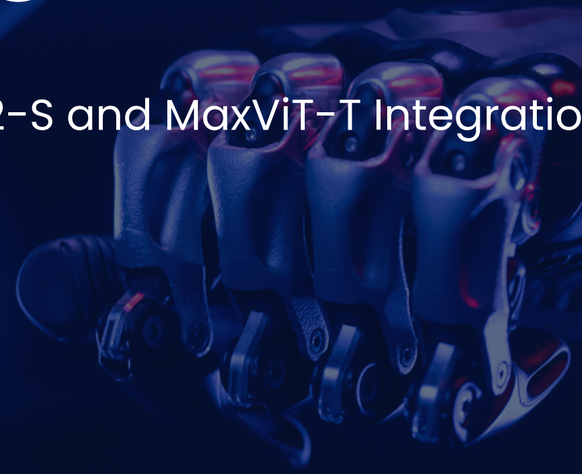
AI OF GOD 4.0

#1 Solution

3-Stage Vision Model Ensemble for Image Classification

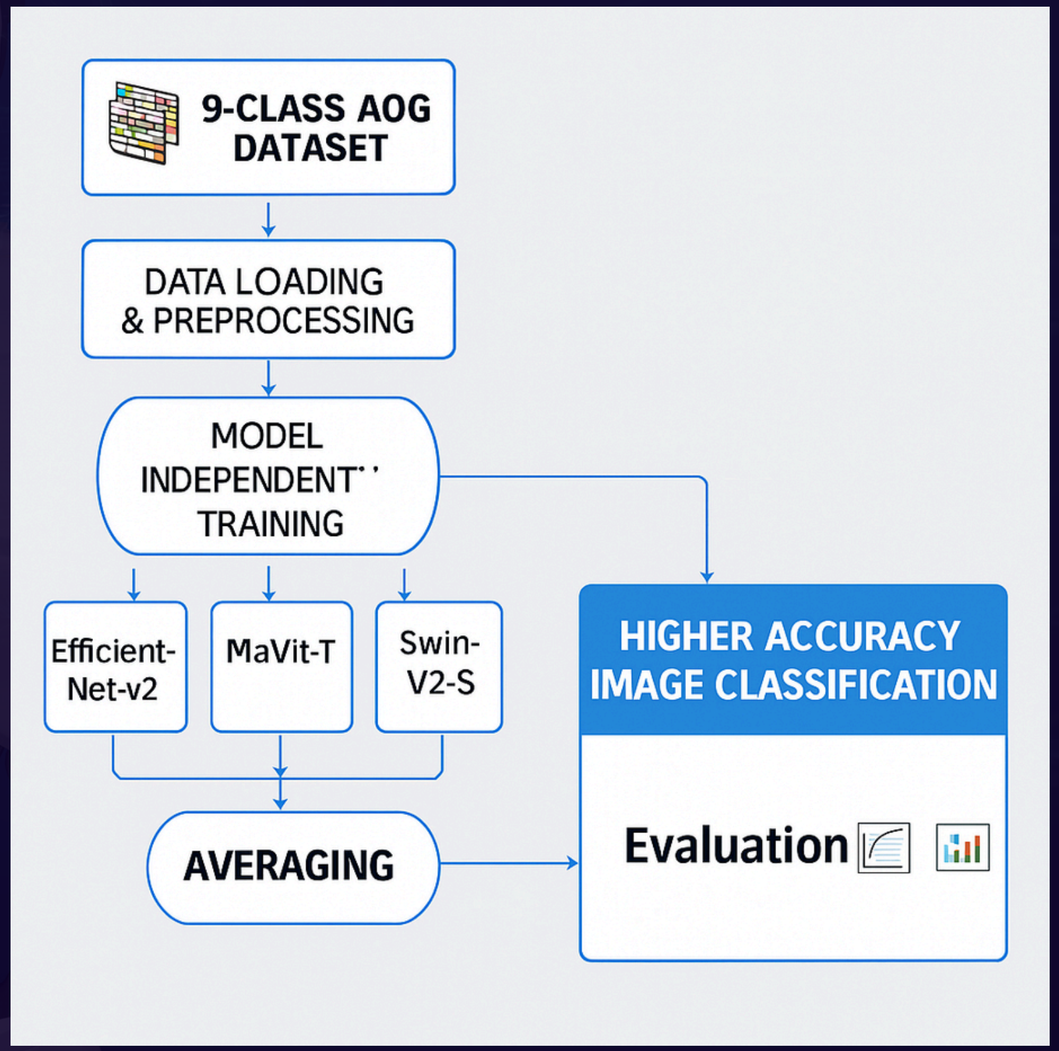
EfficientNetV2-L, Swin Transformer V2-S and MaxViT-T Integration for 9-Class Image Classification



SKYNET

Project Overview

This project aims to build a high-accuracy image classifier using a deep ensemble of CNNs and Vision Transformers on the 9-class AOG dataset. The hybrid ensemble captures both local and global features, implemented in PyTorch with torchvision, PIL, tqdm, and scikit-learn for preprocessing, tracking, and evaluation.



