PROCEEDINGS OF

INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND NEW APPLICATIONS

https://proceedings.icisna.org/

3rd International Conference on Intelligent Systems and New Applications (ICISNA'25), Antalya, December 12-14, 2025.

An Explainable Deep Learning Framework for Agtron-Based Coffee Roast Classification Using Grad-CAM

Havva Hazel ARAS[1], Yusuf ERYESIL[2], Murat KOKLU[2]

[1] Yozgat Vocational School, Yozgat Bozok University, Yozgat, Türkiye
h.hazel.aras@bozok.edu.tr, ORCID: 0000-0002-4179-3188

[2] Department of Computer Engineering, Technology Faculty, Selcuk University, Konya, Türkiye
yusuf.eryesil@selcuk.edu.tr, ORCID: 0000-0001-8735-3666

mkoklu@selcuk.edu.tr, ORCID: 0000-0002-2737-2360

Abstract— **Precise control of the roasting process is a critical determinant of coffee quality, as it governs the chemical transformations that define aroma and flavor profiles. However, traditional quality assessment methods typically rely on subjective manual inspection or expensive colorimetric devices, which are often prone to inconsistency or limited by high operational costs. To address these challenges, this study proposes a robust, automated computer vision framework for fine-grained coffee roast classification based on the Agtron color scale. We utilized a dataset comprising five distinct roast levels (Green, Light, Medium, Dark, and Overbaking) to evaluate the performance of state-of-the-art Convolutional Neural Network architectures, including VGG16, ResNet50, DenseNet201, MobileNetV2, InceptionV3 and Xception. To ensure statistical reliability, all models were trained and tested using a 5-fold cross-validation strategy. Experimental results demonstrated that DenseNet201 achieved superior performance, recording a classification accuracy of 99.84% and an F1-score of 0.9984, outperforming other architectures in both stability and precision. Furthermore, to validate the model's reliability, we employed Gradient-weighted Class Activation Mapping, which visually confirmed that the network focuses on discriminative bean features, such as surface texture and oil expression, rather than background artifacts. These findings indicate that deep learning-based visual inspection can serve as a highly accurate, non-destructive, and cost-effective solution for real-time quality control in the coffee industry.**

Keywords— **Coffee Roast Classification, Deep Learning, DenseNet201, Grad-CAM, Quality Control**

I. INTRODUCTION

The global coffee industry constitutes a multibillion dollar ecosystem that spans agricultural production, processing, and consumer markets, and provides livelihoods for millions worldwide. Among the critical stages in this value chain, the roasting process plays a transformative role by irreversibly altering the physical and chemical structure of green coffee beans from species such as Coffea arabica and Coffea canephora [1]. Far beyond simple heating, roasting involves complex thermodynamic reactions, including the Maillard reaction, caramelization, and pyrolysis, which generate thousands of aromatic compounds that shape the sensory quality of coffee. As the specialty coffee sector continues its rapid expansion, driven by increasingly sophisticated consumer expectations, consistency and precision in controlling the degree of roast have become more essential than ever [2].

Advances in Industry 4.0 and smart agriculture have accelerated the adoption of Computer Vision techniques across food processing pipelines. Traditional image-processing approaches based on handcrafted features (e.g., color histograms or texture descriptors) have proven inadequate for modeling the natural variability of coffee beans and the challenges introduced by uncontrolled lighting [3]. Convolutional Neural Networks (CNNs), with their hierarchical feature-learning capabilities, have replaced these earlier methods by learning both low-level visual cues and high-level semantic representations directly from pixel data. Initially applied to defect detection (e.g., insect damage, broken beans, or immature beans), CNNs have more recently been utilized for finer-grained tasks such as roast-level classification [4].

