Integrating Acoustic and Image Data Features for Melon Ripeness Classification Using Convolutional Neural Network

Endang Purnama Giri[1], Agus Buono[2], Karlisa Priandana[3], Dwi Guntoro[4]
Computer Science Study Program-School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia[1, 2, 3]
Department of Agronomy-Faculty of Agriculture, IPB University, Bogor, Indonesia[4]

***Abstract*—This study evaluates three classification scenarios: image-based only, acoustic-based using Mel Frequency Cepstral Coefficients (MFCC), and a combined multimodal CNN architecture integrating both modalities. The experiments are conducted on a relatively small dataset comprising only 230 samples. To mitigate the risk of overfitting arising from the limited dataset size, data augmentation is applied to both image and audio data, with audio augmentation performed before the construction of the MFCC spectrogram. Experimental results demonstrate that the multimodal CNN with data augmentation achieves the best performance, with precision, recall, and F1-score, respectively, 0.95, 0.94, and 0.94. These results indicate that augmenting both image and audio data effectively enhances data diversity and model robustness, significantly improving classification performance. The findings confirm that combining complementary feature representations from multiple modalities with proper augmentation strategies substantially improves audio-visual classification tasks.**

***Keywords*—CNN; multimodal learning; melon ripeness classification; image classification; acoustic classification; data augmentation**

I. Introduction

Accurate melon ripeness classification is crucial for melon cultivation in Indonesia, as timely harvesting directly affects fruit quality. Being a climacteric fruit, melons continue to ripen after harvest, and their distribution is highly diverse, spanning both domestic and international markets. Therefore, precise classification is essential to ensure optimal harvest timing and maintain quality throughout the supply chain. Accordingly, this study aims to develop a nondestructive melon ripeness classification technique based on CNN.

In recent years, advancements in deep learning techniques have paved the way for innovative applications in agriculture, including the classification of fruit ripeness and sweetness. Melons, a widely cultivated and commercially valuable fruit, come in various varieties with distinct characteristics. However, due to differences in ripeness levels and sweetness, determining the optimal harvest time can be challenging. This challenge has led to the growing interest in nondestructive techniques to assess these qualities. To ensure the study remains focused, this research specifically examines two popular melon varieties: Golden Apollo and Golden Alisha. These varieties are widely cultivated and possess characteristics that make them suitable for classification studies. Additionally, the melons used in this research are assumed to be under controlled environmental conditions, such as consistent greenhouse settings (temperature, humidity, and light exposure) and uniform nutritional treatments, ensuring minimal variation caused by external factors.

The features explored in this study are limited to acoustic-based and image-based features, as they have shown significant potential in previous research for nondestructive classification methods. Furthermore, the classification models are developed using the Convolutional Neural Network (CNN) architecture, which has demonstrated superior performance in image and feature extraction tasks. This study prioritizes improving the accuracy of the classification model, while considerations of time and space complexity are beyond the primary scope. By addressing these aspects, the study aims to push the boundaries of existing research by investigating feature relevance, visualizing feature extraction, and modifying CNN architectures to combine acoustic and visual features. The outcomes of this research are expected to contribute to more accurate and efficient nondestructive methods for classifying melon ripeness and sweetness, providing practical benefits to the agricultural industry.

II. Previous Research

Previous research has consistently demonstrated the effectiveness of Convolutional Neural Networks (CNNs) in addressing classification tasks involving both acoustic and visual features. For speech recognition, CNNs integrated into a hybrid neural network-Hidden Markov Model (NN-HMM) framework achieved an accuracy of 79.93%, surpassing standard neural networks (77.05%) and Deep Belief Network-pretrained models (79.5%) using the TIMIT dataset [1]. Further studies confirmed CNN's superiority in reducing error rates by 6–10% compared to Deep Neural Networks (DNN)-HMM models [2]. Similarly, in phoneme classification using raw audio signals, CNN achieved a significant improvement in accuracy, reaching 67.88% compared to 38.91% with Multi-Layer Perceptron (MLP) models [3]. In music onset detection, CNN also outperformed Recurrent Neural Networks (RNNs), achieving 91.7% accuracy compared to RNN's 89.2% [4]. In environmental sound classification, CNNs showed remarkable performance, achieving 77.85% accuracy on the ESC-10 dataset and 49% on ESC-50, compared to 56% by Tensor Deep Stacking Networks (TDSN) [5]. Another study on forest sound classification, which included bird chirps and chainsaw noises,

reported CNN achieving 85% accuracy using spectrogram features [6]. CNN was further validated as the top-performing model for construction noise classification, achieving 97% accuracy with Mel-spectrograms, outperforming other models such as Random Forest (93%), MLP (92%), k-NN (85%), and SVM (84%) [7].

In visual-based tasks, CNN also excelled. For traffic sign and light classification using RGB-D image features, CNNs achieved 93.33% accuracy [8]. In fruit image classification, CNN achieved an impressive 98.88% accuracy on a dataset comprising 55,244 images of 81 fruit categories using a pure CNN architecture with seven layers and global average pooling [9]. For large-scale image classification, CNN architectures such as GoogLeNet and ResNet achieved an accuracy of 83.12% on the ImageNet dataset, consisting of 1.2 million high-resolution images across 1,000 categories [10]. These studies consistently highlight the capability of CNNs in extracting relevant features, improving classification accuracy, and outperforming traditional machine learning approaches. These findings suggest that CNNs hold great potential for application in challenging tasks like the classification of melon ripeness and sweetness using acoustic and visual features.

