On enhancement of SEVIT for Chest XRay Classification using Defensive Distillation and Adversarial Training

Raza Imam, Salma Alrashdi, and Baketah Alrashdi

Mohamed Bin Zayed University of Artificial Intelligence, AbuDhabi, UAE
{raza.imam, salma.alrashdi, baketah.alrashdi}@mbzuai.ac.ae

Abstract. With the recent adaptation of deep learning in medical imaging problems, elevated vulnerabilities have been explored in reent CNN and ViT based solutions. The vulnerability of ViTs to adversarial, privacy, and confidentiality attacks raise serious concerns about their reliability in medical settings. This work aims to enhance the robustness of existing benchmark solution based on self-ensembling ViTs in tuberculosis chest X-ray classification problem. In the proposed work, we have presented a novel SEVIT-CNN architecture built over SEVIT, that utilizes the CNN modules for improved computational efficiency and robustness utilizing the adversarial training and defensive distillation. The proposed approach leverages the fact that adversarial training when performed with the combination of defensive distillation, presents significantly higher robustness against adversaries. CNN's efficiency in learning spatial features through convolution operations at various levels of abstraction, along with training the model with adversarial examples improves its ability to handle perturbations and generalize better. By creating a distilled model with soft probabilities, uncertainty and variation are introduced into the output probabilities, making it more difficult for privacy attacks like model extraction. Extensive experiments performed with the proposed architecture on publicly available Tuberculosis X-ray dataset shows efficacy in terms of computational efficiency and enhanced robustness.

Keywords: Ensembling · Adversarial Attack · Defensive Distillation

1 Introduction

The use of deep learning in healthcare, particularly in accurately classifying chest X-rays for various lung diseases, is prevalent. Convolutional Neural Networks (CNNs) are commonly used in medical image analysis, but they are vulnerable to adversarial attacks that can cause misleading results. The vulnerabilities of automated medical imaging systems to such attacks are becoming more apparent in recent years [1]. As medical imaging moves towards cloud-based processing, cyberattacks can target the communication channels between the client and cloud, or attack the cloud infrastructure itself, to modify or steal patient data

[2]. With the increased use of AI in healthcare, malicious actors may manipulate medical imaging systems for financial gain through fraudulent billing or insurance claims using adversarial attacks. These makes medical imaging scenarios highly vulnerable to adversarial attacks, and it is imperative to develop a robust defensive strategy to ensure the secure deployment of automated medical imaging systems. Adversarial attacks are a type of security threat where an attacker manipulates input data in a way that is undetectable to humans but can cause a machine learning model to make incorrect predictions. Therefore, developing techniques to detect and defend against adversarial attacks is critical to ensure accurate diagnoses and treatment outcomes [3].

Recently, a ViT based self-ensembling approach was proposed tp defend against such attacks. Self-Ensembling Vision Transformers (SEVITs) is a type of ensembling ViT that have shown promising robust results in medical image analysis compared to Vanilla ViTs [4]. However, SEVIT is still cannot be considered as a benchmark robust solution to adversarial attacks and still could pose several vulnerabilities, which can compromise their accuracy and reliability. It is because SEVIT overlooks the privacy and confidentiality aspect of Trustworthy AI. Attacks like Model Extraction can easily extract such models like SEVIT even though they have showed significant results against adversarial attacks [5]. In addition to adversarial attacks, privacy attacks like model extraction in the medical imaging domain pose a significant threat to patient privacy and confidentiality. In the case of TB chest X-ray classification, for example, an attacker could extract the trained model from a hospital or a clinic and use it for malicious purposes, such as selling the model to third-party vendors, training their own model using the extracted model as a starting point, or using the model to infer sensitive patient information from their X-ray images.

Moreover, medical image datasets often contain sensitive patient information, such as demographics, medical history, and treatment plans. If a model is extracted from a hospital or a clinic without the institution's knowledge or consent, it could potentially expose sensitive patient data, violating their privacy rights [6]. To defend against such attacks, defensive distillation can be used as a countermeasure to model extraction attacks by making it more difficult for an attacker to extract the original model. By training a distilled model that approximates the behavior of the original model, but is less vulnerable to model extraction attacks, the attacker would need to spend more time and resources to extract the model, potentially deterring them from carrying out the attack. SEVIT showed robust performance against adversarial attacks, however, the MLP modules ensembled for each block are comparatively large in terms of parameters, which would make such robust model inefficient when it comes to deploying in real-world. To address this issue, this work aims to improve the robustness of the SEVIT model for Tuberculosis X-ray classification by using defensive distillation and adversarial training techniques, while also improving computational efficiency. The main significance of this research is to provide benchmark robustness of medical imaging systems, which is crucial in achieving better diagnosis and treatment outcomes for patients and healthcare providers. The outline of this work is structured as follows: Section 2 covers relevant literature, Section 3 presents the proposed method, and Section 4 explains the comprehensive experiments conducted and analyses made. The results and findings of the respective experiments are presented in Section 5, while Section 6 concludes the work and suggests possible future research directions.

