AUTOMATED QUALITY ASSESSMENT OF APPLES USING CONVOLUTIONAL NEURAL NETWORKS
/
EVALUAREA AUTOMATĂ A CALITĂȚII MERELOR CU AJUTORUL REȚELELOR NEURONALE CONVOLUTIVE

Adrian IOSIF [1], Edmond MAICAN[1], Sorin BIRIȘ[1], Lucretia POPA[2]
[1] Faculty of Biotechnical Systems Engineering, U.N.S.T POLITEHNICA of Bucharest / Romania
[2] INMA Bucharest / Romania

E-mail: aiosif@yahoo.com
DOI: https://doi.org/10.35633/inmateh-71-42

Keywords: *apple quality assessment, convolutional neural networks, automated fruit sorting, image classification, deep learning applications in agriculture*

ABSTRACT
Quality assessment of apples is a pivotal task in the agriculture and food industries, with direct implications for economic gains and consumer satisfaction. Traditional methods, whether manual, mechanical or electromechanical, face challenges in terms of labor intensity, speed, and quality control. This paper introduces a solution using machine learning algorithms – specifically, Convolutional Neural Networks (CNNs) – for a more nuanced and efficient apple quality assessment. Our approach offers a balance between the high-speed capabilities of electromechanical sorting and the detailed recognition achievable with human evaluation. A dataset consisting of over 2000 apple images, labeled as 'Good' or 'Damaged', was compiled for training and validation purposes. The paper investigates various architectures and hyperparameter settings for several CNN models to optimize performance metrics, such as accuracy, precision, and recall. Preliminary evaluations indicate that the MobileNet and Inception models yield the highest levels of accuracy, emphasizing the potential of machine learning algorithms to significantly enhance apple quality assessment processes. Such improvements can lead to greater efficiency, reduced labor costs, and more rigorous quality control measures.

REZUMAT
Evaluarea calității merelor este o sarcină esențială în industriile agricole și alimentare, cu implicații directe asupra câștigurilor economice și a satisfacției consumatorilor. Metodele tradiționale, fie manuale, mecanice sau electromecanice se confruntă cu provocări în ceea ce privește efortul considerabil de muncă, viteza și controlul calității. Această lucrare propune o soluție care utilizează algoritmi de învățare automată – mai exact, rețele neuronale convolutive (CNN) – pentru o evaluare mai fină și mai eficientă a calității merelor. Abordarea noastră oferă un echilibru între capacitățile de sortare cu viteză mare ale metodelor electromecanice și recunoașterea detaliată realizabilă cu evaluarea umană. A fost compilat în scopuri de antrenare și validare un set de date compus din peste 2000 de imagini cu mere, fiecare măr fiind etichetat ca 'Bun' sau 'Stricat'. Lucrarea investighează diverse arhitecturi și configurări ale hiperparametrilor pentru mai multe modele de CNN în scopul optimizării indicatorilor de performanță (acuratețea, precizia, recall-ul). Evaluările preliminare indică faptul că modelele MobileNet și Inception oferă cele mai înalte niveluri de acuratețe, subliniind potențialul algoritmilor de învățare automată de a îmbunătăți semnificativ procesele de evaluare a calității merelor. Astfel de îmbunătățiri pot conduce la o eficiență mai mare, reducerea costurilor de muncă și tehnici de control al calității mai riguroase.

INTRODUCTION

The agriculture and food industries are among the most critical sectors not only for economic stability but also for ensuring human health and safety. One of the key steps in the food supply chain is the assessment of product quality, especially for fruits like apples, which constitute a significant portion of global fruit consumption. Quality assessment serves multiple purposes, such as grading for market pricing, segregating products for different usage scenarios (e.g., fresh consumption, juice extraction, or processed foods), and ensuring the overall safety and quality of the food. It is a complex but essential activity that is usually the intersection of human expertise, mechanical sorting technologies, and now, emerging computational methods.

In the context of automated quality assessment of apples, it is crucial to recognize the range of issues that can impact apple quality. Among parasitic diseases, Gray Mold, caused by the fungus Botrytis spp., manifests as a grayish fuzzy mold and can occur both in the field and during storage. Blue Mold, attributed to Penicillium expansum Thom, results in soft, water-soaked lesions covered with blue-green spores, commonly manifesting post-harvest. Lenticel Rot, due to the fungus Pezicula malicorticis Jacks Naum., affects the lenticels of apples, leading to dark lesions and a degradation in fruit quality (*Jamwal et al., 2002*).

Equally significant are non-parasitic diseases and damages. Superficial Scald is a physiological disorder that causes brown or dark patches on the apple's skin during long-term cold storage. Jonathan Spot appears as small, dark spots on the fruit's skin and is often associated with mineral imbalances or environmental stress factors. Internal Browning is another post-harvest disorder where the internal tissues of the apple turn brown, often due to storage conditions. Soft Scald is similar to Superficial Scald but results in soft, water-soaked lesions on the skin. Lastly, Soggy Breakdown makes the internal tissue of the apple turn water-soaked and spongy, often as a result of improper storage conditions (*Srivastava et al., 2021*). By understanding these parasitic and non-parasitic diseases and damages, the accuracy and reliability of automated quality assessment systems in categorizing and grading apples can be enhanced (*Nataraj et al., 2018*).

Despite its critical importance, the traditional methods employed for apple quality assessment have significant limitations. Manual sorting and grading are labor-intensive, time-consuming, and prone to human error. On the other hand, mechanical methods, although faster, often lack the nuanced understanding of 'quality,' as they primarily rely on size and weight as the determinant factors. These methods may overlook other essential quality metrics such as skin defects, color uniformity, and internal qualities that are important for grading and consumer satisfaction.

Recent advances in computer vision and machine learning have opened new avenues for automating quality assessment tasks. Convolutional Neural Networks (CNNs) have shown particular promise in image recognition challenges, extending their applicability to the agriculture sector (*Li et al., 2021*). Existing research has explored CNN-based approaches for defect detection in various fruits, vegetable classification, and even crop disease prediction. However, most of these studies have either focused on different fruits or have been limited to smaller datasets. Commercial solutions have begun to integrate machine learning but are often constrained by proprietary algorithms and high operational costs.

In the realm of automated apple sorting, *Kavdir and Guyer (2002*) utilized backpropagation neural networks (BPNNs) for sorting Empire and Golden Delicious apples based on surface quality. They employed both pixel gray values and texture features derived from apple images as inputs to their neural network classifiers. Their study explored two distinct classification scenarios: a 2-class classification, separating apples as either 'defective' or 'good,' and a 5-class classification that included various sub-categories of defects. Interestingly, they found that reducing the image resolution did not adversely affect classification accuracy, thereby offering a faster training and testing phase. Moreover, spectral bands were identified as effective indicators for distinguishing specific surface characteristics, such as bruising, leaf roller defects, and puncture marks. In terms of performance, the 2-class classifier achieved a classification success ranging from 89.2% to 100%. For the 5-class scenario, classification success varied between 89.7% and 100%, depending on the apple variety and the features used. Importantly, the research indicated that BPNN classifiers generally outperformed other methods when using pixel intensity as features, as opposed to extracted texture features. They concluded that neural networks could effectively capture the non-linear relationships between input features and output classes, and that spectral bands beyond 1000 nm improved defect identification in specific apple varieties. Despite the publication date of 2002, this work remains significantly relevant and pioneering in the field of using neural networks for apple sorting. Even in today's context, the findings and methodologies of Kavdir and Guyer stand out for their innovative approach in applying neural networks to agricultural tasks. Their use of BPNNs in this area laid an important work for subsequent research and developments, and the results they achieved continue to be pertinent in contemporary applications.

