International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 07 | Jul 2026 www.irjet.net p-ISSN: 2395-0072

Ensemble Learning Approaches for Aerial Crop Classification Using Deep CNNs, Vision Transformers and Machine Learning Models

[1]MS. URVI KORE, [2]Dr. HIREN DAND, [3]Dr. ANIL CHHANGANI

[1]*Research Scholar, J.J.T. University, Rajasthan, India*
[2]*Principal, Anna Leela College of Commerce and Economics, Mumbai, Maharashtra, India*
[3]*Associate Professor, Dept. of Computer Engineering, Thadomal Engineering college, Mumbai, Maharashtra, India*

---***---

Abstract: *The precise classification of aerial crop imagery is a crucial part of precision agriculture and smart farming technology. Despite the good performance shown by single deep learning model approaches, their predicted outputs are constrained by model-specific limitations in terms of feature extraction and inherent biases. This paper introduces and examines the effectiveness of three ensemble learning techniques for categorization of aerial crop images captured by UAVs. (1) The first ensemble, composed of a custom Convolutional Neural Network (CNN), VGG16 and InceptionV3, performs a soft voting between the models. (2) The second ensemble model consists of a Vision Transformer (ViT), VGG16 and InceptionV3 and employs a probability average as a combination method. (3) The third ensemble is a machine learning based hard voting ensemble made up of XGBoost, Random Forest and Support Vector Machine classifiers. The experimental outcomes demonstrate that the ensemble method based on the custom CNN yields the most accurate classification with an accuracy of 97%, which is followed by the ViT-based ensemble that achieved a classification of 96% whereas the machine learning ensemble was evaluated at 61%. Our findings suggest that ensemble multiple deep learning architectures can substantially enhance classification performance when compared to single deep learning models or existing machine learning methods.*

Key Words: UAV, Ensemble Learning, CNN, Vision Transformer, VGG16, InceptionV3, Precision Agriculture, Crop Classification

1. INTRODUCTION

Advanced and using technology the farming is revolutionizing the way agriculture is practiced. It aims to increase production and efficient use of agricultural resources and sustainable crop management. From the various sensors, UAVs, or unmanned aerial vehicles, are of great interest because they can provide high-quality aerial images used for automatic monitoring and classification of crops [1], [2]. The effectiveness of deep learning together with the use of UAV imagery has also improved the ability to identify crops by enabling the machine to learn the discriminative visual feature from aerial images [3]. Various types of convolutional neural networks, including VGG16 and InceptionV3, as well as transformer architectures have proven to be of high efficacy in crop image analysis due to a method of knowledge transfer [4]– [7].

In our earlier work, we proposed a lightweight Custom Convolutional Neural Network (CNN) for drone-based aerial crop image categorization, which achieved a classification accuracy of 92% and outperformed conventional machine learning approaches including Support Vector Machine (SVM), XGBoost, and AlexNet on the KARC dataset [13]. Subsequently, we extended this work by investigating hybrid deep learning and machine learning models, where deep features extracted using pretrained CNN architectures such as VGG16, EfficientNetB7, and InceptionV3 were classified using SVM, Random Forest, and XGBoost classifiers. The proposed hybrid framework achieved a maximum classification accuracy of 95%, demonstrating the effectiveness of combining transfer learning with classical machine learning techniques [14].

Despite the positive outcomes of these studies, the utilization one deep learning algorithm or CNN based hybrid model could still leave them restricted and lacking in generalization. Ensemble learning becomes one of the well-known techniques for better classification output when combining predictions from different types of classifiers. Voting methods build are capable of minimizing prediction instability and achieving robustness by taking advantage of various strengths of the implemented algorithms [8]– [10]. In this research, three different types of ensemble techniques for aerial crop classification based on UAVs were assessed based on the same KARC dataset. The first one comprises an ensemble of a Custom CNN, VGG16, and InceptionV3 using soft voting. The second ensemble is a combination of the Vision Transformer (ViT), VGG16, and InceptionV3 using probability averaging. Lastly, a machine learning ensemble that functions using hard voting comprised of XGBoost, Random Forest, and Support Vector Machine is introduced for comparisons. It was found that CNN based ensemble produces the best accuracy of 97%, while ViT based ensemble results in 96%, and the machine learning ensemble produces 61% accuracy.

