DLSU RESEARCH CONGRESS 2023
MANILA, PHILIPPINES
JULY 5-7, 2023
Fostering a Humane and Green Future:
Pathways to Inclusive Societies and Sustainable Development

A Computer Vision Approach to Identify Cecid Fly Defects on Mango Fruits Using Vision Transformers

Maria Jeseca C. Baculo[1], Conrado Ruiz Jr. [2] and Aran Oya [3]
[1] *Don Mariano Marcos Memorial State University*
[1,2] *De La Salle University – Manila*
[3] TETAM, Bogazici University, Istanbul, Turkey
**maria_jeseca_baculo@dlsu.edu.ph*

Abstract: Mango producers consider the cecid fly a serious pest because it significantly reduces yields and affects the export market. Mangoes can develop surface defects because of fungal, insect, and cecid fly infestations, among others, which reduce the market value of the fruit. Subjective visual inspection, used in conventional methods to find these defects, can be tedious and unreliable. One possible remedy is computer vision technology, which allows accurate and efficient categorization of defects. In this study, we utilized Vision Transformer to train a binary classifier to identify cecid fly surface defects. The results showed that our method outperformed other CNN architectures, achieving 91% accuracy on the held-out test set. This method could be of great importance for early detection and efficient insect management in mango farms.

Key Words: cecid fly; defect detection; ViT,; transformers

1. INTRODUCTION

The cecid fly, scientifically known as *Dacus ferrugineus*, is a major pest of mangoes in many tropical and subtropical countries, including the Philippines. The female fly lays her eggs on the mango fruit, and after hatching, the larvae bore into the fruit, causing it to rot and drop prematurely. The infested fruit becomes inedible and results in significant yield losses for mango farmers. In addition, the presence of the cecid fly in mango plantations can also have a negative impact on the export market for mangoes, as importing countries often have strict regulations regarding pest infestation.

Cecid fly defects can be observed in the surface of mango fruits, leaves, and flowers. Surface defects of mangoes may include discoloration, scars, and bruises, as well as fungal and insect infestations. These defects can make the fruit unsightly and inedible, reducing its market value. Mango defects are often detected by subjective, time-consuming visual assessment by human specialists, a traditional procedure. In addition, certain defects may be overlooked during inspection, which would give a misleading impression of quality assurance.

Early detection of cecid fly infestation is essential for effective pest control. Timely detection can prevent further spread of the pest, reduce the use of pesticides, and minimize economic damage from yield loss. Therefore, efficient, and reliable methods for detecting cecid fly infestations in mango

plantations need to be developed.

One possible approach for automatically identifying surface defects in mangoes is computer vision technology. Using digital cameras and image processing techniques, computer vision systems capture and analyze images of the fruit surface. The technology can identify defects and categorize them by color, texture and shape, providing accurate and unbiased results.

Previous studies were able to propose autonomous diagnostic systems for rating mangos and identifying their pathologies in recent years thanks to developments in the field of computer vision (Faye et al., 2022; Ansah et al., 2023; Hassoon, 2023; Veling et al,2019). These solutions are based on artificial intelligence and computer vision.

In the study by Nithya et al. (2022), deep convolutional neural networks (CNNs) were used to develop a computer vision system for detecting defects in mango fruit. To train a deep CNN model, the authors collected a dataset of mango photos with various defects, including black spot, powdery mildew, and anthracnose. According to the results of the study, the developed system had an overall accuracy of 98.6% in identifying mango defects. According to the study, the proposed method can be used as a useful tool for identifying mango defects to improve quality assurance and reduce financial losses in the mango business.

The Faster Region-based Convolutional Neural Network (Faster R-CNN) is a popular object detection framework that has been applied in various studies. In the study by Baculo et al. (2021), the Faster R-CNN algorithm was used to detect cecid fly-induced galls in mangoes. The Faster R-CNN architecture used ResNet as its base learner. The authors reported an average precision of 0.92 and an average recall of 0.88, indicating the effectiveness of the framework in detecting cecid fly defects in mangoes. In comparison to this previous work, this study conducted experiments to compare the performance of different base learners in classifying cecid flies.

Vision Transformers (ViT) have recently attracted attention as a powerful tool for image classification (Khan et al., 2022; Zhang et al., 2023). ViT models have shown considerable progress in classification accuracy and effectiveness compared to traditional Convolutional Neural Networks (CNNs) (Tuli et al., 2021). The use of ViT models can improve the classification of various defects, including puncture wounds, galls, and scars, related to the detection of defects by the Cecid fly in mango fruit.

