Enhancing Aerial Scene Classification: A Comparative Study of Vision Models Based on Traditional Machine Learning and Deep Learning Methods

Ash Peng
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5473493@ad.unsw.edu.au

David Hu
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5613921@ad.unsw.edu.au

Yiming Liao
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5506757@ad.unsw.edu.au

Songkai Ma
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5526545@ad.unsw.edu.au

Evelyn Zhou
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5324147@ad.unsw.edu.au

***Abstract*—This study aims to compare the performance of different computer vision methods in classifying aerial scenes from remote sensing images. The comparison is based on the classification of 15 categories of landscape images from a Kaggle dataset. In this study a total of ten models based on machine learning-based methods and deep learning methods are applied. Machine learning-based methods including KNN, Random Forest, and combining the feature descriptors LBP and SIFT for feature extraction and use SVM for classification. Five deep learning methods are also adopted, including ResNet, EfficientNet, VGG, MobileNet, and MLP. By comparing accuracy, class-wise precision, recall, and F1 scores, the study identified ResNet, EfficientNet, MobileNet, and MLP as the best performing models. The result indicates that deep learning-based computer vision methods significantly outperform machine learning-based methods in capturing high-level image features and handling complex scenes. This provides support for large-scale automatic image analysis.**

***Index Terms*—Aerial scenes classification, Machine learning-based methods, Deep learning methods, CNN.**

I. Introduction

Image classification algorithms are a fundamental branch of computer vision. They enable computers to process, analyze, and interpret images to identify targets with various patterns. Image classification is widely regarded as a typical application of deep learning methods. Compared with traditional machine learning models, deep learning models can automatically learn features from images, thus reducing manual effort and minimizing human error. This method not only improves efficiency, but also improves classification accuracy. Currently, deep learning-based image classification techniques have been widely applied in real-world domains such as automated billing systems, medical diagnostics, and intelligent traffic systems.

This study compares the performance of different machine learning based methods and deep learning methods in the task of automatically classifying landform types in aerial remote sensing images. The accurate classification of these landform characteristics is crucial for practical applications such as urban planning, environmental monitoring, and disaster response. In such tasks, manual identification is inefficient and prone to subjectivity. Therefore, there is a need to explore more efficient and reliable automatic image classification methods tailored to these application scenarios. This study uses the SkyView Multi-Landscape Aerial Imagery Dataset, which contains 15 well-balanced and diverse landform categories, with 800 high-resolution remote sensing images per class, totaling 12,000 images. The dataset simulates real remote sensing images captured by satellites or drones. Consequently, it provides a valuable experimental basis for real-world image classification tasks. In the experimental design, 80 percent of the images are used for training and validation, and the remaining 20 percent are used as the test set. During the study, ten different learning models were trained, and their actual performance in the classification task was systematically evaluated. Our objective is to compare the classification performance of these models to identify the most suitable learning method for remote sensing image classification tasks.

II. Literature review

A. Machine learning-based methods

Before the advent of deep learning methods, image classification was based on manually extracted feature descriptors fed into traditional machine learning classifiers. In image analysis, Local Binary Pattern (LBP) and Scale-Invariant Feature Transform (SIFT) were the most used feature descriptors. The LBP operator extracts features by labeling the pixels of an image using thresholding. It compares the 3×3 area around each pixel with the center pixel and turns the result into a binary number [1]. Descriptors are generated by extracting the texture information from these images. A researcher has proposed an improved Extended Local Binary Pattern (ELBP). ELBP enhances the discriminative capability by containing multiple layers of LBP encoding. After extracting features, SVM classifier is used to classify satellite images. Experimental results demonstrated that the method has good feasibility in remote sensing image analysis [2]. SIFT generates descriptors by extracting local invariant features from the image. There was a study used SIFT-based BoVW and SVM, KNN, and Decision Tree classifiers to classify land types in high-resolution aerial images. SVM showed best accuracy in the experiment result [3]. Therefore, the combination of SIFT and SVM is also applicable to aerial image classification. KNN and SVM are both supervised learning classifiers. KNN is sensitive to high-dimensional data. SVM has better classification performance, however its training process is time-consuming on large-scale datasets. Random Forest is a tree-based ensemble method, and it has parallel processing capability and a feature selection mechanism. Fan has compared these traditional classifiers in a study on UAV Remote Sensing Images [4]. The results indicated that RF outperformed the other methods and showed more accurate discrimination capability.

However, these traditional methods rely on handcrafted feature extraction. When dealing with complex aerial image scenes, manually selecting appropriate features is time-consuming and prone to errors [5]. In processing large-scale datasets such as SkyView multi-landform aerial remote sensing images, low computational efficiency and high memory consumption often occur [6]. Due to the above limitations, researchers have gradually shifted to data-driven deep learning methods to improve classification performance and practical adaptability.

