Classification of Watermelons based on Ripeness using Multimodal Data

Akhilesh Joshi[1][0009−0006−8081−1373], Saniya Kadarbhai[2][0009−0003−5976−6449], Atharv Shirgurkar[3][0009−0002−6489−0797], Srushti Padanadi[4][0009−0001−1142−3645], Rajashri Khanai[6][0000−0002−5080−722X], and Prema T Akkasaligar[5][0000−0002−2214−9389]

KLE Technological University's Dr. M. S. Sheshgiri College of Engineering and Technology, Belagavi 590001, Karnataka, India.
rajashrikhanai.mss@kletech.ac.in, premaakkasaligar.mss@kletech.ac.in

Abstract. Qilin watermelon is a unique and sought-after variety known for its rich flavor and high market value. Motivated by SDG 2 (Zero Hunger), this study aimed to improve the efficiency and precision of ripeness classification to reduce post-harvest losses and improve food security. Using deep learning and machine learning techniques, this work explored a two-phase approach using visual data and tapping sounds of watermelons. In the image classification phase, the VGG-16 model is fine-tuned with data enhancement, achieving 85% test precision. For audio-based classification, spectral imaging yielded an accuracy of 84%, while the extraction of MFCC features combined with a random forest classifier achieved an exceptional accuracy of 98%. The novelty lies in the integration of diverse data modalities for ripeness prediction, enhancing reliability. These promising results underscore the potential for scalable applications in agriculture, helping stakeholders optimize harvest and reduce waste. However, challenges remain with the generalizability to other watermelon varieties and its sensitivity to noisy audio data. Future efforts will address these limitations by expanding dataset, improving robustness, and optimizing the model for efficient hardware deployment.

Keywords: Watermelon · Soundwave · Signal · MFCC · Ripeness.

1 Introduction

Watermelon is one of the highly consumed fruits, known for its refreshing taste and nutritional value. However,its quality is often inconsistent due to traditional methods of assessing ripeness. This paper aligns with Sustainable Development Goal (SDG) 2: Under Target 2.4 of the zero hunger initiatives, The quality and sustainability of watermelon can be enhanced by using advanced data analysis techniques. Globally, nearly 1 in 10 people faced hunger in 2022, while 2.4 billion experienced moderate to severe food insecurity. By implementing machine learning techniques, such as appearance and tapping sound analysis, the method seeks to enhance the precision of determining watermelon quality, contributing to reduced waste and improved efficiency in fruit production. Proposed work supports indicator 2.4.1, promoting sustainable agricultural practices and accelerating improvements in diet and nutrition, which are crucial for meeting the SDG target of reducing chronic undernutrition in children.

Recent publications used, the data generated by tapping watermelons to predict its ripeness. The work concluded that signal processing algorithms, such as Fourier transforms, are able to extract key features from the tapping sounds and to classify the watermelons based on ripeness. A thorough examination of the research papers highlights critical drawbacks. Furthermore, the work faced challenges in standardizing the tapping force. This resulted in negative influence on the signal data[1].

The objective of the proposed work is to establish an understanding that multimodal data:(images and sounds) classifies a watermelon into ripe and unripe. By analyzing audio recordings of tapping sounds and images of the watermelons, the present method aims to develop a model that predicts watermelon ripeness. In this regard the major contributions of the work are as follows:

- To demonstrate the efficacy of combining VGG16, Mel-Frequency Cepstral Coefficients (MFCC), and Random Forest for optimal performance in both visual and audio classification tasks.
- To investigate the use of spectral images with the VGG-16 convolutional neural network (CNN) architecture.

The remainder of the paper is organized as follows: Section II offers a brief overview of recent works, while Section III outlines the proposed methodology. Section IV presents the experimental details, along with the results and discussions. Finally, Section V concludes the paper.

2 Literature survey

In this section, the review of recent works related to the watermelon classification, organized into three distinct subsections based on different criteria is discussed. The classification allows for an in-depth analysis of various aspects concerning the, ripeness, visual characteristics, and the analysis of acoustic properties of watermelon.

