DEU FMD 28(83) (2026) 302-310

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

*RESEARCH ARTICLE / ARAŞTIRMA MAKALESI*

# Benchmarking Machine Learning and Transfer Learning Approaches for Fruit Classification Using Explainable Artificial Intelligence

## Açıklanabilir Yapay Zeka Kullanılarak Meyve Sınıflandırması için Makine Öğrenmesi ve Transfer Öğrenme Yaklaşımlarının Kıyaslanması

**Gözde Alp, Fatih Soygazi ***

Aydın Adnan Menderes Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü, Aydın, TÜRKİYE
*Corresponding Author/ Sorumlu Yazar **: fatih.soygazi@adu.edu.tr

**Abstract**

In this study, we investigate traditional machine learning methods with respect to state-of-the-art deep learning approaches for fruit image classification. Conventional classifiers such as Random Forest, and eXreme Gradient Boosting were evaluated alongside neural networks and state-of-the-art transfer learning models, including Residual Network (ResNet), Visual Geometry Group 16 (VGG16), and EfficientNet-B0. The fine-tuned ResNet50 model obtained the maximum classification accuracy as 97.57%, substantially surpassing other traditional solutions as well as shallow networks. One important contribution of this work is in the focus on model explainability. To understand the reasoning from the deep models, we used Gradient-weighted Class Activation Mapping (Grad-CAM) and Grad-CAM++ techniques to visualize what kind of decision is helping the model making predictions as well as by considering what regions in input images are being looked up for the concerned class. This explainability integration provides the transparency and explainability of classification system, being crucial for its real world applicability. The results underline the significance of both model capability and explainability when it comes to model selection, and show that fine-tuned deep networks combined with explainability tools is a strong and useful framework for image based classification.

***Keywords:*** *Deep Learning, Explainable Artificial Intelligence, GradCAM, Residual Networks, Transfer Learning*

**Öz**

Bu çalışmada, meyve görüntüsü sınıflandırması için geleneksel makine öğrenmesi algoritmaları ve modern derin öğrenme tekniklerinin karşılaştırmalı bir analizini sunuyoruz. Random Forest ve eXreme Gradient Boosting gibi geleneksel sınıflandırıcılar, sinir ağları ve Residual Network (ResNet), Visual Geometry Group 16 (VGG16) ve EfficientNet-B0 gibi en güncel transfer öğrenme modelleri ile birlikte değerlendirildi. Bunlar arasında, ince ayarlı ResNet50 mimarisi, %97.57'lik en yüksek sınıflandırma doğruluğunu elde ederek hem geleneksel yaklaşımları hem de sığ ağları açıkça geride bırakmıştır. Bu çalışmanın temel katkılarından biri, model açıklanabilirliğine verilen önemdir. Derin modellerin karar verme sürecini görselleştirmek için Gradyan Ağırlıklı Sınıf Aktivasyon Eşlemesi (Grad-CAM) ve Grad-CAM++ tekniklerini kullandık ve giriş görüntülerindeki ilgi bölgelerine dair içgörüler sunduk. Açıklanabilirliğin bu şekilde bütünleştirilmesi, özellikle gerçek dünyada çalıştırma veya kullanım senaryolarında kritik öneme sahip olan sınıflandırma sisteminin şeffaflığını ve güvenilirliğini artırır. Bulgular, model seçiminde hem doğruluğun hem de yorumlanabilirliğin önemini vurgulamaktadır ve ince ayarlı derin ağların açıklanabilirlik araçlarıyla birleştirilmesinin, görüntü tabanlı sınıflandırma görevleri için sağlam ve bilgilendirici bir çerçeve sağladığını öne sürmektedir.

***Anahtar Kelimeler:*** *Derin Öğrenme, Açıklanabilir Yapay Zeka, GradCAM, Kalıntı Ağlar, Transfer Öğrenmesi*

## 1. Introduction

In recent years, the categorization of fruits by means of computer vision and Machine Learning (ML) has received a great deal of attention because it can be useful for the agriculture, food quality control, and automatic retail systems applications [1]. The accurate classification of fruit categories using only visual features can minimize labor, enhance supply chain and produce homogeneous food products. Additionally, the application of image-based fruit classification systems in the context of smart agriculture and autonomous retail environments is stirring digitalization within the food domain [2]. Robust and efficient fruit classification methods have become a topic of much interest, due to the rise in computational power available along with availability of deep learning-based models [3].

A wide range of machine learning and deep learning models have been proposed for image classification tasks, with the rapid development of artificial intelligence. Classical models such as Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) provide interpretable and computationally efficient solutions, whereas Convolutional Neural Networks (CNN) and transfer learning approaches leveraging architectures like ResNet, EfficientNet, and VGG have shown remarkable performance in extracting spatial features from image data.

In this research we present a comprehensive model benchmarking on Fruits-360 dataset [4] which has been widely adopted in academic research thanks to its diversity and well-structured format. The dataset contains thousands of labeled images representing various fruit categories captured under controlled conditions, offering an ideal platform for evaluating different classification algorithms.

We aim to compare the performance of various machine learning and deep learning models on the Fruits-360 dataset, focusing on both classification accuracy and training time, while also

DOI: https://doi.org/10.21205/deufmd.2026288316

Geliş Tarihi / Received: 23.07.2025
Kabul Tarihi / Accepted: 23.10.2025

benefiting from explainable artificial intelligence (XAI) techniques. We evaluate traditional models including SVM, RF, Light Gradient-Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost) and ANN, as well as CNN and three popular pretrained deep learning architectures ResNet50, EfficientNet-B0 and VGG16 applied in both feature extraction and fine-tuning strategies. The results demonstrate the trade-offs between commonly used performance metrics and computational cost across different approaches, providing insights into the practical deployment of these models in real world fruit classification systems.