B. Deep learning methods

Deep learning enables models to automatically learn features directly from raw pixel data [7]. CNN is a deep learning model composed of multiple convolutional filters. As a trainable feature extractor, it has been widely applied in image classification tasks in computer vision and remote sensing [8]. A study has shown that in aerial image classification, models such as ResNet50, MobileNetV2, and VGG-16 in CNN perform well when processing high-resolution remote sensing images. These models can effectively extract spatial and semantic features of images and are suitable for target recognition and scene classification tasks in high-resolution remote sensing images [9]. In recent years, several classical deep learning architectures have been applied to aerial image classification tasks, achieving significant improvements in performance or efficiency:

The VGG network enhances classification performance by increasing network depth. When the depth of convolutional neural networks expanded to 16–19 layers, it achieved significantly higher accuracy compared to previous shallow architectures in image classification tasks. The VGG network structure is simple and widely used for recognizing landscape types in remote sensing scene classification tasks. However, due to its large number of parameters, computational cost correspondingly increases [10].

The proposal of ResNet solved the gradient degradation problem in deep CNN networks. ResNet introduces residual connections, allowing the network to effectively learn residual functions relative to earlier layers [11]. This design not only enhances training efficiency but also significantly improves the classification accuracy of models. Variants like ResNet-50 and ResNet-101 exhibit excellent accuracy and robustness in high-resolution remote sensing image classification. They also offer strong transfer learning capabilities, which makes them suitable for small-sample remote sensing classification tasks. This capability can greatly reduce the time and computational resources required for training [12].

MobileNet is a lightweight deep neural network with high accuracy. It adopts depthwise separable convolution, which splits standard convolution into depthwise convolution and pointwise convolution to reduce computational cost and the number of parameters [13]. MobileNet also has structurally more complex variants, MobileNet-V2 and MobileNet-V3, but in specific tasks and datasets, the complexity of the model does not necessarily lead to better performance. There was a study shown that MobileNet and its variants perform well in feature extraction and classification tasks for satellite images containing lakes and rivers. In the result of this experiment, MobileNet-V2 has achieved accuracy of 96 percent, which is the best performance among all variants [14]. This also indicates that in satellite remote sensing image classification tasks, MobileNet-V2 might be the most practical model choice.

EfficientNet is an efficient and high-accuracy CNN that achieves excellent image classification performance with less parameters. It uses a compound scaling strategy to efficiently balance depth, width, and input resolution. The model capacity and performance increase progressively from EfficientNet-B0 to B7. Lightweight models like B0 and B1 are suitable for environments with limited computational resources. More complex models like B5 to B7 perform better in tasks required higher precision. Compared with ResNet50, EfficientNet also provides strong generalization capability, but is more efficient and has lower computational cost. Thus, it is more suitable for a wide range of remote sensing image classification tasks [15].

The classic Multilayer Perceptron (MLP) is a simple structured feedforward neural network. MLP excels at capturing

spectral features but lacks the effective spatial structure modeling capability. It has many parameters and overfit easily. Thus, it typically underperforms CNN. However, a study has shown that although CNNs are effective in extracting spatial features, they may overlook spectral details. Therefore, the study proposed combining both through a hybrid MLP-CNN classifier to help improve the performance of remote sensing image classification [16]. Its experimental result has confirmed the feasibility of this approach.

III. Method

This study focuses on the use and comparison of various image classification methods, implemented to evaluate their accuracy on a multiclass aerial landscape dataset. The model comparison was made by using traditional machine learning methods and deep learning architectures. The traditional methods including k-Nearest Neighbors(KNN), Random Forests, as well as methods principally learned feature descriptors like Local binary patterns (LBP), Scale-Invariant Feature Transform (SIFT) that have been then combined with classifiers such as Support Vector Machines (SVM), while the deep learning methods dealt with VGG11, ResNet, EfficientNet, MobileNet, and MLP. The database contains images of 15 scene categories, and each model is trained and tested in the same training and testing splits to preserve the same conditions for a fair comparison. This section defines the methods and tools used for feature extraction, the reason why we used them in the experiment, the structure of models, and the settings of the training process.

