**Tables 1, 2, 2\_1 are before adversarial training. Tables 4, 5, 5\_1 are after adversarial training. All experiments are on perturb=0.03**

**Table 1.** Clean Performance of MLP alternatives ensemble modules for SEVIT (VIT + Ensemble modules). Here abbreviations are as, MW: Majority Voting, A: Averaging, and WA: Weighted Averaging

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **# Trainable Params** | **Ensemble Criteria** | **m = 1** | **m = 2** | **m = 3** | **m = 4** | **m = 5** |
| **ViT** | **-** |  | **96.377%** | | | | |
| MLP | 625219970 | MV | 94.203% | 96.522% | 95.362% | 95.797% | 95.797% |
| A | 95.217% | 95.942% | 95.942% | 95.217% | 95.362% |
| WA | 95.507% | 95.507% | 95.942% | 95.652% | 95.942% |
| CNN | 1032770 | MV | 93.043% | 96.232% | 96.232% | 94.058% | 94.058% |
| A | 95.507% | 96.667% | 95.942% | 94.928% | 91.304% |
| WA | 95.507% | 96.667% | 95.652% | 94.058% | 88.406% |
| ResNet-FT | 13586114 | MV | 95.362% | 93.913% | 93.768% | 90.580% | 90.580% |
| A | 96.667% | 95.942% | 95.217% | 92.029% | 89.710% |
| WA | 96.522% | 95.797% | 94.493% | 91.884% | 89.275% |
| ResNet-TL | 2409602 | MV | 93.043% | 92.464% | 93.333% | 91.014% | 92.029% |
| A | 95.652% | 95.217% | 94.348% | 93.478% | 89.855% |
| WA | 92.754% | 94.493% | 94.058% | 92.899% | 86.087% |
| ResNet-FT-CNN | 14267810 | MV | 96.377% | 92.754% | 93.188% | 85.797% | 85.797% |
| A | 97.101% | 96.667% | 93.333% | 89.420% | 83.478% |
| WA | 96.812% | 95.797% | 92.754% | 86.957% | 77.391% |
| ResNet-TL-CNN | 3091298 | MV | 93.768% | 93.478% | 94.203% | 92.029% | 92.029% |
| A | 95.652% | 94.928% | 94.638% | 94.493% | 91.739% |
| WA | 94.493% | 94.058% | 93.768% | 93.188% | 83.623% |

**Table 2.** Adversarial Attack Performance of MLP alternatives ensemble modules for SEVIT (VIT + Ensemble modules) via Majority Voting. Here FGSM, PGD, BIM, and AutoPGD attack samples are having perturbation budget equal to 0.03

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **m** | **Clean Samples** | **FGSM** | **PGD** | **BIM** | **AutoPGD** | **C&W** |
| **ViT** | **-** | **96.377%** | **55.652%** | **32.323%** | **28.994%** | **23.768%** | **47.836%** |
| MLP | 1 | 94.203% | 70.290% | 62.899% | 94.203% | 58.406% | 94.203% |
| 2 | 96.522% | 83.913% | 80.145% | 96.522% | 78.406% | 96.522% |
| 3 | 95.362% | 84.638% | 82.754% | 95.362% | 81.304% | 95.362% |
| 4 | 95.797% | 89.855% | 88.406% | 95.797% | 87.826% | 95.797% |
| CNN | 1 | 93.043% | 70.870% | 64.783% | 93.043% | 60.435% | 93.043% |
| 2 | 96.232% | 85.217% | 82.319% | 96.232% | 82.029% | 96.232% |
| 3 | 96.232% | 87.391% | 85.217% | 96.232% | 84.928% | 96.232% |
| ~~4~~ | 94.058% | **88.116%** | 86.957% | 94.058% | 86.812% | 94.058% |
| ResNet-FT | 1 | 95.362% | 68.261% | 58.841% | 95.362% | 50.580% | 95.362% |
| 2 | 93.913% | 83.768% | 80.580% | 93.913% | 77.536% | 93.913% |
| 3 | 93.768% | 85.507% | 82.174% | 93.768% | 80.580% | 93.768% |
| 4 | 90.580% | 82.464% | 81.304% | 90.580% | 77.536% | 90.580% |
| ResNet-TL | 1 | 93.043% | 67.391% | 56.957% | 93.043% | 51.449% | 93.043% |
| 2 | 92.464% | 78.406% | 71.449% | 92.464% | 68.261% | 92.464% |
| 3 | 93.333% | 81.884% | 76.522% | 93.333% | 75.652% | 93.333% |
| 4 | 91.014% | 83.333% | 79.130% | 91.014% | 79.710% | 91.014% |
| ResNet-FT-CNN | 1 | 96.377% | 68.116% | 56.377% | 96.377% | 50.580% | 96.377% |
| 2 | 92.754% | 79.710% | 75.507% | 92.754% | 73.043% | 92.754% |
| 3 | 93.188% | 80.870% | 76.957% | 93.188% | 74.638% | 93.188% |
| 4 | 85.797% | 76.957% | 75.217% | 85.797% | 73.188% | 85.797% |
| ResNet-TL-CNN | 1 | 93.768% | 70.580% | 63.333% | 93.768% | 57.536% | 93.768% |
| 2 | 93.478% | 83.043% | 80.435% | 93.478% | 79.130% | 93.478% |
| 3 | 94.203% | 84.928% | 82.174% | 94.203% | 80.725% | 94.203% |
| 4 | 92.029% | 84.203% | 81.449% | 92.174% | 80.000% | 92.029% |

