Report n°1

Non-destructive quality of Dona durian using knock sound-based machine learning and predictive modelling

MAÏSTRE Antoine
TOUPENCE Tom
VIDAL Théo

Table of contents

Introduction

Dona durian, a standout among Vietnamese durians, is valued for its flavor, quality, and yield, making it a key export product. Traditional quality assessments like cutting are destructive and reduce commercial value. This project leverages thermal camera technology and machine learning for nondestructive quality evaluation, boosting the value and global competitiveness of Vietnamese Durians. This project focuses on applying thermal camera technology combined with machine learning and predictive modeling to evaluate the quality of Dona durian in a non-destructive manner. This Advanced solution aims to enhance the commercial value of Vietnamese durians and improve their competitiveness in the global market.

Objective

The objective of this project is to successfully develop an AI-based model and embed it on a device for classifying durians. This system aims to leverage non-destructive technologies, such as Knock sound and predictive modeling, to evaluate and categorize the quality of Dona durians. By achieving this, the project seeks to ensure higher accuracy in quality control, improve efficiency in the sorting process, and meet the stringent standards of both domestic and international markets.

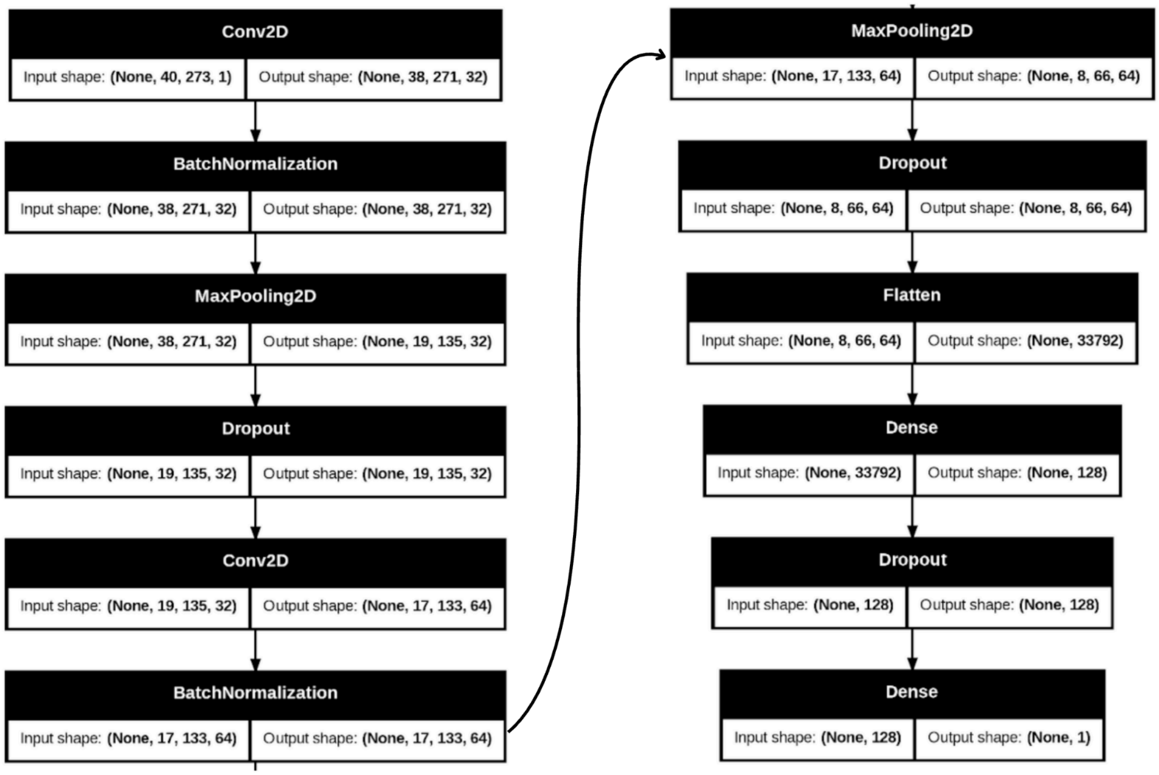
Methodology

The project follows a structured pipeline to design a reliable classification system:

1. **Data Collection**: Audio recordings of knock sounds on Dona durians are gathered using a standardized tapping setup. Each audio file is labeled with its corresponding ground truth quality class.
2. **Preprocessing**: Each audio sample undergoes noise removal and is converted into Mel-Frequency Cepstral Coefficients (MFCCs), a popular audio representation technique. Padding or trimming is applied to ensure uniform input shape.
3. **Dataset Preparation**:
    a. The MFCC matrices are normalized using Euclidean normalization (each mfcc column is being divided by its L2 norm
    b. The dataset is being shuffled, then split into training and test sets (respectively 80% and 20%)
4. **Model Design**: A convolutional neural network (CNN) architecture is defined, consisting of multiple Conv2D layers with ReLU activations, followed by pooling, batch normalization, and dropout for regularization.
5. **Training and Evaluation**: The model is trained on the training set, validated on a subset, and evaluated using accuracy, F1-score, and confusion matrix.

Techniques used

- **MFCC (Mel-Frequency Cepstral Coefficients)**
  Converts audio waveforms into a feature-rich time-frequency representation.
- **Librosa Library**
  Used for audio loading, MFCC extraction, and normalization.
- **Data Normalization and Padding**
  Ensures consistent input dimensions and value ranges.
- **Train/Test Split and Stratification**
  Maintains class balance during model training. We chose 80% for training data, and 20% for the testing data.
- **CNN (Convolutional Neural Networks)**
  Uses the normalized MFCCs like image data to classify durians based on tap sound. The model config is the following



It uses multiple convolutional layers, separated by batch normalization, max pooling and dropout to allow the model to generalize the training data. And two final dense layers, respectively 128 and 1 neurons : To treat CNN latent space results and extract a single value result.