A. Machine learning-based methods

Based on the literature and lectures, we selected three traditional machine learning models. KNN is a non-parametric and instance-based learning algorithm, which classifies a new input by finding the k closest labeled neighbors in the feature space and assigning the class that is most common among them. In prediction, the model would compare the distance between query point and then select k nearest neighbors to conduct a majority vote among labels. KNN has advantage on small scaled task and would be effective in our case, moreover it could be a solid baseline for us to compare with more complex models. The model used is imported from prepackaged model of sklearn, the number of neighbors is set to 10. Images are resized to 64x64 and normalized in data preprocessing step.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the majority vote (for classification) of all of the trees. Each tree is trained on a bootstrap sample with a random subset of features, and it splits nodes also using a random subset of features which can help reduce the possibility of overfitting. Beside this, it can handles high-dimensional feature vectors, and it works well with small to medium datasets. In this experiment, we utilized ResNet50 as a feature extractor to obtain meaningful representations from the images. The Random Forest classifier was configured with 100 estimators and a maximum tree depth of 8, and the batch size was set to 32.

For Support Vector Machine (SVM), it can find the optimal hyperplane that separates classes with the maximum margin in the feature space. It also supports linear and non-linear classification using kernel tricks to map inputs into higher-dimensional spaces. SVM is know as highly effective in high dimension cases, and is commonly used in combination with some special feature descriptor. In our setup, we used the SVC implementation from scikit-learn with a linear kernel. In addition to using SVM independently, we also tested it in combination with LBP and SIFT feature descriptors.

The two feature descriptors employed in this study are Local Binary Patterns (LBP) and Scale-Invariant Feature Transform (SIFT):

- LBP is a texture descriptor that compares each pixel with its 8 neighboring pixels in a local 3×3 window, resulting in an 8-bit binary code (ranging from 0 to 255) that encodes local texture information. Given that aerial landscape images often differ based on their texture characteristics, LBP is particularly effective for distinguishing scene categories with similar global structures but distinct textures.
- SIFT is a local feature descriptor that detects keypoints in an image and describes them in a way that is invariant to scale, rotation, and partially to illumination and affine transformations. Keypoints are identified using the Difference of Gaussians, and descriptors are computed by generating histograms of gradient orientations around each keypoint. This allows SIFT to capture distinctive structural features that are particularly helpful for identifying landmarks in aerial scenes such as buildings, roads, or ports.

Both of descriptors mentioned above are combined with SVM to do prediction in training step, we runned 50 epoch for each descriptor separately.

B. Deep learning methods

ResNet's deep architecture is equipped with a high-level semantic capturing ability, which makes it quite efficient in handling complex special features of aerial landscape images. it is based on the fundamental theory of residual learning. This is where identity shortcut connections come into play. By skipping one or more layers, such shortcuts can solve the vanishing gradient problem and thus allow the training of even deeper networks. The use of the residual connections, additionally, can serve as a preventive measure to stop overfitting and provide a solution to the problem of degradation in deep networks. For the purpose of our experiment, 'Resnet50' was the pre-trained model that we used and imported from torchvision.

EfficientNet is a scalable and efficient CNN that uses compound scaling to uniformly scale depth, width, and resolution using a fixed set of scaling coefficients. It achieves high accuracy with fewer parameters compared to other networks. It can balance between the accuracy and efficiency in prediction

by achieve state-of-the-art (SOTA) performance with fewer parameters. For here we used a pre-packaged model from torchvision.

We runned the method of ResNet and EfficientNet in a same file for comparison, the weights for both of the models are originally pre-trained default weights, and the hyperparameter are set to 10 epoches, batch size of 32, and learning rate of 0.001.

VGG also is a strong baseline model to involve in our experiment, because it's layer architecture simply makes it suitable for visual classification task. VGG networks rely on a uniform architecture, using stacked 3×3 convolution layers followed by max pooling and fully connected layers. Considering the limit of time and computational resources, we selected VGG11, which refers to the version with 8 convolutional layers and 3 fully connected layers. The learning rate is set to 0.0005 and we runned a test of 20 epoches but the result is unexpected, which will be discussed more in result section.

Under the circumstance of limited device, we what to know how is model inferenced on mobile and embedded devices perform on this classification task. MobileNet uses depthwise separable convolutions, which factorize a standard convolution into a depthwise convolution (per channel), and a point wise convolution (1×1), which significantly reduces the computation and model size. We selected the pre-constructed MobileNet-V2 model by Pytorch and using the version with pre-trained weights, with the default learning rate of 0.001 and setting of 50 epoches.

The last model is Multilayer Perceptron (MLP), intended to serve as a lightweight baseline model to compare against convolutional architectures. MLP is a fully connected feed-forward neural network. It maps input features through one or more hidden layers using linear transformations and non-linear activations, and outputs class probabilities using softmax. In practice, we implemented pretrained ResNet50 model to extract key information and added an attention block mechanism to enhace the model by forcing it focus on most informative parts. In the architecture of the MLP model, the hiden dimension is set to 512, and in optimizer, the learning rate is also 0.001.

