Deep Learning Based Approach for Aerial Surveillance System

Rekha Biradar* & Shreya Jere**

DOI: https://doi.org/10.5281/zenodo.21206380

ABSTRACT

Military operations, urban planning, environmental monitoring, and disaster management all benefit from modern aerial observation. In order to identify critical infrastructure, including airports, highways, ports, railroad stations, and defense zones, our work focuses on deep learning-based classification of high-resolution aerial photos. We train and assess three CNN models: MobileNetV2, VGG16, and DenseNet121. VGG16 enhances feature extraction, DenseNet121 increases accuracy by effective feature reuse, and MobileNetV2 offers a lightweight solution for real-time applications. To increase robustness and generalization, data augmentation is used. Accuracy, precision, recall, and F1-score are used to assess model performance. According to experimental results, all models function well, with MobileNetV2 being appropriate for real-time aerial surveillance applications and DenseNet121 offering the highest accuracy.

KEYWORDS: *Aerial Surveillance, Deep Learning, Transfer Learning, Convolutional Neural Networks (CNN), Image Classification.*

Introduction

Intelligent monitoring systems that monitor infrastructure, land-use patterns, environmental changes, and security-sensitive areas rely heavily on aerial surveillance. Large amounts of aerial data are constantly produced due to the quick expansion of satellite photography, UAVs, and drone-based imaging, necessitating the use of automated image processing methods [1,2]. For large-scale real-time applications, traditional manual aerial image interpretation is inefficient and time-consuming. Furthermore, it is difficult to classify aerial scenes due to differences in scale, illumination, and spatial patterns [3].

This study focuses on classifying aerial scenes into multiple classes, including agricultural, port, railway station, desert, city, beach, river, woodland, parking area, residential area, lake, and highway. For precise classification, these datasets' varied urban and natural settings necessitate effective feature extraction techniques [4]. Preprocessing methods like data augmentation, scaling, and standardization are used to enhance classification performance. Because Deep Convolutional Neural Networks (CNNs) efficiently learn spatial and texture-based patterns from images, they are employed for feature extraction [5]. Popular deep learning architectures such as VGGNet [6], ResNet [7], and DenseNet [8] provide strong performance in aerial image classification tasks. The proposed system utilizes transfer learning techniques to improve training efficiency and classification accuracy, especially for moderate-sized datasets [9]. Model optimization is performed using the Adam optimizer [10], and performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.

* **Rekha Biradar,** Assistant Professor, Department of Master of Computer Applications (MCA), SECAB Institute of Engineering and Technology, Vijayapura.

** **Shreya Jere,** Department of Master of Computer Applications (MCA), SECAB Institute of Engineering and Technology, Vijayapura.

Literature Survey

Sl. No.	Author Name	Dataset Used	Method Used	Result
1	Azlan Saleh et al.	UAV datasets	CNN, Transformer, YOLO, Faster R-CNN	Improved accuracy and faster inference DOCX
2	Karen Simonyan et al.	ImageNet, ILSVRC	Deep CNN, Saliency Maps	Competitive localization results DOCX
3	Xu Zeyu et al.	Aerial & satellite image studies	Faster R-CNN, YOLO, ResNet, U-Net	Improved wildlife detection DOCX
4	Wei Chen, Hua & Qili	VisDrone, UAVDT	YOLO, RetinaNet, SSD	Better accuracy using feature fusion DOCX
5	Deng Xinjie et al.	AIDER	Lightweight deep learning model	93.09% Accuracy, 0.939 F1-score DOCX
6	Xiao Zhu et al.	AID, UC-Merced	Fine-tuned CNN	Above 90% Accuracy DOCX
7	Usman Azmat et al.	UAV-Human, Okutama-Action	CNN, GMM	48.60%, 78.01%, 80.03% Accuracy DOCX
8	Teixeira et al.	EuroSAT, Coffee Scenes	CNN, SegNet, U-Net	97–99%, 92–94% Accuracy DOCX
9	Hieu Nguyen et al.	VisDrone, PRAI-1581, UAVDT	Faster R-CNN, SSD, ResNet	28–42% mAP, 80% Rank-1 Accuracy DOCX
10	Carrio et al.	UAV datasets	Faster R-CNN, YOLO	28–65% mAP, 90–95% Accuracy DOCX
11	X. Zhu et al.	Remote sensing datasets	CNN, CNN-LSTM, GAN	Good lightweight performance DOCX
12	Alex Krizhevsky et al.	ImageNet	AlexNet CNN	84.7% Top-5 Accuracy DOCX
13	Ding et al.	DOTA	YOLO, SSD, Faster R-CNN	50–75% mAP DOCX
14	Jan Lohse et al.	Sentinel-1 SAR	GLCM + SVM	80–91% Accuracy DOCX

