electronics

MDPI

Article

Yolo-Papaya: A Papaya Fruit Disease Detector and Classifier Using CNNs and Convolutional Block Attention Modules

Jairo Lucas de Moraes *, Jorcy de Oliveira Neto, Claudine Badue, Thiago Oliveira-Santos and Alberto F. de Souza

LCAD-High Performance Computing Laboratory, University of the State of Espirito Santo, Vitória 29075-910, ES, Brazil; alberto@lcad.inf.ufes.br (A.F.d.S.)
* Correspondence: artsoft.lucas@terra.com.br

Abstract: Agricultural losses due to post-harvest diseases can reach up to 30% of total production. Detecting diseases in fruits at an early stage is crucial to mitigate losses and ensure the quality and health of fruits. However, this task is challenging due to the different formats, sizes, shapes, and colors that the same disease can present. Convolutional neural networks have been proposed to address this issue, but most studies use self-built datasets with few samples per disease, hindering reproducibility and comparison of techniques. To address these challenges, the authors proposed a novel image dataset comprising 23,158 examples divided into nine classes of papaya fruit diseases, and a robust papaya fruit disease detector called Yolo-Papaya based on the YoloV7 detector with the implementation of a convolutional block attention module (CBAM) attention mechanism. This detector achieved an overall mAP (mean average precision) of 86.2%, with a performance of over 98% in classes such as "healthy fruits" and "Phytophthora blight". The proposed detector and dataset can be used in practical applications for fruit quality control and are consolidated as a robust benchmark for the task of papaya fruit disease detection. The image dataset and all source code used in this study are available to the academic community on the project page, enabling reproducibility of the study and advancement of research in this domain.

Keywords: diseases in fruit; papaya; convolutional block attention module; computer vision; Yolov7

Citation: de Moraes, J.L.; de Oliveira Neto, J.; Badue, C.; Oliveira-Santos, T.; de Souza, A.F. Yolo-Papaya: A Papaya Fruit Disease Detector and Classifier Using CNNs and Convolutional Block Attention Modules. *Electronics* **2023**, *12*, 2202. https://doi.org/10.3390/electronics12102202

Academic Editors: Seonah Lee, Jinhyun Kim, Suwon Lee and Iouliia Skliarova

Received: 13 April 2023
Revised: 27 April 2023
Accepted: 3 May 2023
Published: 12 May 2023

1. Introduction

In tropical regions, notably in the coastal areas of Brazil, fruit cultivation has become a vital contributor to crop diversification and enhanced revenue for small-scale farmers, as well as a labor-intensive endeavor [1] regarded as a fundamental activity for the socio-economic advancement of developing nations. Among the diverse fruits produced in these regions, papaya (*Carica Papaya*) holds particular importance as it is cultivated in over 60 countries spanning four continents (Africa, America, Asia, and Oceania) [2].

Brazil, specifically, ranks as the world's second largest papaya producer, boasting an annual output exceeding 1.5 million tons; it is globally recognized as one of the largest exporters of premium quality papaya, surpassed only by Mexico [3,4].

The commercialization of papaya fruit in natura is of paramount economic importance, owing to its high added value, and, thus, making the quality and appearance of the fruit essential factors in this segment. As a climacteric and delicate fruit, papaya is highly susceptible to post-harvest losses, which can amount to 30% to 40% of the total production [1]. As a consequence, early and accurate detection/classification for diseases (for simplification purposes, the term "disease" will be used in this study to refer both to biological diseases and to mechanical damage, scarring, and other non-biological evaluations of the fruit) are critical for ensuring quality control measures and mitigating losses in production activities.

Manual quality control of papaya fruits, however, remains a labor-intensive, time-consuming, and expensive task which demands specialized knowledge and is often unavailable to farmers located in remote regions or small fruit processing facilities. More-

over, the manual inspection process is subject to the level of expertise and psychological-physiological state of the specialist, leading to potential misinterpretations of fruit diseases. Thus, there is an urgent need for computer vision systems that can assist or fully replace human specialists in the task of fruit quality control. Such systems are essential to ensure optimal fruit quality and mitigate the significant losses faced by this industry sector.

The development of an autonomous system for detection and classification of diseases and damages in fruits using image analysis poses significant challenges for computer vision as this system must be capable of addressing the following points: (i) Does the image contain the target fruit? (ii) Where are the coordinates of the fruit located in the image? (iii) Is there any injury, such as disease or from mechanical damage, present on the fruit? (iv) What are the coordinates of the identified injuries? (v) What specific injuries have been detected?

In recent years, convolutional neural networks (CNNs), a type of deep learning technique, have garnered significant attention and application in various research domains, particularly in signal processing related tasks, with computer vision being one of the most prominent. However, according to sources [5–8], the use of CNN-based approaches for fruit quality control tasks, such as visual disease detection, fruit maturation, and measurements, have not yet been fully established. The lack of large, properly annotated public image datasets is one of the reasons as cited by [5,6].

Such approaches require diverse samples from the relevant domain, featuring varied shapes and sizes, complex backgrounds, variable lighting conditions, and different focal lengths. Learning to generalize real-world situations is essential for neural networks, and this can be achieved through training set diversification. However, due to the associated costs, time consumption, and specialized knowledge required, this can be a challenging and often infeasible task.

The FruitNet dataset [9], which includes 14,700 examples of six different fruits and classifies them as "good quality", "poor quality", and "mixed quality", has recently become available on Kaggle. While it offers a reasonable number of samples for research in the field of quality control, its practical use is limited as it does not identify the specific disease affecting the fruit; this is fundamental for an autonomous system in this domain.

