Classification of Ground Targets from Radar Images Using Hybrid CNN-ViT and Few-Shot Learning

Mulubrhan Abebe Nerea*
*University West, Sweden
mulubrhan-abebe.nerea@student.hv.se

***Abstract*—This project investigates the classification of ground targets using radar imagery, with a particular focus on Synthetic Aperture Radar (SAR) data. We explore deep learning methods under the constraints of reduced training data. A hybrid model combining Convolutional Neural Networks (CNN) and Vision Transformer (ViT) is proposed and evaluated, alongside traditional architectures and a few-shot learning approach using Siamese Networks. The Hybrid CNN-ViT model achieved a peak accuracy of 99.89% on the MSTAR dataset. The results demonstrate the hybrid model's superiority and the effectiveness of Siamese learning in limited data scenarios, offering promising solutions for real-time defense surveillance applications.**

***Index Terms*—Radar Image Classification, SAR, CNN, Vision Transformer, Few-Shot Learning, Hybrid Model, Deep Learning, MSTAR**

Admin Part

The source code and implementation details can be accessed on GitHub: https://github.com/Mulubrhan21/Radar_Target_Classification_mstar.

I. Introduction

Radar-based classification of ground targets is crucial for situational awareness in combat scenarios [1]. Optical systems face challenges under poor weather and low-light conditions, whereas radar offers reliable all-weather imaging. The increased use of Synthetic Aperture Radar (SAR) imagery has made automatic target recognition (ATR) using deep learning a key focus [2]. This project tackles the challenge of achieving high classification accuracy with limited training data, common in SAR ATR due to its data scarcity and the inherent noise in radar images.

Ground target classification with SAR is vital for defense surveillance. Unlike optical sensors, which depend on daylight, SAR can capture high-resolution images 24/7, making it ideal for ATR tasks in military and security domains. SAR's ability to penetrate clouds, operate at night, and perform in adverse weather gives it significant advantages over optical imagery [3]. However, SAR images often suffer from graininess and speckle noise, necessitating advanced algorithms for accurate interpretation.

Recent research has also explored SAR for remote sensing, such as flood mapping and agricultural monitoring. Zhan et al. (2021) developed a method for rice mapping using SAR time series [4], and Landuyt et al. (2019) assessed SAR-based flood mapping approaches [1]. These studies highlight SAR's versatility, but also underscore the challenges in accurate target classification due to data complexity and noise.

In SAR ATR, Convolutional Neural Networks (CNNs) have been effective in image classification by capturing fine-grained textures. [2] revolutionized classification with deep CNNs, but they require large datasets, which are often unavailable in SAR ATR. This has led to hybrid models that combine CNNs for local feature extraction with Vision Transformers (ViTs) for global context modeling.

II. Literature Review

Automatic Target Recognition (ATR) refers to classifying objects from sensor imagery [3]. In radar, particularly Synthetic Aperture Radar (SAR), ATR is critical for defense and surveillance. SAR's ability to generate high-resolution imagery regardless of weather or lighting conditions makes the capability is invaluable in military applications where optical sensors often fail. SAR ATR also supports robustness in adverse environments and potential for autonomous decision-making have made SAR ATR a long-standing area of research [5].

A. CNN-Based Target Classification in SAR Imagery

Convolutional Neural Networks (CNNs) have become the cornerstone of SAR ATR due to their strong feature extraction capabilities. Initially, off-the-shelf models like AlexNet, VGG, ResNet, and DenseNet were fine-tuned on SAR datasets such as MSTAR, achieving up to 95% accuracy [6]. Even lightweight networks like MobileNetV3 have performed well under data-limited conditions, offering real-time inference and compact model size [7]. However, standard CNNs are not fully optimized for SAR's unique characteristics (e.g., speckle noise, azimuth variance). Overall, CNNs consistently achieve 92–98% accuracy and remain a strong baseline for SAR ATR applications.

B. Vision Transformers in Image Classification and SAR ATR

Vision Transformers (ViTs) [8] have emerged as a compelling alternative to CNNs in image classification by leveraging self-attention mechanisms to capture long-range dependencies and global context. Unlike CNNs, which rely on local receptive fields, ViTs operate on image patches, allowing them to model relationships across the entire image. With sufficient pre-training, ViTs have achieved performance

comparable to or surpassing CNNs such as ResNet [9] on large-scale benchmarks like ImageNet.