Previous studies have demonstrated that multimodal fusion, particularly the integration of acoustic and visual features within CNN-based frameworks, can outperform unimodal approaches by leveraging complementary temporal and spatial information. Such strategies have been successfully applied to domains including environmental event detection [11], human activity recognition [12] and [13], and object classification [14], resulting in improved robustness and classification accuracy. Nevertheless, despite the effectiveness of audio–visual fusion in these fields, no existing studies have explored its application to melon ripeness and sweetness classification. This research, therefore, addresses a clear gap in the literature by extending established multimodal CNN paradigms to agricultural quality assessment through the joint exploitation of acoustic and visual cues.

III. Methods

In this study, we used CNN as our deep learning architecture to construct a model for melon ripeness classification. In this section, a brief explanation of all methods used in this research is provided.

A. CNN Architecture

CNN is a deep learning architecture specifically designed for processing data with a grid-like topology, such as images and time-series data. It is characterized by its ability to automatically extract relevant features from raw data through layers of convolution and pooling operations. The convolution layers apply filters to the input data, capturing spatial hierarchies and patterns such as edges, textures, and more complex features in deeper layers. Pooling layers, on the other hand, reduce the spatial dimensions of the data, making the model computationally efficient while preserving the most critical information. CNN typically includes several main components: convolutional layers, activation functions (e.g., ReLU), pooling layers, fully connected layers, and an output layer with a softmax or sigmoid function for classification tasks. Advanced CNN architectures often integrate additional mechanisms, such as dropout for regularization or batch normalization to accelerate training.

Convolutional Neural Networks (CNNs) are a type of feed-forward neural network designed to process visual data. In CNNs, individual neurons are arranged to respond to overlapping regions within a visual area [15]. CNNs are hierarchical networks composed of multiple convolutional layers and subsampling (pooling) layers. The architecture can be varied by adjusting the number and type of convolutional and pooling layers. The convolutional layer operates based on parameters such as the size and number of feature maps, kernel size, skipping factors, and connection tables. A kernel slides across the input image, capturing spatial features, while skipping factors determine how many pixels are skipped during this operation. The output map size is calculated using a specific formula. Each map in a layer connects to several maps in the previous layer, and neurons in the same map may share weights.

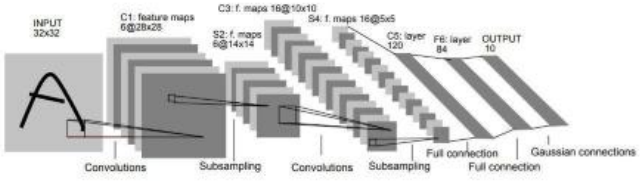

Fig. 1. CNN main architecture [16].

The CNN algorithm consists of two main stages: convolution and subsampling (pooling). For example, in the MNIST dataset architecture (Fig. 1), a grayscale image with dimensions 32x32 pixels is processed through two pairs of convolution and pooling layers, followed by two fully connected layers. Key steps include:

1) First convolution layer: Six 5x5 kernels produce six feature maps of 28x28 pixels using stride 1 and a valid mode (no padding).

2) First pooling layer: Reduces the feature maps to 14x14 pixels with a subsampling factor of 2.

3) Second convolution layer: Sixteen 5x5 kernels create 16 feature maps, every 10x10 pixels.

4) Second pooling layer: Further reduces the feature maps to 5x5 pixels.

5) Fully connected layers: The final pooling layer outputs connect to 120 neurons, followed by 84 neurons, and finally to 10 output neurons representing the classification categories.

6) Activation functions: Non-linear activation functions (e.g., Tanh, Sigmoid, or ReLU) are applied between convolution and pooling layers, as well as between the fully connected layers.

This hierarchical structure allows CNNs to automatically extract and learn relevant features from raw input data, making them highly effective for tasks such as image recognition and classification.

The foundation of Convolutional Neural Networks (CNNs) lies in the architecture of Neural Networks (NNs). However,

CNNs have three main advantages over traditional NNs: sparse connectivity, parameter sharing, and equivariance, which make them more efficient and effective for processing structured data, especially images.

- Sparse Connectivity: unlike traditional NNs, CNNs only connect a neuron to a small region of the input (called a receptive field). This sparse connectivity significantly reduces the number of parameters to train.
- Parameter Sharing: the same set of weights (or filters) is applied across the entire input. This parameter sharing allows CNNs to detect the same feature (e.g., edges, corners) regardless of its position in the input. As a result, the model becomes more efficient in terms of memory and computation, and it can generalize better to unseen data.
- Equivariance: CNNs exhibit equivariance to translation, meaning that a shift in the input (e.g., moving an object in an image) results in a corresponding shift in the feature map output. This property makes CNNs robust to variations in object positions, which is crucial for tasks like image recognition.

These three properties collectively allow CNNs to efficiently handle high-dimensional data like images, reduce overfitting, and achieve better performance in tasks involving spatial or temporal patterns.

B. Spectrogram Image

Although CNN architecture comes in various forms, the most notable and successful is the 2D-CNN architecture. For image data, which is already represented as a two-dimensional matrix, CNN-2D can be applied directly. However, this differs from acoustic data, whose representation is typically in the form of a one-dimensional vector (1-D). A common approach for classifying audio data using CNN-2D is to convert the audio data into a spectrogram. A spectrogram is an image that depicts the magnitude (intensity) of different frequencies over time. This spectrogram image is then used as a Deep Acoustic Feature (DAF) input for the 2D-CNN architecture. The process of creating a spectrogram involves the following steps:

1) Segmenting audio data: The audio data is divided into smaller overlapping chunks. For example, if an audio sample has a size of 1x7, with an overlap of 1 sample and a chunk width of 3 samples, the data is split accordingly.

2) Frequency decomposition using FFT: Assuming that a sound consists of various frequencies, each chunk is transformed using the Fast Fourier Transform (FFT). FFT extracts the frequency components of each chunk of audio at different time intervals, converting the data from the time-amplitude domain to the frequency-time domain.

