JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.3 September 2025: 249-261
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i3.13420

Hybrid CNN-Based Classification of Coffee Bean Roasting Levels Using RGB and GLCM Features

Rico Halim[1*], Mohammad Faisal Riftiarrasyid[2]

[1,2]Computer Science Program, Computer Science Department,
School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
rico.halim001@binus.ac.id, mohammad.riftiarrasyid@binus.ac.id

[*]Correspondence: rico.halim001@binus.ac.id

Abstract – *This study aims to develop a hybrid Convolutional Neural Network (CNN) model for classifying the roasting levels of Coffea arabica beans by integrating RGB color and GLCM texture features. A total of 1,600 high-resolution images were used, consisting of 1,200 training images and 400 testing images, evenly distributed across four roasting levels: Green, Light, Medium, and Dark. Local feature extraction was performed using a sliding window approach to capture fine-grained color and texture information from each image. Three model types were evaluated: a CNN with RGB-only input, a CNN with GLCM-only input, and a hybrid CNN with dual inputs. The hybrid model consistently demonstrated superior performance, achieving a validation accuracy of 99.74%, with minimal misclassification and stable convergence throughout training. Furthermore, six architectural variations of the hybrid model were tested by applying dropout and L2 regularization techniques. The model combining both dropout and L2 regularization achieved the most balanced results in terms of accuracy, generalization, and training stability. This research contributes an effective feature fusion strategy for fine-grained visual classification tasks, particularly in domains where inter-class visual differences are subtle. The proposed approach offers a cost-effective and scalable solution that is well-suited for real-time implementation in small to medium-sized coffee production facilities, and it shows strong potential for broader applications in agricultural product quality assessment.*

Keywords*: coffee bean classification; roasting level; RGB features; GLCM texture; hybrid CNN*

I. INTRODUCTION

Coffee is among the world's most important agricultural products in the global economy, with a projected market value of over US$130 billion (Ngure & Watanabe, 2024). The industry creates jobs for millions of smallholder farmers and builds a complex supply chain spanning production, distribution, and consumption all around many different countries (Samper & Quiñones-Ruiz, 2017). A specialty coffee market giving quality and sustainability top priority has improved the strategic position of the sector in the world economy in recent years (Freitas et al., 2024).

Indonesia is the fourth largest coffee producer in the world after Brazil, Vietnam, and Colombia, providing about 6.51% of the world's coffee exports with a production of about 500,000 tons annually (Ashardiono & Trihartono, 2024; Sutarmin et al., 2022). The sector not only consolidates the national economy through foreign exchange earnings but also sustains the livelihood of more than two million smallholder farmers (Anhar et al., 2021). Nevertheless, it remains confronted by issues of climate variability, low productivity, and poor access to modern technologies and markets (Kumar et al., 2022)

One of the most important phases in the coffee processing chain is roasting, which establishes the final sensory qualities of coffee, such as color, aroma, and taste. Through roasting, beans experience intricate physical and chemical changes like Maillard reactions and loss of moisture that cause drastic modifications in appearance and texture (Wei & Tanokura, 2015; Wang & Lim, 2015). Roasting degrees—from green through light, medium, to dark—directly affect the visual appearance of coffee beans, and hence visual information is a significant quality indicatorRoasting degrees—from green through light, medium, to dark—directly affect the visual appearance of coffee beans, and hence visual information is a significant quality indicator (Baqueta et al., 2020; Pratama et al., 2021).

Due to the strong visual and textural differences between roasting levels, there is substantial potential to automate classification using computer vision techniques. By extracting color and texture features from digital images, such systems can reduce reliance on manual judgment, which remains widespread in the coffee industry despite being subjective, inconsistent, and difficult to replicate (Domingues et al., 2020).

A number of machine learning and imaging-based methods have been proposed, such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Hyperspectral Imaging (HSI), and Near-Infrared Spectroscopy (NIRS). Although useful in certain situations, most of these techniques involve expensive instrumentation, complicated sample preparation, or are not feasible for real-time application—especially in small-scale or resource-limited environments (Castillo et al., 2019). RGB imaging and Gray Level Co-occurrence Matrix (GLCM) texture analysis are also promising low-cost options. RGB features record color differences related to roasting levels, while GLCM defines spatial texture patterns like homogeneity, contrast, and correlation. Both have demonstrated acceptable performance in agricultural classification applications, such as the detection of nutrient deficiency in leaves (Qur et al., 2020).

Earlier research has achieved remarkable classification accuracies of up to 97.22% using RGB, 95.31% using GLCM, and 97.78% with the fusion of both (Noel et al., 2019; Hendrawan et al., 2023). Most of these methods are, however, based on global feature representations, i.e., color histograms or Haralick descriptor aggregations, without regard to local spatial variations. This restricts their performance in fine-grained classification where subtle intra-class variations are important. Furthermore, recent research has shown that color feature bias in CNN models can result in overfitting or weak generalization (Claire et al., 2022), further justifying the necessity of more robust local feature modeling.

To overcome the shortcomings of global feature representations, this research suggests a hybrid method that integrates RGB color and GLCM texture features using a sliding window approach (Dawwd, 2019). By dividing coffee bean images into small local patches, the model can capture fine-grained visual hints concerning both color and surface texture—information that tends to get lost in the global representations.

Besides proposing a dual-input CNN architecture for processing RGB and GLCM features in parallel, this study also presents a direct comparison of the performance of models using RGB-only, GLCM-only, and combined features. The comparison is intended to assess the relative usefulness of each type of feature in discriminating between closely related roasting levels.

The innovation of this study is its focus on local feature extraction in a lightweight CNN architecture, which is especially well-suited for classification in small to medium-sized coffee processing plants. Through the use of inexpensive image-based features, the developed approach provides a scalable and viable alternative to more sophisticated or hardware-reliant approaches.

Thus, the primary aim of this research is to develop, train, and test a hybrid CNN model that combines RGB and GLCM features for efficient and accurate classification of coffee bean roasting levels, prioritizing real-time feasibility and generalizability.

