MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

On enhancement of Self-Ensembling Vision Transformer (SEViT) in Chest XRay Classification using Defensive Distillation & Adversarial Training

2023

Salma Alrashdi, Raza Imam, Baketah Alrashdi

mbzuai.ac.ae

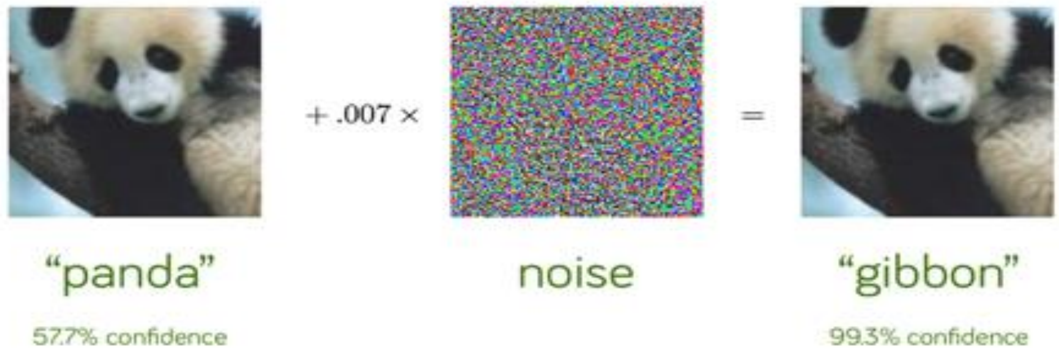
HC701 Final Project presentation

Outline

- Introduction
- Motivation
- SEVIT
- Problem Statement
- Proposed Solution
- Experiments
- Results
- Discussion
- Contribution

Introduction

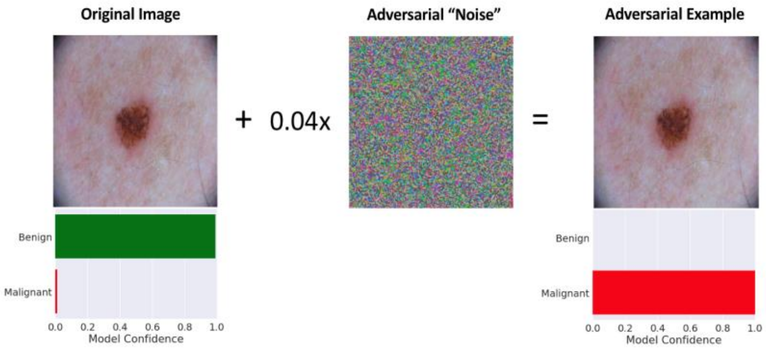
- This paper, is propose a novel self-ensembling method to enhance the robustness of ViT in the presence of adversarial attacks.