This study presents several contributions:

- We evaluate both classical machine learning (SVM, RF, ANN, LightGBM, XGBoost) and deep learning (CNN, fine-tuned/ feature extracted ResNet50, EfficientNet-B0, VGG16) approaches on the Fruits-360 dataset, providing a detailed comparison of accuracy and training time.
- We compare two transfer learning strategies feature extraction versus fine-tuning across three popular architectures (ResNet50, EfficientNet-B0, VGG16), quantifying the performance gains from each.
- We enhance the explainability of the best-performing model (fine-tuned ResNet50) by applying Grad-CAM and Grad-CAM++ techniques, providing visual explanations that reveal class-specific focus regions and support model transparency in decision-making.

The organization of the manuscript is as follows: the related studies from the literature are presented in Section 2. The machine learning algorithms and the CNN architectures are given in Section 3. Data preprocessing steps and performance measures are clarified in Section 4. Experimental results are presented in Section 5. The contribution of the article is summarized in Section 6.

## 2. Literature Review

Nowadays, different deep learning along with hybrid models have been developed [5], to address such measurements namely the performance block such as fruit image classification work using the food image datasets indicating robustness and interpretability among model architectures and dataset settings.

The general trend is based on the application of CNN and transfer learning for improving classification accuracy. Salim et al. observed that AlexNet [6] achieved substantially higher accuracy compared to traditional classifiers like Decision Trees, KNN and SVM, suggesting the effectiveness of deep CNN for image learning. The effectiveness of fine-tuned ResNet50 is shown, especially when preprocessing techniques and data augmentation [7] are applied. Concurrently, a custom CNN was trained for 50 epochs on TF proposed [8], promising results were obtained with such simple network architecture, which supports the importance of tailored CNN models for fruit classification.

Another category of works is the ensemble/hybrid model aiming to enhance the robustness and classification accuracy. A set of EfficientNet-B0, MobileNetV2, and ResNet50V2 [9] was presented and obtained 99.32% in a 24-class subset (97.15% on the full dataset), which indicates improvements in generalization ability. Transfer learning with EfficientNet-B0 also employed [10] and achieved near-perfect accuracy on the test set of a 92-class subset, illustrating the effectiveness of pretrained feature extractors. Moreover, a custom 12-layer CNN, “FruitNet” was compared to 13 pretrained models with exception providing the highest accuracy and MobileNet continuing to be the most efficient [11].

Aside from classification accuracy, robustness and adversarial resistance of the model have especially gained attention in the field. Siddiqi [12] investigated CNN vulnerability to adversarial attacks using the FGSM method and demonstrated that although baseline models achieved high accuracy, they were vulnerable to small input perturbations. The experiments showed that adversarial training is effective in making the model more robust against adversarial examples, which is a desirable property for real world applications.

In addition to the architecture dependent work, application specific and resource sensitive solutions have been explored. The classic image features were combined with MobileNetV2 to construct a lightweight model that can perform quality classification of fruits, i.e., damaged versus healthy [13]. This solution was mainly intended for small scale and embedded applications which had low computational resources. A Modified Cascaded-ANFIS system was introduced capable of classifying all 131 classes in the data set, which outperformed multiple CNN-based alternatives both at accuracy and computational cost, providing a fuzzy logic alternative for deep learning [14]. However, our experimental results reveal that even when fine-tuned on the Fruits-360 dataset (ResNet50 achieves 97.57% classification accuracy) the accuracy of ResNet50 is relatively comparable to other approaches reported in the literature and provides good classification performance for 131 classes of fruits. Unlike many existing works that focus solely on predictive accuracy, our approach emphasizes model explainability techniques.

Last but not least, a few works have investigated the relevance of dataset diversity and specialization. While Fruits-360 remains a widely adopted benchmark, new datasets are being introduced to address limitations in coverage and specificity. Mimma et al. [15], which released a private 8-class fruit dataset for smaller-sized classification tasks. Altaheri et al. [16] concentrated on specific scenarios like harvesting date fruits. Meanwhile, Abayomi-Alli et al. [17] introduced a new dataset for fruit freshness evaluation, which aims at filling the gap of the requirements on real time quality testing and multi-class freshness grading.

## 3. Related Background

This section outlines the theoretical foundations and core principles of the machine learning algorithms and convolutional neural network architectures utilized in this study. The aim is to provide the necessary background for understanding the methodologies adopted.

### 3.1. Machine Learning Algorithms

This part presents the classical machine learning methods used in our work.

#### 3.1.1. Random Forest

RF is a machine learning technique that constructs multiple decision trees during the learning phase and returns the mode of their classifications for classification tasks or their mean for regression. When averaging over a large number of different trees, it reduces greatly overfitting in contrast with individual decision trees. Randomness is used in both feature selection and data sampling to improve the ability of generalization of the algorithm. RF is popular because it is accurate, robust against noise and easily interpretable via feature importance scores.

#### 3.1.2. Support Vector Machine

SVMs are supervised learning models which are well suited for classification problems. They work by determining the best hyperplane to maximally separate data points of different classes

in a high-dimensional space. One of the reasons for it, is due to the robustness in high-dimensionality and ability to shape nonlinear decision boundaries by kernel function. They work well on high dimensional feature space (i.e., when the number of features is greater than the samples) and are widely used in text categorization, bioinformatics and image classification.

#### 3.1.3. Extreme Gradient Boosting

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance in classic problems with structured data. It incorporates regularization methods such as L1 and L2 penalties to prevent overfitting, also optimized for fast parallel computation with reduced training time. XGBoost sequentially builds a collection of decision trees, with the goal that each new tree will correct the mistakes of its predecessors. Its generalization to both benchmark tasks and actual deployment is due to its trade-off of efficiency, flexibility, as well as highly accurate performance.

#### 3.1.4. Light Gradient-Boosting Machine

LightGBM is a gradient boosting framework that uses tree based learning algorithms; it grows trees "leaf-wise" instead of other traditional “level-wise” growing. It is specifically designed to work efficiently with large data sets and can handle both categorical and continuous features naturally. It is based on histogram-based best-first splitting and gradient-based one-side sampling to make full use of data information while accelerating computation which avoids the sacrifice of model performance. It is widely used in credit scoring, ranking as well as recommendation systems for its scalability and predictivity.