II. METHODS

This chapter presents a comprehensive methodological framework adopted in this research for classifying coffee bean roasting levels using hybrid CNN architectures. The process consists of five major stages: (1) dataset collection and preprocessing, (2) visual feature extraction from RGB and GLCM, (3) CNN model design and configuration, (4) experimentation with multiple variations, and (5) model evaluation through quantitative metrics. An overview of the complete workflow is illustrated in Figure 1.

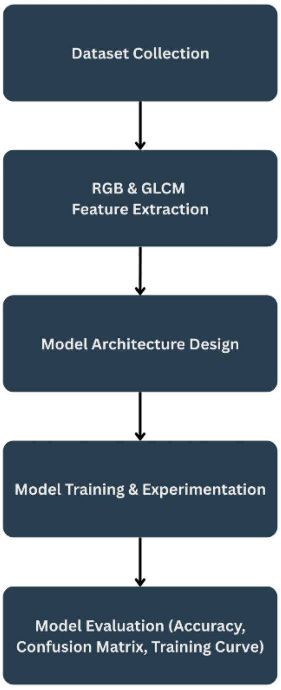

Dataset Collection
RGB & GLCM Feature Extraction
Model Architecture Design
Model Training & Experimentation
Model Evaluation (Accuracy, Confusion Matrix, Training Curve)

Figure 1. Overview of the Research Methodology Framework

2.1 Collecting Dataset

This research utilizes a high-resolution image dataset of roasted coffee beans for multiclass classification based on roasting level. The dataset, downloaded from the Kaggle platform under the title *"Coffee Bean Dataset Resized (224 × 224)"*, is categorized into four roasting levels: Green (unroasted), Light, Medium, and Dark. It comprises 1,200 training images—300 images per class—and 400 testing images—100 images per class. All coffee beans are of the Coffea arabica variety, with details: Laos Typica Bolaven (Green and Light), Doi Chaang (Medium), and Brazil Cerrado (Dark).

The images on the dataset were taken using the iPhone 12 Mini's 12-megapixel rear camera, with Ultra-wide and Wide lenses. The camera was positioned parallel to the plane of the object to maintain the consistency of the image capture angle. To enrich the variety and validity of the data, images were taken under two different lighting conditions, namely natural lighting and LED lights from a light box. Each sample was placed inside a transparent container to produce natural textures and noise that can represent real-world conditions.

All images were saved in PNG format with an original resolution of 3024 × 3032 pixels. Before being used in model training, these images were resized to 128 × 128 pixels and normalized to match the input requirements of the convolutional neural network. The dataset was then divided into two main parts, namely training data and testing data. In addition to the RGB images, each image also had its texture map extracted using the GLCM method.

With high variation in lighting conditions, surface texture, and visual quality, and a combination of systematically extracted local color and texture features, this dataset provides a strong basis for evaluating the effectiveness of a visual classification system on the roasting degree of coffee beans. An example of this dataset can be seen in Figure 2.

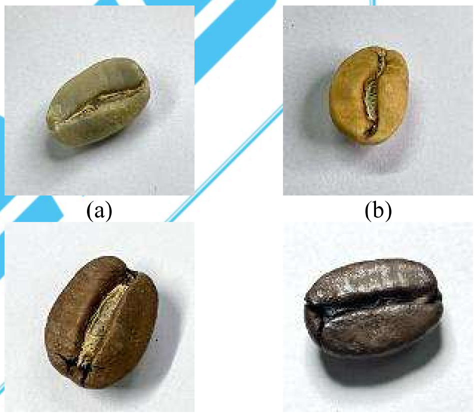

(a) (b) (c) (d)

Figure 2. Example of a coffee bean dataset: (a) Green, (b) Light, (c) Medium, (d) Dark

2.2 Extracting Features

Feature extraction in the current study provides a visual feature of the coffee beans that can be used by the classification algorithm. RGB channel-based color features and Gray Level Co-occurrence Matrix (GLCM) method-based texture features are the two significant categories of features utilized.

Color features are extracted from the RGB images after preprocessing, which includes resizing to the size of 128 × 128 pixels and normalizing the pixel values. Each pixel is described by three independent color channels: Red, Gree, and Blue, with each channel ranging from an intensity of 0 to 255. The choice of RGB features in this study is due to the visually apparent attributes of coffee beans, which present significant color modification throughout the roasting process.

Green coffee beans are green or pale yellow, light roast is light brown, medium roast is medium brown, and dark roast dark brown to blackish. The difference makes color an important attribute in the discrimination of roasting levels, thus RGB color representation is considered effective as an input feature in classification. In addition to color, the data related to texture form an important factor in determining the maturity level of coffee beans, largely due to changes in the surface structure that occur during the roasting process. As a result, the texture features in this study were obtained using the Gray Level Co-occurrence Matrix (GLCM) method, which calculates the spatial distribution of pixel value pairs in a grayscale image.

To mine informative texture features from coffee bean images, a sliding window strategy with a 16×16 pixel block size was adopted in this study. This method splits each grayscale image into tiny, non-overlapping patches so that local texture patterns can be described by the Gray Level Co-occurrence Matrix (GLCM) at the local level. The selection of the 16×16 window represents a practical trade-off: large enough to encode key local structural changes that take place while roasting, and yet small enough to ensure computational efficiency. Window sizes comparable to this have seen extensive application in a range of image classification applications including flame detection, brain tumor diagnosis, and fingerprint enhancement because of their suitability for retaining local context without incurring excessive processing costs (Kumar et al., 2022; Yuzhan et al., 2020)

The computed homogeneity values of all the 16×16 pixel blocks are reconstructed into a GLCM-based texture map. The texture map is subsequently resized to 128 × 128 pixels to be consistent with the dimensions of the RGB image so that both could be utilized in parallel as inputs to the dual-input Convolutional Neural Network (CNN) architecture. Homogeneity was chosen as the only GLCM descriptor owing to its aptness in capturing surface smoothness and local uniformity—textural cues that are most pertinent to the roasting process of coffee beans. As compared to other GLCM features like contrast or energy, homogeneity is less noise-sensitive and has reduced computational complexity, making it amenable to being embedded in lightweight, real-time systems (Azhar & Akbi, 2024; Nugroho et al., 2025; Prabhakar et al., 2024). By augmenting RGB-based color features with GLCM-based homogeneity, this research endeavors to build a richer and more discriminative visual representation that improves classification performance across subtle roasting variations.