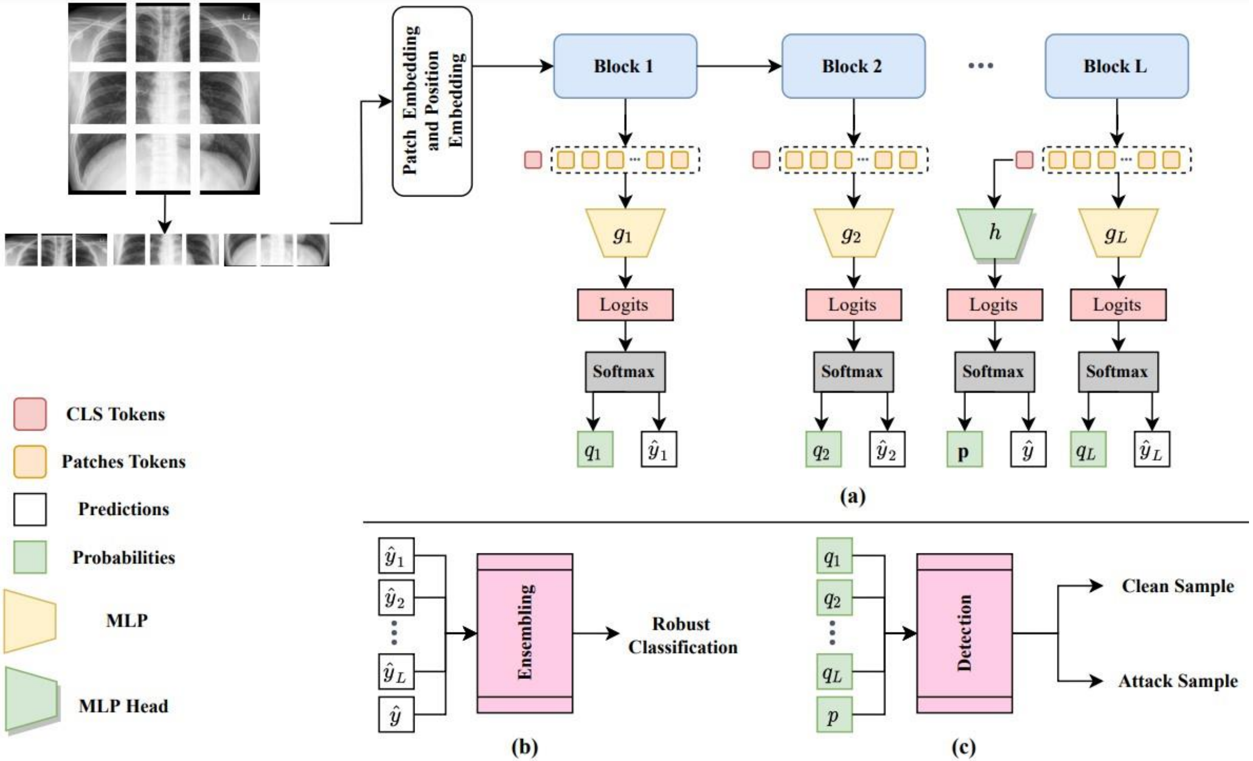
Motivations

The motivation for studying the adversarial robustness of deep learning-based medical imaging systems has been discussed :

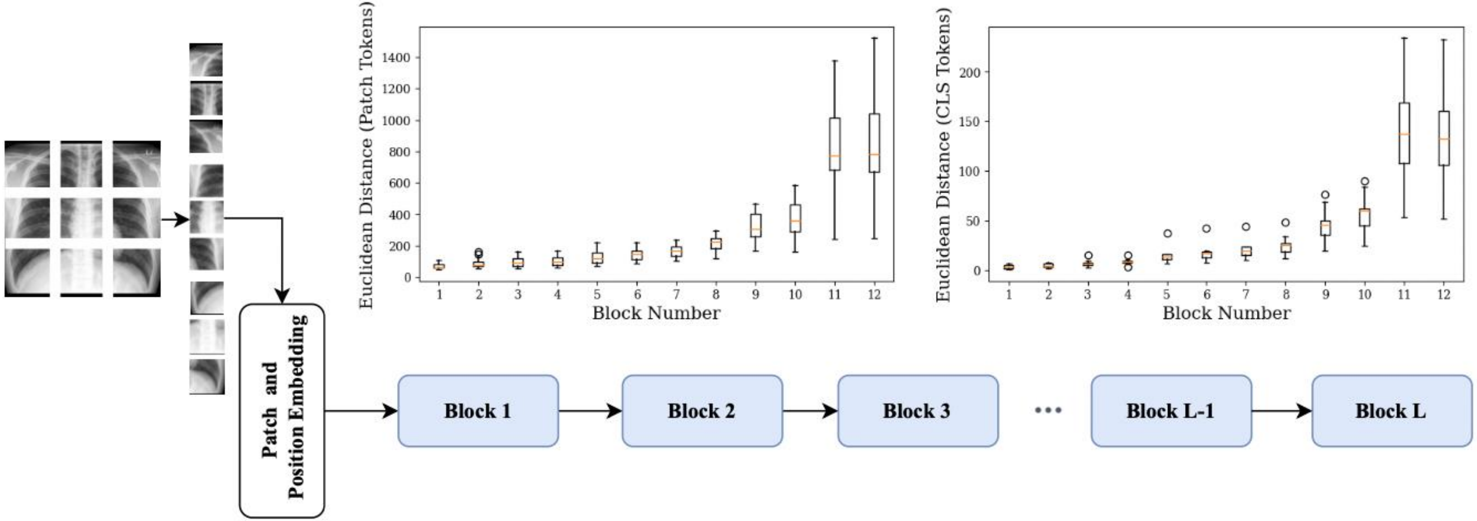
- (i) Automated systems deployed by insurance companies to process reimbursement claims.
- (ii) Automated systems deployed by regulators to confirm results of clinical trials.



