URSI AP-RASC, Sydney, Australia, 17 – 22 August 2025

Entropy-Based Bayesian Model Averaging for Enhancing Classification Resilience Against Adversarial Perturbations Using SAR Images

Amir Hosein Oveis[(1)], Bhaskar Ahuja[(1)(2)], Alessandro Cantelli-Forti[(1)], and Marco Martorella[(1)(3)]
(1) RaSS Laboratory– CNIT, Pisa, Italy.
(2) Department of Physics, University of Trento, Trento, Italy.
(3) Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, UK.

Abstract

Deep learning models for Synthetic Aperture Radar (SAR) image classification are highly vulnerable to adversarial perturbations, which can significantly compromise their reliability in critical applications. In this study, we investigate the impact of adversarial attacks on SAR image classification by generating imperceptible perturbations using the Fast Gradient Sign Method (FGSM). These perturbations are designed to deceive a pre-trained convolutional neural network (CNN) into making incorrect classifications. To mitigate this vulnerability, we employ Entropy-Based Bayesian Model Averaging (EB-BMA), which is an adaptive ensemble learning approach based on model uncertainty. We fine-tune multiple CNN architectures using transfer learning from ImageNet and apply EB-BMA to dynamically adjust model contributions based on prediction confidence. This adaptive weighting mechanism ensures that models with higher certainty contribute more to the final decision, while uncertain models are down-weighted. Our results demonstrate that EB-BMA enhances resilience against adversarial perturbations, as validated on the MSTAR dataset.

1 Introduction

In recent years, deep learning (DL) has achieved impressive success in classifying Synthetic Aperture Radar (SAR) images. SAR is a highly advanced remote sensing technology that enables high-resolution imaging under all-time all-weather conditions. However, despite its effectiveness, DL-based approaches are vulnerable to adversarial perturbations, and this poses a serious threat to the reliability of these models in critical applications.

Adversarial perturbations refer to small, carefully designed noise, added to input data to mislead neural networks into making incorrect predictions [1]. These perturbations can be so small that even domain experts struggle to detect them, but yet they remain highly effective in degrading model performance. A basic strategy to improve model robustness against such perturbations is adversarial training where adversarial examples are incorporated into the training process to help the model learn how to resist them [2]. A particularly concerning aspect of adversarial attacks is their transferability. Adversarial examples designed to deceive one model can often deceive others, even when they have different architectures or are trained on different datasets. This occurs because neural networks tend to learn similar feature representations, and this makes them susceptible to the same types of perturbations [2]. One of the seminal work in adversarial machine learning is the Fast Gradient Sign Method (FGSM) [3]. FGSM is a simple but an effective adversarial attack and it is known for its minimal computational overhead. The perturbation is computed in the direction of the gradient of the loss function with respect to the input and is scaled by a factor to regulate the attack's intensity.

In the field of statistics, Bayesian Model Averaging (BMA) is a well-established framework for enhancing predictive performance by considering model uncertainty and combining multiple models rather than relying on a single one. BMA assigns probabilistic weights to models based on their posterior probabilities given the data. This approach reduces the risk of adversarial perturbations exploiting the weaknesses of any single model and enhances overall robustness.

In this paper, we introduce an adversarial mitigation framework based on Entropy-Based Bayesian Model Averaging (EB-BMA) to enhance the robustness of SAR image classifiers against adversarial attacks. Our approach consists of training multiple convolutional neural network (CNN) architectures using transfer learning (TL) from ImageNet, followed by fine-tuning on our clean, unperturbed training dataset. When exposed to a perturbed test set generated using FGSM, specifically designed to deceive the original model, these TL-based models respond differently to adversarial examples. By using EB-BMA, we dynamically adjust each model's contribution based on its confidence level to ensure that more reliable predictions have a greater influence while uncertain predictions are down-weighted. This ensemble learning strategy significantly improves resilience against adversarial attacks. To validate our theoretical findings, we utilize the publicly available MSTAR dataset (Moving and Stationary Target Acquisition and Recognition) [4]. This dataset consists of X-band SAR images of military targets.

