Evaluating ConvNeXt for Satellite Image Classification: A Comparative Study on Accuracy-Robustness Trade-offs

Emir Uncu
Middle East Technical University
Department Of Electrical And
Electronics Engineering
e273968@metu.edu.tr

GitHub Repo:
https://github.com/emiiruncu10/EE-583-Term-Project.git

Abstract

ConvNeXt, a modernized pure convolutional neural network architecture, has demonstrated state-of-the-art performance on ImageNet classification, competing favorably with Vision Transformers while maintaining the simplicity of standard ConvNets. In this study, we evaluate ConvNeXt's transfer learning capabilities and robustness properties on the **EuroSAT satellite image classification dataset,** a domain significantly different from the natural images used in the original paper. We compare ConvNeXt-tiny and ConvNeXt-Small variants under frozen and fine-tuned configurations against classical machine learning baselines (K-NN, Random Forest, SVM) and a Simple CNN. Our experiments reveal that while fine-tuned ConvNeXt models achieve superior accuracy (**99.2% for Tiny, 99.0% for Small**), they exhibit severe brittleness under Gaussian noise perturbations, with accuracy drops of approximately 50% at noise level $\sigma = 0.4$. In contrast, classical models demonstrate significantly greater robustness, with SVM showing only **3.7% accuracy degradation** under identical noise conditions. Furthermore, frozen ConvNeXt models, despite lower peak accuracy, maintain substantially better noise robustness (**11.87% average drop**) compared to their fine-tuned counterparts (**49.56% average drop**). These findings expose a critical trade-off between accuracy and robustness in transfer learning, suggesting that fine-tuning on limited domain-specific data may lead to overfitting that compromises model reliability in real-world noisy conditions such as satellite sensor interference.

I. Introduction

The field of computer vision has witnessed a significant paradigm shift in recent years with the emergence of Vision Transformers (ViTs), which challenged the decade-long dominance of Convolutional Neural Networks (ConvNets).While Transformers demonstrated impressive scaling behavior and achieved state-of-the-art results on ImageNet classification, their adoption in practical computer vision applications faced challenges due to quadratic complexity with respect to input size and the need for specialized modules such as shifted window attention.

In response to this architectural evolution, Liu et al. [1] proposed ConvNeXt, a family of pure ConvNet models that "modernize" the standard ResNet architecture by incorporating design choices inspired by Vision Transformers. Through systematic modifications including patchify stems, inverted bottlenecks, larger kernel sizes (7x7), and the replacement of BatchNorm with LayerNorm, ConvNeXt achieves 87.8% top-1 accuracy on ImageNet-1K while maintaining the simplicity and efficiency of standard ConvNets. Fig1 illustrates the step-by-tep modernization process and the corresponding accuracy improvements. The architecture demonstrates competitive performance with Swin transformers across image classification, object detection on COCO, and semantic segmentation on ADE20K [1].