Many earlier studies have revealed great promise in coffee bean categorization for both RGB and GLCM characteristics. For instance, a 97.78% accuracy (Hendrawan et al., 2023) was obtained by an ANN-based method combining RGB and GLCM. While a BPNN-based method with GLCM characteristics achieved an accuracy of 97.5% (Nasution & Andayani, 2017), a CNN model with RGB input only recorded an accuracy of up to 96.75% (Marzuki et al., 2025). While an RGB-based CNN model on a separate dataset reported a lower accuracy of 83.3% (Muhlisin et al., 2024), another study using an RGB-based ANN achieved an accuracy of 97.22% (Noel et al., 2019). Previous research related to the classification of coffee bean roasting levels can be seen in Table 1.

Table 1. Research Related to The Classification of Coffee Bean Roasting Level

No	Study	Method	Features Used	Accuracy
1.	(Hendrawan et al., 2023)	ANN	RGB, GLCM	97.78%
2.	(Nasution & Andayani, 2017)	BPNN	RGB	97.5%
3.	(Noel et al., 2019)	ANN	GLCM	97.22%
4.	(Marzuki et al., 2025)	CNN	RGB	96.75%
5.	(Muhlisin et al., 2024)	CNN	RGB	83.3%

2.3 Designing CNN Architecture

Implementing a Convolutional Neural Network (CNN) approach, this paper classifies coffee bean roasting degree based on their visual features. A Convolutional Neural Network (CNN) is a deep learning model that is developed for processing image data based on self-extraction of features from input pixels. This CNN model is described in this paper in two formats: single-input and parallel-input dual-input models.

The single-input model was utilized in order to assess the separate contribution of each category of features individually. Two models were experimented on: the CNN model with RGB input and the CNN model with GLCM map input. The basic design for both models includes two convolutional layers that use a 3 x 3 kernel with a ReLU activation function, followed by a max-pooling layer to reduce the spatial dimensions. Following the feature extraction process, the output of the network is flattened and passed through the dense (fully connected) layer for classification. In order to minimize the risk of overfitting, a dropout layer is added after the flatten layer and prior to the output layer. Designed for two inputs, the model will merge texture features (GLCM) and color features (RGB) at the same time.

Two separate input branches within this architecture individually process RGB and GLCM data using the same convolutional pathways. Within each branch, two convolutional layers, max-pooling, flattening, and dropout are present. The outputs of the two pathways are then merged through a concatenation operation and fed into the dense layer to make a four-class prediction of roasting degrees. The use of two inputs enables the model to learn from complementary multimodal data, in which color aids in assessing the overall degree of maturity, whereas texture contributes additional information regarding physical changes occurring on the surface of the coffee bean. Constructed with the TensorFlow and Keras libraries, the model was constructed with a specific input size of 128 × 128 pixels. Especially in the case of visual data characterized by texture and brightness differences, the parameters on the number of layers, number of filters, and the use of regularization techniques such as dropout and L2 kernel regularization were tuned carefully to balance between accuracy and generalization. The architecture of the Model can be seen in Figure 3.

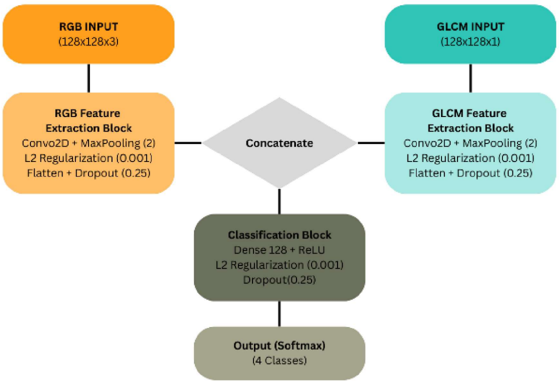


Fig. 3. Proposed Model Architecture

2.4 Conducting Experiments

This research initiates a series of experiments to evaluate the performance of the classification model in classifying the levels of coffee bean roasting depending on the kind of input characteristics utilized. The first three configurations to be studied are: (1) a model with RGB feature input only, (2) a model with GLCM feature input only, and (3) a hybrid model with both RGB and GLCM inputs provided to the system simultaneously through a dual-input CNN architecture.

Comprising 1,600 images, the dataset has 1,200 for training and 400 for testing. As per the input requirement of convolutional models, all images were preprocessed by resizing to 128 × 128 pixels and normalizing pixel values to the range 0 to 1. The training of the models was conducted with the TensorFlow and Keras libraries on the NVIDIA Tesla T4 GPU accelerator in the Google Collaboratory environment. All the models were trained for a total of 50 epochs with a batch size of 16 using the Adam optimizer with its default parameters. Since the classification was among four mutually exclusive classes, the loss function employed was categorical crossentropy.

All experiments used the EarlyStopping method to prevent overfitting and avoid overly long training times. It monitors the validation loss and stops training independently if no improvement is seen over five consecutive epochs. The initial findings of the experiments show that the hybrid RGB and GLCM model using the dual-input CNN architecture had the most consistent and accurate classification performance of the three models. Therefore, testing continued with different modifications to the dual-input model architecture to determine the best setup. The variations that were tested included:

- Adding a dropout layer
- Higher dropout rate in the terminal layer
- L2 regularization use Decreased number of convolutional filters
- Combining L2 regularization with dropout

The evaluation of the model included calculating classification accuracy on the test set and plotting a confusion matrix for analyzing the distribution of class errors. Furthermore, plotting the loss and accuracy curves is a method for ascertaining the convergence and stability of the performance of the model throughout the training process, thereby enabling monitoring of the training process. A summary of the architecture of the six dual-input CNN variations can be seen in Table 2.