Yu et al. (2023) investigates the potential of Convolutional Neural Networks (CNNs) for the task of apple variety classification. Using a total of 7,439 apple images that represent 13 different apple classes, the study employs various deep learning architectures, namely AlexNet, VGG-19, ResNet-18, ResNet-50, and ResNet-101. The study employed two different dataset configurations to test the models, revealing that dataset balance plays a crucial role in classification performance. Specifically, all tested models achieved an accuracy above 96.1% when trained on a dataset with a training-to-testing ratio of 2.4:1, as compared to 89.4–93.9%

accuracy on a dataset with a ratio of 1:1. The VGG-19 model stood out, achieving a perfect 100% accuracy on the first dataset and 93.9% on the second. The study further delves into the impact of network architecture and depth on model size, accuracy, and computational time. It was found that as the number of layers in a model increased, so did its size, accuracy, and the time required for training and testing. Additionally, the authors employed techniques like feature visualization, strongest activations, and Local Interpretable Model-Agnostic Explanations (LIME) to interpret how different models understand and classify apple images. It was discovered that while series networks like AlexNet and VGG-19 focus on apple contours or shapes for classification, Directed Acyclic Graph (DAG) networks like ResNets focus on the entire apple region. The work not only confirms the applicability of CNNs in apple recognition but also contributes to the understanding of their interpretability, thereby providing a foundation for future agricultural applications of deep learning. This article is well-aligned with the broader scope of present research, as it similarly investigates the application of Convolutional Neural Networks (CNNs) for fruit classification, focusing particularly on apple varieties. The insights from this study provide a foundational understanding for the current research, particularly regarding the influence of dataset configuration and model architecture on classification performance. Acknowledging the shortcomings of existing methods, the present paper aims to explore the feasibility of machine learning algorithms, specifically Convolutional Neural Networks (CNNs), for an automated, efficient, and nuanced apple quality assessment. A thorough evaluation of various CNN architectures is conducted to ascertain the most effective in terms of accuracy, precision, and recall metrics. Additionally, this research introduces a custom CNN architecture, uniquely designed for apple sorting, and compares its performance with established models.

Wan and Goudos (2020) offer a refined framework that uses an advanced version of the Faster R-CNN algorithm, specifically designed for multi-class fruit identification tasks. Distinctive in their approach is the development of an extensive image dataset gathered from real-world outdoor orchards, featuring 4,000 images of fruits such as apples, mangoes, and oranges. To augment this dataset and improve model generalizability, they apply various data augmentation strategies. Their architectural enhancements, particularly the alterations made to the convolutional and pooling layers of the Faster R-CNN model, serve dual objectives: they accelerate the fruit detection process and enhance accuracy. Benchmarking studies corroborate the superiority of their method, showing that it surpasses existing models like YOLO, YOLOv2, YOLOv3 and Fast R-CNN in both speed and detection precision. Specifically, the modified Faster R-CNN model significantly outperforms these alternatives and demonstrates a mean Average Precision (mAP) exceeding 91% for multiple fruit types. This work represents a significant stride forward in the automation of agricultural processes. Its practical relevance lies in its potential applications for robotic harvesters, providing them with the capability to detect multiple types of fruits with high accuracy and speed.

Li et al. (2021) present a Convolutional Neural Network (CNN) model specialized for apple quality identification. The model successfully navigates the complexities associated with apple images, particularly when the background resembles the apple's surface. The authors compared their CNN-based approach with Google's Inception v3 and traditional image processing techniques that use features such as Histogram of Oriented Gradient (HOG) and Gray Level Co-occurrence Matrix (GLCM), paired with a Support Vector Machine (SVM) classifier. The CNN model showcased unparalleled training and validation accuracies, peaking at 99% and 98.98% respectively. Further, the model surpassed its competition in an independent test set, demonstrating an accuracy rate of 95.33%. Additionally, the CNN model proved to be more time-efficient, completing its training in just 27 minutes. The study reveals the potential of the proposed CNN model in complex apple quality assessment scenarios, outperforming existing models in both speed and accuracy. The authors plan to extend this model to gauge various apple attributes like color, size, ripeness, and even physiological disorders. They also envision expanding its utility to classify other fruits and integrate it into real-time sorting machinery.

Liu (2020) focuses on the role of deep learning in fruit classification, which is crucial for automating self-checkout and packaging systems in supermarkets. The study introduces a deep convolutional neural network model, called Interfruit, designed to handle the complexities inherent in fruit classification, such as category similarities. A unique dataset of 40 fruit categories was developed to train and test the Interfruit model. The study also uses an improved stack model that integrates the strengths of AlexNet, ResNet, and Inception. Interfruit achieved an overall test accuracy of 92.74%, surpassing other leading models in the domain. The model obviates the need for manual feature extraction and employs various network parameters and data augmentation strategies to bolster its predictive capabilities. Such high performance indicates the robustness and technical validity of the Interfruit model. The study concludes that Interfruit offers a promising and

comprehensive solution for automatically identifying and classifying fruits in supermarkets, aiding in quick retrieval of pricing and other identification information.

Zhang et al (2019) present an innovative paper that addresses the critical need for effective and efficient fruit detection in agricultural robots. Their approach leverages an advanced multi-task cascaded convolutional network (MTCNN) for high-accuracy, real-time performance. One of the contributions of this paper is the introduction of an improved image augmentation method, known as "fusion augmentation," designed to further elevate the performance of the fruit detector. To validate their model, the authors created a robust dataset that includes images from apple orchards as well as additional images from the Internet and ImageNet. The system was not only successful in detecting apples but also showed promising results when extended to other fruits like strawberries and oranges. In terms of efficiency, the system processed 100 images in under 80 seconds, closely approximating real-time performance. Given these results, this study's fruit detection system holds significant promise for broader applications, particularly in automating various agricultural tasks like sorting, grading, and yield estimation.

Keresztes et al. (2018) explore real-time fruit detection for proximal imaging, typically done using tractor-mounted cameras in orchards and vineyards. Their method is a two-step process that combines geometrical pre-processing with deep neural network (DNN) classification. The geometric pre-processing step, which employs a radial Hough-like operator, quickly narrows down probable regions where fruits may be located, making the subsequent DNN classification more efficient. The system was rigorously tested on grapes and apples, with encouraging outcomes that show high correlations with manual counting methods — up to 0.96 for grapes and 0.85 for apples. Moreover, the technology leverages an intelligent camera system, which is both cost-effective and easily adaptable to existing agricultural machinery like tractors. The camera system allows for the acquisition of large datasets while maintaining controlled lighting conditions for consistent fruit detection. This approach not only proves its efficacy in fruit detection but also offers valuable insights into agronomic parameters that could be crucial for advancing precision agriculture. Thus, the method holds potential not just for improving current farm management strategies but also for contributing to the development of more sophisticated decision-support tools in the future.

The Table 1 offers a summarized overview of the state-of-the-art research in the field of automated fruit sorting and classification, with a specific focus on apple sorting and quality assessment. This area of study has seen significant advancements, thanks to machine learning algorithms, particularly Convolutional Neural Networks (CNNs), and other data-driven methodologies. The table categorizes research papers by their authors and key contributions, outlines the methods and algorithms employed, and encapsulates their performance metrics or findings.

Another contribution to the field of automated fruit grading is the review of *Seema et al. (2015).* Their paper delves into the critical role of computer vision in agricultural automation, specifically focusing on fruit grading and sorting. They argue that while traditional human-operated methods are prone to errors and inefficiencies, computer-vision systems offer a more accurate and efficient alternative. The review identifies that the most commonly employed features for computer vision-based sorting are color, texture, and morphological characteristics. These features are typically used to assess factors like diseases, maturity, and overall quality of the fruit. Among the various machine learning techniques explored, Support Vector Machines (SVMs) were found to deliver high accuracy rates. However, Adaptive Neuro Fuzzy Interference Systems (ANFIS) provided the best overall performance. On the color modeling front, the HIS (Hue, Saturation, Intensity) model was highlighted as particularly effective due to its alignment with human perception. The review concludes that machine vision systems hold the potential to significantly improve the efficiency and accuracy of fruit grading, thereby making a strong case for their broader adoption in the agricultural sector.

Table 1

Overview of the state-of-the-art research in the field of automated fruit sorting and classification

Paper	Methods and Algorithms Used	Key Contributions	Performance and Findings
Kavdir and Guyer	Backpropagation Neural Networks (BPNNs)	Surface quality classification into 2 and 5 classes. Explores pixel gray values and texture features.	89.2% to 100% classification success depending on apple variety and features.
Fanqianhui Yu, Tao Lu, Changhu Xue	CNNs (AlexNet, VGG-19, ResNet-18, ResNet-50, ResNet-101)	Focuses on apple variety using deep learning. Explores dataset configurations and model architectures.	Accuracy > 96.1% with a balanced dataset. VGG-19 achieved 100% accuracy.

Paper	Methods and Algorithms Used	Key Contributions	Performance and Findings
Shaohua Wan and Sotirios Goudos	Modified Faster R-CNN	Extensive real-world dataset and architectural enhancements for multi-class fruit detection.	mAP > 91% for multiple fruit types, outperforming YOLO and Fast R-CNN.
Li, Feng, Liu, Han	CNNs	Specialized for apple quality. Compares CNN-based approach with Inception v3 and traditional methods.	Training and validation accuracies peaked at 99% and 98.98%.
Wenzhong Liu	Interfruit (Deep CNN)	Designed for self-checkout and packaging systems. Uses a unique dataset of 40 fruit categories.	Test accuracy of 92.74%, outperforming leading models.
Li Zhang et al.	Multi-Task Cascaded Convolutional Networks (MTCNN)	High-accuracy, real-time performance for agricultural robots. Introduces "fusion augmentation."	Processes 100 images in < 80 seconds, showing real-time performance.
Barna Keresztes et al.	Radial Hough-like operator + DNNs	Two-step process for real-time fruit detection using tractor-mounted cameras.	High correlation with manual counting: up to 0.96 for grapes and 0.85 for apples.