2 Methods and Materials

2.1 FGSM

The Fast Gradient Sign Method (FGSM) [3] is a widely used technique for generating adversarial examples using the gradient of the loss function with respect to the input image. Mathematically, the perturbed image generated using FGSM is defined as:

$$x' = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

where $\nabla_x J$ represents the gradient of the model's loss function J with respect to the input x, θ denotes the model parameters, and y is the true class label. The sign function determines the direction of the gradient, and the perturbation is scaled by a factor ε, which controls the intensity of the attack. The choice of ε is crucial, as it directly impacts the adversarial effectiveness. If ε is too small, the perturbation may be insufficient to fool the network. Conversely, if it is too large, the perturbations become more noticeable, increasing the likelihood of detection by human observers or adversarial defense mechanisms.

2.2 Proposed framework

In the proposed framework, K pre-trained models, $M_k \in \{M_1, M_2, ..., M_K\}$, are fine-tuned on the training set and tested on the perturbed D_{test} that is designed by the FGSM for the original model M_0. Each M_k produces N output scores, where N refers to the number of classes. For each model prediction, entropy is calculated as:

$$H_k = -\sum_{i=0}^{N-1} P(y_i \mid x', M_k) \log P(y_i \mid x', M_k) \quad k \in 1, 2, ..., K \tag{2}$$

where x' is the perturbed input, and $P(y_i \mid x', M_k)$ represents the Softmax probability assigned by model M_k to class y_i given the adversarial input x'. After computing the entropy values, the weights for each model are calculated as:

$$w_k = \frac{\exp(-H_k)}{\sum_{j=1}^{K} \exp(-H_j)} \quad k \in 1, 2, ..., K \tag{3}$$

to obtain a normalized set of weights. These weights ensure that models with lower entropy (i.e., more confident predictions) have a higher influence on the final classification. Finally, model predictions are aggregated class-wise using the computed weights:

$$P(y_i \mid x') = \sum_{k=1}^{K} w_k P(y_i \mid x', M_k), \quad \forall i \in 0, 1, ..., N-1 \tag{4}$$

which prioritizes more confident models while reducing the impact of uncertain predictions, thereby improving robustness against adversarial attacks. The procedure is described in Algorithm 1.

Algorithm 1 Adversarial Mitigation Using Entropy-Based Bayesian Model Averaging (EB-BMA)

Require: Pre-trained CNN models $\{M_1, M_2, ..., M_K\}$, clean training dataset D_{train}, test dataset D_{test}
Ensure: Robust classification predictions under adversarial attacks
Step 1: Fine-tune CNN Models
for each model $M_k \in \{M_1, M_2, ..., M_K\}$ **do**
    Fine-tune M_k on D_{train}
end for
Step 2: Generate Adversarial Test Set
Use FGSM (1), the original model and the test dataset
Step 3: Compute Model Entropies
for each model M_k **do**
    Compute entropy of predicted probabilities H_k using (2)
end for
Step 4: Compute Entropy-Based Weights (3)
Step 5: Compute Final Weighted Prediction using (4)
Step 6: Output Prediction
return $\arg\max P(y \mid x')$ as the final classification result.

3 Real Data Experiment

We conduct our real-data experiments using the well-established MSTAR dataset [4], which consists of SAR images of ten different military ground targets. The images were acquired using an X-band SAR sensor operating at 9.6 GHz in spotlight mode, with a 0.3m $\times$ 0.3m resolution and full 360° aspect angle coverage. Since the original image chips have varying pixel resolutions, we resize all images to 224 $\times$ 224 pixels to ensure consistency with transfer learning models. As an initial step, we train a simple CNN using the MSTAR training set. We then generate FGSM perturbations based on the trained network and apply them to the MSTAR test set. The confusion matrices, t-SNE feature visualizations, and classification reports, shown in Fig. 1, illustrate how performance degrades under an example perturbation level of $\varepsilon = 0.02$. The t-SNE analysis [5] provides further understanding into the feature separability before and after the attack.