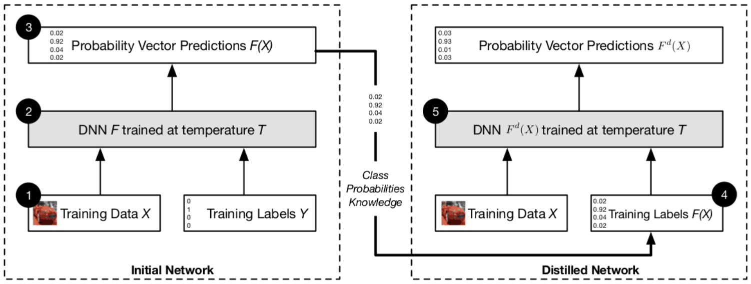


Fig. 1. Defensive distillation: Involves training a teacher model to generate a new dataset of soft targets for a smaller student model to learn from. The student model (distilled model) shows higher robustness due to the introduction of uncertainties in the output, making it more difficult for attackers to generate adversarial examples that can fool the model.

2 Related Works

Ensembling ViTs. ViT still faces challenges, such as vulnerability to adversarial attacks and robustness issues. [7] show that it is possible to generate more powerful adversarial attacks against ViTs by targeting an ensemble of intermediate representations in addition to the final class token. [4] improves the robustness of the ViT classifier as they proposed a novel Self-Ensemble Vision Transformer (SEViT) architecture to enhance the robustness of ViT against adversarial attacks for medical image classification. They propose to add a MLP classifier at the end of each block to utilize the patch tokens of the corresponding block and produce a probability distribution over the class labels. This results in a self-ensemble of classifiers that can be fused to obtain the final classification result.

Adversarial Attacks. The works by [8–12] analyzed a common theme of exploring adversarial attacks in machine learning. They aim to better understand the nature of these attacks and propose new techniques for mitigating their effects on learning models. [8] presents a new technique for creating adversarial examples, while [9] proposes a defense mechanism against them. [10] presents an efficient approach to training models to resist adversarial attacks, and [11] shows how to attack machine learning systems without knowledge of their internal workings. Finally, [12] demonstrates that even state-of-the-art models are vulnerable to adversarial attacks that exploit model extraction. Overall, these papers highlight the importance of understanding and mitigating adversarial attacks in machine learning to ensure the reliability and security of these systems. **Adversarial Defense.** Furthermore, among the few limited works on enhancing the adversarial robustness, the paper proposes a defense mechanism called defensive distillation, which is a form of knowledge transfer from a larger, more accurate model to a smaller, less accurate model. The idea behind defensive distillation is to make the smaller model less vulnerable to adversarial attacks by training it to imitate the outputs of the larger model, which is assumed to be more robust to adversarial perturbations (as shown in Figure 1). The authors demonstrate that defensive distillation can improve the robustness of deep neural networks to adversarial attacks, such as the FGSM and the DeepFool attacks, while maintaining high accuracy on clean inputs.