**Table 2\_1.** Adversarial Attack performance of surrogate and SEVIT models (with MLP alternatives) on attack samples via Majority Voting, where these attack samples are generated on the surrogate model through model extraction attack

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **m** | **Clean Samples** | **FGSM** | **PGD** | **BIM** | **AutoPGD** | **C&W** |
| CNN Surrogate | **-** | **90.14%** | **90.72%** | **90.58%** | **91.45%** | **90.58%** | **91.45%** |
| MLP | 1 | 94.203% | 91.159% | 91.304% | 94.203% | 92.174% | 94.203% |
| 2 | 96.522% | 94.928% | 95.072% | 96.522% | 95.217% | 96.522% |
| 3 | 95.362% | 92.899% | 93.188% | 95.362% | 94.348% | 95.362% |
| 4 | 95.797% | 94.348% | 94.493% | 95.797% | 94.638% | 95.797% |
| CNN | 1 | 93.043% | 90.870% | 90.870% | 93.043% | 91.594% | 93.043% |
| 2 | 96.232% | 94.058% | 94.493% | 96.232% | 95.072% | 96.232% |
| 3 | 96.232% | 94.348% | 94.783% | 96.232% | 95.362% | 96.232% |
| 4 | 94.058% | 92.899% | 93.478% | 94.058% | 93.188% | 94.058% |
| ResNet-FT | 1 | 95.362% | 92.899% | 92.754% | 95.362% | 93.043% | 95.362% |
| 2 | 93.913% | 95.072% | 95.072% | 93.913% | 95.072% | 93.913% |
| 3 | 93.768% | 94.783% | 94.928% | 93.768% | 94.783% | 93.768% |
| 4 | 90.580% | 94.783% | 94.638% | 90.580% | 93.478% | 90.580% |
| ResNet-TL | 1 | 93.043% | 90.145% | 90.000% | 93.043% | 91.304% | 93.043% |
| 2 | 92.464% | 90.725% | 91.159% | 92.464% | 91.304% | 92.464% |
| 3 | 93.333% | 92.029% | 92.174% | 93.333% | 92.029% | 93.333% |
| 4 | 91.014% | 89.130% | 89.710% | 91.014% | 92.029% | 91.014% |
| ResNet-FT-CNN | 1 | 96.377% | 93.768% | 93.623% | 96.377% | 94.638% | 96.377% |
| 2 | 92.754% | 93.768% | 94.058% | 92.754% | 94.058% | 92.754% |
| 3 | 93.188% | 94.203% | 94.203% | 93.188% | 94.348% | 93.188% |
| 4 | 85.797% | 90.725% | 90.435% | 85.797% | 89.710% | 85.797% |
| ResNet-TL-CNN | 1 | 93.768% | 90.725% | 90.290% | 93.768% | 91.739% | 93.768% |
| 2 | 93.478% | 92.029% | 92.029% | 93.478% | 92.029% | 93.478% |
| 3 | 94.203% | 92.174% | 92.319% | 94.203% | 92.464% | 94.203% |
| 4 | 92.029% | 91.594% | 92.029% | 92.029% | 91.884% | 92.029% |

