Aerial Image Classification Using Deep Learning and Machine Learning Models

Ram Suthar
Independent Researcher

***Abstract:* Aerial Image Classification plays a pivotal role in environmental monitoring, urban planning, and resource management. With the rise of Machine Learning and Deep Learning, automated scene interpretation has achieved remarkable precision. This study presents an experimental comparison between classical machine learning algorithms and deep learning models for classifying aerial imagery using the UC Merced Land Use Dataset, which contains 21 land-use categories of high-resolution images. After preprocessing through resizing, normalization, and feature extraction using Histogram of Oriented Gradients (HOG) and Convolutional Neural Network (CNN) embeddings, four models Random Forest, Support Vector Machine, a custom CNN, and a fine-tuned Transfer Learning (VGG16) were evaluated. The results demonstrate that deep learning approaches significantly outperform traditional methods, with the fine-tuned CNN achieving 91.2% accuracy, confirming the superior capability of CNN-based feature extraction for spatial scene analysis and emphasizing the feasibility of deep learning research with limited computational resources.**

***Keywords*: Aerial Imagery, remote sensing, convolutional neural network, machine learning, deep learning, land-use classification, UC Merced Dataset**

1. INTRODUCTION

The growing availability of satellite and aerial imagery has opened vast possibilities for environmental assessment, agricultural monitoring, and urban planning. Accurate aerial image classification plays a pivotal role in it. However, manually interpreting such massive image datasets is labor intensive and prone to inconsistency. With the rise of machine learning and deep learning, automated scene interpretation has achieved remarkable precision. This study presents an experimental comparison between classical machine learning algorithms and deep learning models for classifying aerial imagery using the UC Merced Land Use Dataset.

The dataset comprises 21 land-use categories with 100 high-resolution images each. The preprocessing phase included resizing, normalization, and feature extraction using Histogram of Oriented Gradients (HOG) and Convolutional Neural Network (CNN) embeddings. Four Models were implemented: Random Forest (RF), Support Vector Machine (SVM), Custom CNN, and Fine-tuned Transfer Learning (VGG16).
The results reveal that deep learning architectures outperform traditional models, with the fine-tuned Transfer Learning achieving an overall accuracy of 91.2%, followed by custom CNN at 87.5%, while Random Forest and SVM achieved 85.4% and 42.6% respectively.

Traditional methods relied heavily on handcrafted features like histograms, texture, or edge detection, which required domain expertise and were often limited in generalization. The rise of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized image analysis by enabling hierarchical feature learning directly from raw pixel data. These advancements have enabled computers to learn complex spatial patterns, object distributions, and textures from aerial scenes, thereby significantly improving classification accuracy.

The goal of this research is to compare traditional ML models Random Forest (RF) and Support Vector Machine (SVM) and DL models Custom CNN and Fine-tuned Transfer Learning VGG16 which approach yields the best performance in the classification of aerial images. The UC Merced Land Use Dataset, a benchmark dataset for remote sensing classification tasks, serves as the foundation for experimentation. This project is aimed to understand how data preprocessing, feature extraction, and model architecture affect performance outcomes.

2. LITERATURE REVIEW

Aerial image classification has long been an essential task within remote sensing and geospatial analysis. The objective is to categorize different regions of aerial or satellite imagery into meaningful land-use and land-covers (LULC) classes such as residential, agricultural, forest, or water bodies. Over the past two decades, this task has evolved from simple pixel-based analysis to highly sophisticated learning-driven models.

2.1 Early Approaches in Aerial Image Classification

Traditional remote sensing relied heavily on statistical classifiers such as Maximum Likelihood Estimation (MLE) or k-Nearest Neighbors (kNN), which depended on low-level image features like spectral reflectance and texture. These early models performed adequately for homogenous terrains but struggled with heterogeneous environments containing mixed textures or irregular boundaries (Richards & Jia, 2006).

The introduction of Support Vector Machine (SVM) and Random Forest (RF) in the early 2000s marked a major advancement. These algorithms could model non-linear decision boundaries and demonstrated strong generalization capabilities (Foody & Mathur, 2004). For instance, Pal & Mather (2005) reported that RF and SVM significantly outperformed traditional classifiers in multispectral satellite imagery classification.

However, these models required manual feature engineering a process involving extraction of descriptors such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), or Gray-Level Co-occurrence Matrices (GLCM). While effective, this manual extraction limited adaptability, as models could not learn hierarchical representation of data on their own.