Over the last two decades, research on coffee quality assessment has evolved from simple colorimetric measurements to sophisticated deep learning frameworks. Early studies combined color, morphology, and texture information with traditional machine learning algorithms. For instance, Faridah et al. employed combined texture and RGB-based descriptors to train Artificial Neural Networks, demonstrating the feasibility of digital imaging as an alternative to chemical analysis. Similarly, Turi et al. integrated morphological, color, and texture information to characterize coffee varieties from Ethiopia [5]. Although these studies established the potential of image-based coffee quality assessment, their reliance on manual feature extraction limited robustness under varying illumination or heterogeneous datasets. Earlier chemical analyses by Mazzafera and Ximenes explored the relationship between bean defects and visual indicators, yet these findings could not be easily integrated into automated systems at the time [6, 7].

The emergence of deep learning marked a significant shift in coffee-bean research. Pinto et al. (2016) introduced one of the most influential datasets, which contains over 6,500 beans and 13,000 images, and they applied a CNN model to classify six defect types. Their work demonstrated the necessity of capturing local spatial features rather than relying on global color metrics, revealing the strength of CNNs in detecting subtle textural irregularities [8]. Building on this perspective, Alamanda, Susanto, and Lestari proposed a two-stage pipeline that combines U Net based segmentation with a modified ResNet 50 classifier for post-roast bean analysis. Their method achieved a Dice score of 0.9375 in segmentation and 86% accuracy across six roast levels, although performance degraded in intermediate roast categories because of overlapping visual characteristics. This limitation highlights the difficulty of distinguishing fine-grained roast levels, particularly within narrow Agtron intervals [9].

In parallel, Rivas, Bertarini, and Fernandes explored feature extraction using deep and traditional models, comparing Xception, AdaBoost, Random Forest, and SVM on balanced datasets containing four roast levels. Their experiments reported perfect (100%) accuracy and F1-score for Xception-based feature extraction, attributed to the model's depthwise separable convolutions capturing nuanced texture variations. However, their coarse roast categories (green/light/medium/dark) raise questions regarding generalizability to more granular scales such as Agtron [10].

Roast-level classification presents challenges distinct from defect detection: instead of identifying discrete morphological abnormalities, the task requires differentiating subtle, continuous changes in bean color, oil expression, and surface texture [11]. In this study, we propose a deep learning based methodology to classify five roast levels (Dark, Green, Light, Medium, Overbaking) using an Agtron-based dataset from Kaggle. Six backbone architectures (VGG16, ResNet50, DenseNet201, MobileNetV2, InceptionV3 and Xception) are comparatively evaluated using 5-fold cross-validation to enhance generalization and reduce bias. Furthermore, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the discriminative regions influencing the model's predictions. The goal of this work is to provide an automated, Agtron-standardized roast-level classification framework that supports the digital transformation of quality control processes in the coffee industry.

II. Material and Methods

The overall methodological framework proposed in this study for the automated classification of coffee roast levels is illustrated in Figure 1. This workflow encompasses the sequential stages of data acquisition, preprocessing, data augmentation, transfer learning using CNN, and performance evaluation.

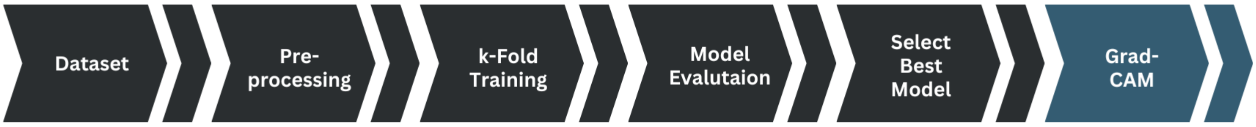


Fig. 1 The general block diagram of the proposed deep learning-based coffee roast classification framework.

A. Dataset

In this study, we utilized the publicly available Coffee Roast–Agtron Scale Dataset hosted on Kaggle, which contains images of roasted coffee beans annotated with Agtron color values. The dataset comprises approximately 2,500 image files, corresponding to roughly 15 GB of data, and was collected specifically for roast-level analysis. The data is categorized into five distinct classes based on roasting and processing status: Green, Light, Medium, Dark, and Overbaking. A balanced distribution was ensured across each class, and this equilibrium was meticulously maintained during both the training and testing phases [12].