3) Creating a frequency matrix: For instance, if each chunk is decomposed into 100 frequency components, each chunk will have 100 values. With 3 chunks, the results will result in a matrix of size 100 rows x 3 columns.

4) Mapping matrix to pixels: Each cell in the matrix is represented as a pixel with a specific color. The colors are assigned using a predefined threshold, where higher values are represented with brighter colors.

5) Generating the spectrogram image: The spectrogram is represented as an image, where the color of each pixel reflects the intensity of a frequency at a specific time. A pixel in the spectrogram at row i, column j represents the intensity of sound at the *i-th* frequency and the *s*-time range. Fig. 2 shows an example of a spectrogram from audio data.

6) Input to CNN-2D architecture: Since the spectrogram is a 2D grid of data (image), it can directly be used as input to a 2D-CNN architecture.

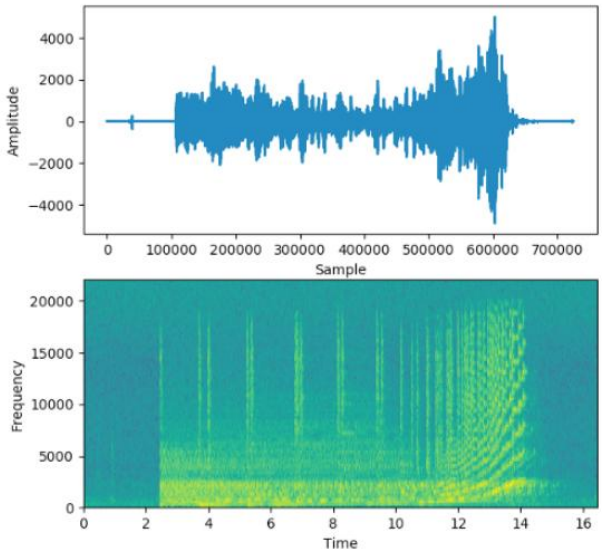

Fig. 2. Spectrogram image [17].

C. Mel Frequency Cepstral Coefficients (MFCC)

In addition to spectrograms, Deep Acoustic Features (DAF) can use Mel Frequency Cepstral Coefficients (MFCC) as input. The detailed MFCC extraction process involves the following steps [18]:

1) Frame blocking (Fig. 3): Audio signals are divided into overlapping frames to prevent information loss at frame transitions.

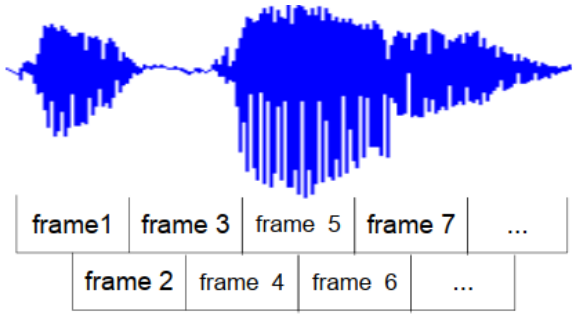

Fig. 3. Frame blocking procedures [18].

2) Windowing: A filter is applied to each frame to minimize distortion between frame blocks. This is done by multiplying each frame's data with a selected windowing

formula, commonly the Hamming window, due to its computational efficiency and effective filtering properties. The Hamming window equation is Eq. (1):

$$W(n) = 0{,}54 - 0{,}46\ cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

Variable descriptions

- $W(n)$ the value of the Hamming window at the n-th sample
- n the sample index within a frame, where n=0, 1, 2, …, N−1
- N the total number of samples in one frame (window length)

3) Fast Fourier Transform (FFT) is applied to the windowed frames, calculating the Discrete Fourier Transform (DFT), which outputs complex numbers. The real and imaginary components are calculated as Eq. (2) and Eq. (3):

$$Real\ X[k] = \sum_{i=0}^{N-1} x[i] cos\left(\frac{2\pi k i}{N}\right) \quad (2)$$

$$Imaginary\ X[k] = -\sum_{i=0}^{N-1} x[i] sin\left(\frac{2\pi k i}{N}\right) \quad (3)$$

Variable description

- $Real\ X[k]$ the real part of the DFT output at index k
- $Imaginary\ X[k]$ the imaginary part of the DFT output at index k
- $x[i]$ The input signal sample at time index i (windowed signal)
- i time-domain sample index, where *i=0, 1, 2 ,..., N−1*
- k frequency bin index of the DFT output
- N total number of samples in one frame (FFT size)

The next step in this stage is to compute $X[k]$ the magnitude of the FFT. The magnitude of a complex number is calculated from its real and imaginary components using Eq. (4):

$$X[k] = \sqrt{Real\ X[k]^2 + Imaginary\ X[k]^2} \quad (4)$$

4) Mel-Frequency Wrapping, human perception of sound frequency is linear for frequencies ≤ 1000 Hz, but logarithmic for higher frequencies. This relationship, represented by the mel-frequency scale, is expressed as Eq. (5):

$$F_{mel} = \begin{cases} 2595 * log_{10}(1 + \frac{F_{Hz}}{700}), & F_{Hz} > 1000 \\ F_{Hz}, & F_{Hz} \leq 1000 \end{cases} \quad (5)$$

Mel-frequency wrapping, then performed using Eq. (6):

$$X_i = log_{10}(\sum_{k=0}^{N-1} X(k).H_i(k)) \quad (6)$$

Variable descriptions

- F_{mel}, the frequency value measured on the Mel scale, which represents the perceived frequency by the human auditory system.
- F_{Hz}, frequency value measured in Hertz (Hz), representing the actual or linear frequency of the signal.
- X_i, frequency wrapping value of the i-th filter
- i, filter index, where i=1,2,…,n, and n is the total number of filters
- $X(k)$, magnitude value of the frequency component at index k
- $H_i(k)$, amplitude (height) of the i-th triangular filter at frequency index k
- k, Frequency index, where k=0, 1, 2,…, N−1
- N, total number of frequency magnitude bins

