



Leveraging Computer vision and NLP for Comparative Analysis of Papaya Crop Disease Classification Using Multimodal Fusion

Eyobed Birhanu Paulos¹  and Mesay Gameda Yigezu² 

¹ Wolaita Sodo University, Wolaita Sodo, Ethiopia
eyobed.birhanu@wsu.edu.et

² University of Colorado, Colorado Springs, USA
myigezu@uccs.edu

Abstract. This study explores the effectiveness of multimodal fusion techniques for papaya disease classification by integrating computer vision and natural language processing (NLP) models. We compare three distinct methods such as single-modal classification using ResNet50, multimodal analysis combining ResNet50 with generative AI-based NLP that is GPT, and multimodal analysis combining VGG19 with GPT. Our findings indicate that the multimodal fusion approach using ResNet-50 and GPT achieves superior accuracy and performance, underscoring the potential of leveraging computer vision and NLP for agricultural disease detection. The study utilized a dataset of 8,000 images across four classes: Black Spot, Healthy, Powdery Mildew, and Ring Spot. Results indicate that the multimodal approach combining ResNet50 with GPT significantly outperformed the other models, achieving 99.7% accuracy in just seven epochs. The study underscores the potential of multimodal learning in enhancing agricultural disease diagnostics.

Keywords: Generative AI, Multimodal learning, deep learning, ResNet-50, GPT, VGG19, papaya disease classification, computer vision, natural language processing

1 Introduction

Agriculture is essential to economic growth; it accounted for 4% of the world GDP and could potentially account for more than 25% in developing countries. Agriculture remains the main engine of Ethiopia's population economy. Crop diseases are a major concern in agriculture, affecting crop yield and quality, and leading to substantial economic losses [8]. Papaya, a widely cultivated fruit in the global economy, particularly for Ethiopia, is susceptible to several diseases such as Black Spot, Powdery Mildew, and Ring Spot, which can devastate crops if not managed properly [8,3]. Traditional disease identification methods rely on manual inspection, which is often subjective and prone to errors. Hence, papaya disease detection is crucial for maintaining crop health and optimizing yield [9].

Accurate and timely identification of diseases can help in taking immediate corrective actions, thus preventing the spread of infections and ensuring better crop management. Traditional image-based methods have proven effective, but integrating textual descriptions of symptoms can enhance classification accuracy by providing additional context. Convolutional Neural Networks (CNNs), such as ResNet50 and VGG19, have shown great potential in image classification tasks, including disease detection [5,11]. However, these models primarily focus on visual features, which may not capture the complete context of disease symptoms.

To address aforementioned limitations, multimodal learning approaches that integrate image data with contextual information, such as textual descriptions, have gained attention [7]. In this study, we explore the efficacy of combining CNN-based image classifiers (ResNet50 and VGG19) with NLP, GPT. From different generative AI approaches, the researchers were chosen GPT-2 for building text model which is open source, because it produces human-like text and also it's state-of-the-art natural language processing model, that can enhance papaya disease classification. This paper provides a detailed comparative analysis of three methods. These methods are a single-modal approach using ResNet50, a multimodal approach (multimodal fusion) combining ResNet50 with GPT-2, and a multimodal approach using VGG19 with GPT-2.

Finally, the performance of these models is evaluated based on classification accuracy, confusion matrices, precision, recall, and F1-scores.

2 Related Work

Previous studies in agricultural disease classification have primarily focused on image-based models. [7] demonstrated the use of deep convolutional neural networks (CNNs) for plant disease identification with significant success. [4] further explored the potential of using deep learning models like ResNet and VGG for detecting and diagnosing plant diseases in various crops. While these studies have laid the foundation for automated plant disease detection, they rely solely on visual data.

Integrating NLP models with image classifiers is a relatively new area. Recent advancements by [1] have shown that using BERT for text classification can enhance understanding in contexts where textual descriptions accompany visual data. However, the potential of generative models like GPT, which can create detailed text descriptions from image features, remains largely under-explored in the context of agricultural disease detection and classification. This study aims to bridge this gap by examining how generative AI can enhance multimodal learning for better disease classification accuracy.

3 Methodology

3.1 Dataset Description

For this study, the researcher gathered relevant image data from a well-known Ethiopian Agricultural Research Institute called Melkasa Agricultural Research

Institute. The dataset consists of 8,000 images of papaya leaves, divided into four classes: Black Spot, Healthy, Powdery Mildew, and Ring Spot. Each class contains 2,000 images with high-resolution visual data that capture various symptoms. The images were gathered from multiple papaya plantations to ensure diversity in environmental conditions, camera angles, and disease manifestations of Ethiopian agricultural research centers. So the class label with their description are as follows. Black Spot Papaya: Leaves with black spots indicating black spot infection. Powdery Mildew: Papaya leaves with powdery mildew, typical of powdery disease. Ring Spot Papaya: Leaves with visible ring spot disease symptoms. And finally, Healthy Papaya: Healthy leaves with no signs of disease. The following figure Fig. 1 depicts the sample diseased leaves of papaya.