2. LITERATURE REVIEW

Identification of agriculture crops using aerial and satellite images has become an important research area with the rapid growth of UAV and remote sensing technologies. Several review studies have summarized recent developments in this field. Teixeira et al. reviewed different deep learning models used for crop classification. They reported that Convolutional Neural Networks (CNNs) have largely replaced traditional methods. Because of CNN's higher accuracy and better feature learning capabilities [1]. Similarly, Zualkernan et al. discussed various machine learning techniques for precision agriculture using UAV image. Highlighted challenges of balancing model accuracy with the limitations of drone-based image acquisition [2]. Bouguettaya et al. also presented a comprehensive review of deep learning techniques for crop classification from UAV imagery. Showed how CNN architectures have evolved to achieve better performance in agricultural applications [3]. Many crop classification methods are based on well-known CNN architectures developed for image recognition. VGG, proposed by Simonyan and Zisserman, demonstrated that using deeper convolutional networks with small filters can significantly improve image classification performance. Because of its strong feature extraction capability, VGG16 is widely used as a transfer learning model for agricultural image classification [4]. Another popular architecture is InceptionV3, introduced by Szegedy et al., which uses multiple convolution filters of different sizes within the same layer to learn features at different scales. This makes it particularly effective for aerial images where crop size and appearance vary with altitude and camera angle [5].

Recently, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs. Instead of using convolution operations, ViTs use self-attention mechanisms to capture relationships between different regions of an image. Dosovitskiy et al. showed that Vision Transformers can achieve performance comparable to or better than CNNs on large image datasets [6]. In agricultural applications, Wang et al. proposed CropFormer, a transformer-based model that achieved strong performance across different crop classification tasks by effectively learning both local and global image features [7]. These studies indicate that transformer models provide complementary information to CNNs and can further improve image classification performance.

Ensemble learning has become an effective approach for improving classification accuracy by combining the predictions of multiple models. Kuncheva explained that ensemble methods generally outperform individual classifiers because different models make different types of prediction errors [8]. Polikar also emphasized that classifier diversity is one of the key factors contributing to the success of ensemble learning methods [9]. Zhou further described popular ensemble techniques such as bagging, boosting, and voting, and explained how these methods improve model robustness and generalization [10]. These studies provide the theoretical foundation for voting-based ensemble methods used in this work.

Several researchers have successfully applied ensemble learning to agricultural image classification. Kalita et al. developed an ensemble of deep convolutional neural networks using multi-filter and multi-scale architectures for aerial crop classification. Their results showed that combining multiple CNN models achieved better accuracy than using a single CNN model [11]. Similarly, Pandey and Jain proposed a dense convolutional neural network for UAV-based crop identification and demonstrated improved classification performance using dense feature learning [12]. Our previous research has also contributed to this area. In our earlier work, we developed a Custom CNN model for drone-based aerial crop image classification and compared its performance with AlexNet and traditional machine learning algorithms, achieving a classification accuracy of 92% [13]. In a subsequent study, we combined pretrained CNN models with machine learning classifiers such as Support Vector Machine (SVM), Random Forest (RF), and XGBoost. The proposed hybrid framework achieved a maximum accuracy of 95%, demonstrating the effectiveness of combining deep feature extraction with classical machine learning techniques [14].

Apart from these studies, several researchers have investigated different CNN architectures for image classification. Dhamala and Acharya compared multiple CNN models and analyzed how network depth and architecture influence classification performance [15]. Sardeshmukh et al. applied CNN models to crop image classification and confirmed that deep learning methods consistently outperform traditional image processing techniques [16]. Kavitha et al. proposed a crop classification method based on spherical contact distribution using remote sensing images, providing an alternative statistical approach to agricultural image analysis [17]. Gawale et al. introduced a lightweight CNN model for satellite image classification that requires fewer computational resources while maintaining competitive accuracy, making it suitable for edge computing applications [18]. Furthermore, Malgi et al. presented a UAV-based precision agriculture system that integrates drone imaging with intelligent image analysis, demonstrating the practical deployment of deep learning models in real agricultural environments [19].