SEVIT Architecture



SEVIT



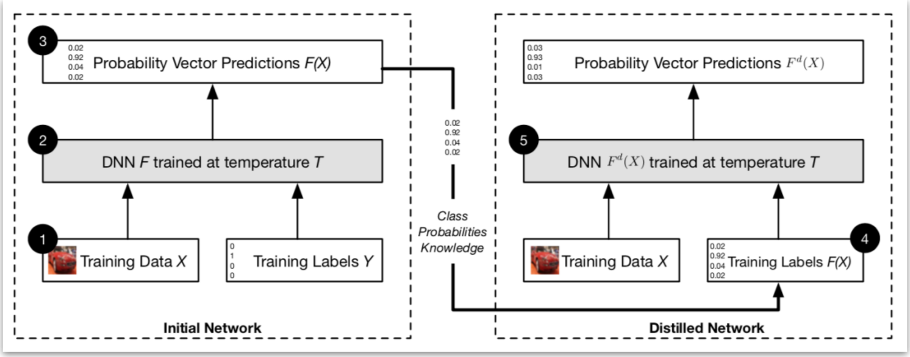
Euclidean Distance (Patch Tokens)
0
200
400
600
800
1000
1200
1400
1
2
3
4
5
6
7
8
9
10
11
12
Block Number
Euclidean Distance (CLS Tokens)
0
50
100
150
200
1
2
3
4
5
6
7
8
9
10
11
12
Block Number
Patch and Position Embedding
Block 1
Block 2
Block 3
...
Block L-1
Block L

Problem Statement

- Not really robust against Model Extraction attacks, which can reveal major vulnerability in Medical Imaging
- Computationally inefficient solutions are not really practically deployable
- Ensembling via Majority Voting is not always the best approach

Defenses

- **Defensive distillation.**

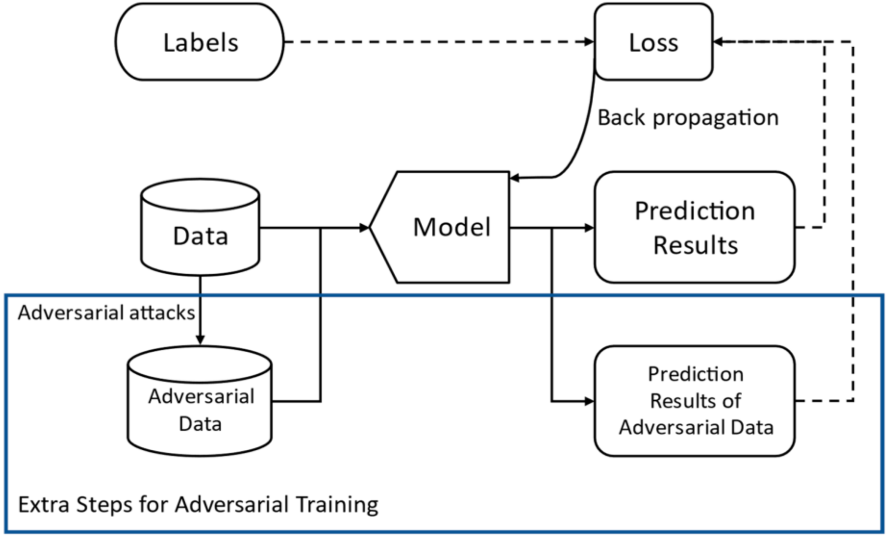


- Involves training a model with **soft** probabilities.
- **Soft** Probabilities makes the model more resistant to small perturbations in the input data, which are a common characteristic of adversarial examples.
- Because small perturbations in the input data are less likely to cause large changes in the output probabilities when they are "softened" than when they are not.

Defenses 2

- **Adversarial Training.**

- Involves training a model on both clean and adversarial examples.
- Improve model's ability to withstand such adversarial attacks.
- Helps the model learn to generalize better to new and unseen examples, including those that may be adversarially crafted.