#### 3.1.5. Artificial Neural Network

ANN is a mode of computation based on the biological structure of the human brain. It is represented by connected nodes (neurons) which are organized in layers, and each neuron takes inputs and forwards them to the next layer. ANN is flexible in the sense that it can approximate complex nonlinear mappings and, there are a wide range of applications involving ANN such as classification, regression, time series prediction etc. However, unlike CNN, it does not explicitly consider spatial or sequential structures, which may limit their effectiveness on image or language data without architectural adjustments.

#### 3.1.6. Convolutional Neural Network

CNNs are a type of deep learning model that receive as input grid-like data, such as images. It employs convolutional layers that can learn spatial hierarchies of features by themselves and therefore take into account patterns like edges, textures, or shapes of objects. CNNs are very efficient in classification, detection and medical imaging. Its strength lies in its ability to reduce the need for manual feature extraction, as it learns meaningful representations directly from raw images.

### 3.2. CNN Architecture Details

Over the past decade, CNN has become the pioneer of computer vision due to its capability of learning hierarchical features from image data. With the development of deep learning, increasingly complex CNN structures were investigated to achieve better performance in accuracy, training speed and generalization. Among these, Residual Networks (ResNet), EfficientNet, and VGG16 have gained significant attention due to their performance on large-scale image classification benchmarks. Each architecture introduces unique innovations such as skip connections, compound scaling, and depth-wise simplicity that enable more effective feature learning in deep neural networks.

#### 3.2.1. Residual Networks

To overcome the issue of vanishing gradients in deep neural networks, ResNets propose skip connections (or simply identity shortcuts). The transformed scores are rescaled using a standard softmax activation, which allows the model to learn residual functions by bypassing one or more layers, so that very deep networks can be trained, including hundreds of layers. The crucial observation inspired by this is that it is much easier to optimize the residual mapping than the original, unreferenced mapping. In consequence, ResNet greatly accelerates the training process and has also achieved state-of-the-art performance on a number of vision tasks, such as image classification, segmentation and object detection.

#### 3.2.2. Efficientnet

EfficientNet is a convolutional neural network that has been designed to scale the efficiency by achieving state-of-the-art accuracy with relatively low computational resources. In contrast, common practice is to scale individual dimensions of the network (such as depth, width and resolution) without maintaining an equal ratio between the three factors. Built upon a baseline architecture called EfficientNet-B0, the model family progressively increases model size with controlled scaling coefficients. It also has depthwise separable convolutions and squeeze-and-excitation blocks, so it is very lightweight but accurate enough. EfficientNet has achieved state-of-the-art accuracy on ImageNet with up to 10x better efficiency (parameters and FLOPs).

#### 3.2.3. Visual Geometry Group 16

VGG16 from the Visual Geometry Group at University of Oxford is a deep convolutional network which is characterized by its simplicity, uniform architecture and high level of accuracy. It has 16 layers small 3x3 convolutional filters, max-pooling and fully connected. The model greatly focuses on depth by increasing number of stacked convolutional layer, so that it extracts state-of-the-art features at multiple scales. Although VGG16 is computationally expensive compared to more modern architectures, it is an attractive option in transfer learning as there are pretrained weights available with stable performance.

#### 3.2.4. Fine-Tuning

Fine-tuning was implemented by unfreezing the selected top layers of the pretrained ResNet, EfficientNet and VGG16 models. These layers were trained simultaneously with the added classification layers. This enables the networks to adapt their high-level feature detectors to extract better patterns and variations from the particular dataset. Careful layer selection and a lower learning rate were used to prevent catastrophic forgetting of previously learned general features, while enabling gradual adaptation to the new task. Hence, fine-tuning can improve the performance of the model by adapting pretrained filters specifically to a given task, especially when the dataset size and complexity support deeper retraining.

#### 3.2.5. Feature Extraction

For all three CNN architectures (ResNet50, EfficientNet-B0, and VGG16) feature extraction was performed by freezing the convolutional base of each pretrained model and using it to generate high-level feature representations from the input images. These pretrained weights, originally optimized on large-scale datasets such as ImageNet, provide robust and transferable visual features. A new classification head composed by fully connected layers was added on top of each frozen model to make predictions specific for the target classes. This approach is

computationally efficient and has the advantage of preventing overfitting, especially when training data are very scarce, while exploiting deep hierarchical prediction features trained on large datasets.

### 3.3. Explainability

Explainability refers to the ability to understand and interpret the internal mechanisms behind a model's predictions. The models in general and deep learning architectures in particular are becoming even more complex, making it difficult to interpret their decision processes. This lack of openness can reduce trust, prevent debugging and can become a concern in sensitive applications, such as medical imaging or quality control on production lines. In answer, a host of post-hoc explanation methods have been devised to either visualize or quantify which input features are affecting the model outputs. Among these, methods like Grad-CAM and Grad-CAM++ are widely used in image classification tasks, as they produce class-specific heatmaps that highlight the most informative regions of an image. Incorporating explainability into model evaluation not only improves interpretability but also facilitates model refinement and increases confidence in real world deployment.

We illustrate an overview of our system in Figure 1. The workflow begins with image preprocessing, followed by the application of both traditional machine learning algorithms and CNN architectures. Transfer learning strategies, including fine-tuning and feature extraction are employed to enhance model performance. Once we have tested and compared the models based on accuracy, we perform explainability analysis with GradCAM and GradCAM++ visualizations to the most accurate model.

