FUSION OF FOUNDATION MODELS FOR IMBALANCED SAR SHIP CLASSIFICATION

Ch Muhammad Awais, Marco Reggiannini, Davide Moroni
Institute of Information Science and Technologies- ISTI
National Research Council of Italy (CNR), Pisa, Italy
{firstname.lastname}@isti.cnr.it

Oktay Karakus
Department of Computer Science,
Cardiff University, Cardiff, UK
o.karakus@cardiff.ac.uk

ABSTRACT

While foundation models pretrained on remote sensing data worked successfully in various advanced tasks, their susceptibility to class imbalance and the potential of targeted mitigation strategies remain underexplored. We present an empirical evaluation of six remote sensing foundation models on two SAR ship classification datasets: OpenSARShip and FUSARShip. We demonstrate that commonly used metrics such as accuracy and weighted F1-score mask poor performance on minority classes, and advocate for macro F1-score as a more reliable measure for imbalanced datasets. Our experiments reveal that several foundation models underperform ImageNet-pretrained baselines (ResNet, VGG, ViT). We investigate four fusion strategies to combine embeddings from multiple foundation models in both full and lightweight configurations, where lightweight variants exclude the two poorest-performing models. Lightweight late fusion with informed weighting achieves up to 5.5 percentage point improvements in macro F1 over the best individual foundation models, while reducing computational overhead.

1 INTRODUCTION

Synthetic Aperture Radar (SAR) imagery enables all-weather, day-night maritime monitoring, making it essential for ship detection and classification applications (Awais et al., 2025). However, applying deep learning to SAR ship classification faces two fundamental challenges. First, the domain gap between optical imagery used in ImageNet pretraining and SAR data limits transfer learning effectiveness. Second, operational SAR ship datasets exhibit severe class imbalance, with majority classes outnumbering minority classes by ratios exceeding 80:1. Foundation models pretrained on large-scale remote sensing data promise to address the domain gap, but their robustness under extreme class imbalance remains unexplored.

We benchmark six foundation models spanning three pretraining paradigms: models pretrained on optical and multispectral data (Prithvi (Jakubik et al., 2023), ScaleMAE (Reed et al., 2023)), models trained on mixed SAR and optical data (DOFA (Xiong et al., 2024), SSL4EO (Wang et al., 2023)), and models trained exclusively on SAR data (SARDet-100K Li et al. (2024b), SARJepa (Li et al., 2024a)). We evaluate these models on two public SAR ship classification datasets, OpenSARShip (subset of 6 classes) (Li et al., 2017) and FUSARShip (subset of 9 classes) (Hou et al., 2020), and compare their performance against ImageNet-pretrained baselines (ResNet, VGG (Mascarenhas & Agarwal, 2021), ViT (Yuan et al., 2021)). Our analysis reveals critical limitations in standard evaluation metrics and demonstrates that late fusion strategies combining complementary model embeddings can substantially improve performance over individual models. Our contributions are:

- We benchmark six foundation models and three ImageNet baselines on SAR ship classification, revealing that several foundation models underperform ImageNet-pretrained models when evaluated with macro F1-score.
- We demonstrate that accuracy and weighted F1-score are unreliable for imbalanced datasets and advocate for macro F1-score as the primary metric, supported by empirical evidence showing large discrepancies between metrics.
- We propose and evaluate four fusion strategies for combining foundation model embeddings, with late fusion methods achieving macro F1 improvements of 5.5% on FUSARShip and 5.4% on OpenSARShip over the best individual model.

- We release precomputed embeddings, fusion code, and Colab notebooks running on free-tier resources to enable lightweight, reproducible experimentation without requiring expensive model retraining. [omitted for blind review]

2 METHODOLOGY

Figure 1 illustrates our evaluation pipeline. We freeze pre-trained encoders to extract embeddings from two SAR ship classification datasets, then evaluate four fusion strategies to combine information from multiple foundation models. All training strategies use these pre-extracted embeddings, enabling efficient experimentation without repeated forward passes through frozen encoders.