The literature shows two major research trends. The first focuses on developing more powerful deep learning architectures, including CNNs and Vision Transformers, for accurate crop classification. The second focuses on ensemble learning, where multiple classifiers are combined to improve robustness and prediction accuracy. Motivated by these developments, this work proposes and compares three ensemble learning strategies for UAV-based aerial crop classification: a CNN-based ensemble, a Vision Transformer-based ensemble, and a machine learning voting ensemble.

3. METHODOLOGY

This study constitutes the third phase in our research aimed at aerial vegetation classification models based on

UAV imagery. The first phase was to developed a custom CNN architecture optimized for classification. The second one pursued the approach of hybrid CNN+ML learning models, supported by transfer learning techniques. In this research paper, we report on the development of the ensemble model, which integrates the deep learning and ML methodologies. Thus, we have achieved a possibility of a complete analysis of individual models, hybrid methods, and ensemble procedures relying on the same aerial crop imagery dataset.

The experiment was carried out using the Kolkata Agricultural Region Crop dataset. This dataset consists of high-resolution aerial images from UAVs featuring twelve crop types. These images were captured in various environmental conditions such as crop growth stages, lighting, and field backgrounds. The images of crop types were divided into train and test set to make sure that evaluation of performance does not bias results.

All aerial crop images were resized to 128 × 128 pixels and normalized by rescaling pixel values to the range [0,1]. Data augmentation was performed using the TensorFlow ImageDataGenerator to improve model generalization and reduce overfitting.

Model 1: CNN-Based Ensemble (Custom CNN + VGG16 + InceptionV3)

The architecture is based on the integration of a light custom-made CNN with two transfer learning pre-trained networks. VGG16 and InceptionV3 are used for enhancing aerial crop classification. The custom CNN model involves two convolutional layers with max-pooling 32 and 64 filters. The flattened fully connected layer comprising 128 ReLU neurons and a softmax layer with twelve classes.128x128 RGB input image dimensions, categorical cross-entropy loss, and Adam optimization are used. VGG16 and InceptionV3 are pre-trained with ImageNet weights, with their base convolutional layers fixed. The heads swapped out for a new one which consists of a flatten layer, a dense layer with 128 neurons, and a softmax layer with twelve classes. The proposed design allows the custom CNN to be capable of learning specific to the dataset. And complemented by hierarchical textures and multi-scale spatial features provided by VGG16 and InceptionV3, respectively. The predictions produced by the three models are aggregated through the soft voting technique. Determine the class with the maximal class probability, thus mitigating the effect of misclassification caused by a single model.

Model 2: Vision Transformer-Based Ensemble (ViT + VGG16 + InceptionV3)

The second model pairs a Vision Transformer (ViT-B16) with the same VGG16 and InceptionV3 networks. This model test whether transformer-based attention can complement convolutional feature extraction. CNNs excel at capturing local spatial detail, ViT uses self-attention to model long-range dependencies and global context across the image. Vit offers a different but complementary representation. The pretrained ViT-B16 has its classification head removed and replaced with a flatten layer, a 256-neuron dense layer, and a twelve-class softmax output. The same fine-tuning approach used for Vit as VGG16 and InceptionV3. As in Model 1, all three networks' convolutional or transformer bases stay frozen while new classifier heads are trained on the crop dataset. Predictions are again fused through soft voting. Averaging the probability distributions from ViT, VGG16, and InceptionV3 to select the class with the highest mean score. This blend of global attention-based context and local convolutional detail helps the ensemble capture richer visual information from UAV imagery. The model achieved 96% accuracy, slightly below the CNN-based ensemble, suggesting transformer integration is competitive but not quite superior on this particular dataset.