Methods

Image Preprocessing and Augmentation

The scale, brightness, and direction of aerial photos taken by UAVs and satellites differ. Pixel scaling (I/255) is used to normalize the input after all photos are scaled to 224 × 224 pixels. To enhance model generalization and lessen overfitting, data augmentation methods like rotation, flipping, zooming, and shifting are used.

Transfer Learning-Based Feature Extraction

MobileNetV2, VGG16, and DenseNet121 are examples of pretrained CNN models that apply transfer learning. These models' top layers are eliminated, and the pretrained networks serve as feature extractors to extract texture-based and spatial information from aerial photos. By using this method, classification performance is improved while training time is reduced.

Fine-Tuning Strategy

The initial layers are frozen and the final layers are retrained using the aerial dataset in order to modify the pretrained models for aerial scene classification. Roads, vegetation, water bodies, and metropolitan areas are examples of surveillance-specific properties that the model learns through fine-tuning.

Classification Head Design

Following feature extraction, a unique classification head is applied. It consists of a Dropout layer (0.5) for regularization, a dense layer with ReLU activation, a Global Average Pooling layer, and a SoftMax output layer for multi-class classification.

Training and Optimization

The Adam optimizer is used to train the models with categorical cross-entropy loss. To avoid overfitting and preserve the best-performing model based on validation accuracy, early halting and model checkpointing are used.

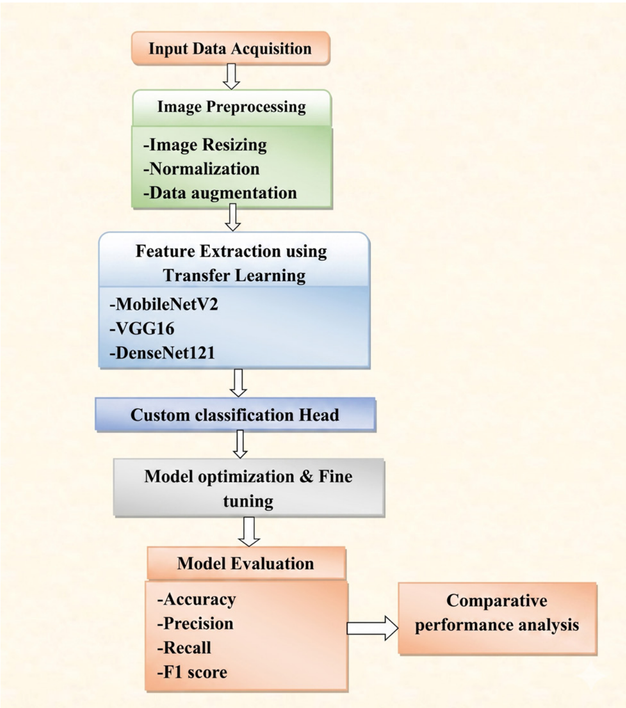
Model Selection and Inference

MobileNetV2, VGG16, and DenseNet121 are evaluated for performance, and the model with the best validation accuracy is chosen. To increase dependability in aerial surveillance applications, predictions with a confidence level less than 0.5 are marked as "Unknown" during inference.