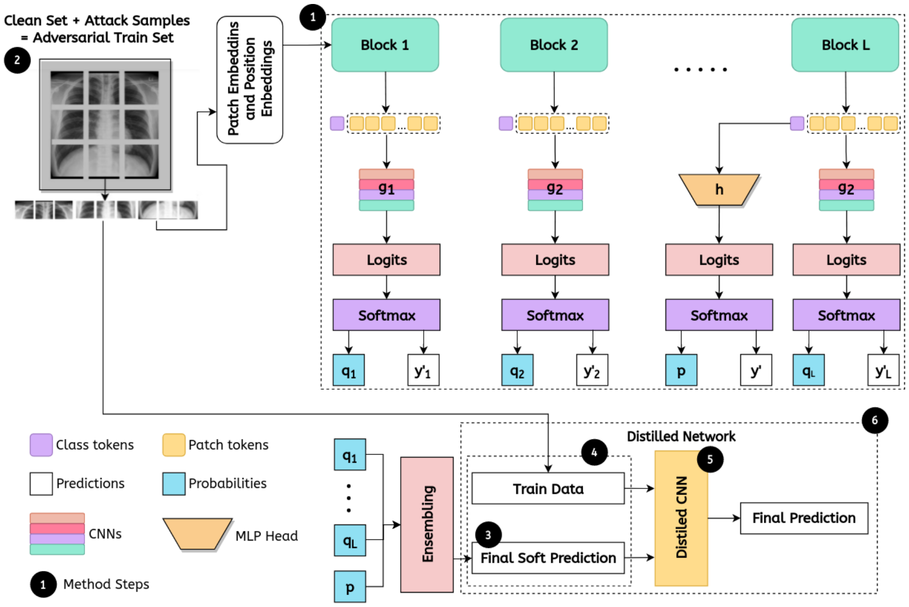


Fig. 2. The proposed SEViT-CNN framework extracts the patch tokens from the initial blocks and trains separate CNNs on Clean+Attack samples as shown in (1) and (2). A self-ensemble of these CNNs with the final ViT classifier goes through the distillation process with the new dataset to obtain a final distilled CNN model as seen in steps (3), (4), (5), and (6)

3 Proposed Method

The objective of an adversarial attack is to generate a perturbed image x', which is similar to the original medical image x within a certain distance metric (L_∞

norm), such that the output of a ViT-based classifier f(x') is different from the true label y with a high probability. The aim of SEVIT's defense mechanism is to obtain a robust classifier f' from the original classifier f. This robust classifier f' should have high accuracy on both clean images (P(f'(x) = y)) and perturbed images (P(f'(x') = y)).

SEVIT Ensemble. The ensembling approach of SEViT model aims to improve adversarial robustness by adding an MLP classifier at the end of each block, utilizing patch tokens to produce a probability distribution over class labels. The intermediate feature representations output by the initial blocks are considered useful for classification and harder to attack. This results in a self-ensemble of L classifiers that can be fused to obtain the final classification result. The SEViT model reduces computational complexity and increases adversarial robustness by adding MLP classifiers only to the first m ($m < L$) blocks and combining their results with the final classification head. The SEViT ensemble can be formed by performing majority voting or randomly choosing only c out of the initial m intermediate classifiers. The original SEVIT defense mechanism also includes a detection mechanism, which aims to distinguish between the original image x and the perturbed image x', especially when the attack is successful (f(x') $\neq$ y). The constraint on the distance metric between x and x' is that the L_∞ norm of their difference should be less than or equal to a predefined value epsilon. Although, we are not proposing any detection mechanism in our enhanced solution as we can use the same detection approach as SEVIT.

Enhanced SEVIT. Our two main hypothesis in context of SEVIT are as: First, small CNN modules would be more computationally efficient alternative to MLP modules for each ViT block. Second, defensive distillation when performed with adversarial training would make SEVIT more robust against adversarial attacks and model extraction attacks. Hence, we propose to enhance the existing SEVIT by modifying 3 major modifications: We propose to substitue MLP blocks with CNN instead to test for efficiency. Next, we perform adversarial training on the SEVIT model instead of training it on just clean samples. And last, we generate soft predictions to train a new distilled model. The proposed modifications are based on the fact that CNNs are more efficient than MLPs in learning spatial features from tokens through convolution operations at different levels of abstraction, which leads to improved generalization performance and reduced overfitting. Additionally, by training the model with adversarial examples, the model becomes better at handling perturbations during inference and generalizes better to new and unseen adversarial examples. The use of soft probabilities during distillation results in a smaller and more efficient model with faster inference time and reduced computational requirements. Moreover, the distilled model is more robust to adversarial attacks since the soft probabilities introduce uncertainty and variation in the output probabilities, making it harder for attackers to generate adversarial examples that can fool the model.

The proposed method architecture is illustrated in Figure 2. The pseudocode for our approach is as follows:

1. Define the SEVIT model architecture with CNN blocks instead of MLP blocks.
2. Train the SEVIT model on the original + adversarial datasets, i.e., adversarial training.
3. Extract the final soft predictions from the SEVIT-CNN model for all images in both datasets.
4. Create a new dataset consisting of the images from both the original and adversarial datasets along with their corresponding soft predictions.
5. Train a new distilled model on the new dataset, where each data point consists of an image and its corresponding soft predictions.
6. Deploy the distilled model instead of SEVIT-CNN model and evaluate the performance of the distilled model on clean and adversarial set.