IV. Experiment Result

A. Evaluation Metrics

Classification performance was assessed based on Precision, Recall, and F1-score, and the outcomes were visualized to facilitate comparative analysis across models.

Indicator	SVM	KNN	LBP	SIFT	VGG11	Random Forest	MLP	MobileNet	RestNet50	EfficientNetB0
Accuracy	36%	38%	45%	49%	66%	85%	95%	96%	97%	98%
Precision	35%	38%	45%	48%	65%	86%	95%	96%	97%	98%
Recall	35%	38%	44%	49%	65%	85%	95%	96%	97%	98%
F1-Score	35%	35%	44%	47%	65%	85%	95%	96%	97%	98%

Fig1. Comparison Table

As shown in Comparison Table (Fig.1), traditional models such as KNN, LBP, and SIFT achieved relatively low performance, with F1-scores below 50%. In contrast, tree-based and neural network models performed significantly better, with EfficientNetB0 achieving the highest scores across all metrics. The results clearly demonstrate the advantage of deep learning models, especially convolutional architectures, in handling complex visual patterns in remote sensing imagery.

B. Model visualization performance

To provide a more comprehensive evaluation, the performance of the deep learning models is further illustrated through confusion matrices and learning curves. These visualizations offer insight into class-wise prediction accuracy as well as model convergence behavior during training. Figures below present the results for MLP, MobileNet, ResNet50, and EfficientNetB0, accompanied by brief analyses.

1) MLP:

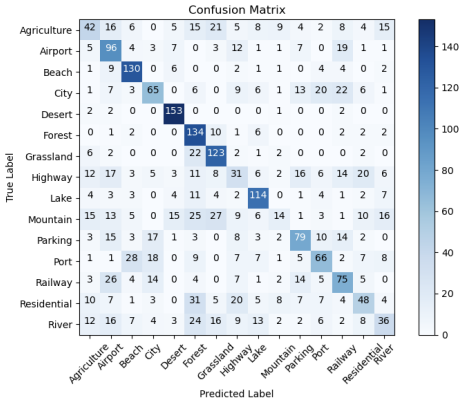

Fig2. MLP without ResNet50

As shown in Figure (Fig2), the classification performance of the MLP model was relatively limited when ResNet50 was not used for feature extraction. The confusion matrix reveals considerable misclassification across several categories, indicating that the model struggled to distinguish between visually similar land cover types based on raw input features.

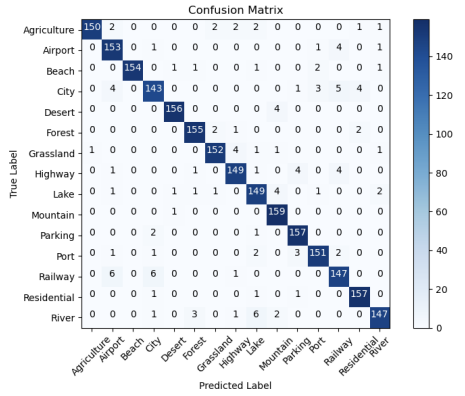

Fig3. MLP with ResNet50

Figure 3 illustrates that the classification performance of the MLP model improved significantly after incorporating feature extraction from ResNet50, achieving noticeably higher accuracy and reduced misclassifications across most categories.

2) MobileNet:

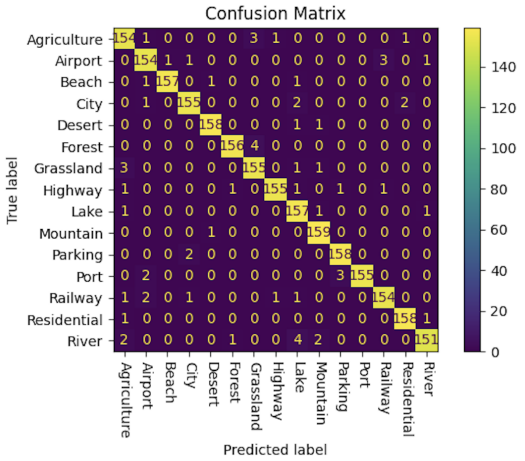

Fig4. Mobile Confusion Matrix

MobileNet, as a lightweight deep learning model, demonstrates strong classification performance with minimal confusion between classes, as shown in Figure 4. The confusion matrix reveals highly concentrated predictions along the diagonal, indicating accurate and consistent results across all land cover categories.