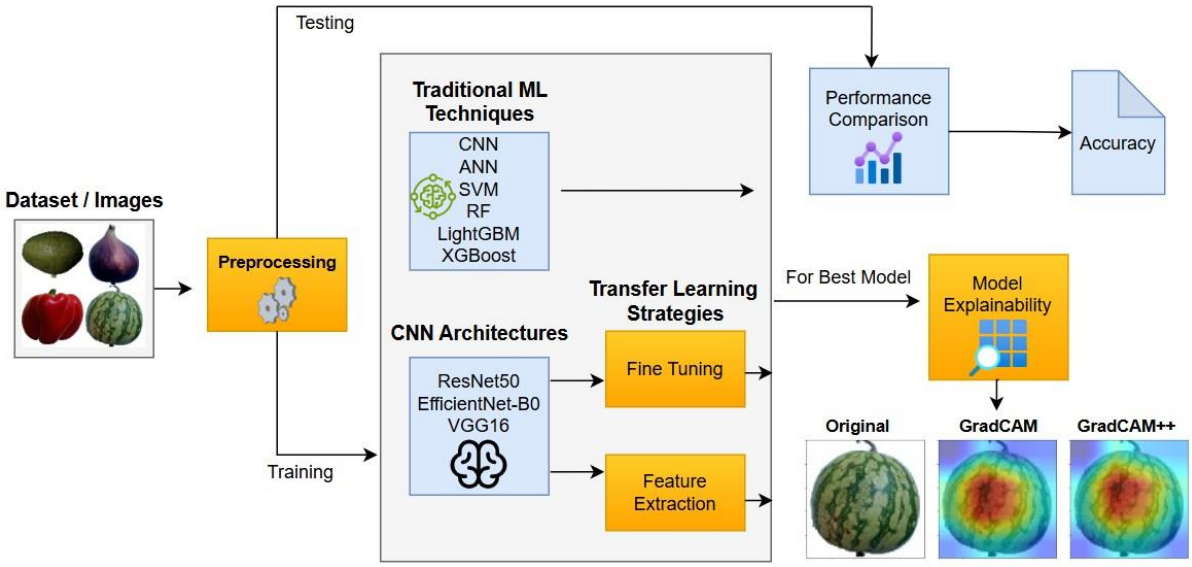


**Figure 1.** The overview of the system.

## 4. Material and Methods

The dataset details, data preprocessing steps and evaluation metrics are explained in this section.

### 4.1. Dataset Details

The Fruits-360 dataset [4] is a popular augmented benchmark data-set for image classification in agriculture and food. This dataset includes 138704 images, belonging to 131 different classes of fruits and either captured as object-detection images under varying conditions or as a 'scan' (strongly limited identity-preserving angle change). Many techniques such as data augmentation were used in its construction to generate additional diversity within the data and reduce overfitting. To augment the data, each original image was subjected to rotation, translation, horizontal flip, vertical flip, zoom in/out and random background removal to mimic real world situations such as lighting changes and viewpoint variations. The dataset includes high resolution images acquired at different angles, making it possible to have a good intraclass variation without interfering with interclass separability. Some samples from the Fruits-360 dataset are illustrated in Figure 2, some fruit categories are labeled with numerical suffixes (e.g., "Apple Golden 1", "Cherry 2") to indicate different varieties, appearances, or sources of the same fruit type, allowing for more fine-grained classification and variability within a single class group. Each fruit class includes hundreds of images, allowing for robust model training and evaluation. Thanks to its diversity and well-balanced distribution, the Fruits-360 dataset has increasingly been used as a benchmark for evaluating the performance and generalization capabilities of traditional machine learning and deep learning models.


**Figure 2.** Image samples from the dataset.

### 4.2. Data Preprocessing

The preprocessing workflow for the dataset was conducted through four consecutive steps (Figure 3). All subcategories were initially categorized into quite broad classes based on folder names, then images were gathered into new structured folders. The images were loaded and resized to the same size in parallel for efficient I/O. In the next step, the image arrays and corresponding labels were stored in compressed. npz format for

efficient reuse. In stage three, constructed image arrays were converted to 1D vectors and batch normalization [18] was used based on the standard scaler for raw feature distribution to avoid model bias related with pixel intensity. The mathematical formula for standard scaler calculation is given in Equation 1, where $\overline{x}$ is the average of feature vector $x$ and $\sigma$ is its standard deviation. The reduction of dimensionality was subsequently performed by PCA where the components that captured most of the variance were kept. Finally, also a lower resolution version of the dataset was produced with images resized to 50×50 and quantized in colors using k-means clustering to decrease the color space complexity and make processing lighter for cheaper models.

$$x' = \frac{x - \overline{x}}{\sigma} \tag{1}$$

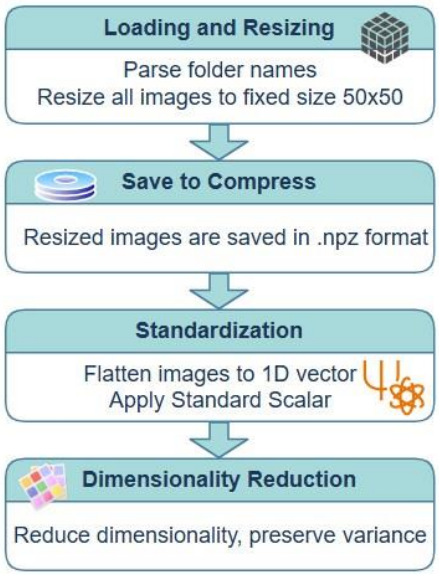


**Figure 3.** Data preprocessing steps.

### 4.3. Evaluation Process

The methods proposed were evaluated in terms of accuracy, precision, recall and F score. For this, True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) metrics were employed. TP is the true positive, which means that it represents cases of positiveness correctly predicted by our prediction model, and TN refers to the predictions for negative cases what has been made right. FP represents the situation where negative samples are mistakenly judged as being positive. In turn, FN refers to the cases where positive samples were incorrectly labeled as negative. The metric values are provided in Equations 2 to 5.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$F\ score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{5}$$

## 5. Experimental Results

The code of the project is written on Intel i7 5500U CPU and 16 GB memory system using Windows 10 operating system with VS code as a compiler in Python programming language. Many machine learning libraries and frameworks are used including PyTorch[1], OpenCV[2], H5py[3], NumPy[4], Pandas[5] and Matplotlib[6]. The experimental setting used to train on 131 classes of Fruits-360 dataset. The splitting ratio was set to 90% training and 10% testing and all models were trained for 50 epochs using the same parameters to be fair when comparing architectures.