**------------------------- AFTER ADVERSARIAL TRAINING-------------------------------**

**Table 4.** Clean Performance of MLP alternatives ensemble modules for SEVIT (VIT + Ensemble modules) after adversarial training. Here abbreviations are as, MW: Majority Voting, A: Averaging, and WA: Weighted Averaging

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **# Trainable Params** | **Ensemble Criteria** | **m = 1** | **m = 2** | **m = 3** | **m = 4** | **m = 5** |
| MLP | 625219970 | MV | 91.449% | 95.507% | 95.652% | 95.507% | 95.507% |
| A | 94.493% | 95.942% | 96.087% | 96.377% | 95.362% |
| WA | 94.783% | 93.768% | 95.217% | 95.652% | 95.507% |
| CNN | 1032770 | MV | 95.507% | 95.652% | 96.087% | 94.203% | 94.348% |
| A | 96.667% | 96.957% | 96.377% | 94.928% | 93.188% |
| WA | 96.522% | 96.667% | 96.087% | 94.928% | 92.174% |
| ResNet-FT | 13586114 | MV | 96.087% | 96.377% | 96.232% | 91.449% | 91.449% |
| A | 96.812% | 96.957% | 95.652% | 93.043% | 89.420% |
| WA | 96.667% | 96.667% | 95.072% | 92.754% | 88.261% |
| ResNet-TL | 2409602 | MV | 91.594% | 90.290% | 93.333% | 91.449% | 91.594% |
| A | 96.377% | 94.348% | 94.638% | 93.333% | 91.739% |
| WA | 93.768% | 92.609% | 94.058% | 91.739% | 90.000% |
| ResNet-FT-CNN | 14267810 | MV | 95.942% | 95.942% | 96.522% | 88.986% | 89.130% |
| A | 96.957% | 96.377% | 94.493% | 93.043% | 91.014% |
| WA | 97.101% | 95.072% | 94.348% | 92.174% | 90.725% |
| ResNet-TL-CNN | 3091298 | MV | 94.203% | 95.362% | 95.217% | 93.333% | 93.478% |
| A | 96.087% | 96.087% | 95.362% | 94.638% | 93.043% |
| WA | 95.942% | 95.797% | 94.783% | 93.913% | 91.739% |

**Table 5.** After adversarial training, Adversarial Attack Performance of MLP alternatives ensemble modules for SEVIT (VIT + Ensemble modules) via Majority Voting. Here FGSM, PGD, BIM, and AutoPGD attack samples are having perturbation budget equal to 0.03

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **m** | **Clean Samples** | **FGSM** | **PGD** | **BIM** | **AutoPGD** | **C&W** |
| MLP | 1 | 91.449% | 70.145% | 62.609% | 91.449% | 58.696% | 91.449% |
| 2 | 95.507% | 84.058% | 81.014% | 95.507% | 80.870% | 95.507% |
| 3 | 95.652% | 86.377% | 84.493% | 95.652% | 84.058% | 95.652% |
| 4 | 95.507% | 91.739% | 91.159% | 95.507% | 90.725% | 95.507% |
| CNN | 1 | 95.507% | 71.304% | 63.768% | 95.507% | 58.841% | 95.507% |
| 2 | 95.652% | 87.826% | 86.087% | 95.652% | 87.536% | 95.652% |
| 3 | 96.087% | 89.420% | 88.696% | 96.087% | 89.710% | 96.087% |
| 4 | 94.203% | 90.580% | 90.000% | 94.203% | 90.000% | 94.203% |
| ResNet-FT | 1 | 96.087% | 70.435% | 62.029% | 96.087% | 56.812% | 96.087% |
| 2 | 96.377% | 86.812% | 80.435% | 96.377% | 79.130% | 96.377% |
| 3 | 96.232% | 87.826% | 81.159% | 96.232% | 80.145% | 96.232% |
| 4 | 91.449% | 81.594% | 77.391% | 91.449% | 74.638% | 91.449% |
| ResNet-TL | 1 | 91.594% | 70.580% | 62.319% | 91.594% | 56.377% | 91.594% |
| 2 | 90.290% | 82.029% | 78.116% | 90.290% | 75.942% | 90.290% |
| 3 | 93.333% | 84.348% | 80.580% | 93.333% | 78.841% | 93.333% |
| 4 | 91.449% | 86.087% | 81.304% | 91.449% | 81.014% | 91.449% |
| ResNet-FT-CNN | 1 | 95.942% | 70.000% | 61.739% | 95.942% | 56.957% | 95.942% |
| 2 | 95.942% | 82.754% | 76.087% | 95.942% | 71.159% | 95.942% |
| 3 | 96.522% | 86.667% | 80.580% | 96.522% | 75.652% | 96.522% |
| 4 | 88.986% | 80.725% | 77.971% | 88.986% | 75.217% | 88.986% |
| ResNet-TL-CNN | 1 | 94.203% | 70.145% | 61.594% | 94.203% | 56.522% | 94.203% |
| 2 | 95.362% | 86.232% | 80.870% | 95.362% | 80.580% | 95.362% |
| 3 | 95.217% | 87.391% | 83.333% | 95.217% | 83.333% | 95.217% |
| 4 | 93.333% | 86.667% | 84.638% | 93.333% | 83.623% | 93.333% |

