different adversarial attacks considered in our experimental analysis.

2.1 Fast Gradient Sign Method (FGSM)

To demonstrate that the high-dimensional linearity of deep neural networks causes adversarial cases to emerge, Goodfellow (Goodfellow et al., 2015) devised the Fast Gradient Sign Method (FGSM) technique. The fundamental idea behind the approach is to produce adversarial perturbations following the deep learning model's maximum gradient change direction, then add the perturbations to the image to create adversarial examples.

$$\mathbf{x}_{adv} = \mathbf{x} + \varepsilon * sign(\nabla_x L(\mathbf{x}, y))$$

This strategy can be seen as a straightforward one-step method for maximizing the inner portion of the saddle point formulation. The FGSM algorithm's benefit is that it is a single-step attack with a quick attack speed. Still, the attack success rate is lesser than iterative attack algorithms like PGD and BIM.

2.2 Project Gradient Descent (PGD)

A more strong adversary is the multi-step variation, which is projected gradient descent (PGD) on the negative loss function. PGD (Madry et al., 2018) perturbs a clean data $\mathbf{x}$ for T steps with smaller step sizes. After each step of perturbation, PGD projects the adversarial example back onto the ε-ball of $\mathbf{x}$ if it goes beyond:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha * sign(\nabla_x L(\mathbf{x}_{i-1}, y))$$

where α is the step size, and $\mathbf{x}_i$ is the adversarial example at the i-th step ($\mathbf{x}_0 = \mathbf{x}$). The step size is usually set to $\varepsilon/T \leq \alpha < \varepsilon$ for overall T steps of perturbations.

2.3 Basic Iterative Method (BIM)

Basic iterative method (Kurakin et al., 2017) is an extension of FGSM. It applies FGSM multiple times with a small step size α while clipping it to keep in the constraint budget. It initializes adversarial example with $\mathbf{x}_0 = \mathbf{x}$

$$\mathbf{x}_i = Clip[\mathbf{x}_{i-1} + \alpha.sign(\nabla_x L(\mathbf{x}_{i-1}, y))]$$

where i denotes the iteration number for iterative attack and the $Clip$ function clips all the values between 0 and 1.

2.4 Momentum Iterative Fast Gradient Sign Method (MIM)

MIM (Akhtar and Mian, 2018) improves the convergence of the PGD algorithm by using momentum. MIM generates adversarial examples by using the momentum-based iterative algorithm. Applying momentum gradient and providing techniques to escape from the poor local maximum during iterations. The momentum gradient m can be calculated as:

$$\mathbf{m}_{i+1} = \mu * \mathbf{m}_i + \frac{(\nabla_\delta L(f_\theta(\mathbf{x}_i + \delta), y_i))}{\|(\nabla_\delta L(f_\theta(\mathbf{x}_i + \delta), y_i))\|_{L_1}}$$

where ∇_δ shows the gradient function and μ is the decay factor. Initially, $\mathbf{x}_{i-1}$ is the original input and m_0 is set to previous iteration value (Mohandas et al., 2022). In each iteration, $\mathbf{x}_i$ is updated as

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \varepsilon * sign(\mathbf{m}_{i+1})$$

The adversarial disturbance produced by the DeepFool assault is relatively minor when compared to FGSM, and other attacks (Liang et al., 2022). We restricted our scope to iterative methods as they have high success rates than others.

The adversarial image is created by subtly altering the original image; as a result, the perturbations appear as noise at the pixel level, obstructing human detection. On the other hand, such noise is visible at the feature level of CNN models. According to (Huang et al., 2017), adversarial perturbations are difficult to detect by human eyes, resulting in significant noise at the feature level. Furthermore, the convolution-pooling techniques performed in CNN models during forward propagation might increase this" noise," resulting in misclassification. On the other hand, because the size of perturbations rises layer by layer, high-level characteristics can easily distinguish between clean and adversarial images (Xie et al., 2019).