EDA

- Observed class imbalance in the dataset.
- Observed that most of the images were taken from the top and contained small plants with mostly greenery.
- Images lacked complex features, making learning easier.
- Image resolution was (256×256).
- So we expected the smaller models to perform as par as comple.x model, due to less complex features.
- Experimented with smaller models before moving to larger versions for better performance.

Data PreProcessing

Transformations Applied

1. Resize
   - Resized Images according to Model Architecture
2. Data Augmentation (Training only)
   - RandomHorizontalFlip → Adds left-right flip invariance.
   - RandomRotation(±10°) → Makes the model robust to slight rotations.
   - ColorJitter (brightness=0.2, contrast=0.2) → Simulates different lighting conditions.
3. Normalize
   - Uses ImageNet mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225].
   - Why?
     - Pre-trained Models expect inputs normalized like ImageNet images.
     - Speeds up convergence and stabilizes training.
     - We did not have enough Data to use mean and std of given dataset

Our Solution

Methodology-Ensemble

EfficientNet V2 L

CNN Model

Swin V2 S

Transformer Model

MaxVit T

CNN + Transformer Model

EfficientNet Models

Since the train data seemed very less complex, we started out with CNNs.
In the EfficientNet family, we took B7 (66.3M params, 37.75 GFLOPS). An ensemble of this gave an accuracy of 95.57% on val and 91.98% on public LB.
Continuing this approach, we took a larger model, EfficientNetV2-L (118.5M params, 56.08 GFLOPS). An ensemble of this gave an accuracy of 97.32% on val and 93.94% on public LB.

- 3-Fold Stratified Cross-Validation (StratifiedKFold).
- Input: 384 × 384 Batch: 8 (with Gradient Accumulation steps=2 → Effective 16).
- Optimizer: AdamW + Cosine Annealing LR Scheduler.
- Metric: Val Acc + Confusion Matrix Per Class.

Swin V2 S Model

Given that CNNs primarily capture local patterns, we shifted to a transformer to leverage its ability to model long-range dependencies and richer feature interactions.

We chose SwinV2-S for its lightweight yet powerful transformer architecture (49.7M parameters, 11.55 GFLOPS), which captures global context effectively, making it well-suited even for relatively simple training data. Ensemble of Fold models gave us an accuracy of 94.218% on public test dataset

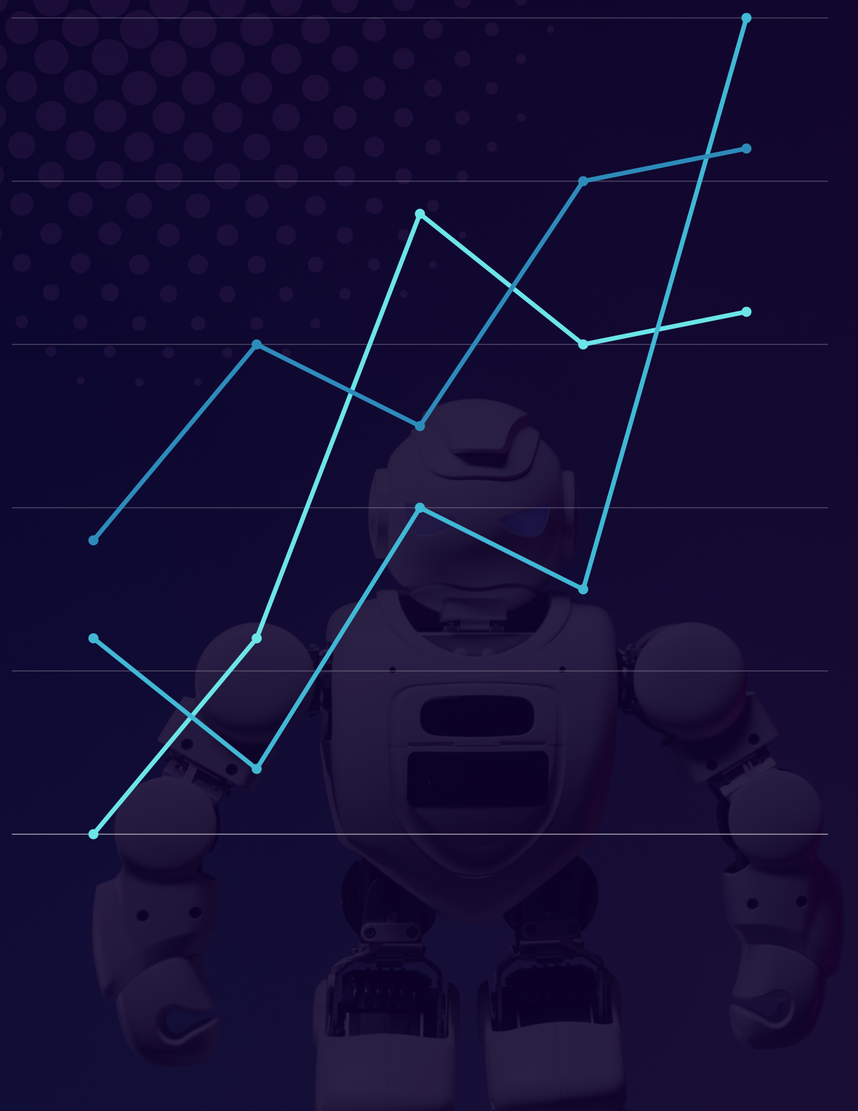
- 3-Fold Stratified Cross-Validation (StratifiedKFold).
- Ensmbeling of Best models from each fold
- Optimzer and Loss Function are Same.
- Metric: Validation Accuracy
- Checkpoint: Best Val Acc per fold.

