J. Agric. Mech., 11 (3):45-65(2026)

Journal of
Agricultural Mechanization

Online ISSN: 2717-4107
Journal homepage: https://jam.tabrizu.ac.ir/?lang=en

Original Article

Evaluation and Comparison of Deep Learning-based Computer Vision Models for Diagnosing Phenological Stages and Health Status of Phalaenopsis Orchids

Ghazaleh Safaralizadeh Harisi [1], Hossein Bakhoda [2] *, Mohammad Ghahdrijani [2]

1-Ph.D student in Biosystems Mechanics, Department of Agricultural Intelligent Systems Engineering, Faculty of Agriculture- Environment and Natural Resources, Science and Research Branch, Islamic Azad University, Tehran, Iran.

2-Assistant Professor, Department of Agricultural Mechanization, Faculty of Agriculture- Environment and Natural Resources,Science and Research branch, Islamic Azad University, Tehran, Iran.

ARTICLE INFO

Keywords:

Phalaenopsis Orchid,

Deep Learning,

Computer Vision,

Vision Transformer (ViT),

YOLO,

Convolutional Neural Network (CNN),

Tissue Culture

Received:
2026/06/21

Revised:
2026/09/19

Accepted:
2026/09/20

***Corresponding Author:**

hossein.bakhoda@gmail.com

ABSTRACT

Phalaenopsis orchid is one of the most important commercially cultivated ornamental plants worldwide, and accurate monitoring of its developmental stages and health status can contribute to improved production management and efficiency. This study aimed to evaluate and compare the performance of three deep-learning architectures, namely YOLOv8, convolutional neural network (CNN), and Vision Transformer (ViT), for analyzing the developmental stages and health status of Phalaenopsis orchids. The study focused on developing an image-based approach that could support objective monitoring of orchid development under tissue-culture conditions and provide a basis for future automation of plant monitoring processes.

Introduction

The expansion of tissue-culture techniques and large-scale orchid production has increased the need for rapid and objective monitoring of plant growth and health. Conventional visual assessment is time-consuming and may be affected by human error and differences in assessor experience. In addition, the morphological characteristics of orchids can vary considerably during different developmental stages, making accurate identification and evaluation more challenging. Computer vision and deep learning can provide non-destructive image-based analysis and may support automated monitoring systems. Deep-learning models have demonstrated considerable potential for plant recognition, classification, object detection, and health assessment. However, their performance may vary according to image complexity, sample size, class distribution, and the specific characteristics of the target structures. Therefore, evaluating different architectures under the same experimental conditions can provide a better understanding of their practical capabilities. In this study, YOLOv8 was considered for object detection and localization, while CNN and ViT were evaluated for image classification across five developmental stages. The comparison was intended to identify the strengths and limitations of each approach and to determine how differences in image characteristics and data distribution affected model performance.

Materials and Methods

A dataset consisting of 127 RGB images of Phalaenopsis orchids under tissue-culture conditions was prepared and organized according to five developmental stages. These stages included undifferentiated embryos, differentiated embryos, proliferation of differentiated embryos, opened embryos containing protocorm-like bodies (PLBs) and leaves, and complete plantlets. The images were preprocessed and divided into training and validation subsets. The

How to cite:
Safaralizadeh Harisi, G., Bakhoda, H., and Ghahdrijani, M. (2026). *Evaluation and Comparison of Deep Learning-based Computer Vision Models for Diagnosing Phenological Stages and Health Status of Phalaenopsis Orchids.* Journal of Agricultural Mechanization, 11(3):45-65. https://doi.org/10.22034/jam.2026.73148.1365

This is an open-access article under the CC BY NC license
(https://creativecommons.org/licenses/by-nc/4.0/)

preprocessing procedure included organizing the images according to developmental stage, preparing the required dataset structure, and generating the annotation files required for model training. For YOLOv8, target structures, including leaves and protocorm-like bodies (PLBs), were annotated using bounding boxes. The annotations were converted into the normalized format required for YOLO training. CNN and ViT were used for image classification, with Stage4 involving the more detailed classification of healthy and unhealthy PLBs and leaves. This Stagerepresented a more challenging classification problem because of greater visual complexity and differences in class distribution. A ResNet18-based CNN with 224 × 224 pixel inputs was trained using the Adam optimizer, a learning rate of 0.001, batch size of 8, and 20 epochs. ViT-Base/16 with 224 × 224 pixel inputs was trained using Adam, a learning rate of 1×10^{-4}, batch size of 8, and 50 epochs. YOLOv8n was trained with 640 × 640 pixel inputs, batch size of 4, and 100 epochs on the CPU. Classification performance was assessed using Accuracy, Precision, Recall, F1-score, and training and validation Loss. YOLOv8 performance was evaluated using Precision, Recall, mAP@0.5, and mAP@0.5:0.95.

Results and Discussion

The results showed that CNN achieved 100% validation accuracy in Stages 1, 2, 3, and 5, whereas its accuracy decreased to 33.3% in Stage4, where visual complexity and class imbalance were greater. This substantial reduction indicates that high classification accuracy in relatively simple stages does not necessarily imply consistent performance under more complex conditions. The result also highlights the importance of considering class distribution and image characteristics when interpreting classification performance. ViT also showed high performance in Stages 1–3 but experienced a substantial decrease in Stage4. Although the architecture is capable of learning complex visual relationships and global image features, its performance was affected by the limited amount of available data and uneven class distribution. Therefore, the results suggest that advanced model architectures cannot completely compensate for limitations in dataset size and quality. YOLOv8 demonstrated variable detection performance across the five stages, with its highest performance in Stage3, where Precision, Recall, mAP@0.5, and mAP@0.5:0.95 reached 99.1%, 100%, 99.5%, and 79.1%, respectively. These findings demonstrate the capability of YOLOv8 to identify and localize the target structures effectively under suitable image conditions. Its object-detection capability also provides information that is complementary to the image-level classification performed by CNN and ViT. Overall, the comparison indicates that model performance was influenced by sample size, class distribution, and image complexity rather than by model architecture alone. The different responses of CNN, ViT, and YOLOv8 across the developmental stages demonstrate that no single model should necessarily be considered optimal for all aspects of orchid monitoring. The combination of object localization and classification can provide a more comprehensive framework for automated analysis. The absence of an independent test set and the limited number of images mean that the reported results should be interpreted within the scope of the available validation data and should not be considered definitive evidence of generalization. In particular, the very high validation accuracies observed in some stages should be interpreted cautiously because a limited dataset may not adequately represent the variability encountered in practical production environments.

Conclusion

Overall, CNN showed strong validation performance in several developmental stages, while ViT was more sensitive to data limitations and uneven class distributions. YOLOv8 provided complementary capability for detecting and localizing PLB and leaf structures. The complementary use of object localization and image classification provides a basis for developing computer-vision systems for automated monitoring of Phalaenopsis orchids in tissue-culture and greenhouse environments. The findings indicate that successful application of deep learning in orchid monitoring depends not only on selecting an appropriate architecture but also on the quality, quantity, balance, and diversity of the training data. Larger and more balanced datasets, independent test sets, and additional images collected under different environmental and production conditions are recommended for future validation of generalizability. Further research can also investigate the integration of detection and classification models into a unified automated monitoring system and its connection with robotic systems for intelligent orchid production management.

نشریه
مکانیزاسیون کشاورزی
شاپا الکترونیکی: 2717-4107
درگاه نشریه: https://jam.tabrizu.ac.ir



مقاله پژوهشی

ارزیابی و مقایسه مدل‌های بینایی ماشین مبتنی بر یادگیری عمیق برای تشخیص مراحل فنولوژیک و وضعیت سلامت ارکیده فلانوپسیس

غزاله صفرعلیزاده هریسی [۱]، حسین باخدا [۲]، محمد قهدریجانی [۲]

تاریخ دریافت: ۱۴۰۵/۰۳/۳۱ تاریخ بازنگری:۱۴۰۵/۰۶/۲۸ تاریخ پذیرش:۱۴۰۵/۰۶/۲۹

۱- دانشجوی دکتری مکانیک بیوسیستم، گروه مهندسی سیستم های هوشمند کشاورزی، دانشکده کشاورزی محیط زیست و منابع طبیعی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.
۲- استادیار، گروه مکانیزاسیون کشاورزی، دانشکده کشاورزی محیط زیست و منابع طبیعی، واحد علوم وتحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

مسوول مکاتبات: hossein.bakhoda@gmail.com E-mail:

چکیده

ارکیده فلانوپسیس (Phalaenopsis) یکی از مهم‌ترین گیاهان زینتی تجاری جهان است و پایش دقیق مراحل رشد و وضعیت سلامت آن می‌تواند در بهبود مدیریت و کارایی تولید نقش داشته باشد. هدف این پژوهش، ارزیابی و مقایسه عملکرد سه معماری یادگیری عمیق شامل YOLO، شبکه عصبی کانولوشنی (CNN) و (ViT) برای تشخیص مراحل رشد و ارزیابی وضعیت سلامت گیاه فلانوپسیس بود. بدین منظور، مجموعه‌ای شامل ۱۲۷ تصویر RGB از گیاهان فلانوپسیس در پنج مرحله مختلف رشد تهیه و پس از پیش‌پردازش، برای آموزش و اعتبارسنجی مدل‌ها سازمان‌دهی شد. مدل YOLOv8 برای تشخیص و مکان‌یابی ساختارهای گیاهی شامل برگ و ساختارهای پروتوکورم‌مانند (PLB) با استفاده از کادرهای محدودکننده به‌کار گرفته شد، در حالی که مدل‌های CNN و ViT برای طبقه‌بندی تصاویر مورد استفاده قرار گرفتند. عملکرد مدل‌های طبقه‌بندی با استفاده از شاخص‌های Accuracy، Precision، Recall و F1-score و روند تغییرات تابع Loss ارزیابی شد و عملکرد YOLOv8 نیز بر اساس معیارهای Precision، Recall، mAP@0.5 و mAP@0.5:0.95 مورد بررسی قرار گرفت. نتایج نشان داد که CNN در مراحل دارای الگوهای تصویری قابل تفکیک عملکرد بالایی داشت و در مراحل ۱، ۲، ۳ و ۵ به دقت اعتبارسنجی ۱۰۰ درصد دست یافت، در حالی که عملکرد آن در مرحله ۴ به حدود ۳۳ درصد کاهش یافت. مدل ViT نیز در برخی مراحل عملکرد بالایی نشان داد، اما در مرحله ۴ با کاهش عملکرد مواجه شد .YOLOv8 نیز در تشخیص و مکان‌یابی ساختارهای هدف عملکرد مناسبی نشان داد و در بهترین حالت به mAP@0.5 برابر با ۶۷ درصد و Precision حدود ۹۷ درصد دست یافت. با توجه به محدود بودن تعداد نمونه‌ها و نبود مجموعه آزمون مستقل، نتایج این پژوهش باید در چارچوب عملکرد مدل‌ها بر روی داده‌های اعتبارسنجی تفسیر شوند. در مجموع، یافته‌ها نشان می‌دهد که استفاده مکمل از مکان‌یابی ساختارهای گیاهی و طبقه‌بندی تصاویر می‌تواند مبنایی برای توسعه سامانه‌های بینایی ماشین جهت پایش خودکار مراحل رشد و وضعیت سلامت فلانوپسیس در محیط‌های کشت‌بافت و گلخانه‌ای فراهم کند.

کلمات کلیدی: ارکیده فلانوپسیس، یادگیری عمیق، بینایی ماشین، Vision Transformer، YOLO، شبکه عصبی کانولوشنی، کشت بافت.

This is an open-access article under the CC BY NC license
(https://creativecommons.org/licenses/by-nc/4.0/)

۱-مقدمه

ارکیده فلانوپسیس (Phalaenopsis)[1] یکی از ارزشمندترین و پرکاربردترین گیاهان زینتی در جهان است که به دلیل ویژگی‌های زیبایی‌شناختی، تنوع ژنتیکی و ارزش اقتصادی، جایگاه ویژه‌ای در صنعت گل و گیاه دارد. موفقیت در تولید تجاری این گیاه به شناسایی دقیق مراحل رشد، ارزیابی مستمر وضعیت سلامت و مدیریت مناسب شرایط تولید وابسته است. در سال‌های اخیر، گسترش فناوری کشت‌بافت و تولید انبوه گیاهچه‌های ارکیده، نیاز به روش‌های دقیق، سریع و غیرمخرب برای پایش رشد و سلامت گیاه را افزایش داده است. در روش‌های متداول، ارزیابی وضعیت رشد و سلامت نمونه‌ها عمدتاً بر مشاهده بصری کارشناسان متکی است که علاوه بر زمان‌بر بودن، می‌تواند تحت تأثیر خطای انسانی، تفاوت در تجربه ارزیابان و محدودیت‌های عملیاتی قرار گیرد. این چالش‌ها در سامانه‌های تولید انبوه و کشت‌بافت که نیازمند پردازش تعداد زیادی نمونه هستند، اهمیت بیشتری پیدا می‌کنند. پیشرفت‌های اخیر در حوزه کشاورزی هوشمند، رباتیک کشاورزی و بینایی ماشین، امکان توسعه سامانه‌های خودکار برای پایش و مدیریت تولید گیاهان را فراهم کرده است. پردازش تصویر و یادگیری ماشین می‌توانند به‌عنوان ابزارهای ادراکی در این سامانه‌ها، اطلاعات بصری مربوط به ویژگی‌های گیاه، مرحله رشد و وضعیت سلامت را به‌صورت سریع و غیرمخرب استخراج کنند. این اطلاعات می‌تواند به‌عنوان ورودی برای مراحل بعدی تصمیم‌گیری، پایش و اتوماسیون مورد استفاده قرار گیرد و در نتیجه، نقش بینایی ماشین را از صرفاً تشخیص ویژگی‌های ظاهری به یک مؤلفه قابل استفاده در سامانه‌های هوشمند تولید ارتقا دهد. مطالعات مختلف نیز نشان داده‌اند که ترکیب بینایی ماشین با الگوریتم‌های یادگیری عمیق می‌تواند دقت تحلیل تصاویر گیاهی را افزایش داده و وابستگی به ارزیابی انسانی را کاهش دهد (Sungheetha and Sharma, 2021).

پژوهش‌های انجام‌شده در حوزه پردازش تصاویر گیاهی نشان داده‌اند که مدل‌های یادگیری عمیق قادرند ویژگی‌های پیچیده و چندبعدی تصاویر زیستی را استخراج و تحلیل کنند (Kuznetsova et al.,2020). در میان روش‌های یادگیری عمیق، شبکه‌های عصبی کانولوشنی (CNN)[2] به دلیل توانایی در استخراج ویژگی‌های محلی و طبقه‌بندی تصاویر، کاربرد گسترده‌ای در مسائل بینایی ماشین و کشاورزی پیدا کرده‌اند. این شبکه‌ها در کاربردهایی مانند تشخیص بیماری‌ها، طبقه‌بندی محصولات و ارزیابی ویژگی‌های گیاهی عملکرد موفقی داشته‌اند. با این حال، محدودیت نسبی در مدل‌سازی وابستگی‌های بلندبرد میان بخش‌های مختلف تصویر و حساسیت عملکرد به حجم و تنوع داده‌های آموزشی از جمله چالش‌های این معماری محسوب می‌شود(Krichen,2023).