MATERIALS AND METHODS

The prevailing trend in automated fruit sorting research leans toward intricate classification schemes that involve complex feature sets and multiple quality metrics. While such approaches have merit, they often bypass a critical, foundational decision point in the agricultural supply chain: the initial binary classification of fruit as either acceptable or subpar for consumption or sale. The importance of this primary bifurcation cannot be overstated, as it serves as an essential presorting step for more granular quality evaluations, packaging, distribution, and other downstream processes.

The integration of a binary classification system as a presorting stage prior to multi-class classification offers a multitude of advantages. One of the immediate benefits is the quick filtration of subpar produce, enabling only quality apples to proceed to the more resource-intensive multi-class classification stage, thereby enhancing its accuracy and effectiveness. This method is computationally efficient, as it ensures that the greater computational resources required for multi-class classification are expended only on apples that have already passed a basic quality threshold. This initial classification also addresses the issue of class imbalance, which can otherwise disproportionately skew the performance of multi-class classifiers. The binary filter removes low-quality apples from the outset, mitigating this problem. Moreover, multi-class classifiers frequently encounter boundary issues that a preliminary binary sort can effectively alleviate by widening the distinctions between 'damaged' and 'good' classes, thereby improving the efficiency and accuracy of the multi-class system. In addition, this binary presorting technique enables a modular approach to fruit grading. It can function as a stand-alone unit that feeds into more specialized multi-class systems, offering greater flexibility in deployment and scalability. Training machine learning models for multi-class classification often takes longer; however, this initial binary sort streamlines the process by reducing the data set size and complexity, potentially accelerating training times. Furthermore, the simplicity of a binary model lowers the risk of overfitting compared to more complex multi-class models, serving as a regularization technique that enhances the generalization performance of subsequent models. Troubleshooting is also simplified; when errors occur, it is easier to determine whether they originated during the binary or multi-class classification stage. Finally, this presorting stage yields financial benefits by reducing both computational and labor costs, as fewer apples need to pass through the resource-intensive multi-class classification process. Overall, the use of a binary presorting stage substantially optimizes the apple grading process, rendering subsequent multi-class classifications more efficient, accurate, and cost-effective.

The current study addresses this often-overlooked area by introducing a robust method for this initial categorization. By establishing a reliable and efficient binary classification system, the research lays the groundwork for subsequent, more nuanced quality assessments. This initial filtration stage can have substantial implications for the agricultural supply chain, including the potential for reduced labor costs and minimized error rates in later stages of quality assessment.

The dataset initially comprises a total of 2468 apple images, all of which were personally captured in a single apple orchard. To improve the dataset's robustness and diversity, the original set of 2468 images was augmented, expanding the total number of images to 4200. The advantage of sourcing all images from the same area lies in the consistency it brings to the data. The dataset was collected during the peak harvest season to ensure representative sampling. Variability due to different lighting conditions, apple varieties, and background interference is minimized, thereby allowing the model to learn more efficiently and accurately. The

images are divided into two categories: ”Good” apples and ”Damaged” apples. These categories were created by manually inspecting each apple, ensuring a reliable and consistent labeling process.

In the orchard from which the apples used in this dataset were harvested, various cultivars were present, each exhibiting unique characteristics, most notably in the color of their skin when fully matured. These cultivars included the Golden Delicious, the Jonathan, and the ”Rabbit's Snout” (”Bot de iepure” in Romanian). The aim of this study was not to discriminate between these cultivars, but rather to conduct binary classification without taking skin color into account. While the inclusion of multiple cultivars with different mature skin colors undoubtedly complicates the classification task and may inhibit achieving high accuracy, this approach was intentionally chosen. The objective was to evaluate the efficacy of an apple sorting process that is cultivar-agnostic and does not rely on the specific hue of the skin at maturity for classification. This sets the groundwork for a more universally applicable sorting algorithm that can handle apples from diversified sources.

For easier data processing and model training, the images are organized into two separate folders corresponding to these categories. The original resolution of each image is 3648x2736 pixels. However, for computational efficiency and to facilitate faster training, these images were resized to a uniform shape of 150x150 pixels with 3 color channels (RGB). The input shape for the Convolutional Neural Network was set to (150, 150, 3), corresponding to the height, width, and the number of color channels of the resized images. An exception was the MobileNet network, for which the input shape was 224x224x3. The choice of image size is a trade-off between computational efficiency and model performance. Smaller images like 128x128 are faster to process and require less memory, but they might lack detail that could be important for classification. Larger images retain more detail but are computationally more expensive to process and may require more memory. With a larger number of pixels, the model might learn noise or insignificant variations in the data, rather than truly relevant features. This can lead to overfitting, especially if the training dataset is relatively small. During training, a tendency towards overfitting was observed, which led to the decision to use the 150x150 format, except for the MobileNet network (*Simonyan and Zisserman, 2015; Pen at al., 2023; Ke et al., 2023; Abdo et al., 2023*).

In our dataset, a class distribution where 55% of the apples fall under the ”Good” category and 45% are categorized as ”Damaged” can be observed. While the classes are not perfectly balanced, the distribution is relatively even, with only a 10% difference between the two categories. Such a slight imbalance is generally not substantial enough to significantly bias most machine learning algorithms. Given the near-balanced nature of our dataset, the question arises whether resampling methods are necessary or not. In cases with severe class imbalance, resampling is often recommended to create a more balanced dataset, thereby improving model performance. However, in this case, the 10% disparity between the classes is quite minimal, so the need for resampling methods like under-sampling or over-sampling becomes less pressing.

For a binary classification task like distinguishing between “Good” apples and “Damaged” apples, a 150x150 image size may be sufficient if the damage is usually large-scale and easily visible. However, if the types of damage are subtle or require high-resolution to spot, using larger images could be beneficial. Optimal image size in convolutional neural networks is influenced by computational resources, the subtlety of features in the data, risks of overfitting with smaller datasets (sometimes using larger images can lead to overfitting, especially if the dataset is small), and compatibility requirements with advanced architectures. Empirical determination of the most appropriate image size for this specific classification problem was achieved through training the convolutional neural network models on multiple image dimensions. Through experimentation, aided by available computational resources, an optimum image size was identified for yielding the best performance metrics for this particular application (Simonyan and Zisserman, 2015). The dataset employed for this study exhibits a well-balanced distribution between the two classes, “Good” and “Damaged” apples. This balance enhances the reliability of the model's predictions and mitigates the risk of class bias, thereby improving the generalizability of the findings.

Figure 1 offers selected samples from the dataset utilized in this study, showcasing the marked visual differences between the “Good” apples and the “Damaged” apples. The depicted apples belong to two distinct varieties, underlining the study's diversity. The first two images present apples from the “Good” apples category. These samples demonstrate a consistent color distribution, smooth skin texture, and an absence of discernible defects. In contrast, the following two images represent the “Damaged” apples category. Both of these apples display characteristics of a common fungal infection, evident through dark spots and altered skin texture. This specific infection is prevalent in the dataset. It is worth noting that the dataset did not feature apples with noticeable insect damage.

Fig. 1 - Samples from the dataset

These images serve to highlight the key distinguishing features that the convolutional neural network model aims to learn for accurate classification.

Choosing the optimal convolutional network architecture for evaluating apple quality depends on a variety of elements, such as the complexity of the visual cues in the dataset, the computational resources that can be allocated, and the urgency for real-time inference (*Ke et al., 2023; Abdo et al., 2023; Theng et al., 2023*).

LeNet, one of the foundational architectures in Convolutional Neural Networks, was introduced by Yann LeCun in the 1990s and played a pivotal role in the proliferation of deep learning in image recognition tasks. Its architecture is designed with a series of convolutional and subsampling layers followed by fully connected layers, setting the stage for many subsequent CNN designs (*Simonyan and Zisserman, 2015*). It is particularly well-suited for straightforward visual tasks. Given its design simplicity and relatively fewer parameters compared to modern networks, LeNet can be efficient and less prone to overfitting on smaller datasets. Therefore, it could be an excellent fit if the apple dataset does not possess overly complex features (see mode used in fig. 2).