Table 2. A summary of the architecture of the six Dual-Input CNN variations

Variation	Key Modification	Notes
1.	Baseline	Standard dual-input CNN
2.	Dropout added to both input branches	Dropout (0.25) in RGB & GLCM
3.	Higher dropout rate	Dropout increased to 0.5 (post-merge)
4.	Reduced filters	Fewer filters, simpler architecture
5.	L2 regularization	L2 applied to all main layers
6.	L2 + Dropout combined	Most regularized configuration

2.5 Evaluating Model Performance

The model performance evaluation in this study was conducted using two main metrics, namely classification accuracy and confusion matrix. Accuracy was chosen because it provides an overview of the model's ability to correctly classify coffee bean roasting levels. Mathematically, classification accuracy is defined as shown in Eq. (1), where it measures the proportion of correctly predicted instances out of the total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Meanwhile, the confusion matrix is used to identify patterns of prediction error between classes, especially in the case of visually similar roasting levels such as medium and dark. In a systematic review by (Anto et al., 2023), accuracy was reportedly used in more than 58% of the visual classification studies analyzed, confirming its relevance in this context.

However, in the fine-grained visual classification literature, some studies mentioned that the fine differences between classes may make this metric less sensitive to minor errors that have a significant impact on model interpretation. However, accuracy and confusion matrix are still considered representative in this study because the dataset is balanced and each class has an equal number of classes. The use of confusion matrix also helps in observing the tendency of model errors in certain classes, as discussed in the study by (Krstinić et al., 2024).

The initial approach in this study tested three types of models: RGB-based only, GLCM-based only, and RGB-GLCM combined (Hybrid). The training results showed that the hybrid model consistently provided higher validation accuracy and more stable classification performance than the other two approaches. This finding is in line with previous research which shows that the combination of color and texture features can improve classification performance because it produces a richer visual representation (Jinguang et al., 2009; Liu et al., 2007).

Based on the preliminary results showing the advantages of the hybrid approach, six variations of CNN architecture were explored on the hybrid model with modifications such as the addition of dropouts, L2 regularization, simplification of the number of filters, and a combination of both. The model without additional regulation showed relatively high loss fluctuations and overfitting tendencies. In contrast, the sixth model-which combined dropout and L2 regularization-showed high validation accuracy with the most stable and convergent loss, and the least amount of misclassification, so it can be considered as the most optimal performing model in this experiment.

Evaluation of the training-validation curve is used to observe the stability of the training and convergence process of the model. When the training and validation curves show a uniform and stable pattern, it can be concluded that the model is not overfitting, and the training is optimal. The study by (Tang et al., 2025) showed that training stability can be improved through a dynamic training strategy based on Lyapunov analysis, which adaptively adjusts the learning rate to maintain performance and avoid parameter oscillations. Thus, training curve analysis is not only useful for monitoring, but also an important part of evaluating the performance of reliable models.

III. RESULTS AND DISCUSSION

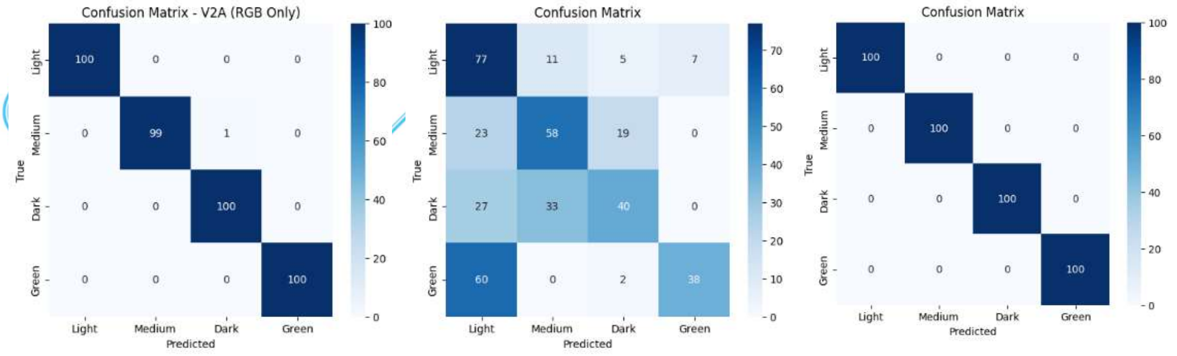
This chapter marks the outcome of the experiments conducted, along with an in-depth analysis of the results achieved. The analysis compares the classification performance in convolutional neural network (CNN) models for the classification of coffee beans' roasting levels based on three input approaches: sole RGB, sole GLCM, and a hybrid approach of both RGB and GLCM features. All of the models were analyzed on the basis of two main parameters: classification accuracy and the confusion matrix, as outlined in the Evaluation Matrix section. These measures were selected on the grounds of interpretability and ease of use, more particularly to a balanced data set with visually separable but subtly changing classes.

The findings are separated into two major phases: (1) initial assessment via RGB, GLCM, and hybrid model comparison for identifying the best feature representation; and (2) extensive assessment of six CNN architecture variants employed under the hybrid model to scrutinize the performance of varied regularization methods. Graphic representations of confusion matrices along with training-validation graphs are included to complement the analysis and pinpoint the classification character and training integrity of each model variant.

3.1 Comparison Results of RGB, GLCM, and Hybrid

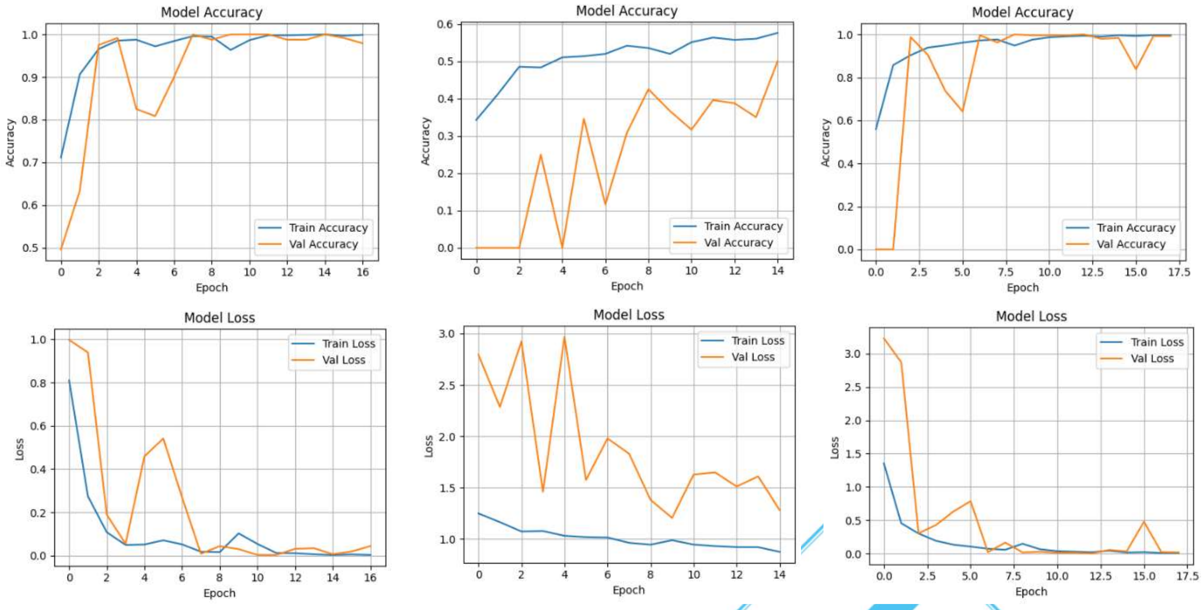
To evaluate the effectiveness of feature representation in coffee bean roasting level classification, three initial models were developed using different input approaches: an RGB image-based model only, a model with GLCM-based texture features only, and a hybrid model combining both. All models were trained using a uniform CNN architecture and parameters to ensure a fair comparison.