2.1 Based on Ripeness

Baki *et. al.,*[2] developed an automated watermelon grading system using K-means clustering for color and shape detection, achieving 84.62 percent accuracy. Future work includes adding features like size and texture. Nazulan *et. al.,*[3] used MFCC and MLP neural networks to classify watermelon ripeness via acoustic signals, reaching 77.25 percent accuracy.

2.2 Based on visual characteristics

Vorobyev *et. al.,*[4] developed a machine vision application using K-means and OpenCV for watermelon quality assessment, available for public use. Varying lighting conditions may affect accuracy. J. Ho *et. al.,*[5] used Faster R-CNN on aerial photos for watermelon yield estimation, achieving 99 percent accuracy. Post-processing reduced error rates, yet initial detection missed some fruits.

2.3 Based on analysis of Acoustic properties

Lin *et. al.,*[6] analyzed watermelon ripeness through tapping sounds, linking sound patterns to ripeness stages. Consistent tapping needed for reliability. Rahim *et. al.,*[7] correlated sucrose levels in watermelon samples using NIR spectroscopy, achieving high accuracy. Calibration models necessary for sample variability.

From the study, the identified summary is presented in Table 1. It identifies the implementation of different visual and acoustic techniques on various watermelon datasets. The identified gaps in existing research include, limited use of multimodal data (combining image and audio) for watermelon ripeness classification and the lack of robust noise-handling techniques in audio-based solutions. By integrating image and audio modalities, the proposed model addresses these gaps, offering adaptability in diverse real-world scenarios.

Table 1. Comparison of Various Visual and Acoustic Techniques used in Existing Literature.

Reference	**Year of publication**	**Approach**	**Dataset**	**Integration of visual and acoustic properties.**
Daosawang *et. al.,*[8]	2020	Fast Fourier Transform (FFT) for sound Classification.	30 watermelon weighing 1.5-2 kg.	Acoustic property was used.
H. Yi *et. al.,*[9]	2024	Used GTR-Net with stereo vision and RioU loss.	watermelon growth data with occlulsion and variable lighting.	Used Visual properties for feature extraction.
Sanchez-Galan *et. al.,*[10]	2022	K-Means clustering, Fuzzy Logic, Artificial Neural Networks, and Support Vector Machine.	320 watermelon images are labeled into three categories.	Visual properties have been used to categorize watermelons.
Martinez *et. al.,*[11]	2024	Canny edge detection and Convolutional Neural Networks.	Total 750 images, Images are categorized into 4 Categories.	Visual properties have been used.

3 Proposed Methodology

Qilin watermelons are a unique variety known for their exceptional taste, vibrant color, and quality. Originating from regions with optimal climate and soil conditions, they are known for their sweet taste and juicy flesh, making them popular among the consumers. Distinguished by their visual patterns and bright green skin color, Qilin watermelons require careful cultivation techniques, including proper irrigation and pest management, to ensure optimal sweetness. As consumer demand for high-quality produce grows, understanding the factors that influence the quality of Qilin watermelons becomes increasingly important.

The quality assessment of watermelons often relies on subjective methods of evaluation, which can lead to inconsistencies in determining its ripeness. Traditional techniques do not effectively correlate visual characteristics and acoustic properties. The proposed model is to develop a correlation between images of Qilin watermelon and sounds when stroked the watermelon with an external force.

The objective of this work is to establish an understanding that multimodal data:(images and sounds) classifies a watermelon into ripe and unripe. By analyzing audio recordings of tapping sounds and images of the watermelons, the method aims to develop a model that predicts watermelon ripeness.