Proposed Method

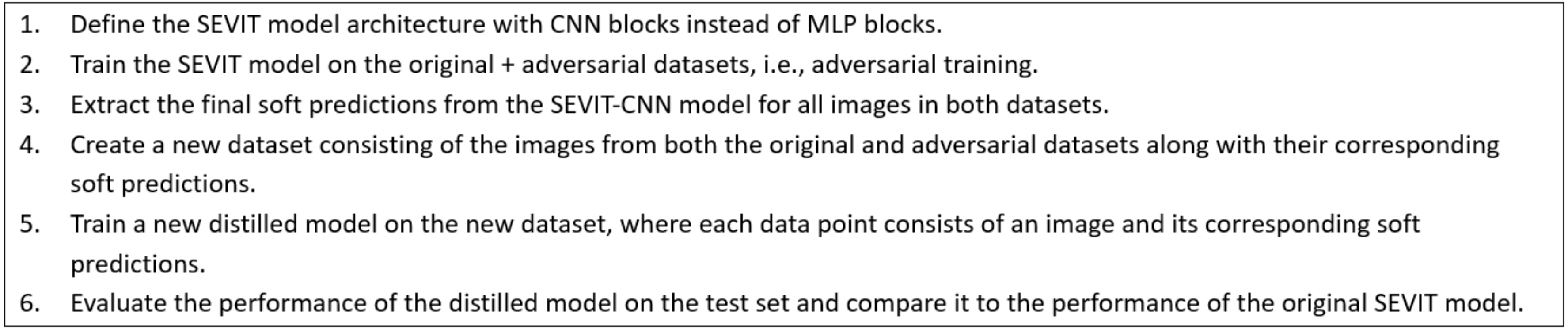
- Instead of using MLP blocks, use CNN or Fine-Tuned Model with additional CNN layers
- Adversarial Training increases the overall robust accuracy
- Surrogate Model generated using Defensive Distillation can be deployed against Extraction

Proposed Method - Details

Hypothesis

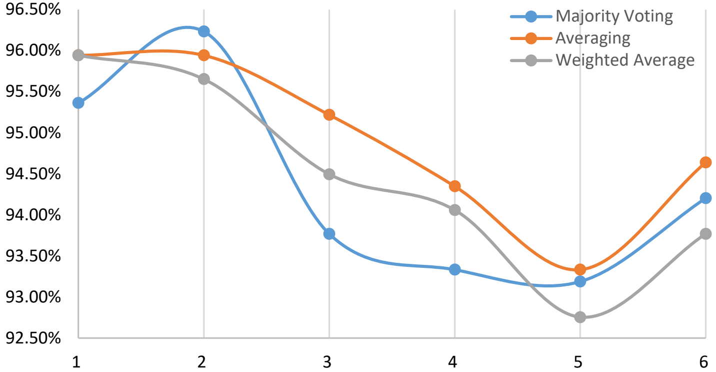
Defensive Distillation when performed with Adversarial training would make SEVIT more robust against Adversarial attacks and Model Extraction Attacks

Pseudocode for our approach

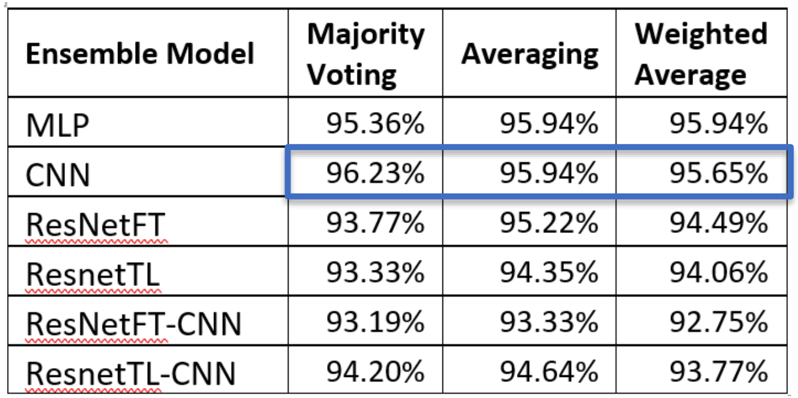


1. Define the SEVIT model architecture with CNN blocks instead of MLP blocks.
2. Train the SEVIT model on the original + adversarial datasets, i.e., adversarial training.
3. Extract the final soft predictions from the SEVIT-CNN model for all images in both datasets.
4. Create a new dataset consisting 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. Evaluate the performance of the distilled model on the test set and compare it to the performance of the original SEVIT model.