4 Experimentation

Dataset. The experiments are performed on a medical imaging dataset by [13] that includes 7,000 chest X-ray images, and the classification task is binary, where the images are categorized as either Normal or Tuberculosis. The dataset was split randomly, where 80% of the images were allocated for training, 10% were assigned for validation, and the remaining 10% were used for testing. This random split of the dataset ensures that the training, validation, and testing sets are mutually exclusive, and it enables the evaluation of the model's generalization ability to new, unseen images.

Attack Types. In our study, we utilize the Foolbox library [14] to create 3 different types of L_∞ adversarial attacks such as FGSM [1], PGD [15] and AutoPGD [16] attacks. For generating attack samples with these algorithms, we use a value of perturbation budget ϵ=0.03 while keeping all other parameters at their default values.

MLPs alternatives. To implement the Vision Transformer (ViT) model, we utilized the ViT-B/16 architecture pretrained on ImageNet [5]. In order to create intermediate classifiers that take patch tokens as input, we trained 12 MLPs alternatives such as CNN and ResNet variants (Section 4.2), for every block. We used the fine-tuned ViT used in the original SEVIT method. Our experiments were conducted on a single Nvidia Quadro RTX 6000 GPU with 24 GB memory. The resulting ViT model achieved an accuracy of 96.38% on the original clean test set for chest X-rays. The accuracy of the intermediate classifiers on the original clean test set can be seen in Figure 3.

4.1 Ensembling Criterias

The proposed method is based on the working of SEVIT, where SEVIT utilizes the majority voting criteria for ensembling the predictions outputted by the intermediate MLP modules stemming from each VIT block. Instead of using just majority voting for ensembling, our improved work also provides additional insight across other ensembling criterias that can be implemented instead, depending on factors including the number of ensembling models, diversity among the predictions of individual modules, and so on. Based on these factors, we have experimented with two other voting criterias in addition to majority voting, i.e., Averaging and Weighted Averaging. Majority Voting can be sensitive to outliers, whereas Averaging and Weighted Averaging, on the other hand, can reduce the impact of outliers as they consider all the predictions along with better utilization of diversity in predictions. The findings are shown in Figure 3.

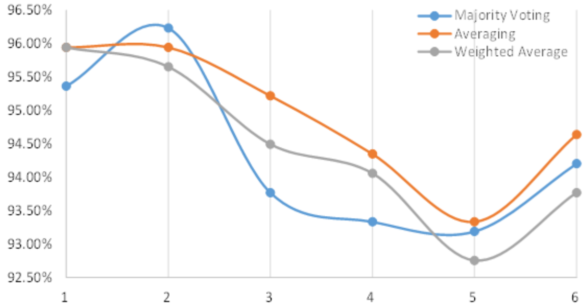
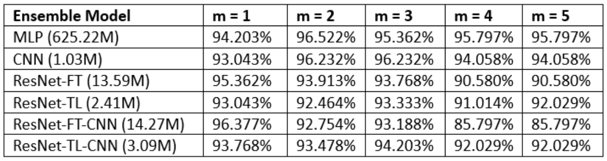


Fig. 3. Comparison of clean performance of CNN modules when ensembling is done with different voting criterias in the enhanced SEVIT-CNN. The x-axis is the number of ensembles while y-axis is the test accuracy

4.2 Efficient SEVIT with MLP alternatives

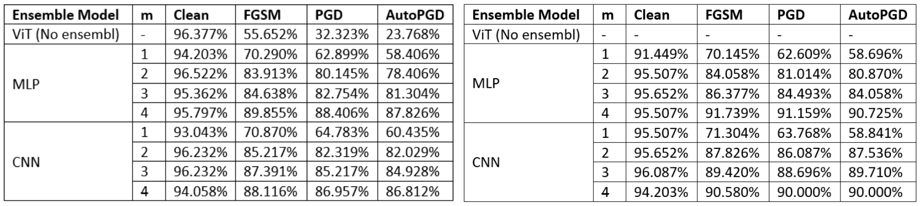
The computational efficiency of a machine learning model can be evaluated based on factors such as the number of parameters and training time. While the original MLP block in SEVIT is accurate, it requires a significant amount of computational resources to train and deploy with having about 625M parameters [17]. In our research, we enhanced the SEVIT model by incorporating several MLP alternatives that maintain the same level of clean and robust accuracy while reducing computational requirements. We experimented with different MLP alternatives, including a 2 convolution layer CNN, Fine-tuned ResNet50, Transfer-learned ResNet50, Fine-tuned ResNet50 with 2 additional convolution layers, and Transfer-learned ResNet50 with 2 additional convolution layers. Fine-tuning involves training the ResNet50 model on our specific dataset, while Transfer-learning involves using the pre-trained weights from ImageNet and adapting the model to our specific task. By comparing the performance (refer to Table 1 and Table 2) of these alternatives to the original MLP block, we are able to determine the most efficient option for TB classification with maintained performance while having reduction in parameters, space complexity and computations.