در سال‌های اخیر، معماری‌های مبتنی بر ترنسفورمر و به‌ویژه (ViT)[3] به‌عنوان رویکردی جدید برای تحلیل تصاویر مورد توجه قرار گرفته‌اند. این مدل‌ها با استفاده از سازوکار توجه[4] قادرند روابط میان نواحی مختلف تصویر را مدل‌سازی کرده و ویژگی‌های سراسری تصویر را مورد استفاده قرار دهند. مطالعات مختلف نشان داده‌اند که معماری‌های مبتنی بر ترنسفورمر می‌توانند در مسائل طبقه‌بندی تصویر عملکرد رقابتی یا برتری نسبت به برخی معماری‌های متداول ارائه دهند در سال‌های اخیر، معماری‌های مبتنی بر ترنسفورمر و به‌ویژه (ViT) به‌عنوان رویکردی جدید برای تحلیل تصاویر مورد توجه قرار گرفته‌اند. این مدل‌ها با استفاده از سازوکار توجه قادرند روابط میان نواحی مختلف تصویر را مدل‌سازی کرده و ویژگی‌های سراسری تصویر را مورد استفاده قرار دهند. مطالعات مختلف نشان داده‌اند که معماری‌های مبتنی بر ترنسفورمر می‌توانند در مسائل طبقه‌بندی تصویر عملکرد رقابتی یا برتری نسبت به برخی معماری‌های متداول ارائه دهند (Naranjo-Torres et al.,2022)

از سوی دیگر، الگوریتم YOLO[5] به‌عنوان یکی از معماری‌های شناخته‌شده تشخیص شیء، علاوه بر شناسایی، امکان مکان‌یابی هم‌زمان ساختارهای هدف در تصویر را فراهم می‌کند. این قابلیت، YOLO را از مدل‌های مورد استفاده صرفاً برای طبقه‌بندی تصویر متمایز می‌سازد و آن را برای شناسایی موقعیت ساختارهای گیاهی مانند برگ و ساختارهای پروتوکورم‌مانند (PLB) مناسب می‌کند. سرعت پردازش و قابلیت مکان‌یابی این مدل، کاربرد آن را در سامانه‌های بینایی ماشین، رباتیک و کشاورزی هوشمند امکان‌پذیر ساخته است. از این رو، استفاده هم‌زمان از YOLO، CNN و ViT در این پژوهش با هدف مقایسه مستقیم یک معیار واحد میان سه مدل انجام نشده است، بلکه هر معماری متناسب با قابلیت اصلی خود در یک وظیفه مشخص به‌کار گرفته شده است؛ به‌طوری که YOLO برای تشخیص و مکان‌یابی ساختارهای هدف و CNN و ViT برای طبقه‌بندی تصاویر مورد استفاده قرار گرفته‌اند.کاربرد یادگیری عمیق در شناسایی و طبقه‌بندی گونه‌های مختلف ارکیده نیز در سال‌های اخیر مورد توجه قرار گرفته است. با استفاده از مدل DenseNet121[6]، دقت بالایی در طبقه‌بندی گونه‌های ارکیده گزارش شده و توانایی شبکه‌های عمیق در استخراج ویژگی‌های پیچیده تصاویر گیاهی نشان داده شده است(Sari and Akbar, 2025). همچنین روش‌های مبتنی بر بازیابی تصویر بر اساس محتوا (CBIR)[7] برای شناسایی ارکیده مورد

[1] Phalaenopsis → Phalaenopsis: genus of flowering plants in the orchid family (Orchidaceae)
[2] CNN → Convolutional Neural Network
[3] ViT → Vision Transformer
[4] سازوکار توجه Attention Mechanism→
[5] YOLO → You Only Look Once
[6] DenseNet121 → Densely Connected Convolutional Network 121
[7] CBIR → Content-Based Image Retrieval

استفاده قرار گرفته‌اند که قابلیت استفاده از ویژگی‌های تصویری برای تشخیص این گیاهان را نشان می‌دهد(Apriyanti et al., 2013) . علاوه بر این، در چارچوب Orchid، مدلی مبتنی بر یادگیری عمیق قابل توضیح ارائه شده است که علاوه بر انجام طبقه‌بندی، امکان شناسایی نواحی مؤثر تصویر در فرآیند تصمیم‌گیری مدل را فراهم می‌کند (Sangiovanni et al., 2025). پژوهش‌های دیگری نیز با استفاده از ویژگی‌های قابل تفسیر نظیر شکل و رنگ گل، قابلیت توضیح‌پذیری سامانه‌های تشخیص ارکیده را بررسی کرده‌اند (Apriyanti et al., 2023). در زمینه یادگیری با داده‌های محدود نیز روش‌های یادگیری عمیق مورد توجه قرار گرفته‌اند. با استفاده از روش Few-Shot Learning[1] و سازوکار Dual-Attention[2]، عملکرد مناسبی در شناسایی گونه‌های ارکیده گزارش شده است .Lee et al) (2025. همچنین با ترکیب شبکه‌های عصبی عمیق و شبکه‌های بیزی، چارچوب‌هایی برای افزایش قابلیت تفسیر نتایج مدل‌های یادگیری عمیق ارائه شده‌اند (Apriyanti et al., 2025). در حوزه حفاظت و پایش تنوع زیستی نیز با توسعه سامانه IUC-NN[3]، ظرفیت یادگیری عمیق را برای شناسایی خودکار گونه‌های ارکیده و کاربردهای مرتبط با حفاظت از تنوع زیستی نشان دادند. مجموعه این مطالعات بیانگر ظرفیت بالای روش‌های یادگیری عمیق در تحلیل تصاویر ارکیده است؛ با این حال، تمرکز بخش قابل توجهی از پژوهش‌های پیشین بر شناسایی گونه، گل یا ویژگی‌های ظاهری بوده است. با وجود پیشرفت‌های مذکور، بررسی مراحل مختلف رشد Phalaenopsis در شرایط کشت‌بافت و ارزیابی هم‌زمان وضعیت سلامت ساختارهای گیاهی، به‌ویژه ساختارهای پروتوکورم‌مانند (PLB) و برگ‌ها، کمتر مورد توجه قرار گرفته است. در محیط کشت‌بافت، تشخیص مرحله رشد و شناسایی نمونه‌های سالم و ناسالم می‌تواند اطلاعات مهمی برای پایش روند تکوین گیاهچه‌ها فراهم کند. همچنین، در مراحل پیچیده‌تر رشد که ساختارهای گیاهی از نظر ظاهری شباهت بیشتری دارند، تشخیص خودکار می‌تواند با چالش بیشتری مواجه شود. بنابراین، بررسی رفتار مدل‌های مختلف تحت چنین شرایطی می‌تواند به انتخاب مناسب‌تر روش بینایی ماشین برای توسعه سامانه‌های پایش هوشمند کمک کند. بر این اساس، هدف پژوهش حاضر، ارزیابی و مقایسه نظام‌مند سه رویکرد یادگیری عمیق شاملYOLO، شبکه عصبی کانولوشنی (CNN) و (ViT) برای پایش تصویری مراحل رشد و وضعیت سلامت ارکیده فلانوپسیس در شرایط کشت‌بافت است. در این پژوهش، YOLO برای تشخیص و مکان‌یابی ساختارهای هدف، به‌ویژه برگ و PLB[4] و CNN و ViT برای طبقه‌بندی تصاویر در پنج مرحله رشد مورد استفاده قرار گرفتند. در Stage4، به دلیل پیچیدگی بیشتر مسئله، طبقه‌بندی در چهار کلاس شامل PLB سالم، PLBناسالم، برگ سالم و برگ ناسالم انجام شد .(Zizka et al., 2020). نوآوری پژوهش حاضر در ارائه یک معماری جدید یا معرفی یک الگوریتم یادگیری عمیق جدید نیست، بلکه در طراحی و ارزیابی یک چارچوب بینایی ماشین مبتنی بر سه رویکرد مکملYOLO، CNN و (ViT) برای پایش تصویری مراحل رشد و وضعیت سلامت ارکیده فلانوپسیس در شرایط کشت‌بافت است. در این چارچوب، مسئله پایش گیاه به وظایف مشخص بینایی ماشین تفکیک شد؛ به‌گونه‌ای که YOLOv8 برای شناسایی و مکان‌یابی ساختارهای قابل مشاهده گیاه، به‌ویژه برگ و ساختارهای پروتوکورم‌مانند(PLB) ، به کار گرفته شد، در حالی که CNNو ViT برای طبقه‌بندی تصاویر و بررسی تفاوت‌های موجود میان مراحل رشد و وضعیت سلامت نمونه‌ها مورد استفاده قرار گرفتند. برای این منظور، مجموعه‌ای شامل ۱۲۷ تصویر RGB[5] از ارکیده فلانوپسیس در پنج مرحله رشد تهیه و داده‌ها به صورت مرحله‌به‌مرحله سازمان‌دهی شدند. عملکرد هر سه رویکرد در شرایط یکسان از نظر تفکیک داده‌های آموزشی و اعتبارسنجی و با در نظر گرفتن تفاوت در حجم نمونه، پیچیدگی تصویری و توزیع کلاس‌ها بررسی شد. در مرحله چهارم، که از نظر تشخیص ساختارهای گیاهی و وضعیت سلامت پیچیده‌تر بود، علاوه بر بررسی طبقه‌بندی، شناسایی و مکان‌یابی PLB و برگ با استفاده از YOLOv8 مورد توجه قرار گرفت. بدین ترتیب، پژوهش حاضر صرفاً به تشخیص یک ویژگی ظاهری اکتفا نکرده، بلکه قابلیت استخراج اطلاعات تصویری، طبقه‌بندی مرحله رشد و سلامت و مکان‌یابی ساختارهای گیاهی را به عنوان اجزای مکمل یک سامانه پایش هوشمند بررسی کرده است. اهمیت این چارچوب در آن است که خروجی مدل‌های بینایی ماشین می‌تواند به عنوان لایه ادراک تصویری در سامانه‌های هوشمند آینده مورد استفاده قرار گیرد؛ به این معنا که اطلاعات حاصل از تصویر، مانند مرحله رشد، وضعیت سلامت و موقعیت ساختارهای گیاهی، می‌تواند در مراحل بعدی برای تصمیم‌گیری، جداسازی نمونه‌ها، پایش روند رشد، مدیریت شرایط تولید و در نهایت ارتباط با سامانه‌های اتوماسیون و رباتیک به کار گرفته شود. بنابراین، هدف پژوهش حاضر توسعه یک سامانه رباتیک کامل نبوده است، بلکه ایجاد و ارزیابی مؤلفه بینایی و اطلاعات تصویری مورد نیاز برای چنین سامانه‌ای و تعیین نقاط قوت و محدودیت هر معماری در شرایط واقعی داده‌های کشت‌بافت بوده است. این ارزیابی می‌تواند مبنایی تجربی برای انتخاب مدل مناسب در توسعه سامانه‌های

[1] Few-Shot Learning → Few-Shot Learning: A machine learning approach that enables models to learn from a small number of labeled examples

[2] Dual-Attention → Dual-Attention Mechanism: An attention mechanism that simultaneously focuses on complementary feature representations to improve feature extraction.

[3] IUC-NN → IUC-NN: A deep-learning-based system developed for automated conservation assessment and identification of orchid species.

[4] PLB → Protocorm-Like Body: A protocorm-like structure formed during orchid tissue culture and micropropagation

[5] RGB → Red, Green, Blue: A color model representing images using red, green, and blue color channels

آینده پایش و مدیریت خودکار تولید گیاهچه‌های فلانوپسیس فراهم کند.

۲- مواد و روش‌ها

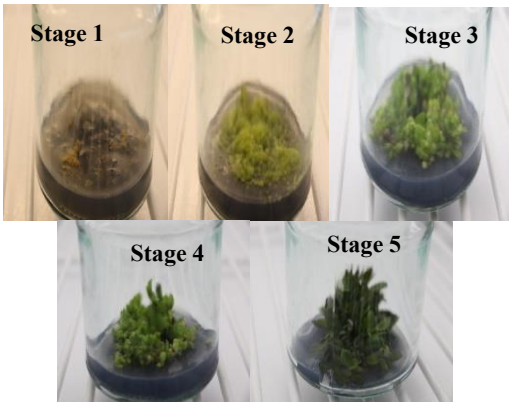
۲-۱- مجموعه داده (Dataset)

در این پژوهش، مجموعه‌ای از تصاویر مربوط به گیاه ارکیده (Phalaenopsis) در شرایط کشت‌بافت مورد استفاده قرار گرفت. تصاویر شامل مراحل مختلف رشد گیاه و وضعیت‌های متفاوت سلامت نمونه‌ها بودند و با هدف ارزیابی قابلیت روش‌های بینایی ماشین مبتنی بر یادگیری عمیق در تشخیص مراحل رشد و ارزیابی سلامت گیاه گردآوری شدند. داده‌های تصویری پنج مرحله تکوینی، شامل Stage1 تا Stage5، را دربرگرفتند که روند رشد گیاه را از مراحل اولیه تشکیل ساختارهای پروتوکورم‌مانند (Protocorm-Like Bodies ; PLBs)[1] تا مراحل پیشرفته‌تر رشد گیاهچه نشان می‌دادند. در مجموع، مجموعه داده اولیه شامل ۱۲۷ تصویر رنگی (RGB) بود. تصاویر در پنج مرحله تکوینی توزیع شدند؛ به‌طوری‌که ۹ تصویر مربوط به Stage1، ۲۳ تصویر مربوط به Stage2، ۲۵ تصویر مربوط به Stage3، ۴۷ تصویر مربوط به Stage4 و ۲۳ تصویر مربوط به Stage5 بود. بیشترین تعداد تصاویر به Stage4 اختصاص داشت؛ زیرا در این مرحله، ساختارهای PLB و برگ‌ها به‌صورت هم‌زمان در تصاویر مشاهده می‌شدند و علاوه بر نوع ساختار، وضعیت سلامت نمونه‌ها نیز مورد بررسی قرار گرفت. به‌منظور آماده‌سازی داده‌ها برای مدل‌های یادگیری عمیق، تصاویر هر مرحله به‌صورت جداگانه سازمان‌دهی و برچسب‌گذاری شدند. برای مدل‌های CNNو ViT، تصاویر در پوشه‌های اختصاصی مربوط به مرحله و زیرمجموعه آموزش و اعتبارسنجی قرار گرفتند. در مدل YOLOv8، تصاویر و برچسب‌های مربوط به ساختارهای هدف به‌صورت داده‌های تشخیص شیء و با استفاده از جعبه‌های محدودکننده (Bounding Box)[2]سازمان‌دهی شدند. بنابراین، ساختار مجموعه داده متناسب با نوع وظیفه هر مدل، یعنی طبقه‌بندی تصویر برای CNN و ViT و تشخیص و مکان‌یابی شیء برای YOLOv8، تنظیم شد. برای هر مرحله، تصاویر به دو زیرمجموعه آموزش (Train)[3] و اعتبارسنجی (Validation)[4]تقسیم شدند. در این پژوهش مجموعه آزمون مستقل (Independent Test Set)[5] در نظر گرفته نشد؛ بنابراین، نتایج گزارش‌شده مربوط به عملکرد مدل‌ها روی مجموعه اعتبارسنجی بوده و نباید به‌عنوان عملکرد روی یک مجموعه آزمون مستقل تفسیر شوند. تصاویر مجموعه اعتبارسنجی در فرآیند به‌روزرسانی پارامترهای مدل مورد استفاده قرار نگرفتند و پس از آموزش برای ارزیابی عملکرد مدل‌ها به کار گرفته شدند. این جداسازی از استفاده هم‌زمان یک تصویر در آموزش و اعتبارسنجی جلوگیری کرد. در میان مراحل مورد بررسی، Stage4 از نظر پیچیدگی تصویری و طبقه‌بندی، چالش‌برانگیزتر بود. در این مرحله، علاوه بر حضور هم‌زمان PLB و برگ، وضعیت سلامت این ساختارها نیز در نظر گرفته شد. ازاین‌رو، مدل‌ها در این مرحله با تنوع ساختاری و تفاوت‌های ظاهری بیشتری مواجه بودند. این ویژگی Stage4 را به مرحله‌ای مناسب برای بررسی محدودیت مدل‌های یادگیری عمیق در شرایط حجم محدود داده، تفاوت در توزیع کلاس‌ها و افزایش پیچیدگی بصری تبدیل کرد. توزیع تصاویر میان مراحل مختلف رشد و زیرمجموعه‌های آموزش و اعتبارسنجی در جدول ۱ ارائه شده است. همچنین، نمونه‌ای از تصاویر اصلی مورد استفاده در پنج مرحله رشد ارکیده فلانوپسیس در شکل ۱ نشان داده شده است.

جدول ۱- توزیع تصاویر مجموعه داده در مراحل مختلف رشد و زیرمجموعه‌های آموزش و اعتبارسنجی

Table 1. Distribution of dataset images across different growth stages and training and validation subsets.

مرحله رشد	تعداد کل تصاویر	آموزش (Train)	اعتبارسنجی (Validation)	آزمون مستقل (Test)
Stage1	9	6	2	0
Stage2	23	18	5	0
Stage3	25	20	5	0
Stage4	47	37	10	0
Stage5	23	18	5	0
مجموع	127	99	27	0



شکل ۱- تصاویر گیاه ارکیده Phalaenopsis در پنج مرحله رشد

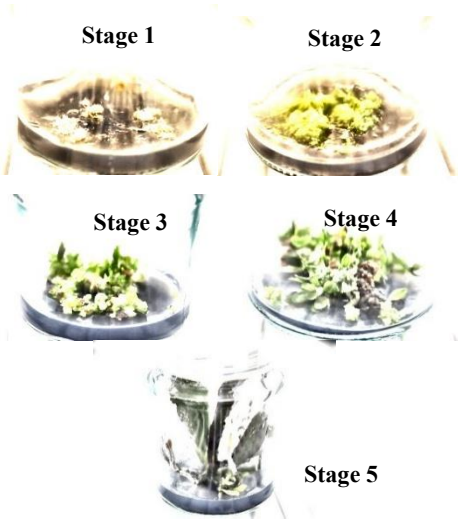
Fig 1. Phalaenopsis orchid plant pictures in 5 stages

[1] Protocorm-Like Bodies (PLBs) → Protocorm-Like Bodies (structures resembling protocorms that develop during orchid tissue culture and can regenerate into plantlets).
[2] Bounding Box → Bounding Box: a rectangular box used to define the location and extent of an object in an image.
[3] Train → Training: the process of teaching a machine-learning model using training data.
[4] Validation → Validation: the process of evaluating a trained model on a separate validation dataset during model development.
[5] ndependent Test Set → Independent Test Set: a separate dataset used to evaluate the final performance and generalization of a trained model.

۲-۲-پیش‌پردازش داده‌ها (Data Preprocessing)

پس از انجام مراحل پیش‌پردازش تصاویر، داده‌های پردازش‌شده برای آماده‌سازی و آموزش سه معماری YOLOv8، CNN و (ViT) مورد استفاده قرار گرفتند. هدف از این مرحله، بررسی و در صورت امکان افزایش تنوع نمونه‌های آموزشی و کاهش حساسیت مدل‌ها به تغییرات ظاهری تصاویر بود.