3 PROPOSED METHOD - FEATURES NORMALISATION AND STANDARDISATION (FNS)

We propose a new adversarial image detection module for the medical image classification system, independent of the attacker's method, and doesn't require model retraining. Figure 1 shows the proposed method where a Multivariate Gaussian Model (MGM) is created with the extracted features of clean (original) images just before the classification layer of DenseNet-121 (Huang et al., 2017). We have considered the output of the last Dense block for modeling MGM as the high-level characteristics can be distinguished easily between clean and adversarial images as per (Xie et al., 2019). As mentioned in Figure 1, these extracted features were normalized and standardized before the creation of MGM. Once the model is created, it is used to identify the clean and adversarial images during the testing phase, as shown in Figure 2. Only clean images are passed to the classification layer of DensetNet-121 for disease classification.

Before being modeled using MGM: $y \sim \mathcal{N}(\mu, \Sigma)$, where $y = H(\mathbf{x})$ represents the feature extracted using the final fully connected layer given a clean input image $\mathbf{x}$, the high-level feature distribution of clean images is normalized and standardized. The $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ are mean vector and covariance matrix, where d represents the dimension of MGM. Considering features extracted from clean training images $Y = \{y_1, ..., y_n\}$, estimate $\mu = \left(\frac{1}{n}\right) \sum_{i=1}^{n} y_i$ and

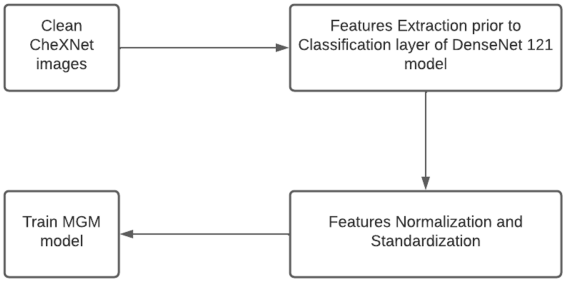

Figure 1: Proposed model for training MGM.

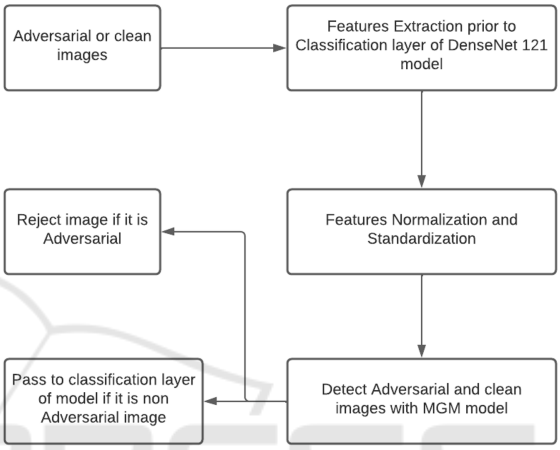

Figure 2: Testing phase.

$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^{\mathrm{T}} (y_i - \mu) + \lambda I$, where λI is the non-negative regularization added to the diagonal of the covariance matrix and n is the number of input samples. After modeling MGM, for a given image (can be clean or adversarial), compute the probability of $y^* = H(\mathbf{x}^*)$ belonging to the clean image distribution by

$$p(y^*) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{d}{2}}} \exp\left(\frac{-1}{2}(y_i - \mu)^{\mathrm{T}} \Sigma^{-1} (y_i - \mu)\right).$$

However, in reality, p(y^*) is computationally expensive due to the high dimension (d = 1024) and because its value is so near to zero, arithmetic underflow results. We re-parameterize the covariance matrix using Cholesky decomposition to get around these technical challenges, i.e., $\Sigma = \mathrm{RR}^T$ and rewrite the probability density function into log form:

$$\log p(y^*) = \frac{1}{2}[2 \times (\sum_{i=1}^{d} R_{ii}) + \|R^{-1}(y^* - \mu)\|^2 + d \log 2\pi].$$

Finally, as shown in Figure 2, $\mathbf{x}^*$ will be detected as an adversarial image and rejected if $\log p(y^*)$ is lower than a threshold. The threshold value is determined by considering 95% clean images during training.