- **TensorFlow/Keras Framework**
  Were used to build and train the CNN.
- **Evaluation Metrics**
  Accuracy, confusion matrix, precision, recall, and F1-score.

Results

The MFCC + CNN model achieves an overall **test accuracy of 0.84** and a **macro-averaged F1-score of 0.84** on the held-out data. As shown in **Figure 1**, the training loss plummets from ~22 to ~0.8 over 12 epochs, while the validation loss remains nearly flat around 0.6 – 0.7, signalling moderate over-fitting due to the limited dataset and the model's ≈ 0.5 M parameters. The corresponding accuracy curves in Figure 1 also highlight a gap between training and validation performance.

The confusion matrix in **Figure 2** reveals that **9 out of 11 "unripe"** samples (82 %) and **7 out of 8 "ripe"** samples (88 %) are correctly classified, for **16 correct predictions out of 19**. The full breakdown in the classification report (see **Table 1**) confirms these results.

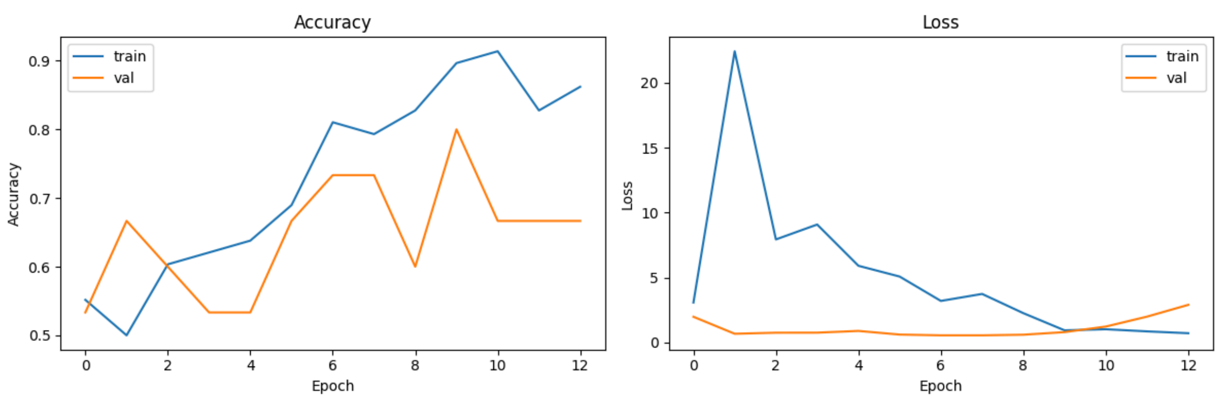


Figure 1 : Training vs. validation accuracy and loss curves.

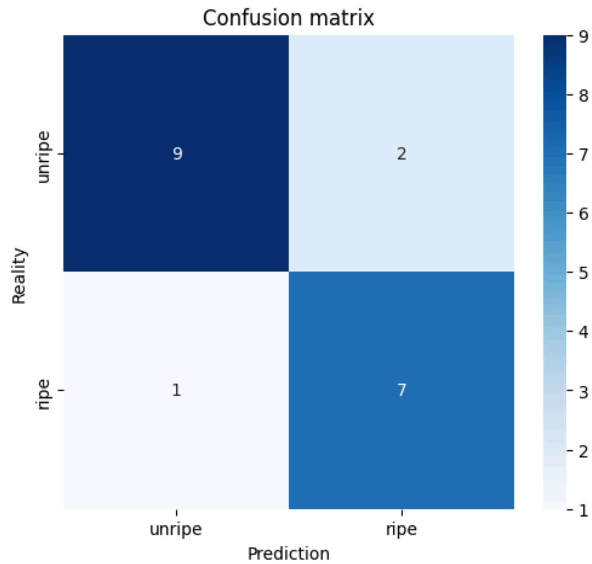


Figure 2 : Confusion matrix on the test set.

	precision	recall	f1-score	support
unripe	0.90	0.82	0.86	11
ripe	0.78	0.88	0.82	8
accuracy			0.84	19
macro avg	0.84	0.85	0.84	19
weighted avg	0.85	0.84	0.84	19

Table 1 : Precision, recall and F1-scores by class.

Comparison with previous work

Sanwibhuk & Boonma, in “Durian Ripeness Prediction through Knocking Sound”, trained CNNs on several time-frequency representations and reported the following accuracies:

Representation	Model	Accuracy
MFCC	CNN	0.65
Mel-Spectrogram	CNN	0.71

STFT (magnitude)	CNN	0.92

Their findings suggest that the finer time and frequency resolution of the short-time Fourier transform (STFT) captures resonance and damping patterns that MFCCs partially smooth out. Consequently, STFT-based features set a high baseline (0.92 accuracy) for purely acoustic, non-destructive durian grading and provide a clear target for future improvements.

Future Work

- **Audio augmentation** – Apply time-stretching, pitch-shifting, additive white noise, and impulse-response convolution to mimic real-world acquisition conditions.
- **Data augmentation** - Utilize IA tools to generate new dataset data from existing ones
- **Tweaking hyperparameters**
    - Evaluate lighter backbones (e.g., MobileNetV3-Small) or 1-D CNN/Transformer hybrids.
    - Increase regularisation (higher dropout, weight decay).

Implementing these directions should push accuracy beyond the industrial target of > 0.85 while securing the robustness required for automated sorting lines.