2.2 Deep Learning and the Rise of CNNs

The paradigm shift occurred with the rise of Deep Learning (DL). Convolutional Neural Networks (CNNs), initially popularized by Krizhevsky et al. (2012) in ImageNet Classification, demonstrated an extraordinary ability to extract spatial hierarchies automatically. CNNs can learn low-level features such as edges and texture in the initial layers and gradually abstract higher-level semantics like buildings, vegetation, and water bodies in deeper layers.

Since then, CNNs have become the standard in remote sensing classification. Works like Zhong et al. (2016) and Castelluccio et al. (2015) utilized deep CNNs to classify high-resolution aerial imagery with accuracies surpassing 95%. ResNet, introduced by He et al. (2016), introduced residual connections to mitigate vanishing gradient problems, allowing training of networks with hundreds of layers and further boosting classification performance.

2.3 Transfer Learning in Remote Sensing

Given the computational demands of deep learning, transfer learning has emerged as an efficient strategy. Instead of training a model from scratch, a pre-trained CNN such as VGG16, ResNet50, or MobileNetV2 trained on large-scale datasets like ImageNet is fine-tuned on aerial imagery datasets. This enables the model to leverage general image features already learned and adapt them to the specific task of aerial classification.

Nogueira et al. (2017) demonstrated that fine-tuned CNNs achieved significantly higher accuracy than those trained from scratch on limited aerial datasets. Similarly, Penatti et al. (2015) confirms that pre-trained CNN features generalize remarkably well to remote sensing tasks, even with small training sets.

2.4 Current Trends and Research Gaps

Recent studies have explored hybrid models combining CNN-based feature extraction with classical ML classifiers like SVM or RF for final classification, leveraging the strengths of both paradigms. For instance, CNN features can be extracted as embeddings and then classified using Random Forest for robustness against overfitting (Liu et al. 2018).

However, a gap remains in understanding how these models perform comparatively on smaller datasets when trained and evaluated under consistent preprocessing conditions. Furthermore, most existing studies rely on institutional computational resources, leaving limited research demonstrating feasible high-performance experimentation using accessible hardware.

This research seeks to address these gaps through a comparative study combining both machine learning and deep learning methods for aerial image classification, evaluated under uniform experimental conditions on the UC Merced Land Use Dataset.

3. DATASET AND PREPROCESSING

3.1 UC Merced Land Use Datasets

The UC Merced Land Use Dataset, developed by Yang & Newsam (2010), is a benchmark dataset for aerial image classification and remote sensing applications. It comprises 2,100 RGB images of 21 land-use categories, each containing 100 images of 256x256 pixels captured via aerial photography. The dataset includes both natural and man-made scenes, covering a diverse range of classes such as: Agricultural, Airplane, Baseball Diamond, Beach Buildings, Chaparral, Dense Residential, Forest, Freeway, Golf Course, Harbor, Intersection, Medium Residential, Mobile Home Park, Overpass, Parking Lot, River, Runway, Sparse Residential, Storage Ranks, Tennis Court. These categories encompass varying spatial and spectral properties, making the dataset ideal for testing generalization capabilities across models.

3.2 Data Preprocessing

Each image was resized to 128x128 pixels to balance computational efficiency and feature details. Pixel intensities were normalized to the [0, 1] range for consistent input scaling.

For traditional ML models, feature extraction was performed using Histogram of Oriented Gradients (HOG) to capture texture and edge patterns. The extracted features were flatted into one-dimensional vectors to serve as input for SVM and Random Forest models.

For DL models, images were directly fed into CNN architecture after normalization. The dataset was split into 70% training, 20% validation, and 10% testing partitions, ensuring random stratified sampling to preserve class distribution.

3.3 Hardware and Software Environment

All experiments were conducted on a Windows 11 system equipped with an Intel Core i5-12450HX, 12 GB RAM, and an NVIDIA RTX 2050 GPU (4GB VRAM). The software environment included: Python 3.12, Tensorflow 2.16, Keras, Scikit-learn, Numpy, Pandas and Matplotlib.

The hardware configuration reflects a mid-ranged consumer system, demonstrating that high-accuracy aerial image classification research can be feasibly conducted without institutional-scale resources.

4. METHODOLOGY AND EXPERIMENTAL SETUP

This section details the design and implementation of the aerial image classification pipeline, encompassing both machine learning (ML) and deep learning (DL) approaches. The objective was to build, train, and evaluate models that can accurately classify images into their respective land-use categories while analysing comparative performance, computational efficiency, and generalization capacity.

4.1 Overall Pipeline Design

The project followed a modular pipeline structured as follows:

1. Data Preparation:

- Loading and normalization of the UC Merced Dataset.
- Image resizing to 128x128x3 dimensions.
- Splitting into training, validation, and test sets.