However, SAR ATR tasks typically involve small datasets such as MSTAR, which limits the effectiveness of standard ViTs due to their data-hungry nature. To address this, several lightweight or adapted ViT architectures have been proposed. In particular, the SS-ViT model achieved 94.72% accuracy on the MSTAR dataset, outperforming ResNet50 in the same benchmark setting [6]. These results highlight ViTs' potential for SAR ATR, but their performance often requires architectural modifications or hybridization with CNNs due to limited SAR datasets.

C. Hybrid CNN-ViT Models for SAR ATR

Hybrid CNN-ViT models combine the strengths of CNNs for local feature extraction and ViTs for global context, improving classification accuracy in SAR ATR. CNNs capture fine-grained textures, while ViTs offer global attention. Hybrid models, such as those proposed in [10] and [11], have shown promising results in similar domains such as PolSAR and wetland classification, particularly with limited data. While direct applications to MSTAR are rare.

Based on these advancements, we propose a CNN-ViT hybrid model for SAR classfication on MSTAR, leveraging both CNN and ViT strengths to enhance accuracy, especially with limited training data. Our approach combines CNNs for local feature extraction and ViTs for global context, leveraging the strengths of both to improve SAR target recognition, particularly on small datasets like MSTAR.

III. Requirement and Data Analysis

A. Dataset

We evaluated our models using the standard Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [12], a widely used public benchmark for synthetic aperture radar (SAR) target classification. It contains X-band SAR image chips of various military ground targets, collected at different depression angles and target orientations.

Our selected subset includes eight distinct vehicle classes: *2S1* (self-propelled artillery), *BRDM_2* (armored scout car), *BTR_60* (armored personnel carrier), *D7* (bulldozer), *SLICY* (synthetic calibration target), *T62* (main battle tank), *ZIL131* (cargo truck), and *ZSU_23_4* (anti-aircraft system). These classes represent a realistic mix of combat, support, and calibration vehicles, offering a challenging classification problem.

In total, the dataset includes 9,466 grayscale SAR images, with approximately 7,572 used for training and 947 each for validation and testing across the 8 classes. Other MSTAR variants [13], may contain 10 classes, including additional targets or viewing angles. Our subset focuses on the well-labeled and balanced 8-class configuration suitable for SAR-based automatic target recognition (ATR).

As shown in Fig. 1, Each sub-image (128×128 pixels) corresponds to a different vehicle type (as labeled). In these SAR images, the targets appear as bright white blobs or shapes against a dark background. The brightness indicates strong

Fig. 1. MSTAR dataset

radar reflection off parts of the vehicle (such as metallic surfaces and corners). As shown in Fig. 2, illustrates the class

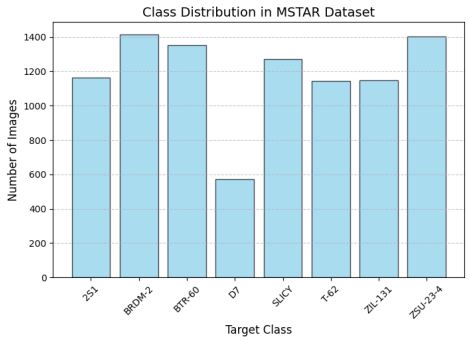

Fig. 2. Class Distribution in MSTAR Dataset

distribution in the chosen MSTAR dataset split, showing the number of image samples per target class. We observe a mild class imbalance: several classes (D7, ZIL131, 2S1, ZSU23/4, BRDM2, T62) have 572–573 images each (combining both depression angles), whereas others like T72, BMP2, BTR70 have 428–433 images each, and BTR60 has 451. This imbalance stems from the original data collection—certain vehicles had more variants or were imaged more frequently.