Table 1. Clean performance of different MLP alternatives across several ensembles using Majority Voting



Ensemble Model	m = 1	m = 2	m = 3	m = 4	m = 5
MLP (625.22M)	94.203%	96.522%	95.362%	95.797%	95.797%
CNN (1.03M)	93.043%	96.232%	96.232%	94.058%	94.058%
ResNet-FT (13.59M)	95.362%	93.913%	93.768%	90.580%	90.580%
ResNet-TL (2.41M)	93.043%	92.464%	93.333%	91.014%	92.029%
ResNet-FT-CNN (14.27M)	96.377%	92.754%	93.188%	85.797%	85.797%
ResNet-TL-CNN (3.09M)	93.768%	93.478%	94.203%	92.029%	92.029%

Table 2. VIT vs SEVIT vs SEVIT-CNN in terms of Clean and Robust accuracy across different number of intermediate ensembles. Pre-adversarial training performance (left) vs Post-adversarial training (right)



Ensemble Model	m	Clean	FGSM	PGD	AutoPGD
ViT (No ensembl)	-	96.377%	55.652%	32.323%	23.768%
MLP	1	94.203%	70.290%	62.899%	58.406%
	2	96.522%	83.913%	80.145%	78.406%
	3	95.362%	84.638%	82.754%	81.304%
	4	95.797%	89.855%	88.406%	87.826%
CNN	1	93.043%	70.870%	64.783%	60.435%
	2	96.232%	85.217%	82.319%	82.029%
	3	96.232%	87.391%	85.217%	84.928%
	4	94.058%	88.116%	86.957%	86.812%

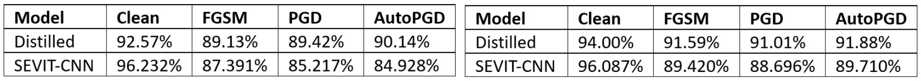
Ensemble Model	m	Clean	FGSM	PGD	AutoPGD
ViT (No ensembl)	-	-	-	-	-
MLP	1	91.449%	70.145%	62.609%	58.696%
	2	95.507%	84.058%	81.014%	80.870%
	3	95.652%	86.377%	84.493%	84.058%
	4	95.507%	91.739%	91.159%	90.725%
CNN	1	95.507%	71.304%	63.768%	58.841%
	2	95.652%	87.826%	86.087%	87.536%
	3	96.087%	89.420%	88.696%	89.710%
	4	94.203%	90.580%	90.000%	90.000%

4.3 Defensive Distillation vs Extraction Attack

Model extraction attacks can put patient privacy and confidentiality at risk by allowing attackers to extract the trained model from healthcare institutions and use it for malicious purposes. To combat this, a distilled model can be trained to approximate the original model's behavior while being less vulnerable to extraction attacks, making it more difficult for attackers to extract (steal) the model [18]. The above experiments conducted in this study suggest that the SEVIT-CNN is a better alternative than the original SEVIT with MLP blocks. As a defender, we perform defensive distillation and obtain the distilled model, which was generated from the original SEVIT-CNN using its soft probabilities. The distilled model having a smaller architecture of 5 convolution layer CNN, is compared to the original model in terms of efficiency and robustness and the findings are presented in Table 3. The results (as described in Results section) indicate that the distilled model is a superior option for deployment due to its lower susceptibility to attack samples [19].

In the next step of our experiment, we evaluate the performance of SEVIT-CNN and the distilled model against model extraction attacks. Both models are tested in a black-box setting where the attacker can only input queries and receive outputs. In this experiment, the attacker utilizes this input-output relationship to create a replica model (a three convolution layer CNN) of the original model. The aim is to compare the extracted model's performance in two scenarios: one where SEVIT-CNN is deployed, and the other where the distilled model is deployed (Table 4). This comparison allows us to determine that deploying which model would be more resilient to model extraction attacks.