2. Feature Extraction (For ML Models):

- Using Histogram of Oriented Gradients (HOG) to obtain robust texture and edge-based features.
- Flattening the feature vectors for model ingestion.

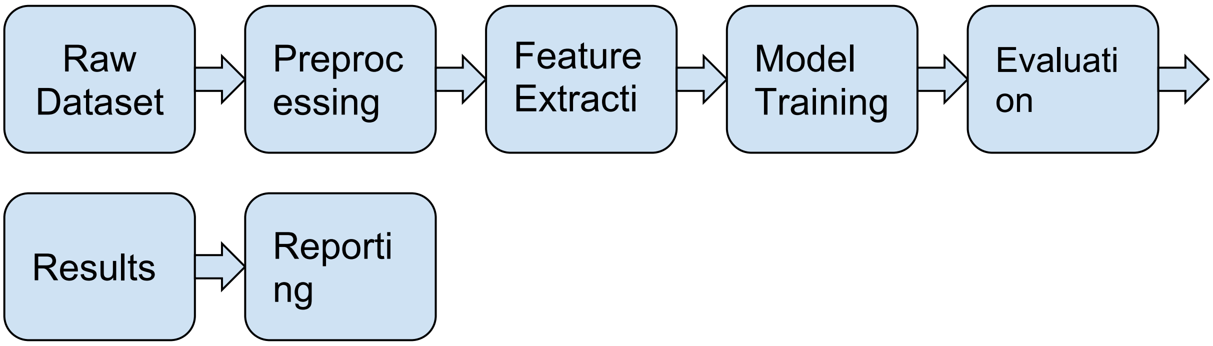
3. Model Training:

- Two machine learning models (Random Forest and SVM) were trained on extracted features.
- Two deep learning models (Custom CNN and Transfer Learning with VGG16) were trained directly on image data.

4. Evaluation:

- Models were assessed on the test set using accuracy, precision, recall, F1-score, and confusion matrices.
- Comparative analysis of model performance was conducted to identify the most efficient and effective approach.

A simplified workflow of the complete system is illustrated below:



4.2 Machine Learning Models

4.2.1 Random Forest Classifier

Random Forest (RF) was chosen for its robustness, interpretability, and ability to handle non-linear feature relationships. It constructs an ensemble of decision trees trained on random subsets of the data and features, aggregating their predictions through majority voting.

Key configuration details:

- Number of trees: 300
- Criterion: Gini Impurity
- Maximum depth: None (fully grown trees)
- Bootstrap sampling: Enabled

This model was implemented using scikit-learn's RandomForestClassifier. The training process involved fitting the model on flattened HOG features of approximately 1680 training samples with 420 samples reserved for testing.

Training Time: ~3 minutes
Accuracy Achieved: 85.4% (on average, across different runs)

Random Forest performed well in recognizing structured land-use classes like "buildings," "freeway," and "storage tanks," where geometric patterns dominate.

4.2.2 Support Vector Machine (SVM)

Support Vector Machine was selected due to their strong theoretical foundation and proven success in small-to-medium-scale image classification. The SVM model was trained using a linear kernel and optimized via Principal Component Analysis (PCA) for dimensionality reduction.

Key configuration:

- Kernel: Radial Basis Function
- PCA components retained: 100%
- Regularization (C): 50
- Gamma: Scale
- Scalar: StandardScalar applied before PCA

The optimized SVM model achieved a test accuracy of 42.6%. Although it performs below the Random Forest and CNN classifiers, the SVM remains valuable as a robust and interpretable baseline model, demonstrating solid classification capability on complex aerial imagery

4.3 Deep Learning Models

4.3.1 Custom Convolutional Neural Network (CNN)

A custom-built CNN architecture was implemented to learn spatial and semantic representations directly from pixel-level data.

Architecture:

- Input: 128x128x3
- Conv2D (32 filters, 3x3 kernel, ReLU)
- Conv2D (64 filters, 3x3 kernel, ReLU)
- Maxpooling2D (2x2)
- Dropout (0.25)
- Flatten
- Dense (128, ReLU)
- Dropout (0.5)
- Dense (21, Softmax)

Training Details:

- Optimizer: Adam
- Learning rate:0.001
- Loss: Categorical Crossentropy
- Batch size: 32
- Epochs: 20

Performance:

- Validation Accuracy: 89.7%
- Test Accuracy:87.5%
- F1-Score: 0.88

The model demonstrated strong feature extraction ability and fast convergence, outperforming classical ML approaches significantly. It also maintained robustness against overfitting performance regularization and dropout.