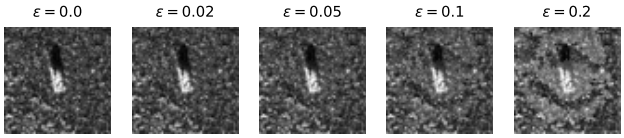
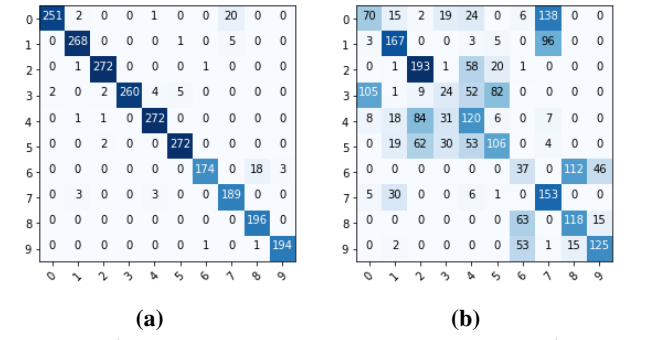
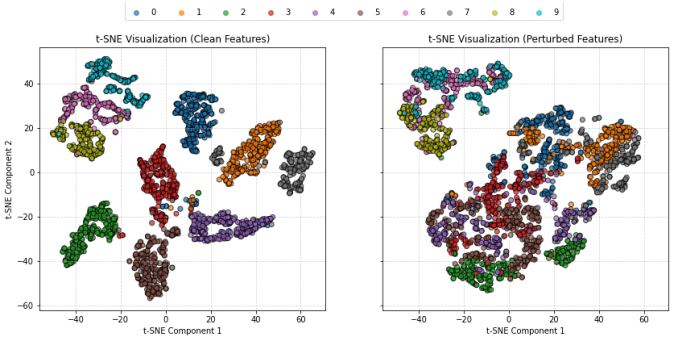


Figure 2. The effects of adding FGSM perturbations





(c)

Class	Without Attack Precision	Without Attack Recall	FGSM Perturbation Precision	FGSM Perturbation Recall
0	0.99	0.92	0.37	0.26
1	0.97	0.98	0.66	0.61
2	0.98	0.99	0.55	0.70
3	1.00	0.95	0.23	0.09
4	0.97	0.99	0.38	0.44
5	0.98	0.99	0.48	0.39
6	0.99	0.89	0.23	0.19
7	0.88	0.97	0.38	0.78
8	0.91	1.00	0.48	0.60
9	0.98	0.99	0.67	0.64
Accuracy	0.97		0.46	

(d)

Figure 1. Comparison of classification performance: (a) Confusion matrix (b) Confusion matrix under FGSM perturbation with $\varepsilon = 0.02$, (c) t-SNE visualization before and after adding the perturbations, (d) Classification report

As an example, Fig. 2 illustrates the effect of adding FGSM-generated perturbations to a SAR image. As the perturbation intensity ε increases, the adversarial noise becomes more noticeable. Notably, for $\varepsilon = 0$ and $\varepsilon = 0.02$, the image is correctly classified as class 0. However, when $\varepsilon \geq 0.05$, the perturbation becomes strong enough to deceive the neural network, causing it to misclassify the image as class 7 with high confidence scores.

Next, we utilize four well-known CNN architectures, namely MobileNet [6], VGG16 [7], ResNet-50 [8], and DenseNet-121 [9], all pre-trained on the ImageNet optical dataset, and evaluate them on the MSTAR test dataset

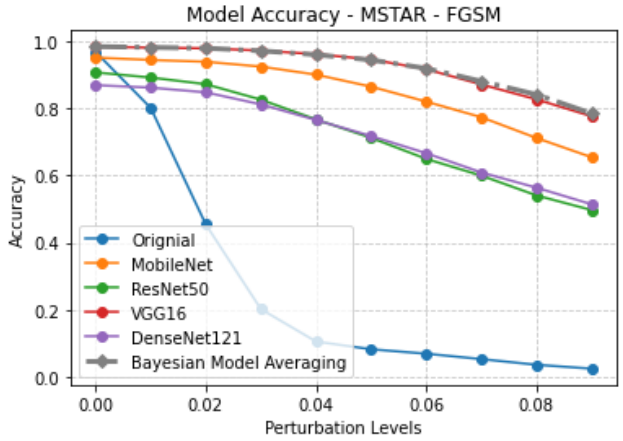


Figure 3. Overall accuracies of the "original" model, four transfer-learning models, and the proposed EB-BMA framework