در فرآیند آماده‌سازی داده‌ها، روش‌های متداول افزایش داده شامل چرخش تصادفی تصاویر چرخش تصادفی[1](Random Rotation) ، وارونگی افقی[2](Horizontal Flip) ، برش تصادفی (Random Crop)[3]، تغییر مقیاس[4](Scaling)، تغییر روشنایی و کنتراست و تغییرات رنگی [5](Color Jittering) به‌عنوان روش‌های متداول قابل استفاده در نظر گرفته شدند. با این حال، در پیکربندی نهایی آزمایش‌های انجام‌شده، نمونه‌های جدید به‌صورت مصنوعی از طریق این تبدیل‌ها به مجموعه داده افزوده نشدند. بنابراین، تعداد تصاویر گزارش‌شده در جدول مجموعه داده، تعداد تصاویر واقعی مورد استفاده در آزمایش‌ها را نشان می‌دهد.در مقابل، پیش‌پردازش تصاویر بخشی از فرآیند آماده‌سازی داده‌ها برای هر سه معماری بود و نباید با Data Augmentation[6] یکسان در نظر گرفته شود. پیش‌پردازش با هدف ایجاد قالب یکنواخت برای تصاویر، آماده‌سازی آن‌ها برای ورود به مدل و بهبود سازگاری داده‌ها انجام شد؛ در حالی که افزایش داده با هدف تولید نمونه‌های جدید از طریق اعمال تبدیل‌های مختلف بر تصاویر موجود انجام می‌شود. پس از پیش‌پردازش، تصاویر مطابق با وظیفه هر مدل آماده شدند. برای CNN و ViT، تصاویر با اندازه ورودی ۲۲۴ × ۲۲۴ پیکسل آماده‌سازی شدند و برای YOLOv8، تصاویر در قالب مورد نیاز مدل تشخیص شیء و با اندازه ورودی ۶۴۰ × ۶۴۰ پیکسل مورد استفاده قرار گرفتند. در YOLOv8 همچنین برای ساختارهای هدف، شامل PLB و برگ، برچسب‌های مکانی به‌صورت Bounding Boxتهیه شد. از آنجا که در این پژوهش افزایش مصنوعی تعداد نمونه‌ها انجام نشد، عملکرد مدل‌ها بر اساس تصاویر واقعی موجود در مجموعه داده و تقسیم‌بندی از پیش تعیین‌شده آموزش و اعتبارسنجی ارزیابی شد. این موضوع به‌ویژه با توجه به حجم محدود داده اهمیت دارد و می‌تواند یکی از عوامل مؤثر بر تفاوت عملکرد مدل‌ها در مراحل مختلف، به‌خصوص Stage4، باشد. در نتیجه، در این پژوهش باید میان دو مفهوم پیش‌پردازش تصویر و افزایش داده تمایز قائل شد: پیش‌پردازش برای آماده‌سازی و استانداردسازی تصاویر انجام شد، در حالی که افزایش داده برای تولید نمونه‌های مصنوعی جدید در آزمایش‌های نهایی مورد استفاده قرار نگرفت. بررسی اثر روش‌های مختلف افزایش داده بر عملکرد مدل‌ها می‌تواند در مطالعات آینده با مجموعه داده بزرگ‌تر و متوازن‌تر انجام شود. نمونه‌ای از تصاویر حاصل از مراحل پردازش در پنج مرحله رشد ارکیده فلانوپسیس در شکل (۲) ارائه شده است که تمامی مراحل در نرم افزار پایتون انجام شده است.



شکل۲- پردازش تصویر هر ۵ مرحله رشد گل ارکیده فلانوپسیس توسط نرم افزار پایتون

Fig 2. Image processing of all 5 stages of Phalaenopsis orchid flower growth using Python software

۲-۴- مدل‌های یادگیری عمیق و روش تحقیق

در این پژوهش، سه معماری یادگیری عمیق شامل YOLOv8، شبکه عصبی کانولوشنی (CNN) و (ViT) برای تحلیل تصاویر ارکیده Phalaenopsis مورد بررسی قرار گرفتند. انتخاب این سه معماری با هدف بررسی قابلیت‌های مکمل روش‌های مختلف بینایی ماشین در دو وظیفه اصلی تشخیص و مکان‌یابی ساختارهای گیاه و طبقه‌بندی تصاویر انجام شد. ازاین‌رو، مقایسه مدل‌ها بر اساس یک وظیفه کاملاً یکسان انجام نشده است؛ بلکه هر معماری متناسب با قابلیت اصلی خود در چارچوب سامانه پایش تصویری ارکیده مورد استفاده قرار گرفت. YOLOv8 به‌عنوان مدل تشخیص شیء برای شناسایی و مکان‌یابی ساختارهای هدف گیاه، شامل برگ و ساختارهای پروتوکورم‌مانند(PLB)، به کار گرفته شد. در این بخش، تصاویر با استفاده از کادرهای محدودکننده (Bounding Box) برای ساختارهای هدف برچسب‌گذاری شدند و مدل وظیفه تعیین محل و نوع ساختارهای موردنظر در تصویر را بر عهده داشت. بنابراین، خروجی YOLOv8علاوه بر اطلاعات مربوط به کلاس، شامل موقعیت مکانی ساختار شناسایی‌شده نیز بود. این قابلیت به‌ویژه برای مرحله ۴ اهمیت داشت؛ زیرا در این مرحله، حضور هم‌زمان PLB و برگ و تفاوت وضعیت سلامت نمونه‌ها، پیچیدگی تصویری بیشتری ایجاد می‌کرد. عملکرد YOLOv8 با معیارهای اختصاصی تشخیص شیء، از جمله

[1] Random Rotation → Random Rotation
[2] Horizontal Flip → Horizontal Flip
[3] Random Crop → Random Crop
[4] Scaling → Scaling
[5] Color Jittering → Color Jittering
[6] Data Augmentation → Data Augmentation

معیارهای دقت مثبت (Precision)[1]، بازخوانی (Recall)[2] و میانگین دقت متوسط Precision (mAP)[3] ، ارزیابی شد.

CNN برای انجام وظیفه طبقه‌بندی تصاویر در مراحل مختلف رشد و همچنین ارزیابی وضعیت سلامت در مرحله ۴ مورد استفاده قرار گرفت. در این پژوهش، معماری ResNet18[4] به‌عنوان مدل CNNانتخاب شد. مدل با استفاده از وزن‌های ازپیش‌آموزش‌دیده بر مجموعه ImageNet[5] آغاز شد و لایه نهایی آن متناسب با تعداد کلاس‌های هر مرحله تنظیم گردید. تصاویر ورودی برای این مدل به اندازه 224 × 224 پیکسل تغییر اندازه داده شدند. در مراحل ۱ تا ۳ و ۵، طبقه‌بندی بر اساس کلاس‌های تعریف‌شده برای هر مرحله انجام شد؛ در حالی که مرحله ۴ به‌صورت یک مسئله طبقه‌بندی جزئی‌تر شامل چهار کلاس PLBسالم، PLB ناسالم، برگ سالم و برگ ناسالم در نظر گرفته شد. بنابراین، CNN در این پژوهش علاوه بر تشخیص مرحله رشد، برای بررسی توانایی مدل در تفکیک وضعیت سلامت ساختارهای گیاهی نیز مورد استفاده قرار گرفت.

(ViT) به‌عنوان یک معماری مبتنی بر ترنسفورمر برای بررسی قابلیت یادگیری روابط سراسری میان بخش‌های مختلف تصویر به کار گرفته شد. در این پژوهش از معماری ViT-Base/16[6] با وضوح ورودی ۲۲۴ × ۲۲۴ پیکسل استفاده شد. برخلاف CNN که استخراج ویژگی را عمدتاً بر اساس عملیات کانولوشنی و الگوهای محلی انجام می‌دهد، ViTتصویر را به قطعات کوچک‌تر (Patch)[7] تقسیم کرده و با استفاده از سازوکار Self-Attention[8] روابط میان این بخش‌ها را مدل‌سازی می‌کند. هدف از به‌کارگیری ViT در این پژوهش، بررسی رفتار یک معماری ترنسفورمری در مقایسه با یک معماری کانولوشنی، به‌ویژه در شرایطی با حجم محدود داده و تفاوت در پیچیدگی مراحل رشد بود. به‌منظور جلوگیری از برداشت نادرست از مقایسه مدل‌ها، لازم به ذکر است که YOLOv8و مدل‌های CNN و ViT یک وظیفه یکسان را انجام نمی‌دهند YOLOv8برای تشخیص و مکان‌یابی اشیاء، و CNN و ViT برای طبقه‌بندی تصاویر به کار گرفته شدند. بنابراین، معیارهای عملکرد این مدل‌ها متناسب با وظیفه مربوطه تفسیر شدند و مقادیر Accuracy[9] مربوط به مدل‌های طبقه‌بندی مستقیماً با mAP مربوط به YOLOv8 مقایسه نشدند. هدف اصلی، بررسی این موضوع بود که هر یک از این معماری‌ها در بخش متفاوتی از یک چارچوب بینایی ماشین چه عملکردی ارائه می‌دهد و چگونه می‌توان از قابلیت‌های مکمل آن‌ها در توسعه سامانه‌های آینده پایش خودکار رشد و سلامت ارکیده استفاده کرد. در این چارچوب، CNN و ViT اطلاعات طبقه‌بندی مربوط به مرحله رشد و وضعیت سلامت را فراهم می‌کنند، در حالی که YOLOv8 اطلاعات مکانی مربوط به ساختارهای قابل مشاهده در تصویر را استخراج می‌کند .ترکیب این اطلاعات می‌تواند در مراحل بعدی توسعه سامانه، به‌عنوان ورودی برای تصمیم‌گیری، پایش خودکار و در آینده اتصال به سامانه‌های اتوماسیون و رباتیک مورد استفاده قرار گیرد. از این منظر، پژوهش حاضر یک مدل جدید معرفی نمی‌کند، بلکه با ارزیابی نظام‌مند سه رویکرد مکمل، نقاط قوت، محدودیت‌ها و شرایط عملکرد آن‌ها را در پنج مرحله رشد ارکیده Phalaenopsis و تحت شرایط حجم محدود داده بررسی می‌کند.

۲-۵- برچسب‌گذاری داده‌ها

به‌منظور آماده‌سازی داده‌ها برای آموزش و ارزیابی مدل‌های یادگیری عمیق، تصاویر پس از تفکیک بر اساس مراحل رشد، متناسب با وظیفه هر مدل سازمان‌دهی و برچسب‌گذاری شدند. برچسب‌گذاری تصاویر به‌صورت دستی و با استفاده از نرم‌افزار LabelImg[10] انجام گرفت. در این فرآیند، ساختارهای قابل مشاهده گیاه، شامل برگ‌ها و پروتوکورم‌مانندها(PLB)، با توجه به نوع ساختار و وضعیت سلامت آن‌ها مشخص شدند.

برای وظیفه تشخیص شیء با YOLOv8، ساختارهای هدف موجود در تصاویر با استفاده از Bounding Box مشخص شدند. برای هر ناحیه موردنظر، یک کادر محدودکننده در اطراف ساختار مربوطه ترسیم و کلاس آن تعیین شد. برچسب‌های مورد استفاده در مجموعه داده تشخیص شامل دو کلاس اصلی PLB و Leaf بودند که به‌ترتیب با شناسه‌های عددی 0 و ۱ در فایل‌های برچسب YOLO ذخیره شدند. هر فایل برچسب مربوط به یک تصویر بوده و اطلاعات مربوط به کلاس و موقعیت مکانی شیء شناسایی‌شده را در قالب استاندارد YOLO شامل مختصات نرمال‌شده کادر محدودکننده ثبت می‌کرد. در مرحله ۴، به دلیل اهمیت بررسی هم‌زمان ساختار و سلامت نمونه‌ها، داده‌ها با جزئیات بیشتری سازمان‌دهی شدند. در این مرحله، نمونه‌های PLB و برگ بر اساس وضعیت سلامت به دو وضعیت سالم و ناسالم تفکیک شدند. بنابراین، برای وظیفه طبقه‌بندی مرحله ۴، چهار کلاس شامل PLB سالم (Healthy PLB)[11]، PLB ناسالم (Unhealthy PLB)[12]، برگ سالم (Healthy Leaf)[13] و برگ ناسالم (Unhealthy Leaf)[14] در نظر گرفته شد. این تفکیک امکان بررسی توانایی مدل‌های

[1] Precision → Precision
[2] Recall → Recall
[3] mAP → mean Average Precision
[4] ResNet18 → Residual Network 18
[5] ImageNet → ImageNet dataset
[6] ViT-Base/16 → Vision Transformer Base model with 16×16 pixel image patches
[7] Patch → Image Patch
[8] Self-Attention → Self-Attention Mechanism
[9] Accuracy → Classification Accuracy
[10] LabelImg → Graphical Image Annotation Tool
[11] Healthy PLB → Healthy Protocorm-Like Body
[12] Unhealthy PLB → Unhealthy Protocorm-Like Body
[13] Healthy Leaf → Healthy Leaf
[14] Unhealthy Leaf → Unhealthy Leaf

CNNو ViT در تشخیص هم‌زمان نوع ساختار و وضعیت سلامت آن را فراهم کرد.

برای CNN و ViT، تصاویر بر اساس وظیفه طبقه‌بندی در پوشه‌های مجزا و ساختار مرحله‌ای سازمان‌دهی شدند و کلاس هر تصویر از طریق ساختار پوشه مربوط به آن تعیین شد. در نتیجه، روش برچسب‌گذاری در این دو مدل با YOLO متفاوت بود؛ زیرا CNN و ViT برای طبقه‌بندی تصویر استفاده شدند، در حالی که YOLOv8 برای تشخیص و مکان‌یابی ساختارهای مشخص درون تصویر به برچسب‌های مکانی نیاز داشت. فرآیند برچسب‌گذاری و سازمان‌دهی داده‌ها پیش از مرحله آموزش انجام شد و مجموعه‌های آموزش و اعتبارسنجی به‌صورت جداگانه نگهداری شدند. تصاویر مجموعه اعتبارسنجی در فرآیند به‌روزرسانی پارامترهای مدل مورد استفاده قرار نگرفتند و پس از آموزش برای ارزیابی عملکرد مدل‌ها به کار رفتند. همچنین، از آنجا که در این پژوهش مجموعه آزمون مستقل (Independent Test Set)[1] ایجاد نشده بود، ارزیابی نهایی گزارش‌شده مربوط به مجموعه اعتبارسنجی است و نباید به‌عنوان عملکرد روی یک مجموعه آزمون مستقل تفسیر شود. با توجه به حجم محدود مجموعه داده، دقت در تعیین کلاس و مکان ساختارهای هدف اهمیت ویژه‌ای داشت؛ زیرا خطا در برچسب‌گذاری می‌تواند مستقیماً بر فرآیند یادگیری مدل و نتایج ارزیابی اثر بگذارد. ازاین‌رو، تصاویر و برچسب‌های مربوط به هر مرحله پیش از استفاده در فرآیند آموزش بررسی و مطابق ساختار داده موردنیاز هر مدل سازمان‌دهی شدند. ساختار نهایی داده‌ها و تعداد تصاویر آموزش و اعتبارسنجی هر مرحله در جدول مربوط به مجموعه داده ارائه شده است.

۲-۶- معماری مدل‌ها

در این پژوهش، سه معماری YOLOv8، CNN مبتنی بر ResNet18 و (ViT-Base/16) با توجه به وظیفه موردنظر برای تحلیل تصاویر ارکیده Phalaenopsis به کار گرفته شدند. از آنجا که این مدل‌ها برای وظایف متفاوتی استفاده شدند، ساختار و خروجی آن‌ها نیز متناسب با وظیفه مربوطه در نظر گرفته شد YOLOv8.برای تشخیص و مکان‌یابی ساختارهای گیاهی و CNN و ViT برای طبقه‌بندی تصاویر مورد استفاده قرار گرفتند.

مدل YOLOv8n به‌عنوان مدل تشخیص شیء برای شناسایی و مکان‌یابی ساختارهای هدف، شامل PLB و Leaf، استفاده شد. در این روش، تصویر ورودی توسط شبکه پردازش شده و برای هر شیء شناسایی‌شده، کلاس و موقعیت مکانی آن به‌صورت کادر محدودکننده (Bounding Box)پیش‌بینی می‌شود. بنابراین، برخلاف مدل‌های طبقه‌بندی، خروجی YOLOv8 تنها به تعیین کلاس تصویر محدود نبوده و محل قرارگیری ساختار موردنظر در تصویر را نیز مشخص می‌کند. در ساختار YOLOv8، فرآیند تشخیص بر پایه بخش‌های اصلی شبکه شامل استخراج ویژگی، پردازش ویژگی‌های چندمقیاسی و سر تشخیص (Detection Head)[2] انجام می‌شود. تابع هزینه مدل از مؤلفه‌های مربوط به مکان‌یابی جعبه، طبقه‌بندی و Distribution Focal Loss (DFL)[3] تشکیل می‌شود. این مؤلفه‌ها به‌ترتیب در یادگیری موقعیت دقیق اشیاء، تعیین کلاس و بهبود دقت پیش‌بینی مرز جعبه‌های محدودکننده نقش دارند. در این پژوهش، YOLOv8 به‌طور ویژه برای تشخیص ساختارهای PLB و برگ مورد استفاده قرار گرفت و نتایج آن با معیارهای اختصاصی تشخیص شیء، از جمله Precision، Recall وmAP، ارزیابی شد. بنابراین، عملکرد YOLOv8 از نظر ماهیت وظیفه با عملکرد طبقه‌بندی CNN و ViT متفاوت است.