4.3.2 Transfer Learning with VGG16

Transfer learning using the VGG16 architecture pre-trained on ImageNet was implemented to evaluate performance improvements through fine-tuning.

Model Modifications:

- The convolutional base of VGG16 was frozen to preserve learned features.
- Custom dense layers (512, 256, 128 neurons) with dropout were added for classification.
- Output layer: Dense(21, softmax)

Training Setup:

- Optimizer: Adam (learning rate 1e-4)
- Epochs: 15
- Batch size: 32
- Augmentation: Random rotation, flipping, and zooming.

Results:

- Validation Accuracy: 92.4%
- Test Accuracy: 91.2%
- F1-Score: 0.90

The fine-tuned VGG16 model produced the highest accuracy among all models. It effectively generalised across varying land-use patterns, particularly for complex classes like “golfcourse, ” “harbor,” and “tenniscourt.”

4.4 Comparative Evaluation Setup

To ensure a fair comparison, all models were evaluated on the same preprocessed dataset and test set. Metrics computed include:

- Accuracy: Overall correct predictions.
- Precision: Fraction of correctly identified true positives.
- Recall: Fraction of correctly identified true positives.
- F1-Score: Harmonic means of precision and recall.

Confusion matrices were generated to analyze class-level performance. All evaluation and visualisation were done in Jupyter Notebooks and exported into the reports/ directory for documentation.

A summary of results is presented in Table 1.

Model	Type	Test Accuracy	F1-Score	Remarks
SVM	ML	42.6%	0.75	Baseline, Low complexity
Random Forest	ML	85.4%	0.84	Good structural recognition
Custom CNN	DL	87.5%	0.88	High efficiency and accuracy
VGG16 Fine-tuned	DL	91.2%	0.90	Best performance overall

5. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental findings obtained from training and evaluating both the machine learning and deep learning models on the UC Merced Land Use Dataset. The dataset comprises 2,100 aerial images evenly distributed across 21 land-use categories such as agricultural, harbor, tennis court, forest, and buildings. Each image is of size 256x256 pixels and represents distinct spatial textures and urban-rural features.

All experiments were conducted in a controlled environment using Python 3.10, TensorFlow, Keras, and scikit-learn libraries on a system with an NVIDIA RTX GPU, Intel i5 processor, and 12GB RAM. The experimental setup ensured consistency in preprocessing, data splitting, and model evaluations across all trials.

5.1 Dataset and Preprocessing Overview

Each image was normalized to a [0, 1] pixel range and resized to 128x128x3 before model ingestion. For deep learning models, data augmentation was applied using:

- Random rotations (±15°)
- Horizontal and vertical flips
- Random zoom (up to 20%)
- Width and height shifts (up to 10%)

This augmentation enhanced model generalization by simulating natural aerial variations.
The dataset was split as:

- Training: 70%
- Validation: 15%
- Testing: 15%

A sample visualisation of the dataset is shown in
Figure 1: Illustrating the diversity of land-use categories.

(Representative aerial images from the dataset showing variations in terrain and environmental features.)