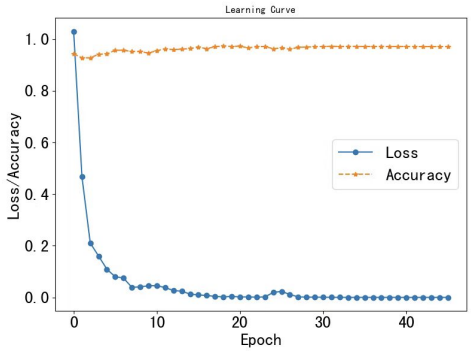

Fig5. MobileNet Learning Curve(Loss&Accuracy)

3) ResNet50:

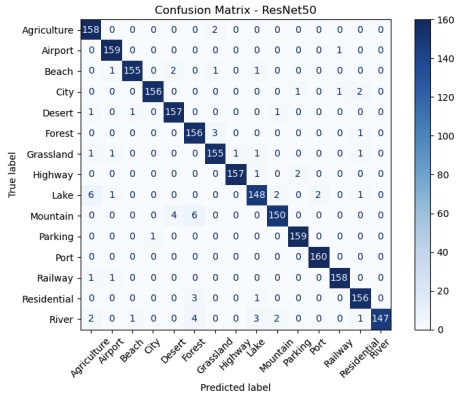

Fig6. ResNet50 Confusion Matrix

As shown in Figure 5, the confusion matrix of ResNet50 demonstrates outstanding classification performance. The predictions are highly concentrated along the diagonal, and the model achieves an overall classification accuracy of 97%. Only minimal misclassifications are observed, indicating strong robustness and feature discrimination capability.

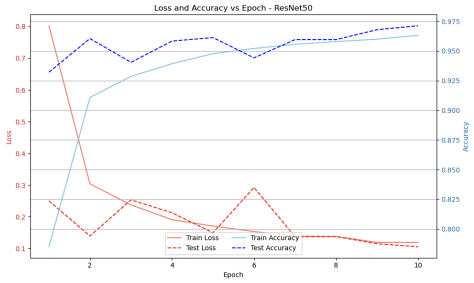

Fig7. ResNet50 Learning Curve(Loss&Accuracy)

4) EfficientNetB0:

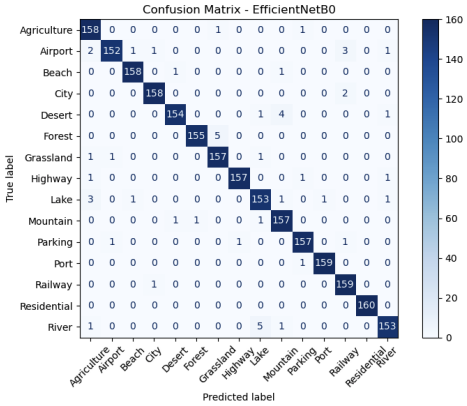

Fig8. EfficientNetB0 Confusion Matrix

EfficientNetB0 achieves near-perfect classification, with all predictions almost entirely aligned along the diagonal of the confusion matrix (Fig8). This result indicates exceptional generalization and fine-grained class distinction, confirming EfficientNetB0 as the top-performing model in this study.

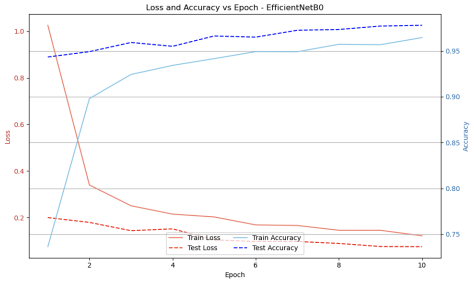

Fig9. EfficientNetB0 Learning Curve(Loss&Accuracy)

V. Discussion

A. MLP with ResNet50 Features

While the standalone MLP model underperformed, incorporating deep feature representations extracted from ResNet50 significantly improved its classification capability. As shown in Figure 3, the updated confusion matrix exhibits much tighter diagonal alignment, suggesting that the model was able to leverage semantic features captured by ResNet50 to enhance its decision boundaries. This architecture achieved over 94% accuracy, and notably reduced confusion among similar land

cover types such as Forest, Mountain, and Grassland. Compared with the baseline MLP without deep feature extraction, the enhanced version using ResNet50 features showed significantly reduced confusion across visually similar classes.

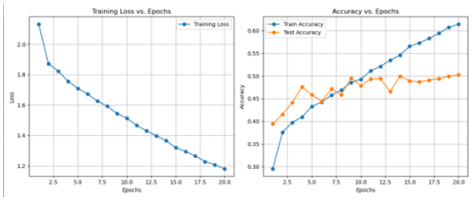

Fig10. Accuracy and loss curve of MLP without Resnet50

As shown in Figure 10, in the baseline MLP model without ResNet50 features,although training loss decreases steadily, the training accuracy plateaus below 62%, and test accuracy fluctuates around 50%, showing no substantial improvement. This suggests that the model struggles to capture meaningful patterns directly from raw input, lacking the ability to effectively distinguish between land cover categories.