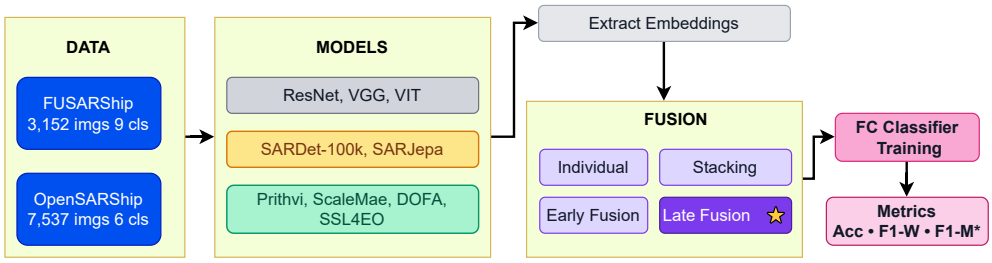

Figure 1: Evaluation pipeline for SAR ship classification using frozen foundation model encoders and fusion strategies.

Datasets and Preprocessing: We evaluate models on two SAR ship classification datasets. FUSARShip contains 3,152 images across 9 classes (Cargo, Fishing, Bulk, Tanker, Container, Dredger, Tug, GeneralCargo, Passenger) with a 46:1 majority-to-minority ratio (Cargo: 1,693 samples; Passenger: 37). OpenSARShip comprises 7,537 images across 6 classes (Cargo, Tanker, Dredging, Fishing, Passenger, Tug) with an 85:1 ratio (Cargo: 5,303; Tug: 62). All images are resized to 224×224 pixels and normalized to [0, 1]. Single-channel SAR images are duplicated to three channels for optical-pretrained models. Datasets are split 80/10/10 (train/val/test), stratified by class.

Models and Feature Extraction: We evaluate three ImageNet-pretrained baselines (VGG-16, ResNet-50, ViT-B/16) and six foundation models: Prithvi-100M (ViT-Base, optical multispectral, 768-dim), ScaleMAE (ViT-Large, optical, 1024-dim), DOFA (ViT-Base, mixed SAR/optical, 768-dim), SSL4EO-S12 (ViT-Small, Sentinel-2, 384-dim), SARDet-100K (ResNet-50, SAR detection, 2048-dim), and SAR-JEPA (ViT-Base, SAR JEPA, 768-dim). All encoders are frozen and penultimate-layer embeddings are pre-extracted to enable efficient experimentation.

Fusion Strategies. We investigate four methods: (1) *Early Fusion* concatenates feature embeddings from all models before feeding them to a single classifier; (2) *Stacking* employs a meta-classifier that learns to combine the class probability distributions (logits) produced by individual models; (3) *Late Fusion (Informed)* assigns each model a weight proportional to its validation macro F1 score, thereby giving stronger influence to better-performing models; (4) *Late Fusion (Grid Search)* determines optimal prediction weights through exhaustive search over the weights space. We additionally evaluate *Lite variants* that exclude the two poorest-performing models and fuse only the top four, reducing computational overhead while improving ensemble quality by eliminating weak predictors.

All classifiers are trained for 100 epochs using Adam optimizer with learning rate 0.0005 and batch size 64. We apply early stopping with patience of 90 epochs based on validation macro F1-score and load the best weights. We use a three-layer fully connected classifier with batch normalization and dropout (p=0.2) for all experiments. Models are evaluated using accuracy, weighted F1-score ($F1\text{-}W$), and macro F1-score ($F1\text{-}M$), with macro-F1 as the primary metric due to severe class imbalance.

3 RESULTS

Table 1 presents individual model performance. On FUSARShip, VGG-16 achieves the highest accuracy among ImageNet baselines (64.1%), while DOFA and SARDet-100K tie at 63.2%. However, macro F1 reveals a different picture: SARDet-100K achieves 47.7%, substantially outperforming VGG-16 (40.5%). This discrepancy demonstrates how accuracy masks poor minority class

performance. For instance, ResNet achieves 52.6% accuracy but only 7.7% macro F1, a 45% gap indicating near-zero performance on minority classes.

Table 1: Individual model performance on FUSARShip (9 classes) and OpenSARShip (6 classes).