The proposed methodology is implemented using VGG-16, MFCC and random forest. Fig. 3 shows the system model of classification of watermelon based on ripeness using multimodal data. The convolutional neural network (CNN) architecture, specifically VGG16, has proven effective in image classification tasks due to its hierarchical feature extraction capabilities. This enables the capture of complex spatial patterns in images, leveraging pre-trained weights for transfer learning. The Mel-Frequency Cepstral Coefficients (MFCC) are a widely adopted for feature extraction method, effectively capturing perceptual characteristics of sound. Ensemble learning techniques like random forest have demonstrated superior performance in classification tasks involving smaller datasets, such as those derived from MFCC features. The dataset has been categorised into Ripe and Unripe, based on sugar levels provided, allowing us to analyze and differentiate the data effectively. The synergy between VGG16, MFCC, and random forest enables optimal performance in both visual and audio classification tasks, making this combination a viable approach for multimodal classification problems.

3.1 VGG-16

Working with the watermelon images the proposed work is implemented a deep learning pipeline for classifying images of ripe and unripe fruits using the pre-trained VGG-16 model. It starts by splitting the dataset into training and testing sets(70% train, 30% test). The images are preprocessed with ImageDataGenerator to normalize pixel values and augment the training set. The VGG16 model is loaded with pre-trained weights, and its layers are frozen to retain the learned features. Custom fully connected layers are added for binary classification.The model is trained for 40 epochs, evaluated on the test set and analyzed using a classification report and confusion matrix visualized as a heatmap. Fig. 1 signifies the workflow of VGG-16 model used in the proposed work.

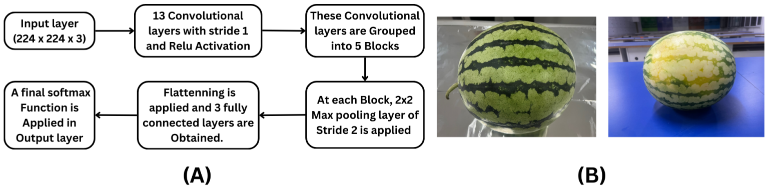
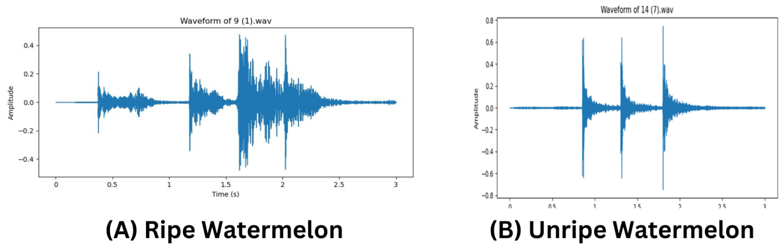

Input layer (224 x 224 x 3)
13 Convolutional layers with stride 1 and Relu Activation
These Convolutional layers are Grouped into 5 Blocks
At each Block, 2x2 Max pooling layer of Stride 2 is applied
Flattenning is applied and 3 fully connected layers are Obtained.
A final softmax Function is Applied in Output layer
(A)
(B)

Fig. 1. (A)Proposed VGG-16 Architecture for Watermelon Classification and (B) Instance of images used in Proposed Model.

3.2 Spectogram

The proposed model processed watermelon tapping audio files in .wav format to generate and save their waveform images as .png files. Each audio file read from the folder, loaded the audio data using librosa and plotted the waveform to show amplitude over time. Finally, the waveform is saved in an output folder. Then applied the VGG-16 model to the spectral images, following the same process as with the watermelon image classification. This included organizing the spectral data, preprocessing it, and fine-tuning the model to classify the spectral patterns effectively. Fig. 2 shows the different spectograms generated using the Librosa framework to visualize and train the .wav files.



(A) Ripe Watermelon

(B) Unripe Watermelon

Fig. 2. Spectograms of Tapping sounds of the Watermelons.

3.3 Mel-Frequency Cepstral Coefficient

In this step audio features are extracted from ripe and unripe audio files using Librosa. Features like MFCC's, spectral centroid, spectral bandwidth,spectral rolloff, and other features are computed for each audio file. These features are stored in a CSV file. Once features for both ripe and unripe audio are extracted, labeled the data accordingly and merged the labeled datasets into a single CSV file. Using this merged dataset, random forest classifier is applied to classify the audio into ripe and unripe categories. The data is split into training (70%) and testing (30%) sets and the random forest model is trained on the training set. Predictions on the test set achieved a specific accuracy,this pipeline demonstrated and end-to-end workflow for audio feature extraction and classification using machine learning.