B. Preprocessing

All SAR image chips were single-channel grayscale by nature; we normalized the pixel values and applied a uniform resizing depending on the model input requirements. For CNN-based models, the images were resized to 128×128 (matching the original resolution). For the Vision Transformer (ViT) model (and the CNN-ViT hybrid), images were resized to 224×224, since the ViT architecture we employed expects 224×224 inputs (a standard size for ViT pre-trained on ImageNet). Data augmentation included rotation, zoom, shift, brightness variation, and flipping. After preprocessing, the dataset was split into training, validation, and test sets.

C. Training Setup

All models were trained on Google Colab with a Tesla T4 GPU. We used Adam optimizer with a learning rate scheduler and mixed precision training.

D. Specification Requirements:

Based on the problem scope, and to guide model selection and design, we defined the following specifications:

- **Accuracy:** The primary metric was classification accuracy. We aimed for test accuracy above 95%, in line

Fig. 3. System pipeline for radar target classification using CNN and ViT.

with state-of-the-art benchmarks for SAR classification on MSTAR.

- **Robust Generalization:** The model should remain effective under slight variations in input (like aspect angle, speckle noise).
- **Inference Speed:** The model should be capable of real-time inference on GPUs (ideally under 100 ms/image), enabling use in time-critical applications. Lightweight models like MobileNetV2 were evaluated for this purpose.

In summary, the requirements phase defined performance and deployment goals appropriate for SAR-based ground target recognition.

IV. System Engineering

Based on the identified requirements, we designed a modular and scalable radar target classification system. The architecture emphasizes robust feature extraction from SAR imagery using both convolutional and transformer-based components, optimized for small datasets and low-shot learning scenarios.

The proposed pipeline consists of four stages: image input, preprocessing, feature extraction (CNN + ViT), and a classification head. ResNet50 is used to extract local spatial features, while the Vision Transformer (ViT) captures global contextual information. Challenges such as class imbalance and high computational demands are mitigated through data augmentation, mixed precision training, and a carefully designed feature fusion mechanism.

Although Fig. 3 shows a standard classification pipeline, we extend this design further to support pairwise similarity learning. Specifically, we embed the hybrid CNN-ViT model within a Siamese network, allowing the system to compare two SAR images and predict whether they belong to the same target class or not. This extension supports low-data generalization and few-shot recognition tasks.

The full Siamese-based architecture is illustrated in Fig. 4. It processes a pair of radar image inputs through two identical hybrid branches (sharing weights), and extracts their embeddings. These embeddings are then compared using an absolute difference function, and a final fully connected layer with sigmoid activation predicts the similarity score.

Internally, the hybrid CNN-ViT backbone used in each branch of the Siamese model is detailed in Fig. 5. This component combines CNN-based spatial feature extraction and ViT-based global attention modeling, and outputs a fused feature vector that captures both local and global radar signature cues. This hybrid structure improves robustness and adaptability under data constraints, and supports modular experimentation with other backbones in the future.

The overall pipeline follows a modular sequence:

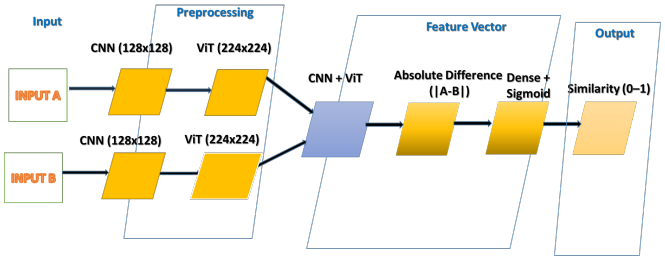

Fig. 4. Siamese architecture for similarity learning using shared hybrid CNN-ViT backbones.

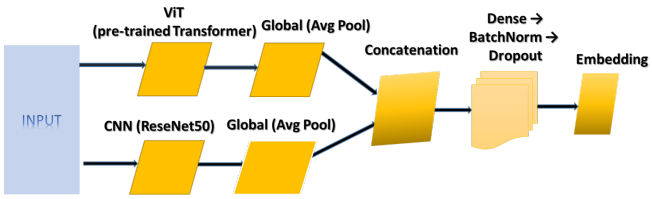

Fig. 5. Internal design of the hybrid CNN-ViT feature extractor.