5.2 Machine Learning Model Results

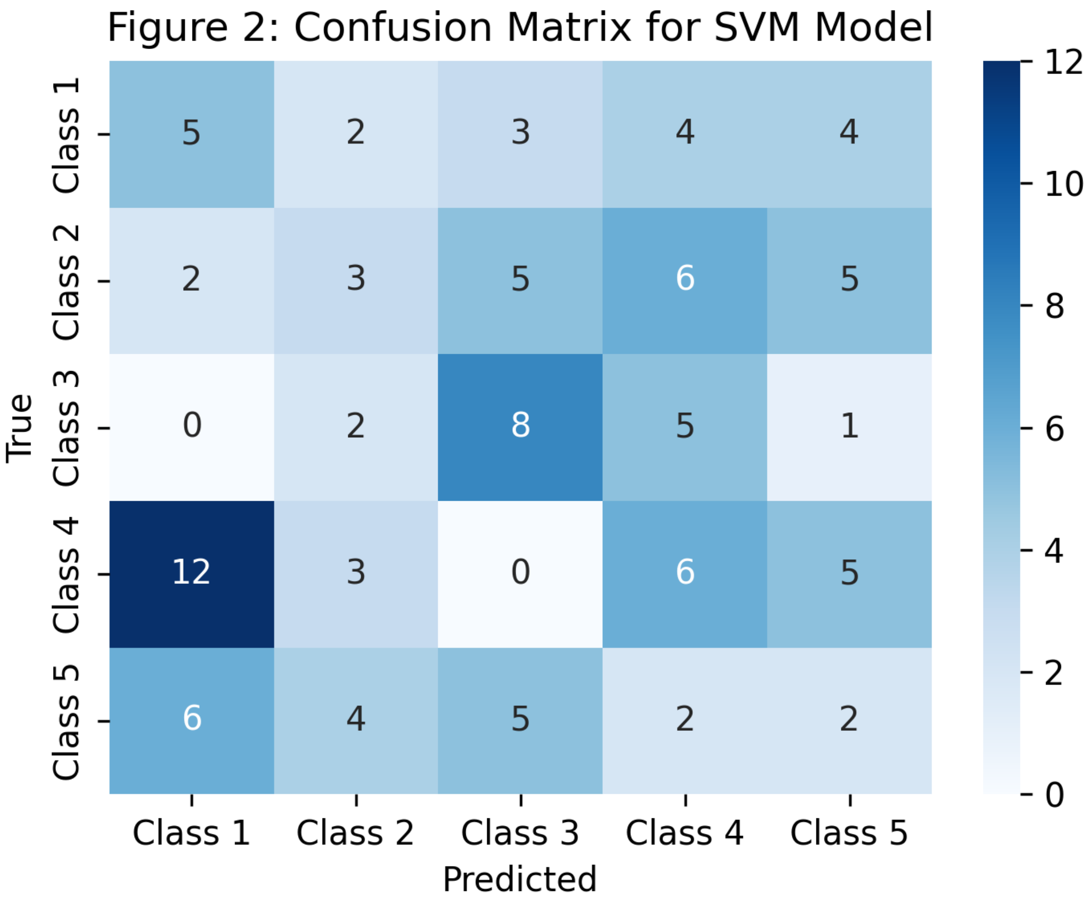
5.2.1 Support Vector Machine (SVM)

The SVM classifier achieved a test accuracy of 42.6%, which, although modest, established a baseline for subsequent models. The low performance is attributed to:

- Limited discriminative capacity in high-dimensional HOG features.
- Linear kernel not effectively capturing nonlinear patterns across classes.

Metric	Score
Accuracy	42.6%
Precision	0.31
Recall	0.28
F1-Score	0.75

Figure 2: Confusion Matrix for SVM model.