5) Cepstrum Calculation, in this final step, the mel-frequency is converted into the time domain using the Discrete Cosine Transform (DCT), Eq. (7):

$$C_j = \sum_{i=1}^{M} X_i\ cos\left(\frac{j(i-1)}{2}\ \frac{\pi}{M}\right) \quad (7)$$

Variable descriptions

- C_j, j-th cepstral coefficient
- j, Cepstral coefficient index, where j=1,2,…, the desired number of coefficients, for audio classification, we use maximum j=40
- M, number of Mel filter banks. On these study, we used 64 Mel filters
- X_i, The Mel-frequency wrapping value at frequency index i, where i=1,2,…,n. The n, we set it similar to the number of M

In this study, acoustic data were used for Mel-Frequency Cepstral Coefficients (MFCC).

IV. Data

A. New Dataset Acquisition

From 11th May 2021, the procedure for obtaining new sample data has begun at the Agro Technology Park (ATP) Cikarawang. The melon data acquired is from the Golden Alisha melon variety. In general, the morphological shape of the Golden Alisha tends to be round, has a smooth skin texture, and has a bright golden yellow colour. In data acquisition, four different days after planting (DAP), class 47 DAP on 11th May, 53 DAP (17th May), 60 DAP (24th May) and 67 DAP (30th May). There were 12 melons used as samples. From each measured sample, five acoustic data and image data were obtained, respectively. So that when the data acquisition is completed, 60 pieces of data will be obtained for each different DAP and in total, it was hoped that 240 pieces of data would be obtained in all. In the greenhouse as the data location, there are four melon planting lines, each planting position has a different light intensity (Fig. 4). Based on these conditions, in order to be sufficiently representative, the measured melons were spread over four different planting lines. Twelve melons that were used as samples were spread out three each in each planting lane.

Fig. 4. Planting lines at greenhouse, ATP Cikarawang.

The tools used and measurement scenarios in data acquisition procedures are:

- Samsung A20 Smartphone camera, with 30 cm portrait distance, 3:4 dimension ratio, and 2.4x zoom aspect, with an image resolution of 3096x4128 pixels.
- Flux meter to measure light intensity during capturing data.

Some results of data acquisition procedures are shown in Fig. 5.

Fig. 5. Some data acquisition results.

As a result of the acquisition, finally obtained a dataset for ripeness classification are 230 acoustic knock (audio) and image data packets from 4 classes of DAP (47, 53, 60, and 67). Audio data in .m4a format, Image data used smartphone camera with dimensions of 3096x4128 pixels, Image data used Intel RealSense Depth camera with dimensions of 640x480 pixels. The proportion of data in each class is 55 records. The total amount of data is only 230 pieces of data and not 240 pieces of data because the acoustic data of a melon with ID 4 has been picked before 67 DAP and declared lost.

The dataset used in this study is relatively small (230 samples), which raises concerns about potential overfitting, particularly when employing deep CNN architectures. To mitigate this risk, data augmentation techniques were applied, including rotations, scaling, and flips, to artificially increase the diversity of training samples. These measures help ensure that the reported performance reflects the model's true capability rather than overfitting to the limited dataset.

V. Proposed Methods

In the case of classifying melons using two types of data: acoustic data from tapping the melon and image data of the melon, the following procedure is proposed for combining features from both data types. The integration is performed by merging the MFCC spectrogram values with the feature map values derived from the output of the last layer in the feature extraction stage (illustrated in Fig. 6, stage 2, represented by the rightmost purple neurons).

Challenges in Combining Spectrogram and Feature Maps to merge spectrogram and feature map data include:

- Dimensional Mismatch: The dimension of the spectrogram may vary for different acoustic data samples. Additionally, the spectrogram's dimensions may not match the dimensions of the feature maps.
- Value Range Differences: The value range (or domain) of the spectrogram will differ from that of the feature maps.

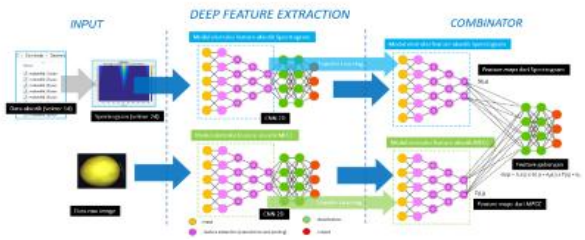

Fig. 6. Proposed method scheme.

Proposed Solutions to address these challenges, processed as these following steps. Combination Procedure - the procedure for combining the features is outlined in the diagram in Fig. 6. The value of the spectrogram, whose different dimensions can be combined by adding them with the trained weights. The vector of the combined feature G from the spectrogram S with weight A1, combined with the feature maps F with weight A2, can then be formulated using Eq. (8). This method is obtained by performing feature concatenation within a flatten layer, which allows the model to integrate information from both sources effectively before passing it to subsequent layers.

$$G(i,j) = A1(i,j)x\,S(i,j) + A2(i,j)xF(i,j) + b0 \qquad (8)$$

The weighted summation step runs after aligning the spectrogram's dimensions and normalizing its value range; it is combined with the feature map values by performing a weighted summation. Furthermore, the values of A1 and A2 are trained using the feed-forward and backpropagation approaches to determine the correct values. As for b0, it is the bias value.

VI. Experiments

A. Experimental Scenario

This section presents various experimental results conducted on this study. The CNN architecture used is illustrated in Fig. 7.