	FUSARShip (9)			OpenSARShip (6)		
Model	**Accuracy (Acc) (%)**	**F1-M (%)**	**F1-W (%)**	**Acc (%)**	**F1-M (%)**	**F1-W (%)**
ResNet	52.6	7.7	47.0	70.2	13.8	64.2
VGG	64.1	40.5	69.2	71.3	19.9	68.2
ViT-224	53.3	22.5	60.4	73.7	27.0	72.9
Foundation Models						
DOFA	63.2	37.0	63.6	59.9	32.1	63.3
SSL4EO	60.1	42.4	60.4	56.7	31.7	61.9
ScaleMAE	48.0	34.4	49.5	50.2	24.2	55.9
Prithvi	48.0	33.2	50.2	35.0	20.1	44.3
SARDet-100K	63.2	**47.7**	62.9	68.9	**33.4**	69.4
SARJepa	60.7	41.4	62.3	52.2	27.0	58.9

On OpenSARShip, ImageNet models achieve over 70% accuracy (ViT-224: 73.7%), yet fail on macro F1 with scores below 27%. In contrast, three foundation models exceed 30% macro F1, with SARDet-100K reaching 33.4%. This confirms that accuracy and weighted F1 are unreliable metrics for imbalanced datasets, as they disproportionately reflect majority class performance.

Table 2: Fusion strategy performance. Late fusion substantially outperforms individual models (Table 1). Lite variants (using only top 4 models) reduce computation while maintaining performance.

	FUSARShip (9)			OpenSARShip (6)		
Fusion Method	**Acc (%)**	**F1-M (%)**	**F1-W (%)**	**Acc (%)**	**F1-M (%)**	**F1-W (%)**
Full Fusion (6 models)						
Early Fusion	61.3	42.8	61.0	65.3	35.2	67.6
Stacking	21.4	7.8	25.3	20.2	9.8	26.5
Late Fusion Informed	68.1	52.0	67.5	67.9	38.2	70.1
Late Fusion Grid Search	68.1	52.8	67.8	66.9	33.9	68.9
Lite Fusion (4 models: DOFA, SSL4EO, SARDet-100K, SARJEPA)						
Early Fusion	65.6	43.9	66.2	68.3	38.4	69.6
Stacking	8.0	5.8	9.7	26.7	12.2	35.9
Late Fusion Informed	**69.0**	**53.2**	**68.5**	**68.2**	**38.8**	**70.1**
Late Fusion Grid Search	67.5	48.2	66.9	67.5	34.8	69.2

Table 2 presents fusion results. Late fusion methods substantially outperform individual models. On FUSARShip, Late Fusion Informed (Lite) achieves 53.2% macro F1, a 5.5% improvement over the best individual model (SARDet-100K: 47.7%). On OpenSARShip, the same method reaches 38.8% macro F1, a 5.4% gain. Notably, lite variants maintain or improve performance while reducing computational cost by eliminating poorly performing models. Stacking performs poorly (7.8% and 9.8% macro F1), likely due to overfitting on validation data. Early Fusion provides modest gains (42.8% and 35.2%), while Late Fusion methods consistently achieve the highest scores by weighting models based on minority class performance.

4 Discussion

Our results reveal critical limitations in standard evaluation practices for imbalanced SAR classification. ImageNet baselines achieve high accuracy (70-73% on OpenSARShip) but catastrophically fail on minority classes, with ResNet and VGG scoring 0% F1 on four to five classes. This occurs because models overfit to the majority class (Cargo: 70% of OpenSARShip samples), achieving high overall accuracy while ignoring rare ship types. Accuracy and weighted F1 mask this failure by overweighting majority class performance. Macro F1, which treats all classes equally, reveals the true capability: ResNet achieves 52.6% accuracy but only 7.7% macro F1 on FUSARShip, a 45% discrepancy indicating near-complete minority class neglect.

Foundation models pretrained on SAR data (SARDet-100K, SARJepa) substantially outperform optically-pretrained models (Prithvi, ScaleMAE). SARDet-100K achieves non-zero F1 on all classes, including rare categories with fewer than 15 samples (Tug: 17.1% F1, Passenger: 13.8% F1 on OpenSARShip). In contrast, Prithvi achieves 0% F1 on Tug despite 35.0% overall accuracy. This demonstrates that domain-specific pretraining provides superior feature representations for minority classes. The poor performance of Prithvi may also reflect architectural mismatch: it expects 6-channel multispectral input but receives 3-channel duplicated SAR, potentially degrading learned features.