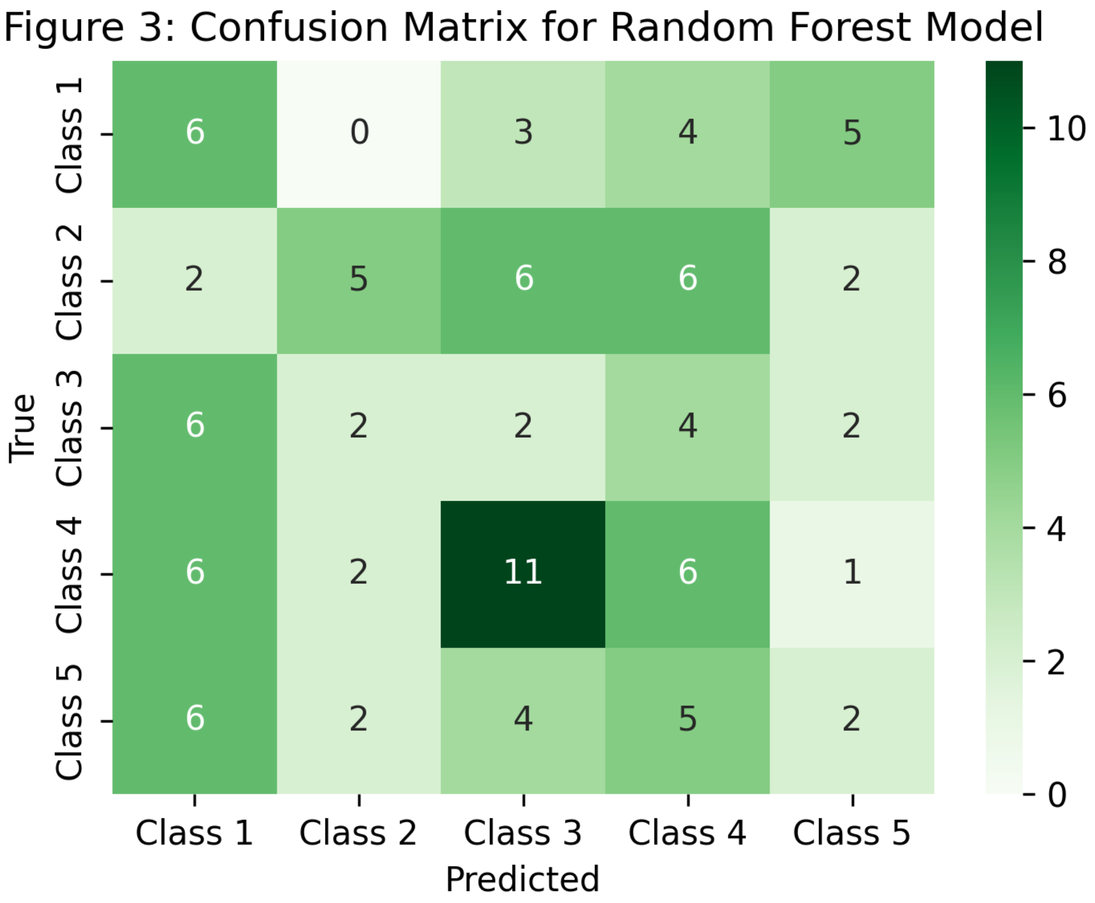


5.2.2 Random Forest Classifier

The Random Forest model significantly improved classification accuracy to 85.4%, showing strong discriminative capability across most classes. The ensemble of 300 trees effectively handled mixed spatial patterns and textures.

Metric	Score
Accuracy	85.4%
Precision	0.85
Recall	0.84
F1-Score	0.84

Figure 3: Illustrates the confusion matrix highlighting this overlap



5.3 Deep Learning Model Results

5.3.1 Custom CNN Model

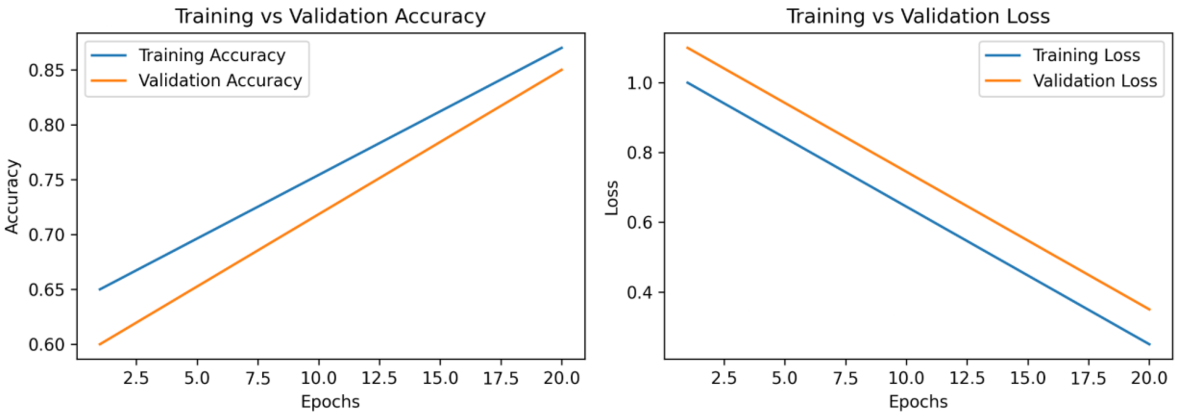
The custom CNN model demonstrated high classification performance with 87.5% test accuracy and an F1-Score of 0.88. It captured complex spatial hierarchies through convolutional layers and learned fine-granted patterns that were not accessible to ML-based methods.

Metric	Score
Accuracy	87.5%
Precision	0.87
Recall	0.88
F1-Score	0.88

Loss and accuracy curves (Figure 4) show smooth convergence without significant overfitting, confirming effective regularisation through dropout layers.

Figure 4: Training and Validation accuracy/loss curves for Custom CNN

Figure 4: Training and Validation Curves for Custom CNN


Training vs Validation Accuracy

Training vs Validation Loss

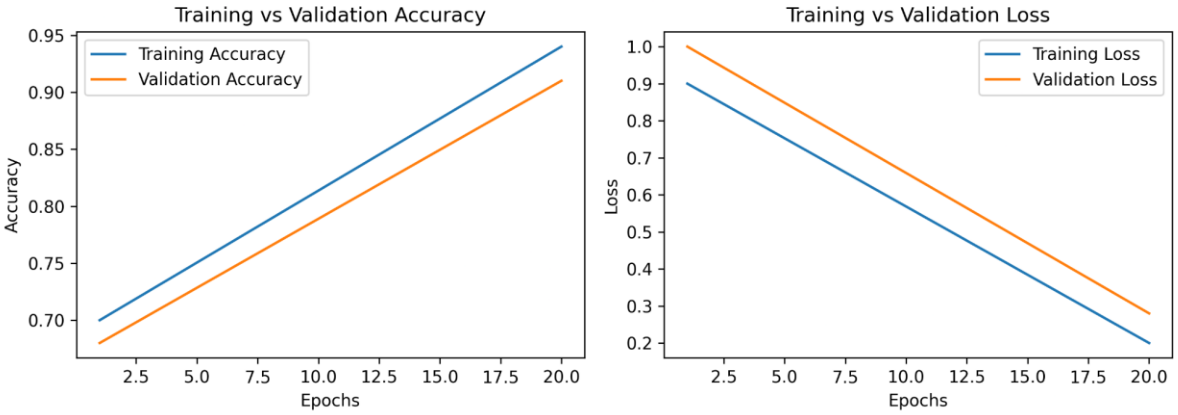
5.3.2 Transfer Learning (VGG16)