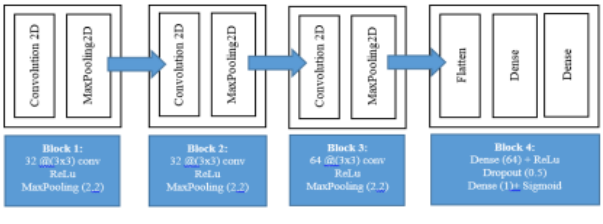

Fig. 7. CNN architecture on experiment.

In this study, there are three main experiment scenarios: (1) image-based only, (2) acoustic-based only using MFCC features, and (3) a combined multimodal CNN architecture. Data augmentation is applied to both image and audio data in

all scenarios. Augmentation of the audio data is applied before the MFCC spectrogram is constructed.

From a preliminary study, especially for acoustic-based classification model was investigated by comparing four experimental scenarios. The first scenario employed a CNN with Mel Spectrogram features and achieved moderate performance, with a precision of 0.62, recall of 0.55, F1-score of 0.55, and accuracy of 0.55. The second scenario utilized a CNN with MFCC Spectrogram features, which resulted in a consistent performance improvement, reaching a precision of 0.72, a recall of 0.65, an F1-score of 0.64, and an accuracy of 0.65. The third scenario incorporated data augmentation into the Mel Spectrogram-based CNN. However, this configuration did not yield a significant performance gain, as the accuracy remained at 0.55. In contrast, the fourth scenario, which combined a CNN with MFCC Spectrogram features and data augmentation, demonstrated a substantial improvement across all evaluation metrics. This approach achieved the highest precision (0.82), recall (0.80), F1-score (0.80), and accuracy (0.80) among all tested configurations. Based on these comparative results, it can be concluded that MFCC-based representations augmented through data augmentation provide a more robust and discriminative characterization of acoustic patterns. Consequently, the CNN with MFCC Spectrogram and data augmentation configuration is identified as the most effective approach and is therefore adopted as the baseline model in this study.

The hardware and software specifications used for this study are as follows:

- Hardware Environment Processor AMD Ryzen 9 5900HX (16 CPUs, 3.3 GHz) with Radeon Graphics. Intel Quad Core i7-4700HQ with a clock speed of 2.4 GHz, 16GB DDR5, and an SSD with 1 TB capacity.
- Software Environment and Supporting Applications: Operating System: Windows 11 Home Single Language 64-bit (Version 10.0, Build 26100). Python 3.7 and the Keras library with TensorFlow backend. Anaconda with Jupyter Notebook environment.

B. Experimental Results

This section presents the results of the classification experiments based on three different input modalities: image-based CNN, acoustic-based CNN, and multimodal CNN combining both image and audio features. The performance of each scenario is evaluated using precision, recall, and F1-score for each ripeness class (DAP47, DAP53, DAP60, and DAP67), as shown in Fig. 8- Fig. 10.

	precision	recall	f1-score	support
DAP47	1.00	0.75	0.86	20
DAP53	0.69	1.00	0.82	20
DAP60	0.89	0.80	0.84	20
DAP67	1.00	0.90	0.95	20
accuracy			0.86	80
macro avg	0.89	0.86	0.87	80
weighted avg	0.89	0.86	0.87	80

Fig. 8. Image-based CNN scenario result.

	precision	recall	f1-score	support
DAP47	0.88	0.70	0.78	20
DAP53	0.63	0.85	0.72	20
DAP60	0.89	0.85	0.87	20
DAP67	0.89	0.80	0.84	20
accuracy			0.80	80
macro avg	0.82	0.80	0.80	80
weighted avg	0.82	0.80	0.80	80

Fig. 9. Acoustic-based CNN scenario result.

Classification Report:

	precision	recall	f1-score	support
DAP47	0.95	1.00	0.98	20
DAP53	1.00	0.75	0.86	20
DAP60	0.83	1.00	0.91	20
DAP67	1.00	1.00	1.00	20
accuracy			0.94	80
macro avg	0.95	0.94	0.94	80
weighted avg	0.95	0.94	0.94	80

Fig. 10. Multimodal CNN scenario result.

TABLE I. EXPERIMENTAL RESULT SUMMARY

Scenario	Prec.	Rec.	F1-score	Acc.
Image-Based CNN	0.89	0.86	0.87	0.86
Acoustic-Based CNN	0.82	0.80	0.80	0.80
Multimodal CNN	0.95	0.94	0.94	0.94

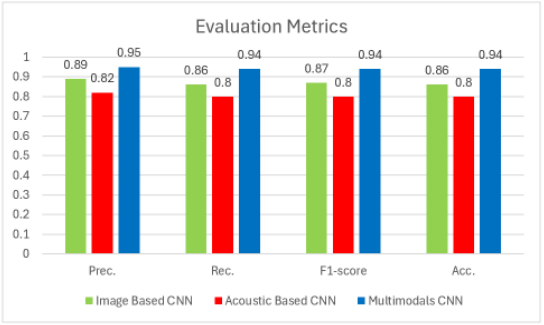

Fig. 11. Evaluation metrics.

For image-based CNN scenarios (Fig. 8), the model achieved an overall accuracy of 86%. The highest performance was observed in the DAP67 and DAP47 classes, with F1-scores of 0.86 and 0.86, respectively. However, the model struggled with class DAP53, which obtained a relatively low F1-score of 0.72. This indicates that visual features alone, while generally effective, may be insufficient for distinguishing between the earlier ripeness stages where external appearance changes are subtler.