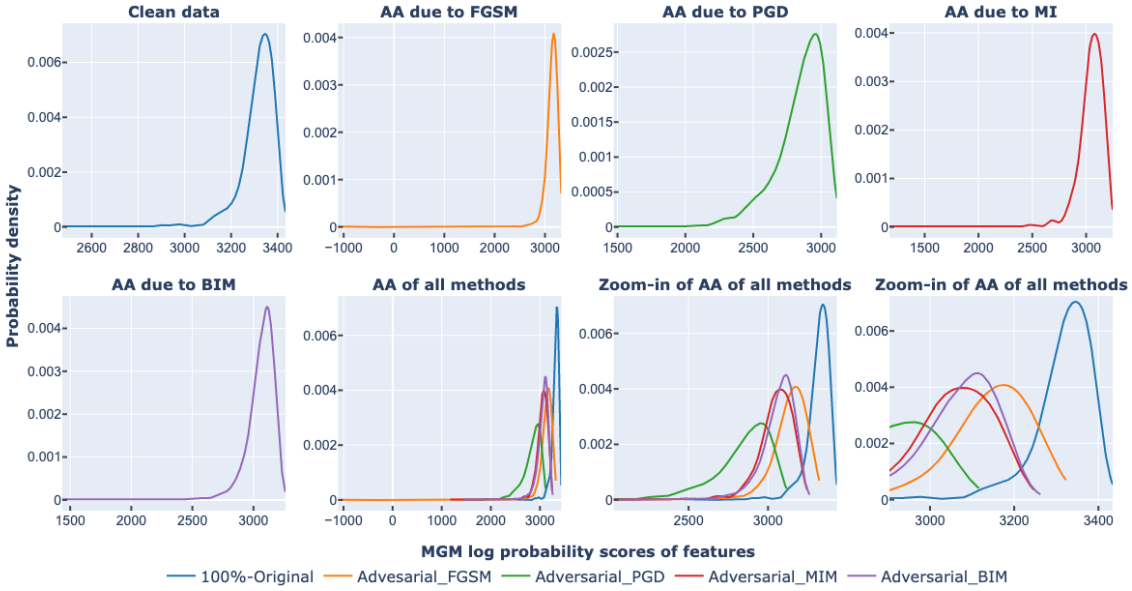

Figure 3: MGM model trained without FNS using 95% clean data (AA - Adversarial Attack) on X-ray dataset.

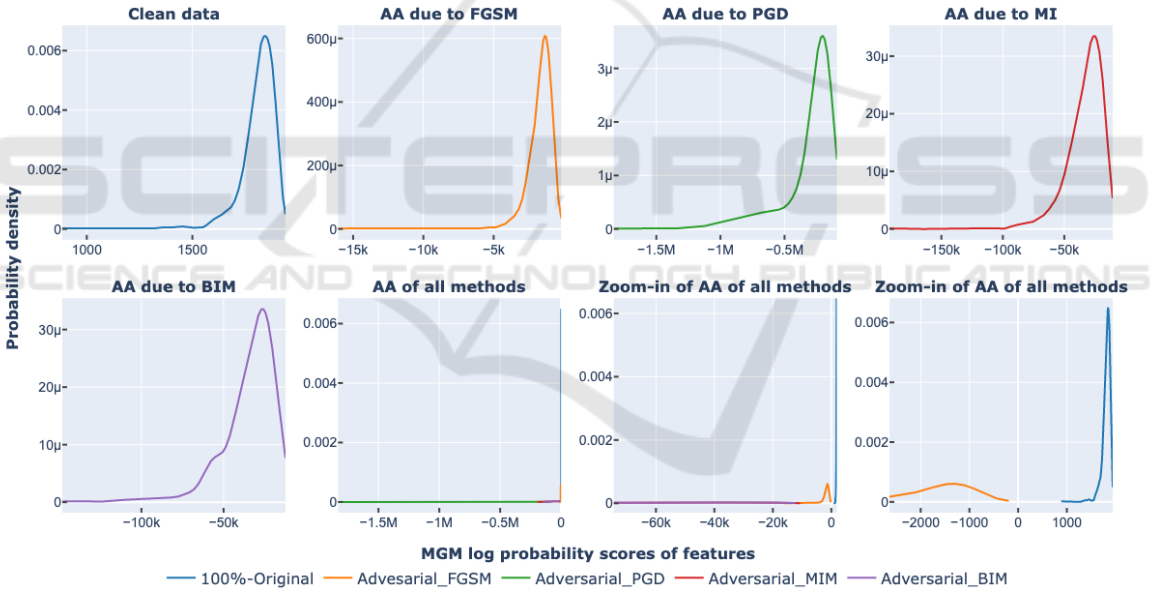

Figure 4: MGM model trained with FNS using 95% clean data (AA - Adversarial Attack) on X-ray dataset.