**Table 5\_1.** Adversarial Attack performance of surrogate and SEVIT models after adversarial training (with MLP alternatives) on attack samples via Majority Voting, where these attack samples are generated on the surrogate model through model extraction attack.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ensemble Model** | **m** | **Clean Samples** | **FGSM** | **PGD** | **BIM** | **AutoPGD** | **C&W** |
| CNN Surrogate | **-** | **90.71%** | **90.29%** | **90.00%** | **92.03%** | **90.43** | **92.03%** |
| MLP | 1 | 91.449% | 89.420% | 89.275% | 91.449% | 90.145% | 91.449% |
| 2 | 95.507% | 93.623% | 93.768% | 95.507% | 94.203% | 95.507% |
| 3 | 95.652% | 93.768% | 93.913% | 95.652% | 94.348% | 95.652% |
| 4 | 95.507% | 94.493% | 94.638% | 95.507% | 94.783% | 95.507% |
| CNN | 1 | 95.507% | 92.464% | 92.464% | 95.507% | 93.188% | 95.507% |
| 2 | 95.652% | 94.203% | 94.493% | 95.652% | 94.638% | 95.652% |
| 3 | 96.087% | 94.928% | 95.072% | 96.087% | 95.362% | 96.087% |
| 4 | 94.203% | 94.348% | 94.493% | 94.203% | 94.348% | 94.203% |
| ResNet-FT | 1 | 96.087% | 93.043% | 92.899% | 96.087% | 94.058% | 96.087% |
| 2 | 96.377% | 95.507% | 95.507% | 96.377% | 95.652% | 96.377% |
| 3 | 96.232% | 95.072% | 95.072% | 96.232% | 95.362% | 96.232% |
| 4 | 91.449% | 94.058% | 93.333% | 91.449% | 93.188% | 91.449% |
| ResNet-TL | 1 | 91.594% | 89.420% | 89.565% | 91.594% | 90.145% | 91.594% |
| 2 | 90.290% | 88.841% | 88.986% | 90.290% | 89.130% | 90.290% |
| 3 | 93.333% | 91.739% | 92.174% | 93.333% | 91.884% | 93.333% |
| 4 | 91.449% | 90.290% | 90.000% | 91.449% | 90.290% | 91.449% |
| ResNet-FT-CNN | 1 | 95.942% | 93.188% | 93.188% | 95.942% | 93.913% | 95.942% |
| 2 | 95.942% | 95.652% | 95.942% | 95.942% | 95.652% | 95.942% |
| 3 | 96.522% | 95.362% | 95.652% | 96.522% | 96.087% | 96.522% |
| 4 | 88.986% | 93.043% | 92.754% | 88.986% | 91.884% | 88.986% |
| ResNet-TL-CNN | 1 | 94.203% | 92.319% | 92.319% | 94.203% | 92.899% | 94.203% |
| 2 | 95.362% | 93.768% | 93.768% | 95.362% | 93.768% | 95.362% |
| 3 | 95.217% | 93.913% | 94.203% | 95.217% | 94.493% | 95.217% |
| 4 | 93.333% | 91.884% | 92.174% | 93.333% | 91.884% | 93.333% |