(1) **Input:** A raw SAR image chip (typically 128×128) is provided to the system.

(2) **Preprocessing:** Images are normalized and resized depending on model requirements (224×224 for ViT models). Since MSTAR images are already well-calibrated, this step mainly involves resizing and normalization.

(3) **Feature Extraction:** Our hybrid model combines a ResNet50 backbone (for local texture and edge detection) and a Vision Transformer (for capturing long-range spatial context). The CNN processes spatial features, while ViT encodes global relationships by treating the image as a patch sequence. Their outputs are fused to form a comprehensive representation.

(4) **Classification:** The fused features are passed through a fully connected classifier (with softmax) that assigns a label to the input. The highest scoring class is selected as the final prediction.

Internally, CNN and ViT modules operate in parallel, with a dedicated fusion layer ensuring output compatibility. The design is modular—components can be swapped or fine-tuned independently, allowing scalability and future adaptation.

During system design, several challenges were addressed to ensure robust SAR target recognition. Due to the limited size of the MSTAR dataset, we applied data augmentation, transfer learning, and explored lightweight models like MobileNet for comparison. Class imbalance was mitigated through stratified sampling and targeted augmentation.

The hybrid CNN-ViT model, while powerful, introduced computational challenges, requiring mixed precision and gradient checkpointing for training optimization.

V. Algorithm Design using Spiral Approach

Following the system engineering phase, we adopted a spiral design methodology with two primary iterations. Each iteration involved evaluating model performance against the

defined specifications and improving the training strategy based on the outcome. Additionally, we explored few-shot learning via Siamese networks to support generalization in low-data scenarios.

A. Iteration 1: Baseline Models

Custom CNN (ResNet50-based): We constructed a convolutional neural network inspired by the ResNet50 architecture. Leveraging ImageNet pretraining, we removed the top classification layer and appended a dense head customized for 8-class classification on MSTAR. The model achieved a strong test accuracy of **97.4%**, confirming that deep convolutional features can effectively separate SAR signatures. While ResNet50 was efficient at inference and delivered high accuracy, it lacked the capacity to capture global contextual information inherent in SAR images.

MobileNetV2: To evaluate a lightweight model suitable for edge deployment, we trained a MobileNetV2 using transfer learning. Although fast and memory-efficient, the model achieved only **83.2%** accuracy, failing to meet our performance threshold. This highlights the trade-off between speed and representational power.

Key outcomes: This iteration established ResNet50 as a solid baseline. We also validated training routines involving early stopping, data augmentation (random rotations, flips, Gaussian noise), and one-cycle learning rate schedules.

B. Iteration 2: Optimized Hybrid CNN-ViT Architecture

To enhance performance, we developed a hybrid model that fused a ResNet50 CNN with a Vision Transformer (ViT-B/16). The CNN extracted local spatial features, while the ViT provided global context modeling. Their outputs were concatenated and passed through dense layers for final classification. A two-stage training strategy was employed: initially freezing the ViT, then fine-tuning its final layers. Mixed precision training, dropout, and batch normalization were used to stabilize learning.

This hybrid model achieved a test accuracy of **99.89%**, significantly outperforming both prior models. It also demonstrated balanced precision and recall across all classes. Despite a slightly higher computational load, the accuracy gains justify its use in high-stakes radar classification tasks.

C. Few-Shot Learning: Siamese Network with Hybrid Backbone

We further explored a few-shot recognition setup by employing a Siamese network architecture. The hybrid CNN-ViT model was used as a shared backbone, and the network was trained to distinguish whether two SAR images belong to the same class using contrastive loss.

The Siamese model achieved **95.5%** accuracy on unseen pairs. This result shows the viability of using hybrid embeddings for similarity-based learning, particularly when the number of training examples is low. It enables flexible deployment in scenarios with dynamic target sets or limited labels.

VI. Results

A. Model Comparison

As shown in Table I, summarizes the classification accuracies of all evaluated all models(Basline and Optimized) on the 8-class MSTAR dataset.

