Deep Learning Enhanced Hand Gesture Recognition for Efficient Drone use in Agriculture

Phaitoon Srinil[1], Pattharaporn Thongnim*[2]
Applied Artificial Intelligence and Smart Technology, Faculty of Science and Arts,
Burapha University, Chanthaburi, Thailand[1]
Statistics, Department of Mathematics, Faculty of Science, Burapha University, Chonburi, Thailand[2]

***Abstract*—The use of deep learning in unmanned aerial vehicles (UAVs), or drones, has greatly improved various technologies by making complex tasks easier, faster, and requiring less human help. This study looks into how artificial intelligence (AI) can be used in farming, especially through creating a system where drones can be controlled by hand gestures to support agricultural activities. By using a special type of AI called a Convolutional Neural Network (CNN) with an EfficientNet B3 model, this research developed a gesture recognition system. It was trained on 1,393 pictures of different hand signals taken under various light conditions and from three different people. The system was evaluated based on its training and testing performance, showing very high scores in terms of loss, accuracy, F1 score, and the Area Under the Curve (AUC), which means it can recognize gestures accurately and work well in different situations. This has big implications for farming, as it gives farmers an easy way to control drones for tasks like checking on crops and spraying them precisely, which also helps keep them safe. This study is an important step towards smarter farming practices. Moreover, the system's ability to perform well in different settings shows it could also be useful in other areas like construction, where drones need to operate precisely and flexibly.**

***Keywords*—Deep learning; Convolutional Neural Network; hand gesture recognition; drone; agriculture**

I. Introduction

Drones, also known as unmanned aerial vehicles (UAVs), have moved beyond their military beginnings to become essential tools in many industries, not just for recreation. Drones are used in many areas, such as security, defense, farming, energy, insurance, and water management [1]. This variety shows how drones are and their potential to improve traditional methods. Drones can reach difficult area, carry out detailed aerial survey, and provide immediate data, improving decision making and operational efficiency in many fields. The growing popularity of drones is driven by continuous technological improvement, making them more user friendly and effective for both professional and personal use [2]. Technological advancements in drone capabilities have significantly broadened their applications, enabling them to contribute to environmental monitoring, search and rescue operations, and infrastructure inspection, among others. Innovations such as increased autonomy through AI integration [3], extended battery life, and enhanced payload capacities allow drones to perform complex tasks more efficiently and reliably. For instance, in agriculture, drones equipped with advanced sensors can monitor crop health [4], optimize water usage [5], and manage resources more sustainably [6]. Similarly, in emergency response, drones provide invaluable assistance in locating victims and assessing damage in disaster stricken areas, demonstrating their critical role in saving lives and managing crises [7].

The integration of drones into agriculture is poised to enhance crop health monitoring, reduce environmental impact, protect farmer health and increase the efficiency of farming operations. Farmers are progressively turning to drones to oversee their crops and enhance precision agriculture practices, a trend that is expected to significantly fuel the growth of the drone market in agriculture over the next decade. These drones have the ability to monitor vast fields, capture intricate images, and provide data that is not affected by cloud cover, offering a clear advantage over traditional monitoring methods. As drone technology continues to advance, becoming more efficient and cost-effective, their adoption in agriculture is set to increase [8]. Therefore, these developments promise to revolutionize farming by improving yield predictions, optimizing resource use, and enabling more precise application of water, fertilizers, and pesticides [9].

In the context of agriculture, drones equipped with Artificial Intelligence (AI) extend their utility beyond monitoring and analysis to include actionable interventions, such as precise spraying. Spraying drones leverage AI to optimize the application of pesticides, herbicides, and fertilizers. They can autonomously navigate over fields, applying substances directly where needed and in the correct amounts, protect farmer health, significantly reducing waste and environmental impact. This targeted approach ensures that crops receive the exact treatment they require, enhancing growth conditions and potentially increasing yield efficiency. The combination of drones and deep learning is transforming how tasks are performed and redefining the possibilities for innovation and efficiency in global industries [10]. Transitioning to the development of a hand gesture recognition system, this technology further amplifies efficiency in agriculture by enabling farmers to control drones and other automated equipment.

Before reach into the development of a hand gesture recognition system, it is important to understand the context in which such technology could be particularly beneficial in agriculture. Farmers often face the challenge of applying spray fertilizers to their crops at various times, depending on the crop's growth stage, weather conditions, and the type of fertilizer being used. The timing and amount of fertilizer application are critical to ensure optimal crop health and yield. Traditional methods can be imprecise and labor intensive, requiring manual labor to cover large areas and sometimes

*Corresponding authors

leading to uneven distribution of the fertilizer. The implementation of a wide array of intuitive and easy to perform gestures requires a user centric design approach. This involves conducting extensive user research to identify natural and comfortable gestures for different commands, considering both ergonomic principles and cultural differences. By ensuring that the gestures are easily performable by a broad spectrum of users, including those with physical disabilities, the system becomes more inclusive and user friendly.