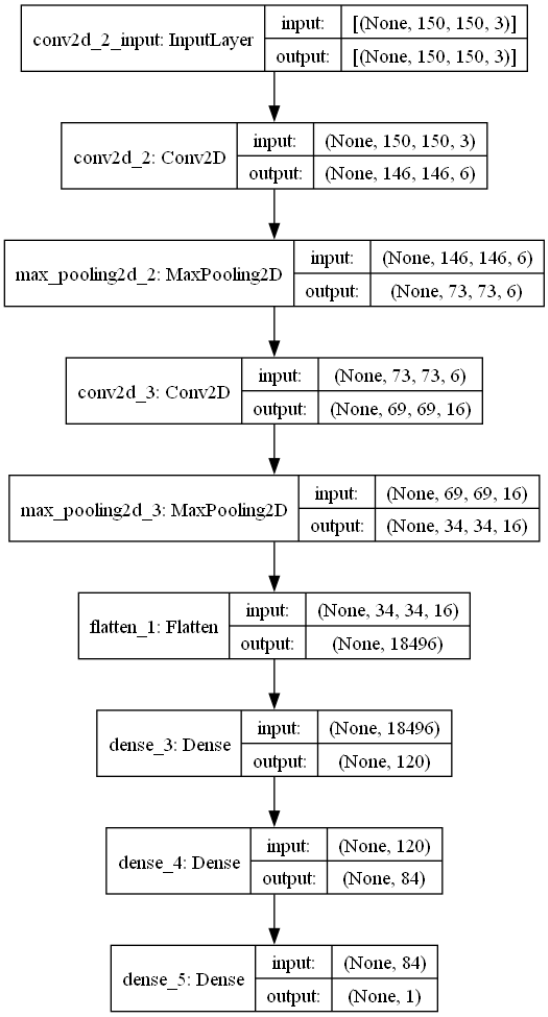

Fig. 2 – LeNet Implementation

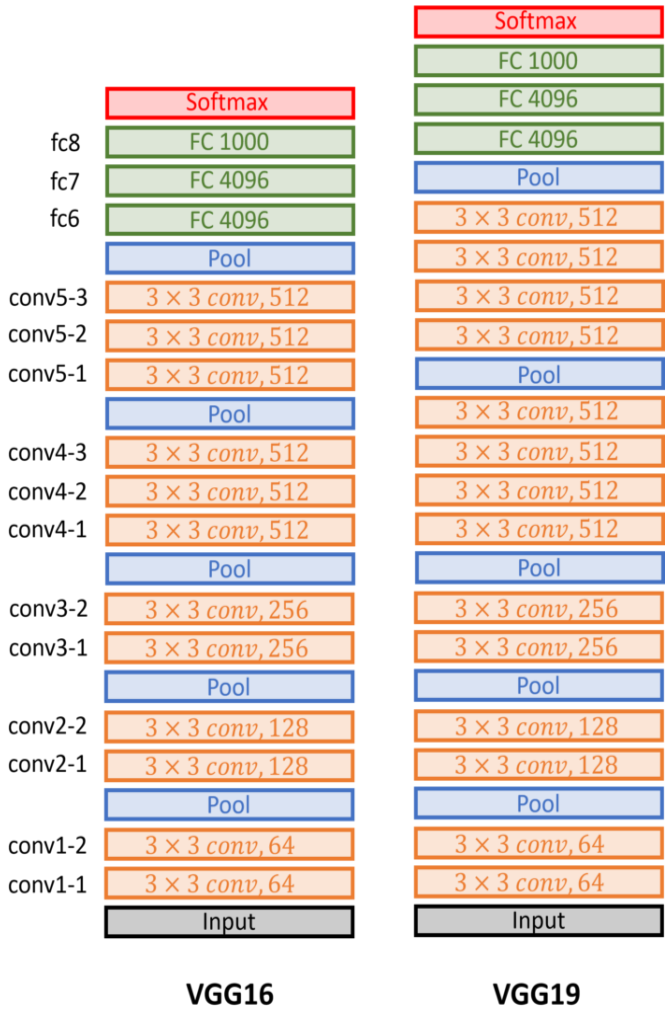

Fig. 3 - VGG16 and VGG19 Implementation

The VGG (Visual Geometry Group) architectures, specifically VGG16 and VGG19, are convolutional neural networks developed by the Visual Geometry Group at the University of Oxford. They were designed to be simple yet highly effective for a wide range of image recognition tasks. The primary attribute that distinguishes VGG architectures is their depth—specifically, the number of weight layers in the network. VGG16 and VGG19 have 16 and 19 weight layers, respectively. VGG16 consists of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. Each convolutional layer employs a small receptive field using 3x3 kernels, which is a key innovation contributing to the model's effectiveness. Max-Pooling layers are interspersed among the convolutional layers and utilize 2x2 kernels with a stride of 2 to reduce the spatial dimensions. Three Fully Connected layers are used at the end, the first two having 4096 channels and the final one having 1000 channels corresponding to the number of classes in the ImageNet dataset. ReLU (Rectified Linear Unit) is used as the activation function throughout the model (fig. 3). VGG19 is very similar to VGG16 but has 3 additional convolutional layers, increasing the depth to 19 weight layers. This makes VGG19 slightly more complex and computationally intensive but also offers a modest improvement in performance. Like VGG16, it also uses 3x3 convolutional kernels, max-pooling layers with 2x2 kernels and stride of 2, and fully connected layers toward the end of the architecture *(Nguyen et al., 2022; Ong et al., 2023; Anuar et al., 2023; Abdo et al., 2023*).

The Inception architecture, particularly its third version, InceptionV3, is a product of Google's research and has been one of the standout models for image classification tasks. Built on the foundational idea of 'network within a network', InceptionV3 employs multiple kernel sizes in parallel, rather than in series, to capture various spatial hierarchies of features within an image (fig. 4). This unique design choice facilitates the extraction of both local features using smaller convolutions and more global features with larger convolutions. Furthermore, the model incorporates techniques like factorization and efficient grid size reduction to keep the computational cost manageable. These sophisticated configurations ensure that the architecture learns a wide array of features at different scales. As such, InceptionV3 offers a versatile and robust approach, making it particularly apt for the multi-feature quality assessment in apples, where varied scales of features may be crucial (*Ong et al., 2023; Abdo et al., 2023*).

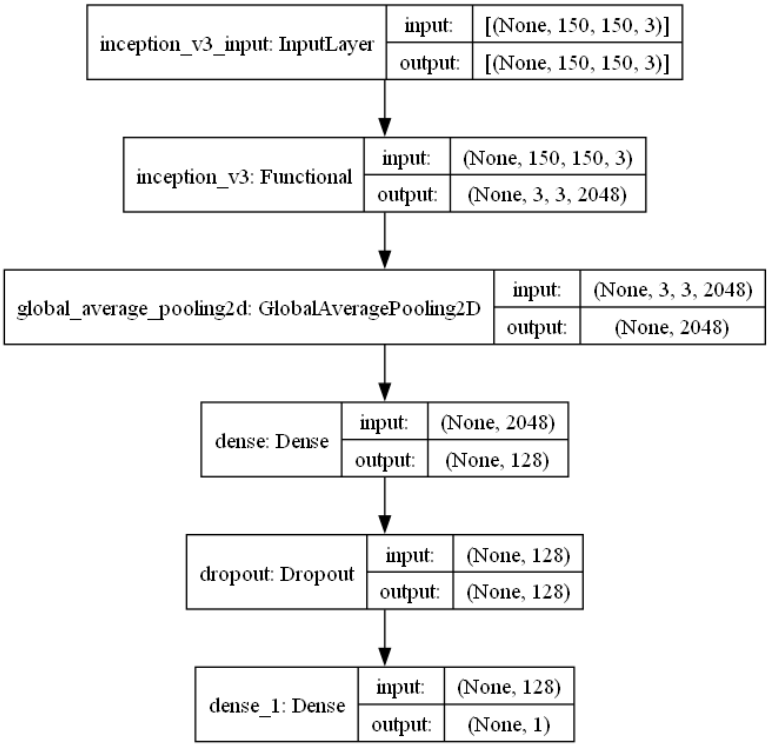

Fig. 4 - Inception Implementation

MobileNet, developed by Google, is specifically designed for mobile and embedded vision applications, striking a balance between computational efficiency and model performance. Utilizing depthwise separable convolutions, it factors standard convolutions into depthwise and pointwise operations, thereby significantly reducing the number of parameters and computational overhead without sacrificing the ability to capture meaningful features.