On the other hand, the acoustic-based CNN scenario (Fig. 9) yielded a slightly lower overall accuracy of 80%. The best-performing classes were DAP60 and DAP67, with F1-scores of 0.87 and 0.84, respectively. Class DAP53 again presented the most difficulty, with an F1-score of just 0.72. These results suggest that audio features extracted from tapping sounds can represent internal fruit characteristics but may still

be ambiguous for similar ripeness levels, especially between DAP47 and DAP53.

Furthermore, for the last scenario, multimodal CNN (Fig. 10), which integrates both image and audio features, achieved the best overall performance with an accuracy of 94% and a macro-average F1-score of 0.94. All classes showed strong and balanced results, with F1-scores above 0.90, including DAP53, which previously had the lowest performance in other scenarios. The fusion of visual and acoustic information enabled the model to capture both external and internal cues associated with fruit ripeness, resulting in significantly improved classification performance.

In summary (Table I and Fig. 11), the multimodal CNN approach outperforms single-modality models, confirming that combining complementary features from different domains can enhance the accuracy and robustness of ripeness classification systems. This finding supports the potential of multimodal learning in agricultural quality assessment tasks, especially when the classification target involves subtle, multi-dimensional characteristics.

VII. Discussion

In this section, further discussion will be provided regarding the experimental results obtained. It includes confusion matrix analysis for all scenarios. At the end of this chapter, a statistical analysis of the experimental results is also conducted.

The confusion matrices presented in Fig. 11 to Fig. 13 illustrate the classification performance of three different CNN models using image data, acoustic data, and a multimodal fusion of both modalities. These matrices provide deeper insights into how well each model is able to distinguish between the four ripeness stages of melon: DAP47, DAP53, DAP60, and DAP67.

Fig. 12 shows the confusion matrix for the image-based CNN model, which demonstrates relatively strong performance overall. The model perfectly classified the DAP53 and DAP67 classes, while DAP47 was correctly predicted in 15 out of 20 cases, with 5 samples misclassified as DAP53. Class DAP60 showed a minor confusion with DAP53, as 4 samples were misclassified. This suggests that while image-based features are effective in capturing external ripeness characteristics, early ripeness stages such as DAP47 still tend to overlap visually with neighboring classes like DAP53.

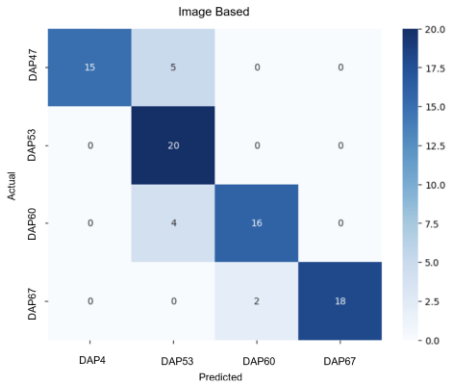

Fig. 12. Image-based CNN confusion matrix.

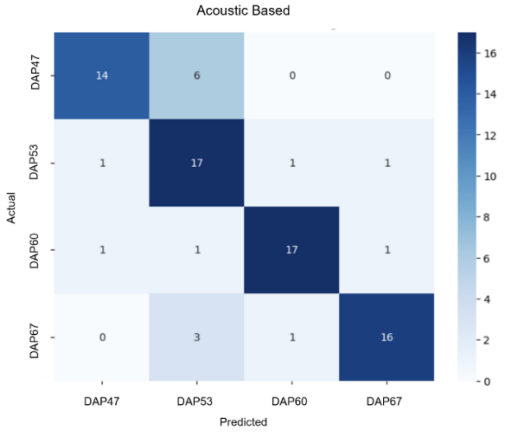

Fig. 13. Acoustic-based CNN confusion matrix.

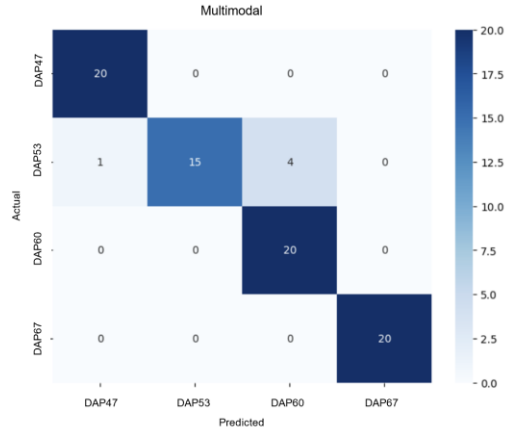

Fig. 14. Multimodal CNN confusion matrix.

In contrast, the acoustic-based CNN model with MFCC features and augmentation (Fig. 13) exhibited slightly lower precision across several classes. Class DAP47 was correctly identified 14 times but misclassified 6 times as DAP53. Similarly, DAP53 had 17 correct predictions, while the remaining 3 samples were confused with DAP47, DAP60, and DAP67. Notably, even DAP67, which was generally easy to classify, showed minor confusion in this setting. These results indicate that while acoustic features enriched with augmentation improve general robustness, they still present challenges in precisely distinguishing between classes with similar internal acoustic signatures, such as DAP47 and DAP53.

The most significant improvement is evident in Fig. 14, representing the multimodal CNN model that combines both image and audio features. This model achieves near-perfect classification: all samples from DAP47, DAP60, and DAP67 were classified correctly, while only one sample from DAP53 was misclassified as DAP47 and four as DAP60. The confusion in DAP53 remains the only notable weakness, which is consistent across models and suggests that this class may lie at a transitional point in both acoustic and visual characteristics. One notable difficulty during data acquisition was that the greenhouse was located near a rural road frequently used by vehicles. Consequently, acoustic recordings were often

contaminated with environmental noise, which likely contributed to the relatively lower performance of the acoustic-based model. This issue also affected the multimodal model, particularly in classifying samples from DAP53, where the presence of background noise and overlapping features made accurate classification more challenging. Despite this, the multimodal model achieved the best overall balance across all classes, demonstrating that fusing external (image) and internal (acoustic) cues significantly enhances the model's ability to discriminate between subtle ripeness stages.