To enhance the efficiency of hand gesture control, some approaches include the wearable device or IoT device controller placed on the back of the hand to intend hand motion and control the UAV with hand gesture recognition [11] [12]. Multi modal control is another technique for overcoming UAV gesture control. A multi modal control system integrates multiple interactions, such as hand gestures, eye movements, and voice interactions [13]. The multi channel joint interaction promotes high UAV control efficiency. It is crucial to use advanced technologies with deep learning. By combining computer, and cameras, the system's capability to capture and understand gestures in varied lighting will be greatly improved.

In this study, the focus primarily on hand gesture control, as it has shown promising results in previous research [12] [13]. Therefore, this study aims to create a hand gesture recognition system that works effectively in different environmental settings and individuals, such as under direct sunlight, on cloudy days, and shady. The given model first detects the hand and then draws the hand skeleton. Next, the model is generated by using the detected hand as a training set for a deep convolutional neural network. These technologies are excellent at picking up slight movements, which is essential for the system to tell apart purposeful gestures from accidental ones. Moreover, applying machine learning algorithms and deep learning to process the data from the study will enhance the system's precision and flexibility. This will allow it to accurately recognize a broad array of gestures.

II. Method

A. Data Collection and Preprocessing

A dataset consisting of 1,393 images was compiled around a farm in Thailand, capturing both indoor and outdoor settings. This collection aims to advance posture trajectory analysis and includes shots taken under a variety of lighting conditions; sunlight, cloudy, and in shade. Participation from three individuals ensured a wide range of imagery. The dataset features eight specific gesture types: ascending, descending, pitch forward, pitch backward, roll left, roll right, yaw left, and yaw right. Each contributor supplied images for every gesture, photographed under three distinct lighting scenarios. Cameras were employed to take these pictures, which were then stored in JPG format. During the image preprocessing phase, the sizes of the collected images were standardized to a uniform dimension of 300 × 300 pixels. These images were divided into eight classes: ascending, descending, pitch forward, pitch backward, roll left, roll right, yaw left, and yaw right, as shown in Fig. 1 Subsequently, the dataset underwent a division into training, validation and testing sets, allocating 880 images for training purposes, 320 images for validation, and the remaining 192 images for testing.

B. The Proposed Model

The proposed model presents a sophisticated hand gesture recognition model designed to enhance the operational efficiency of drones in agricultural settings. This innovation is made possible through the integration of a Convolutional Neural Network (CNN) with an EfficientNet B3 architecture, tailored to interpret various hand signals under diverse environmental conditions.

The integration of MediaPipe, Hand Landmark, TensorFlow, Keras in TensorFlow, and the EfficientNet B3 model within this method provides a robust framework for accurate hand gesture recognition tailored for drone control in agricultural applications. MediaPipe offers a real-time, efficient hand tracking solution, utilizing the Hand Landmark model to precisely identify the positions of key points on the hand, essential for recognizing complex gestures. TensorFlow serves as the backbone for deep learning operations, enabling scalable and efficient model training and execution. By leveraging Keras, a high-level API within TensorFlow, the process of building and training deep learning models is simplified, making it more accessible while maintaining flexibility and performance. The choice of the Convolutional Neural Network (CNN) architecture, specifically EfficientNet B3, is strategic for its ability to handle image data effectively, utilizing compound scaling to optimize accuracy and computational efficiency. This combination of technologies and models ensures the system's ability to accurately interpret a wide range of hand gestures under various environmental conditions, making it a powerful tool for enhancing drone operations in agriculture.

1) MediaPipe: Numerous deep learning frameworks and libraries are available for hand gesture recognition, among which MediaPipe stands out. MediaPipe is a framework tailored for the deployment of deep learning solutions ready for production [14], [15]. It facilitates the construction of pipelines necessary for performing inference on various types of sensory data. Moreover, MediaPipe supports the publication of code alongside research efforts and aids in the development of technological prototypes. As an open-source tool, it is accessible to developers worldwide and supports a wide range of platforms, ensuring its versatility and broad applicability. Its lightweight nature enhances its performance and ease of integration into various software and hardware environments, making it a preferred choice for real-time applications.

2) Hand landmark: In the MediaPipe framework (see Fig. 2), the hand is modeled using 21 distinct 3D landmarks to represent the joints and tips of the fingers [16]. For each finger, there are four landmarks: the Carpometacarpal (CMC) joint is marked as Landmark 1 for the thumb, followed by the Metacarpophalangeal (MCP) joint as Landmark 2, the Interphalangeal (IP) joint as Landmark 3, and the fingertip as Landmark 4. This pattern is consistent across the hand, with the MCP joint for the index, middle, ring, and pinky fingers designated as Landmarks 5, 9, 13, and 17, respectively. The Proximal Interphalangeal (PIP) and Distal Interphalangeal (DIP) joints follow in sequence for each finger, culminating with the fingertip, or Landmark 8 for the index, Landmark 12 for the middle, Landmark 16 for the ring, and Landmark 20 for the pinky finger, providing a comprehensive mapping of the hand's articulations for gesture recognition.