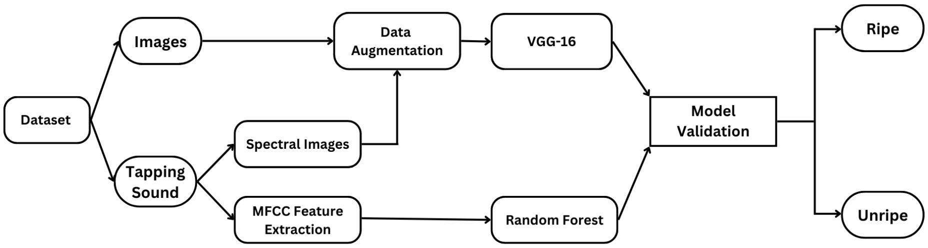


Fig. 3. Architecture for the Proposed Work

4 Results and Discussions

The proposed work is carried out on a device equipped with an AMD Ryzen 5 5600H processor, which operates at 3.30 GHz, along with 8 GB of RAM. The system runs a 64-bit version of Windows 11 Home and features an NVIDIA GTX 1650 GPU.

For the implementation, we utilized several Python libraries. For data manipulation and analysis, Pandas and NumPy are utilized for handling and performing numerical computations. For machine learning tasks, Scikit-learn is employed for building and evaluating predictive models, while TensorFlow is used for any deep learning models. For audio processing, Librosa is used to analyze audio features and extract relevant acoustic properties from the wav files. MFCC (Mel-frequency Cepstral Coefficients) facilitates feature extraction from wav files. Additionally, Matplotlib and Seaborn are used for data visualization.

The Qilin Watermelon Dataset [12] contains 19 folders, Each folder consists of 4 directories: "picture" contains the Watermelon images clicked from 9 different angles, "audio" contains wave recordings of tapping sounds produced when

striking the watermelons and a "chu" directory that contains processed data of watermelon image and wave. Each file follows a naming convention of "unique identifier sweetness measurement". For example, "001 12.jpg" represents watermelon sample number 001 with a sweetness measurement of 12. The data is artificially increased to enhance the model's accuracy. Operations such as flipping and rotations are applied in order to augment the data. Furthermore, the dataset is split into 70%-15%-15% for training, validation, and testing, respectively. The model is trained for 40 epochs.
In other words the dataset consists of 171 watermelon images resized to 128x128 pixels and audio tapping sounds converted to Mel spectrograms with 128 Mel bands and 100 time frames. Preprocessing included normalization of image pixel values to [0, 1], audio feature standardization, and data augmentation for images (rotation, shifts, and flips) to increase diversity. The Number of images after augmentation are 513.

The implementation of the proposed workflow of classification of watermelons based on ripeness using multimodal data is discussed. This work is organized into 2 different phases: Phase 1 and Phase 2. Phase 1 focuses on classifying the labeled data into 2 categories: Ripe and Unripe. Whereas, Phase 2 addresses the classification of watermelon tapping sounds into Ripe and Unripe categories.

4.1 Watermelon Images Classification

A pre-trained VGG-16 model is employed to classify the Qilin watermelon images. After obtaining high training accuracy, the model highlighted need for strategies to improve generalization. To mitigate overfitting, L2 Regularization technique is implemented. The application of data augmentation and regularization techniques improved the model's generalizability. The Training time is approximately 20 minutes for the proposed work. Fig. 4(A) shows that, the