برای وظیفه طبقه‌بندی، از معماری ResNet18 به‌عنوان مدل CNNاستفاده شد. این معماری با استفاده از لایه‌های کانولوشنی، ویژگی‌های تصویری را به‌صورت سلسله‌مراتبی استخراج می‌کند و از ساختار باقی‌مانده (Residual)[4] برای تسهیل یادگیری شبکه استفاده می‌کند. در این پژوهش، مدل ResNet18 با وزن‌های ازپیش‌آموزش‌دیده روی مجموعه ImageNet آغاز شد و لایه نهایی آن متناسب با تعداد کلاس‌های هر مرحله تنظیم گردید. تصاویر ورودی CNN به اندازه 224 × 224 پیکسل تغییر اندازه داده شدند. مدل برای هر مرحله رشد بر اساس ساختار داده همان مرحله آموزش داده شد و خروجی شبکه، کلاس مربوط به تصویر را تعیین می‌کرد. در مرحله ۴، به دلیل وجود هم‌زمان اطلاعات مربوط به نوع ساختار و وضعیت سلامت، مسئله طبقه‌بندی با چهار کلاس PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم تعریف شد. در نتیجه، CNN در این پژوهش برای بررسی توانایی یک معماری کانولوشنی در استخراج ویژگی‌های مورفولوژیک و طبقه‌بندی تصاویر در مراحل مختلف رشد و همچنین تفکیک وضعیت سلامت ساختارهای مرحله ۴ مورد استفاده قرار گرفت.

برای بررسی عملکرد معماری‌های مبتنی بر ترنسفورمر، از مدل ViT-Base/16 با وضوح ورودی 224 × 224 پیکسل استفاده شد. در این معماری، تصویر ورودی به مجموعه‌ای از قطعات کوچک‌تر (Image Patches)[5] تقسیم می‌شود. هر قطعه پس از تبدیل به

[1] Independent Test Set → Independent Test Set: a separate dataset used to evaluate the final generalization performance of a trained model.

[2] Detection Head → Detection Head: the final component of an object-detection network that predicts object locations, classes, and confidence scores.

[3] Distribution Focal Loss (DFL) → Distribution Focal Loss: a loss function used to improve the accuracy of bounding-box localization by modeling the distribution of object-boundary positions.

[4] Residual → Residual: the difference or shortcut connection between the input and output of a neural-network layer or block.

[5] Image Patches → Image Patches: small, fixed-size regions into which an input image is divided for processing by a Vision Transformer.

نمایش برداری (Patch Embedding)[1] به‌عنوان یک ورودی برای بخش ترنسفورمر در نظر گرفته می‌شود. پس از ایجاد نمایش برداری قطعات، سازوکار Self-Attention[2] روابط میان بخش‌های مختلف تصویر را مدل‌سازی می‌کند. این ویژگی به مدل اجازه می‌دهد وابستگی میان نواحی مختلف تصویر را در فرآیند استخراج ویژگی در نظر بگیرد. در این پژوهش، ViT برای همان وظیفه کلی طبقه‌بندی مورد استفاده قرار گرفت و مانندCNN ، داده‌ها به‌صورت مرحله‌ای سازمان‌دهی شدند.

هدف از استفاده هم‌زمان CNN و ViT، بررسی تفاوت رفتار یک معماری مبتنی بر استخراج ویژگی‌های کانولوشنی با یک معماری مبتنی بر توجه سراسری در شرایط یکسان داده‌ای و در مراحل مختلف رشد بود. به‌ویژه، عملکرد این دو معماری در مرحله ۴ که دارای پیچیدگی بیشتر و کلاس‌های مرتبط با سلامت نمونه‌ها بود، مورد توجه قرار گرفت. در مجموع، معماری‌های مورد استفاده مکمل یکدیگر در نظر گرفته شدند YOLOv8: اطلاعات مربوط به مکان و نوع ساختارهای گیاهی را فراهم کرد، در حالی که CNN و ViT برای طبقه‌بندی تصاویر و بررسی مرحله رشد و وضعیت سلامت به کار گرفته شدند .بنابراین، مقایسه نتایج این سه مدل با توجه به وظیفه اختصاص‌یافته به هر معماری انجام شد و معیارهای ارزیابی هر مدل متناسب با همان وظیفه تفسیر گردید.

۲-۷- تنظیمات آموزش مدل‌ها

به‌منظور بررسی عملکرد سه معماری یادگیری عمیق، مدل‌های YOLOv8، CNN و ViT به‌صورت مستقل و متناسب با وظیفه مربوط به هر مدل آموزش داده شدند. داده‌های هر مرحله به مجموعه‌های آموزش و اعتبارسنجی تقسیم شدند و تصاویر مجموعه اعتبارسنجی در فرآیند به‌روزرسانی پارامترهای مدل مورد استفاده قرار نگرفتند. از آنجا که در این پژوهش مجموعه آزمون مستقل ایجاد نشده بود، ارزیابی عملکرد مدل‌ها بر اساس مجموعه اعتبارسنجی انجام شد.

برای مدل CNN مبتنی برResNet18، تصاویر با اندازه ۲۲۴ × ۲۲۴ پیکسل به شبکه وارد شدند. مدل با استفاده از وزن‌های از پیش‌آموزش‌دیده روی مجموعه ImageNet مقداردهی اولیه شد و با استفاده از بهینه‌ساز Adam[3]، نرخ یادگیری ۰/۰۰۱، اندازه دسته ۸ و به مدت 20 epoch[4] آموزش داده شد. تابع Cross-Entropy Loss[5] به‌عنوان تابع هزینه طبقه‌بندی مورد استفاده قرار گرفت. در طول آموزش، مقدار خطای آموزش، خطای اعتبارسنجی و دقت اعتبارسنجی ثبت و برای بررسی روند یادگیری مدل استفاده شد. برای مدل-ViT) (Base/16 نیز اندازه ورودی تصاویر ۲۲۴ × ۲۲۴ پیکسل در نظر گرفته شد. مدل با استفاده از وزن‌های ازپیش‌آموزش‌دیده مقداردهی اولیه و با بهینه‌ساز Adam، نرخ یادگیری ۱×۱۰^{-۴} ، اندازه دسته ۸ و به مدت 50 epoch آموزش داده شد. تابع Cross-Entropy Loss برای بهینه‌سازی مدل مورد استفاده قرار گرفت و مقادیر Loss آموزش و اعتبارسنجی در طول فرآیند آموزش ثبت شدند.

برای YOLOv8n، تصاویر با اندازه ۶۴۰ × ۶۴۰ پیکسل مورد استفاده قرار گرفتند. مدل با وزن‌های از پیش‌آموزش‌دیده YOLOv8n آغاز شد و با اندازه دسته ۸ و به مدت 100 epoch روی CPU آموزش داده شد. در این مدل، تنظیمات بهینه‌سازی پیش‌فرض YOLOv8 مورد استفاده قرار گرفت و مقدار مشخصی برای بهینه‌ساز و نرخ یادگیری در اسکریپت آموزش تعیین نشده بود؛ بنابراین، از نسبت دادن یک مقدار ثابت برای این پارامترها در گزارش پژوهش خودداری شد. آموزش YOLOv8 بر اساس ساختار استاندارد داده‌های تشخیص شیء شامل تصاویر و فایل‌های برچسب مربوط به Bounding Box انجام گرفت. پارامترهای اصلی آموزش سه مدل در جدول ۲ خلاصه شده‌اند. این جدول با هدف افزایش قابلیت بازتولیدپذیری پژوهش، امکان مشاهده مستقیم تنظیمات مورد استفاده در هر معماری را فراهم می‌کند.

جدول۲- پارامترهای اصلی آموزش سه مدل

Table 2. Key training parameters for the three models

مدل	وظیفه	اندازه ورودی	Batch Size	Epoch	Optimizer	Learning Rate
CNN (ResNet18)	طبقه‌بندی	224 × 224	8	20	Adam	0.001
ViT-Base/16	طبقه‌بندی	224 × 224	8	50	Adam	0.0001
YOLOv8n	تشخیص شیء	640 × 640	4	100	تنظیمات پیش‌فرض YOLOv8	تنظیمات پیش‌فرض YOLOv8

با توجه به تفاوت وظایف سه معماری، تنظیمات آموزشی نیز بر اساس ماهیت هر مدل تعیین شد. بنابراین، هدف از این تنظیمات ایجاد شرایط یکسان عددی برای تمام مدل‌ها نبود، بلکه حفظ یک فرآیند آموزش مشخص، قابل گزارش و قابل بازتولید برای هر یک از وظایف تشخیص و طبقه‌بندی بود. آموزش مدل‌ها در محیط CPU و با استفاده از پردازنده Intel(R) Core(TM) i5-2450M CPU @ 2.50 GHz و Python نسخه 3.12.10 انجام شد و از GPU استفاده نشد. مدل CNN (ResNet18) باBatch Size=8، نرخ یادگیری 0.001 و Epoch 20، مدل ViT-Base/16 با Batch Size=8، نرخ یادگیری ۱×۱۰^{-۴} و Epoch 50، و مدل YOLOv8n با اندازه ورودی ۶۴۰×۶۴۰، Batch Size=4 و Epoch 100 آموزش

[1] Patch Embedding → Patch Embedding: the process of converting image patches into vector representations that can be processed by a Vision Transformer.
[2] Self-Attention → Self-Attention Mechanism
[3] Adam → Adaptive Moment Estimation
[4] Epoch → One complete pass through the training dataset
[5] Cross-Entropy Loss → Cross-Entropy Loss Function

داده شـدند. برای CNN و ViT از Cross-Entropy Loss و برای YOLOv8 از تابع هزینه پیش‌فرض این معماری اسـتفاده شـد. با توجه به تفاوت وظایف، YOLOv8 برای تشـخیص و مکان‌یابی و CNN و ViT برای طبقه‌بندی تصـاویر به‌کار گرفته شـدند؛ بنابراین نتایج آن‌ها به‌صورت مستقیم و هم‌وظیفه با یکدیگر مقایسه نشد.

۲-۸- معیارهای ارزیابی عملکرد

برای ارزیابی عملکرد مدل‌ها، از شاخص‌های متناسب با نوع وظیفه هر معماری استفاده شد. از آنجا که YOLOv8 به‌عنوان مدل تشخیص و مکان‌یابی شیء و CNN و ViT به‌عنوان مدل‌های طبقه‌بندی تصویر به‌کار گرفته شدند، معیارهای ارزیابی آن‌ها متناسب با وظیفه مربوطه در نظر گرفته شد. برای مدل YOLOv8، معیارهای Precision، Recall، mAP@0.5 و mAP@0.5:0.95 برای ارزیابی توانایی مدل در تشخیص صحیح و مکان‌یابی ساختارهای هدف شامل PLB و برگ استفاده شدند Precision .بیانگر نسبت تشخیص‌های صحیح به کل تشخیص‌های مثبت مدل و Recall نشان‌دهنده توانایی مدل در شناسایی نمونه‌های هدف موجود است. معیارهای mAP نیز برای ارزیابی عملکرد کلی مدل در تشخیص و مکان‌یابی اشیاء در آستانه‌های مختلف هم‌پوشانی جعبه‌های مرزی مورد استفاده قرار گرفتند. برای مدل‌های CNN و ViT، معیارهای Accuracy، Precision، Recall و F1-score[1] برای ارزیابی عملکرد طبقه‌بندی محاسبه شدند. همچنین Training Loss و Validation Loss در طول فرآیند آموزش ثبت و بررسی شدند. مقایسه روند این دو مقدار برای بررسی رفتار یادگیری مدل و شناسایی نشانه‌های احتمالی بیش‌برازش مورد استفاده قرار گرفت؛ به‌ویژه در مواردی که دقت اعتبارسنجی بالا یا برابر با ۱۰۰ درصد مشاهده شد. با توجه به تعداد محدود تصاویر و نبود مجموعه آزمون مستقل، Accuracyهای بالا، از جمله مقدار ۱، صرفاً بیانگر عملکرد مدل بر روی مجموعه اعتبارسنجی موجود بوده و به‌عنوان اثبات قطعی تعمیم‌پذیری مدل تفسیر نشدند. علاوه بر این، ماتریس درهم‌ریختگی (Confusion Matrix) برای بررسی الگوی پیش‌بینی‌های صحیح و نادرست مدل‌های طبقه‌بندی مورد استفاده قرار گرفت. این ماتریس امکان بررسی عملکرد مدل در سطح هر کلاس و شناسایی کلاس‌هایی را که بیشترین میزان اشتباه در تشخیص آن‌ها رخ داده است، فراهم کرد. این موضوع به‌ویژه برای Stage4 اهمیت داشت، زیرا در این مرحله، چهار کلاس شامل PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم مورد بررسی قرار گرفتند. در نهایت، نتایج هر مدل با در نظر گرفتن نوع وظیفه، تعداد محدود نمونه‌ها، توزیع کلاس‌ها و پیچیدگی متفاوت مراحل رشد تفسیر شدند. بنابراین، معیارهای طبقه‌بندی CNN و ViT مستقیماً با معیارهای تشخیص YOLOv8مقایسه نشدند و هر مدل بر اساس وظیفه‌ای که برای آن طراحی و آموزش داده شده بود ارزیابی شد.

۲-۹- محیط پیاده‌سازی

کلیه مراحل آماده‌سازی داده‌ها، آموزش و ارزیابی مدل‌های یادگیری عمیق در محیط برنامه‌نویسی Python 3.12.10 انجام شد. آموزش و اجرای مدل‌ها با استفاده از پردازنده مرکزی (CPU) انجام گرفت و در این پژوهش از پردازنده گرافیکی (GPU) استفاده نشد. پردازنده مورد استفاده از نوع Intel(R) Core(TM) i5-2450M CPU @ 2.50GHz بود. برای پیاده‌سازی مدل CNN و انجام پردازش‌های مرتبط با داده‌ها از کتابخانه PyTorch[2] و ابزارهای مرتبط با آن، از جمله Torchvision[3]، استفاده شد. مدل Vision Transformerنیز با استفاده از چارچوب PyTorch و کتابخانه timm[4] پیاده‌سازی و آموزش داده شد. مدل YOLOv8 با استفاده از کتابخانه Ultralytics[5] اجرا گردید. همچنین از کتابخانه‌های NumPy[6]، OpenCV[7]، Pillow[8]، Pandas[9] و Matplotlib[10] برای پردازش و آماده‌سازی تصاویر، مدیریت داده‌ها، محاسبه و ذخیره نتایج و ترسیم نمودارهای مورد نیاز استفاده شد. تمامی مدل‌ها در همین محیط محاسباتی آموزش و ارزیابی شدند؛ بنابراین نتایج گزارش‌شده تحت شرایط یکسان سخت‌افزاری و نرم‌افزاری به دست آمده‌اند. با توجه به استفاده از CPU[11]، زمان آموزش مدل‌ها نسبت به اجرای آن‌ها بر روی GPU[12] بیشتر بوده است.