Each sample consists of whole coffee beans arranged on a flat background and photographed after roasting to a target Agtron value. In the original dataset, images were captured with different camera devices, including consumer mobile phones and a Canon EOS R50 camera, under controlled indoor lighting. For each capture session, a reference photograph of a blank white sheet is also provided to facilitate illumination normalization and color calibration across devices and roasting batches [12].

B. Deep Learning Models

VGG16 was selected as a baseline architecture due to its deep yet straightforward stack of standard 3×3 convolutions and max-pooling layers, which have historically provided strong performance in image classification tasks [13, 14]. Its uniform architecture enables controlled comparison with more advanced models and serves as a reference point for evaluating feature-learning improvements in later architectures [15].

ResNet50 introduces residual connections, a mechanism that allows gradients to propagate through identity mappings without attenuation [16]. This design effectively mitigates the vanishing-gradient problem observed in deep CNNs and enables the training of substantially deeper networks without degradation in accuracy. Owing to its stability and strong representational power, ResNet50 was included as a robust mid-level architecture [17].

DenseNet201 extends the idea of connectivity by establishing direct feed-forward links from each layer to all subsequent layers [18]. This feature-reuse strategy reduces redundant computations, lowers the total number of parameters, and encourages richer gradient flow [19]. DenseNet models have demonstrated excellent performance in fine-grained visual tasks, making DenseNet201 an appropriate choice for capturing subtle texture and color variations in roasted coffee beans [20].

MobileNetV2 employs depthwise separable convolutions and an inverted residual structure with linear bottlenecks, enabling high representational efficiency while significantly reducing computational cost [21]. This lightweight yet expressive design makes MobileNetV2 particularly suitable for resource-constrained environments such as mobile and embedded systems [22]. Its efficient architecture provides a complementary baseline to heavier convolutional networks, offering insights into performance–complexity trade-offs within the model comparison framework [23].

InceptionV3 incorporates factorized convolutions, asymmetric kernels, and multi-branch processing modules that capture spatial information at multiple receptive-field scales simultaneously [24]. These architectural innovations substantially improve computational efficiency while preserving high representational capacity [25]. InceptionV3's ability to extract both coarse and fine-grained features makes it a strong candidate for complex visual recognition tasks, justifying its inclusion as a diverse architectural alternative within the comparison set [26].

Xception extends the Inception paradigm by fully replacing standard convolutions with depthwise separable convolutions, thereby decoupling spatial and channel-wise feature extraction [27]. This "extreme" version of Inception increases model efficiency and expressiveness, enabling the network to learn richer feature representations with fewer parameters [28]. Due to its strong performance in various image-classification benchmarks and its elegant structural simplicity, Xception was selected as a high-performing architecture emphasizing efficient feature disentanglement [29].

C. *K-Fold Cross-Validation*

K Fold Cross Validation is a widely adopted resampling strategy used to obtain a reliable estimate of a model's generalization performance [30]. In this approach, the dataset is partitioned into K equally sized subsets, and during each iteration, one subset is designated as the validation set while the remaining subsets are used for training [31]. Through this rotation, every sample in the dataset is evaluated at least once during validation, enabling a comprehensive and statistically robust assessment of model performance [32]. A key advantage of K Fold Cross Validation is its ability to reduce the bias and variance associated with a single random division of the data into training and validation sets. In image classification tasks, variations in class distribution, lighting conditions, and intra class diversity can cause a model to perform disproportionately well or poorly when evaluated on a single subset of the data. By evaluating the model across multiple partitions, K Fold provides a more stable estimate and ensures that performance is not overly dependent on any particular way of dividing the dataset [33].

In the context of roast level classification, the visual differences between classes are subtle and often exist along a continuous spectrum defined by the Agtron scale [34]. As such, it is essential to assess whether the model can consistently discriminate between classes that differ in very small visual details across different subsets of the data. For this reason, K Fold Cross Validation was employed to ensure that the reported results reflect robust generalization rather than behavior specific to a particular data division [30].