Proposed Methodology

Dataset Description

In this methodology, 15 categories including agriculture, ports, cities, forests, rivers, and highways—are monitored for land use and land cover using multi-class RGB aerial photos. Because scale, lighting, season, and structural patterns vary, the dataset has a significant degree of intra-class variance and inter-class similarities. To make the images CNN compatible, they are shrunk to 224 × 224 pixels, normalized, and arranged into training and testing bins.



The Figure 1 describes the proposed aerial surveillance framework for multi-class aerial scene classification. The system integrates image preprocessing, data augmentation, transfer learning, fine-tuning, and performance evaluation to achieve accurate scene recognition across 15 land-use categories.

Image Preprocessing and Data Augmentation

To maintain consistent processing and transfer learning compatibility, all images are downsized in accordance with the fixed input size of 224 × 224 pixels required for pretrained CNN architectures including VGG16, MobileNetV2, and DenseNet121. The normalization of pixel values from 0–255 to 0–1 enhances training stability, speeds up convergence, and enables effective gradient updates.

$$_{omlzd} = \frac{}{255} \qquad (1)$$

where represents the original pixel intensity value.

To boost dataset diversity and decrease overfitting, data augmentation techniques including rotation, zooming, flipping, and shifting are used during training. By simulating actual aerial differences in scale, angle, and object position, these adjustments enhance the accuracy, resilience, and generalization of the model on previously unseen images.

Transfer Learning-Based Feature Extraction

By eliminating their last layers, pretrained CNN models like DenseNet121, VGG16, and MobileNetV2 are used as feature extractors in transfer learning. In order to improve performance and shorten training times for aerial image classification, deeper layers learn semantic features while earlier layers capture edges and textures.

Fine-Tuning Strategy

Using a fine-tuning technique, pretrained models are adjusted to domain-specific aerial images. In order to maintain generic feature representations, the network's first 30 layers are frozen; the remaining layers are then unfrozen and retrained.

Let denote frozen layers and denote trainable layers:

$$\text{Model} = \{\} + \{\} \qquad (2)$$

This strategy allows the model to specialize in recognizing aerial-specific patterns such as road networks, agricultural textures, coastal boundaries, and urban structures.

Classification Head Design

Each pretrained model is given a unique classification head for aerial picture classification following feature extraction. After reducing spatial dimensions with a Global Average Pooling (GAP) layer, a Dense layer (128 neurons, ReLU) is used to learn task-specific characteristics. A final Dense layer with SoftMax generates class probabilities, while a Dropout layer (0.5) guards against overfitting:

$$P(y_i) = e^{(z_i)} / \left(\sum_{j=1}^{C} e^{(z_j)}\right) \qquad (3)$$

where C is the total number of classes and z_i is the output logits that the network produced prior to activation. The total of the expected probability for every class is guaranteed by this function. The model is optimized using the Adam optimizer, which offers effective and adaptive gradient updates, and has a learning rate of 1×10^{-4}. Categorical Cross-Entropy is the loss function that is employed, and it is defined as:

$$L = -\sum_{(i=1)}^{C} y_i \log(\hat{y}_i) \quad (4)$$

The projected likelihood for class i is denoted by Þ_i, whereas the true label is represented by y_i. By calculating the discrepancy between the actual and predicted distributions, this loss function helps the model minimize classification errors while it is being trained.

Training Strategy and Optimization

The models are trained for 10 epochs with a batch size of 16. To prevent overfitting and ensure optimal model selection, the following techniques are implemented:

- Early Stopping: Training is halted if validation loss does not improve for three consecutive epochs. The best-performing weights are restored automatically.
- Model Checkpointing: The model achieving the highest validation accuracy is saved for final evaluation.