This design choice enables MobileNet to run efficiently even on devices with limited computational resources. Its architecture is not only compact but also modular, allowing for varying levels of granularity based on the application's requirements. If the computational budget is limited or rapid inference is required, MobileNet emerges as the prime choice. Its lightweight and fast processing characteristics make it suitable for real-time applications or environments where latency is a concern (*Khazalah et al., 2023*). However, like many efficient models, there might be a trade-off in accuracy when compared to larger, more resource-intensive networks (fig. 5).

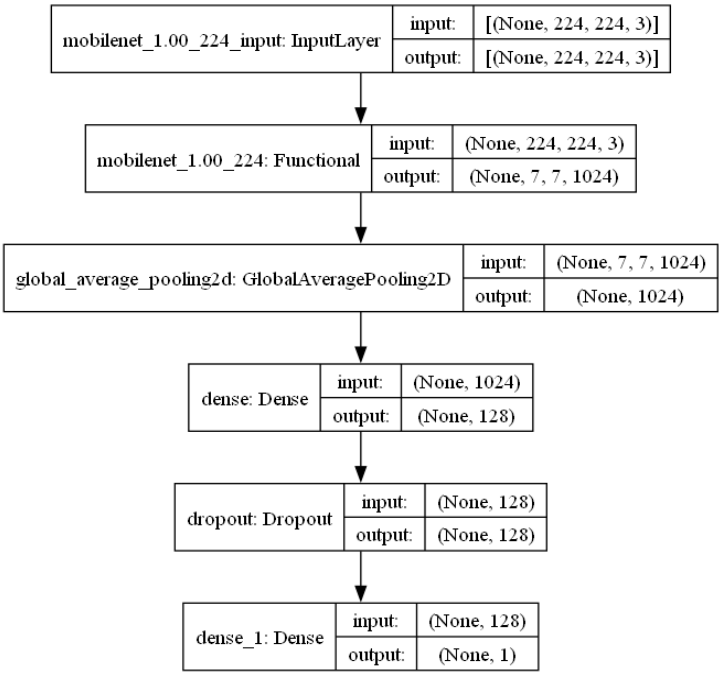

Fig. 5 - MobileNet Implementation

EfficientNet, designed by Google researchers, rethinks the way network scaling is done by proportionally adjusting depth, width, and resolution. This unique, coordinated scaling ensures improved performance with fewer parameters (*Tan and Le, 2019*). Born from an optimization process, its base model sets a foundation, which can be expanded based on computational needs. EfficientNet balances computational efficiency with performance, making it a prime choice in scenarios where both factors are vital (figure 6).

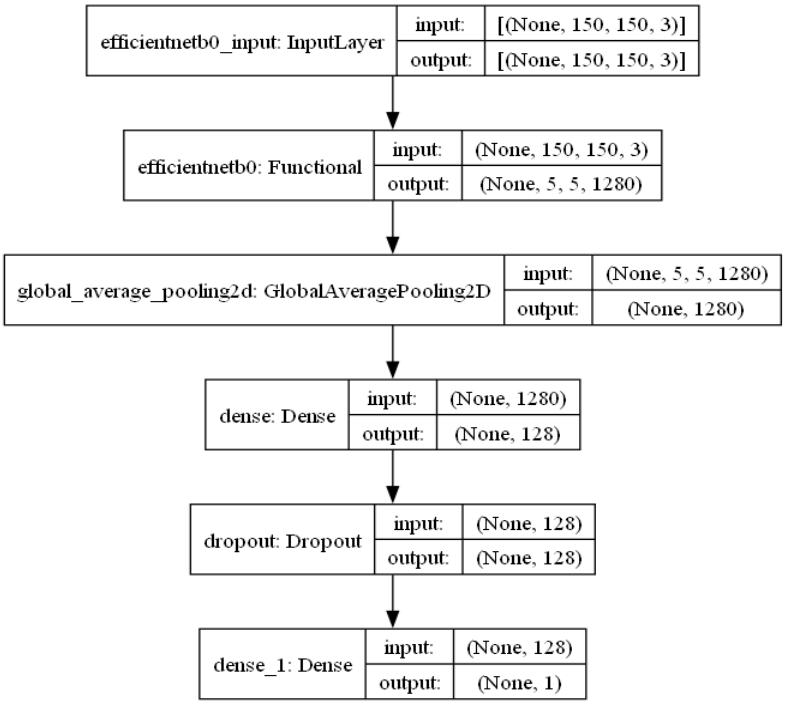

Fig. 6 - EfficientNet Implementation

For the unique demands of apple quality assessment, a custom CNN architecture may offer the most direct approach to achieving good performance (table 2). For example, a heightened emphasis on texture differentiation might necessitate a greater number of convolutional layers, while color-sensitive assessments could benefit from an increased number of filters in the initial convolutional layers (*Ong et al., 2023; Ke et al., 2023; Abdo et al., 2023; Anuar et al., 2023*).

All of these architectures were tested, as much as possible, under similar conditions, initially on the original dataset and then on the augmented dataset (batch size of 32, 30 training epochs, binary cross-entropy as the loss function, and the Adam optimizer with a learning rate of 0.0001).

Table 2

Custom CNN Model Implementation

Layer (type)	Output Shape	Param #
conv2d_6 (Conv2D)	(None, 148, 148, 32)	896
activation_10 (Activation)	(None, 148, 148, 32)	0
max_pooling2d_6 (MaxPooling2)	(None, 74, 74, 32)	0
conv2d_7 (Conv2D)	(None, 72, 72, 64)	18496
activation_11 (Activation)	(None, 72, 72, 64)	0
max_pooling2d_7 (MaxPooling2)	(None, 36, 36, 64)	0
conv2d_8 (Conv2D)	(None, 34, 34, 128)	73856
activation_12 (Activation)	(None, 34, 34, 128)	0
max_pooling2d_8 (MaxPooling2)	(None, 17, 17, 128)	0
flatten_2 (Flatten)	(None, 36992)	0
dense_4 (Dense)	(None, 64)	2367552
activation_13 (Activation)	(None, 64)	0
dropout_2 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 1)	65
activation_14 (Activation)	(None, 1)	0

Total params: 2,460,865

Trainable params: 2,460,865

The primary metric used for assessing the performance of the binary classification models in this study is Accuracy, defined as:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

or

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

where:

TP = true positive (the items that were correctly identified as belonging to the positive class),

TN = true negative (the items that were correctly identified as belonging to the negative class),

FP = false positive (the items that were incorrectly identified as belonging to the positive class when they actually belong to the negative class - in many contexts, this is known as a "type I error" or a "false alarm"),

FN = false negative (the items that were incorrectly identified as belonging to the negative class when they actually belong to the positive class - this is sometimes referred to as a "type II error" or a "miss").

This metric is well-suited for binary classification tasks where the classes are approximately balanced. It gives us a simple, interpretable measure of the model's overall performance (*Seema et al., 2015*).

In addition to accuracy as the primary metric, the F1 score was also employed as a supplementary evaluation measure. For binary classification tasks such as ours, the F1 score is especially pertinent because it provides a balanced harmonic mean of precision and recall. This ensures that both false positives and false negatives are taken into account, making it a more comprehensive metric than accuracy alone, especially in situations where class imbalances exist (*Larner, 2023*). The F1 score is defined as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \quad (3)$$

where:

precision = $\frac{TP}{TP+FP}$ (the number of correct positive results divided by the number of all positive results), and

recall = $\frac{TP}{TP+FN}$ (the number of correct positive results divided by the number of positive results that should have been returned).

In this analysis, the Binary Cross-Entropy (BCE) loss function is utilized, which is particularly suited for binary classification tasks. Binary Cross-Entropy quantifies the difference between two probability distributions: the true distribution and the predicted distribution. It measures the "distance" between the ground truth and our predictions, aiming to minimize this distance as the model learns. The rationale behind employing BCE for binary classification lies in its ability to handle probabilistic predictions. When predicting the two classes, it is essential not only to classify but also to measure the confidence of the model in its prediction. Binary Cross-Entropy penalizes the model significantly when it is confident and wrong, and to a lesser extent when it is unsure. This makes it particularly effective in driving the model towards making more accurate predictions with higher confidence *(Larner, 2023*).

Mathematically, the Binary Cross-Entropy loss for a set of predictions p with respect to true labels y is given by:

$$BCE = -\sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i), \quad (4)$$

where:

n is the number of samples, y_i is the true label (0 or 1), p_i is the predicted probability of the sample belonging to class 1.

By optimizing this loss during the training process, the model endeavors to improve its binary classification performance, generating predictions that align closely with the ground truth.