MaxVit T Model

Given the promising results from CNN-based and Transformer based architecture ,we experimented with hybrid architectures that combine convolution and attention based mechanisms for improved spatial and global understanding.

Within this family, we selected MaxViT-T (30.9 M Parameters, 5.56 GFLOPS). This Achieved an accuracy of 93.83% on Private Hidden Dataset and

- 5-Fold Stratified Cross-Validation (StratifiedKFold).
- Input: 224 x 224 Batch: 16.
- Optimizer: AdamW + Cosine Annealing LR Scheduler.
- Metric: Validation Accuracy and Loss.
- All other stuffs are same as previous models.



Ensemble Stage 1 (Swin +EfficientNet)

By fusing predictions from both models, we mitigate the individual limitations of each, improve generalization, and deliver more robust and precise predictions compared to a single-model approach.

With this combined ensemble, we obtained **94.53%** accuracy on the Private LB and **94.40%** on the Public LB

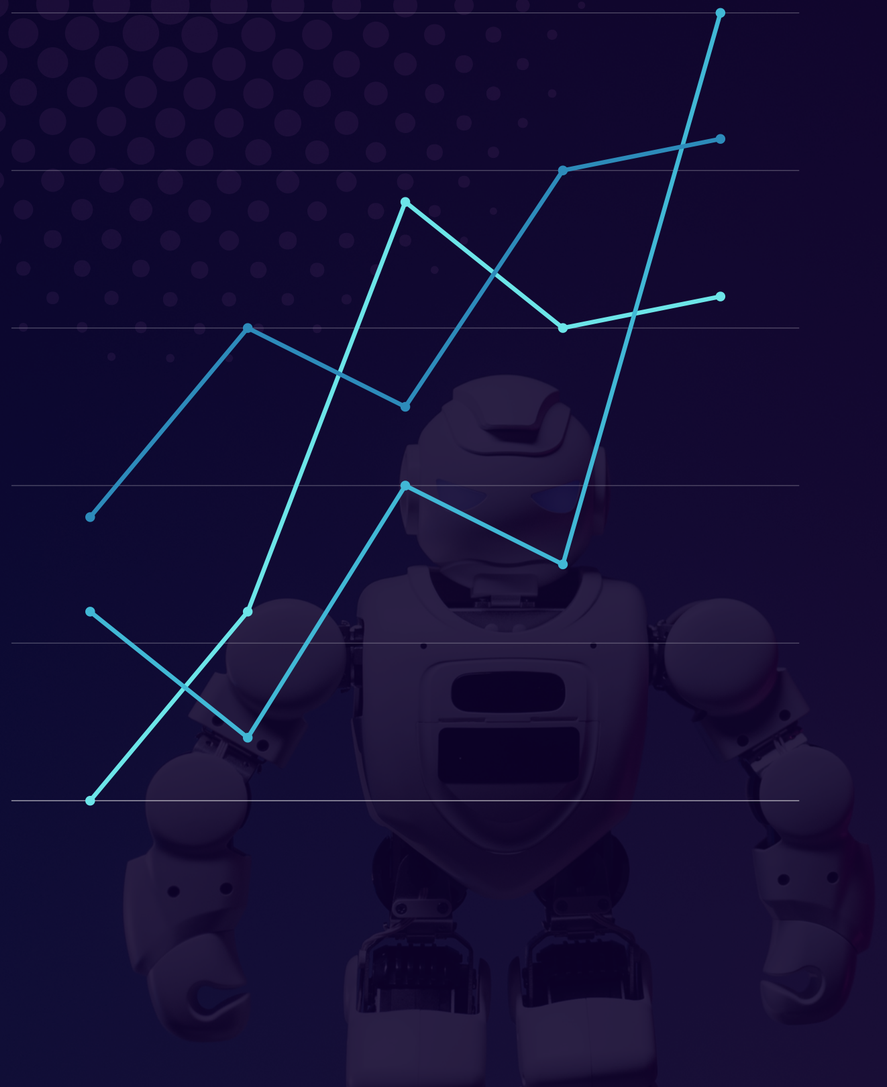
- EfficientNetV2-L: Excels at capturing fine-grained local textures and subtle features.
- SwinV2-S: Effectively models hierarchical and long-range dependencies within the image.