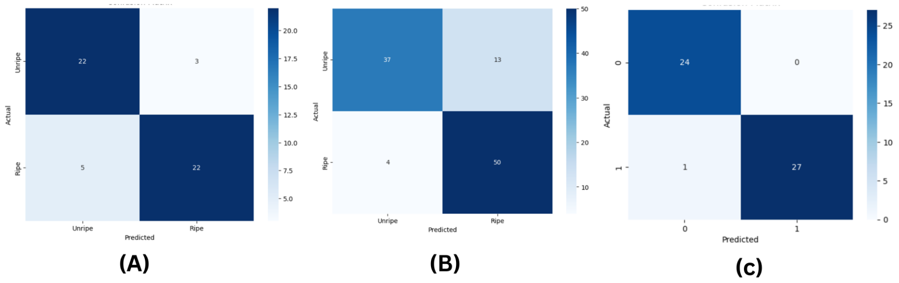


Fig. 4. (A) Confusion matrix for Audio Classification using Spectrogram + VGG-16, (B)Confusing Matrix for Image classification using VGG-16 and (C)Confusion matrix for Audio Classification using MFCC + Random Forest.

model correctly classified 22 images of unripe watermelons as unripe, 3 unripe watermelons are misclassified as ripe, it correctly classified 22 ripe watermelons as ripe. However, 5 ripe watermelons are misclassified as unripe.

4.2 Watermelon Audio Classification

The following are two different approaches that are employed for audio classification.

Spectral images with VGG-16: Spectrograms are generated from the audio files to convert them into visual representations. Using a similar methodology as the image classification phase, the data is divided into training, validation, and testing splits. Fig. 4(B) shows that, 37 unripe watermelons are correctly classified as unripe, while 13 are incorrectly classified as ripe. Similarly, 50 ripe watermelons were correctly classified as ripe, while 4 are misclassified as unripe. Standardizing of tapping for is one of the challenges faced during the preprocessing of pipeline standardized audio inputs by normalizing the Mel spectrograms and truncating or padding them to a fixed size.

MFCC Features with Random Forest: MFCC features are extracted from the audio files and saved in a CSV file, with labels indicating ripe or unripe categories. A random forest classifier is trained on the extracted features, achieving a significantly higher performance. Ripe audio samples exhibit lower average energy, smoother transitions in higher-order MFCCs, and more consistent variances, reflecting softer and more uniform textures. In contrast, unripe samples show higher energy, sharper MFCC variations, and greater variability, indicating harder and less uniform material properties. Spectral features like centroid values also highlight softer textures in ripe samples and stiffer surfaces in unripe ones.

Fig. 4(C) shows us that the model classified 24 unripe samples correctly as unripe, with no misclassification in this category. Only one unripe sample is misclassified as ripe. For ripe samples, the model correctly classified 27 as ripe and misclassified 1 as unripe.

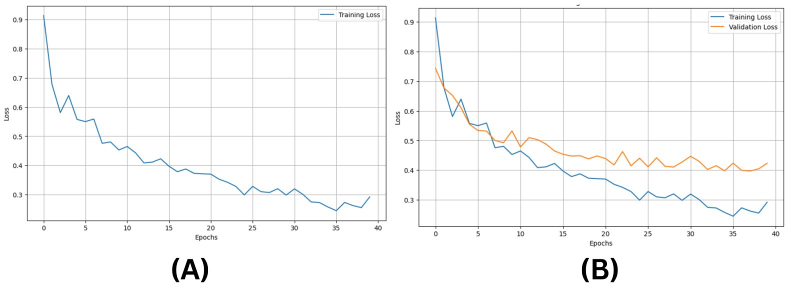
Table 2. Comparison of Results in proposed model with state-of-the-art works.

Reference	**Approach**	**Ripe**	**Unripe**	**Accuracy**
Karjagi *et. al.*,[1]	VGG-16	0.90	0.91	0.93
	Random Forest	0.76	0.74	0.89
Proposed Work	VGG-16	0.81	0.88	0.85
	VGG-16 using Spectogram images	0.90	0.79	0.84
	Random forest using MFCC features	0.96	0.98	0.98

The Table 2 shows comparison of results of proposed model with state-of-the-art works. For images, VGG-16 is used with data augmentation and regularization, achieving an accuracy of 85%. For audio, two approaches are evaluated: spectrograms processed using VGG-16 achieved a precision of 84%, while MFCC features combined with a random forest classifier achieved the highest accuracy of 98%. The MFCC-based approach demonstrated superior performance compared to the others. It is observed that the proposed model performed better using the multimodel approach than in the existing literature.