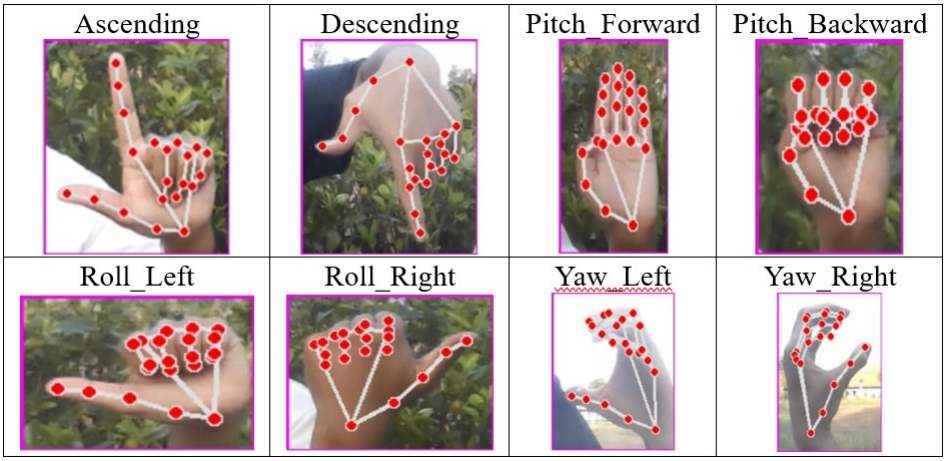
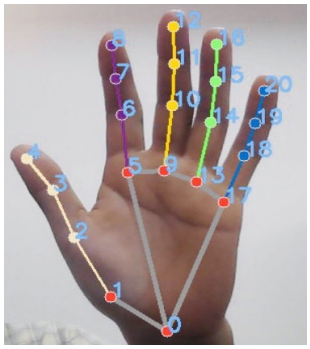


Fig. 1. Effects of selecting different switching under dynamic condition.

The model predicts the (x, y, z) coordinates of these landmarks in the image, with: x and y representing the landmark's position on the plane of the image, and z indicating the landmark's relative depth from the camera. To put the trained model into action, it is integrated into an OpenCV-based workflow that handles real-time data processing. This involves using MediaPipe to detect and track hand landmarks in each data stream. The information about the landmarks is then input into the TensorFlow model, which determines what gesture is being made.

3) TensorFlow: The trained model is then implemented within an OpenCV pipeline to process data sets in real time. As the data stream flows through the pipeline, MediaPipe extracts the hand landmarks from each data, and these are instantly passed to the TensorFlow model for gesture prediction [17].



Wrist (0)
Thumb: CMC (1), MCP (2), IP (3), and Tip (4).
Index Finger: MCP (5), PIP (6), DIP (7), and Tip (8).
Middle Finger: MCP (9), PIP (0), DIP (11), and Tip (12).
Ring Finger: MCP (13), PIP (14), DIP (15), and Tip (16).
Pinky (Little Finger): MCP (17), PIP (18), DIP (19), and Tip (20).

Fig. 2. The 21 landmarks (0-20) of hand gestures in MediaPipe.

This seamless process allows for the recognition of gestures as they occur, enabling real-time interaction. The system can be further tailored to recognize a wide array of gestures, enhancing its utility in various applications.

In TensorFlow, computations are represented as graphs, where nodes in the graph represent mathematical operations, and the edges represent the tensors that flow between these operations. The core concept of TensorFlow can be encapsulated in how it handles these tensors and performs operations on them [18]. The concept of Gradient Descent is implemented through optimizers that automatically adjust the model's parameters (weights and biases) to minimize the loss function:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla_\theta J(\theta),$$

where, θ represents the model parameters, $J(\theta)$ is the loss function, α is the learning rate, $\nabla_\theta J(\theta)$ is the gradient of the loss function with respect to the parameters. TensorFlow abstracts and simplifies the implementation of gradient descent, making it accessible and flexible for optimizing a wide variety of models. By adjusting the model parameters (weights and biases), optimizers improve the model's accuracy over time.

Neural networks, including those built with TensorFlow, rely heavily on linear algebra [19]. One fundamental operation is matrix multiplication, used in fully connected layers:

$$Y = XW + b,$$

where, X represents the input matrix, W is the weights matrix, b is the bias vector, and Y is the output matrix. Linear algebra operations in TensorFlow are used behind the scenes in training machine learning models, especially in operations like forward and backward propagation in neural networks, where weights and inputs are represented as matrices and vectors. Operations such as convolution in CNNs can also be understood in terms of linear algebra.

4) Keras in TensorFlow: TensorFlow provides a comprehensive, scalable platform for building and deploying machine learning models, with Keras serving as the high-level interface that simplifies model development through its focus on ease

of use and modularity. The combination of TensorFlow's scalability and Keras's user-friendliness makes it an excellent toolkit for both beginners and experts in machine learning. Integrating Keras directly into TensorFlow as tf.keras offers a streamlined workflow for designing and training machine learning models with TensorFlow's robust capabilities for scaling and deployment. This integration provides a high-level, user-friendly API for TensorFlow, without sacrificing flexibility and performance [20].