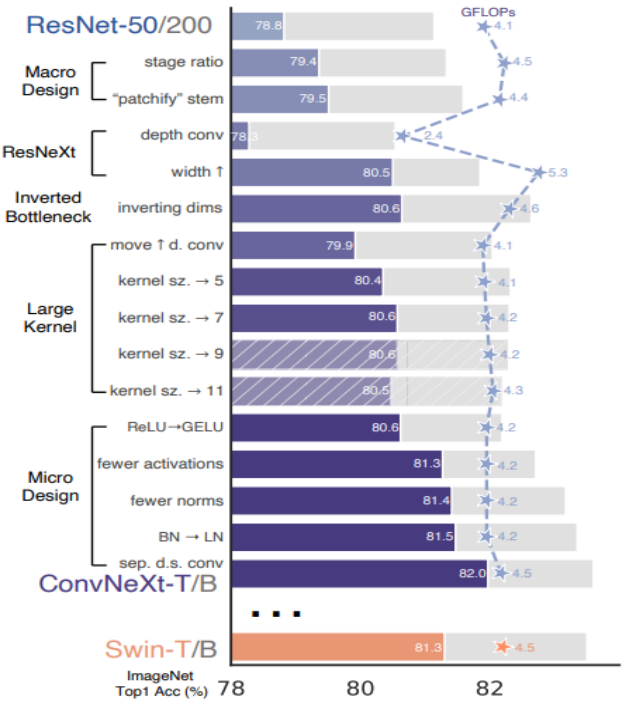

Figure 1: Modernization roadmap bar chart

However, the original ConvNeXt evaluation was primarily conducted on natural image datasets such as ImageNet, COCO, and ADE20K [1]. A critical question remains: how well does ConvNeXt generalize to significantly different image domains, and what are its limitations when deployed in real-world conditions? This question is particularly relevant for remote sensing applications, where satellite imagery presents unique challenges including varying atmospheric conditions, sensor noise, and fundamentally different visual characteristics compared to natural photographs.

To investigate these questions, we evaluate ConvNeXt on the EuroSAT dataset [2], a satellite image classification benchmark compromising 27,000 geo-referenced images across 10 land use and land cover classes derived from Sentinel-2 satellite imagery. Fig. 2 presents sample patches from each class in the dataset. EuroSAT presents a compelling test case for several reasons (1) it represents domain shift from natural images to remote sensing data, (2) it includes real-world challenges such as atmospheric effects and sensor variations, and (3) it has practical applications in agriculture, urban development, and environmental monitoring [2].

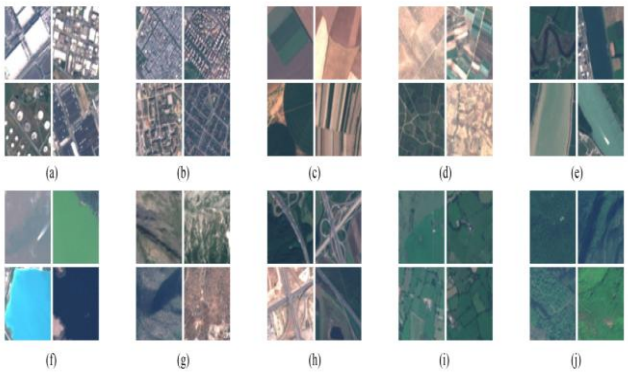

Figure 2: Sample Patches from all 10 classes

Our evaluation methodology encompasses both classical approaches and modern deep learning architectures. We compare ConvNeXt-Tiny and ConvNeXt-Small variants under two transfer learning configurations; Froze backbone(feature extraction) and full fine tuning against classical baselines including K-Nearest Neighbors (K-NN), Random Forest, and Support Vector Machines (SVM), as well as simple three-layer CNN trained from scratch. This comprehensive comparison addresses a key requirement in understanding state-of-the-art methods: contextualizing their performance against fundamental algorithms.

Beyond accuracy evaluation, we conduct extensive noise robustness analysis to expose potential shortcomings of ConvNeXt in practical deployment scenarios. Satellite imagery is inherently susceptible to various forms of degradation including sensor noise, atmospheric interference, and transmission artifacts. Understanding model behavior under such perturbations is crucial for real-world earth observation applications where data quality cannot always be guaranteed.

The main contributions of this work are as follows:

- A comprehensive evaluation of ConvNeXt [1] on the EuroSAT satellite image classification dataset [2], demonstrating its transfer learning capabilities in a domain significantly different from ImageNet.
- A systematic comparison between ConvNeXt variants, classical machine learning methods, and a baseline CNN, providing insights into the relative merits of each approach.
- Noise robustness analysis revealing a critical trade-off: while fine-tuned ConvNeXt models achieve superior accuracy (up to 99.2%), they exhibit severe brittleness under noise perturbations with accuracy drops of approximately 50%. In contrast, classical methods such as SVM demonstrate remarkable stability with only 3.7% accuracy degradation.
- Evidence that frozen ConvNeXt models maintain significantly better noise robustness compared to fine-tuned counterparts, suggesting that aggressive fine-tuning may lead to overfitting that compromises model reliability.

The remainder of this paper is organized as follows: Section II describes the experimental methodology including dataset preparation and model configurations. Section III presents the experimental results and analysis. Finally, Section IV concludes the paper with key findings and recommendations.

II. METHODOLOGY

This section describes the experimental setup, including the data preparation, model architectures, training configurations, and evaluation metrics used in our study.