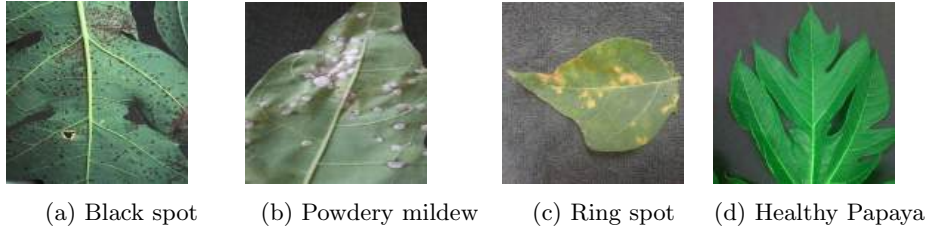


Fig. 1: Sample image dataset collected for experiment

3.2 Preprocessing Techniques

- **Image Preprocessing:** Images were resized to 224x224 pixels to standardize input size, normalized to have zero mean and unit variance to facilitate better convergence, and augmented with random rotations (up to 30 degrees), flips (horizontal and vertical), and brightness adjustments ($\pm 20\%$) to mimic real-world scenarios and enhance model generalization.
- **Text Preprocessing:** Text descriptions were tokenized using the GPT tokenizer, with padding added to ensure uniform input length. Special tokens like '[CLS]' and '[SEP]' were used to denote the start and end of sequences. Additionally, we applied stop-word removal and lowercasing to reduce noise in the data.

3.3 Dataset Description

ResNet-50 for Image Classification: ResNet-50, a deep CNN with 50 layers, is known for its ability to mitigate the vanishing gradient problem through residual connections. This model employs shortcut connections that bypass one or more layers, allowing gradients to flow more smoothly during backpropagation. This architectural innovation enables ResNet-50 to train deeper networks

effectively, capturing intricate patterns in image data. It is particularly suited for image classification tasks like disease detection in plants due to its robustness and relatively low computational cost [6].

In this study, the ResNet50 was fed 224x224x3 photos of leaves together with their corresponding class labels. This pre-trained model utilizes a softmax classifier and contains 1000 neurons with the output of the class probability and 50 layers for feature extraction. This state-of-art algorithm algorithm's default settings were left alone. The classification problem domain of this study is modified into the last layer, which is fully linked and then two dense layers were added and fine tuned.

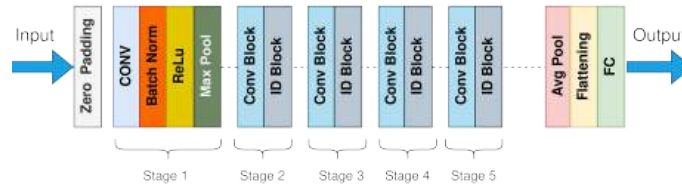


Fig. 2: Resnet 50 architecture

VGG19 for Image Classification: VGG19 is another popular deep learning model consisting of 19 layers, mainly using sequential blocks of convolutional and pooling layers without residual connections [2]. Despite being simpler and more straightforward than ResNet50, VGG19 is known for its consistent performance in image classification due to its deep architecture and large receptive fields. However, it often suffers from slower convergence and higher computational costs compared to ResNet50 due to its lack of residual connections and deeper stack of sequential layers. In this model, the default settings were left alone. Then this model is modified into the last layer, which is fully linked layer and then two dense layers were added and fine tuned. In this study, VGG19 was used as a comparative baseline to assess the effectiveness of ResNet50 in the multimodal approach.

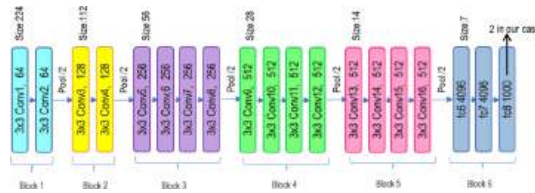


Fig. 3: VGG19 architecture

GPT for Text Generation: GPT (Generative Pretrained Transformer) is a generative model that predicts the next word in a sequence, making it effective for generating descriptive text based on image input [10]. GPT focuses on generating coherent and contextually relevant text [10]. In this study, GPT was fine-tuned to generate disease descriptions from image features, leveraging its ability to produce human-like text. The text descriptions were fed into GPT-2 to generate contextual embeddings, capturing the semantic information related to disease symptoms. This allowed the model to leverage textual information as an additional modality, enhancing its overall predictive power that could be useful for agricultural experts in identifying diseases.