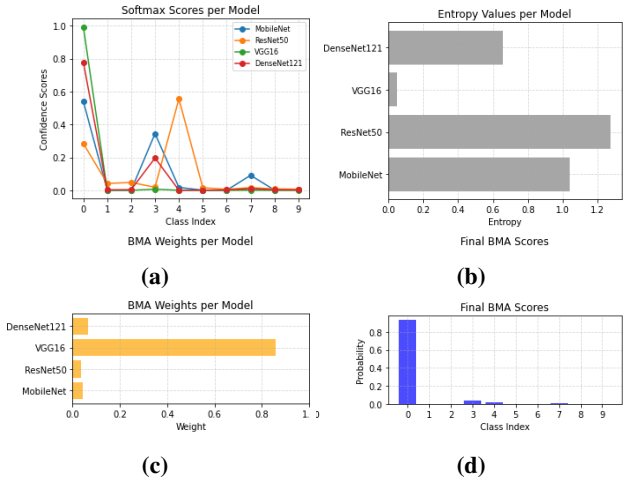


Figure 4. Given the first test image belonging to class 0: (a) Softmax scores per model, (b) Entropy values per model, (c) BMA weights per model, and (d) Final BMA scores

under various perturbation conditions. The results, presented in Fig. 3, indicate that while the adversarial attack is transferable to these networks, its severity is reduced. Notably, the "original" model (shown in blue) exhibits a significantly greater performance degradation compared to the other four models. This is expected, as the FGSM attack was specifically designed for the original model and not the transfer-learning networks. Among these four, VGG16 demonstrates superior robustness compared to other models.

To illustrate the method, consider the first image from the 2,425 test images under $\varepsilon = 0$ (i.e., no perturbation). Given that we have $N = 10$ classes and $K = 4$ models, the softmax outputs of the four models form a matrix $\hat{y} \in \mathbb{R}^{(4,10)}$, as depicted in Fig. 4(a). The corresponding entropy values for MobileNet, ResNet-50, VGG16, and DenseNet-121 are shown in Fig. 4(b). Following entropy normalization, the BMA weights w_k are computed by (3) and displayed in Fig.

4(c). Finally, for each class, the confidence scores from the four models are weighted (w_k) and summed based on (4) , and the final BMA probability distribution, as shown in Fig. 4(d), is produced. The overall accuracy of the BMA method across all perturbation levels is shown in Fig. 3 (gray curve), and it demonstrates its superior performance compared to all other methods.

4 Conclusion

In this paper, we proposed an entropy-based Bayesian Model Averaging (EB-BMA) framework to enhance the robustness of deep learning classifiers for SAR image classification against adversarial perturbations. Unlike traditional single-model classifiers, which are highly vulnerable to adversarial attacks, our method utilizes multiple pre-trained CNN architectures and adaptively adjusts their contributions based on their confidence levels. A key distinction between our dynamic entropy-based weighting and static weight averaging is that static methods assign fixed importance to each model, regardless of their reliability on a given input. This rigid approach does not account for variations in model confidence and can lead to suboptimal decisions when some models are uncertain. In contrast, EB-BMA dynamically reweights model contributions based on their entropy, and it ensures that more confident models insert a greater influence while it reduces the impact of uncertain predictions. This adaptive strategy improves classification accuracy and robustness, particularly in adversarial settings where model uncertainty varies significantly across different samples. In future work, we plan to explore other dynamic ensemble learning approaches with more advanced adversarial attacks.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[2] Y. Xu *et al.*, "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1604–1617, 2021.

[3] I. J. Goodfellow, J. J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[4] "Mstar public targets dataset," [online] Available: https://www.sdms.afrl.af.mil/index.php?collection=mstar, Sep 1995.

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[6] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint: 1704.04861*, 2017.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] H. Kaiming *et al.*, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.