The confusion matrix visualization of each model is presented in Figure 4, while Figure 5 shows the training and validation curves for accuracy and loss. The evaluation results show that the RGB-only based model (V2A) has a very high validation accuracy of 0.9974, with a near-perfect confusion matrix-only one misclassification in the medium class. The GLCM-only model recorded a much lower validation accuracy of 0.6342, with prediction errors spread across all classes. Meanwhile, the Hybrid (RGB + GLCM) model performed the best with a validation accuracy of 1.0000, with no misclassification at all.



(a) (b) (c)

Fig. 4. Visualization of confusion matrix: (a) RGB Only, (b) GLCM Only, (c) Hybrid



(a) (b) (c)

Fig. 5. Training and validation curves for accuracy and loss: (a) RGB Only, (b) GLCM Only, (c) Hybrid

The training-validation curves in Figure 4 support this finding: the hybrid model not only achieves the highest accuracy, but also shows a stable and convergent training process, with low validation loss and minimal fluctuation. In contrast, the GLCM model shows unstable accuracy and loss curves, reflecting the difficulty in generalization. A summary of the accuracy metrics and stability analysis of the three models is presented in Table 3.

Table 3. Summary of Accuracy, Stability, and Observations for Initial Models

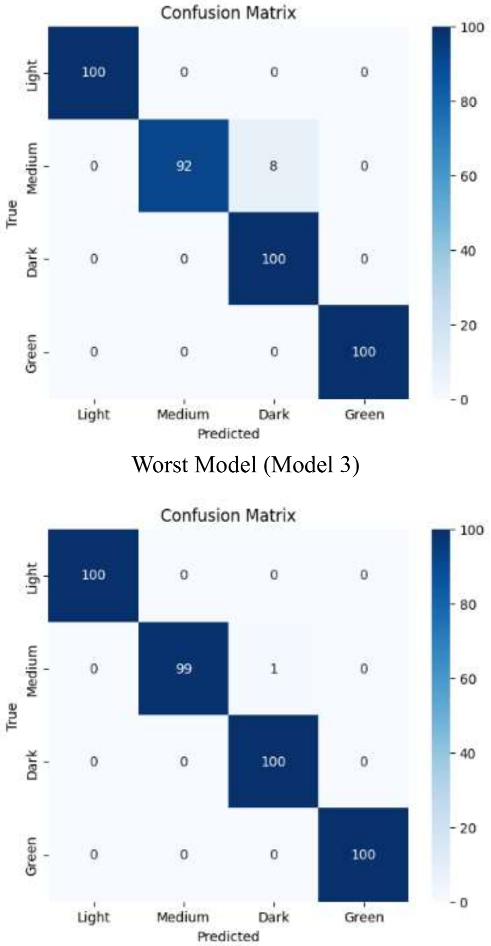
Model	Accuracy / Loss	Remarks
RGB Only	0.9974 / 0.0062	Stable; minor error in "Medium" class
GLCM Only	0.6342 / 0.8770	Unstable; frequent misclassification
Hybrid (RGB+GLCM)	**1.0000 / 0.0096**	**Exceptionally stable; perfect classification**

3.2 Performance Evaluation of Six Variations of Hybrid Architecture

After the demonstration that the hybrid method achieved better accuracy than the RGB or GLCM models individually, this study went on further with an analysis of six various versions of the hybrid-based CNN architecture (RGB + GLCM). Each version was obtained with specific architectural modifications, including the addition of dropout layers, use of L2 regularization, decreasing the number of filters, and the fusion of both regularization methods. The aim of these variations is to examine the influence of regularization methods on classification performance, stability of training, and model generalization.

The results of the experiments show that all the models have better performance, but Model 6, which combines dropout and L2 regularization, shows the most balanced performance in evaluations. Despite Model 2 having the highest numerical accuracy and the least loss, Model 6 can still hold a very high validation accuracy of 0.9974 at a comparatively low loss of 0.1221 with two regularization techniques combined. This characteristic qualifies Model 6 as the most stable and trustworthy model in relation to generalization without any compromise on the classification accuracy. In contrast, Model 3 with the highest dropout performed the worst (accuracy 0.9808) with the highest number of misclassifications, especially in the medium class.

To emphasize the performance contrast between variations, Figure 6 and Figure 7 only show the confusion matrix and training-validation curves of Model 6 (best) and Model 3 (worst). A full summary of the accuracy, loss, and observations for each variation is shown in Table 4.