۲-۱۰- طراحی آزمایش

به‌منظور بررسی قابلیت مدل‌های یادگیری عمیق در تحلیل تصاویر ارکیده فلانوپسیس، یک طراحی آزمایشی چندمدلی بر پایه سه معماری YOLOv8، CNN و (ViT) در نظر گرفته شد. هدف از این طراحی، ایجاد یک مدل واحد یا ادغام مستقیم سه شبکه نبود؛ بلکه هر معماری متناسب با قابلیت اصلی خود در یک وظیفه مشخص به‌صورت مستقل آموزش داده شد تا توانایی، محدودیت‌ها و رفتار آن در شرایط مختلف داده مورد بررسی قرار گیرد. در این طراحی، YOLOv8 به‌عنوان مدل تشخیص شیء برای شناسایی و مکان‌یابی ساختارهای قابل مشاهده گیاه، شامل پروتوکورم‌مانند (PLB) و برگ، استفاده شد. این بخش به‌ویژه در مرحله چهارم رشد اهمیت داشت؛ زیرا در این مرحله ساختارهای PLB و برگ به‌صورت هم‌زمان در تصاویر وجود داشتند و امکان بررسی توانایی مدل در تعیین محل این

[1] F1-score → F1-score: Harmonic mean of Precision and Recall
[2] PyTorch → PyTorch: An open-source deep learning framework
[3] Torchvision → Torchvision: A computer vision library for PyTorch
[4] timm → timm: PyTorch Image Models
[5] Ultralytics → Ultralytics: A framework for developing and training YOLO models
[6] NumPy → NumPy: Numerical Python
[7] OpenCV → Open-Source Computer Vision Library
[8] Pillow → Pillow: Python Imaging Library
[9] Pandas → Pandas: Python Data Analysis Library
[10] Matplotlib → Matplotlib: A Python library for data visualization
[11] CPU → Central Processing Unit
[12] GPU → Graphics Processing Unit

ساختارها فراهم بود. در مقابل، CNN مبتنی بر ResNet18 و Vision Transformerبرای وظیفه طبقه‌بندی تصاویر به‌کار گرفته شدند و عملکرد آن‌ها در مراحل مختلف رشد و، در مرحله چهارم، در طبقه‌بندی وضعیت سلامت ساختارهای مورد بررسی ارزیابی شد .برای CNNوViT، هر مرحله رشد به‌صورت مستقل مورد آموزش و ارزیابی قرار گرفت. بدین ترتیب، عملکرد مدل‌ها در پنج مرحله تکوینی شامل Stage1تا Stage5 بررسی شد و اثر تفاوت در ویژگی‌های ظاهری، پیچیدگی ساختاری، تعداد نمونه‌ها و توزیع کلاس‌ها بر عملکرد مدل مورد ارزیابی قرار گرفت. در مرحله چهارم، به دلیل پیچیدگی بیشتر تصاویر و وجود کلاس‌های مرتبط با نوع ساختار و وضعیت سلامت، مسئله طبقه‌بندی با جزئیات بیشتری بررسی شد. داده‌های هر مرحله پیش از آموزش به دو زیرمجموعه آموزش و اعتبارسنجی تقسیم شدند و این جداسازی در طول آزمایش‌ها حفظ شد. تصاویر مجموعه اعتبارسنجی در به‌روزرسانی پارامترهای مدل استفاده نشدند و پس از آموزش برای ارزیابی عملکرد مدل به کار رفتند. از آنجا که مجموعه داده مستقل آزمون در این پژوهش وجود نداشت، نتایج گزارش‌شده به‌عنوان عملکرد روی مجموعه اعتبارسنجی تفسیر شدند و به‌عنوان شاخص قطعی تعمیم‌پذیری مدل به داده‌های دیده‌نشده در نظر گرفته نشدند. طراحی آزمایش به‌گونه‌ای انجام شد که تفاوت میان وظایف تشخیص و طبقه‌بندی نیز به‌صورت روشن حفظ شود. بنابراین، معیارهای ارزیابی YOLOv8 بر اساس ماهیت مسئله تشخیص شیء، شاملPrecision ، Recall و mAP، و معیارهای CNN و ViT بر اساس ماهیت مسئله طبقه‌بندی، شامل Accuracy، Precision، Recall و F1-score در نظر گرفته شدند. از این رو، عملکرد YOLOv8 مستقیماً با دقت طبقه‌بندی CNN و ViT مقایسه نشد؛ بلکه نتایج سه معماری از نظر قابلیت‌های مکمل برای تشکیل اجزای بینایی یک سامانه پایش خودکار بررسی شدند. این طراحی آزمایشی امکان شناسایی نقاط قوت و محدودیت‌های هر معماری را در شرایط یکسان داده‌ای فراهم کرد. در نهایت، نتایج به‌عنوان مبنایی برای انتخاب و ترکیب مؤلفه‌های مناسب در توسعه آتی سامانه‌های هوشمند پایش رشد و سلامت ارکیده در شرایط کشت‌بافت و تولید گیاهچه در نظر گرفته شد. در چنین سامانه‌ای، اطلاعات حاصل از تشخیص و طبقه‌بندی تصویری می‌تواند در مراحل بعدی به‌عنوان ورودی برای تصمیم‌گیری، مدیریت نمونه‌ها و توسعه سامانه‌های خودکار و رباتیک مورد استفاده قرار گیرد.

۳-نتایج و بحث

در این پژوهش، عملکرد سه معماری یادگیری عمیق شامل YOLOv8، شبکه عصبی کانولوشنی (CNN) و (ViT) برای تحلیل تصاویر ارکیده Phalaenopsis در پنج مرحله تکوینی مورد بررسی قرار گرفت. با توجه به تفاوت ماهیت وظایف این مدل‌ها، YOLOv8 برای تشخیص و مکان‌یابی ساختارهای گیاهی و مدل‌های CNN و ViT برای طبقه‌بندی تصاویر مورد استفاده قرار گرفتند؛ بنابراین، نتایج این سه معماری به‌عنوان عملکرد مدل‌های یکسان در یک وظیفه مشترک تفسیر نمی‌شوند، بلکه قابلیت‌های مکمل آن‌ها برای تحلیل تصویری گیاه بررسی شده است. نتایج نشان داد که عملکرد مدل‌ها در مراحل مختلف رشد یکسان نبود و تحت تأثیر عواملی از جمله پیچیدگی مورفولوژیک تصاویر، تعداد نمونه‌های آموزشی، تعداد و توزیع کلاس‌ها و ویژگی‌های تصویری هر مرحله قرار داشت. در مراحل ابتدایی و انتهایی که ساختارهای تصویری متمایزتر بودند، مدل‌های طبقه‌بندی توانستند عملکرد بالاتری در مجموعه اعتبارسنجی نشان دهند؛ در حالی که مرحله چهارم، به دلیل وجود هم‌زمان ساختارهای PLBو برگ و همچنین افزایش پیچیدگی طبقه‌بندی وضعیت سلامت، چالش‌برانگیزتر بود. از آنجا که در این پژوهش مجموعه داده مستقل آزمون وجود نداشت، مقادیرAccuracy ، Precision، Recall و F1-score گزارش‌شده برای CNN و ViT بیانگر عملکرد مدل‌ها روی زیرمجموعه اعتبارسنجی هستند و نباید به‌عنوان شواهد قطعی از تعمیم مدل به داده‌های کاملاً دیده‌نشده تفسیر شوند. این موضوع به‌ویژه با توجه به حجم محدود داده‌ها در تفسیر نتایج مراحل مختلف اهمیت دارد. در ادامه، نتایج هر یک از سه معماری به تفکیک ارائه و تحلیل می‌شود. ابتدا عملکرد CNN در پنج مرحله رشد بررسی می‌شود، سپس نتایج ViT و در نهایت عملکرد YOLOv8 در تشخیص و مکان‌یابی ساختارهای هدف ارائه خواهد شد. در پایان، عملکرد سه رویکرد از نظر قابلیت‌ها، محدودیت‌ها و نقش احتمالی آن‌ها در توسعه سامانه‌های آینده پایش خودکار ارکیده در شرایط کشت‌بافت مورد بحث قرار می‌گیرد.

۳-۱- عملکرد مدل YOLO

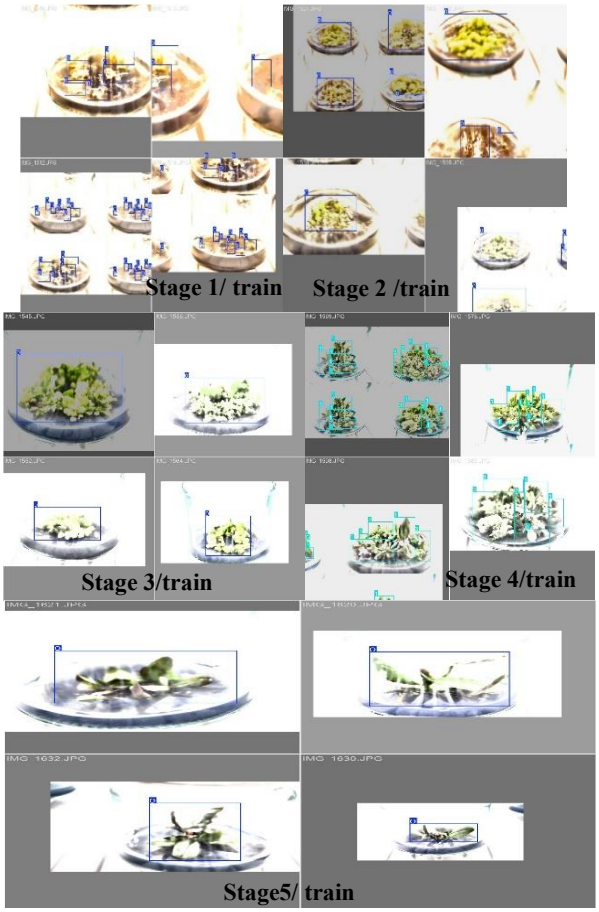
به‌منظور بررسی توانایی مدل YOLOv8 در تشخیص و مکان‌یابی ساختارهای تصویری ارکیده Phalaenopsis، مدل برای هر یک از پنج مرحله رشد به‌صورت مستقل آموزش داده شد. در این آزمایش، YOLOv8برای شناسایی و مکان‌یابی ساختارهای هدف، شامل برگ و پروتوکورم‌مانند(PLB)، مورد استفاده قرار گرفت و عملکرد آن با استفاده از معیارهای Precision، Recall، mAP@0.5 و mAP@0.5:0.95 ارزیابی شد. نتایج حاصل از ارزیابی مدل در پنج مرحله رشد در جدول ۳ ارائه شده است.

جدول۳- مقایسه عملکرد YOLOv8 در پنج مرحله رشد

Table 3- Comparison of YOLOv8 Performance Across Five Growth Stages

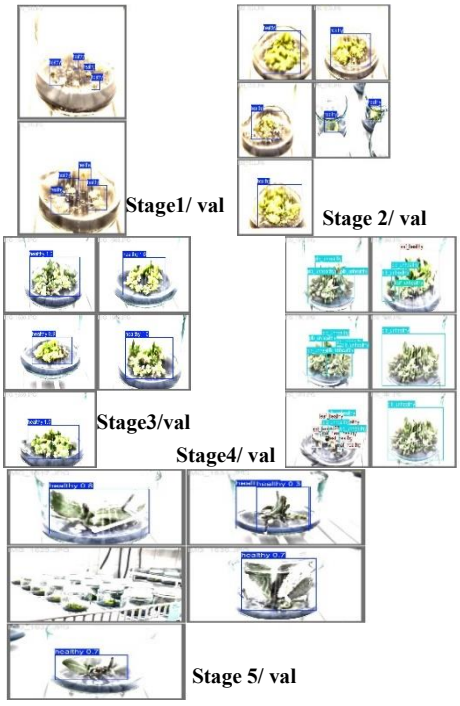
Stage	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Stage1	0.420	0.275	0.307	0.069
Stage2	0.970	0.667	0.675	0.452
Stage3	0.991	1.000	0.995	0.791
Stage4	0.388	0.146	0.072	0.029
Stage5	0.750	0.490	0.500	0.181

نتایج جدول ۳ نشان می‌دهد که عملکرد YOLOv8 در مراحل مختلف رشد یکسان نبوده و تفاوت قابل توجهی میان مراحل مشاهده می‌شود. بهترین عملکرد کلی مدل در Stage3 به دست آمد؛ به‌طوری‌که مقدار Precision برابر با ۰/۹۹۱، Recall برابر با ۱/۰۰۰، mAP@0.5 برابر با ۰/۹۹۵ و mAP@0.5:0.95 برابر با ۰/۷۹۱ بود. این نتایج نشان می‌دهد که در این مرحله، مدل توانایی بالایی در شناسایی و مکان‌یابی ساختارهای هدف داشته است. در مقابل، Stage4ضعیف‌ترین عملکرد را نشان داد. مقادیرPrecision، Recall، mAP@0.5 و mAP@0.5:0.95 در این مرحله به‌ترتیب ۰/۳۸۸، ۰/۱۴۶، ۰/۰۷۲ و ۰/۰۲۹ به دست آمدند. کاهش هم‌زمان Precision و Recall و مقادیر پایین mAP نشان می‌دهد که تشخیص و مکان‌یابی ساختارهای هدف در این مرحله برای مدل دشوارتر بوده است. این موضوع می‌تواند با پیچیدگی بیشتر تصاویر Stage4، وجود هم‌زمان ساختارهای PLB و برگ، تفاوت اندازه و شکل اهداف، و شباهت یا هم‌پوشانی برخی ساختارهای تصویری مرتبط باشد. Stage1 نیز عملکرد پایین‌تری نسبت به مراحل ۲ و ۳ نشان داد و مقادیر Precision برابر با ۰/۴۲۰، Recall برابر با ۰/۲۷۵، mAP@0.5 برابر با ۰/۳۰۷ و mAP@0.5:0.95 برابر با ۰/۰۶۹ به دست آمد. پایین بودن Recall در این مرحله نشان می‌دهد که بخشی از ساختارهای واقعی موجود در تصاویر توسط مدل شناسایی نشده‌اند. حجم محدود داده‌های این مرحله نیز می‌تواند یکی از عوامل مؤثر در عملکرد پایین‌تر باشد؛ با این حال، با توجه به نبود مجموعه آزمون مستقل، این موضوع به‌عنوان یک تفسیر احتمالی در نظر گرفته می‌شود. در Stage2، مقادیرPrecision، Recall، mAP@0.5 و mAP@0.5:0.95 به‌ترتیب ۰/۹۷۰، ۰/۶۶۷، ۰/۶۷۵ و ۰/۴۵۲ به دست آمد. در Stage5 نیز این مقادیر به‌ترتیب ۰/۷۵۰، ۰/۴۹۰، ۰/۵۰۰ و ۰/۱۸۱ بودند. بنابراین، نتایج نشان می‌دهد که توانایی YOLOv8 در تشخیص و مکان‌یابی ساختارهای هدف به مرحله رشد وابسته بوده و احتمالاً تحت تأثیر ویژگی‌های ظاهری، پیچیدگی ساختاری و حجم داده‌های موجود در هر مرحله قرار گرفته است. با توجه به اینکه YOLOv8 در این پژوهش برای تشخیص و مکان‌یابی و نه طبقه‌بندی مستقیم وضعیت سلامت استفاده شد، معیارهای عملکرد آن مستقیماً با Accuracy مدل‌های CNN و ViT قابل مقایسه نیستند. با این حال، نتایج YOLOv8 نشان می‌دهد که این معماری می‌تواند به‌عنوان یک مؤلفه مکمل برای استخراج موقعیت ساختارهای گیاهی در یک سامانه بینایی ماشین مورد استفاده قرار گیرد. چنین اطلاعات مکانی می‌تواند در مراحل بعدی توسعه سامانه‌های پایش خودکار، به‌عنوان ورودی برای تحلیل دقیق‌تر ساختارهای گیاهی و تصمیم‌گیری استفاده شود.



شکل ۳- تصاویر آموزشی هر ۵ مرحله رشد ارکیده با پردازش yolo

Fig 3. Training images for all five orchid growth stages used in the YOLO model



شکل۴- تصاویر ارزیابی هر ۵ مرحله رشد ارکیده با پردازش yolo

Fig 4. Photos of all 5 stages of orchid growth processed using YOLO.

۳-۲- عملکرد مدل CNN

مدل CNN برای طبقه‌بندی تصاویر مربوط به پنج مرحله رشد ارکیده Phalaenopsis به‌صورت مستقل آموزش داده شد و عملکرد آن بر اساس مجموعه‌های اعتبارسنجی (Validation) ارزیابی گردید. نتایج نشان داد که عملکرد مدل در مراحل مختلف رشد یکسان نبود و به میزان تمایزپذیری ویژگی‌های تصویری، تعداد نمونه‌ها و پیچیدگی طبقه‌بندی وابسته بود. در Stage1، Stage2، Stage3 و Stage5، مدل به دقت اعتبارسنجی ۱۰۰ درصد دست یافت. با توجه به تعداد محدود تصاویر اعتبارسنجی و نبود مجموعه آزمون مستقل، این مقادیر باید به‌عنوان عملکرد روی مجموعه اعتبارسنجی تفسیر شوند و نمی‌توانند به‌تنهایی بیانگر قابلیت تعمیم مدل به داده‌های جدید باشند. بنابراین، دستیابی به Accuracy برابر با ۱۰۰ درصد در این مراحل، لزوماً به معنای تعمیم کامل مدل نیست و می‌تواند تحت تأثیر حجم کم نمونه‌های اعتبارسنجی و قابل تفکیک بودن الگوهای تصویری قرار گرفته باشد. در مقابل، Stage4 پیچیده‌ترین مرحله برای طبقه‌بندی با CNN بود.

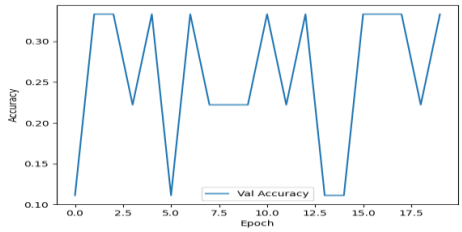
در این مرحله، مدل باید چهار کلاس شامل برگ سالم(leaf_healthy)، برگ ناسالم (leaf_unhealthy)، PLB سالم (plb_healthy) و PLB ناسالم (plb_unhealthy) را از یکدیگر تفکیک می‌کرد. بر اساس گزارش طبقه‌بندی، دقت اعتبارسنجی مدل برابر با ۰/۳۳۳ (۳۳/۳ درصد) به دست آمد. تحلیل جزئی‌تر نتایج نشان داد که از مجموع ۹ نمونه اعتبارسنجی Stage4 ، سه نمونه مربوط به کلاس plb_healthy بودند که هر سه نمونه به‌درستی شناسایی شدند؛ بنابراین Recall این کلاس برابر با ۱/۰۰ بود. با این حال، Precision این کلاس تنها ۰/۳۳۳ و F1-score برابر با ۰/۵۰ بود. برای سه کلاس دیگر، یعنی leaf_healthy، leaf_unhealthy و plb_unhealthy=، مقادیر Precision، Recall و F1-score برابر با صفر گزارش شد. این الگو نشان می‌دهد که مدل در Stage4 توانایی تفکیک مناسب چهار کلاس را نداشته و پیش‌بینی‌های آن به‌شدت به سمت یک کلاس متمایل شده است.