### 5.1. Model Performance Analysis

The experimental results provide valuable insights into the comparative performance of various traditional machine learning models, neural networks, and deep learning architectures for the task at hand. The accuracy, precision, recall, F score results are given in Table 1, based on the cross-validated performance of the model. Among traditional models, the CNN achieved the highest accuracy of 95.82%, indicating its strong ability to capture spatial features inherent in image data. Although slightly lower in accuracy, the ANN (95.22%) and the SVM (94.7%) demonstrated competitive performance with considerably shorter training times, particularly in the case of the ANN (135 seconds), making it an efficient alternative for real time or resource-constrained scenarios.

RF and gradient boosting methods such as LightGBM and XGBoost yielded comparatively lower accuracies (90.18%, 87.78%, and 83.88% respectively), though they maintained advantages in simplicity. Notably, RF required only 16 seconds for training, highlighting its efficiency despite the moderate performance.

For transfer learning methods, we found fine-tuned deep CNNs to deliver better performance than feature extraction-based ones. The four convolutional blocks of the model (Figure 4) commonly referred to as the convolutional base, are responsible for extracting hierarchical feature representations from input images. In the context of CNN, fine-tuning is a process in which part of the convolutional base is selectively unfrozen and retrained alongside the classifier. Typically, the top one layer of the base network that capture more abstract and task-specific features is made trainable, while the three lower-level layers remain frozen to preserve general feature extraction capabilities. This gradual adaptation allows the model to refine its internal representations to better suit the target domain, leading to improved accuracy and generalization. Often, fine-tuning is applied only after the classifier has been trained on top of the frozen base, ensuring that the model does not forget previously learned representations and converges more stably.

[1] https://pytorch.org/
[2] https://opencv.org/
[3] https://www.h5py.org/
[4] https://numpy.org/
[5] https://pandas.pydata.org/
[6] https://matplotlib.org/

**Table 1.** The accuracy, precision, recall, F score and the training time outcomes in the experiments.

| Model | Phase | Accuracy | Precision | Recall | F Score | Training Time Seconds |
|---|---|---|---|---|---|---|
| CNN | | 95.82 | 0.9605 | 0.9582 | 0.957 | 1155 |
| ANN | | 95.22 | 0.9533 | 0.9522 | 0.9514 | 135 |
| SVM | | 94.70 | 0.9504 | 0.947 | 0.946 | 136 |
| RF | | 90.18 | 0.9089 | 0.9018 | 0.8999 | 16 |
| LightGBM | | 87.78 | 0.8819 | 0.8778 | 0.8756 | 216 |
| XGBoost | | 83.88 | 0.8441 | 0.8388 | 0.8366 | 102 |
| ResNet50 | fine-tuning | **97.57** | 97.80 | 97.60 | 97.60 | 188 |
| EfficientNet-B0 | fine-tuning | 96.23 | 96.40 | 96.20 | 96.20 | 128 |
| VGG16 | fine-tuning | 94.85 | 95.00 | 94.80 | 94.80 | 355 |
| EfficientNet-B0 | feature extraction | 93.91 | 94.10 | 93.90 | 93.80 | 39 |
| ResNet50 | feature extraction | 93.71 | 93.90 | 93.70 | 93.70 | 71 |
| VGG16 | feature extraction | 87.62 | 88.00 | 87.60 | 87.50 | 120 |

Feature extraction involves utilizing the pretrained convolutional layers of a well-established network which have already learned to detect generic visual patterns like edges, textures, shapes, and other abstract features from large-scale datasets. Instead of training the entire model from scratch, the convolutional base is reused as a fixed feature extractor: the new dataset is passed through these frozen layers (four convolutional blocks), and the resulting output feature maps are fed into a newly added classifier (Fully Connected (FC) Layer - Figure 4) that is trained to perform the specific task at hand. This approach offers the advantages of reduced training time and a lower risk of overfitting while leveraging the powerful, general-purpose representations learned by deep networks.

ResNet50 attained the best accuracy with fine-tuning (97.57%) while EfficientNet-B0 (96.23%) and VGG16 (94.85) followed closely thereafter. These findings highlight the benefits of end-to-end fine-tuning in accommodating pretrained models for any given datasets. In comparison, freezing the pretrained backbone and training only feature extractors led to worse performance with a top-1 accuracy of 93.91% and 93.71% respectively for EfficientNet-B0 and ResNet50. The VGG16 model, functioning as a feature extractor, greatly underperformed (87.62%), suggesting that these kinds of architectures may be less suited for transfer without full retraining.

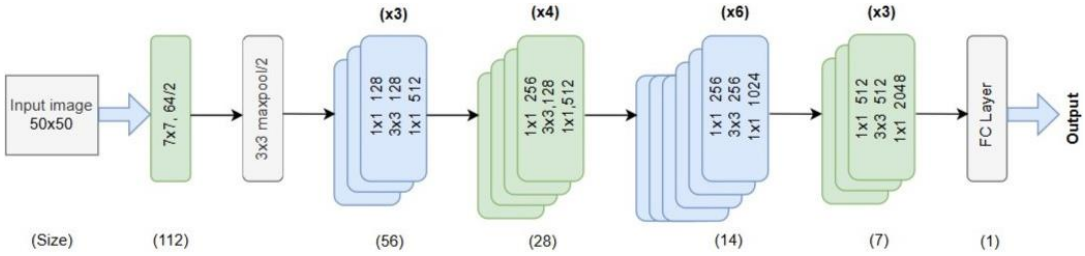


**Figure 4.** Fine-tuned ResNet50 architecture.

The better results of the best fine-tuned ResNet50 on Fruits-360 dataset is likely due to its deep residual network architecture that easily resolves the vanishing gradient problem and therefore facilitates learning of deeper hierarchical representations. Its skip connections allow the model to preserve important low-level information while refining higher-level representations, leading to more accurate fruit classification. Moreover, fine-tuning the pretrained ImageNet weights by unfreezing the last hidden layer of ResNet50 provides a strong initialization, facilitating efficient adaptation to the visual characteristics of the Fruits-360 dataset.