Therefore, the research is defined a neural network architecture using TensorFlow. This could be a Convolutional Neural Network (CNN) for processing image data or a custom model suited for sequential data like time-series of landmarks. The model is trained on the preprocessed hand landmark data, using labeled gestures to teach the model the corresponding gesture for each set of landmarks.

C. Convolutional Neural Network and EfficientNet B3 Model

The framework is notable for its collection of pre-trained machine learning models, which serve as a foundation for advanced applications in computer vision and augmented reality. Among its offerings are highly accurate face detection algorithms that can identify and track multiple hand in real time. Convolutional Neural Networks (CNNs) are at the heart of image recognition and processing tasks [21], [22]. In the context of using CNNs for recognizing hand gestures, a key operation is the convolution, applied to the input image using filters or kernels to extract features:

$$G[j,k] = \sum_m \sum_n F[m,n] \cdot H[j-m, k-n],$$

where G is the output feature map, F is the input image, H is the filter/kernel, j, k are indices in the output feature map, and m, n are indices in the filter/kernel.

In the process of applying convolution operations within CNNs, an input image or feature map from a previous layer, denoted as $F[m,n]$ undergoes a transformation through a convolutional filter, $H[j-m, k-n]$. This filter, a small matrix, traverses the input, focusing on extracting specific features by learning relevant patterns during the model's training phase. The convolution between the input image and the filter results in an output feature map, represented by $G[j,k]$, where each element signifies the convolution operation's output at distinct locations across the input. This output encapsulates the detected features, such as edges or textures, effectively capturing the input's essential characteristics for further processing or classification tasks, like hand gesture recognition, where the input can range from grayscale to color (RGB) images. Therefore, in hand gesture recognition, convolution allows the model to learn to identify key features of hand gestures. This capability is crucial for accurately classifying different gestures based on visual input.

Moreover, efficientNet B3 is part of the EfficientNet family, which is a group of Convolutional Neural Network (CNN) models designed for efficient performance [23], [24], [25]. The EfficientNet models use a systematic approach to scaling called compound scaling, which uniformly scales the network depth, width, and resolution with a set of fixed scaling coefficients. This approach is different from traditional scaling methods that independently scale these dimensions, often leading to suboptimal performance.

The compound scaling method used in EfficientNet involves scaling the network's depth, width, and resolution with a compound coefficient ϕ, according to the following formulas: $d = \alpha^\phi$, $w = \beta^\phi$ and $r = \gamma^\phi$ where depth (d) is the number of layers in the network, width (w) is the number of channels in the layers, resolution (r) is the size of the input image, α, β, and γ are constants that determine the scaling of depth, width, and resolution, respectively and ϕ is the compound coefficient that controls the overall resource increase of the network. Higher values of ϕ result in larger, potentially more accurate networks. The idea is to find a balance between depth, width, and resolution that leads to the best performance improvement for a given increase in model size and computational cost.

Therefore, incorporating the EfficientNet B3 architecture into the study of performance metrics for deep learning in hand gesture recognition models further illustrates the model's advanced technical capabilities and its practical utility in augmenting drone operations for agricultural purposes. EfficientNet B3 is part of the EfficientNet family, which represents a series of Convolutional Neural Network (CNN) architectures designed to provide higher accuracy with fewer parameters than previous models, making them both powerful and efficient. The use of EfficientNet B3 in the hand gesture recognition model capitalizes on its ability to scale model size in a more balanced and effective manner, optimizing for accuracy, latency, and resource utilization.

D. Evaluation Metrics for Hand Gesture Recognition Model

An evaluation of the hand gesture recognition model across eight distinct sign classes was conducted, employing metrics such as precision, recall, and the F1-score for a comprehensive analysis, detailed as follows:

Precision, also referred to as the positive predictive value, is determined by the following formula:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall, measured as the percentage of correctly predicted instances out of all actual instances of the class, is given by the equation:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

The F1-score, also known as the F-measure, encapsulates the harmonic mean of precision and recall, thereby reflecting their equilibrium. Improvement in the F1-score is observed only with simultaneous increases in both precision and recall. This score spans from 0 to 1, with values closer to 1 denoting greater accuracy in classification. The formula for calculating the F1-score is as follows:

$$F1\text{-}score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is quantified as the ratio of accurate predictions to the total number of predictions made. The calculation for accuracy is represented by the following formula:

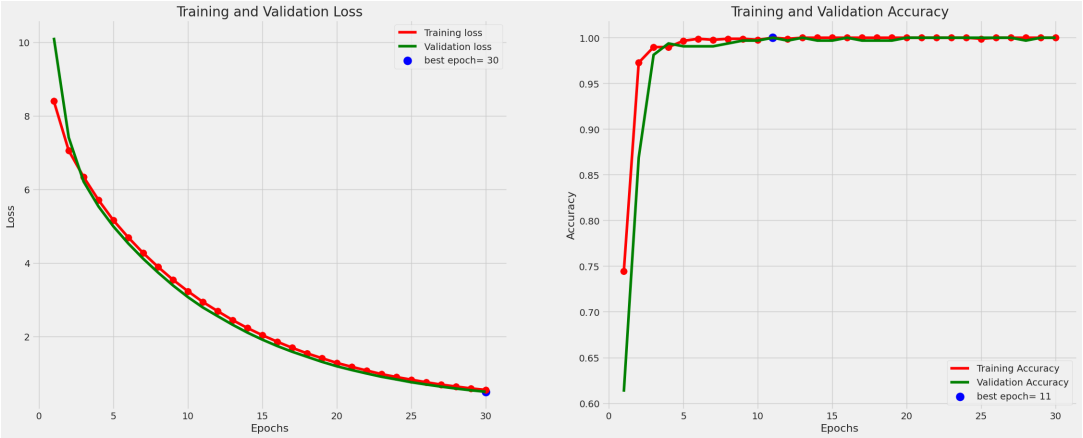


Fig. 3. Training and validation curves for hand gesture recognition model: The left graph displays validation loss and the right graph displays validation accuracy.

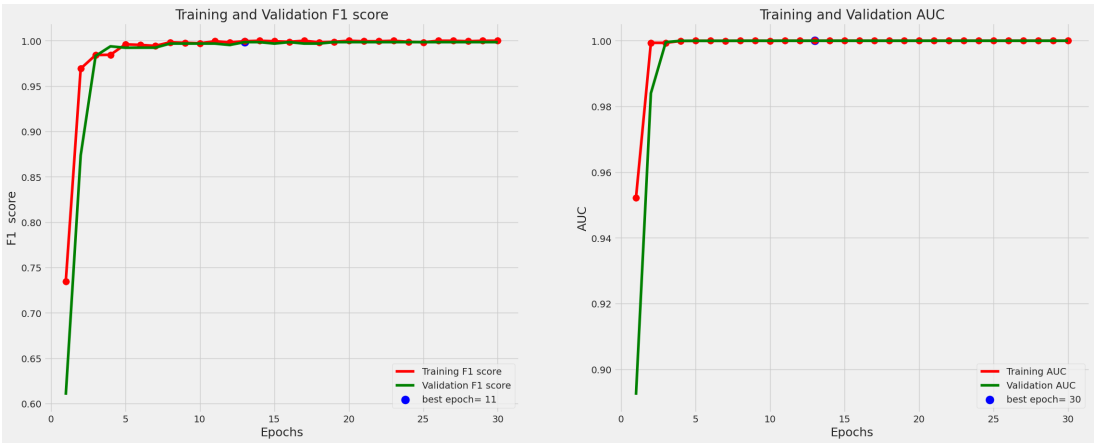


Fig. 4. Training and validation curves for hand gesture recognition model: The left graph displays validation F1 score and the right graph displays validation AUC.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

To effectively use these metrics, it is important to have a well-defined test dataset that accurately represents the real-world scenarios in which the model will be deployed. Comparing these metrics after integrating EfficientNet B3 can also provide insights into how this architecture improves the model's performance.

III. Results

Fig. 3 shows two plots side by side, on the left is the Training and Validation Loss, and on the right is the Training and Validation Accuracy over 30 epochs of efficientNet B3 model training. The left plot indicates that both training and validation loss decrease sharply initially and then level off, converging to a low value, with the best epoch marked at 30. On the right plot, the accuracy of both training and validation rapidly increases and plateaus close to 1.0, indicating high effectiveness of the model, with the best epoch for accuracy marked at 11. These plots suggest that the model quickly learned the task and achieved a stable and high performance early in the training process, with minimal overfitting as indicated by the close convergence of training and validation lines.

Fig. 4 showcases two performance metric plots for a machine learning model over the course of 30 training epochs. On the left is the Training and Validation F1 Score plot, which represents the harmonic mean of precision and recall. The plot shows both training and validation F1 scores quickly converging to a value close to 1.0, indicating excellent model performance with a peak F1 score at epoch 11. This suggests that the model maintains a balanced precision-recall relationship and is neither overfitting nor underfitting.

On the right is the Training and Validation AUC (Area Under the ROC Curve) plot, which is used to evaluate the performance of a binary classification system. The AUC values are consistently high and also converge to a score near 1, with the best AUC score achieved at epoch 30. This high AUC value indicates a high degree of separability, meaning the model is very capable of distinguishing between classes. The close proximity of the training and validation lines in both plots suggests that the model is generalizing well to unseen data.