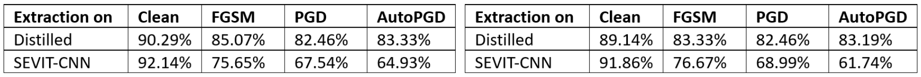
Table 3. Comparison of distilled model and SEVIT-CNN (with m=3, i.e., number of intermediate ensembling blocks = 3) performance against Clean Samples and Attack Samples. Pre-adversarial training performance (left) vs Post-adversarial training (right)



Model	Clean	FGSM	PGD	AutoPGD
Distilled	92.57%	89.13%	89.42%	90.14%
SEVIT-CNN	96.232%	87.391%	85.217%	84.928%

Model	Clean	FGSM	PGD	AutoPGD
Distilled	94.00%	91.59%	91.01%	91.88%
SEVIT-CNN	96.087%	89.420%	88.696%	89.710%

Table 4. Comparison of extracted model when model extraction is performed on distilled model V/s when extraction is performed on SEVIT-CNN (with m=3). The performance of extracted models are against Clean Samples and Attack Samples. Pre-adversarial training performance (left) vs Post-adversarial training (right)



Extraction on	Clean	FGSM	PGD	AutoPGD
Distilled	90.29%	85.07%	82.46%	83.33%
SEVIT-CNN	92.14%	75.65%	67.54%	64.93%

Extraction on	Clean	FGSM	PGD	AutoPGD
Distilled	89.14%	83.33%	82.46%	83.19%
SEVIT-CNN	91.86%	76.67%	68.99%	61.74%

4.4 Adversarial Training

In order to make SEVIT-CNN more robust against adversarial attacks, we utilized multiple defense strategies, however, there was still a noticeable difference between the robust and clean accuracy in the previous experiments, which is not desirable as a good defender should aim to minimize the gap between them. To improve the robustness, we employed adversarial training, which is a technique that has been shown to enhance the robustness of deep neural networks against various adversarial attacks in the literature. By generating adversarial examples and adding them to the training set, the model learns to better handle these perturbations during inference. This technique effectively increases the model's ability to generalize to new and unseen adversarial examples. With this aim, we repeated the previous experiments with adversarial training and compared the results side by side with the experiments conducted without adversarial training in Section 4.2 and Section 4.3. The outcomes of the pre-adversarial training and post-adversarial training experiments are presented in Table 3 and 4.

5 Results and Discussions

5.1 Results on voting criteria

In our experimentation with different voting criteria, as can be seen in Figure 3, we observed that Majority Voting works well when the number of models in the ensemble is small, typically 2 or 3 individual CNN modules each stemming from corresponding VIT block in our proposed case. This is because it is easier to reach a consensus among fewer models, leading to more reliable predictions. Moreover, Averaging was effective in combining more diverse opinions from individual models, and is useful when the predictions of individual modules are diverse, as Averaging better utilizes this diversity and lead to improved performance. In addition, Weighted averaging, lead to overfitting when the weights are not selected carefully, especially when the number of models is large. Therefore, it does not work well when there are too many models in the ensemble. Based on these observations (Figure 3), we can state that different voting criteria based on the number of ensembling models should be selected.

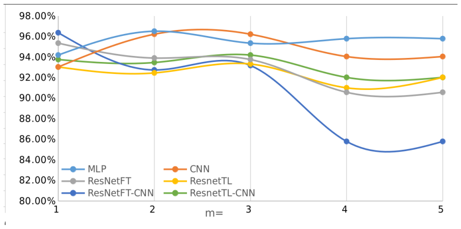


Fig. 4. Comparison of clean performance of different MLP alternatives across various ensembles

5.2 Results on Computational Efficiency

We have compared the computational efficiency of our proposed approach with the original MLP block by ensembling consistently on 3 models (i.e., m=3), as illustrated in Figure 4 and 5. Our experiments conclude that the 2 convolution layer CNN alternative outperforms the original MLP block in terms of clean accuracy and maintains high accuracy even against adversarial attack samples. This can be attributed to CNN's ability to learn spatial features from tokens through convolution operations at different levels of abstraction, leading to reduced overfitting and improved generalization performance. For example, when subjected to attack samples generated by FGSM algorithm, the CNN alternative achieved 87.4% robust accuracy, which is higher than the 84.6% robust accuracy achieved by the 4-layered MLP block. This increment in robust accuracy can be noticed consistently on other attacks as well. Other alternatives like ResNet-FT and ResNet-TL-CNN have moderate performance, while ResNet-FT-CNN and ResNet-TL show poor performance in terms of robustness. Although adding convolution layers can increase model expressiveness, it also increases the risk of overfitting and makes the model more vulnerable to adversarial attacks. The reason for lower performance for later alternatives could be that since transfer learning can help in learning useful features for the task, but it may also inherit irrelevant or harmful features, leading to decreased robustness. Based on these findings, we conclude that the CNN alternative is the best choice for SEVIT considering clean and robust accuracy, as well as computational efficiency.