Model Selection and Inference

The performance of MobileNetV2, VGG16, and DenseNet121 is compared based on validation accuracy. The best-performing model is selected for deployment. During inference, a confidence threshold mechanism is applied. If the predicted probability is below 0.5, the output is classified as "Unknown" to improve reliability in real-world aerial surveillance scenarios.

Result and Discussion

Experimental Setup

The suggested system classifies fifteen categories of aerial scenes using CNN models based on transfer learning. Rotation, zooming, flipping, and shifting are used to enhance, normalize, and resize images to 224 × 224. The Adam optimizer and Categorical Cross-Entropy loss are used to train models. To increase accuracy and decrease overfitting, techniques like fine-tuning, dropout, early stopping, and model checkpointing are used.

Performance of Transfer Learning Models

Three pretrained CNN models—MobileNetV2, VGG16, and DenseNet121—were used to assess aerial surveillance classification. DenseNet121 gets the best accuracy with greater feature reuse, VGG16 offers steady feature extraction, and MobileNetV2 is lightweight and appropriate for real-time UAV applications. For airborne surveillance activities, these models work together to balance precision, stability, and efficiency.

Model Generalization and Overfitting Control

Early Stopping, Dropout (0.5), and data augmentation were used as strategies to prevent overfitting. Dropout decreased an excessive dependence on neurons during training, while data augmentation increased dataset diversity and model generalization. The optimal model weights were restored and validation loss was tracked by early stopping. For aerial scene categorization, stable convergence of training and validation curves demonstrated successful fine-tuning using ImageNet pretrained weights.

Prediction and Confidence Analysis

Unseen aerial pictures were used to assess the trained models. Class probability scores were produced by resizing, normalizing, and classifying input photos using the SoftMax layer. Predictions that fell below the 0.5 confidence level were classified as "Unknown". In practical aerial surveillance applications, our method decreases misclassification and increases robustness.

Discussion

According to experimental findings, transfer learning greatly enhances the performance of aerial scene classification. Models learn aerial-specific properties more efficiently as the higher layers are refined. Of the evaluated architectures, MobileNetV2 offered lower computational complexity appropriate for real-time surveillance, but DenseNet121 obtained superior accuracy and generalization. For automated aerial surveillance applications, confidence-based prediction decreased ambiguous classifications, increasing overall system resilience and dependability.

Table 1: Proposed Methodology Models Accuracy

MODELS	ACCURACY
Mobile Net V2	99%
VGG16	35%
Densnet121	96%

MobileNetV2 is the best model for this dataset, with the greatest accuracy of 99%. Effective feature extraction is made possible by its lightweight and effective architecture, which reduces overfitting. Due to its thick connections, which improve feature reuse and learning, DenseNet121 also performed well, achieving 96% accuracy; however, the tiny performance difference may be explained by its higher complexity. However, VGG16 only obtained 35% accuracy, perhaps due to its high data requirements and numerous parameters, which restricted its capacity to adapt to the dataset.

Proposed methodology confusion matrix

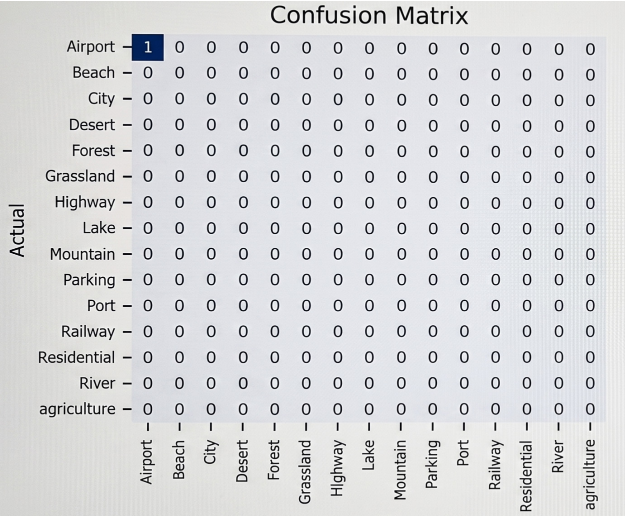


Fig.2. Confusion Matrix of MobileNet Model for Aerial Surveillance

The model's performance on the UCM land-use dataset was assessed using a confusion matrix, in which rows stand for real classes and columns for anticipated classes. While other values show incorrect classifications, the diagonal values show accurate classifications. In this instance, the image is accurately and flawlessly categorized as "Airport". Additionally, the confusion matrix assists in identifying often confused classes, like Residential and City or River and Lake.