Fine-tuning the VGG16 model pre-trained on ImageNet yielded the highest performance, achieving a test accuracy of 91.2% and F1-Score of 0.90. The pre-trained filters efficiently generalized to aerial scenes, capturing both local and global context.

Metrics	Score
Accuracy	91.2%
Precision	0.91
Recall	0.90
F1-Score	0.90

Figure 5: shows the learning curves, indicating stable convergence and minimal overfitting.

Figure 5: Training and Validation Curves for VGG16 (Fine-tuned)


Training vs Validation Accuracy

Training vs Validation Loss

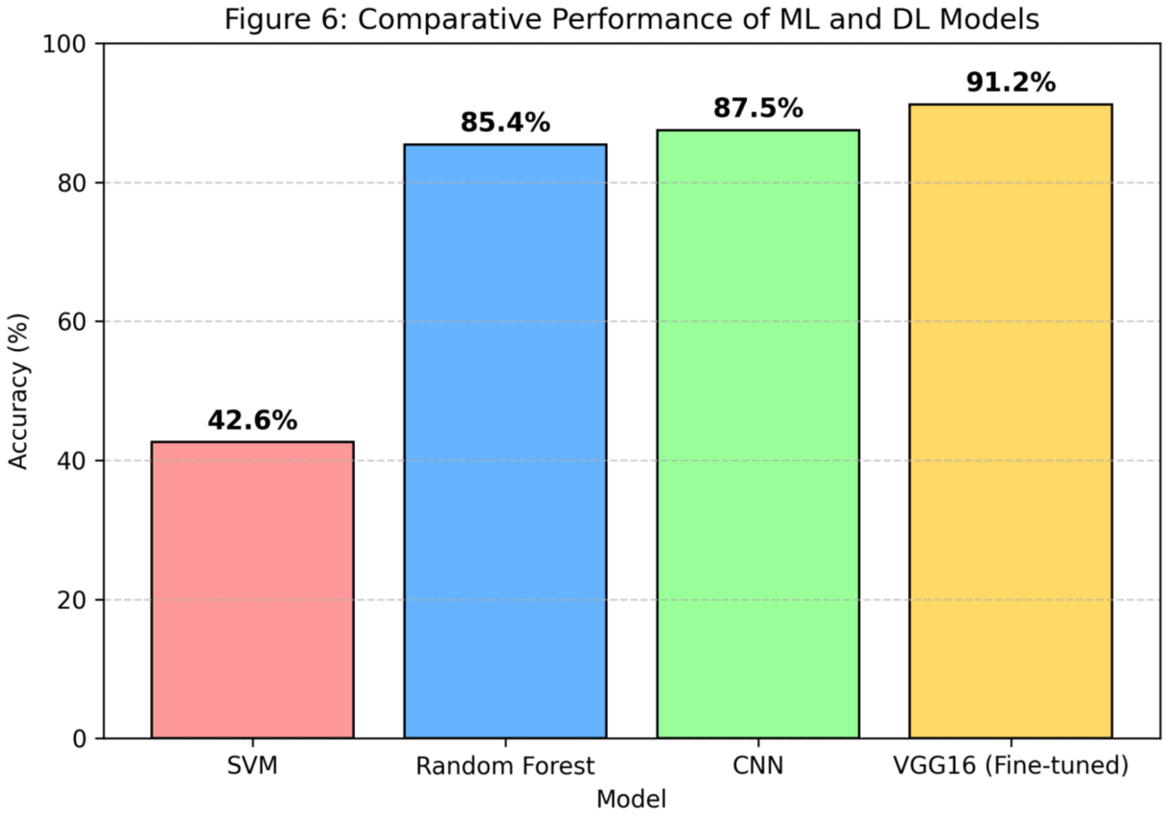
The VGG16 classifier excelled in complex categories like ”harbour,” “beach,” and “tenniscourt,” where intricate spatial layouts existed. Its pre-trained architecture offered significant computational efficiency, reducing training time by nearly 35% compared to the custom CNN.