Ensemble Stage 2 (Swin + MaxViT)

We perform a transformer pair fusion by combining the predictions of SwinV2-S and MaxViT-T to leverage their complementary attention mechanisms. Using this Model we were able to achieve a accuracy of **94.65%** on Private LB and **94.35%** of accurayc on Public LB

Benefits of Swin + MaxViT fusion:

- Local + Grid Attention Synergy: Swin captures hierarchical local context, while MaxViT integrates global attention patterns.
- Reduced Overfitting: Averaging predictions smooths out individual model biases.
- Enhanced Generalization: Combines complementary strengths to better handle subtle inter-class differences.



Final 3-Way Ensemble

- Combined Strengths:
- EfficientNet → Captures fine local texture fidelity.
- SwinV2 → Models hierarchical context effectively.
- MaxViT → Provides global attention synergy, capturing long-range dependencies.
- By averaging their predictions, we reduce individual model biases, enhance generalization, and achieve more accurate and reliable results than any single model.
- Using this approach we achieved **94.65%** of accuracy on Private LB and 94.53% on Public LB and this is the top performer of AI OF GOD 4.0


AI OF GOD 4.0
Some Other
Approachs

OverSampling Approach: EfficientNet B2 / V2 L

We considered the edge case that the 40% public test (which felt similar to train data) have a different distribution than hidden 60% (always there was 3% accuracy drop in public LB), hence for a robust method we used Oversamping , we first started with EfficientNet B2 and were able to achieve Accuracy of 92.55% on Private Hidden Test Dataset

When we got Extra 5 Hours, we use **EfficinetNet V2 L** with oversampling and transformations + This is trained only on 2/3rd train data and validated on rest 1/3rd due to time constraints, so we had to submit it halfway but still we got an accuracy of 93.56% on hidden private test dataset. And we expect an ensemble of this to achieve even better performance while improving class generalisation.

Efficient Approach: EfficientNet B2

To improve the efficiency of our models while maintaining high predictive performance, we transitioned from larger architectures to the EfficientNet-B2 model, which comprises only 9.1 million parameters and reduces computational cost to 1.09 GFLOPS.
This adjustment enabled a favorable balance between accuracy and inference speed.

Despite the substantial reduction in model complexity, the EfficientNet-B2 maintained strong performance, achieving a validation accuracy of 95.30%, a public leaderboard score of 92.39%, and a private leaderboard score of 92.74%, demonstrating competitive results relative to larger models across the leaderboard.

Stacking Apporach: Final 3 Way Ensemble

After the competition was extended, we experimented with adding a four-layer MLP [128, 64, 32, 9] with a softmax activation on top of the predicted class probabilities.

The MLP was trained on 80% of the training data and validated on the remaining 20% holdout set

Achieving a public leaderboard accuracy of 94.11%.

While promising, this approach was considered exploratory and not part of the final submission.

AI OF GOD 4.0

Key Takeaways

- Test distribution closely matches train distribution; SKF-trained models performed consistently well on Public Leaderboard.
- Multi-architecture ensembles showed strong generalization across the dataset.
- This approach can be further extended to stacking for improved performance.
- CNN + Transformer hybrids achieved superior results by combining local and global feature learning.

AI OF GOD 4.0

Thank You!

Arnav Tripathi, Rajat Nandkumar Shedshyal, Ritesh Kumbhare

arnavtripathi01, rajatshedhyal, riteshkumbhare11