Experiments- Ensembling Criterias



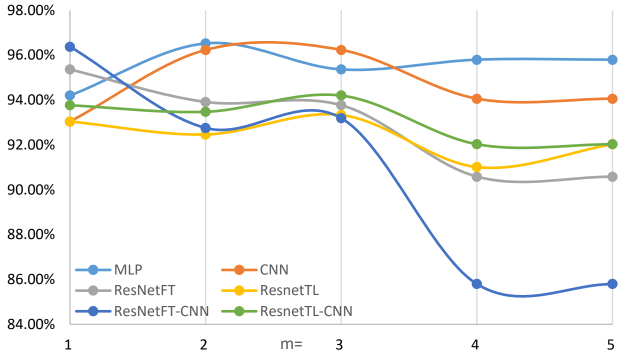
Clean Performance of MLP alts with different ensembling criterias



Ensemble Model	Majority Voting	Averaging	Weighted Average
MLP	95.36%	95.94%	95.94%
CNN	96.23%	95.94%	95.65%
ResNetFT	93.77%	95.22%	94.49%
ResnetTL	93.33%	94.35%	94.06%
ResNetFT-CNN	93.19%	93.33%	92.75%
ResnetTL-CNN	94.20%	94.64%	93.77%

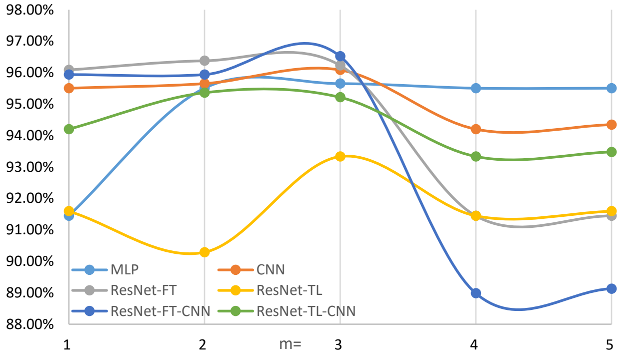
- Majority voting works well when the number of models in the ensemble is small, typically 2 or 3. It is because it is easier to reach a consensus among fewer models.
- Averaging allows more diverse opinions to be combined.
- Weighted averaging can lead to overfitting if the weights are not selected carefully, especially when the number of models is large. Therefore, it may not work well when there are too many models in the ensemble.

Experiments-Computational Efficient



Clean Performance of MLP alts **Before** Adv training

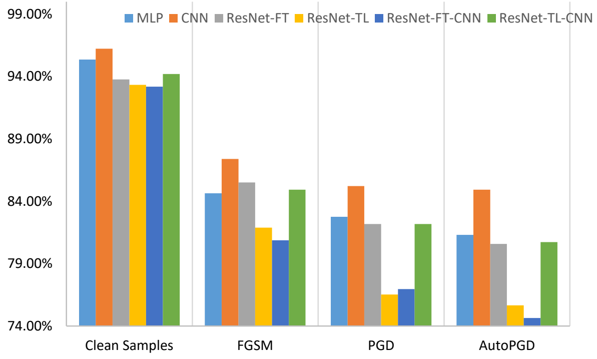
Ensemble Model	**m = 1**	**m = 2**	**m = 3**	**m = 4**	**m = 5**
MLP (625.22M)	94.203%	96.522%	95.362%	95.797%	95.797%
CNN (1.03M)	93.043%	96.232%	96.232%	94.058%	94.058%
ResNet-FT (13.59M)	95.362%	93.913%	93.768%	90.580%	90.580%
ResNet-TL (2.41M)	93.043%	92.464%	93.333%	91.014%	92.029%
ResNet-FT-CNN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