In summary, the results affirm that while single-modality models (image-only or audio-only) are capable of capturing certain class characteristics, their performance is limited by the inherent ambiguities in either visual or auditory features alone. The multimodal CNN approach stands out as the most effective, offering a holistic representation of fruit ripeness and minimizing inter-class confusion. This supports the hypothesis that multimodal learning is highly beneficial for tasks requiring fine-grained classification, particularly in agricultural applications where the features are complex and multi-dimensional.

VIII. Statistical Analysis

This section presents a statistical analysis of the experimental results, with a particular focus on accuracy values. Statistical significance of performance differences among the three classifiers was evaluated using pairwise McNemar's tests on the same test samples. In addition, 95% confidence intervals and variance of accuracy were reported to assess the reliability and stability of the models.

The statistical analysis of the experimental results begins with the computation of the Standard Error (SE) and the Confidence Interval (CI). SE represents the magnitude of the sampling error of the accuracy estimate, indicating the extent to which the observed accuracy obtained from the test sample may deviate from the true performance on the broader data population. A smaller SE indicates a more stable and reliable performance estimate. CI provides a range within which the true model performance is expected to lie at a given confidence level (commonly 95%). The CI enables a more cautious interpretation of performance metrics, particularly for small datasets, as it reflects the lower and upper bounds of the possible performance values [19]. The formulations of SE and CI are presented in Eq. (9) and Eq. (10). In this analysis, the variance is also computed. The variance is used to measure the degree of variability or instability of the model performance across test data or multiple experimental runs, Eq. (11).

$$SE = \sqrt{\frac{p(1-p)}{n}} \tag{9}$$

$$CI = p \pm 1.96 \times SE \tag{10}$$

$$Var(p) = \frac{p(1-p)}{n} \tag{11}$$

Variable descriptions

- SE, Standard Error
- CI, Confidence Interval
- Var, variance
- p, denotes the accuracy value
- n, denotes the number of test samples
- 1.96 is the constant corresponding to the 95% confidence level

TABLE II. Stability Level of Classification

Scenario	Acc.	Var.	SE	CI
Image-Based CNN	0.86	0.00148	0.039	[0.79, 0.94]
Acoustic-Based CNN	0.80	0.00200	0.045	[0.71, 0.89]
Multimodal CNN	0.94	0.00073	0.027	[0.88, 0.99]

Based on the variance, SE, and CI values shown in Table II, the multimodal CNN exhibits the most stable and reliable performance. It achieves the lowest variance (0.00073) and SE (0.027), along with the narrowest CI range, indicating more consistent performance across test samples. Therefore, multimodal CNN can be considered the best-performing and most reliable model among the evaluated approaches. In contrast, the acoustic-based CNN shows the highest variance and SE, suggesting greater performance variability and lower stability. Pairwise, statistical significance between classifiers was assessed using McNemar's test [20] on the same test samples with continuity correction. Pairwise comparisons were conducted between the multimodal model and each unimodal baseline (image-based and acoustic-based) compute using Eq. (12) and data in Table III. In McNemar's test, n_{01} and n_{10} denote the number of discordant pairs where only one classifier produces a correct prediction, and the continuity-corrected chi-square statistic is computed to assess the significance of performance differences. If the Chi-Square χ^2.

$$\chi^2 = \frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}} \tag{12}$$

TABLE III. Samples with False Prediction

Scenario	Samples with false prediction
Image-Based CNN	DAP47_03, DAP47_7, DAP47_12, DAP47_18, DAP47_19 DAP60_03, DAP60_08, DAP60_17, DAP60_19 DAP67_06, DAP67_11
Acoustic-Based CNN	DAP47_02, DAP47_05, DAP47_09, DAP47_14, DAP47_19, DAP47_20 DAP53_02, DAP53_08, DAP53_15 DAP60_01, DAP60_04, DAP60_17 DAP67_03, DAP67_08, DAP67_12, DAP67_17
Multimodal CNN	DAP53_02, DAP53_11, DAP53_15, DAP53_17, DAP53_18

The McNemar test statistic was compared with the chi-square critical value of 3.84 (df = 1, α = 0.05) to assess statistical significance. For the comparison between the Multimodal CNN and the Image-based CNN, $n_{01} = 11$ denotes samples misclassified by the image-based model but correctly classified by the multimodal model, while $n_{10} = 5$ denotes samples correctly classified by the image-based model but misclassified by the multimodal model. The resulting test statistic is χ^2=1.56, which is lower than the critical value

(3.84). In contrast, for the comparison between the Multimodal CNN and the Acoustic-based CNN, $n_{01} = 14$ and $n_{10} = 3$, yielding a test statistic of χ^2=5.88, which exceeds the critical value. These results indicate that the multimodal CNN significantly outperforms the acoustic-based CNN. whereas the improvement over the image-based CNN is not statistically significant. Nevertheless, it achieved an accuracy 8% higher than the image-based model.

IX. Conclusion

This study evaluated three different scenarios for audio-visual classification: image-based only, acoustic-based using MFCC features, and a combined multimodal CNN architecture. In all scenarios, data augmentation was applied to both image and audio data, with audio augmentation performed prior to the construction of the MFCC spectrogram.

The experimental results demonstrate that the multimodal approach, which integrates both image and acoustic features, consistently outperforms the single-modality models. Specifically, the multimodal CNN with data augmentation achieved the highest performance across all evaluation metrics, confirming the effectiveness of combining complementary feature representations. Furthermore, applying data augmentation on both modalities contributed to improved model generalization and robustness, highlighting the importance of augmentation strategies in multimodal learning tasks.