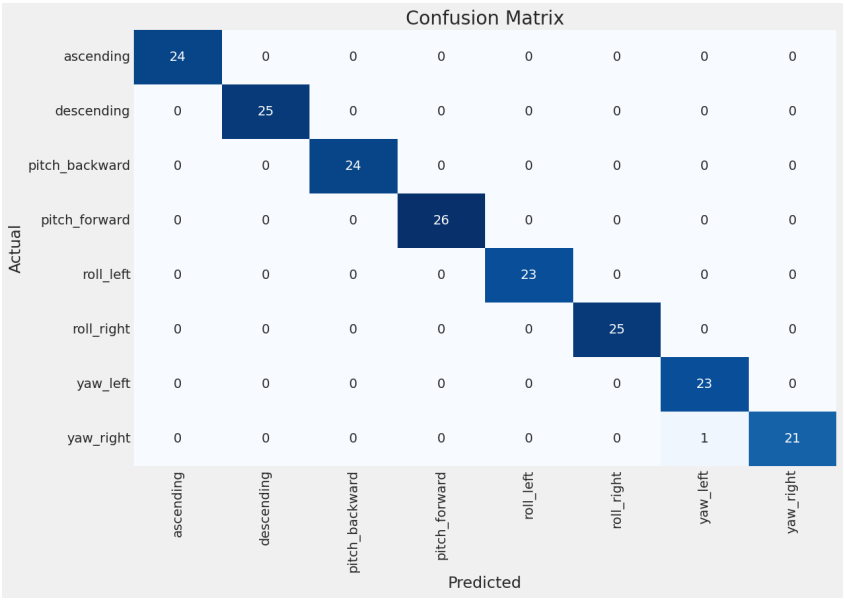


Fig. 5. Confusion matrix depicting the performance of the hand gesture recognition model for drone control in agricultural applications.

Fig. 5 shows visualizes the confusion matrix of a hand gesture recognition model used for controlling drones in an agricultural setting. It is structured with actual gestures along the y-axis and predicted gestures along the x-axis. The matrix contains eight different hand gestures: ascending, descending, pitch backward, pitch forward, roll left, roll right, yaw left, and yaw right. The diagonal from the top left to bottom right represents instances where the predicted gesture matches the actual gesture, signifying a correct prediction by the model. The numbers within these diagonal cells, 24 for ascending, 25 for descending, 24 for pitch backward, 26 for pitch forward, 23 for roll left, 25 for roll right, 23 for yaw left, and 21 for yaw right, indicate a high rate of accurate classifications for each respective gesture. Non-diagonal cells would show misclassifications, but in this matrix, almost all non-diagonal cells are zero, demonstrating that there are very few errors made by the model. Notably, there is only one misclassification observed, where a gesture was actually yaw right but was predicted by the model as yaw left. This could be attributed to the potential similarity in the appearance of these two gestures to the model. Overall, the high count of True Positives (TP) and the sparse misclassifications underscore the model's robustness and reliability in interpreting hand gestures for drone operation under the tested conditions.

The data-driven approach, employing a Convolutional Neural Network (CNN) with EfficientNet B3 architecture, confirms its suitability for visual tasks such as hand gesture recognition. The EfficientNet B3 model's performance signifies that its application in the agricultural domain, controlling drones via hand gestures, can be both feasible and effective. This holds promise for increasing operational efficiency and the democratization of technology use in the field, allowing for more intuitive and natural human-machine interaction without the need for complex controllers or extensive training.

IV. Discussion

The research presented herein marks a notable advancement in leveraging artificial intelligence, particularly convolutional neural networks (CNNs) with EfficientNet B3 architecture, for hand gesture recognition aimed at drone control in agriculture. This integration showcases a substantial leap in precision, robustness, and dependability in gesture recognition technology, as demonstrated by superior performance metrics including loss accuracy, F1 score, and Area Under the Curve (AUC). Such achievements signal the potential for transformative enhancements in agricultural methodologies, optimizing operational efficiency and elevating safety standards.

The exceptional performance of the hand gesture recognition system is rooted in meticulous dataset preparation, encompassing a diverse array of lighting conditions and subjects, in conjunction with deploying the EfficientNet B3 model within the CNN framework [26]. The scalability and efficiency inherent to this model were instrumental in achieving a balanced and effective learning process, thereby facilitating the system's ability to recognize gestures with high accuracy under varying environmental conditions and across different individuals [27]. Moreover, by integrating EfficientNet B3, the hand gesture recognition model achieves superior performance in recognizing and interpreting complex hand gestures, translating them into precise commands for drone control.

Practical implications of this advancement are manifold, primarily offering a simplified and intuitive means for farmers to control drones, thus circumventing the complexities associated with traditional control mechanisms. This innovation significantly diminishes the learning curve associated with drone technology, making it more accessible and user friendly for agricultural applications [28]. Incorporating hand gesture recognition into agricultural drone operations could revolutionize crop monitoring processes, enable precise application of pesticides and fertilizers, and reduce the reliance on manual labor. Moreover, this technology promises to enhance safety

by reducing human exposure to potentially harmful chemicals and facilitating crop inspection in otherwise inaccessible areas [29]. Traditional gesture recognition systems often faced difficulties when used in poor lighting or with subjects that moved quickly [30]. This study overcomes these challenges by utilizing advanced image processing methods and machine learning algorithms. These enhancements improve the system's ability to recognize gestures in a variety of lighting situations and from different viewpoints, making it more flexible and dependable.