To ensure robustness and generalization capabilities, it is essential to prevent overfitting in neural network models. Overfitting occurs when a model excels on the training data but struggles to generalize effectively to unseen or new data, capturing the noise and specificities of the training set rather than the general patterns. To address this challenge, L1 and L2 regularization techniques are applied across the convolutional and fully connected layers. These techniques introduce penalties to the loss function, acting as constraints that discourage the model from fitting too closely to the peculiarities of the training data.

The underlying principle of L1 and L2 regularization is the penalization of large values of model weights. While L1 regularization encourages sparsity by driving certain weights to zero, thereby leading to feature selection, L2 regularization aims to shrink weights towards zero without making them exactly zero, ensuring the values remain small and distributed.

Mathematically, the regularization term added to the loss function can be defined as:

$$Regularization\ Term = \lambda_1 \sum|w| + \lambda_{12} \sum w^2 \quad (5)$$

Here:

λ_1 and λ_2 represent the regularization coefficients for L1 and L2, respectively, and w signifies the model weights.

Incorporating these regularization terms in neural network training ensures a balance between adequately fitting the training data and retaining the capability to generalize to new instances. By incorporating these regularization terms, it is ensured that our models achieve a balance between fitting the training data and retaining their ability to generalize to new dataIn initial training sessions. Early Stopping was employed to halt training when the validation performance ceased to improve. It was observed that Early Stopping acted for all models up to the 30th epoch. As a result, subsequent training sessions were adjusted to have a fixed number of epochs set at 30, streamlining the training process and preventing overfitting.

The learning rate is adaptively adjusted during training using a "Reduce On Plateau" strategy. When the validation loss reaches a plateau, the learning rate is reduced by a factor, facilitating the model to escape local minima and potentially leading to better generalization (*Ke et al., 2023; Larner, 2023*).

RESULTS

The application of various Convolutional Neural Network (CNN) architectures for the binary classification of apples produced diverse outcomes in terms of accuracy and F1 score across both training and validation datasets (Table 3). The results elucidate the potential and limitations of each model with respect to their performance metrics.

Table 3

Performance Metrics of Various CNN Architectures (%)

Architecture	Train Accuracy	Train F1 Score	Validation Accuracy	Validation F1 Score
Custom CNN	0.91	0.91	0.92	0.90
LeNet	0.96	0.96	0.95	0.95
VGG16	0.91	0.91	0.91	0.91
VGG19	0.90	0.90	0.90	0.90
MobileNet	0.98	0.98	0.99	0.99
InceptionV3	0.98	0.98	0.98	0.99
EfficientNet	0.48	0.52	0.50	0.66

The following subsections provide a visual representation of the performance metrics for each CNN architecture (figures 7-13).