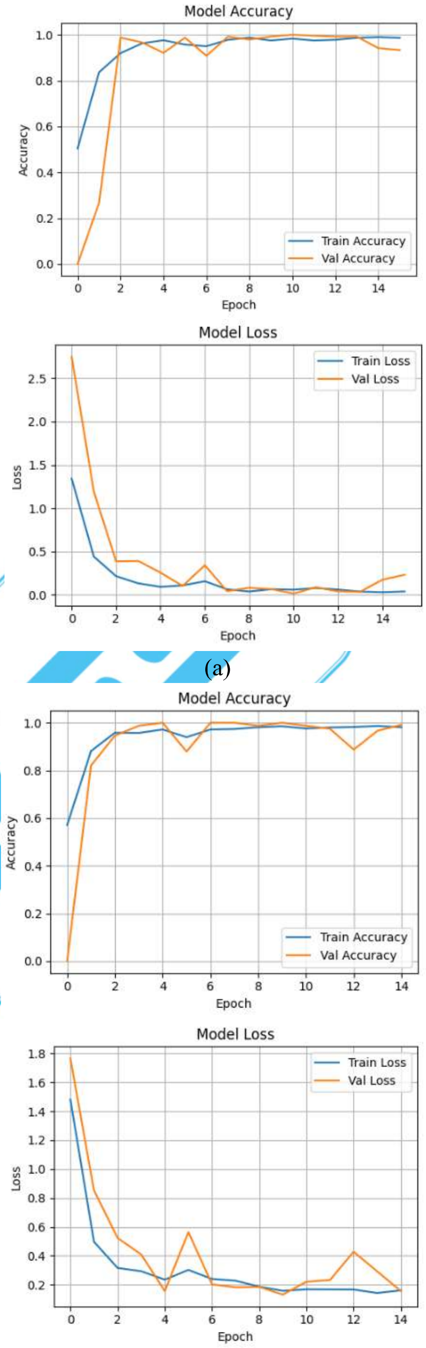
Worst Model (Model 3)

Best Model (Model 6)

Figure 6. Confusion matrices

Table 4. Performance Summary of Hybrid CNN Model Variations

Variation	Accuracy / Loss	Remarks
1	1.0000 / 0.0096	Perfect accuracy; no regularization applied
2	1.0000 / 0.0045	Excellent convergence; minimal loss
3	0.9808 / 0.0466	Most unstable; highest misclassification rate
4	0.9995 / 0.0191	Slight loss increase; still performs reliably
5	0.9953 / 0.1073	Tendency toward overfitting in later epochs
6	**0.9974 / 0.1221**	**Most balanced; best generalization and stability**



(a)

(b)

Figure 7. Confusion matrices: (a) Worst Model (Model 3), (b) Best Model (Model 6).

3.3 Comparison with Previous Research

The results from this study show that the CNN approach with hybrid RGB and GLCM inputs yields the highest validation accuracy of 99.74% in the best variation (Model 6), surpassing the accuracy reported in previous studies. In Table 5, studies with ANN methods combining RGB and GLCM recorded an accuracy of 97.78%, while other approaches using only RGB or GLCM recorded accuracies ranging from 83.3% to 97.5%. This indicates that combining color and texture features through a dual-input CNN approach can significantly improve classification quality.

In addition, this study also makes a novel contribution by conducting an in-depth evaluation of six hybrid CNN architecture variations, not only in terms of accuracy, but also training stability and loss convergence patterns. This approach has not been found in previous studies which generally only focus on the comparison of methods or features used. By using combined regularization (dropout and L2) in Model 6, this study successfully demonstrates that the model is not only accurate, but also stable and generalizable-important qualities in the visual classification of details such as the roasting degree of coffee beans.

3.4 Visual Interpretability using Class Activation Map (CAM)

To gain additional insights into the reasoning behind the decision-making of the hybrid CNN model, we applied Class Activation Mapping (CAM) to visualize each input image's most salient contribution to the model's decision-making process. This method of providing interpretability is critical for both researchers and practitioners in algorithms, such as when using a hybrid model like this for fine-grained classification tasks such as recognizing specific coffee roasting levels, where subtle differences in color or texture allow one class to be distinguished from another.

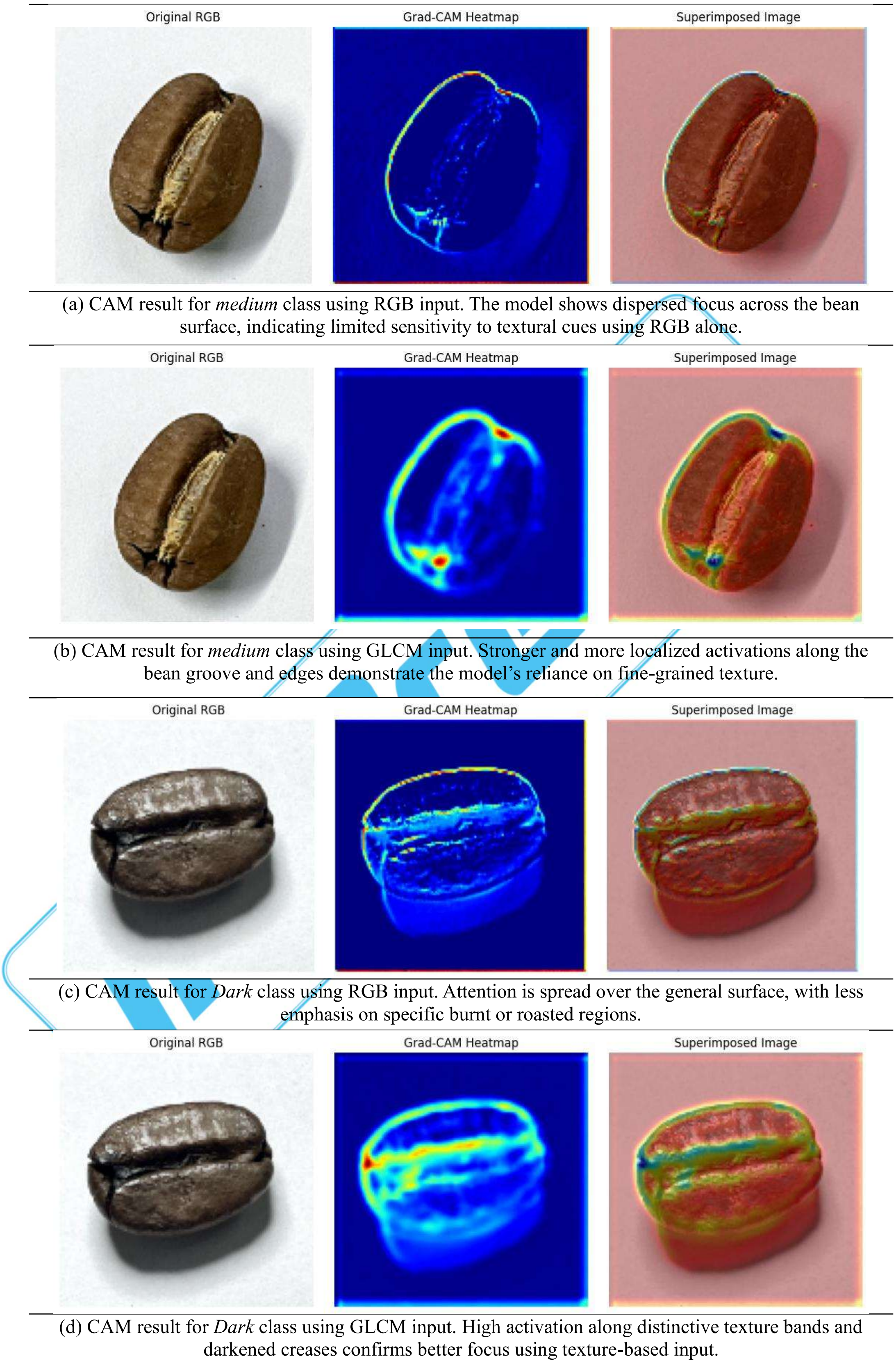
In Figure 8, we present partner CAM visualizations for two representative classes of coffee, Medium, and Dark for both RGB and GLCM input branches of our hybrid model. The visualizations indicate that the model produces attention maps that focus in ways representative of the most salient differences. Model activations almost always correspond to the areas of surface texture and the central groove of coffee beans. For GLCM, activation was more localized to discrete texture features of the beans, and for RGB it was more general to the color gradient of the beans, and structural edges of the beans. These visualizations helped confirm that the dual-branched architecture is adequately tapping complementary information from each input.