D. *Grad-CAM*

To enhance the interpretability of the deep learning models used in this study, Gradient-Weighted Class Activation Mapping (Grad-CAM) was employed [35]. Grad-CAM is an explainable AI technique that generates localization heatmaps by leveraging the gradients of a target class with respect to the activations of the final convolutional layers [36]. These heatmaps highlight the spatial regions within an input image that most strongly influence the model's prediction. In the context of roasted coffee–bean classification, interpretability is essential because the visual differences between roast levels—such as slight variations in color saturation, texture smoothness, or surface oil expression—are often subtle and may not be immediately distinguishable to the human eye [37]. Grad-CAM enables qualitative assessment of whether the model focuses on relevant visual cues, such as the bean surface, tonal transitions, and texture patterns, rather than irrelevant background regions [38].

Applying Grad-CAM to the predictions of each backbone architecture allows us to verify that the models make decisions based on semantically meaningful regions. This not only supports the reliability of the classification results but also provides insights into the discriminative features associated with each roast level [39]. Furthermore, Grad-CAM serves as an important diagnostic tool for identifying misclassifications and analyzing failure cases, offering a clearer understanding of model behavior beyond numerical accuracy metrics [40].

III. Experimental Results

All experiments were carried out in the Google Colab environment equipped with a high-performance NVIDIA A100 GPU, which substantially accelerated the training process. The deep learning models were implemented using Python 3.9 and the TensorFlow Keras framework. To ensure compatibility with the pre trained backbone architectures VGG16, ResNet50, DenseNet201, MobileNetV2, InceptionV3 and Xception all input images were resized to 224 x 224 pixels using bicubic interpolation. Preprocessing steps specific to each architecture were applied through the corresponding Keras preprocessing utilities. Given the importance of model generalization in roast level classification, a data augmentation pipeline was integrated into the training phase to increase sample variability and mitigate overfitting. The applied transformations included random rotations, stochastic horizontal flips and brightness and contrast adjustments. A summary of these augmentation operations is provided in Table 1.

TABLE 1

SUMMARY OF APPLIED DATA AUGMENTATION OPERATIONS

Augmentation Type	Description
Random Rotation	Rotation of images by fixed angles
Random Horizontal Flip	Stochastic flipping along the horizontal axis
Brightness and Contrast Adjustment	Random perturbations to simulate lighting variations

The training procedure was governed by a set of hyperparameters and optimization strategies designed to improve stability and convergence. The Adam optimizer with an initial learning rate of 1 x 10^{-4} was employed, while classification error across the five roast levels was measured using the Sparse Categorical Cross Entropy loss function. All models were trained with a batch size of 16 for a maximum of 30 epochs.

Training efficiency and robustness were further enhanced through the integration of an Early Stopping mechanism, which monitored the validation accuracy and terminated training if no improvement was observed for five consecutive epochs. In addition, a ReduceLROnPlateau scheduler dynamically reduced the learning rate by a factor of 0.5 when validation accuracy stagnated over three consecutive epochs.

The classification performance of the six deep learning models was evaluated using 5-fold cross-validation. Table 2 summarizes the average Accuracy, F1-Score, Area Under the Curve (AUC), and training time per fold for each architecture. The results demonstrate that deep learning models can distinguish between fine-grained roast levels with exceptional precision.