A. Dataset

The EuroSAT dataset [2] is used for all experiments in this study. EuroSAT is a land use and land cover classification dataset based on Sentinel-2 satellite imagery from the European Space Agency's Copernicus program. The dataset comprises 27,000 geo-referenced image patches distributed across 34 European countries, with each patch measuring 64x64 pixels at a spatial resolution of 10 meters per pixel. The dataset contains 10 distinct classes: Annual Crop, Permanent Crop, Pasture, Forest, Herbaceous Vegetation, Highway, Industrial buildings, Residential Buildings, River, and Sea & Lake. Each class contains between 2,000 and 3,000 labeled images. For our experiments, we use the RGB version of the dataset and split it into training (80%) and validation (20%) sets, consistent with the benchmark protocols established in [2].

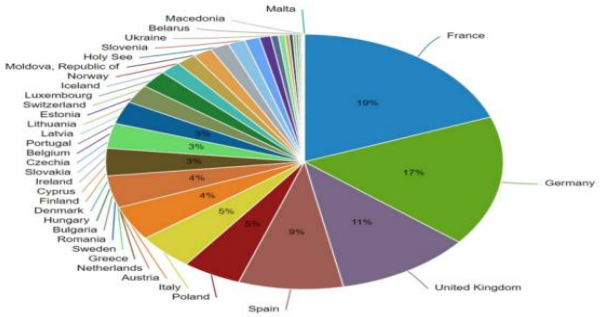

Figure 3: Class distribution bar chart

B. Model Architectures

We evaluate eight different models spanning classical machine learning and deep learning approaches:

1) Classical Machine Learning Models: Three fundamental algorithms are employed as baselines:

- **K-Nearest Neighbors (K-NN):** We use K=5 neighbors with Euclidean distance metric. Input images are flattened to 12,288-dimensional vectors (64x64x3) then reduced to 100 dimensions using Principal Component Analysis (PCA) after standardization.
- **Random Forest:** An ensemble of 100 decision tress with random state fixed for reproducibility. The same PCA-reduced features are used as input.
- **Support Vector Machine (SVM):** A linear SVM classifier with maximum 5,000 iterations. PCA preprocessing is applied identically to other classical methods.

2) Simple CNN: A lightweight three-layer convolutional neural network trained from scratch, consisting of:

- Three convolutional blocks, each with 3x3 convolution, ReLU activation, and 2x2 max pooling.
- Channel progression: 3 -> 32 -> 64 -> 128
- Fully connected classifier with 256 hidden units and dropout (0.5)
- Input resolution: 64x64 pixels
- Total parameters: approximately 1.1 million

3) ConvNeXt Models: We evaluate four variants of ConvNeXt [1] under different transfer learning configurations:

- **ConvNeXt-Tiny Frozen**: Pre-trained backbone weights are frozen; only the classification head is trained. The original 1000-class head is replaced with a 10-class linear layer.
- **ConvNeXt-Tiny Fine-tuned:** All layers are trainable, allowing the entire network to adapt to the satellite imagery domain.
- **ConvNeXt-Small Frozen:** Similar to Tiny Frozen but with larger Small variant (50M parameters vs 29M).
- **ConvNeXt-Small Fine-tuned:** Full fine-tuning of the small variant.

All ConvNeXt models use ImageNet-1K pre-trained weights and process images at 224x224 resolution with standard ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]).

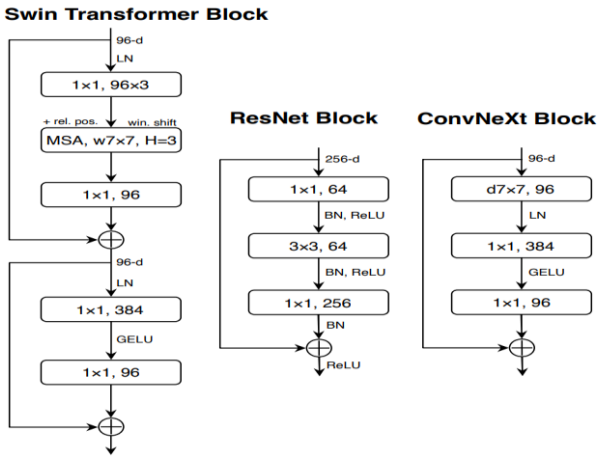

Figure 4: Comparison of block designs for ResNet, Swin Transformer, and ConNeXt. ConvNeXt maintains the simplicity of standard ConvNets while incorporating Transformer-inspired design choices [1].