Clean Performance of MLP alts **After** Adv training

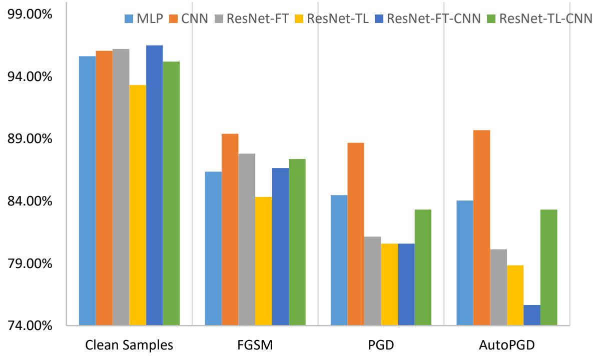
Ensemble Model	**m = 1**	**m = 2**	**m = 3**	**m = 4**	**m = 5**
MLP (625.22M)	91.449%	95.507%	95.652%	95.507%	95.507%
CNN (1.03M)	95.507%	95.652%	96.087%	94.203%	94.348%
ResNet-FT (13.59M)	96.087%	96.377%	96.232%	91.449%	91.449%
ResNet-TL (2.41M)	91.594%	90.290%	93.333%	91.449%	91.594%
ResNet-FT-CNN (14.27M)	95.942%	95.942%	96.522%	88.986%	89.130%
ResNet-TL-CNN (3.09M)	94.203%	95.362%	95.217%	93.333%	93.478%

Experiments- Adversarial Performance



Adversarial Performance of MLP alts **Before** Adv training

Ensemble Model	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
ViT	**96.377%**	**55.652%**	**32.323%**	**28.994%**	**23.768%**	**47.836%**
MLP	95.362%	84.638%	82.754%	95.362%	81.304%	95.362%
CNN	96.232%	87.391%	85.217%	96.232%	84.928%	96.232%
ResNet-FT	93.768%	85.507%	82.174%	93.768%	80.580%	93.768%
ResNet-TL	93.333%	81.884%	76.522%	93.333%	75.652%	93.333%
ResNet-FT-CNN	93.188%	80.870%	76.957%	93.188%	74.638%	93.188%
ResNet-TL-CNN	94.203%	84.928%	82.174%	94.203%	80.725%	94.203%

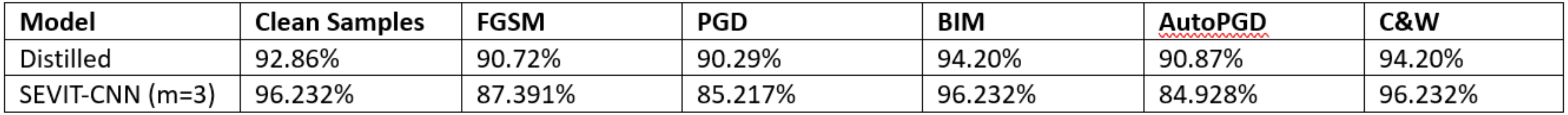


Adversarial Performance of MLP alts **After** Adv training

Ensemble Model	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
ViT	-	-	-	-	-	-
MLP	95.652%	86.377%	84.493%	95.652%	84.058%	95.652%
CNN	96.087%	89.420%	88.696%	96.087%	89.710%	96.087%
ResNet-FT	96.232%	87.826%	81.159%	96.232%	80.145%	96.232%
ResNet-TL	93.333%	84.348%	80.580%	93.333%	78.841%	93.333%
ResNet-FT-CNN	96.522%	86.667%	80.580%	96.522%	75.652%	96.522%
ResNet-TL-CNN	95.217%	87.391%	83.333%	95.217%	83.333%	95.217%

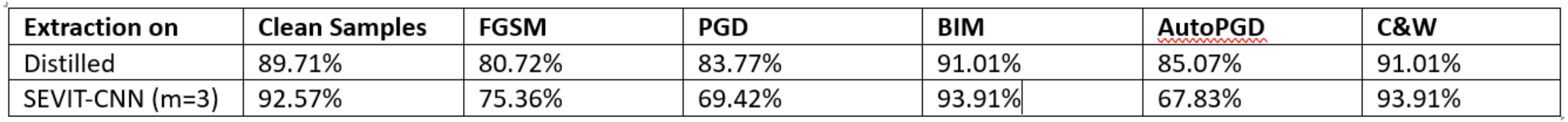
Experiments- Distillation & Extraction

Distilled Model from SEVIT-CNN (Testing is on Attack Samples generated by SEVIT-CNN) (**Before** Adv Training)