The self-attention mechanism of ViT models allows them to capture spatial correlations more accurately between features, which increases their ability to detect minute differences in cecid fly injuries. ViT models are also able to handle images of different sizes, which increases their adaptability in real-world scenarios. The use of ViT models in identifying mango defects has the potential to increase the accuracy and efficiency of the classification process, which would ultimately lead to better quality control and financial gains for the fruit trade.

2. METHODOLOGY

2.1 Dataset

The dataset utilized for both training and evaluating the model was sourced from the authors' previous study (Baculo et al., 2021). The dataset consisted of 1200 cecid fly surface defect images, while 1200 individual non-cecid defects were manually annotated to represent the non-cecid images. To fit the input layer of the training architecture, each RGB image was rescaled to 224 x 224. Fig 1 and 2 shows the sample images in the two classes.

Fig. 1. Sample Cecid Fly Defect Images

Fig. 2. Sample Non-Cecid Fly Defect Images.

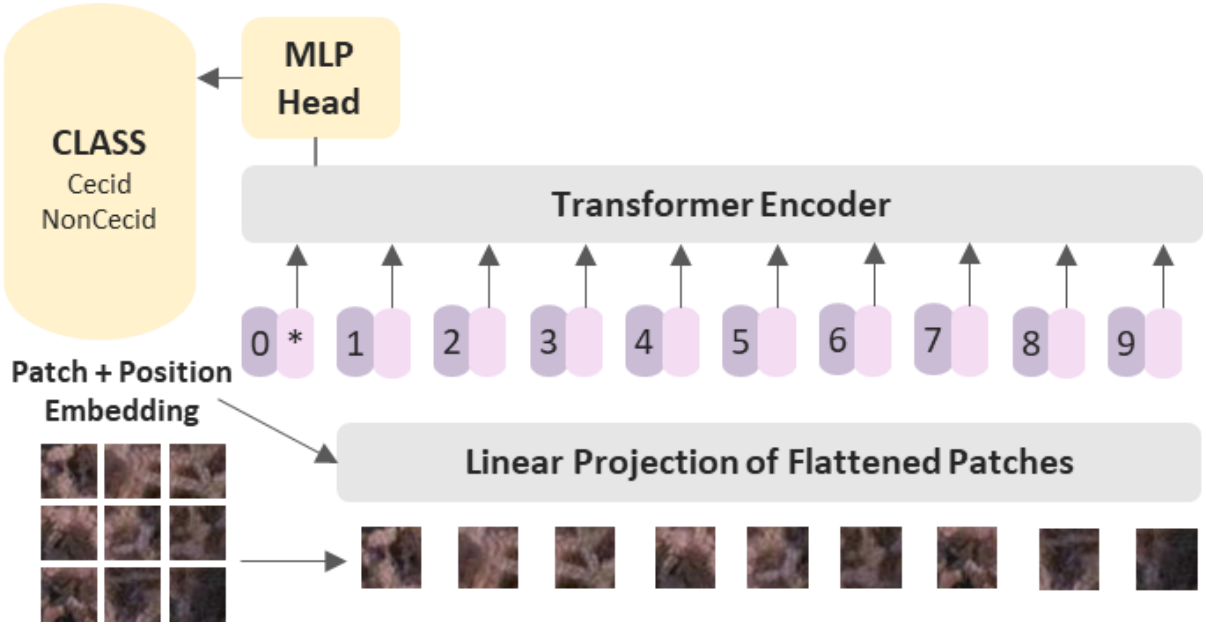


Fig 3. Vision Transformer Architecture

The dataset was divided into train and test sets prior to training. In this study, we allocated 240 images to the test set, which accounted for 10% of the total dataset.

2.2 Model Training

This study used Vision Transformer to try to improve the performance of previous models in classifying mango surface defects, specifically, those infested by cecid flies

The experiments were conducted using pytorch's implementation of ViT. The original paper's design concepts (Dosovitskiy, 2020) guide the construction of ViT in PyTorch. The classification head, the transformer encoder, and the patch embedding layer make up its three primary components.