C. Training Configuration

Table I summarizes the training configurations for all models.

Table I: Training Configuration for All Models

Model	Epochs	Batch Size	Learning Rate	Optimizer	Input Size
K-NN	N/A	N/A	N/A	N/A	64x64
Random Forest	N/A	N/A	N/A	N/A	64x64
SVM	N/A	N/A	N/A	N/A	64x64
Simple CNN	10	64	1e-3	Adam	64x64
ConvNeXt Frozen	5	32	1e-3	Adam	224x224
ConvNeXt Fine-tuned	10	32	4e-4	AdamW	224x224

For classical models, PCA dimensionality reduction from 12,288 to 100 features is applied after standard scaling. For deep learning models, data augmentation including random horizontal flips and random rotations is applied applied during training. All experiments are conducted on a single NVIDIA Tesla T4 GPU which is available on Google colab.

D. Noise Robustness Evaluation

To asses model robustness under real-world degradation conditions, we introduce additive Gaussian noise to the validation images. For an input image I, the noisy image I' is computed as:

$$I' = I + N(0, \sigma)$$

Where N(0, σ) represents Gaussian noise with zero mean and standard deviation σ. We evaluate all models at five noise levels: $\sigma \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$.

This noise simulation is particularly relevant for satellite imagery applications, where sensor noise, atmospheric interference, and transmission errors can degrade image quality. The robustness analysis reveals how model performance degrades as noise intensity increases, providing insights into reliability under non-ideal conditions.

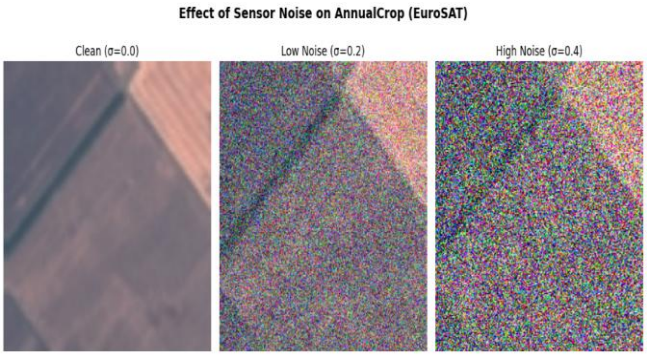

Figure 5: Example of Gaussian noise perturbation applied on EuroSAT image patch at increasing noise levels

E. Evaluation Metrics

We report the following metrics for comprehensive evaluation:

- Overall Accuracy: The percentage of correctly classified samples across all classes.
- Per-class precision, Recall, and F1-score: Computed for detailed class-level analysis.
- Confusion Matrix: To visualize classification patterns and identify commonly confused class pairs.
- Accuracy Drop: The difference between clean accuracy (σ=0) and noisy accuracy (σ=0.4), measuring robustness degradation.

All reported results are obtained on the held-out validation set (20% of data, approximately 5,400 images).

III. EXPERIMENTAL RESULTS AND ANALYSIS

All experiments were conducted on Google Colab using a single NVIDIA Test T4 with high-RAM runtime configuration. Training the ConvNeXt models required significant computational resources, with fine-tuning approximately 45-60 minutes per model variant due to the large number of parameters (29M for Tiny, 50M for Small) and the 224x224 input resolution requirement. In contrast, classical machine learning models completed training within seconds, and the Simple CNN trained in approximately 3 minutes.