5.3 Results on Robustness via Distillation and Adversarial Training

We conducted a comparison of the original SEVIT-CNN and the distilled model in terms of their clean and robust performance. Our findings, as presented in

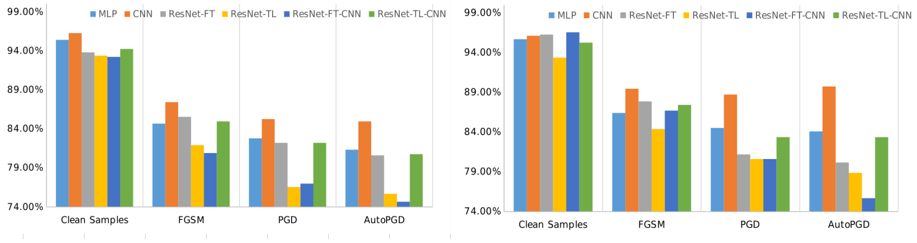


Fig. 5. Clean and adversarial performance of MLP alternatives Before adversarial training (left) vs After adversarial training (right). For each MLP alternative, the number of ensembles are m=3 where the attack samples have $\epsilon = 0.03$

Figure 6, indicate that the distilled model, which was created through distillation from the original SEVIT-CNN model using soft probabilities, outperformed the original model in terms of robust accuracy despite having a smaller architecture. However, there was a slight decline of around 4% in the clean accuracy of the distilled model, which could be viewed as a reasonable trade-off between clean and robust accuracy. Distillation using soft probabilities leads to a compact and smaller architecture, which results in faster inference time and lower computational requirements, making the model more efficient. Furthermore, the distilled model is less susceptible to adversarial attacks as soft probabilities impart a smoothing effect during the distillation process. This is because soft probabilities introduce some degree of uncertainty and variation in the output probabilities of the model, making it more difficult for an attacker to generate adversarial examples that can fool the model. Hence, the distilled model is robust against adversarial attacks than the original SEVIT-CNN model with a slight trade-off with clean accuracy.

Moreover, in this experiment, we have also evaluated the vulnerability of two models- the original SEVIT-CNN model with a 3-member ensemble and the distilled model- to model extraction attacks. The attacker was able to extract both models and compare their accuracy. The results are presented in Figure 7, which shows that the attacker's clean accuracy in reproducing the distilled model (90.29%) is lower compared to that of the original SEVIT-CNN model (92.14%). This is because the distilled model was trained using soft probabilities generated by the original model, which resulted in a smoothing effect that made it less susceptible to model extraction attacks. The attacker was less likely to recover the exact parameters of the original model from the smoothed probabilities. These findings indicate that using defensive distillation can be a useful technique for improving the security of SEVIT against model extraction attacks. Thus, we can conclude that deploying the distilled model is a more secure option against model extraction attacks than the SEVIT-CNN model.

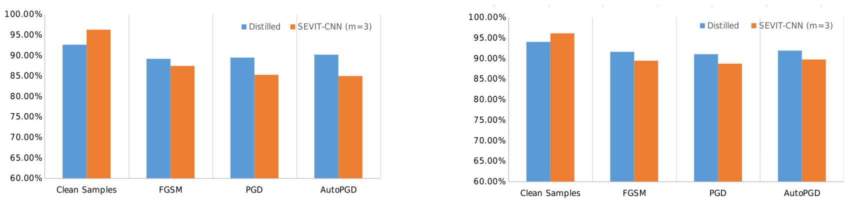


Fig. 6. Comparison of distilled model and SEVIT-CNN performance against Clean Samples and Attack Samples in the case of pre-adversarial training (left) vs post-adversarial training (right). The above table is the Pre-adversarial training performance while the down table is Post-adversarial training case

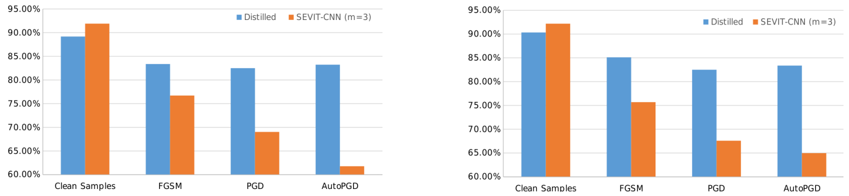


Fig. 7. Comparison of extracted model when model extraction is performed on distilled model V/s when extraction is performed on SEVIT-CNN. The performance of extracted models are against Clean Samples and Attack Samples. Pre-adversarial training (left) V/s Post-adversarial training (right)