این رفتار را می‌توان با تعداد بسیار محدود نمونه‌ها، عدم توازن بین کلاس‌ها و شباهت ظاهری برخی نمونه‌های سالم و ناسالم مرتبط دانست. بنابراین، Accuracy برابر با ۳۳/۳ درصد به‌تنهایی تصویر کاملی از عملکرد مدل ارائه نمی‌دهد و استفاده هم‌زمان از Precision، Recall، F1-score و ماتریس درهم‌ریختگی برای تحلیل عملکرد Stage4 ضروری است. در مجموع، نتایج نشان داد که CNN در مراحل دارای الگوهای تصویری متمایز، عملکرد اعتبارسنجی بالایی داشت، اما با افزایش پیچیدگی طبقه‌بندی و ورود هم‌زمان ویژگی نوع بافت و وضعیت سلامت در Stage4، عملکرد آن به‌شدت کاهش یافت (شکل‌های ۵ و ۶). این یافته نشان می‌دهد که عملکرد بالای CNN در مراحل ساده‌تر را نباید بدون درنظرگرفتن اندازه مجموعه اعتبارسنجی و پیچیدگی مسئله به کل سامانه تعمیم داد. از این رو، افزایش تعداد نمونه‌ها، ایجاد توزیع متوازن‌تر بین کلاس‌ها و استفاده از مجموعه آزمون مستقل در مطالعات آینده برای ارزیابی قابلیت تعمیم مدل ضروری است. خلاصه عملکرد مدل در پنج مرحله در جدول ۴ ارائه شده است.

جدول۴- مقایسه عملکرد CNN در پنج Stage

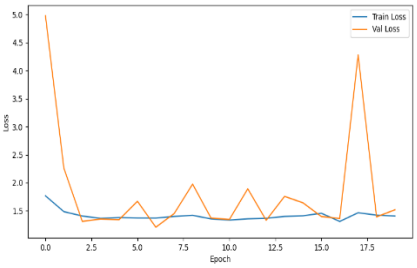
Table 4- Comparison of CNN Performance Across Five Stages

Stage	Validation Accuracy	وضعیت عملکرد	تفسیر
Stage1	100%	بسیار بالا	تفکیک مناسب الگوهای تصویری
Stage2	100%	بسیار بالا	تفکیک مناسب الگوهای تصویری
Stage3	100%	بسیار بالا	همگرایی و تفکیک مناسب
Stage4	33.3%	ضعیف	مشکل در تفکیک چهار کلاس سلامت و نوع بافت
Stage5	100%	بسیار بالا	تفکیک مناسب الگوهای تصویری



شکل۵- نمودار روند تغییرات دقت (Accuracy) مدل CNN در طول آموزش Stage4

Fig 5. Trend of CNN model accuracy changes during Stage4 training.



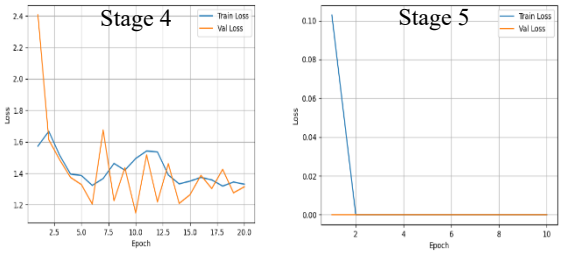
شکل۶- نمودار روند تغییرات تابع خطا (Loss) مدل CNN در مرحله آموزش Stage4

Fig 6. Trend of the CNN model's loss function during the Stage4 training phase.

۳-۳- عملکرد مدل Vision Transformer (ViT)

مدل (ViT) به‌منظور بررسی عملکرد معماری‌های مبتنی بر ترنسفورمر در طبقه‌بندی تصاویر ارکیده Phalaenopsis در پنج مرحله رشد مورد ارزیابی قرار گرفت. عملکرد مدل در مراحل مختلف یکسان نبود و تغییرات آن با پیچیدگی تصویری، تعداد نمونه‌های

آموزشی و توزیع کلاس‌ها همراه بود. در Stage1، Stage2 و Stage3، مدل عملکرد بسیار بالایی در مجموعه‌های اعتبارسنجی نشان داد و در برخی مراحل به دقت اعتبارسنجی کامل دست یافت. با این حال، با توجه به تعداد محدود تصاویر موجود در هر مرحله و نبود یک مجموعه آزمون مستقل، این نتایج باید صرفاً به‌عنوان عملکرد مدل بر روی داده‌های اعتبارسنجی موجود تفسیر شوند و نمی‌توانند به‌تنهایی بیانگر قابلیت تعمیم مدل به تصاویر جدید باشند. بنابراین، دستیابی به Accuracy برابر با ۱/۰۰ در این مراحل نباید به‌عنوان شواهد قطعی از عملکرد عمومی مدل در نظر گرفته شود. در Stage4، عملکرد ViT به‌طور محسوسی کاهش یافت (شکل ۷). دقت اعتبارسنجی مدل در حدود ۳۳/۳۳ درصد باقی ماند و روند تغییرات تابع خطا نیز نسبت به مراحل ساده‌تر ناپایدارتر بود. این مرحله شامل تفکیک چهار کلاس PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم بود و بنابراین از نظر تعداد کلاس‌ها و شباهت ویژگی‌های تصویری، مسئله طبقه‌بندی پیچیده‌تری نسبت به مراحل دیگر ایجاد می‌کرد. محدود بودن تعداد نمونه‌ها، عدم توازن کلاس‌ها و شباهت ظاهری میان نمونه‌های سالم و ناسالم می‌تواند از عوامل مؤثر بر کاهش عملکرد مدل در این مرحله باشد. در Stage5، مدل رفتار آموزشی پایدارتری نشان داد. کاهش هم‌زمان Training Loss و Validation Loss و نزدیک بودن روند دو منحنی نشان داد که مدل در این مرحله به همگرایی مناسبی دست یافته است. با این حال، با توجه به حجم محدود داده‌ها، این رفتار به‌تنهایی برای رد قطعی بیش‌برازش کافی نیست و ارزیابی قابلیت تعمیم مدل به داده‌های جدید نیازمند مجموعه آزمون مستقل و بزرگ‌تر است. به‌منظور بررسی دقیق‌تر رفتار آموزشی ViT و جلوگیری از افزایش غیرضروری حجم مقاله، نمودارهای مربوط به Stage4 و Stage5 ارائه شدند (شکل ۷). Stage4 به‌عنوان چالش‌برانگیزترین مرحله از نظر عملکرد طبقه‌بندی و Stage5 به‌عنوان مرحله‌ای با رفتار آموزشی پایدارتر انتخاب شدند. مقایسه این دو مرحله امکان مشاهده تفاوت رفتار مدل در شرایط داده‌ای و پیچیدگی متفاوت را فراهم می‌کند. در مجموع، نتایج نشان داد که ViT در مراحل دارای الگوهای تصویری متمایز عملکرد اعتبارسنجی بالایی داشت، اما در Stage4 با افزایش پیچیدگی طبقه‌بندی و محدودیت داده‌ها، عملکرد آن کاهش یافت. بنابراین، نتایج این آزمایش نشان می‌دهد که عملکرد ViT تنها به معماری مدل وابسته نیست و حجم داده، توزیع نمونه‌ها، توازن کلاس‌ها و پیچیدگی ویژگی‌های تصویری نیز نقش مهمی در عملکرد آن دارند. این یافته با نتایج CNN در Stage4 نیز هم‌راستا است و نشان می‌دهد که این مرحله، به دلیل ماهیت پیچیده‌تر مسئله سلامت و نوع بافت، چالش اصلی طبقه‌بندی در مجموعه داده حاضر بوده است.



شکل۷- نمودار عملکرد vit در مرحله ۴ و ۵

Fig 7. VIT performance chart for stages 4 and 5

۳-۴- مقایسه عملکرد مدل‌ها

مقایسه نتایج سه معماری مورد بررسی نشان داد که عملکرد مدل‌ها به نوع وظیفه، ویژگی‌های داده، پیچیدگی تصاویر، تعداد نمونه‌ها و توزیع کلاس‌ها وابسته است. از آنجا که YOLOv8 به‌عنوان مدل تشخیص و مکان‌یابی شیء و CNN و ViT به‌عنوان مدل‌های طبقه‌بندی تصویر مورد استفاده قرار گرفتند، مقایسه عملکرد آن‌ها بر اساس یک شاخص واحد، مانندAccuracy، مناسب نیست. بنابراین، هر مدل بر اساس معیارهای متناسب با وظیفه خود مورد بررسی قرار گرفت YOLOv8. برای تشخیص و مکان‌یابی ساختارهای هدف شامل برگ و PLB استفاده شد. نتایج نشان داد که عملکرد این مدل در مراحل مختلف رشد تفاوت قابل توجهی دارد. بهترین عملکرد YOLOv8 در Stage3 مشاهده شد؛ به‌طوری‌که Precision برابر با ۰/۹۹۱، Recall برابر با ۱/۰۰۰، mAP@0.5 برابر با ۰/۹۹۵ و mAP@0.5:0.95 برابر با ۰/۷۹۱ به دست آمد. در مقابل، عملکرد مدل در Stage4 به‌طور محسوسی کاهش یافت و مقادیر Precision، Recall، mAP@0.5 و mAP@0.5:0.95 به‌ترتیب ۰/۳۸۸، ۰/۱۴۶، ۰/۰۷۲ و ۰/۰۲۹ گزارش شد. این تفاوت نشان می‌دهد که توانایی YOLOv8 در مکان‌یابی ساختارهای گیاهی به ویژگی‌های تصویری و پیچیدگی ساختارهای موجود در هر مرحله وابسته است. بنابراین، بر اساس داده‌های این پژوهش، YOLOv8 را می‌توان به‌عنوان یک مؤلفه مناسب برای مکان‌یابی و استخراج نواحی مورد نظر در یک سامانه بینایی ماشین در نظر گرفت، اما عملکرد آن در همه مراحل یکسان نبود. در مورد CNN، نتایج نشان داد که این مدل در مراحل دارای الگوهای تصویری متمایز عملکرد بسیار بالایی داشت. مدل در Stage1، Stage2، Stage3 و Stage5 به Accuracy برابر با ۱۰۰ درصد در مجموعه اعتبارسنجی دست یافت. با این حال، با توجه به تعداد محدود نمونه‌های اعتبارسنجی و نبود مجموعه آزمون مستقل، این مقادیر نباید به‌عنوان شواهد قطعی قابلیت تعمیم مدل تفسیر شوند. در مقابل، Stage4 چالش‌برانگیزترین مرحله برای CNN بود و Accuracy برابر با ۰/۳۳۳ به دست آمد. همچنین Macro F1-score برابر با ۰/۱۲۵ و Weighted F1-score برابر با ۰/۱۶۷ گزارش شد. تحلیل کلاس‌به‌کلاس نشان داد که مدل برای کلاس plb_healthy دارای Recall برابر با ۱ بود، اما Precision این کلاس

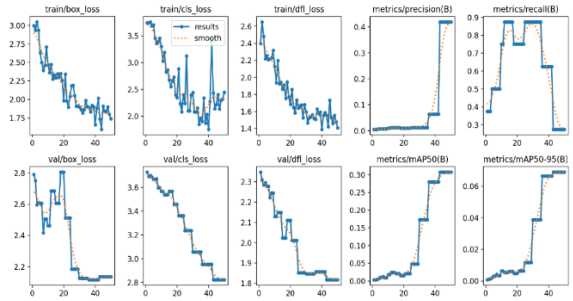
تنها ۰/۳۳۳ و F1-score برابر با ۰/۵۰ بود؛ در سه کلاس دیگر، مقادیر Precision، Recall و F1-score برابر صفر گزارش شدند. این الگو نشان می‌دهد که مدل در Stage4 قادر به تفکیک مناسب چهار کلاس سلامت و نوع بافت نبوده و پیش‌بینی‌های آن به‌شدت به سمت یک کلاس متمایل شده است. مدل (ViT) نیز رفتار متفاوتی در مراحل مختلف نشان داد. در برخی مراحل، مدل به عملکرد اعتبارسنجی بسیار بالایی دست یافت، در حالی که در Stage4 عملکرد آن کاهش محسوسی داشت. این رفتار نشان می‌دهد که عملکرد ViT در مجموعه داده حاضر به حجم محدود داده، توزیع کلاس‌ها و پیچیدگی ویژگی‌های تصویری حساس بوده است. بنابراین، اگرچه ViTتوانایی یادگیری روابط سراسری میان بخش‌های مختلف تصویر را دارد، نتایج این پژوهش از برتری مطلق آن نسبت به CNN یا YOLOپشتیبانی نمی‌کند. در عوض، یافته‌ها نشان می‌دهند که قابلیت هر معماری باید با توجه به وظیفه موردنظر و شرایط داده ارزیابی شود. یکی از مهم‌ترین یافته‌های این پژوهش، تأثیر ساختار و توزیع داده‌ها بر عملکرد مدل‌ها بود. در مراحلی که ویژگی‌های تصویری نمونه‌ها تمایز بیشتری داشتند، مدل‌های طبقه‌بندی عملکرد بسیار بالایی نشان دادند؛ در حالی که Stage4، که شامل تفکیک هم‌زمان چهار کلاس شامل PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم بود، برای مدل‌های طبقه‌بندی چالش‌برانگیزتر بود. تعداد محدود نمونه‌ها و توزیع نامتوازن کلاس‌ها نیز می‌تواند در این کاهش عملکرد مؤثر باشد. بنابراین، نتایج نشان می‌دهد که در توسعه سامانه‌های بینایی ماشین برای پایش گیاه، انتخاب معماری به‌تنهایی تعیین‌کننده عملکرد نیست و کیفیت، حجم، تنوع و توزیع مناسب داده‌ها نیز اهمیت اساسی دارند. نتایج پژوهش حاضر با روند مطالعات اخیر در زمینه کاربرد یادگیری عمیق در کشاورزی هوشمند همخوانی دارد. در مطالعه (De Silva & Brown,2023) ترکیب CNN و ViT برای تشخیص بیماری‌های گیاهی بررسی شد و نتایج نشان داد که ViT در استخراج روابط سراسری تصویر و CNN در استخراج ویژگی‌های محلی قابلیت‌های متفاوتی دارند. این یافته با نتایج پژوهش حاضر هم‌راستا است؛ زیرا در این مطالعه نیز رفتار CNN و ViT در مراحل مختلف یکسان نبود و تغییرات عملکرد آن‌ها با پیچیدگی و ساختار داده‌ها همراه بود .همچنین (Thakur et al,.2022) گزارش کردند که مدل‌های مبتنی بر ViT در شناسایی بیماری‌های گیاهی قادر به دستیابی به دقت‌های بالا بوده و در استخراج الگوهای پیچیده تصویری قابلیت مناسبی دارند. یافته‌های پژوهش حاضر نیز نشان داد که ViT در برخی مراحل دارای عملکرد اعتبارسنجی بالا بود؛ با این حال، کاهش عملکرد آن در Stage4 نشان داد که استفاده از این معماری در مجموعه‌داده‌های کوچک و دارای کلاس‌های نامتوازن نیز می‌تواند با محدودیت همراه باشد. در زمینه تشخیص و مکان‌یابی، (Tripath et al., 024) عملکرد معماری‌های YOLOv8 و YOLOv9 را در تشخیص بیماری‌های گیاهی بررسی کردند و قابلیت مدل‌های YOLOرا برای تشخیص و مکان‌یابی نواحی مورد نظر نشان دادند. در پژوهش حاضر نیز YOLOv8 به‌عنوان یک مدل تشخیص شیء، قابلیت مناسبی برای مکان‌یابی ساختارهای هدف نشان داد، هرچند عملکرد آن بین مراحل مختلف تفاوت داشت و در Stage4 کاهش محسوسی مشاهده شد .همچنین (Khanal et al,.2024) نشان دادند که مدل‌های مبتنی بر ViT برای وظایف طبقه‌بندی و مدل‌های مبتنی بر YOLO برای شناسایی هم‌زمان و مکان‌یابی نواحی آلوده می‌توانند کاربردهای متفاوت و مکملی داشته باشند. این موضوع با طراحی پژوهش حاضر سازگار است؛ زیرا YOLOv8 در این مطالعه برای مکان‌یابی ساختارهای گیاه و CNN و ViT برای طبقه‌بندی تصاویر به کار گرفته شدند. بنابراین، این سه مدل را می‌توان به‌عنوان معماری‌هایی با وظایف مکمل در نظر گرفت، نه سه مدل کاملاً هم‌وظیفه که بر اساس یک معیار واحد با یکدیگر رتبه‌بندی شوند. از سوی دیگر، (Özden, 2025) نشان داد که اگرچهCNN ها می‌توانند در مجموعه‌داده‌های کنترل‌شده عملکرد مطلوبی داشته باشند، اما داده‌های پیچیده و نامتوازن می‌توانند عملکرد آن‌ها را کاهش دهند و معماری‌های ترنسفورمری یا مدل‌های ترکیبی در برخی شرایط عملکرد پایدارتری ارائه کنند. یافته‌های پژوهش حاضر نیز اهمیت ساختار داده را نشان داد؛ به‌ویژه در Stage4 که عملکرد CNN به‌شدت کاهش یافت. با این حال، به دلیل محدودیت حجم مجموعه داده حاضر، نمی‌توان بر اساس این آزمایش نتیجه گرفت که ViT در همه شرایط قابلیت تعمیم بیشتری از CNN دارد. در مجموع، مقایسه نتایج با مطالعات پیشین نشان می‌دهد که یافته‌های این پژوهش در چارچوب تحقیقات جدید بینایی ماشین و یادگیری عمیق در کشاورزی قرار می‌گیرد. با این حال، وجه تمایز پژوهش حاضر، ارزیابی نظام‌مند سه رویکرد با وظایف مکمل شامل YOLOv8 برای مکان‌یابی، CNN برای طبقه‌بندی و ViT برای تحلیل تصویری در پنج مرحله رشد ارکیده Phalaenopsis تحت شرایط حجم محدود داده و پیچیدگی متفاوت مراحل است .این ارزیابی امکان شناسایی نقاط قوت و محدودیت‌های هر رویکرد را فراهم می‌کند و می‌تواند برای انتخاب مؤلفه‌های مناسب در توسعه سامانه‌های آینده پایش خودکار کشت‌بافت و تولید گیاهچه مورد استفاده قرار گیرد. به‌طور کلی، نتایج این پژوهش نشان می‌دهد که:

YOLOv8: مناسب برای تشخیص و مکان‌یابی ساختارهای هدف مانند برگ وPLB، با عملکرد متفاوت در مراحل مختلف رشد.