IV. CONCLUSION

This study demonstrated that a hybrid Convolutional Neural Network (CNN) model incorporating both RGB color features and GLCM-based texture features offers superior performance in classifying coffee bean roasting levels. Compared to RGB-only and GLCM-only models, the hybrid approach consistently achieved the highest classification accuracy, with the most stable training and convergence behavior. Among the six architectural variations tested, the model combining dropout and L2 regularization (Model 6) achieved the best balance between accuracy (99.74%), loss stability, and generalization, making it the most optimal configuration for fine-grained visual classification in this domain.

These findings highlight the significance of multimodal feature fusion—especially in applications where visual differences between classes are subtle yet critical, such as coffee quality assessment. The use of training-validation curves and confusion matrices also proved essential in identifying the most reliable and interpretable model. Future research could explore other color spaces, advanced texture descriptors, and transfer learning techniques to further improve model robustness and scalability across diverse datasets and real-world imaging conditions.

In spite of the encouraging findings, there are some limitations in this study that must be noted. Firstly, just one specific GLCM descriptor—homogeneity—was utilized for texture extraction, which might restrict the richness of texture representation. Secondly, the model depends on RGB images taken in controlled lighting environments, and how it performs with changing real-world conditions has not been tested. Lastly, although the hybrid model performed extremely well in terms of validation accuracy, the fact that no runtime or deployment tests were conducted means that overfitting cannot be totally discounted. In follow-up work, these gaps need to be filled through the inclusion of more texture features, the investigation of other color spaces (e.g., HSV, Lab), and testing the model in real-time scenarios to determine its robustness and scalability.



(a) CAM result for *medium* class using RGB input. The model shows dispersed focus across the bean surface, indicating limited sensitivity to textural cues using RGB alone.

(b) CAM result for *medium* class using GLCM input. Stronger and more localized activations along the bean groove and edges demonstrate the model's reliance on fine-grained texture.

(c) CAM result for *Dark* class using RGB input. Attention is spread over the general surface, with less emphasis on specific burnt or roasted regions.

(d) CAM result for *Dark* class using GLCM input. High activation along distinctive texture bands and darkened creases confirms better focus using texture-based input.

Figure 8. CAM visualizations of coffee roasting level

REFERENCES

Anhar, A., Rasyid, U. H. A., Muslih, A. M., Baihaqi, A., Romano, & Abubakar, Y. (2021). Sustainable Arabica coffee development strategies in Aceh, Indonesia. *IOP Conference Series: Earth and Environmental Science*, *667*(1), 012106. https://doi.org/10.1088/1755-1315/667/1/012106

Anto, I. A. F., Munandar, A., Wibowo, J. W., Salim, T. I., & Mahendra, O. (2023). Coffee Bean Roasting Levels Detection: A Systematic Review. *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 146–151. https://doi.org/10.1109/ICITISEE58992.2023.10404775

Ashardiono, F., & Trihartono, A. (2024). Optimizing the potential of Indonesian coffee: a dual market approach. *Cogent Social Sciences*, *10*(1). https://doi.org/10.1080/23311886.2024.2340206

Azhar, Y., & Akbi, D. R. (2024). Performance Comparison of GLCM Features and Preprocessing Effect on Batik Image Retrieval. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1339. https://doi.org/10.62527/joiv.8.3.2179

Baqueta, M. R., Coqueiro, A., Março, P. H., Valderrama, P., & Driscoll, S. (2020). *Arabica coffee evaluation concerning the degree of roasting: An approach by using smartphone and projection pursuit* (D. Kalschne, M. Corso, & R. Dias, Eds.). Nova Science Publishers, Inc.

Castillo, A. M., Aradanas, R. D., Arboleda, E. R., Dizon, A. A., & Dellosa, R. M. (2019). Coffee Type Classification Using Gray Level Co-Occurrence Matrix Feature Extraction And The Artificial Neural Network Classifier. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *8*(10). www.ijstr.org

Claire, A., Ardison, I., Josheba, M., Hutagalung, R., Chernando, R., & Cenggoro, T. W. (2022). Observing Pre-Trained Convolutional Neural Network (CNN) Layers as Feature Extractor for Detecting Bias in Image Classification Data. In *CommIT Journal* (Vol. 16, Issue 2).