TABLE I
ACCURACY COMPARISON OF MODELS ON THE 8-CLASS MSTAR DATASET

Model	Test Accuracy
Custom CNN (ResNet)	97.43%
MobileNetV2 (Fine-Tuned)	83.21%
Vision Transformer (ViT)	99.26%
Hybrid CNN-ViT	**99.89%**
Siamese Hybrid (Pairwise)	95.51%

The custom CNN model, based on ResNet, achieved 97.43% accuracy, while MobileNetV2 delivered 83.21%, highlighting the trade-off between model size and accuracy. The Vision Transformer (ViT) model attained 99.89% accuracy, demonstrating its ability to generalize well on radar data.

As shown in Fig. 7 the hybrid CNN-ViT model reached the highest accuracy of 99.89%, combining local spatial features and global context through a two-stage training strategy and regularization.

We also explored a Siamese network using the hybrid backbone for pairwise similarity, achieving 95.51% accuracy. This model's primary goal was to evaluate the feasibility of few-shot radar target recognition, not to surpass the hybrid classifier.

B. Best Model Evaluation

To further assess the performance of the hybrid CNN-ViT model, we present the confusion matrix and classification report on the test set.

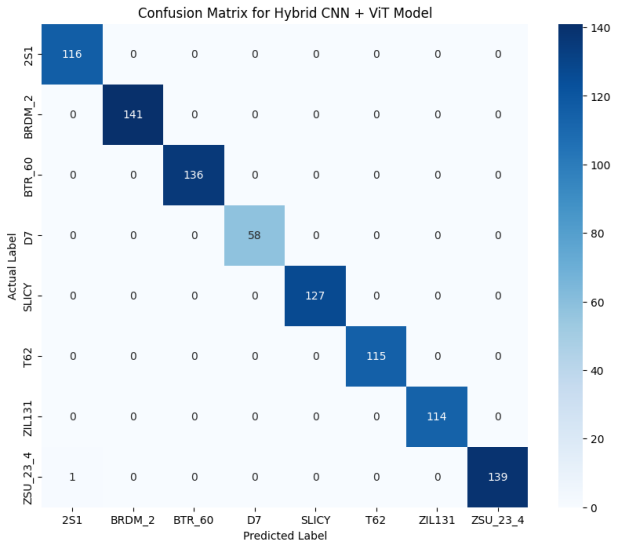

Fig. 6. Confusion Matrix for Hybrid CNN-ViT Model

As shown in Figure 6, the hybrid model correctly classified all classes with minimal confusion.

TABLE II
CLASSIFICATION REPORT FOR HYBRID CNN-VIT

Class	Precision	Recall	F1-score
2S1	0.99	1.00	1.00
BRDM_2	1.00	1.00	1.00
BTR_60	1.00	1.00	1.00
D7	1.00	1.00	1.00
SLICY	1.00	1.00	1.00
T62	1.00	1.00	1.00
ZIL131	1.00	1.00	1.00
ZSU_23_4	1.00	0.99	1.00
Avg / Total	1.00	1.00	1.00

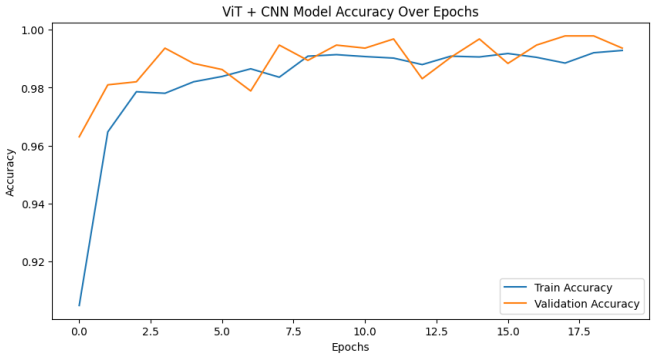

Fig. 7. Hybrid ViT + CNN Model Accuracy.

The metrics indicate strong generalization across all classes, with particularly high performance in precision and recall. Minor confusion was noted in class ZSU_23_4, though the F1-score remained high.

VII. CONCLUSION

In this study, we investigated deep learning models for ground target classification using SAR imagery, particularly focusing on the MSTAR dataset. Our hybrid CNN-ViT model achieved a peak accuracy of **99.89%**, outclassing both the baseline CNN (97.43%) and MobileNetV2 (83.21%). This demonstrates the power of combining CNN's local feature extraction with ViT's global context.