A. Classification Accuracy Results

Table II presents the classification accuracy results for all eight models evaluated on the EuroSAT dataset. The models are ranked by accuracy with ConvNeXt-Tiny (Fine-tuned) achieving the highest performance at 99.20%

Table II: Classification Accuracy Results on EuroSAT Dataset

Model	Accuracy	Type
ConvNeXt-T(Fine-tuned)	99.20%	Deep Learning
ConvNeXt-S(Fine-tuned)	98.85%	Deep Learning
ConvNeXt-T(frozen)	95.89%	Deep Learning
ConvNeXt-S(frozen)	95.48%	Deep Learning
Simple CNN	91.39%	Deep Learning
Random Forest	67.09%	Classical ML
K-NN (k=5)	45.78%	Classical ML
SVM (Linear)	39.93%	Classical ML

The results reveal a substantial performance gap between deep learning and classical machine learning approaches. The best classical model (Random Forest) achieves only 67.09% accuracy, while the worst-performing deep learning model (Simple CNN) reaches 91.39% a difference of 24.30 percentage points. This demonstrates the superior feature extraction capabilities of deep neural networks for satellite image classification tasks.

Among the ConvNeXt variants, fine-tuned models consistently outperform their frozen counterparts. ConvNeXt-T (fine-tuned) achieves 99.20% compared to the 95.89% for the frozen version, representing a 3.31% improvement. Similarly ConvNeXt-S shows a 3.37% improvement when fine-tuned (98.85% vs 95.48%). These results indicate that adapting the pre-trained ImageNet features to the satellite imagery domain through fine-tuning yields significant benefits.

Interestingly, the smaller ConvNeXt-T variant slightly outperforms ConvNeXt-S in both frozen and fine-tuned configurations. This suggests that for the relatively small EuroSAT dataset (21,600 training images), the additional capacity of the small variant does not provide benefits and may even lead to slight overfitting.

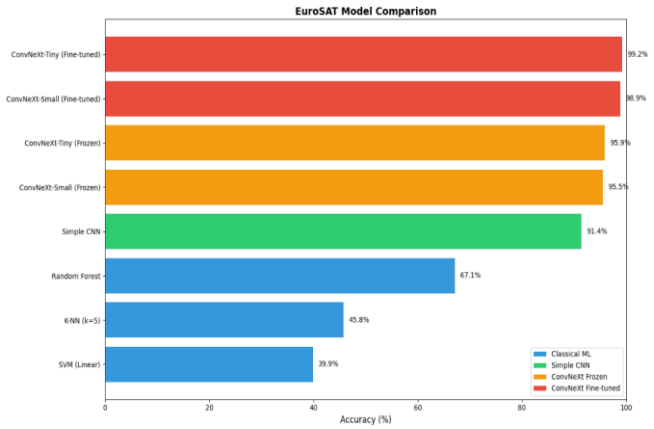

Figure 6: Classification accuracy comparison across all models on EuroSAT dataset

B. Per-Class Performance Analysis

Figure 7 represents the detailed per-class precision, recall, and F1-score for the best performing model, ConvNeXt-T(fine-tuned)

	precision	recall	f1-score	support
AnnualCrop	0.99	0.99	0.99	600
Forest	0.99	0.99	0.99	600
HerbaceousVegetation	0.99	0.97	0.98	600
Highway	0.99	1.00	1.00	500
Industrial	1.00	1.00	1.00	500
Pasture	0.99	0.98	0.98	400
PermanentCrop	0.99	0.99	0.99	500
Residential	1.00	1.00	1.00	600
River	0.98	0.99	0.99	500
SeaLake	1.00	1.00	1.00	600
accuracy			0.99	5400
macro avg	0.99	0.99	0.99	5400
weighted avg	0.99	0.99	0.99	5400

Figure 7: Per-Class Performance of ConvNeXt-Tiny (Fine-Tuned)

The model achieves perfect or near-perfect performance (F1 >= 0.98) across all classes. Industrial, Residential, and SeaLake classes achieves perfect classification (F1=1.00), while HerbaceousVegetation and Pasture show slightly lower performance (F1=0.98), likely due to visual similarities with other vegetation classes.