Despite the promising outcomes, this study acknowledges certain limitations. The dataset's diversity, while extensive, was limited to images from three individuals. Augmenting the dataset with a broader spectrum of gesture variations from a more diverse demographic could significantly improve the model's generalizability and performance in real-world settings. Moreover, the controlled environment of the study may not fully capture the complexity and unpredictability of actual agricultural environments, where factors like fluctuating lighting conditions, background clutter, and weather variations could impact system performance.

Looking towards the future, the integration of these intuitive drone control systems with artificial intelligence and data analytics heralds a new era of precision agriculture. Future research could focus on developing fully autonomous drones capable of real-time monitoring and management of crops, pest control, and targeted nutrient application, thus optimizing crop health and yield. Additionally, exploring the synergy between drones and other technological innovations in agriculture, such as robotic ground vehicles and sensor networks, could lead to the creation of comprehensive, interconnected farm management systems. This could revolutionize agricultural practices, making them more efficient, sustainable, and tailored to specific environmental and crop needs, thereby supporting global human and food security challenges.

V. Conclusion

This study represents a significant advancement in the application of artificial intelligence (AI) within the realm of agriculture, showcasing a system that leverages Convolutional Neural Networks (CNNs), specifically the EfficientNet B3 model, for the purpose of hand gesture recognition to control drones. The system's training involved a dataset of 1,393 images featuring diverse hand signals captured under various lighting conditions and from three distinct individuals, demonstrating its robust ability to accurately interpret gestures with high performance metrics such as loss, accuracy, F1 score, and Area Under the Curve (AUC). This breakthrough provides a tangible solution to enhancing agricultural productivity and safety by enabling farmers to effortlessly manage drones for critical tasks through intuitive hand gestures. The successful application of hand gesture model in agriculture demonstrates the potential for its adoption in construction scenarios where drones can operate in more structured environments. In addition, the precision of the hand gesture recognition system will be crucial for ensuring accurate delivery of materials, especially in high or hard-to-reach areas. Future research will aim to enhance the model's robustness against the diverse environmental conditions typically found on agricultural sites with farms. This would entail further data collection and model training to ensure the system can accurately interpret hand gestures even in less than ideal conditions. Additionally, integrating the use of drones for spraying will be explored, potentially enabling precise and efficient delivery of substances in various farming scenarios.

References

[1] K. Natarajan, T.-H. D. Nguyen, and M. Mete, "Hand gesture controlled drones: An open source library," in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*. IEEE, 2018, pp. 168–175, doi: 10.1109/ICDIS.2018.00035.

[2] B. Latif, N. Buckley, and E. L. Secco, "Hand gesture and human-drone interaction," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2022, pp. 299–308.

[3] B. Hu and J. Wang, "Deep learning based hand gesture recognition and uav flight controls," *International Journal of Automation and Computing*, vol. 17, no. 1, pp. 17–29, 2020, doi: 10.1007/s11633-019-1194-7.

[4] S. A. Shah, G. M. Lakho, H. A. Keerio, M. N. Sattar, G. Hussain, M. Mehdi, R. B. Vistro, E. A. Mahmoud, and H. O. Elansary, "Application of drone surveillance for advance agriculture monitoring by android application using convolution neural network," *Agronomy*, vol. 13, no. 7, p. 1764, 2023, doi: 10.3390/agronomy13071764.

[5] B. S. Acharya, M. Bhandari, F. Bandini, A. Pizarro, M. Perks, D. R. Joshi, S. Wang, T. Dogwiler, R. L. Ray, G. Kharel *et al.*, "Unmanned aerial vehicles in hydrology and water management: Applications, challenges, and perspectives," *Water Resources Research*, vol. 57, no. 11, p. e2021WR029925, 2021, doi: 10.1029/2021WR029925.

[6] P. Thongnim, V. Yuvanatemiya, E. Charoenwanit, and P. Srinil, "Design and testing of spraying drones on durian farms," in *2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*. IEEE, 2023, pp. 1–6, doi :10.1109/ITC-CSCC58803.2023.10212524.

[7] E. Lin-Greenberg, "Wargame of drones: remotely piloted aircraft and crisis escalation," *Journal of Conflict Resolution*, vol. 66, no. 10, pp. 1737–1765, 2022, doi: 10.1177/00220027221106960.

[8] C.-J. Chen, Y.-Y. Huang, Y.-S. Li, Y.-C. Chen, C.-Y. Chang, and Y.-M. Huang, "Identification of fruit tree pests with deep learning on embedded drone to achieve accurate pesticide spraying," *IEEE Access*, vol. 9, pp. 21 986–21 997, 2021, doi: 10.1109/ACCESS.2021.3056082.

[9] P. Thongnim, V. Yuvanatemiya, and P. Srinil, "Smart agriculture: Transforming agriculture with technology," in *Asia Simulation Conference*. Springer, 2023, pp. 362–376, doi : 10.1007/978-981-99-7240-129.