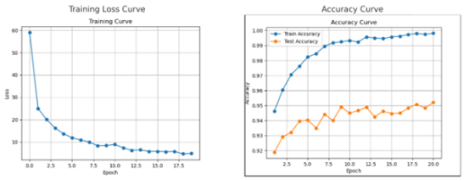

Fig11. Accuracy and loss curve of MLP with Resnet50

In contrast, when ResNet50 features are introduced (Fig11), the training process improves drastically. Training accuracy surpasses 98% within just a few epochs, and stabilizes above 99%. Test accuracy also increases consistently, reaching approximately 95% by epoch 20. This significant performance gain confirms that pretrained deep features greatly enhance the model's ability to generalize on complex image classification tasks. This result highlights the effectiveness of combining handcrafted classifiers with pretrained CNN-based embeddings.

B. MobileNet

MobileNetV2 demonstrates outstanding classification performance, especially considering its lightweight architecture. As illustrated in the confusion matrix (Fig4), the model maintains strong prediction consistency across all categories, with only minor misclassifications in a few complex classes like River and Port. The corresponding learning curve (Fig5) shows a rapid drop in loss and stable convergence of accuracy, achieving a final test accuracy of 96.58%. The early stabilization of accuracy around epoch 5 reflects MobileNet's computational efficiency and fast learning capability, making it suitable for deployment in resource-constrained environments.

C. ResNet50

ResNet50, a deeper convolutional network with residual connections, also achieved high accuracy at 97.25%. As observed in the confusion matrix (Fig6), the model performs exceptionally well across most categories, with very few off-diagonal elements. Its ability to maintain gradient flow across layers enables better representation learning, particularly for more visually subtle categories such as Lake and Residential. The learning curve (Fig7) demonstrates a steady decrease in loss and consistent accuracy gain over training epochs, indicating a well-generalized and stable training process.

D. EfficientNetB0

EfficientNetB0 outperforms all other models in this study, achieving a classification accuracy of 97.96%. As shown in Figure 8, the confusion matrix exhibits an almost ideal structure with diagonal dominance across all classes. The model displays exceptional generalization and is particularly effective in differentiating between fine-grained classes that are typically prone to confusion, such as River versus Lake, and Grassland versus Forest. This performance can be attributed to its compound scaling strategy, which uniformly balances network depth, width, and resolution. EfficientNetB0's superior results affirm the value of utilizing state-of-the-art convolutional architectures for remote sensing image classification.

E. Ablation Study: Overfitting Behavior of Random Forest

To evaluate the impact of feature dimensionality on Random Forest performance, we conducted an ablation study comparing training on full versus PCA-reduced feature sets. Figures illustrate the learning curves of both accuracy and log loss under each configuration. Without dimensionality reduction (Fig12 & Fig13), the model exhibits nearly perfect training accuracy (approaching 100%) and near-zero log loss, indicating overfitting. However, validation accuracy remains low and validation loss high, suggesting poor generalization.

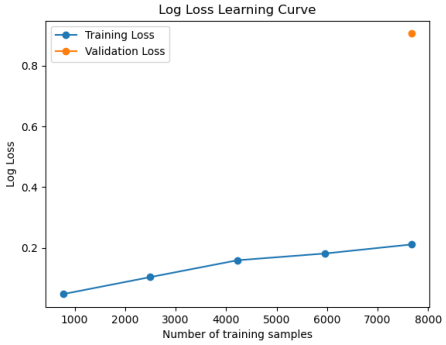

Fig12. Log loss learning curve of Random Forest without remove overfitting

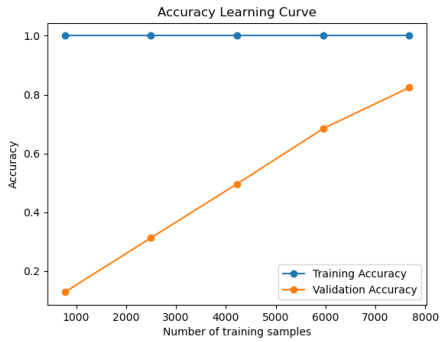

Fig13. Accuracy learning curve of Random Forest without remove overfitting

In contrast, when ResNet features are reduced from 2048 to 256 dimensions via PCA (Fig14 & Fig15), the model achieves better balance: training accuracy drops moderately, but validation accuracy steadily increases and validation loss consistently decreases. The convergence curves are more stable, and the gap between training and validation narrows.