Due to the large number of classes in the dataset (131 in total), presenting detailed confusion matrix for each class individually would exceed space constraints. Instead, the performance of the best-performing model (ResNet50-fine-tuned) across a sample of nine representative classes is illustrated using a confusion matrix in Figure 5. The samples for selected classes are presented in Table 2.

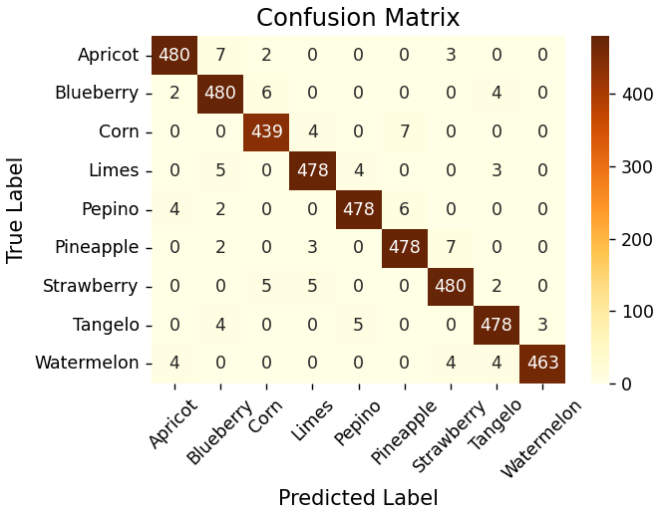


**Figure 5.** Confusion matrix for nine classes classification.

**Table 2.** Sample distribution of the Fruit-360 dataset among some of the classes.

| Class Label | Number of Samples |
|---|---|
| Apricot | 492 |
| Blueberry | 462 |
| Corn | 450 |
| Limes | 490 |
| Pepino | 490 |
| Pineapple | 490 |
| Strawberry | 492 |
| Tangelo | 490 |
| Watermelon | 475 |

**Table 3.** ResNet50 fine-tuning hyperparameters.

| Hyperparameter | Value |
|---|---|
| Weights | 'imagenet' (pretrained) |
| Optimizer | Adam |
| Learning Rate | 0.00001 |
| Loss Function | 'categorical_crossentropy' |
| Batch Size | 64 |
| Epoch | 50 |

Fine-tuned ResNet50 architecture hyperparameters are given in Table 3. A small learning rate (0.00001) ensured gradual fine-tuning of pretrained layers without degrading feature representations. A batch size of 64 provided a good balance between convergence speed and memory efficiency, and 50 epochs were sufficient for the model to converge during fine-tuning without overfitting. Loss and accuracy graphs for the best performing ResNet50-fine-tuning are given in Figure 6 and Figure 7, respectively.

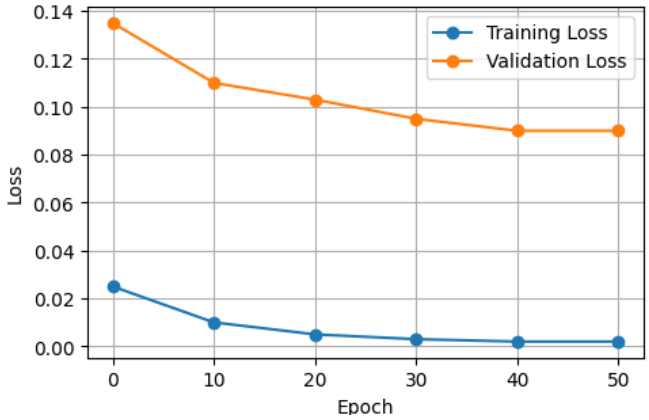


**Figure 6.** Validation and training loss for ResNet50-fine-tuning.

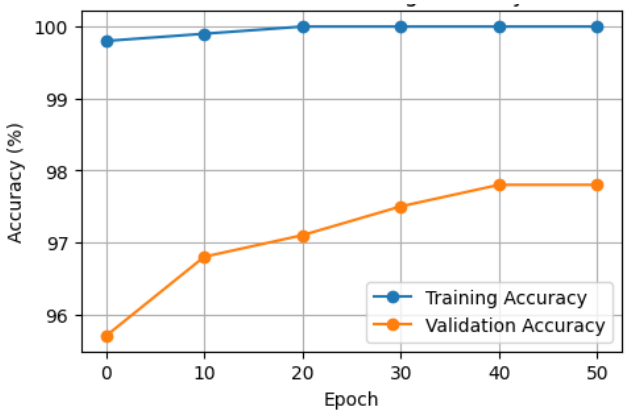


**Figure 7.** Validation and training accuracy for ResNet50-fine-tuning.

Numerous studies have leveraged the Fruits-360 dataset to evaluate the performance of traditional and deep learning-based classification methods, as summarized in Table 4. As can be observed, deep architectures such as AlexNet, MobileNetV2, and EfficientNet-B0 have achieved notably high accuracy scores, often surpassing 99%, even with a relatively low number of epochs or classes. While methods like VGG-16 show comparatively lower performance, this may be attributed to the lack of fine-tuning or limited training epochs.

Compared to existing methods, our ResNet50-based fine-tuned model achieves competitive accuracy (97.57%) with only 50 epochs and the full set of 131 fruit classes, indicating a favorable balance between performance and training efficiency.