Vision Transformer is a deep neural network used for computer vision tasks such as object identification and image classification. The Transformer architecture, originally developed for problems related to natural language processing, serves as the foundation for ViT. Unlike conventional convolutional neural networks (CNNs), which use convolutional layers to extract features from images,

ViT models compute the relationships between different elements of an image using a self-attention process. ViT models can now learn more complicated and precise patterns in images, which has brought their performance on many computer vision benchmarks up to date. In large-scale image categorization, ViT models have demonstrated significant potential, outperforming conventional CNNs with much smaller models and less CPU performance.

In PyTorch, ViT implementation involves breaking down the input image into smaller patches, which are then transformed into vectors. These patch vectors are projected to a lower-dimensional space and combined with positional information. The transformed patches are fed into a Transformer Encoder, which learns contextual relationships between different patches using self-attention and feed-forward networks. Finally, a classification head maps the learned features to the desired output

classes. During training, the model is optimized using backpropagation and standard optimization algorithms. PyTorch provides convenient modules and tools for efficient ViT implementation, making it easier to train and evaluate the model.

In the previous experiments, the dataset was trained using CNN architectures such as GoogleNet and ResNet. In this study, we investigate whether the performance can be further improved by using Vision Transformers.

3. RESULTS AND DISCUSSION

To assess the performance of the models, accuracy, precision, and recall were computed as the performance metrics. The ratio of correct predictions to all predictions is called accuracy. It serves as a measure of how accurately the model predicts the corresponding class. Precision can be defined as the proportion of correct positive predictions to the total number of correct positive and false positive predictions. It is a measure of how accurately the model predicts the positive cases. Recall is the ratio of true positives to the total true positives plus true negatives. It is a measure of how well the model can detect positive occurrences.

Two ViT configurations were used in the experiments: ViT_50, trained with 50 epochs, and ViT_100, trained with 100 epochs. The hyperparameters, including the initial learning rate of 0.0002 and batch size of 8, were tuned to optimize the classification performance. Additionally, the patch size was fine-tuned to 16.

Table 1 shows the summary of the performance of the trained classifiers.

Table 1. Performance of the Binary Classifiers

Model	Training	Test		
	Accuracy	Accuracy	Precision	Recall
GoogleNet	.90	.75	.75	.75
ResNet	.96	.81	.82	.82
ViT_50	.92	.83	.84	.83
ViT_100	.97	.89	.89	.89

The experimental results demonstrate that the ViT models achieved the highest classification performance compared to the other classifiers. Specifically, the ViT model trained for 100 epochs exhibited the best performance with higher accuracy, precision, and recall on the held-out test set than the other models. The table indicates that while the CNN models achieved at least 90% training accuracy, their performance significantly declined when tested on the held-out dataset. This decline in performance can be attributed to the fine-grained characteristics of the input images, which lack definitive color, shape, and size, making it challenging for the CNN models to accurately classify them.

The confusion matrices shown in Fig 4 and 5 compare the expected labels with the actual labels and provide valuable insights into the model's prediction accuracy. It is evident from the matrices that the ViT architecture trained for a longer duration has significantly improved the classification of the cecid fly defect.