Late fusion methods achieve substantial improvements over individual models. On FUSARShip, Late Fusion Informed (Lite) reaches 53.2% macro F1, surpassing the best individual foundation model (SARDet-100K: 47.7%) by 5.5% and the best ImageNet model (VGG: 40.5%) by 12.7%. Similar gains occur on OpenSARShip (38.8% vs 33.4% and 27.0%). Fusion succeeds because it weighs models by validation macro F1, prioritizing those with better minority class performance. Lite variants maintain or improve performance while reducing computational cost by eliminating the two poorest-performing models, demonstrating that not all foundation models contribute positively to fusion. Stacking performs poorly (5.8-12.8% macro F1), falling below random guessing for FUSARShip. This likely results from meta-classifier overfitting on small validation sets (10% of 3,152 and 7,537 samples) and bias toward majority classes. The meta-classifier learns to weight models that achieve high accuracy rather than balanced performance, reinforcing majority class predictions. Alternative meta-classifier architectures or class-balanced training may address this failure.

While demonstrated on SAR ship classification, our fusion methodology is model-agnostic and applicable to any domain where: (1) multiple pretrained foundation models exist, (2) class imbalance is present, and (3) combining complementary representations may improve minority class performance. Such methodology could potentially find application in many scenarios, including imbalanced medical imaging, stock assessment and monitoring of biological species (including rare species), and minority demographic classification.

Future directions. Future work will examine adaptation of foundation models beyond frozen encoders with linear heads, including parameter-efficient and full fine-tuning, to quantify performance gains relative to the additional computational cost. We will also replace the current late-fusion member selection based on a single validation ranking with more systematic subset selection procedures to assess whether improved ensembles can be obtained. For early fusion, we will investigate alternative feature-combination mechanisms beyond concatenation, such as element-wise interactions and learned fusion modules (e.g., attention-based approaches). In addition, we will extend the study to larger SAR ship datasets to clarify how dataset scale influences the relative effectiveness of different pretrained encoders and fusion strategies. Finally, we will evaluate the generality of the proposed embedding-and-fusion pipeline on additional SAR tasks and, where applicable, other remote sensing domains to determine the extent to which the conclusions transfer across settings.

5 Conclusion

We benchmark six foundation models and three ImageNet baselines on two imbalanced SAR ship classification datasets. Our evaluation reveals three critical findings. First, accuracy and weighted F1 scores mask severe minority class failures; for instance, ViT-224 achieves 73.7% accuracy on OpenSARShip but only 27.0 % macro F1, demonstrating the necessity of macro F1 for imbalanced evaluation. Second, foundation model performance varies substantially: while the best foundation model (SARDet-100K) outperforms all ImageNet baselines on both datasets, several foundation models underperform ImageNet baselines, with optically-pretrained models (ScaleMAE, Prithvi) showing particularly weak performance. Third, lightweight late fusion with informed weighting improves macro F1 by 5.5 and 5.4 percentage points over the best individual model on FUSARShip and OpenSARShip respectively, while reducing computational cost by excluding the two poorest-performing models. These results underscore the importance of appropriate evaluation metrics for imbalanced data and demonstrate that strategic fusion of complementary foundation models can substantially enhance minority class recognition.

REFERENCES

Ch Muhammad Awais, Marco Reggiannini, Davide Moroni, and Emanuele Salerno. A survey on sar ship classification using deep learning. *arXiv preprint arXiv:2503.11906*, 2025.

Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63(4):140303, 2020.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Boying Li, Bin Liu, Lanqing Huang, Weiwei Guo, Zenghui Zhang, and Wenxian Yu. Opensarship 2.0: A large-volume dataset for deeper interpretation of ship targets in sentinel-1 imagery. In *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*, pp. 1–5. IEEE, 2017.

Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024a.

Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *Advances in Neural Information Processing Systems*, 37:128430–128461, 2024b.

Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pp. 96–99. IEEE, 2021.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *CoRR*, 2024.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.