CNN: مناسب برای طبقه‌بندی تصاویر در شرایطی که الگوهای تصویری متمایز باشند، اما حساس به محدودیت تعداد نمونه‌ها و عدم توازن کلاس‌ها، به‌ویژه در Stage4 .

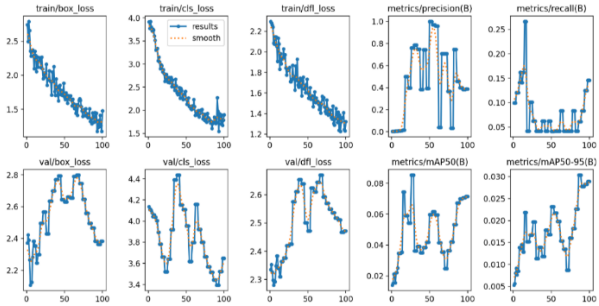
ViT: دارای قابلیت مناسب برای یادگیری روابط و ویژگی‌های سراسری تصویر، اما عملکرد آن در مجموعه داده حاضر نیز به حجم، توزیع و پیچیدگی داده‌ها وابسته بود.

بنابراین، نتایج این پژوهش نشان نمی‌دهد که یک معماری در همه شرایط «بهترین» است؛ بلکه نشان می‌دهد هر معماری برای یک وظیفه مشخص مزیت متفاوتی دارد .ترکیب این قابلیت‌ها می‌تواند در مطالعات آینده به طراحی یک چارچوب چندمرحله‌ای برای پایش خودکار رشد و سلامت ارکیده Phalaenopsis کمک کند؛ برای مثال، YOLOv8 می‌تواند ابتدا نواحی هدف را مکان‌یابی کند و سپس مدل‌های طبقه‌بندی برای تحلیل وضعیت یا مرحله رشد مورد استفاده قرار گیرند. چنین چارچوبی می‌تواند در آینده به‌عنوان بخش بینایی یک سامانه پایش خودکار در محیط‌های کشت‌بافت و تولید گیاهچه توسعه یابد . طبق شکل ۸ تا ۱۰ و جدول ۵ ارائه شده است.



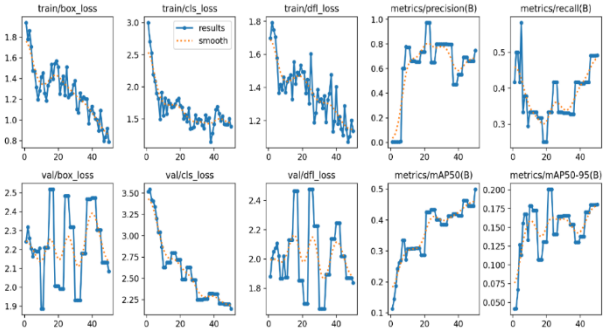
شکل۸- نمودار مدل yolo : trin/box-loss , val/box-loss مرحله ۱

Fig 8. yolo model chart: trin/box-loss, val/box-loss Stage1



شکل۹- نمودار مدل yolo: trin/box-loss , val/box-loss مرحله ۴

Fig 9. yolo model chart: trin/box-loss, val/box-loss Stage4



شکل۱۰- نمودار مدل yolo: trin/box-loss , val/box-loss مرحله ۵

Fig 10. yolo model chart: trin/box-loss, val/box-loss Stage5

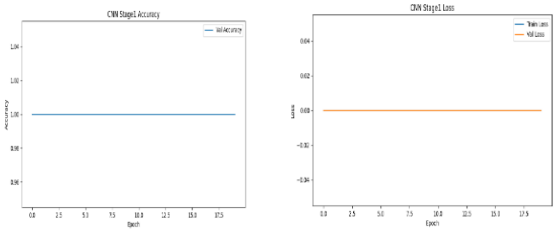
جدول۵- نتایج مدل yolo برای تمامی مرحله های مراحل رشد

Table 5 - Results of the YOLO model for all stages of growth

Stage	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50-95 (%)	Error Rate (%)
stage1	41.97	27.46	30.74	6.89	58.03
stage2	96.99	66.67	67.55	45.18	3.01
stage3	99.1	100	99.5	79.06	0.9
stage4	38.77	14.58	7.16	2.89	61.23
stage5	74.65	49.19	49.9	18.05	25.35

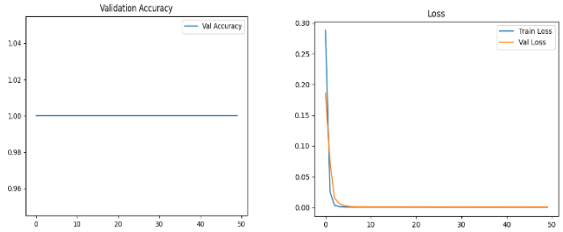
نتایج مدل CNN برای طبقه‌بندی مراحل رشد و در Stage4، برای تفکیک وضعیت سلامت ساختارهای گیاهی مورد بررسی قرار گرفت. عملکرد مدل در مراحل مختلف رشد یکسان نبود و تفاوت مشاهده‌شده را می‌توان با تعداد محدود نمونه‌های هر مرحله، اندازه کوچک مجموعه اعتبارسنجی و تفاوت در پیچیدگی و تفکیک‌پذیری ویژگی‌های تصویری مرتبط دانست. در Stage1، Stage2، Stage3وStage5، مدل CNN به Accuracy برابر با ۱۰۰ درصد در مجموعه اعتبارسنجی دست یافت. همچنین، در این مراحل کاهش تابع خطای آموزش و اعتبارسنجی و رسیدن آن‌ها به مقادیر پایین، نشان‌دهنده همگرایی مدل با داده‌های موجود بود. با این حال، از آنجا که مجموعه داده پژوهش محدود بوده و تعداد تصاویر اعتبارسنجی در هر مرحله کم است، دستیابی به Accuracy برابر با ۱۰۰ درصد در این مراحل به‌تنهایی نمی‌تواند نشان‌دهنده تعمیم‌پذیری قطعی مدل به تصاویر جدید باشد. بنابراین، این نتایج باید به‌عنوان عملکرد مدل روی مجموعه اعتبارسنجی موجود تفسیر شوند و نه به‌عنوان اثبات عملکرد عمومی مدل. در Stage1 و Stage2، تفکیک‌پذیری مناسب ویژگی‌های تصویری مراحل رشد موجب شد مدل بتواند الگوهای موجود در تصاویر را با سرعت بیشتری یاد بگیرد. در Stage3 نیز مدل پس از طی چند دوره آموزش به عملکرد کامل در مجموعه اعتبارسنجی رسید. در Stage5، به دلیل قابل تشخیص‌تر بودن ویژگی‌های مورفولوژیک گیاه، مدل عملکرد بالایی نشان داد و Accuracy اعتبارسنجی به ۱۰۰ درصد رسید. در مقابل، Stage4 پیچیده‌ترین مسئله طبقه‌بندی در این پژوهش بود .در این مرحله، مدل CNN باید چهار کلاس شامل PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم را از یکدیگر تفکیک می‌کرد. برخلاف مراحل دیگر، شباهت ظاهری میان نمونه‌های سالم و ناسالم، تفاوت‌های ظریف بین کلاس‌ها و توزیع نامتوازن نمونه‌ها، مسئله طبقه‌بندی را دشوارتر کرد . Accuracyمدل در مجموعه اعتبارسنجی برابر با ۰/۳۳۳ (۳۳/۳ درصد) به دست آمد. بررسی جزئی‌تر نتایج Stage4 نشان داد که مدل عمدتاً در شناسایی کلاس plb_healthy موفق بوده است. برای این کلاس، Recall برابر با ۱/۰۰ و F1-score برابر با ۰/۵۰ به دست آمد، در حالی که برای کلاس‌های leaf_healthy، leaf_unhealthy و plb_unhealthy مقادیر Precision ، Recall و F1-score برابر با صفر گزارش شد. همچنین Macro F1-score برابر با ۰/۱۲۵ و Weighted F1-score برابر با ۰/۱۶۷ بود. این نتایج نشان می‌دهد

که Accuracy برابر با ۳۳/۳ درصد در Stage4 به معنی عملکرد مناسب در هر چهار کلاس نیست و بخش عمده پیش‌بینی‌های صحیح مربوط به یک کلاس بوده است. بنابراین، نتایج Stage4 اهمیت بررسی معیارهای Precision، Recall و F1-score را در کنار Accuracy نشان می‌دهد؛ به‌ویژه زمانی که تعداد نمونه‌ها محدود و توزیع کلاس‌ها نامتوازن باشد. عملکرد ضعیف مدل در این مرحله احتمالاً تحت تأثیر هم‌زمان محدودیت حجم داده، عدم توازن کلاس‌ها و شباهت ویژگی‌های ظاهری نمونه‌های سالم و ناسالم قرار گرفته است. ازاین‌رو، Accuracy بالای حاصل‌شده در مراحل ساده‌تر نباید به‌تنهایی به‌عنوان شاخص عملکرد عمومی CNN در نظر گرفته شود. نمودارهای Accuracy و Loss مربوط به Stage1 تا Stage5 در شکل‌های ۱۱ تا ۱۵ ارائه شده‌اند. این نمودارها روند تغییرات عملکرد مدل در طول آموزش و تفاوت رفتار یادگیری آن در مراحل مختلف رشد را نشان می‌دهند. بررسی هم‌زمان منحنی‌های آموزش و اعتبارسنجی، علاوه بر ارزیابی روند همگرایی، امکان بررسی نشانه‌های احتمالی بیش‌برازش را نیز فراهم می‌کند. در مجموع، نتایج نشان داد که CNN در مراحل دارای ویژگی‌های تصویری متمایز، عملکرد بالایی روی مجموعه اعتبارسنجی موجود ارائه کرد، اما در Stage4 که مسئله طبقه‌بندی پیچیده‌تر و داده‌ها نامتوازن‌تر بودند، عملکرد آن به‌طور محسوسی کاهش یافت. این یافته تأکید می‌کند که عملکرد یک مدل یادگیری عمیق تنها به معماری آن وابسته نیست و حجم، کیفیت، تنوع و توزیع کلاس‌های داده نقش مهمی در نتیجه نهایی دارند.



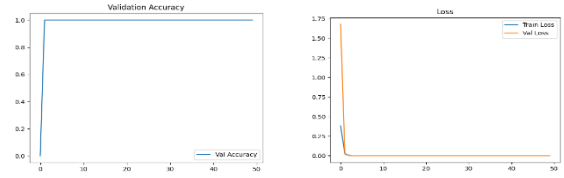
شکل۱۱- نمودارهای مدل CNN: Accuracy و Loss مرحله۱

Fig 11. CNN model chart: trin/box-loss, val/box-loss stage1



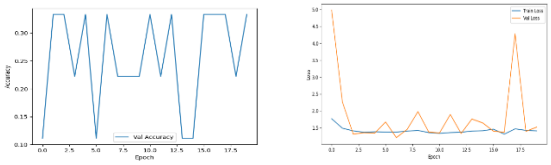
شکل۱۲- نمودار های مدل CNN: Accuracy و Loss مرحله۲

Fig 12. CNN model chart: trin/box-loss, val/box-loss stage2



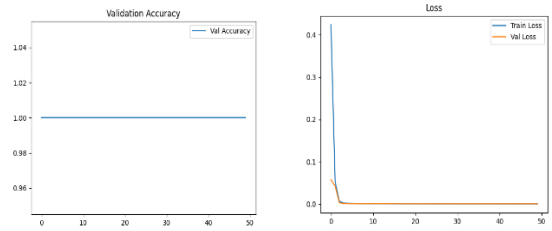
شکل۱۳- نمودار های مدل CNN: Accuracy و Loss مرحله۳

Fig 13. CNN model chart: trin/box-loss, val/box-loss stage3



شکل۱۴- نمودار های مدل CNN: Accuracy و Loss مرحله۴

Fig 14. CNN model chart: trin/box-loss, val/box-loss stage4

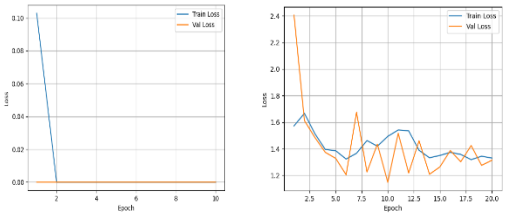


شکل۱۵- نمودار های مدل CNN: Accuracy و Loss مرحله۵

Fig 15. CNN model chart: trin/box-loss, val/box-loss stage5

مدل (ViT) به‌منظور بررسی توانایی معماری‌های مبتنی بر ترنسفورمر در استخراج ویژگی‌های پیچیده و یادگیری روابط سراسری میان نواحی مختلف تصاویر ارکیده Phalaenopsis مورد ارزیابی قرار گرفت. نتایج نشان داد که عملکرد ViT در مراحل مختلف رشد یکسان نبوده و به پیچیدگی تصاویر، تعداد نمونه‌های آموزشی و اعتبارسنجی و توزیع کلاس‌ها وابستگی قابل توجهی داشته است .در Stage1، Stage2 و Stage3، مدل در مجموعه اعتبارسنجی به Accuracy برابر با ۱۰۰ درصد دست یافت و مقادیر Precision، Recall و F1-score نیز برابر با ۱/۰۰ گزارش شدند. این نتایج نشان‌دهنده توانایی مدل در تفکیک نمونه‌های موجود در مجموعه اعتبارسنجی این مراحل است. با این حال، با توجه به تعداد محدود تصاویر و کوچک بودن مجموعه اعتبارسنجی، دستیابی به عملکرد ۱۰۰ درصدی را نمی‌توان به‌عنوان شواهد قطعی از تعمیم‌پذیری مدل به تصاویر جدید در نظر گرفت. بنابراین، این مقادیر باید به‌عنوان عملکرد ViTروی مجموعه اعتبارسنجی موجود تفسیر شوند. در Stage4، کاهش محسوسی در عملکرد مدل مشاهده شدAccuracy . اعتبارسنجی در طول فرآیند آموزش در محدوده ۱۱/۱۱ تا ۳۳/۳۳ درصد قرار داشت و مدل در تفکیک چهار کلاس شامل PLB سالم،PLB ناسالم، برگ سالم و برگ ناسالم با مشکل مواجه شد. این کاهش عملکرد با پیچیدگی بیشتر تصاویرStage4، شباهت ظاهری میان نمونه‌های سالم و ناسالم، تعداد محدود نمونه‌ها و توزیع نامتوازن کلاس‌ها قابل ارتباط است. همچنین، اختلاف و نوسان میان منحنی‌های Train Loss و Validation Loss در این مرحله نشان می‌دهد که رفتار یادگیری مدل در مقایسه با مراحل ساده‌تر ناپایدارتر بوده و توانایی تعمیم آن در این شرایط محدود شده است. در مقابل، در Stage5، روند آموزش مدل پایدارتر بود. مقادیر Train Lossو Validation Lossدر طول آموزش کاهش یافتند و در پایان به مقادیر

بسیار کوچک، در حدود 10^{-8}، رسیدند. هم‌زمانی روند کاهش این دو تابع خطا نشان‌دهنده همگرایی مناسب مدل با داده‌های موجود است. با این حال، به دلیل نبود مجموعه آزمون مستقل، این کاهش شدید Lossرا نمی‌توان به‌تنهایی به‌عنوان اثبات تعمیم‌پذیری بسیار بالای مدل در نظر گرفت. برای بررسی دقیق‌تر رفتار آموزشی ViT، نمودارهای مربوط به Stage4 و Stage5 در شکل ۱۶ ارائه شده‌اند. انتخاب این دو مرحله با هدف نشان دادن دو رفتار متفاوت مدل انجام شده است؛ به‌طوری‌که Stage4 بیانگر چالش‌برانگیزترین و ناپایدارترین وضعیت یادگیری و Stage5 نشان‌دهنده همگرایی پایدارتر مدل است. مقایسه این دو مرحله، تأثیر پیچیدگی داده‌ها و تعداد و توزیع نمونه‌ها بر رفتار آموزشی ViT را به‌صورت واضح‌تر نشان می‌دهد. در مجموع، نتایج نشان می‌دهد که ViT در مراحل دارای الگوهای تصویری متمایز، عملکرد بسیار بالایی روی مجموعه اعتبارسنجی موجود داشته است، اما در Stage4 با افزایش پیچیدگی مسئله و محدودیت داده‌ها، عملکرد آن به‌طور محسوسی کاهش یافته است. بنابراین، نتایج این مدل نیز نشان می‌دهد که انتخاب معماری یادگیری عمیق به‌تنهایی تضمین‌کننده عملکرد مطلوب نیست و حجم، کیفیت، تنوع و توزیع داده‌های آموزشی نقش اساسی در عملکرد و قابلیت تعمیم مدل دارند.