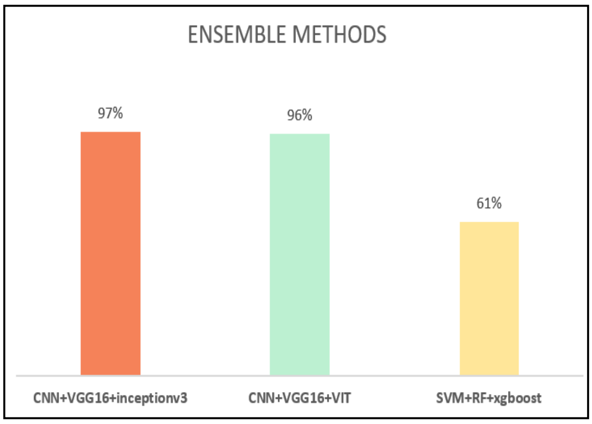
Model 3: Machine Learning-Based Ensemble (XGBoost + Random Forest + SVM)

The third model combining three classical machine learning classifiers, XGBoost, Random Forest, and Support Vector Machine. It evaluates traditional ensemble methods on extracted image features rather than raw pixels. Each algorithm brings a distinct learning strategy. XGBoost's gradient boosting captures complex nonlinear patterns. Random Forest's bagging approach improves stability by averaging many decision trees. SVM finds optimal separating boundaries in high-dimensional feature space. The three classifiers are trained independently on the same dataset. Each algorithm is tuned with its own hyperparameters before being merged using Scikit-learn's VotingClassifier. Rather than averaging probabilities, this ensemble uses hard voting, where each model casts a single class prediction and the majority label wins. If at least two classifiers agree on a crop category, that category becomes the final output for the image. This majority-rule approach leverages the diversity among boosting, bagging, and margin-based methods to reduce the risk of any one classifier's bias skewing results. This model offers a computationally lighter alternative to the deep learning ensembles, relying on feature-based classification rather than automatic hierarchical feature learning.

4. RESULTS AND DISCUSSION

Chart-1. presents the accuracy trends of CNN, VGG16, InceptionV3, and ViT over 10 training epochs. The CNN model showed a gradual increase in performance, reaching approximately 81% accuracy at the final epoch. InceptionV3 achieved better results, attaining around 91% accuracy with stable convergence throughout training. VGG16 consistently outperformed CNN and InceptionV3, reaching nearly 96% accuracy by the end of the training process. Among all individual models, ViT demonstrated the highest performance, rapidly converging and achieving close to 99% accuracy. As illustrated in Figure 2, ensemble learning further enhanced classification performance. The CNN + VGG16 + InceptionV3 ensemble achieved the highest accuracy of 97%, followed by CNN + VGG16 + ViT with 96% accuracy. The traditional SVM + RF + XGBoost ensemble obtained only 61% accuracy, indicating the superiority of deep learning-based ensemble approaches.

Chart-1. Accuracy comparison of CNN, VGG16, InceptionV3, and Vision Transformer (ViT) models



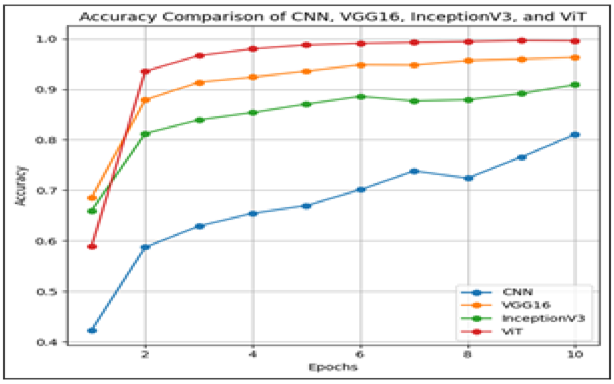


Fig.-1. Accuracy comparison of different ensemble methods

The results indicate that deep learning models significantly outperform traditional machine learning methods for the given classification task. ViT demonstrated superior feature extraction and learning efficiency, resulting in the highest individual model accuracy. VGG16 also exhibited strong and consistent performance due to its deep convolutional architecture, while InceptionV3 achieved competitive accuracy with efficient feature representation. CNN provided acceptable results but lacked the discriminative capability of the more advanced architectures. Interestingly, the ensemble of CNN, VGG16, and InceptionV3 slightly exceeded the individual performance of VGG16 by combining complementary features from multiple models. These findings suggest that carefully designed deep learning ensembles can improve classification reliability and predictive performance.