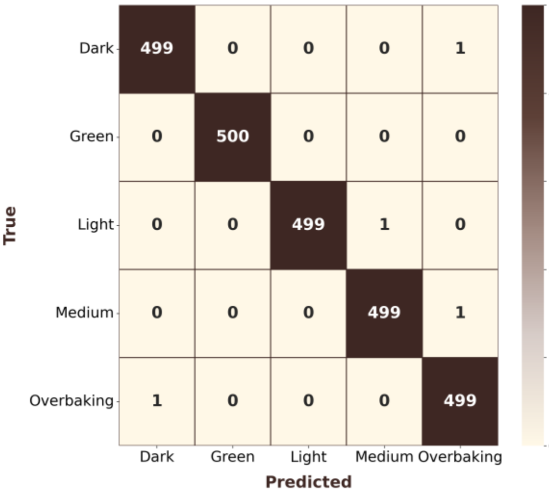
TABLE 2

SUMMARY OF 5-FOLD CROSS-VALIDATION RESULTS

Model	Accuracy (%)	F1-Score (%)	AUC (%)	Training Time (s)
DenseNet201	99.84 ± 0.22	0.9984 ± 0.0022	0.9999 ± 0.0000	322 ± 36
ResNet50	99.64 ± 0.50	0.9963 ± 0.0052	0.9999 ± 0.0001	331 ± 82
Xception	99.36 ± 0.46	0.9935 ± 0.0046	0.9999 ± 0.0001	411 ± 101
VGG16	98.92 ± 0.72	0.9894 ± 0.0070	0.9998 ± 0.0002	614 ± 188
MobileNetV2	98.64 ± 0.97	0.9859 ± 0.0103	0.9997 ± 0.0003	411 ± 131
InceptionV3	98.48 ± 0.82	0.9849 ± 0.0080	0.9997 ± 0.0003	442 ± 110

Among the evaluated architectures, DenseNet201 achieved the state-of-the-art performance, recording the highest average accuracy of 99.84% and an F1-score of 0.9984. Its low standard deviation (±0.22%) across folds indicates superior stability and generalization capability compared to other models. This performance can be attributed to the feature reuse mechanism in the DenseNet architecture, which effectively captures subtle textural variations in coffee beans without suffering from the vanishing gradient problem. ResNet50 also demonstrated highly competitive results with 99.64% accuracy, confirming that residual connections are effective for this task. While VGG16, MobileNetV2, and InceptionV3 performed slightly lower, all models surpassed the 98% accuracy threshold, validating the robustness of the proposed deep learning framework for roast-level classification.In terms of computational efficiency, DenseNet201 was unexpectedly the most efficient, requiring approximately 322 seconds per fold for training. Conversely, VGG16 was the most computationally expensive model (614 seconds), primarily due to its large number of parameters in the fully connected layers. This finding suggests that DenseNet201 offers the optimal balance between classification accuracy and computational cost for industrial deployment.

To further analyze the classification behavior of the best-performing model, we examined the cumulative confusion matrix of DenseNet201 aggregated across all five folds (Figure 2). The confusion matrix provides a detailed breakdown of true positives versus false positives for each roast category.



True \ Predicted	Dark	Green	Light	Medium	Overbaking
Dark	499	0	0	0	1
Green	0	500	0	0	0
Light	0	0	499	1	0
Medium	0	0	0	499	1
Overbaking	1	0	0	0	499

Fig. 2 Confusion matrix of the DenseNet201 model, illustrating classification performance across the five roast levels.

As indicated by the dominant diagonal elements, DenseNet201 exhibited exceptional sensitivity and specificity across all classes. The model achieved perfect classification accuracy for the 'Green' and 'Overbaking' classes. This result is significant because these stages represent the extremes of the roasting spectrum, and their correct identification is critical for basic quality control. Minor misclassifications were negligible, which aligns with the high overall accuracy of 99.84%. The few errors observed in other architectures typically occur between adjacent roast levels (e.g., Light vs. Medium or Medium vs. Dark) due to the continuous nature of the Maillard reaction, which creates subtle visual transitions. However, DenseNet201's feature reuse capability allowed it to effectively discriminate even these closely related categories, minimizing inter-class confusion. The sparsity of off-diagonal values confirms that the model does not suffer from significant bias toward any specific class, making it highly reliable for automated industrial inspection systems.