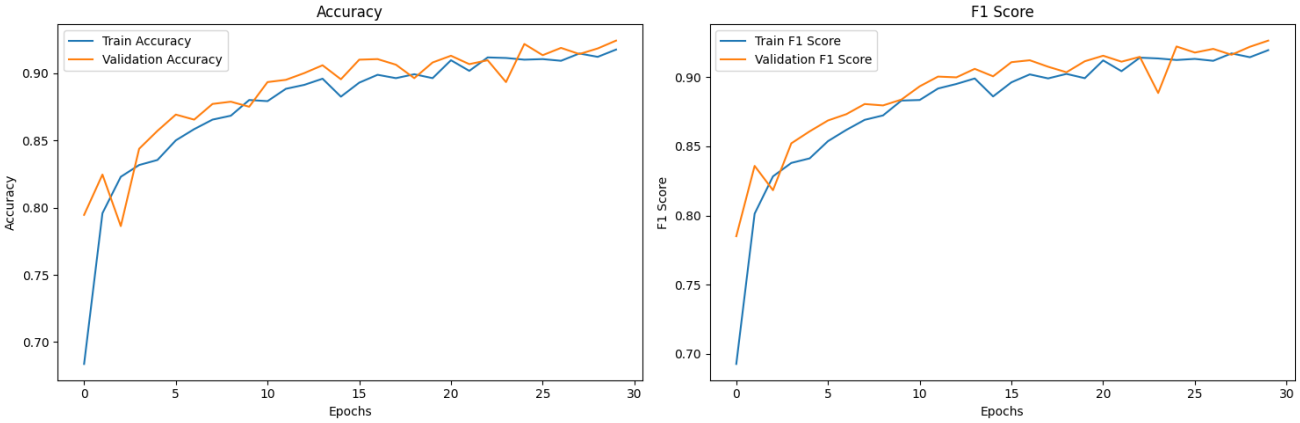

Fig. 7 - Custom CNN Architecture Performance Metrics: Accuracy and F1 Score

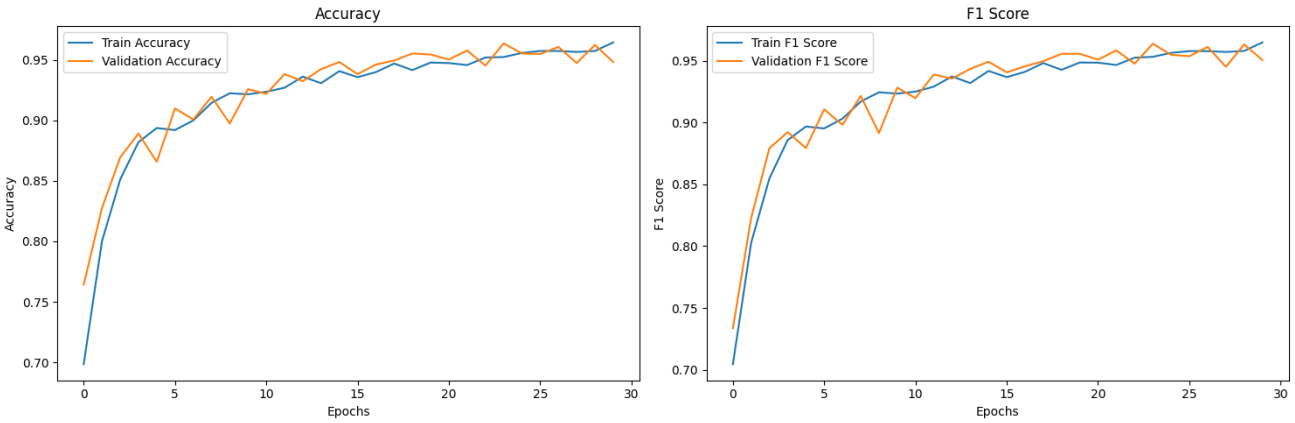

Fig. 8 – LeNet Architecture Performance Metrics: Accuracy and F1 Score

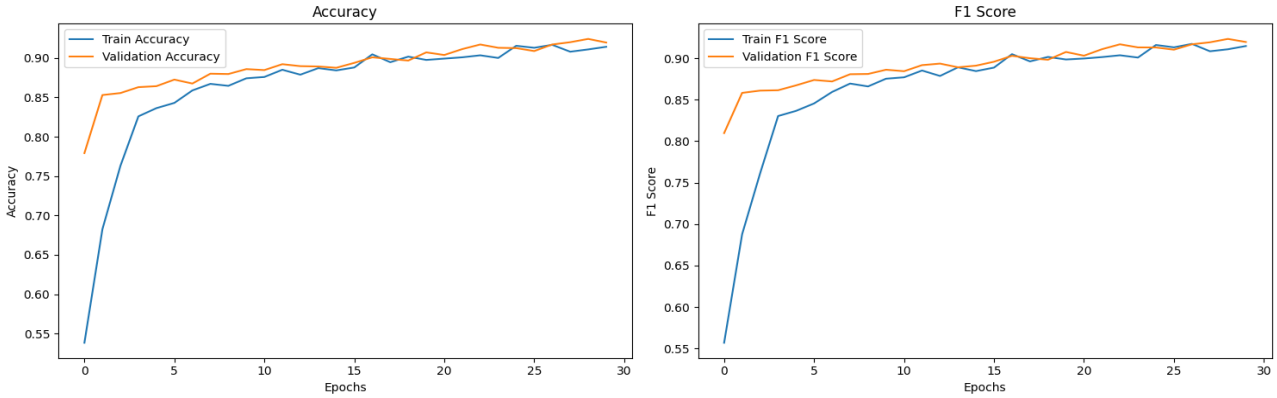

Fig. 9 - VGG16 Architecture Performance Metrics: Accuracy and F1 Score

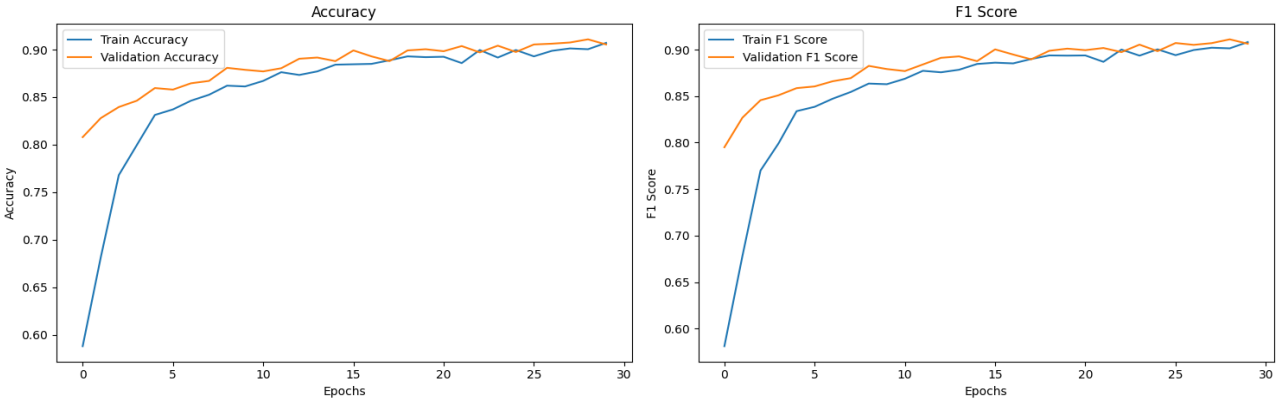

Fig. 10 - VGG19 Architecture Performance Metrics: Accuracy and F1 Score

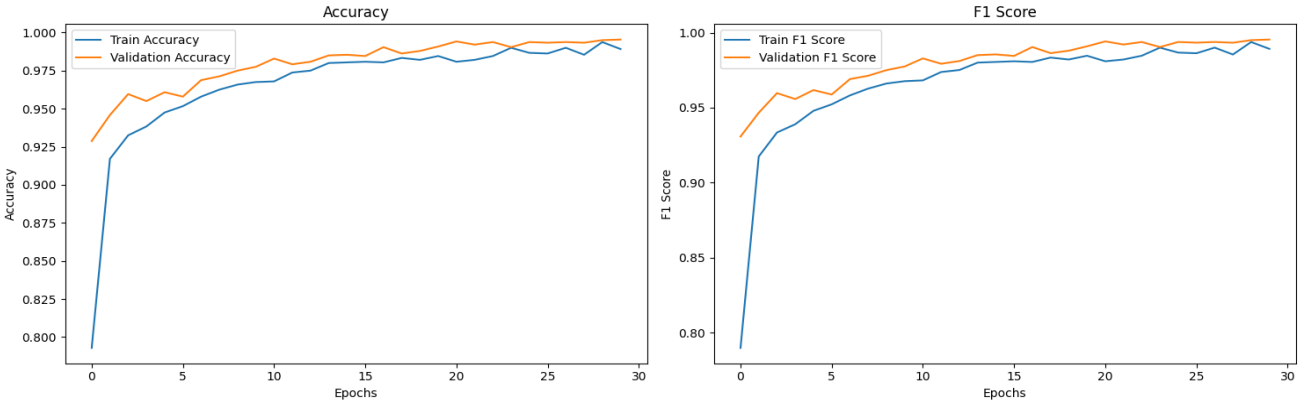

Fig. 11 – MobileNet Architecture Performance Metrics: Accuracy and F1 Score

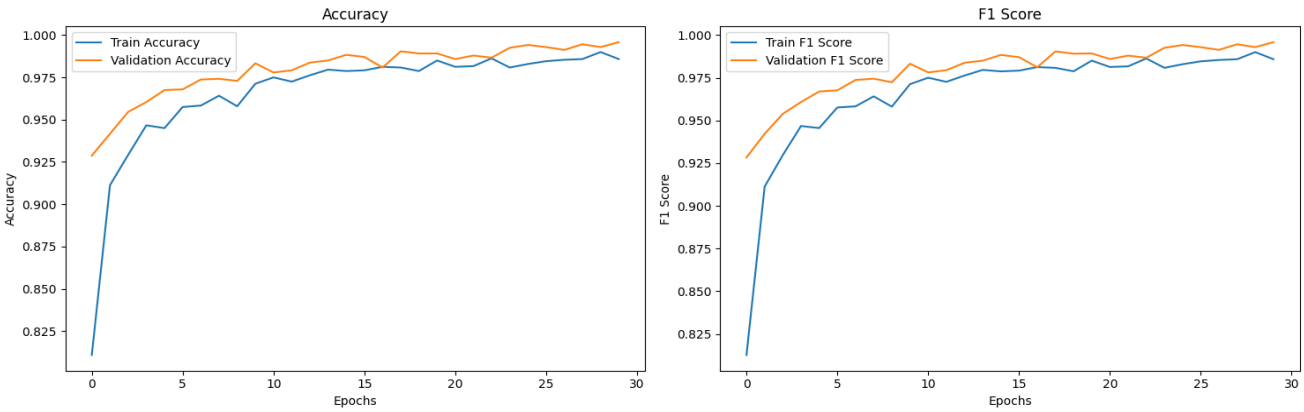

Fig. 12 - InceptionV3 Architecture Performance Metrics: Accuracy and F1 Score

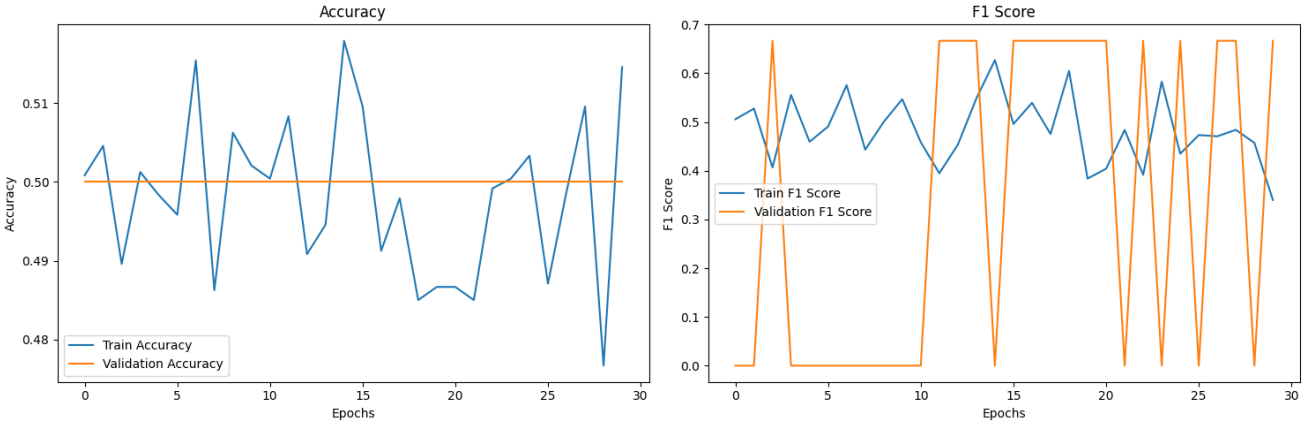

Fig. 13 – EfficientNet Architecture Performance Metrics: Accuracy and F1 Score

Each figure provides a visual trajectory of how the respective models performed across epochs, allowing for a comprehensive understanding of the models' learning patterns and stability. The results obtained provide a foundation upon which the subsequent conclusions are drawn, aiming to comprehend the practical applicability and limitations of the explored CNN architectures in the context of apple sorting through binary classification.

CONCLUSIONS

The empirical exploration of different Convolutional Neural Network (CNN) architectures has provided a basic understanding of their potential applicability in the binary classification of apples for sorting purposes.

Analyzing the data, it is evident that different architectures have offered various levels of performance, with MobileNet and InceptionV3 indicating a relatively high proficiency in this particular application, as opposed to the less consistent results derived from EfficientNet. The underperformance of EfficientNet might be attributed to various factors, such as potential misalignment between network configurations (like width, depth, and resolution scaling) and the particularities of the dataset, or possibly a deficiency in the volume or diversity of training data to adequately leverage the model's capabilities.

Interestingly, LeNet also demonstrated commendable performance, achieving 95% accuracy in validation. LeNet, one of the earlier and simpler CNN architectures, has often been acclaimed for its efficiency and lightweight nature, which can be particularly advantageous in scenarios where computational resources are limited or optimization is pivotal. Its relative simplicity, compared to more complex networks like InceptionV3 and MobileNet, allows for easier implementation and can often be less prone to overfitting when dealing with smaller datasets. This architecture, despite its age and simplicity, illustrates that enhanced complexity is not always synonymous with superior performance and underscores the importance of aligning the network architecture with the specific application and dataset.