5. CONCLUSIONS

The comparative analysis shows that advanced deep learning architectures achieve higher classification accuracy than conventional machine learning ensemble techniques. Among the individual models, ViT delivered the best overall performance with nearly 99% accuracy, followed by VGG16 and InceptionV3. CNN achieved the lowest accuracy among the deep learning models. But it contributed useful complementary features in ensemble learning. The CNN + VGG16 + InceptionV3 ensemble produced the best overall result with 97% accuracy, demonstrating the effectiveness of combining multiple deep learning models. The traditional SVM, Random Forest, and XGBoost ensemble showed considerably lower performance. The results highlight that deep learning-based ensemble approaches provide a robust and accurate solution for complex image classification tasks.

REFERENCES

[1] I. Teixeira, R. Morais, J. J. Sousa, and A. Cunha, "Deep Learning Models for the Classification of Crops in Aerial Imagery: A Review," Agriculture, vol. 13, no. 5, p. 965, 2023.

[2] D. A. Zualkernan, M. H. Hussain, J. Khan, and M. ElMohandes, "Machine Learning for Precision Agriculture Using Imagery from Unmanned Aerial Vehicles (UAVs): A Survey," Drones, vol. 7, no. 6, p. 382, 2023.

[3] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Deep learning techniques to classify agricultural crops through UAV imagery: A review," Neural Computing and Applications, vol. 34, pp. 9511–9536, 2022.

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[6] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[7] H. Wang et al., "CropFormer: A generalized deep learning classification approach for multi-scenario crop classification," Frontiers in Plant Science, vol. 14, p. 1130659, 2023.

[8] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Hoboken, NJ, USA: Wiley, 2004.

[9] R. Polikar, "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–45, 2006.

[10] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.

[11] I. Kalita, G. P. Singh, and M. Roy, "Crop classification using aerial images by analyzing an ensemble of DCNNs under multi-filter & multi-scale framework," Multimedia Tools and Applications, vol. 82, pp. 18409–18433, 2023.

[12] A. Pandey and K. Jain, "An intelligent system for crop identification and classification from UAV images using conjugated dense convolutional neural network," Computers and Electronics in Agriculture, vol. 192, p. 106543, 2022.

[13] U. Kore, H. Dand, and A. Chhangani, "Deep Learning Classifiers for Drone-based Aerial Crop Image Categorization," JJTU Journal of Renewable Energy Exchange, vol. 13, no. 1, pp. 230–236, 2025.

[14] U. Kore, H. Dand, and A. Chhangani, "Aerial Vegetable Crop Classification Using CNN Feature Extraction and Machine Learning," International Journal of Creative Research Thoughts (IJCRT), vol. 14, no. 6, 2026.

[15] N. Dhamala and K. P. Acharya, "Classification of images using CNN model and its variants," International Journal of Computer Applications, vol. 10, no. 7, pp. 1062–1068, 2023.

[16] M. M. Sardeshmukh, M. Chakkaravarthy, S. Shinde, and D. Chakkaravarthy, "Crop image classification using convolutional neural network," Multidisciplinary Science Journal, vol. 5, no. 4, 2023, doi: 10.31893/multiscience.2023039.

[17] Kavitha, A. Srikrishna, and C. Satyanarayana, "Crop image classification using spherical contact distributions from remote sensing images," Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 2, pp. 534–545, 2022, doi: 10.1016/j.jksuci.2019.02.008.

[18] R. D. Gawale, N. S. Patil, V. B. Lokhande, and A. Khandare, "A lightweight CNN architecture for land classification on satellite images," International Research Journal of Modernization in Engineering Technology and Science, vol. 6, no. 4, Apr. 2024, doi: 10.56726/IRJMETS36987.

[19] G. D. Malgi, D. Tejaswani, D. Sharma, K. Singh, and Jyothi, "Precision agriculture drone equipped with image analysis," JETIR, vol. 11, no. 5, pp. 277–282, May 2024.