To ensure that the high accuracy of the DenseNet201 model is driven by relevant visual features rather than background noise or artifacts, we employed Gradient-weighted Class Activation Mapping. This technique generates heatmaps where red and yellow regions indicate areas of high importance for the model's classification decision. As illustrated in Figure 3, the model demonstrates a strong focus on semantically meaningful regions for distinct roast levels. For the Dark roast class (Fig. 3-a), the activation maps concentrate intensely on the bean surface, suggesting that the network has learned to identify specific characteristics such as oiliness and deep color saturation. In the case of Green coffee (Fig. 3-b), the model effectively highlights the unique texture and pale coloration of raw beans, clearly distinguishing them from roasted variants. Similarly, for the Overbaking class (Fig. 3-c), the focus shifts to surface irregularities and carbonized areas, which are key indicators of excessive roasting. Crucially, in all instances, the heatmaps are strictly confined to the coffee beans themselves, ignoring the white background. This visual evidence confirms that the model relies on intrinsic features like color intensity and surface texture rather than spurious correlations, thereby validating the robustness of the proposed framework for real-world quality control scenarios.

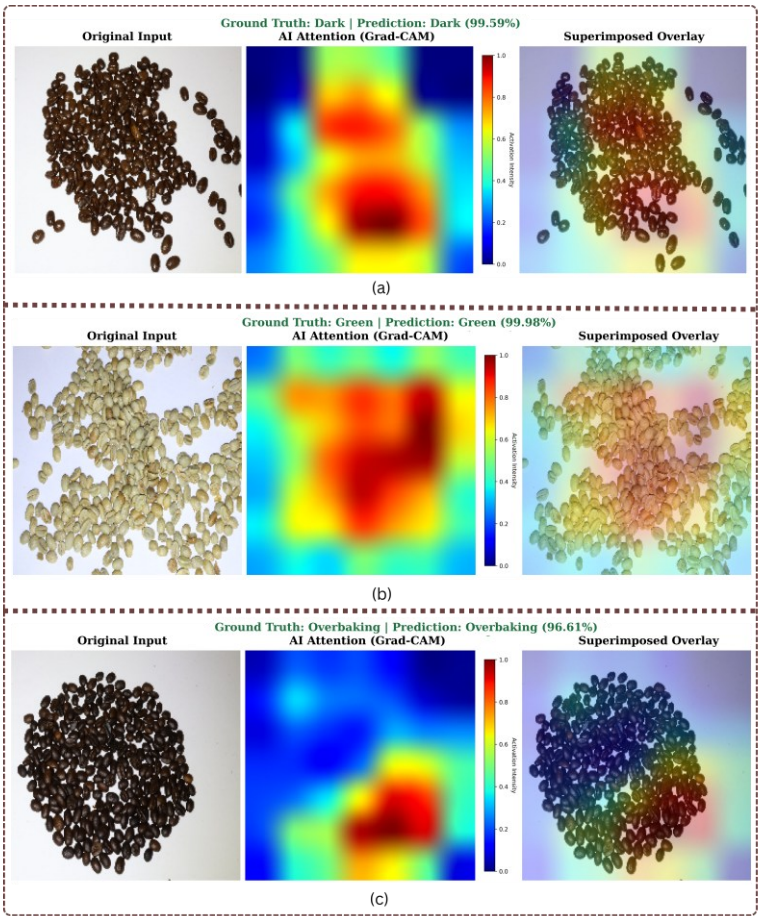


Fig. 3 Grad-CAM visualization results for different roast classes: (a) Dark, (b) Green, and (c) Overbaking, demonstrating the model's focus on relevant bean surface features while ignoring the background.

IV. Conclusion

This study presented a robust deep learning framework for the automated classification of coffee roast levels, addressing the limitations of subjective manual inspection and expensive colorimetric analysis in the coffee industry. By evaluating six state-of-the-art Convolutional Neural Network (CNN) architectures on an Agtron-standardized dataset, we demonstrated that computer vision techniques can achieve near-perfect accuracy in distinguishing fine-grained roasting stages. Experimental results obtained through 5-fold cross-validation revealed that DenseNet201 provided the state-of-the-art performance, achieving a classification accuracy of 99.84% and an F1-score of 0.9984. This architecture outperformed other robust models such as ResNet50 and Xception, primarily due to its efficient feature reuse mechanism which effectively captured subtle textural and color variations across the five roast categories (Green, Light, Medium, Dark, and Overbaking). Furthermore, the confusion matrix analysis confirmed the model's reliability, showing negligible misclassification even between adjacent roast levels. Beyond numerical performance, the integration of Grad-CAM provided critical visual interpretability. The activation heatmaps validated that the model's decision-making process is driven by intrinsic bean features, such as surface oiliness and color saturation, rather than irrelevant background noise. This explainability is crucial for building trust in automated quality control systems.