On the other hand, while MobileNet and InceptionV3 have demonstrated higher metrics, particularly in validation accuracy, extending these results to wider or differing contexts requires additional scrutiny and validation. Employing networks that have achieved over 98% in validation accuracy, such as MobileNet and InceptionV3, brings with it both merits and challenges. Their high performance in this study indicates a capability to adeptly handle the specific classification problem, potentially offering reliable and accurate sorting in practical deployments. However, it is vital to consider possible limitations, such as a risk of overfitting due to their model complexity and depth, or the computational demands for deploying these networks, especially in scenarios where resources are limited, or optimization is crucial. The variation in results among the different CNN models implies that there is no one-size-fits-all solution, and choosing an appropriate model requires a balance between accuracy, computational cost, and implementation complexity. Additionally, it is crucial to consider that despite the quantitative results obtained, practical implementation in a real-world scenario encompasses a multitude of other variables that may affect performance, such as lighting conditions, apple varieties, and computational resources available.

This preliminary investigation into employing various CNN architectures for apple sorting in the agricultural sector provides an initial step and preliminary insights into a potentially broader application. While the performance metrics varied among the models tested, these discrepancies underscore the importance of continued research and development to refine these technologies for practical, field-based applications. The

findings from this study present an intriguing initial viewpoint for forthcoming research. Nevertheless, while these results offer a noteworthy starting point, the authors intend to further refine the system by conducting additional testing on an operational conveyor belt tasked with sorting apples, to further evaluate and enhance its practical application. Looking forward, the authors intend to expand the database with a larger and more diverse set of images and venture into more intricate classification challenges, exploring the realm of multiclass classification to provide a more comprehensive and nuanced sorting mechanism within the agricultural domain.

ACKNOWLEDGEMENT

Mr. Oprea Daniel, from Fundeni, Buzau County, to whom the authors extend their appreciation, has showcased a notable willingness to support our study. His provision of access to his apple orchard proved beneficial for obtaining essential photographs for our dataset. We additionally express our gratitude for his continued support, as he has kindly offered to assist with forthcoming data collection efforts.

REFERENCES

[1] Abdo, A., Hong, C.J., Kuan, L.M., Pauzi, M.M., Sumari, P., Abualigah, L, Zitar, R.A., Oliva, D. (2023). Markisa/Passion Fruit Image Classification Based Improved Deep Learning Approach Using Transfer Learning. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 143-190). Springer. doi:10.1007/978-3-031-17576-3

[2] Anuar, N.A., Muniandy, L., Bin Jaafar, K.A., Lim, Y., Sabeeh, A.L., Sumari, P., Abualigah, L., Abd Elaziz, M., Alsoud, A.R.,Ahmad MohdAziz Hussein, A.M. (2023). Rambutan Image Classification Using Various Deep Learning Approaches. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 23-44). Springer. doi:10.1007/978-3-031-17576-3

[3] Jamwal, A., Srivastava, J.N., Dutta, U. (2022). Important Diseases of Apple (Malus domestica L.) and Their Management. In J. S. Srivastava, *Diseases of Horticultural Crops: Diagnosis and Management* (Vol. I, pp. 31-60). Apple Academic Press.

[4] Kavdir, I., Guyer, D.E. (2002, November). Apple Sorting Using Artificial Neural Networks and Spectral Imaging. *Transactions of the ASAE. American Society of Agricultural Engineers, 45*(6).

[5] Ke, C., Weng, N.T., Yang, Y., Yang, Z.M., Sumari, P., Abualigah, L., Kamel, S., Ahmadi, M., Al-Qaness, M., Forestiero, A., Alsoud, A.R. (2023). Mango Varieties Classification-Based Optimization with Transfer Learning and Deep Learning Approaches. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 45-66). Springer. doi:10.1007/978-3-031-17576-3

[6] Keresztes, B., Abdelghafour, F., Randriamanga. D., da Costa, J.-P., Germain, C. (2018). Real-time Fruit Detection Using Deep Neural Networks. *14th International Conference on Precision Agriculture.* Montréal. Retrieved from https://hal.science/hal-02518559

[7] Khazalah, A., Prasanthi, B., Thomas, D., Vello, N., Jayaprakasam, S., Sumari, P., Abualigah, L., Ezugwu, A.E., Hanandeh, E.S., Khodadadi, N. (2023). Image Processing Identification for Sapodilla Using Convolution Neural Network (CNN) and Transfer Learning Techniques. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 108-128). Springer. doi:10.1007/978-3-031-17576-3

[8] Larner, A. (2023). *The 2x2 Matrix. Contingency, Confusion and the Metrics of Binary Classification.* Springer. doi:10.1007/978-3-030-74920-0

[9] Li, Y., Feng, X., Liu, Y., Han, X. (2021). Apple quality identification and classification by image processing based on convolutional neural networks. *Nature (Scientific Reports), 11*. doi:10.1038/s41598-021-96103-2

[10] Liu, W. (2020). Interfruit: Deep Learning Network for Classifying Fruit Images. *bioRxiv*. doi:10.1101/2020.02.09.941039

[11] Nataraj K. B, Manohar M., Poornima K., Niharika, U. (2018, February). Automated System for Detection of Apple Purity and Its Grading. *International Journal on Future Revolution in Computer Science & Communication Engineering, IV*(2), 100-103.

[12] Nguyen, T.-H., Nguyen, T.-N., Ba-Viet Ngo, B-V. (2022). A VGG-19 Model with Transfer Learning and Image Segmentation for Classification of Tomato Leaf Disease. *AgriEngineering*(4), 871–887. doi:10.3390/agriengineering4040056

[13] Ong, S.-Q., Nair, G., Al Dabbagh, R.D., Aminuddin, N.F., Sumari, P, Abualigah, L, Jia, H., Mahajan, S., G. Hussien, A.G., Abd Elminaam, D.S. (2023). Comparison of Pre-trained and Convolutional Neural Networks for Classification of Jackfruit Artocarpus integer and Artocarpus heterophyllus. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 129-142). Springer. doi:10.1007/978-3-031-17576-3

[14] Pen, L.Z., Xian, K.X., Yew, C.F., Hau, O.S., Sumari, P., Abualigah, L., Ezugwu A.E., Al Shinwan, M., Gul, F., Mughaid, A. (2023). Artocarpus Classification Technique Using Deep Learning Based Convolutional Neural Network. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 1-22). Springer. doi:10.1007/978-3-031-17576-3

[15] Seema, Kumar, K., Gill, G.S. (2015). Automatic Fruit Grading and Classification System Using Computer Vision: A Review. *Second International Conference on Advances in Computing and Communication Engineering* (pp. 598-603). Dehradun: IEEE. doi:10.1109/ICACCE.2015.15

[16] Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks. *ICLR 2015.* Retrieved from https://arxiv.org/abs/1409.1556v6

[17] Srivastava, J.N., Singh, A.K., Sharma, R.K. (2021). Diseases of Apples and Their Management. In G. A. Chand, *Diseases of fruits and vegetable crops: recent management approaches* (pp. 19-40). Apple Academic Press.

[18] Tan, M., Le, Q. v. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, PMLR 97. Retrieved from arXiv:1905.11946v5 [cs.LG] 11 Sep 2020

[19] Theng, L.W., San, M.M., Cheng, O.Z., Shen, W.W., Sumari, P., Abualigah, L., Zitar, R.A., Izci, D., Jamei, M., Al-Zu'bi, S. (2023). Salak Image Classification Method Based Deep Learning Technique Using Two Transfer Learning Models. In L. Abualigah (Ed.), *Classification Applications with Deep Learning and Machine Learning Technologies* (pp. 67-106). Springer. doi:10.1007/978-3-031-17576-3

[20] Wan, S., Goudos, S. (2020). Faster R-CNN for multi-class fruit detection using a robotic vision system. *Computer Networks, 168*(107036). doi:10.1016/j.comnet.2019.107036

[21] Yu, F., Lu, T., Xue, C. (2023, December). Deep Learning-Based Intelligent Apple Variety Classification System and Model Interpretability Analysis. *Foods, 885*. doi:10.3390/foods12040885

[22] Zhang, L., Gui, G., Khattak, A.M. (2019). Multi-Task Cascaded Convolutional Networks Based Intelligent Fruit Detection for Designing Automated Robot. *IEEE Access, 7*. doi:0.1109/ACCESS.2019.2899940