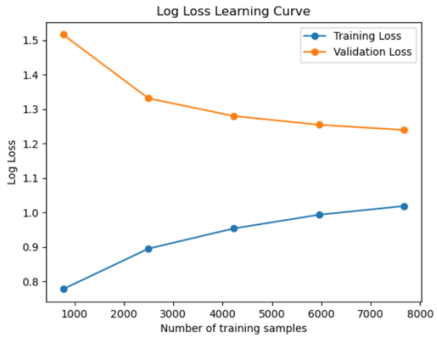

Fig14. Log loss learning curve of Random Forest after remove overfitting

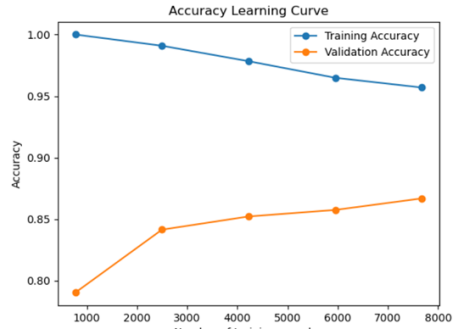

Fig15. Accuracy Learning curve of Random Forest after remove overfitting

These results confirm that dimensionality reduction plays a key role in mitigating overfitting for tree-based models and enhances their robustness for image classification tasks involving high-dimensional visual features.

F. Underfitting Models

We also noticed models that were not performing well in this task. For SVM itself, it relies on too much information from feature input and is limited when using a linear kernel. After we introduced two kinds of feature descriptors, the prediction result is better (36% vs. 45%), but it is still below 50%. One reason may be because the feature descriptor LBP encodes local grayscale texture, which is more struggling to detect structural variations, and for SIFT, it only able to detect a limited number of key points, which would lead to a lack of important information from background in aerial images.

For KNN, without powerful learned features, it uses all dimensions equally, even if many are noisy or uninformative. Moreover, we have 12000 images in sample dataset, which is relatively large for traditional machine learning methods, and the training could be inconsistent in complex images because KNN needs to store all data and compare all of the training samples.

For VGG11, without attention mechanism, VGG cannot perform very well on cluttered or fine-grained scenes, because it treats all regions equally. VGG11 also has fewer layers than ResNet or EfficientNet, limiting its ability to capture deep hierarchical features. The figure 16 is VGG11 confusion matrix, it is not hard to notice that the model did well on first 10 categories and performed not ideally on parking, port, Railway, residential, and river. This is because VGG11 has limited layers and lack of architectural features to handle the similarity of those five categories. While deeper versions such as VGG16 or VGG19 might yield better results, we were unable to test them due to limited computational resources and time constraints.

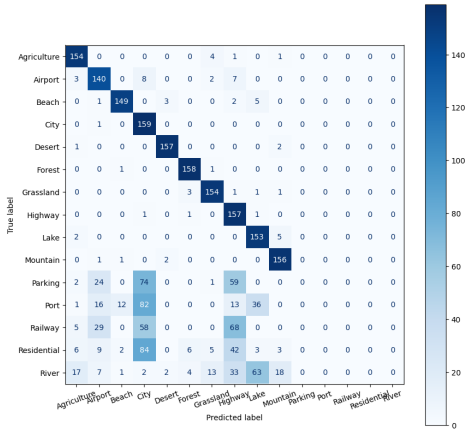

Fig16. VGG11 Confusion Matrix

G. Impact of Implementation Techniques

1) Data enhancement: Data augmentation was applied in MobileNet, ResNet, and EfficientNet models through random horizontal flipping, cropping, and resizing. These strategies significantly improved the model's ability to generalize and prevented overfitting. In contrast, the models without augmentation (e.g., baseline MLP or traditional methods) showed clear signs of performance degradation on the test set.

2) Feature selection: Feature selection played a critical role in models like Random Forest and KNN, where high-dimensional CNN features were reduced using PCA. Although this improved computational efficiency, it sometimes resulted in performance loss. Conversely, MLP and SVM models

benefited greatly from using ResNet50 feature vectors as input rather than raw pixels.

3) attention module: The MLP model implemented an attention module via a learnable channel-wise weighting layer. This mechanism selectively enhanced informative feature dimensions and suppressed irrelevant ones. The accuracy gain observed in the MLP with attention supports the effectiveness of this strategy in compensating for the model's lack of spatial understanding.