شکل۱۶- نمودارهای مدل vit: Accuracy و Loss مرحله۵

Fig 16. Vit model chart: trin/box-loss, val/box-loss Stage5

۴-نتیجه گیری

در این پژوهش، قابلیت سه معماری یادگیری عمیق شامل YOLOv8، شبکه عصبی کانولوشنی (CNN) و (ViT) برای تحلیل تصاویر ارکیده Phalaenopsis در پنج مرحله رشد و در شرایط متفاوت از نظر پیچیدگی تصویری و توزیع کلاس‌ها مورد ارزیابی و مقایسه قرار گرفت. نتایج نشان داد که عملکرد مدل‌ها در مراحل مختلف یکسان نبوده و حجم و توزیع داده‌ها، پیچیدگی ویژگی‌های تصویری و توازن کلاس‌ها نقش مهمی در رفتار و عملکرد مدل‌ها داشته‌اند. مدل YOLOv8 با هدف تشخیص و مکان‌یابی ساختارهای گیاهی، به‌ویژه برگ‌ها و ساختارهایPLB، مورد استفاده قرار گرفت. نتایج نشان داد که عملکرد این مدل بین مراحل مختلف رشد متفاوت است و در مراحل دارای ویژگی‌های تصویری متمایز، قابلیت مناسبی در تشخیص و مکان‌یابی نواحی هدف دارد. بهترین عملکرد YOLOv8در Stage3 با Precision برابر با ۰/۹۹۱، Recall برابر با ۱/۰۰، mAP@0.5 برابر با ۰/۹۹۵ و mAP@0.5:0.95 برابر با ۰/۷۹۱ مشاهده شد، در حالی که عملکرد آن در Stage4 به دلیل پیچیدگی بیشتر داده‌ها و محدودیت نمونه‌ها کاهش یافت. مدل CNN در مراحل ۱، ۲، ۳ و ۵ به Accuracy برابر با ۱۰۰ درصد در مجموعه اعتبارسنجی دست یافت، اما این نتایج با توجه به تعداد محدود نمونه‌های اعتبارسنجی و نبود مجموعه آزمون مستقل، به‌عنوان شواهد قطعی تعمیم‌پذیری مدل به داده‌های جدید تفسیر نمی‌شوند. درStage4 ، که شامل چهار کلاس PLB سالم، PLB ناسالم، برگ سالم و برگ ناسالم بود، عملکرد CNN به‌طور محسوسی کاهش یافت و Accuracy برابر با ۳۳ درصد و Macro F1-score برابر با ۰/۱۲۵ به دست آمد. این نتیجه اهمیت تعداد نمونه‌ها، توازن کلاس‌ها و پیچیدگی مسئله را در عملکرد مدل نشان داد. مدل ViT نیز در مراحل دارای الگوهای تصویری متمایز، عملکرد بسیار بالایی در مجموعه اعتبارسنجی نشان داد، اما در Stage4 کاهش محسوسی در عملکرد آن مشاهده شد. درStage5 ، کاهش هم‌زمان Train Loss و Validation Lossو رسیدن آن‌ها به مقادیر بسیار کوچک، در حدود 10^{-8}، بیانگر همگرایی مناسب مدل با داده‌های موجود بود. با این حال، مشابهCNN ، به دلیل نبود مجموعه آزمون مستقل، نتایج به‌دست‌آمده نمی‌توانند به‌تنهایی تعمیم‌پذیری مدل را در شرایط واقعی اثبات کنند. یکی از مهم‌ترین یافته‌های این پژوهش آن است که انتخاب معماری مدل به‌تنهایی عامل تعیین‌کننده عملکرد سامانه‌های بینایی ماشین نیست و کیفیت، حجم، تنوع و توزیع داده‌ها نقش اساسی در عملکرد مدل دارند. به‌ویژه نتایج Stage4 نشان داد که افزایش پیچیدگی تصویری و وجود کلاس‌های متعدد با نمونه‌های محدود می‌تواند عملکرد مدل‌های طبقه‌بندی را به‌طور محسوسی کاهش دهد. از دیدگاه کاربرد در کشاورزی هوشمند و سامانه‌های خودکار تولید گیاه، یافته‌های این پژوهش نشان می‌دهد که خروجی مدل‌های بینایی ماشین می‌تواند به‌عنوان یک مؤلفه ادراکی و اطلاعات ورودی برای مراحل بعدی تصمیم‌گیری و اتوماسیون مورد استفاده قرار گیرد. در این چارچوب، YOLOv8می‌تواند اطلاعات مکانی مربوط به برگ‌ها و PLB را فراهم کند، در حالی که CNN و ViT می‌توانند اطلاعات مربوط به طبقه‌بندی مرحله رشد و وضعیت نمونه را استخراج کنند. این اطلاعات در آینده می‌تواند به سامانه‌های تصمیم‌یار، مدیریت خودکار شرایط کشت، پایش مداوم گیاه و در مراحل توسعه‌یافته‌تر، سامانه‌های رباتیک کشت‌بافت و جابه‌جایی گیاهچه منتقل شود. بنابراین، نوآوری پژوهش حاضر در معرفی یک معماری جدید یا جایگزینی یک مدل موجود نیست، بلکه در ارزیابی نظام‌مند و مقایسه رفتار سه رویکرد بینایی ماشین با وظایف مکمل در پنج مرحله رشد ارکیده Phalaenopsis و در شرایط حجم محدود داده و تفاوت در پیچیدگی و توزیع کلاس‌ها است .نتایج این ارزیابی می‌تواند مبنایی برای انتخاب و ترکیب مؤلفه‌های مناسب بینایی ماشین در توسعه سامانه‌های آینده پایش خودکار کشت‌بافت و تولید گیاهچه باشد. در نهایت، برای افزایش قابلیت اطمینان و تعمیم‌پذیری این چارچوب، افزایش حجم مجموعه داده، ایجاد توزیع متوازن‌تر میان کلاس‌ها، تهیه

مجموعه آزمون مستقل و ارزیابی مدل‌ها با داده‌های جدید و شرایط تصویربرداری متنوع در مطالعات آینده ضروری است. همچنین، اتصال خروجی سامانه بینایی ماشین به یک سامانه تصمیم‌گیری و سپس تجهیزات اتوماسیون یا رباتیک می‌تواند مسیر توسعه کاربرد عملی این پژوهش در کشاورزی هوشمند و تولید خودکار گیاهچه‌های ارکیده را فراهم کند.

۵- پیشنهادها

با توجه به نتایج به‌دست‌آمده و محدودیت‌های موجود در مجموعه داده و عملکرد مدل‌ها، پیشنهادهای زیر برای ادامه پژوهش و توسعه کاربردهای عملی ارائه می‌شود.

افزایش حجم و تنوع مجموعه داده:
تهیه و ثبت تصاویر بیشتر از ارکیده Phalaenopsis در مراحل مختلف رشد و تحت شرایط متنوع نور، زاویه تصویربرداری و شرایط محیطی، به‌منظور افزایش تنوع داده‌ها و بررسی بهتر قابلیت تعمیم مدل‌های یادگیری عمیق.

بهبود توازن کلاس‌ها:
افزایش تعداد نمونه‌های مربوط به کلاس‌های کم‌نمونه، به‌ویژه کلاس‌های مربوط به وضعیت سلامت درStage4 ، و ایجاد توزیع متوازن‌تر میان کلاس‌ها برای کاهش سوگیری مدل و بهبود معیارهایی مانند Precision، Recall و F1-score .

استفاده هدفمند از روش‌های افزایش داده:
بررسی و مقایسه روش‌های مختلف Data Augmentation متناسب با ویژگی‌های واقعی تصاویر کشت‌بافت، با هدف افزایش تنوع داده‌های آموزشی و کاهش احتمال بیش‌برازش، بدون ایجاد تغییرات غیرواقعی در ویژگی‌های مورفولوژیک گیاه.

ایجاد مجموعه آزمون مستقل:
در مطالعات آینده، بخشی از تصاویر به‌صورت مستقل به‌عنوان مجموعه آزمون (Test Set) در نظر گرفته شود تا عملکرد نهایی مدل‌ها روی تصاویر مشاهده‌نشده و قابلیت تعمیم آن‌ها با اطمینان بیشتری ارزیابی شود.

توسعه چارچوب‌های ترکیبی:
بررسی چارچوب‌های ترکیبی مبتنی بر YOLO، CNN و ViT برای استفاده هم‌زمان از قابلیت مکان‌یابی ساختارهای گیاهی، طبقه‌بندی و استخراج ویژگی‌های تصویری و ارزیابی اینکه آیا ترکیب این قابلیت‌ها می‌تواند عملکرد سامانه پایش را بهبود دهد یا خیر.

اتصال سامانه بینایی ماشین به تصمیم‌گیری خودکار:
در مراحل بعدی، خروجی مدل‌های بینایی ماشین می‌تواند به‌عنوان ورودی یک سامانه تصمیم‌یار مورد استفاده قرار گیرد. برای مثال، اطلاعات مربوط به مرحله رشد، موقعیت برگ و PLB و وضعیت سلامت نمونه می‌تواند برای تعیین اقدامات بعدی در فرآیند پایش و مدیریت گیاه مورد استفاده قرار گیرد.

توسعه سامانه بینایی ماشین برای کشت‌بافت و گلخانه:
طراحی و پیاده‌سازی یک سامانه تصویربرداری و بینایی ماشین که بتواند به‌صورت دوره‌ای تصاویر گیاهان را ثبت و مدل‌های آموزش‌دیده را برای پایش خودکار رشد و سلامت نمونه‌ها به کار گیرد. چنین سامانه‌ای می‌تواند در آینده به سیستم‌های کنترل و مدیریت محیط کشت‌بافت یا گلخانه متصل شود.

توسعه سامانه‌های رباتیک:
در مراحل توسعه‌یافته‌تر، خروجی مدل تشخیص و طبقه‌بندی می‌تواند با یک سامانه رباتیک یکپارچه شود تا اطلاعات تصویری حاصل از مدل به‌عنوان مبنایی برای انجام عملیات خودکار مانند شناسایی، انتخاب و جابه‌جایی گیاهچه‌ها مورد استفاده قرار گیرد.

بررسی انتقال‌پذیری مدل‌ها:
ارزیابی مدل‌ها بر روی ارقام مختلف ارکیده Phalaenopsis و در صورت امکان سایر گیاهان زینتی کشت‌بافتی، به‌منظور بررسی میزان انتقال‌پذیری دانش آموخته‌شده و شناسایی محدودیت‌های مدل در مواجهه با نمونه‌های جدید.

به‌طور کلی، نتایج این پژوهش نشان می‌دهد که توسعه یک سامانه هوشمند پایش گیاه، علاوه بر انتخاب معماری مناسب، به داده‌های کافی و متوازن، مجموعه آزمون مستقل و ارتباط میان بخش بینایی ماشین و فرآیند تصمیم‌گیری نیاز دارد. بنابراین، مسیر پیشنهادی برای ادامه این پژوهش از افزایش و بهبود داده‌ها، ارزیابی مستقل مدل‌ها، ترکیب قابلیت‌های تشخیص و طبقه‌بندی، اتصال خروجی بینایی ماشین به تصمیم‌گیری، و در نهایت یکپارچه‌سازی با سامانه‌های اتوماسیون و رباتیک تشکیل می‌شود. این مسیر می‌تواند زمینه تبدیل نتایج حاصل از تحلیل تصاویر ارکیده Phalaenopsis به یک فناوری کاربردی برای پایش و مدیریت خودکار تولید در کشت‌بافت و گلخانه را فراهم کند.

منابع

Akbar, F. A. and Sari, C. A. (2025). *Orchid species classification using the DenseNet121 deep learning model with a data imbalance handling approach*. Journal of Applied Informatics and Computing, 9(6), 3118–3129. https://doi.org/10.30871/jaic.v9i6.11458.

Apriyanti, D. H., Arymurthy, A. M. and Handoko, L. T. (2013). *Identification of orchid species using content-based flower image retrieval*. Proceedings of the 2013 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), 53–57. https://doi.org/10.1109/IC3INA.2013.6819148.

Apriyanti, D. H., Spreeuwers, L. J. and Lucas, P. J. F. (2023). *Deep neural networks for explainable feature extraction in orchid identification*. Applied Intelligence, 53(21), 26270–26285. https://doi.org/10.1007/s10489-023-04880-2.

Apriyanti, D. H., Spreeuwers, L. J. and Lucas, P. J. F. (2025). *Explainable automated wild-orchid identification combining deep neural networks and Bayesian networks*. Engineering Applications of Artificial Intelligence, 161, 111961. https://doi.org/10.1016/j.engappai.2025.111961.

Arkin, E., Yadikar, N., Xu, X., Aysa, A. and Ubul, K. (2023). *A survey: Object detection methods from CNN to transformer*. Multimedia Tools and Applications, 82(14), 21353–21383. https://doi.org/10.1007/s11042-022-13801-3.

De Silva, M. and Brown, D. (2023). *Multispectral plant disease detection with Vision Transformer–Convolutional Neural Network hybrid approaches*. Sensors, 23(20), 8531. https://doi.org/10.3390/s23208531.

Hemalatha, S. and Jayachandran, J. J. B. (2024). *A multitask learning-based Vision Transformer for plant disease localization and classification*. International Journal of Computational Intelligence Systems, 17, 188. https://doi.org/10.1007/s44196-024-00597-3.

Khanal, B., Poudel, P., Chapagain, A., Regmi, B., Pokhrel, S. and Khanal, S. R. (2024). *Paddy disease detection and classification using computer vision techniques: A mobile application to detect paddy disease*. arXiv, 1–21. https://doi.org/10.48550/arXiv.2412.05996.

Krichen, M. (2023). *Convolutional neural networks: A survey*. Computers, 12(8), 151. https://doi.org/10.3390/computers12080151.

Kuznetsova, A., Maleva, T. and Soloviev, V. (2020). *Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot*. Agronomy, 10(7), 1016. https://doi.org/10.3390/agronomy10071016.

Lee, S.-H., Ku, H.-C. and Zhang, Y.-C. (2025). *Few-shot learning based on dual-attention mechanism for orchid species recognition*. International Journal of Data Science and Analytics, 20, 3439–3452. https://doi.org/10.1007/s41060-024-00671-1.

Naranjo-Torres, J., Mora, M., Hernández-García, R., Barrientos, R. J., Fredes, C. and Valenzuela, A. (2020). *A review of convolutional neural network applied to fruit image processing*. Applied Sciences, 10(10), 3443. https://doi.org/10.3390/app10103443.

Özden, C. (2026). *Vision Transformer and CNN-based models for image analysis of plant diseases: An approach for agricultural decision support systems*. ISPEC Journal of Agricultural Sciences, 10(1), 108–121. https://doi.org/10.5281/zenodo.18228928.

Sangiovanni, M., Starace, L. L. L., Di Pierro, D., Cozzolino, S. and Addonizio, E. (2025). *OrchID: Explainable deep learning for Ophrys orchid biodiversity monitoring*. Array, 27, 100480. https://doi.org/10.1016/j.array.2025.100480.

Sungheetha, A. and Sharma, R. R. (2021). *3D image processing using machine learning based input processing for man-machine interaction*. Journal of Innovative Image Processing, 3(1), 1–6. https://doi.org/10.36548/jiip.2021.1.001.

Thakur, P. S., Khanna, P., Sheorey, T. and Ojha, A. (2022). *Explainable Vision Transformer enabled Convolutional Neural Network for plant disease identification: PlantXViT*. arXiv, 1–21. https://doi.org/10.48550/arXiv.2207.07919.

Tripathi, A., Gohokar, V. and Kute, R. (2024). *Comparative analysis of YOLOv8 and YOLOv9 models for real-time plant disease detection in hydroponics*. Engineering, Technology & Applied Science Research, 14(5), 17269–17275. https://doi.org/10.48084/etasr.8301.

Zizka, A., Silvestro, D., Vitt, P. and Knight, T. M. (2021). *Automated conservation assessment of the orchid family with deep learning*. Conservation Biology, 35(3), 897–908. https://doi.org/10.1111/cobi.13616.