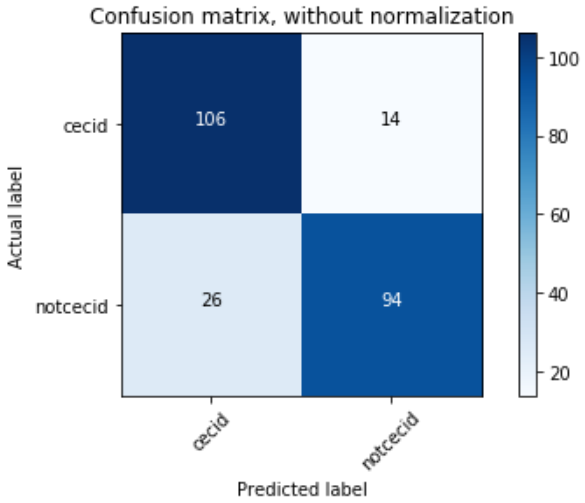


Fig. 4. Confusion Matrix for ViT_50

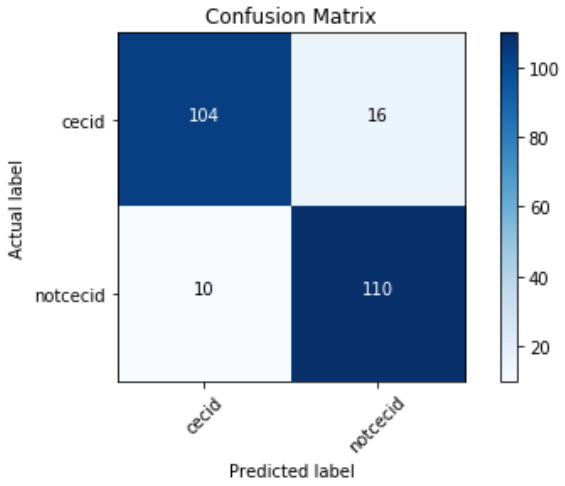


Fig. 5. Confusion Matrix for ViT_100

Fig. 6. Correctly Classified Instances

Fig 6 displays a set of sample test images with their corresponding correct predictions made by the model. These images demonstrate the model's ability to generalize its predictions accurately across a wide range of image shapes and colors, indicating its robustness in recognizing different objects in the image.

On the other hand, Fig 7 presents a few instances of images that were misclassified by the model. These misclassifications could be attributed to poor lighting conditions of the test images. In particular, the images share similarities in their darkened appearance and reduced image details, which may have made it difficult for the model to accurately distinguish the objects in the image.

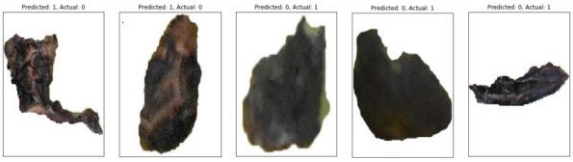


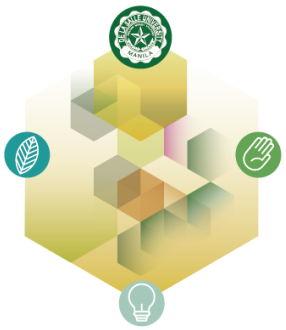
Fig. 7. Incorrect Classifications

4. CONCLUSIONS

Overall, the results suggest that the model performs well in recognizing a diverse set of objects, but its performance may be affected by variations in lighting conditions and image quality. Further improvements could be made to enhance the model's ability to handle such challenges and improve its accuracy under different real-world scenarios.

5. ACKNOWLEDGMENTS

The authors would like to express gratitude to the Commission on Higher Education (CHED) for providing the scholarship that enabled the research to be conducted. The author would also like to thank the National Mango Research and Development Center, and Oro Verde farm, Guimaras that generously allowed the gathering of the dataset, without which



the research would not have been possible. Their contributions are deeply appreciated and have been crucial in advancing the understanding of the subject matter.

6. REFERENCES (use APA style for citations)

Ansah, F. A., Boateng, M. A., Siabi, E. K., & Bordoh, P. K. (2023). Location of Seed Spoilage in Mango Fruit using X-ray Imaging and Convolutional Neural Networks. Scientific African, e01649.

Baculo, M. J. C., Ruiz, C., & Aran, O. (2021). Cecid Fly Defect Detection in Mangoes Using Object Detection Frameworks. In Advances in Computer Graphics: 38th Computer Graphics International Conference, CGI 2021, Virtual Event, September 6–10, 2021, Proceedings 38 (pp. 205-216). Springer International Publishing.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Faye, D., Diop, I., & Dione, D. (2022). Mango Diseases Classification Solutions Using Machine Learning or Deep Learning: A Review. Journal of Computer and Communications, 10(12), 16-28.

Hassoon, I. M. (2022). Classification and Diseases Identification of Mango Based on Artificial Intelligence: A Review. Journal of Al-Qadisiyah for computer science and mathematics, 14(4), Page-39.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s), 1-41.

Nithya, R., Santhi, B., Manikandan, R., Rahimi, M., & Gandomi, A. H. (2022). Computer Vision System for Mango Fruit Defect Detection Using Deep Convolutional Neural Network. Foods, 11(21), 3483.

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision?. arXiv preprint arXiv:2105.07197.

Veling, P. S., Kalelkar, R. S., Ajgaonkar, L. V., Mestry, N. V., & Gawade, N. N. (2019). Mango disease detection by using image processing. International Journal for Research in Applied Science and Engineering Technology, 7(4), 3717-3726.

Zhang, Q., Xu, Y., Zhang, J., & Tao, D. (2023). Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. International Journal of Computer Vision, 1-22.