For low-data scenarios, the Siamese network with a shared hybrid backbone reached **95.5%** accuracy, showing promise for flexible deployment in situations with limited data or new targets.

To address challenges like data scarcity, class imbalance, and overfitting, we used transfer learning, data augmentation, and mixed precision training. Our approach supports future model optimizations for real-time deployment on constrained devices.

VIII. FUTURE CONSOLIDATION OF ATPS AND FUTURE PLAN:

The hybrid CNN-ViT model achieved 99.89% accuracy on the MSTAR dataset, demonstrating its effectiveness for SAR-based target recognition. This success was made possible by combining CNNs for local feature extraction and ViTs for global context, addressing challenges of limited data.

Future Work:

- Implement self-supervised contrastive learning to enhance few-shot learning.
- Apply model pruning and quantization for real-time edge deployment.
- Explore model compression techniques for efficient and scalable deployment.

These improvements aim to increase model efficiency and facilitate real-time deployment in defense and surveillance applications.

ACKNOWLEDGMENTS

We thank our supervisor and project owner, Amit Mishra, for guidance and support.

REFERENCES

[1] L. Landuyt, A. V. Wesemael, G. J.-P. Schumann, R. Hostache, N. E. C. Verhoest, and F. M. B. V. Coillie, "Flood mapping based on synthetic aperture radar: An assessment of established approaches," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 722–739, 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[3] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 9, pp. 4–22, 2022.

[4] P. Zhan, W. Zhu, and N. Li, "An automated rice mapping method based on flooding signals in synthetic aperture radar time series," *Remote Sensing of Environment*, vol. 252, p. 112112, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425720304855

[5] J. Yin, C. Duan, H. Wang, and J. Yang, "A review on the few-shot sar target recognition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–16, 2024, license: CC BY 4.0.

[6] J. Fein-Ashley, T. Ye, R. Kannan, V. Prasanna, and C. Busart, "Benchmarking deep learning classifiers for sar automatic target recognition," 2022.

[7] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," *arXiv preprint arXiv:2206.01191*, 2022. [Online]. Available: https://arxiv.org/abs/2206.01191

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[10] W. Wang, J. Wang, D. Quan, M. Yang, J. Sun, and B. Lu, "Polsar image classification via a multigranularity hybrid cnn-vit model with external tokens and cross-attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, no. 5, pp. 1464–1475, 2023.

[11] A. Radman, F. Mohammadimanesh, and M. Mahdianpari, "Wet-convit: A hybrid convolutional–transformer model for efficient wetland classification using satellite data," *Remote Sensing*, vol. 16, no. 14, p. 2673, 2024, this article belongs to the Special Issue Satellite-Based Climate Change and Sustainability Studies.

[12] A. Majumdar, "MSTAR Dataset on Kaggle," https://www.kaggle.com/datasets/atreyamajumdar/mstar-dataset-8-classes, 2022, [Online; accessed March 21, 2025].

[13] H. Wang, S. Chen, F. Xu, and Y.-Q. Jin, "Application of deep-learning algorithms to mstar data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*. IEEE, 2015, pp. 3743–3745.

EU AI Regulation Impact on Radar Image Classification

Radar-based classification models used in defense and surveillance may fall under the "High-Risk" category of the EU AI Act. This includes AI systems used in military applications, requiring adherence to specific regulations.

A. AI Safety Regulations

Implications for this project:

- **Explainability:** The model must justify its decisions, ensuring transparency in critical defense contexts.
- **Bias Mitigation:** The model should perform fairly across all target types, preventing biases in military or civilian vehicle classification.
- **Data Privacy & Security:** Sensitive radar data requires strong protection against unauthorized access.

B. Bias and Ethical Considerations

Bias in Radar Data: Limited or biased datasets, such as those collected from specific regions, may affect the model's ability to generalize.

Misclassification Risks: Misclassifications, such as confusing civilian vehicles with military targets, could have serious consequences.

Solution: Data augmentation and bias detection techniques should be used to mitigate these issues and ensure fairness.