Model	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
Distilled	92.86%	90.72%	90.29%	94.20%	90.87%	94.20%
SEVIT-CNN (m=3)	96.232%	87.391%	85.217%	96.232%	84.928%	96.232%

***Model Extraction Attack** on Distilled Model and Original Model (**Before** Adversarial Training)



Extraction on	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
Distilled	89.71%	80.72%	83.77%	91.01%	85.07%	91.01%
SEVIT-CNN (m=3)	92.57%	75.36%	69.42%	93.91%	67.83%	93.91%

Experiments- Distillation & Extraction 2

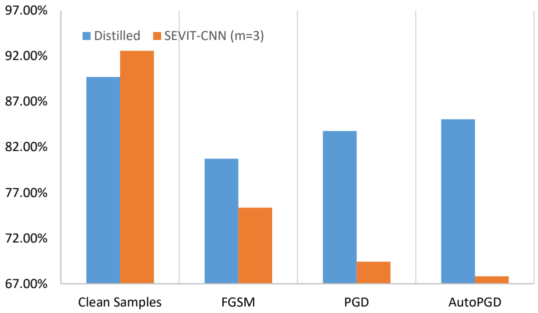
Distilled Model from SEVIT-CNN (Testing is on Attack Samples generated by SEVIT-CNN) (**After** Adv Training)

Model	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
Distilled	92.00%	88.99%	88.70%%	93.33%%	88.55%	93.33%
SEVIT-CNN (m=3)	96.087%	89.420%	88.696%	96.087%	89.710%	96.087%

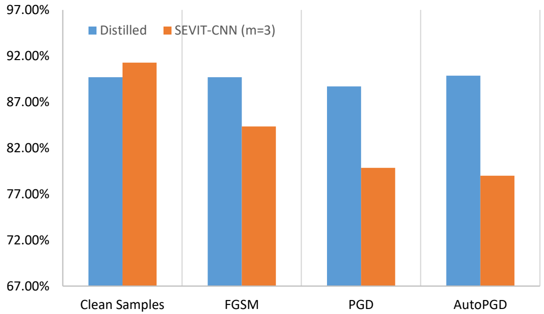
***Model Extraction Attack** on Distilled Model and Original Model (**After** Adversarial Training)

Extraction on	Clean Samples	FGSM	PGD	BIM	AutoPGD	C&W
Distilled	89.71%	89.71%%	88.70%	91.01%	89.86%	91.01%
SEVIT-CNN (m=3)	91.29%	84.35%	79.86%	92.61%	78.99%	92.61%

Model Extraction **Before** Adersarial Training →

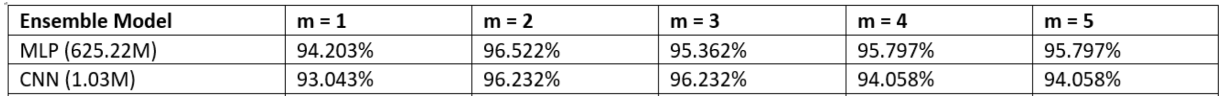


Model Extraction **After** Adversarial Training →



Results

- Among several alternatives for MLP modules in SEVIT, CNN performed the best in terms of clean and adversarial performance (Efficiency)



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%

- It’s a better choice to ensemble using different voting criteria based on the no. of ensembling models.

Results

- Against Model Extraction attack, we notice a trade-off between clean accuracy and robust accuracy (in both case of pre-adv. training and post-adv. training)
- Trade-off is the result of competing objectives of the attacker and defender.
- Better to deploy Distilled version of SEVIT-CNN as it will more robust than SEVIT to adv. attacks.

Conclusion

- MLP alternatives will be a better choice for computational efficiency
- Enhanced the overall robustness of SEVIT using the combination of Defensive Distillation and Adversarial Training
- Different voting criteria based on the no. of ensembling models

Future goals

- Evaluate enhanced SEViT in the context of natural images
- Explore the possibility for mobile deployment

Team Member Contribution

- Raza: final report, organized meets, defensive experiments, presentation
- Salma: final report, defensive experiment, literature review, presentation
- Baketah: contribute on the final report, extraction experiments, presentation