In conclusion, the proposed DenseNet201-based framework offers a cost-effective, non-destructive, and highly accurate alternative to traditional quality assessment methods. Future work will focus on deploying this model into a real-time mobile application for small-scale roasters and expanding the dataset to include a wider variety of coffee bean origins (e.g., Arabica vs. Robusta) to further enhance generalization.

Conflicts of Interest: The authors declare no conflict of interest

Funding: This research received no external funding.

Disclosure Statement: Generative artificial intelligence tools were employed for grammar refinement, linguistic clarity, and improvements in academic writing quality. These tools served as language-editing assistance within the manuscript preparation process.

References

[1] D. Giacalone, T. K. Degn, N. Yang, C. Liu, I. Fisk, and M. Münchow, "Common roasting defects in coffee: Aroma composition, sensory characterization and consumer perception," *Food quality and preference,* vol. 71, pp. 463-474, 2019.

[2] D. Seninde and E. Chambers, "Coffee flavor: a review. Beverages 6 (3): 44," ed, 2020.

[3] S.-J. Chang and C.-Y. Huang, "Deep learning model for the inspection of coffee bean defects," *Applied Sciences,* vol. 11, no. 17, p. 8226, 2021.

[4] I. M. Pakaya, R. Radi, and B. Purwantana, "Classification of Roasting Level of Coffee Beans Using Convolutional Neural Network with MobileNet Architecture for Android Implementation," *Jurnal Teknik Pertanian Lampung (Journal of Agricultural Engineering),* vol. 13, no. 3, p. 924, 2024.

[5] B. Turi, G. Abebe, and G. Goro, "Classification of Ethiopian coffee beans using imaging techniques," *East African Journal of Sciences,* vol. 7, no. 1, pp. 1-10, 2013.

[6] P. Mazzafera, "Chemical composition of defective coffee beans," *Food chemistry,* vol. 64, no. 4, pp. 547-554, 1999.

[7] M. A. Ximenes, "A tecnologia pós-colheita e qualidade física e organoléptica do café arábica de Timor," Universidade Tecnica de Lisboa (Portugal), 2010.

[8] C. Pinto, J. Furukawa, H. Fukai, and S. Tamura, "Classification of Green coffee bean images basec on defect types using convolutional neural network (CNN)," in *2017 international conference on advanced informatics, concepts, theory, and applications (ICAICTA)*, 2017: IEEE, pp. 1-5.

[9] F. Alamanda, R. Susanto, and W. Lestari, "Visual Segmentation and Classification of Coffee Beans After Roasting," *Journal of Applied Informatics and Computing,* vol. 9, no. 4, pp. 1354-1362, 2025.

[10] R. E. G. Rivas, P. L. L. Bertarini, and H. Fernandes, "Automated Coffee Roast Level Classification Using Machine Learning and Deep Learning Models," *Journal of Food Science,* vol. 90, no. 9, p. e70532, 2025.

[11] K. Przybył *et al.*, "Application of machine learning to assess the quality of food products—case study: Coffee bean," *Applied Sciences,* vol. 13, no. 19, p. 10786, 2023.

[12] J. A. S. Sarango. *Coffee Roast-Agtron Scale Dataset*, Kaggle, doi: https://doi.org/10.34740/KAGGLE/DSV/13456783.

[13] R. Yang, "Convolutional Neural Network for Image Classification Research-Based on VGG16," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2024: IEEE, pp. 213-217.

[14] R. Kursun, E. T. Yasin, and M. Koklu, "Machine learning-based classification of infected date palm leaves caused by dubas insects: a comparative analysis of feature extraction methods and classification algorithms," in *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2023: IEEE, pp. 1-6.