[10] A. T. Meshram, A. V. Vanalkar, K. B. Kalambe, and A. M. Badar, "Pesticide spraying robot for precision agriculture: A categorical literature review and future trends," *Journal of Field Robotics*, vol. 39, no. 2, pp. 153–171, 2022, doi: 10.1002/rob.22043.

[11] S. S. Y, K. Y. W, and K. Y. G, "Hand gesture-based wearable human-drone interface for intuitive movement control," in *2019 IEEE International Conference on Consumer Electronics, ICCE 2019 Article 8662106 (2019 IEEE International Conference on Consumer Electronics, ICCE 2019)*. IEEE, 2019.

[12] W. Lee, J and H. Yu, K, "Wearable drone controller: Machine learning-based hand gesture recognition and vibrotactile feedback," *Sensors*, vol. 23, no. 5, p. 2666, 2023.

[13] A. Zhou, L. Han, and Y. Meng, "Multimodal control of uav based on gesture, eye movement and voice interaction," in *Advances in Guidance, Navigation and Control (ICGNC 2022)*. Springer, 2023, pp. 3765–3774.

[14] J. Bora, S. Dehingia, A. Boruah, A. A. Chetia, and D. Gogoi, "Real-time assamese sign language recognition using mediapipe and deep learning," *Procedia Computer Science*, vol. 218, pp. 1384–1393, 2023, doi: 10.1016/j.procs.2023.01.117.

[15] M. Peral, A. Sanfeliu, and A. Garrell, "Efficient hand gesture recognition for human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 272–10 279, 2022, doi: 10.1109/LRA.2022.3193251.

[16] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020, doi: 10.48550/arXiv.2006.10214.

[17] D. Someshwar, D. Bhanushali, V. Chaudhari, and S. Nadkarni, "Implementation of virtual assistant with sign language using deep learning and tensorflow," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, pp. 595–600, doi : 10.1109/ICIRCA48905.2020.9183179.

[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[19] B. Ramsundar and R. B. Zadeh, *TensorFlow for deep learning: from linear regression to reinforcement learning*. " O'Reilly Media, Inc.", 2018.

[20] N. K. Manaswi and N. K. Manaswi, "Understanding and working with keras," *Deep learning with applications using Python: Chatbots and face, object, and speech recognition with TensorFlow and Keras*, pp. 31–43, 2018, doi : 10.1007/978-1-4842-3516-42.

[21] G. Elliott, K. Meehan, and J. Hyndman, "Using cnn and tensorflow to recognise 'signal for help'hand gestures," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2021, pp. 0515–521, doi : 10.1109/UEMCON53757.2021.9666484.

[22] R. Patel, J. Dhakad, K. Desai, T. Gupta, and S. Correia, "Hand gesture recognition system using convolutional neural networks," in *2018 4th international conference on computing communication and automation (ICCCA)*. IEEE, 2018, pp. 1–6.

[23] S. Alquzi, H. Alhichri, and Y. Bazi, "Detection of covid-19 using efficientnet-b3 cnn and chest computed tomography images," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 1*. Springer, 2022, pp. 365–373.

[24] S. Abd El-Ghany, M. Elmogy, and A. A. El-Aziz, "Computer-aided diagnosis system for blood diseases using efficientnet-b3 based on a dynamic learning algorithm," *Diagnostics*, vol. 13, no. 3, p. 404, 2023.

[25] A. A. Nafea, M. S. Ibrahim, M. M. Shwaysh, K. Abdul-Kadhim, H. R. Almamoori, and M. M. AL-Ani, "A deep learning algorithm for lung cancer detection using efficientnet-b3," *Wasit Journal of Computer and Mathematics Science*, vol. 2, no. 4, pp. 68–76, 2023.

[26] M. Islam, M. Aloraini, S. Aladhadh, S. Habib, A. Khan, A. Alabdulatif, and T. M. Alanazi, "Toward a vision-based intelligent system: A stacked encoded deep learning framework for sign language recognition," *Sensors*, vol. 23, no. 22, p. 9068, 2023.

[27] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[28] V. Moysiadis, D. Katikaridis, L. Benos, P. Busato, A. Anagnostis, D. Kateris, S. Pearson, and D. Bochtis, "An integrated real-time hand gesture recognition framework for human–robot interaction in agriculture," *Applied Sciences*, vol. 12, no. 16, p. 8160, 2022.

[29] A. Anagnostis, L. Benos, D. Tsaopoulos, A. Tagarakis, N. Tsolakis, and D. Bochtis, "Human activity recognition through recurrent neural networks for human–robot interaction in agriculture," *Applied Sciences*, vol. 11, no. 5, p. 2188, 2021.

[30] V. A. Shanthakumar, C. Peng, J. Hansberger, L. Cao, S. Meacham, and V. Blakely, "Design and evaluation of a hand gesture recognition approach for real-time interactions," *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 17 707–17 730, 2020.