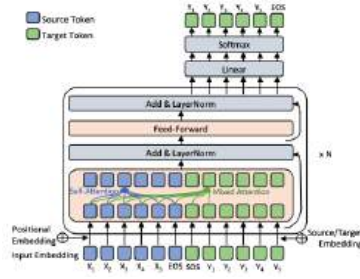


Fig. 4: GPT-2 architecture

Combined Approach: ResNet-50 with GPT and VGG19 with GPT

The combined approach leverages the strengths of ResNet-50 and VGG19 for feature extraction from images and GPT for generating descriptive text based on these features. The generated text was then used alongside image features for final disease classification, creating a multimodal fusion model. This integration allows the model to learn complementary features from both modalities, leading to better performance and robustness in disease detection tasks.

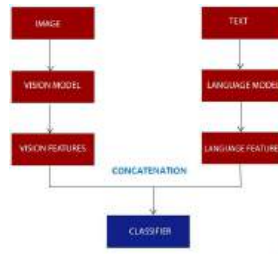


Fig. 5: Combined Multimodal architecture of computer vision & NLP

4 Experiments and Results

In this section, the researcher tries to explain the experimental results generated from comparative study during training and testing the model.

4.1 Experimental Settings

The models were trained using train and test datasets. The experiments were conducted by using percentage split of total image datasets associating with their respective image description with 80% train split, 20% test split and then the model is validated using 10% of training data. In addition, the training was conducted using the Adam optimizer with a learning rate of 0.001, chosen for its efficiency in handling sparse gradients commonly found in NLP tasks. The batch size was set to 32, balancing memory constraints and convergence speed. Early stopping was employed to prevent overfitting, with a patience parameter set to 5 epochs. The primary evaluation metric was accuracy, with additional metrics including precision, recall, and F1-score to assess the models' performance. The training process was conducted on python 3.10.4 on Google colab L4 GPU to accelerate computation and AMD Ryzen 3 5300U with Radeon Graphics 2.60 GHz HP laptop to run on local machine for comparative analysis.

4.2 Results of Different Methods

Single Modal approach

- **Single-Modal Classification Using ResNet50:** ResNet50, a widely used CNN model, was employed as a standalone classifier in the first approach. The architecture of ResNet50 is characterized by its use of residual blocks, which allow the network to learn deep representations without suffering from vanishing gradients. The model was initialized with weights pre-trained on ImageNet and fine-tuned on the papaya disease dataset. During training, the input images were resized to 224x224 pixels, and data augmentation techniques such as rotation, flipping, and scaling were applied to improve the model's generalization. The model was trained for 30 epochs using the Adam optimizer and categorical cross-entropy loss, achieving an accuracy of 92.5%. Thus, the model struggled with differentiating between visually similar diseases.

Multimodal approach

- **ResNet-50 with GPT (Multimodal Fusion):** In the second approach, a multimodal framework was implemented by combining ResNet50 with GPT-2, a generative AI language model. This architecture leverages ResNet50 for extracting image features and GPT-2 for encoding textual descriptions of each disease class. The ResNet50 outputs a feature vector of size 2048,

which is concatenated with a 768-dimensional embedding generated by GPT-2. The combined feature vectors are then fed into a dense neural network for classification. This model setup allowed the integration of visual and contextual information, greatly enhancing the model’s ability to distinguish between disease types. This multimodal approach achieved a remarkable accuracy of 99.7% within just seven epochs, demonstrating the power of combining visual and textual data for more nuanced disease classification.

- **Multimodal Classification Using VGG19 and GPT-2:** The third approach also employed a multimodal strategy but replaced ResNet50 with VGG19, another popular CNN model. VGG19 is known for its simplicity and uniform architecture, consisting of 19 layers with small 3x3 convolutional filters. Similar to the ResNet50-GPT combination, VGG19 was used for image feature extraction, while GPT-2 provided textual embeddings. The feature vectors from VGG19 (4096 dimensions) were concatenated with GPT-2 embeddings and passed through a fully connected layer for final classification. While this model achieved 94.5% accuracy in 10 epochs, it required more training time and was less efficient at capturing complex disease features compared to the ResNet50-GPT model, largely due to VGG19’s deeper architecture without residual connections, which slowed convergence.