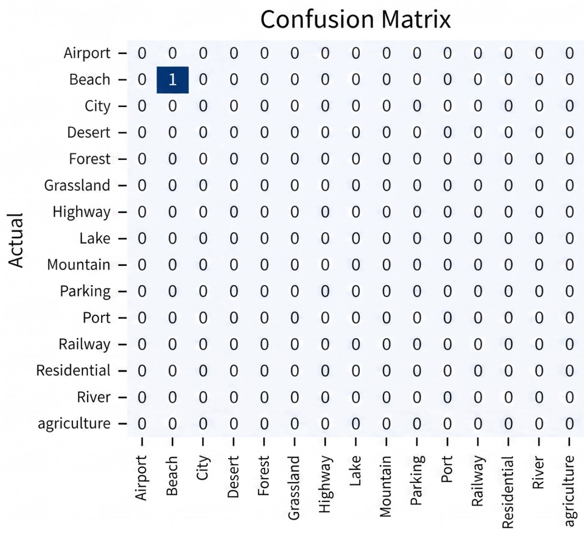
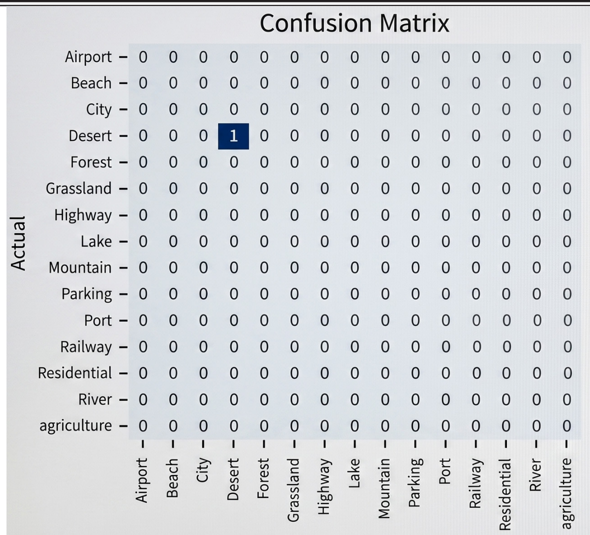


Fig.3. Confusion Matrix of VGG16 Model

The confusion matrix shows the 15 land-use groups classified by the VGG16 model. Rows exhibit real labels and columns display expected labels when the diagonal is accurately predicted. The Beach–Beach diagonal value confirms precise and devoid of errors in classification. In addition to F1-score, recall, accuracy, and precision, it also helps to detect misunderstanding between similar classes.



(Fig.4. Confusion Matrix of DenseNet Model)

The DenseNet model categorizes aerial photos into 15 land-use groups, as shown in the confusion matrix. With accurate predictions on the diagonal and errors off-diagonal, rows show real labels and columns show expected labels. There is no misclassification, as confirmed by the Beach–Beach diagonal value. Accuracy, precision, recall, and F1-score should all be taken into account for a thorough assessment, as it aids in assessing confusion between similar classes on bigger datasets.

Conclusion

There is a trade-off between accuracy, efficiency, and computing cost when analyzing CNN architectures for aerial surveillance image classification. Because of its extensive feature connections, DenseNet was able to discern intricate patterns in aerial photos with the best accuracy. Using lightweight depthwise separable convolutions, MobileNet achieved a speed-accuracy balance that made it appropriate for real-time and resource-constrained applications. Because of its deeper architecture, VGG16 required more processing power yet provided stable and dependable performance. In general, the model selection is based on the needs of the application: VGG16 for dependable performance with enough hardware resources, MobileNet for lightweight real-time applications, and DenseNet for optimal accuracy. For effective airborne surveillance systems, future research can concentrate on increasing accuracy while lowering processing complexity.