4 EXPERIMENTS

We verify the performance of our proposed defense approach by conducting experiments on a large public chest X-ray dataset. The NIH ChestX-ray14 (Goodfellow et al., 2015) contains 112,120 frontal-view chest X-rays taken from 30,805 patients, where around 46% images are labeled with at least one of 14 pathologies. The features extracted from the complete clean training and validation datasets are used for training and validating our proposed detection module. We created an MGM (detection model) with the features of 95% clean images and extracted the features of FGSM, BIM, PGD, and MIM-based adversarial attack images for the whole dataset. For generating PGD and BIM adversarial attack images, we ran the iteration count of 7 (7 PGD steps). For generation MIM adversarial attack images, we considered a decay factor of 0.1 and an iteration count of 3 (Mohandas et al., 2022).

Table 1: F1 score in detecting Adversarial image for X-ray dataset, * refers the results captured from (Li and Zhu, 2020).

Approach	Threshold value	FGSM	PGD	MIM	BIM
MGM without feature normalization	3260	88.5	92.18	92.13	92.18
MGM with feature normalization (our method)	882	100	100	100	100
Isolation Forest*	-	83.8	87.4	87.4	87.4
One class Support Vector Machine*	-	87	93.1	93.1	93.1

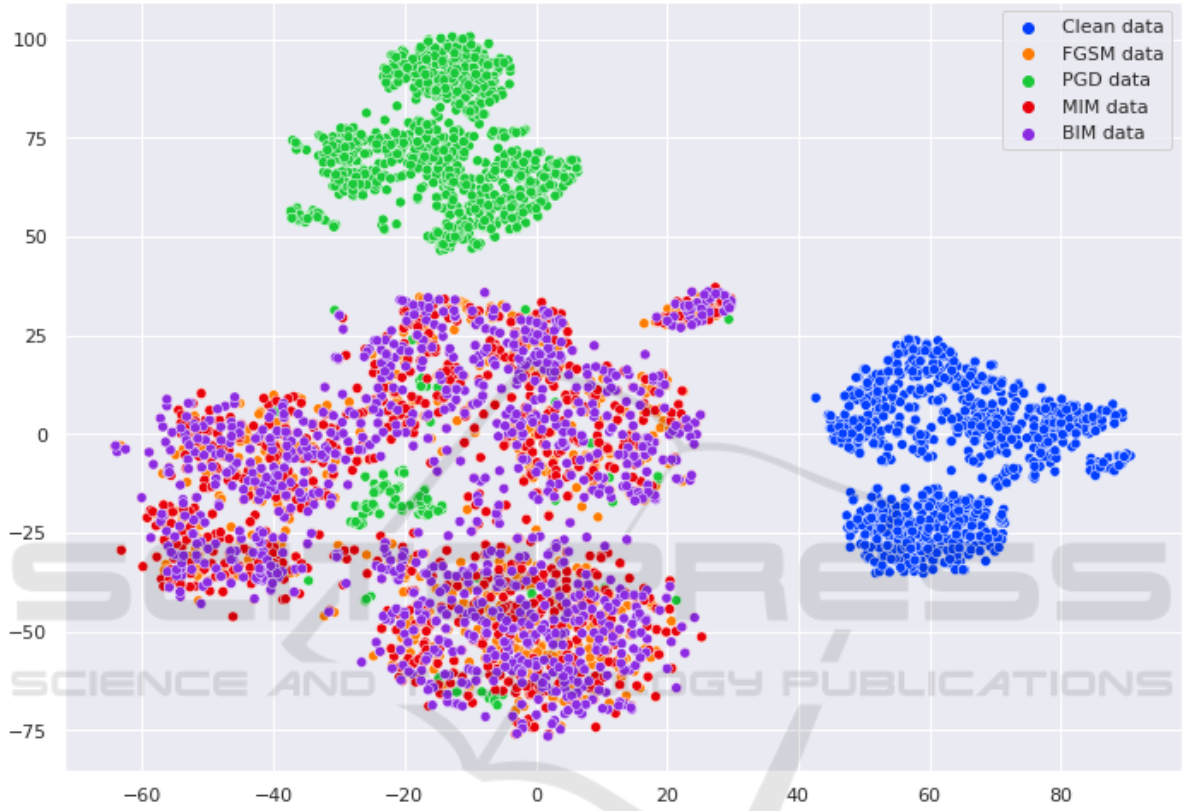

Figure 5: 2D t-SNE visualization of features X-ray images.