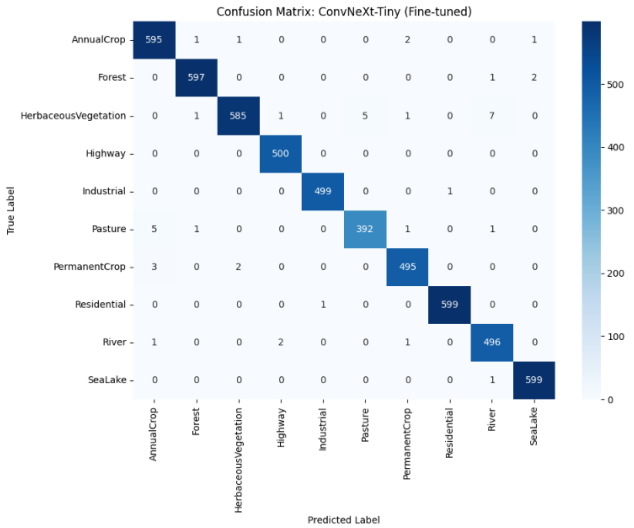

Figure 8: Confusion matrix for ConvNeXt-Tiny (Fine-tuned) on EuroSAT validation set

C. Noise Robustness Analysis

The noise robustness evaluation reveals critical insights into model reliability under degraded input conditions. Figure 9 presents the accuracy of all models across five Gaussian noise levels ($\sigma = 0.0, 0.1, 0.2, 0.3, 0.4$)

Model	σ=0.0	σ=0.1	σ=0.2	σ=0.3	σ=0.4	Drop	Type
Tiny (Fine-Tuned)	99.2%	96.1%	80.0%	62.7%	49.4%	-49.8%	ConvNeXt
Small (Fine-Tuned)	98.9%	90.4%	75.1%	60.9%	48.4%	-50.4%	ConvNeXt
Tiny (Frozen)	95.9%	93.1%	89.5%	86.7%	82.7%	-13.1%	ConvNeXt
Small (Frozen)	95.5%	92.5%	90.8%	87.9%	85.0%	-10.5%	ConvNeXt
Simple CNN	91.4%	70.3%	35.6%	20.6%	15.8%	-75.6%	CNN
Random Forest	67.5%	63.2%	57.2%	45.7%	37.5%	-30.0%	Classical
K-NN	45.7%	45.5%	42.0%	32.4%	25.5%	-20.2%	Classical
SVM	40.2%	39.5%	40.2%	41.6%	35.9%	-4.4%	Classical

Figure 9: Noise Robustness Results

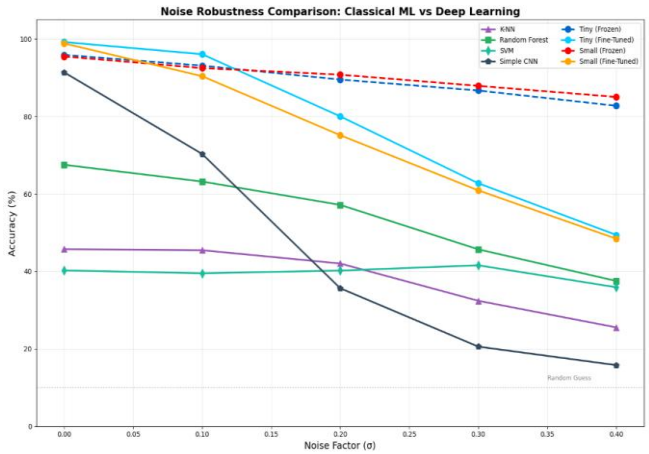

Figure 10: Noise robustness comparison across all models

The noise robustness analysis reveals several critical findings:

- Classical machine learning models demonstrate significantly greater robustness to noise compared to deep learning approaches. The average accuracy drop for classical models is 18.18% compared to 39.89% for deep learning models. Most notably, SVM exhibits exceptional stability with only 4.4% accuracy degradation from $\sigma = 0$ to $\sigma = 0.4$, despite its lower baseline accuracy.