References:

1. X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8–36, 2017. Doi: 10.1109/MGRS.2017.2762307
2. G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," Proceedings of the IEEE, vol. 105, no. 10, pp. 1865–1883, 2017. Doi: 10.1109/JPROC.2017.2675998
3. L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," IEEE Geoscience and Remote Sensing Magazine, vol. 8, no. 2, pp. 22–40, 2020. Doi: 10.1109/MGRS.2020.2979780
4. G. S. Xia et al., "AID: A benchmark dataset for performance evaluation of aerial scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3965– 3981, 2017. Doi: 10.1109/TGRS.2017.2685945
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, 2012. Doi: 10.1145/3065386
6. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations (ICLR), 2015. doi: 10.48550/arXiv.1409.1556
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of CVPR, 2016. Doi: 10.1109/CVPR.2016.90
8. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," Proceedings of CVPR, 2017. Doi: 10.1109/CVPR.2017.243
9. S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010. Doi: 10.1109/TKDE.2009.191
10. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations (ICLR), 2015. Doi: 10.48550/arXiv.1412.6980
11. A. Saleh, M. A. Zulkifley, H. H. Harun, F. Gaudreault, I. Davison, and M. Spraggon, "Forest fire surveillance systems: A review of deep learning methods," Heliyon, vol. 10, no. 1, p. e23127, Dec. 2023, Doi: 10.1016/j.heliyon. 2023.e23127.
12. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in Proc. Workshop at Int. Conf. Learning Representations (ICLR), 2014
13. Z. Xu, T. Wang, A. K. Skidmore, and R. Lamprey, "A review of deep learning techniques for detecting animals in aerial and satellite images," Int. J. Appl. Earth Obs. Geoinformation, vol. 128, p. 103732, Apr. 2024, Doi: 10.1016/j.jag.2024.103732.
14. W. Hua and Q. Chen, "Comprehensive survey on small object detection in aerial images," [Journal/Conference], 2024
15. X. Deng et al., "Lightweight deep learning architecture for UAV-based disaster image classification," [Journal/Conference], 2024,
16. X. Zhu, L. Mou, D. Tuia, G. S. Xia, L. Zhang, F. Xu, and F. Frauendorfer, "Deep-learning-based remote sensing image scene understanding: A survey," [Journal/Conference], 2024,
17. U. Azmat, S. S. Alotaibi, N. Al Mudawi, B. I. Alabduallah, M. Alonazi, A. Jalal, and J. Park, "A novel framework for human action recognition in aerial surveillance using elliptical modelling," [Journal/Conference], 2025,
18. R. Teixeira et al., "Deep learning approaches for crop classification using aerial imagery," [Journal/Conference], 2024,
19. T. Nguyen et al., "Human-centric aerial surveillance: Detection, tracking, identification and behavior recognition," [Journal/Conference], 2025,
20. S. Carrio, M. Sampedro, J. Rodriguez-Ramos, and R. Campoy, "Deep learning methods for Unmanned Aerial Vehicles (UAVs): Perception, planning, control and surveillance," [Journal/Conference], 2017,
21. X. Zhu, L. Mou, D. Tuia, G.-S. Xia, L. Zhang, F. Xu, and F. Frauendorfer, "Deep learning in remote sensing: A data-intensive perspective," [Journal/Conference], 2024,
22. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. NIPS, 2012, pp. 1097–1105,
23. J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," [Journal/Conference], 2021.
24. A. Lohse et al., "Effects of incident angle on texture features for sea ice classification in SAR imagery," [Journal/Conference], 202*,