**Table 4.** Overview of literature on fruit image classification using deep learning models.

| Method | Dataset | Classes | Epochs | Accuracy (%) |
|---|---|---|---|---|
| AlexNet [6] | Fruits-360 | 40 | 5 | 99.85 |
| VGG-16 [7] | Fruits-360 | - | 30 | 90.27 |
| CNN [8] | Fruits-360 | 131 | 50 | 94.35 |
| Multi-fused CNN [9] | Fruits-360 | 24 - 131 | 500 | 99.32- 97.15 |
| CNN - EfficientNet-B0 [10] | ImageNet - Fruits-360 | 11 - 92 | 110 - 110 | 96.77 - 99.9 |
| IndusNet[12] | IndusFruits dataset [19] | 18 | 97.57 | 95.67 |
| MobileNetV2 [14] | Fruits-360 | - | 50 | 100.0 |
| Cascaded-ANFIS [19] | Fruits-360 | 131 | - | 98.40 |
| **Our Study - ResNet50-fine-tuning** | **Fruits-360** | **131** | **50** | **97.57** |

To further understand how our model makes classification decisions, the next section provides detailed explainability techniques.

### 5.2. Explainability of the Model Decisions

Deep learning models, in particular CNNs, have made impressive performance on image classification. Among these, our experiments on the Fruits-360 dataset returned best results for the fine-tuned ResNet50 model. Nevertheless, despite its high prediction performance, the model's decision-making process remains largely opaque, reflecting the so-called "black box" nature of deep learning where we cannot easily understand why certain data patterns are classified.

To provide better explainability of the ResNet50 model and to know what visual features it depends on for prediction, we have used Gradient-weighted Class Activation Mapping (Grad-CAM) and gradient-weighted Grad-CAM++. These techniques create heatmaps that represent the most relevant parts of an input image to make a classifying decision.

Figure 8, Figure 9 and Figure 10 show the Grad-CAM and Grad-CAM++ visualizations of three example classes; Watermelon Lemon and Corn, respectively. The watermelon example represents the best performance, producing the most accurate and dense heatmaps of both methods; the model demonstrated robust and reliable judgment by focusing sharply on the entire object. The lemon performs moderately; the model successfully captures the key distinguishing feature color and central mass, but focuses less on the lemon's contours and corners, suggesting a reliance on dominant color rather than geometric shape. The corn example demonstrates the weakest situation, as the model pays a diffuse and very broad (where its performance is poor) attention. The heatmap extends beyond the object to irrelevant white background regions, suggesting that the model bases its judgment on broader contextual cues rather than object-specific features.

These visual explanations validate that the ResNet50 model can extract class-informative features for well-performing categories. But at the same time, they express an opportunity of improvement in classes where classification accuracy is lower. Explainability tools like Grad-CAM can help to provide an understanding of model functioning and thereby ensure transparency and trust when deploying deep learning in real world applications for food categorization.

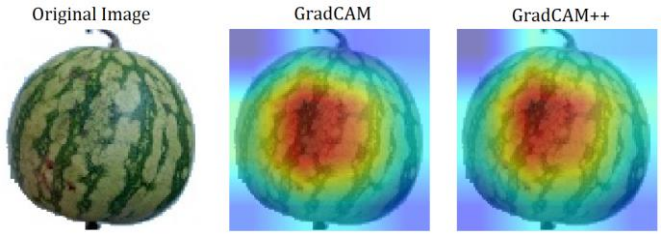


**Figure 8.** Grad-CAM visualization for the best class "Watermelon".

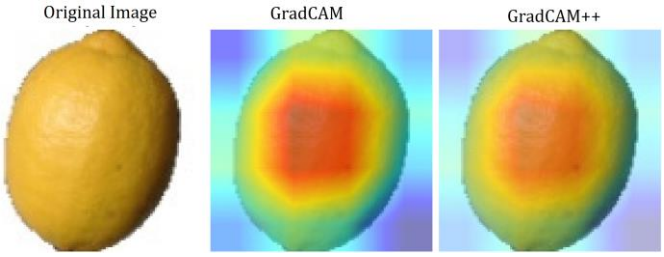


**Figure 9.** Grad-CAM visualizations for an average class "Lemon".

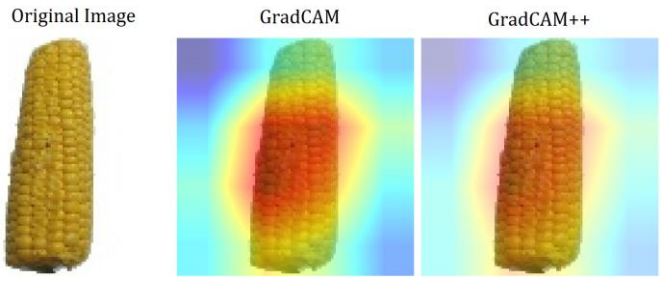


**Figure 10.** Grad-CAM visualizations for the worst class "Corn".

Overall, the findings suggest that fine-tuned deep learning models, particularly ResNet50, offer the best trade-off between accuracy and training efficiency for this classification task. However, in settings where computational resources or time are limited, ANN or traditional machine learning models like SVM may serve as viable alternatives without substantial loss in performance.

## 6. Conclusions

In this study, a comprehensive comparison of traditional machine learning models, neural networks, and deep convolutional architectures was conducted on the Fruits-360 dataset to evaluate their effectiveness in fruit image classification. The results show a significant performance improvement by the deep fine- tuning models, notably ResNet50, which obtained an accuracy of 97.57%, surpassing traditional classifiers and shallow neural networks. The better performance of ResNet50 well demonstrates the significance of deep residual connections and the convenience from end-to-end finetuning towards image-rich datasets.

Among the traditional methods the CNN has been able to achieve a comparable result, meaning that even very lightweight convolutional architectures can compete on this task. In the meantime, ANN and SVM also performed comparably well with much shorter training times, so they can be compared to good candidates in challenged environments of low computation capacity. Nevertheless, the tree-based methods (RF, LightGBM and XGBoost) achieved lower accuracies but have the merits of being faster in training process and interpretability.