Figure 3 shows the distribution of log probability scores of MGM (trained on the features of 95% clean images) for the features from clean (100 % originals) images, FGSM, BIM, MIM, and PGD-based adversarial images. The zoomed-in version of images gives more clarity about the interference of adversarial attacks on original images. Figure 5 shows the visualization of features extracted before the classification layer of Densenet-121 for Original images, FGSM, PGD, BIM, and MIM adversarial attacks.

To the best of our knowledge, data normalization and standardization are applied at the input level (layer 0) or the intermediate layer but not on the features extracted before the classification layer of a neural network. We extracted the features (1000 features) just before the classification layer of the DenseNet-121 and performed normalization and standardization of the features before generating an MGM. In our proposed method, an MGM model is created with the above normalized and standardized features with 95% of clean images and randomly tested with 1000 clean images, FGSM, BIM, MIM, and PGD adversarial attack images. Figure 4 shows the distribution of log probability scores of above MGM on clean images, FGSM, BIM, PGD, and MIM attack images. From the zoomed-in version of the images, it can noticed that there is no interference of adversarial attack images on the original images. With a proper threshold value, we can identify the adversarial images 100%, as shown in Table 1. All the above experiments are performed on an Nvidia T4 machine with CUDA Version 11.2.

5 CONCLUSIONS

This paper proposes a Feature Normalisation and Standardisation unsupervised approach for detecting adversarial images. This is very useful in real-life scenarios where it doesn't require the attacker's method or retraining the model. We provide an experimental comparison of the iterative adversarial attack algorithms on the X-ray dataset. The results show that our proposed algorithm accurately determines adversarial images. This can be extended for other medical image datasets where one can use different models than GMM to model the extracted features.

REFERENCES

Akhtar, N. and Mian, A. S. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430.

Daniels, Z. A. and Metaxas, D. N. (2019). Exploiting visual and report-based information for chest x-ray analysis by jointly learning visual classifiers and topic models.

Finlayson, S. G., Bowers, J., Ito, J., Zittrain, J., Beam, A., and Kohane, I. S. (2019). Adversarial attacks on medical machine learning.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

He, X., Yang, S., Li, G., Li, H., Chang, H., and Yu, Y. (2019). Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *AAAI*.

Huang, G., Liu, Z., and Weinberger, K. Q. (2017). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2017). Adversarial machine learning at scale. *ArXiv*, abs/1611.01236.

Li, X. and Zhu, D. (2020). Robust detection of adversarial attacks on medical images. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1154–1158.

Liang, H., He, E., Zhao, Y., Jia, Z., and Li, H. (2022). Adversarial attack and defense: A survey. *Electronics*.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083.

Mohandas, S., Manwani, N., and Dhulipudi, D. P. (2022). Momentum iterative gradient sign method outperforms pgd attacks. In *ICAART*.

Reda, I., Ayinde, B. O., Elmogy, M. M., Shalaby, A. M., El-Melegy, M. T., El-Ghar, M. A., El-Fetouh, A. A., Ghazal, M., and El-Baz, A. S. (2018). A new cnn-based system for early diagnosis of prostate cancer.

Shaffie, A., Soliman, A., Khalifeh, H. A., Ghazal, M., Taher, F., Elmaghraby, A. S., Keynton, R. S., and El-Baz, A. S. (2019). Radiomic-based framework for early diagnosis of lung cancer.

Shi, X., Peng, Y., Chen, Q., Keenan, T. D. L., Thavikulwat, A. T., Lee, S., Tang, Y., Chew, E. Y., Summers, R. M., and Lu, Z. (2022). Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognit.*, 132:108923.

Taghanaki, S. A., Das, A., and Hamarneh, G. (2018). Vulnerability analysis of chest x-ray image classification against adversarial attacks. In *MLCN/DLF/iMIMIC@MICCAI*.

Xie, C., Wu, Y., van der Maaten, L., Yuille, A. L., and He, K. (2019). Feature denoising for improving adversarial robustness. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 501–509.

Zhang, H. R., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. *ArXiv*, abs/1901.08573.

Zheng, Z. and Hong, P. (2018). Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In *NeurIPS*.