Dawwd, S. (2019). GLCM Based Parallel Texture Segmentation using A Multicore Processor. In *The International Arab Journal of Information Technology* (Vol. 16, Issue 1).

Domingues, L. O. C., Garcia, A. de O., Ferreira, M. M. C., & Morgano, M. A. (2020). Sensory quality prediction of coffee assessed by physicochemical parameters and multivariate model. *Coffee Science*, *15*(1), 1–11. https://doi.org/10.25186/cs.V15I.1654

Freitas, V. V., Borges, L. L. R., Vidigal, M. C. T. R., dos Santos, M. H., & Stringheta, P. C. (2024). Coffee: A comprehensive overview of origin, market, and the quality process. *Trends in Food Science & Technology*, *146*, 104411. https://doi.org/10.1016/j.tifs.2024.104411

Hendrawan, Y., Ramadhan, T. F., Sutan, S. M., Al-Riza, D. F., Damayanti, R., & Hermanto, M. B. (2023). *Identification of Arabica coffee roasting levels using Nikon D3100 commercial camera and optimized neural network*. 040002. https://doi.org/10.1063/5.0166769

Jinguang, S., Zhipeng, W., & Da, Y. (2009). Research of Image Retrieval Based on Uniting Features. *2009 International Forum on Information Technology and Applications*, 603–607. https://doi.org/10.1109/IFITA.2009.318

Krstinić, D., Skelin, A. K., Slapničar, I., & Braović, M. (2024). Multi-Label Confusion Tensor. *IEEE Access*, *12*, 9860–9870. https://doi.org/10.1109/ACCESS.2024.3353050

Kumar, K. S. V., Maheen, A., & Devulapalli, P. K. (2022). Classification of Brain Tumours from The MR Images Using Neural Network and Central Moments. *2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, 1–4. https://doi.org/10.1109/ICAITPR51569.2022.9844188

Liu, P., Jia, K., & Wang, Z. (2007). An Effective Image Retrieval Method Based on Color and Texture Combined Features. *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, 169–172. https://doi.org/10.1109/IIH-MSP.2007.79

Marzuki, K., Apriani, & Qulub, M. (2025). Coffee Roaster Maturity Level Classification Based on Convolutional Neural Network. *Mathematical Modelling of Engineering Problems*, *12*(1), 46–54. https://doi.org/10.18280/mmep.120106

Muhlisin, E. I., Nurmalasari, R. R., Kamelia, L., & Sururie, R. W. (2024). Implementation Of Convolutional Neural Network (CNN) in The Android-Based Application for Detecting Coffee Bean Maturity. *2024 10th International Conference on Wireless and Telematics (ICWT)*, 1–5. https://doi.org/10.1109/ICWT62080.2024.10674676

Nasution, T. H., & Andayani, U. (2017). Recognition of Roasted Coffee Bean Levels using Image Processing and Neural Network. *IOP Conference Series: Materials Science and Engineering*, *180*, 012059. https://doi.org/10.1088/1757-899X/180/1/012059

Ngure, G. M., & Watanabe, K. N. (2024). Coffee sustainability: leveraging collaborative breeding for variety improvement. *Frontiers in Sustainable Food Systems*, *8*. https://doi.org/10.3389/fsufs.2024.1431849

Noel, J., Sarino, C., Bayas, M. M., Arboleda, E. R., Guevarra, E. C., & Dellosa, R. M. (2019). Classification Of Coffee Bean Degree Of Roast Using Image Processing And Neural Network. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *8*(10). www.ijstr.org

Nugroho, H., Pramudito, W. A., & Laksono, H. S. (2025). Gray Level Co-Occurrence Matrix (GLCM)-based Feature Extraction for Rice Leaf Diseases Classification. *Buletin Ilmiah Sarjana Teknik Elektro*, *6*(4), 392–400. https://doi.org/10.12928/biste.v6i4.9286

Prabhakar, A., Jena, P., & Pati, U. C. (2024). A New Framework for Brain Tumor Feature Extraction and Classification Using Localized Global Feature Patches. *2024 International Conference on Smart Systems for Applications in Electrical Sciences (ICSSES)*, 1–6. https://doi.org/10.1109/ICSSES62373.2024.10561327

Pratama, Y., Dirgayussa, I. G. E., Simarmata, P. F., & Tambunan, M. H. (2021). Detection roasting level of Lintong coffee beans by using euclidean distance. *Bulletin of Electrical Engineering and Informatics*, *10*(6), 3072–3082. https://doi.org/10.11591/eei.v10i6.3153

Qur, A., Harsani, P., Ayu Wulandhari, L., & Agung Santoso Gunawan, A. (2020). Color Extraction and Edge Detection of Nutrient Deficiencies in Cucumber Leaves Using Artificial Neural Networks. In *Communication & Information Technology) Journal* (Vol. 14, Issue 1).

Samper, L., & Quiñones-Ruiz, X. (2017). Towards a Balanced Sustainability Vision for the Coffee Industry. *Resources*, *6*(2), 17. https://doi.org/10.3390/resources6020017

Sutarmin, Mukroji, Rastuti, U., Yunanto, A., Suliyanto, & Jatmiko, D. P. (2022). Increasing the Additional Value of Coffee Cultivation Results in Brebes Regency with a Value Chain Analysis Approach. *Quality - Access to Success*, *23*(188). https://doi.org/10.47750/QAS/23.188.13

Tang, D., Yang, N., Deng, Y., Zhang, Y., Sani, A. S., & Yuan, D. (2025). *Stability-Driven CNN Training with Lyapunov-Based Dynamic Learning Rate* (pp. 58–70). https://doi.org/10.1007/978-981-96-1242-0_5

Wei, F., & Tanokura, M. (2015). Chemical Changes in the Components of Coffee Beans during Roasting. In *Coffee in Health and Disease Prevention* (pp. 83–91). Elsevier. https://doi.org/10.1016/B978-0-12-409517-5.00010-3

Yuzhan, M., Abdullah Jalab, H., Hassan, W., Fan, D., & Minjin, M. (2020). Recaptured Image Forensics Based on Image Illumination and Texture Features. *2020 The 4th International Conference on Video and Image Processing*, 93–97. https://doi.org/10.1145/3447450.3447465