[15] P. Wang, H.-W. Tseng, T.-C. Chen, and C.-H. Hsia, "Deep Convolutional Neural Network for Coffee Bean Inspection," *Sensors & Materials,* vol. 33, 2021.

[16] O. Kilci, Y. Eryesil, and M. Koklu, "Classification of Biscuit Quality With Deep Learning Algorithms," *Journal of Food Science,* vol. 90, no. 7, p. e70379, 2025.

[17] A. Shabbir *et al.*, "Satellite and scene image classification based on transfer learning and fine tuning of ResNet50," *Mathematical Problems in Engineering,* vol. 2021, no. 1, p. 5843816, 2021.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[19] K. Tutuncu, E. T. Yasin, and M. Koklu, "Enhancing quality control: defect state classification of taralli biscuits with MobileNet-v2 and DenseNet-201," in *2023 IEEE 12th international conference on intelligent data acquisition and advanced computing systems: technology and applications (IDAACS)*, 2023, vol. 1: IEEE, pp. 718-723.

[20] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review,* vol. 53, no. 8, pp. 5455-5516, 2020.

[21] E. T. Yasin and M. Koklu, "Using pretrained models in ensemble learning for date fruits multiclass classification," 2025.

[22] Y. Eryeşil, H. Kahramanli Örnek, and Ş. Taşdemir, "Optimizing solid waste classification using deep learning and grey wolf optimizer for recycling efficiency," *International Journal of Environmental Science and Technology,* vol. 23, no. 1, p. 44, 2026.

[23] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314-1324.

[24] M. M. Saritas, R. Kursun, and M. Koklu, "Detection of Bone Fractures in X-ray Images with Machine Learning Methods Using InceptionV3 Deep Features," 2025.

[25] M. Koklu, I. Cinar, Y. S. Taspinar, and R. Kursun, "Identification of sheep breeds by CNN-based pre-trained InceptionV3 model," in *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, 2022: IEEE, pp. 01-04.

[26] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697-8710.

[27] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data,* vol. 8, no. 1, p. 53, 2021.

[28] B. Gencturk, E. T. Yasin, and M. Koklu, "Maturity Classification of Dragon Fruits Using Deep Learning Methods," *AGRI-INTELLIGENCE,* p. 182.

[29] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with

convolutional neural networks," *Physical and engineering sciences in medicine,* vol. 43, no. 2, pp. 635-640, 2020.

[30] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 8, pp. 1586-1594, 2019.

[31] B. Isgor and M. Koklu, "Lightweight Hybrid Model for Bone Fracture Detection Using MobileNetV2 Feature Extraction and Ensemble Learning," *Journal of Future Artificial Intelligence and Technologies,* vol. 2, no. 3, pp. 521-533, 2025.

[32] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one,* vol. 14, no. 11, p. e0224365, 2019.

[33] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics,* vol. 187, no. 1, pp. 95-112, 2015.

[34] S.-C. Vanegas-Ayala, D.-D. Leal-Lara, and J. Barón-Velandia, "Roasted coffee beans characterization through optoelectronic color sensing," *Coffee Science-ISSN 1984-3909,* vol. 18, pp. e182156-e182156, 2023.

[35] R. Kursun and M. Koklu, "Enhancing Explainability in Plant Disease Classification using Score-CAM: Improving Early Diagnosis for Agricultural Productivity," in *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2023, vol. 1: IEEE, pp. 759-764.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618-626.

[37] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018: IEEE, pp. 839-847.

[38] N. Nigar, H. M. Faisal, M. Umer, O. Oki, and J. M. Lukose, "Improving plant disease classification with deep-learning-based prediction model using explainable artificial intelligence," *IEEE access,* vol. 12, pp. 100005-100014, 2024.

[39] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one,* vol. 10, no. 7, p. e0130140, 2015.

[40] K. Gopalan, S. Srinivasan, M. Singh, S. K. Mathivanan, and U. Moorthy, "Corn leaf disease diagnosis: enhancing accuracy with resnet152 and grad-cam for explainable AI," *BMC Plant Biology,* vol. 25, no. 1, p. 440, 2025.