Dan's SAR Project Strategy

Initial Exploratory Data Analysis (EDA) provided several key insights:

1. We first confirmed the dataset's validity, noting a balanced class distribution (approx. 52% Ship, 48% Iceberg). Critically, we performed a covariate shift analysis by overlaying the Train and Test distributions for the incidence angle, Band 1, and Band 2; all features were well-aligned, ruling out external data leakage between the partitions.
2. Mean backscatter analysis shows Ships are slightly brighter in Band 1 (HH) than Icebergs, requiring the model to rely on complex spatial features rather than simple intensity.
3. The most critical finding was that the discrete Incidence Angle Distribution shows a pronounced peak in the 37-40 range that is almost purely Iceberg, suggesting the data is not sparse and suffers from a statistical artifact. We identified this extreme purity as a data leak related to the limited collection geometry, which ultimately guided our decision to build a robust model rather than relying on this unstable pattern.

Given the data is provided in decibels, the image bands were normalized via Standard Scaling. The overall normalization strategy, augmentation strategy (e.g., flipping, rotation) and modal architecture choices were guided by the paper provided in the assignment.

Before beginning the main training, a small subset of the training data was set aside as a temporary validation set for immediate evaluation and comparison between model checkpoints. Due to the relatively small size of the dataset, the primary training was then conducted using a K-Fold Cross-Validation. strategy to ensure robust generalization and prevent overfitting to any single data split.

Further Improvement

1.**Adaptive Normalization and Augmentation:** We would move beyond fixed scaling by implementing learnable normalization layers (e.g., Batch Normalization) within the CNN, allowing the network to dynamically adapt the feature distributions. This adaptive approach will be coupled with SAR-specific augmentations to robustly expand the limited dataset.

2.Architecture Depth and Connectivity: We would test alternative and larger architectures, such as adapting a ResNet or Inception model for the small input. Incorporating skip connections is essential for building deeper, more performant networks without increasing the vanishing gradient problem.

3.Deep Dive in the Angle Problem in the EDA: The angle distribution would be further investigated using specialized EDA tools to precisely isolate the behavioral patterns within the core peak. This involves zooming in repeatedly to characterize exactly how the purity changes with minute angular variations, guiding feature engineering or targeted data-level interventions.