VI. Conclusion

In this work, we carried out a comprehensive evaluation of fifteen landscape categories using both traditional machine-learning pipelines and modern deep-learning architectures. Our experiments showed that convolutional neural networks—namely EfficientNetB0, ResNet50 and MobileNetV2—consistently outperformed handcrafted-feature approaches (LBP+SVM, SIFT+SVM, KNN and Random Forest), achieving top-tier accuracies of 97.96%, 97.25% and 96.58%, respectively. The confusion matrices for these CNN models exhibit strong diagonal dominance, even in challenging pairs such as River vs. Lake or Grassland vs. Forest, confirming their robust generalization across fine-grained aerial scenes.

By contrast, pipelines relying solely on texture or keypoint descriptors plateaued below 50% F1-score, highlighting the limitations of manual feature design in complex remote-sensing imagery. Notably, a hybrid strategy—feeding pretrained ResNet50 embeddings into a lightweight MLP—elevated accuracy from approximately 50% to over 94%, demonstrating the practical benefit of combining deep representations with simple classifiers under constrained compute.

Our ablation studies further revealed that random flipping, cropping and resizing substantially mitigate overfitting in deep models, while PCA-based dimensionality reduction aids tree-based methods by narrowing the gap between training and validation performance. Incorporating channel-wise attention in the MLP also improved focus on discriminative features, compensating for its lack of spatial inductive bias.

Looking ahead, addressing class imbalance with resampling or loss re-weighting, exploring vision transformers and multi-modal fusion (e.g. spectral + spatial data), and integrating explainability tools such as Grad-CAM will be crucial. In summary, state-of-the-art deep networks are the most effective solution for large-scale aerial image classification, while carefully engineered hybrids and augmentation techniques remain valuable when resources are limited.

References

[1] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: https://doi.org/10.1109/tpami.2002.1017623.

[2] R. Chourasiya and S. Khaparkar, "Satellite Image Processing Using SVM classifier and ELBP-ML Features," International Journal for Research Trends and Innovation, vol. 6, no. 6, 2021.

[3] Ò. Corominas, I. Riera, S. Aditya, and S. Singh, "Image Classification with Classic and Deep Learning Techniques." Available: https://arxiv.org/pdf/2105.04895

[4] C. L. Fan, "Ground surface structure classification using UAV remote sensing images and machine learning algorithms," Applied Geomatics, vol. 15, no. 4, pp. 919–931, Oct. 2023, doi: https://doi.org/10.1007/s12518-023-00530-x.

[5] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 6, May 2018, doi: https://doi.org/10.1002/widm.1264.

[6] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 157, pp. 155–170, Nov. 2019, doi: https://doi.org/10.1016/j.isprsjprs.2019.09.009.

[7] C. Liu, "Research on image classification leveraging deep convolutional neural networks and visual cognition," Applied and Computational Engineering, vol. 32, no. 1, pp. 200–209, Jan. 2024, doi: https://doi.org/10.54254/2755-2721/32/20230212.

[8] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li, "A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition," Remote Sensing, vol. 15, no. 3, p. 827, Jan. 2023, doi: https://doi.org/10.3390/rs15030827.

[9] B. A. Wijaya, P. J. Gea, A. D. Gea, A. Sembiring, and C. Mitro, "Satellite Images Classification using MobileNet V-2 Algorithm," Sinkron, vol. 8, no. 4, pp. 2316–2326, Oct. 2023, doi: https://doi.org/10.33395/sinkron.v8i4.12949.

[10] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," Machine Learning with Applications, vol. 6, p. 100134, Aug. 2021, doi: https://doi.org/10.1016/j.mlwa.2021.100134.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Jun. 2016, doi: https://doi.org/10.1109/cvpr.2016.90.

[12] H. Dastour and Q. K. Hassan, "A Comparison of Deep Transfer Learning Methods for Land Use and Land Cover Classification," Sustainability, vol. 15, no. 10, pp. 7854–7854, May 2023, doi: https://doi.org/10.3390/su15107854.

[13] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv.org, Apr. 17, 2017. https://arxiv.org/abs/1704.04861

[14] S. Ramadasan, K. Vijayakumar, V. Raju, and S. Prabha, "Automatic Lake and River Detection from Satillite Images Using MobileNet Scheme," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), pp. 1–5, Apr. 2024, doi: https://doi.org/10.1109/icstem61137.2024.10560754.

[15] S. Bobba, "Leveraging Pre-trained Deep Learning Models for Remote Sensing Image Classification: A Case Study with ResNet50 and EfficientNet," American Journal of Science, Engineering and Technology, vol. 9, no. 3, pp. 150–162, Aug. 2024, doi: https://doi.org/10.11648/j.ajset.20240903.11.

[16] C. Zhang et al., "A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 140, pp. 133–144, Jun. 2018, doi: https://doi.org/10.1016/j.isprsjprs.2017.07.014.