4.3 Performance analysis

The plot in Fig. 5 (A) shows that: with the increase in number of epochs, the value of error(cost) decreases. Also, the plot in Fig. 5 (B) signifies the decrease in loss of the training and validation with respect to increase in number of epochs. Whereas Table 3 gives a brief evaluation highlighting the good perfomance of the proposed model.



(A)

(B)

Fig. 5. (A)Epoch vs Error Graph of Proposed Work and (B)Loss of Proposed Model

Error Analysis The error analysis reveals that misclassifications primarily occur in samples with poor image quality, such as blurry or underexposed images, and audio data affected by background noise or inconsistent tapping force. These factors can obscure the distinguishing features used by the model. Addressing these issues with better data preprocessing, noise reduction techniques, and controlled data collection environments can improve accuracy. Some misclassifications are predominantly raised from data inconsistencies as seen in the confusion matrix in Fig. 4, such as overlapping visual features between ripe and unripe watermelons or distorted audio signals caused by tapping force variations. Investigating these cases as future work would highlight the need for better feature engineering and advanced preprocessing methods, such as augmenting data with synthetic samples and employing noise-resilient audio processing algorithms.

Table 3. Evaluation Report of the Proposed Model

Model	Classification	Precision	Recall	F1-Score
RandomForest (MFCC Features)	Ripe	0.96	0.98	0.96
	Unripe	0.98	0.98	0.96
Watermelon Images (V66-16)	Ripe	0.88	0.85	0.91
	Unripe	0.81	0.85	0.88
Spectral Images (V66-16)	Ripe	0.79	0.85	0.91
	Unripe	0.90	0.81	0.78

4.4 Limitations and Future Work

The model may face challenges due to dataset imbalance, sensitivity to noise in audio recordings, and dependency on high-quality inputs from both modalities. Additionally, it may require retraining for other watermelon varieties or scenarios where either image or audio data is compromised.

Potential real-world deployment challenges Real-world deployment challenges include variability in image quality and ambient noise affecting audio data, which could impact prediction accuracy. Scalability may require optimizing the multimodal model to reduce computational resource demands, such as using lightweight architectures or deploying the model on edge devices with hardware constraints. Additionally, integrating automated data collection methods for diverse environments would support broader applicability. The proposed solution has significant potential in agricultural settings, offering farmers a cost-effective, automated method to classify watermelon ripeness. For scalability, lightweight model architectures and integration with mobile or edge devices could facilitate widespread adoption. Additionally, deploying the solution in diverse agricultural environments would require adaptations to handle varying lighting, audio conditions, and watermelon varieties.

Future Work The model's generalizability to other watermelon varieties and external factors such as lighting conditions and tapping variability is a recognized challenge. To address this, future iterations could include a more diverse dataset with various watermelon types and environmental conditions. Additionally, implementing controlled lighting environments for image capture and calibrated tapping mechanisms for audio recording can help mitigate variability and improve model robustness.
The Proposed model can be optimized for Raspberry Pi by employing techniques

like model quantization, pruning, and enhancing hardware acceleration for improved computational efficiency. Real-time processing capabilities and integration with edge AI frameworks such as TensorFlow Lite can improve performance and scalability can also be achieved by testing the model on diverse datasets and applying transfer learning techniques.

5 Conclusion

The multimodal classifies a watermelon into Ripe and Unripe. The evaluation of performance of VGG-16 model is done by categorizing the watermelons as ripe or unripe based on the images of the watermelon, tapping sounds of the image and images of the audio wave. The dataset of images is augmented so that various angles of images. The watermelon image dataset is split into 70% training, 15% validation and 15% testing. The model was trained for 40 epochs resulting in significat accuracy. For the audio, there are two methods. In the first method, The images of the audio wave is trained. In the second method, the model extracted the MFCC from the audio files and saved them in a csv file. Furthermore, the proposed work applied random forest algorithm for classification, which enhanced the accuracy to 98%. The trained model demonstrates exceptional predictive accuracy and stability. Future works involve the hardware implementation of this model and enhancing the precision of the model on real time input.