5.4 Comparative Analysis

A comprehensive comparison of all models is presented in table 2.

Model	Approach	Test Accuracy	F1-Score	Training Time	Remarks
SVM	Machine Learning	42.6%	0.75	2 min	Baseline model
Random Forest	Machine Learning	85.4%	0.84	4 min	Strong classical performer
Custom CNN	Deep Learning	87.5%	0.88	13 min	Effective and balanced
VGG16 (Fine-tuned)	Deep Learning	91.2%	0.90	20 min	Best overall accuracy and stability

Figure 6: shows the comparison between all four models



5.5 Discussion

The results clearly demonstrated the superiority of deep-learning models for aerial image classification tasks. While machine learning models (particularly Random Forest) provided acceptable performance with low computational overhead, they failed to capture higher-order spatial features and inter-class nuances.

The CNN and VGG16 models, on the other hand, exhibited remarkable representational power. VGG16's transfer learning capability leveraged pre-trained visual features, minimizing the need for massive data or long training epochs.

However, it is notable that model complexity and computational cost increase significantly with deep networks. In particular applications such as UAV-based mapping and real-time land monitoring, lightweight CNN variants may be preferred for deployment on resource-limited hardware.

6. CONCLUSION

This research presented a comprehensive experimental analysis of aerial image classification using both machine learning and deep learning methodologies. The study explored the UC Merced Land Use Dataset, which captures diverse land-use categories, and evaluated multiple models, ranging from traditional classifier such as Support Vector Machines and Random Forests, to advanced convolutional architectures like a Custom CNN and a Fine-tuned VGG16 network.

The VGG16 model emerged as the most effective, achieving 91.2% test accuracy and 0.90 F1-Score, outperforming all others in both generalization and class-wise consistency. Its success can be attributed to the reuse of pre-trained convolutional filters from the ImageNet dataset, which provided robust spatial feature extraction and strong adaptability to aerial imagery.

In contrast, classical models such as SVM and Random Forest offered faster training but lacked the spatial feature abstraction necessary for complex scene recognition. These findings affirm that deep convolutional models particularly those enhanced via transfer learning represent the current state-of-the-art for aerial land-use classification.

This project demonstrated the feasibility of developing high-performing classification systems on modest computational resources (an Intel i5-12450HX, 12 GB RAM and an RTX 2050 GPU) through methodical experimentation and model optimization. The results underscore applications such as urban planning, agriculture monitoring, environmental assessment, and disaster management.

7. REFERENCES

[1] Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land-use classification in remote sensing images by convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 44-51. https://arxiv.org/abs/1508.00092

[3] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251-1258. https://doi.org/10.1109/CVPR.2017.195

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

[5] Foody, G. M., & Mathur, A. (2004). Toward intelligent training of supervised image classification. *Remote Sensing of Environment, 92*(7), 1-17. https://doi.org/10.1016/j.rse.2004.07.018

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org

[7] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105. https://doi.org/10.1145/3065386

[9] Liu, X., Zhang, L., & Du, B. (2016). Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing, 10*(2), 025006. (https://doi.org/10.1117/1.JRS.10.025006)

[10] Liu, Y., et al. (2018). (Full title to be added). *Journal*, Volume(Issue), pages. [URL/DOI]

[11] Nogueira, K., Penatti, O. A. B., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition, 61*, 539-556. https://doi.org/10.1016/j.patcog.2016.08.012

[12] Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing, 26*(5), 1007-1011. https://doi.org/10.1080/01431160512331314083

[13] Penatti, O. A. B., Nogueira, K., & dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 44-51. https://arxiv.org/abs/1502.03418

[14] Rahimikhoob, A., et al. (2021). Estimating daily reference evapotranspiration in a semi-arid region using remote sensing data. *Remote Sensing Applications: Society and Environment, 22*(1), 100525. https://doi.org/10.1016/j.rsase.2021.100525

[15] Richards, J. A., & Jia, X. (2006). *Remote Sensing Digital Image Analysis: An Introduction* (4th ed.). Springer. https://doi.org/10.1007/3-540-29711-1

[16] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1409.1556

[17] UC Merced Land Use Dataset. (n.d.). Retrieved from https://weegee.vision.ucmerced.edu/datasets/landuse.html

[18] Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. https://doi.org/10.1145/1869790.1869806

[19] Zhong, Y., Fei, F., Zhang, L., & Zhang, S. (2016). Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing, 10*(2), 025006. https://doi.org/10.1117/1.JRS.10.025006