Based on statistical analysis, the multimodal CNN achieved the highest accuracy (0.94), outperforming the image-based CNN (0.86) and acoustic-based CNN (0.80). The 95% confidence interval of the multimodal model ([0.88, 0.99]) is narrower than those of the unimodal models, indicating higher reliability. Furthermore, the variance of the multimodal CNN (0.00073) is lower than that of the image-based (0.00148) and acoustic-based (0.00200) models, demonstrating more stable performance. Pairwise McNemar's tests indicate that the performance improvement of the Multimodal CNN over the acoustic-based CNN is statistically significant ($p < 0.05$), whereas the improvement over the image-based CNN is not statistically significant.

Overall, the findings suggest that leveraging multimodal data with appropriate augmentation techniques is a promising direction for improving classification accuracy in audio-visual applications. Future work may explore more advanced fusion techniques and augmentation methods to further enhance model performance.

Limitations and Future Work: The current study is limited by the small dataset size, controlled greenhouse conditions that do not reflect natural variability, and the lack of external validation. Future research should focus on collecting larger and more diverse datasets under real-world conditions to improve model generalizability and robustness.

Acknowledgment

This study and the research behind it would not have been possible without the exceptional support of my supervisor, Prof. Agus Buono, Dr. Karlisa Priandana, and Prof. Dwi Guntoro. Their enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track from my first draft to the final revision of this study.

References

[1] Hamid A. S., Mohamed A., Jiang H., & Pen G. 2012. Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4277-4280 25-30 March 2012, Kyoto, Japan.

[2] Hamid A. S., Mohamed A. R., Jiang H., Deng L., Pen G. & Yu D. 2014. Convolutional Neural Networks for Speech Recognition. IEEE/ACM Transactions on Audio, Speech, And Language Processing, VOL. 22, NO. 10, October 2014.

[3] Palaz D., Collobert R., dan Doss M.M. 2013. Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks. arXiv preprint arXiv:1304.1018.

[4] Schlüter J. dan Böck S. 2014. Improved Musical Onset Detection With Convolutional Neural Networks. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014).

[5] Khamparia A, Gupta D, Nguyen NG, Khanna A, Pandey B and Tiwari P. 2019. Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," in IEEE Access, vol. 7, pp. 7717-7727, 2019, doi: 10.1109/ACCESS.2018.2888882..

[6] Jaiswal K and Patel DK. 2018. Sound Classification Using Convolutional Neural Networks. 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, 2018, pp. 81-84, doi: 10.1109/CCEM.2018.00021.

[7] Maccagno A, Mastropietro A, Mazziotta U, Scarpiniti M, Lee YC, and Uncini A. 2020. A CNN Approach for Audio Classification in Construction Sites. In: Esposito A., Faundez-Zanuy M., Morabito F., Pasero E. (eds) Progresses in Artificial Intelligence and Neural Systems. Smart Innovation, Systems and Technologies, vol 184. Springer, Singapore. https://doi.org/10.1007/978-981-15-5093-5_33

[8] Jmour N, Zayen S, and Abdelkrim A. 2018. Convolutional Neural Networks for image classification. 2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET), Hammamet, 2018, pp. 397-402, doi: 10.1109/ASET.2018.8379889

[9] Kausar A, Sharif M, Park J, and Shin DR. 2018. Pure-CNN: A Framework for Fruit Images Classification. 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 404-408, doi: 10.1109/CSCI46756.2018.00082.

[10] Sultana F., Sufian A, & Dutta P. 2018. Advancements in Image Classification using Convolutional Neural Network. 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2018, pp. 122-129, doi: 10.1109/ICRCICN.2018.8718718.

[11] González A. A. & Santiago A. M. 2025. CNN-Based Road Event Detection Using Multiaxial Vibration and Acceleration Signals," *Applied Sciences*, vol. 15, no. 18, p. 10203, Sep. 2025. doi: 10.3390/app151810203.

[12] P. Thottempudi, V. Kumar, and R. Kumar. 2025. Dynamic multi-modal attention network for robust and real-time through-wall human activity recognition. Results in Engineering, vol. 28, p. 107632, Dec. 2025, doi: 10.1016/j.rineng.2025.107632.

[13] M. Cao, J. Wan, and X. Gu, "CLEAR: Multimodal Human Activity Recognition via Contrastive Learning Based Feature Extraction Refinement," *Sensors*, vol. 25, no. 3, p. 896, 2025, doi: 10.3390/s25030896

[14] C. Miron, A. Pasarica and R. Timofte. 2021. Efficient CNN Architecture for Multi-modal Aerial View Object Classification," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021, pp. 560-565, doi: 10.1109/CVPRW53098.2021.00068.

[15] Ciresan CD, et al. 2011. Flexible, High Performance Convolutional Neural Networks for Image Classification. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI'11), Vol. Two, Pages 1237-1242.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[17] K. Doshi, 2021. Audio Deep Learning Made Simple: Sound Classification, Step-by-Step, *Medium*, Mar. 19, 2021. [Online]. Available: https://medium.com/data-science/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5

[18] Clara, 2011. Application of Mel Frequency Cepstrum Coefficients (MFCC) as Feature Extraction on Phoneme Recognition with Probabilistic Neural Network (PNN) as Classifier, Bachelor's thesis, IPB University, Bogor, Indonesia.

[19] T. Hastie, R. Tibshirani, and J. Friedman, 2013. An Introduction to Statistical Learning: With Applications in R, 1st ed. New York, NY, USA: Springer.

[20] Q. McNemar, 1947. Note on the sampling error of the difference between correlated proportions and percentages, Psychometrika, vol. 12, no. 2, pp. 153–157