5.4 Results on Adversarial Training

We evaluated the performance of enhanced SEVIT that utilizes MLP alternative like CNN, along with deploying distilled model would a robust choice. Furthermore, to achieve even higher robustness, we also perform adversarial training in Section 4.4. The results indicate the effectiveness of adversarial training in improving the robustness against adversarial attacks as shown in Figure 5. Among the two cases, in the case of pre-adversarial training, both models (Distilled and SEVIT-CNN) have high accuracy on clean samples, but their performance on adversarial examples is increased even higher by about +2% after adversarial training. Additionally, the models' vulnerability to model extraction attacks is high in this pre-adversarial training case, with the extracted models' accuracy on clean and adversarial examples being considerably high (Figure 6 and 7). However, after adversarial training in case of post-adversarial training, both models show a considerable improvement in their robustness against adversarial attacks. The models' accuracy on clean and adversarial examples generated by FGSM, PGD, and AutoPGD attacks is significantly improved, indicating that the models have learned to generalize better on adversarial examples. Furthermore, the models' vulnerability to model extraction attacks is significantly reduced, with the accuracy of extracted models on clean and adversarial examples being considerably lower compared to Case of pre-adversarial training. These results suggest that adversarial training has improved the models' robustness and made it slightly more challenging to extract the model.

6 Conclusion

The vulnerabilities of automated medical imaging systems to adversarial attacks are becoming more apparent in recent years. With such elevation, more and more attacks could pose greater threat to deployed automated medical imaging models. A recent ViT based approach, Self-Ensemble Vision Transformer shows robust performance against a range of adversarial attacks. However, it still doesn't provide any benchmark robustness along with posing susceptibility against privacy attacks like model extraction. To overcome these issues, this work have proposed and improved architecture to overcome computational bottleneck and improved defensive approach by modifying 3 major modifications: We propose to substitute MLP blocks with CNN instead to test for efficiency. Next, we perform adversarial training on the SEVIT model instead of training it on just clean samples. And last, we generate soft predictions to train a new distilled model. These techniques have been tested in combination to provide a stronger defense against adversarial attacks and model extraction attacks. We prove the effectiveness of our novel enhanced approach on SEViT-CNN using extensive set of experimentation on publicly available Tuberculosis X-ray dataset.

In the future, we aim to further enhance the proposed SEVIT-CNN model using (i) defensive approaches like differential privacy (ii) explore the trade-off between adversarial robustness and model accuracy when performed defensive distillation (iii) exploring diverse set of alternatives for each block rather than just having CNN for every block. This would increase the diversity among the ensemble models and further improve overall performance. By continuing to research and develop new defensive mechanisms, we can improve the adversarial robustness of medical imaging models and ensure the accuracy and reliability of medical diagnoses and treatments.

References

1. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
2. Qi-Xian Huang, Wai Leong Yap, Min-Yi Chiu, and Hung-Min Sun. Privacy-preserving deep learning with learnable image encryption on medical images. *IEEE Access*, 10:66345–66355, 2022.
3. Sara Kaviani, Ki Jin Han, and Insoo Sohn. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. *Expert Systems with Applications*, page 116815, 2022.
4. Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 376–386. Springer, 2022.

5. Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.
6. Raihan Ur Rasool, Hafiz Farooq Ahmad, Wajid Rafique, Adnan Qayyum, and Junaid Qadir. Security and privacy of internet of medical things: A contemporary review in the age of surveillance, botnets, and adversarial ml. *Journal of Network and Computer Applications*, page 103332, 2022.
7. Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021.
8. Hashmat Shadab Malik, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Adversarial pixel restoration as a pretext task for transferable perturbations. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.
9. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.
10. Boxi Wu, Jindong Gu, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Towards efficient adversarial training on vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 307–325. Springer, 2022.
11. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
12. Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. Model extraction and adversarial transferability, your bert is vulnerable! *arXiv preprint arXiv:2103.10013*, 2021.
13. Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.
14. Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020.
15. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
16. Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.
17. Xin Zhou, Zhepei Wang, Xiangyong Wen, Jiangchao Zhu, Chao Xu, and Fei Gao. Decentralized spatial-temporal trajectory planning for multicopter swarms. *arXiv preprint arXiv:2106.12481*, 2021.
18. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.
19. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.