5 Discussion

The comparative analysis of the three approaches revealed significant variations in performance, underscoring the advantages of multimodal learning over single-modal methods. The single-modal approach using ResNet50 demonstrated solid performance with 92.5% accuracy over 30 epochs. Although effective, the model faced difficulties in distinguishing diseases with subtle visual differences, such as Black Spot and Ring Spot, due to its reliance solely on image features. This limitation highlights the challenge of single-modal classifiers in complex disease identification tasks.

The integration of generative AI in the multimodal approaches led to marked improvements in classification accuracy. The ResNet50-GPT multimodal approach achieved an impressive 99.7% accuracy within just seven epochs, outperforming the single-modal ResNet50 model by a significant margin. The fusion of image and text data allowed the model to utilize contextual disease information, enhancing its ability to differentiate between visually similar classes. The rapid convergence of this model suggests that the addition of textual descriptions provides crucial information that aids the learning process, enabling the model to generalize better to unseen data.

In contrast, the VGG19-GPT multimodal approach, while still superior to the single-modal ResNet50, lagged behind the ResNet50-GPT model, achieving 94.5% accuracy in 10 epochs. The VGG19 model’s deeper but less efficient architecture required more training iterations to reach optimal performance. The absence of residual connections, which are a hallmark of ResNet50, likely contributed to slower training and reduced classification performance. Despite its

lower accuracy, the VGG19-GPT combination still demonstrated the value of multimodal learning, particularly when compared to the single-modal approach.

Performance metrics such as confusion matrices, precision, recall, and F1-score further illustrated the strengths of the ResNet50-GPT approach. The confusion matrix revealed near-perfect classification with minimal misclassifications, especially for disease classes that were challenging for other models. Precision, recall, and F1-scores exceeded 99%, indicating that the ResNet50-GPT model not only classified diseases accurately but also maintained consistency across all classes.

Overall, the results prove the superiority of multimodal learning in agricultural disease classification, particularly when combining powerful CNN architectures like ResNet50 with advanced generative AI models. The enhanced performance of the ResNet50-GPT model underscores the potential of integrating visual and contextual data, paving the way for more accurate and reliable disease diagnostics in agriculture sectors.

6 Conclusion

This study presents a novel approach to papaya disease classification by integrating computer vision with NLP. It presents a comprehensive comparative analysis of three methods for papaya disease classification, demonstrating the effectiveness of multimodal learning (multimodal fusion) approaches. The single-modal ResNet50 model, while competent, struggled with differentiating visually similar diseases, highlighting the limitations of purely visual classifiers. The integration of NLP in multimodal approaches, particularly the ResNet50-GPT model, significantly enhanced classification accuracy, achieving 99.7% accuracy in just seven epochs. The results affirm the potential of combining visual and contextual data for more accurate and reliable disease diagnostics in agriculture. Future work could explore the scalability of this approach to other crops and the development of real-time disease detection systems for field applications.

References

1. Dai, G., Fan, J., Dewi, C.: Itf-wpi: Image and text based cross-modal feature fusion model for wolfberry pest recognition. *Computers and Electronics in Agriculture* **212**, 108129 (2023)
2. Dey, N., Zhang, Y.D., Rajinikanth, V., Pugalenth, R., Raja, N.S.M.: Customized vgg19 architecture for pneumonia detection in chest x-rays. *Pattern Recognition Letters* **143**, 67–74 (2021)
3. Gabrekiristos, E., Dagne, A.: A newly emerging disease of papaya in ethiopia: Black spot (*asperisporium caricae*). *Disease and Management Options. J Plant Pathol Microbiol* **11**, 488 (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)

5. Hossen, M.S., Haque, I., Islam, M.S., Ahmed, M.T., Nime, M.J., Islam, M.A.: Deep learning based classification of papaya disease recognition. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). pp. 945–951. IEEE (2020)
6. Koonce, B., Koonce, B.E.: Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization. Springer (2021)
7. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. *Frontiers in plant science* **7**, 1419 (2016)
8. Morris, D.: Tropical fruits. *American Journal of Pharmacy (1835-1907)* **58**(9), 444 (1886)
9. Villegas, V.: Edible fruits and nuts-carica papaya l. EWM Verheij, RE Coronel, Eds **2** (1997)
10. de Vries, W., Nissim, M.: As good as new. how to successfully recycle english gpt-2 to make models for other languages. arXiv preprint arXiv:2012.05628 (2020)
11. Zhang, Z., He, X., Sun, X., Guo, L., Wang, J., Wang, F.: Image recognition of maize leaf disease based on ga-svm. *Chemical Engineering Transactions* **46**, 199–204 (2015)