Overall, the study underscores the importance of selecting an appropriate model based on the specific requirements of the application, such as accuracy, training time, and available hardware. Fine-tuned deep learning models give the best performance but take a longer time for training and have higher computational requirement. In future work, we plan to visualize how much each region contributes from the heat map by displaying it through the numerical importance levels of the heat map [20] and compare the model with saliency maps for better understanding [21]. Furthermore, to continue improving the effectiveness of our model, we would like to perform hyperparameter tuning with popular frameworks such as Optuna [22], which have been effective for improving deep learning models in image classification tasks.

Furthermore, in the future more attention should be paid to enhance efficiency via model compression, and quantization, pruning for reducing computational cost and memory usage so as to facilitate transferred learning algorithm execution on edge devices. Approaches such as Post-Training Quantization, Quantization-Aware Training, and hardware-aware optimizations will be investigated to tailor models such as ResNet50 for resource-constrained scenarios so that inference is faster and more energy-efficient [23].

### Ethics committee approval and conflict of interest statement

This article does not require ethics committee approval. This article has no conflicts of interest with any individual or institution.

## References


[1] Song X, Zhang X, Dong G, Ding H, Cui X, Han Y, et al. AI in food industry automation: applications and challenges. Frontiers in Sustainable Food Systems 2025;9. doi:10.3389/fsufs.2025.1575430.

[2] Thapa A, Nishad S, Biswas D, Roy S. A comprehensive review on artificial intelligence assisted technologies in food industry. Food Bioscience 2023;56. doi:10.1016/j.fbio.2023.103231.

[3] Zhu L, Spachos P, Pensini E, Plataniotis KN. Deep learning and machine vision for food processing: A survey. Current Research in Food Science 2021;4:233-249. doi:10.1016/j.crfs.2021.03.009.

[4] Muresan H, Oltean M. Fruit recognition from images using deep learning. Acta Universitatis Sapientiae, Informatica 2018;10(1):26-42. doi:10.2478/ausi-2018-0002.

[5] Ukwuoma CC, Zhiguang Q, Bin Heyat MB, Ali L, Almaspoor Z, Monday HN. Recent advancements in fruit detection and classification using deep learning techniques. Mathematical Problems in Engineering 2022;2022(1):9210947. doi:10.1155/2022/9210947.

[6] Salim NO, Mohammed AK. Comparative Analysis of Classical Machine Learning and Deep Learning Methods for Fruit Image Recognition and Classification. Traitement du Signal 2024;41(3). doi:10.18280/ts.410322.

[7] Zhang D. Fruit 360 classification based on the convolutional neural network. Applied and Computational Engineering 2023;15:219-222.

[8] Joseph JL, Kumar VA, Mathew SP. Fruit classification using deep learning. In: Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021. Springer Singapore; 2021, p. 807-817.

[9] Sinha BB, Dhanalakshmi R. A multi-fused convolutional neural network model for fruit image classification. International Journal of Cognitive Computing in Engineering 2024;5:416-424. doi:10.1016/j.ijcce.2024.09.003.

[10] Bongulwar DM, Talbar SN. Robust Convolutional Neural Network Model For Recognition of Fruits. Indian Journal of Science and Technology 2021;14(45):3318-3334. doi:10.17485/IJST/v14i45.1493.

[11] Siddiqi R. Comparative performance of various deep learning based models in fruit image classification. In: Proceedings of the 11th International Conference on Advances in Information Technology; 2020, p. 1-9. doi:10.1145/3406601.3406619.

[12] Siddiqi R. Fruit-classification model resilience under adversarial attack. SN Applied Sciences 2022;4(1):31. doi:10.1007/s42452-021-04917-6.

[13] Zárate V, Hernández DC. Simplified Deep Learning for Accessible Fruit Quality Assessment in Small Agricultural Operations. Applied Sciences 2024;14(18):8243. doi:10.3390/app14188243.

[14] Rathnayake N, Rathnayake U, Dang TL, Hoshino Y. An efficient automatic fruits-360 image identification and recognition using a novel modified cascaded-ANFIS algorithm. Sensors 2022;22(12):4401. doi:10.3390/s22124401.

[15] Mimma NEA, Ahmed S, Rahman T, Khan R. Fruits classification and detection application using deep learning. Scientific Programming 2022;2022(1):4194874. doi:10.1155/2022/4194874.

[16] Altaheri H, Alsulaiman M, Muhammad G, Amin SU, Bencherif M, Mekhtiche M. Date fruit dataset for intelligent harvesting. Data in Brief 2019;26:104514. doi:10.1016/j.dib.2019.104514.

[17] Abayomi-Alli OO, Damaševičius R, Misra S, Abayomi-Alli A. FruitQ: a new dataset of multiple fruit images for freshness evaluation. Multimedia Tools and Applications 2024;83(4):11433-11460. doi:10.1007/s11042-023-16058-6.

[18] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 2011;12:2825-2830.

[19] Wang SH, Chen Y. Fruit category classification via an eight-layer convolutional neural network with parametric rectified linear unit and dropout technique. Multimedia Tools and Applications 2020;79(21):15117-15133. doi:10.1007/s11042-018-6661-6.

[20] Marmolejo-Saucedo JA, Kose U. Numerical grad-cam based explainable convolutional neural network for brain tumor diagnosis. Mobile Networks and Applications 2024;29(1):109-118. doi:10.1007/s11036-022-02021-6.

[21] Hu B, Tunison P, RichardWebster B, Hoogs A. Xaitk-saliency: An open source explainable ai toolkit for saliency. In: Proceedings AAAI Conference on Artificial Intelligence; 2023;37(13):15760-15766. doi:10.1609/aaai.v37i13.26871.

[22] Benedichuk M, Bashkova P, Tuchinov B. Computer Vision and Explainable Approaches for Chest Tuberculosis Screenings. In: 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). IEEE; 2024, p. 820-823. doi:10.1109/piere62470.2024.10804967.

[23] Nahshan Y, Chmiel B, Baskin C, Zheltonozhskii E, Banner R, Bronstein AM, Mendelson A. Loss aware post-training quantization. Machine Learning 2021;110(11):3245-3262. doi:10.1007/s10994-021-06053-z.