References

1. S. Karjagi, S. Neelappagol, S. P. S, V. S and V. Karjigi, "Watermelon Ripeness Detector Using Signal Processing," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-6, doi: 10.1109/PuneCon55413.2022.10014898.
2. S. R. M. Shah Baki, M. A. Mohd Z., I. M. Yassin, A. H. Hasliza and A. Zabidi, "Non-destructive classification of watermelon ripeness using Mel-Frequency Cepstrum Coefficients and Multilayer Perceptrons," The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 2010, pp. 1-6, doi: 10.1109/IJCNN.2010.5596573.
3. W. N. S. W. Nazulan, A. L. Asnawi, H. A. M. Ramli, A. Z. Jusoh, S. N. Ibrahim and N. F. M. Azmin, "Detection of Sweetness Level for Fruits (Watermelon) With Machine Learning," 2020 IEEE Conference on Big Data and Analytics (ICBDA), Kota Kinabalu, Malaysia, 2020, pp. 79-83, doi: 10.1109/ICBDA50157.2020.9289712.
4. G. Vorobyev, D. A. Subbotin and E. Losevskaya, "Watermelon Quality Determining from a Photo Using Machine Vision," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 2286-2289, doi: 10.1109/ElConRus51938.2021.9396492.
5. M. -J. Ho, Y. -C. Lin, H. -C. Hsu and T. -Y. Sun, "An Efficient Recognition Method for Watermelon using Faster R-CNN with Post-Processing," 2019 8th International Conference on Innovation, Communication and Engineering (ICICE), Zhengzhou, China, 2019, pp. 86-89, doi: 10.1109/ICICE49024.2019.9117374.

6. Y. -W. Lin, Y. -B. Lin, W. -L. Chen, C. -H. Chang and H. -K. Li, "Watermelons Talk: Predicting Ripeness through Tapping," in IEEE Internet of Things Magazine, vol. 7, no. 4, pp. 154-161, July 2024, doi: 10.1109/IOTM.001.2300251.
7. A. A. A. Rahim et al., "A numerical analysis of correlation between sucrose level measurement and near-infrared (NIR) for various grades of watermelon ripeness," 2013 International Conference on Technology, Informatics, Management, Engineering and Environment, Bandung, Indonesia, 2013, pp. 180-185, doi: 10.1109/TIME-E.2013.6611988.
8. W. Daosawang, K. Wongkalasin and N. Katewongsa, "A Study Sound Absorption for Ripeness and Unripe Classification of Watermelon," 2020 8th International Electrical Engineering Congress (iEECON), Chiang Mai, Thailand, 2020, pp. 1-4, doi: 10.1109/iEECON48109.2020.229521.
9. H. Yi, K. Song and X. Song, "Watermelon Detection and Localization Technology Based on GTR-Net and Binocular Vision," in IEEE Sensors Journal, vol. 24, no. 12, pp. 19873-19881, 15 June15, 2024, doi: 10.1109/JSEN.2024.3393916.
10. J. E. Sánchez-Galán, A. Henry-Royo, K. -H. Jo and D. Cáceres-Hernández, "Automatic Feature Detection and Classification for Watermelon (Citrillus lanatus)," 2022 International Workshop on Intelligent Systems (IWIS), Ulsan, Korea, Republic of, 2022, pp. 1-7, doi: 10.1109/IWIS56333.2022.9920868.
11. M. R. B. Martinez, K. M. D. Dayrit and A. N. Yumang, "Classification of Red Watermelon Varieties Using Canny Edge Detection and CNN," 2024 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 2024, pp. 47-52, doi: 10.1109/I2CACIS61270.2024.10649622.
12. Ryan Chen, Aria Fan, Evange He, Mor Ning, Jason Tang, Linhua Zhang, July 10, 2024, "Watermelon appearance and knock correlate data sets with sugar content", IEEE Dataport, doi: https://dx.doi.org/10.21227/wb3s-h174.