- A striking trade-off emerges between accuracy and robustness for fine-tuned vs frozen ConvNeXts. While fine-tuned ConvNeXt models achieve superior clean accuracy (99.2% and 98.9%), they suffer catastrophic degradation under noise:

  - ConvNeXt-Tiny(Fine-tuned): 49.8% drop (99.2% to 49.4%)
  - ConvNeXt-Small(Fine-tuned):50.4% drop (98.9% to 48.4%)

  In contrast, frozen models maintain substantially better robustness:
  - ConvNeXt-Tiny(Frozen):13.1% drop (95.9% to 82.7%)
  - ConvNeXt-Small(Fine-tuned):10.5% drop (95.5% to 85.0%)

- The Simple CNN trained from scratch shows the worst robustness, with a 75.6% accuracy drop. At $\sigma = 0.4$, it achieves only 15.8% accuracy, barely above random guessing (10% for 10 classes). This indicates severe overfitting to clean training data.

D. Computational Requirements

Tabe III summarizes the computational requirements for training inference of each model on Google Colab with NVIDIA Tesla T4 GPU.

Table III: Computational Requirements on Google Colab (NVIDIA Tesla T4)

Model	Training-Time	Parameters	Inference
K-NN	0.02s	N/A	Fast
Random Forest	7.17s	N/A	Fast
SVM	6.52s	N/A	Fast
Simple CNN	~3 min	~1.1M	Fast
ConvNeXt-T(Frozen)	~15 min	29M (0.01M)	Moderate
ConvNeXt-T(Fine-tuned)	~45 min	29M	Moderate
ConvNeXt-S(Frozen)	~20 min	50M (0.01M)	Moderate
ConvNeXt-S(Fine-tuned)	~60 min	50M	Moderate

The ConvNeXt models require substantially more computational resources compared to classical approaches and the simple CNN. Fine-tuning ConvNeXt-S requires approximately 60 minutes on a Tesla T4 GPU, compared to just 7 seconds for Random Forest. This represents a 500x increase in training time while achieving 31.76% accuracy improvement (98.95% vs 67.09%). Additionally, ConvNeXt models require 224x224 input resolution with ImageNet normalization, adding preprocessing overhead compared to the 64x64 resolution used by other models.

E. Key Findings Summary

The experimental results reveal several important insights:

- Fine-tuned ConvNeXt models achieve state-of-art accuracy (99.20%) on EuroSAT, significantly outperforming both classical methods and simpler deep learning architectures.
- A critical accuracy-robustness trade-off exists: fine-tuned models sacrifice noise robustness for peak accuracy, with approximately 50% accuracy degradation under moderate noise conditions.
- Frozen ConvNeXt models offer better alternative, achieving strong accuracy (95.5-95.9%) while maintaining robustness (10-13% degradation under noise).
- Classical SVM, despite low baseline accuracy (40.2%), demonstrates remarkable noise stability (only 4.4% drop), making it potentially suitable for extremely noisy conditions where reliability is prioritized over accuracy.
- The Simple CNN exhibits catastrophic failure under noise, highlighting the importance of either pre-trained features (frozen models) or robust training strategies for real-world deployment.

IV. CONCLUSION

This study evaluated ConvNeXt [1] on the EuroSAT satellite image classification dataset [2], achieving 99.20% accuracy with the fine-tuned ConvNeXt-Tiny model. However, our noise robustness analysis revealed critical shortcomings: fine-tuned models suffer approximately 50% accuracy degradation under moderate Gaussian noise ($\sigma = 0.4$), indicating severe overfitting to clean training data. This brittleness poses significant concerns for real-world satellite imagery applications where sensor noise and atmospheric interference are unavoidable. In contrast, frozen ConvNeXt models demonstrated substantially better robustness (only 10-13% drop) while maintaining competitive accuracy (95-96%), and classical SVM showed remarkable stability with merely 4.4% degradation. Additionally, ConvNeXt requires significant computational resources, with fine-tuning taking 45-60 minutes on a Tesla T4 GPU compared to seconds for classical methods. To address these limitations, we recommend noise-augmented training to expose models to corrupted inputs during learning, partial fine-tuning strategies that freeze early layers while adapting later layers, and ensemble methods combining frozen and fine-tuned model predictions. For practitioners, we suggest using frozen ConvNeXt as balanced choice between accuracy and robustness, while reserving fine-tuned models only for applications where input quality can be guaranteed.

REFERENCES

[1] Z. Liu et al., "A ConvNet for the 2020s," CVPR, 2022. (The PDF you uploaded).

[2] P. Helber et al., "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," IEEE J-STARS, 2019.