The insufficient availability of samples is not limited to the papaya cultivation domain, as many published works in this area rely on image datasets created by the authors themselves. Such sets typically have very few samples and limited variation in symptoms, making them inadequate for state-of-the-art approaches in computer vision [5,6]. Therefore, this lack of standardization in terms of sample size, number of diseases detected, and training and validation set size among the image sets described in the literature, poses a challenge to making a fair comparison between them.

In his work, Barbedo [5] examined the factors that affect the detection of plant diseases and identified some of the same issues mentioned earlier that lead to unrealistic results. In addition to those points, Arsenovic [6] noted that most current methods are limited in their ability to detect multiple instances of the same disease or multiple diseases in a single image. While both studies were conducted on images of leaves and plants, their findings can be also extended to the domain of fruits.

It is important to note that, recently, neural networks based on transformers have gained notoriety in the field of computer vision, particularly with the development of vision transformers (ViTs) [10,11]. However, we have chosen not to consider these approaches in our work for the following reasons: (i) Studies comparing CNNs and ViTs have shown that transformer-based approaches require significantly more sample data to achieve comparable or better accuracy than CNN networks [12]. In the domain of fruit quality control, despite the recently presented database with adequate annotations, sample data remain a very limited resource; (ii) Training state-of-the-art ViT networks still requires extremely high computational resources, but this enterprise level of raw computing power is out of reach for most research labs worldwide. As one of the goals of our work is to establish a reliable benchmark for disease detection in fruits and encourage research in

this area, we have limited ourselves to solutions that could be replicated using moderate computational resources, often requiring only a single GPU on a single training node, which is considered quite accessible to most labs and researchers in this domain.

In the present study, we aimed to overcome the aforementioned limitations by creating a new image database that includes multiple instances of various diseases affecting the papaya tree. This dataset contains over 23,000 examples of eight different types of lesions, in addition to healthy fruit samples. We also propose a novel method for detecting papaya fruit diseases, utilizing a combination of the convolutional block attention module (CBAM) attention mechanism and a YOLOv7 network. Our results demonstrate the high accuracy and robustness of the proposed detector in identifying injuries in papaya fruits.

The main contributions of this work are as follows:

1. Development of a new method for detecting diseases in papaya fruits using a combination of convolutional block attention modules (CBAM) and YOLOv7 frameworks. The implemented detector can efficiently and accurately detect eight diseases affecting papaya fruits as well as healthy fruits. It can detect multiple diseases and/or multiple instances of the same disease in a single image, and its results surpass other methods tested on large datasets in this domain, thus establishing a new state-of-the-art (SotA) performance level;
2. Availability to the academic community of the source code necessary for the implementation of the proposed attention mechanism, the convolutional block attention module (CBAM), and the required modifications to the YOLOv7 framework to incorporate what has been proposed;
3. The provision of a new version of the Sisfrutos Papaya image dataset [13], comprising 23,158 examples of eight distinct diseases in addition to examples of healthy fruits. Notably, the annotations of the new dataset were performed while considering the cases of multiple diseases and/or multiple instances in the same image.

As secondary, but still relevant, contributions we may cite the following:

4. Implementation and testing of four additional state-of-the-art (SotA) detectors, thus establishing a solid benchmark for future work;
5. Public availability of pre-trained weights for the proposed detector, enabling researchers to use them in conjunction with their own datasets via transfer learning approaches [14];
6. Provision of the image annotations in two different formats (TXT and JSON), aiming to reach a broader audience of researchers.

The rest of this study is organized as follows: Section 1.1 provides a brief review of related works in the domain of interest, while Section 2 and its subsections describe the steps involved in image acquisition and annotation, dataset creation, attention module development, and the main experiments and their results. Section 4 discusses the key findings.

1.1. Related Studies

To the date of this research (December 2022), we have found few studies that focus on the detection of diseases in fruits and report their results on large sets of publicly available data. Thus, we have also examined the most recent studies published in the field of fruit quality control for various crops in which researchers have used small datasets. These are described in detail below.

In [15], the authors evaluated the performance of several network models (MobileNetV2, EfficientNetB0, ResNet50, and VGG16) for the task of classifying fruit quality into three categories: “good”, “poor”, and “mixed”. They used 5553 images from the FruitNet [5] dataset and applied various data augmentation techniques, such as horizontal flip, rotation, width shift, height shift, and zooming, to balance the samples. They also created a training set of their own, consisting of 200 samples from 18 different classes, where each class represented a combination of a fruit with its possible quality states (e.g.,

orange "bad", orange "good", orange "mixed", apple "bad", apple "good", apple "mixed"). The best accuracy for classifying fruits was achieved by the EfficientNetB0 model, with a value of 95%. However, it should be noted that the dataset used in the aforementioned study did not include information on the specific diseases affecting the fruits, which limits its applicability in disease detection and classification tasks.

In their pursuit of developing effective disease detection and classification techniques for papaya fruits, Habib et al. [16,17] have conducted two studies employing computer vision methodologies. In the first study [16], the authors presented an expert system for detecting and classifying diseases in papaya fruit images captured through mobile devices. The proposed method entails several steps, including (i) bicubic interpolation to standardize the image size to 300 × 300; (ii) histogram equalization to improve contrast; (iii) color space conversion from RGB to L × a × b space; (iv) image segmentation using k-means clustering; (v) the actual disease detection and classification using the SVM classifier. The authors utilized 129 papaya images in their study, with 84 images allocated for training and 45 for testing. However, they did not report the use of any validation set in their work. The proposed method achieved a precision of 90.15%. Both the dataset and source code utilized in the study were not made publicly available.

In their subsequent work, Habib et al. [17] evaluated the performance of nine distinct classifiers for papaya disease detection and classification, including k-nearest neighbors (KNN), logistic regression, repeated incremental pruning to produce error reduction (RIPPER), naive Bayes, random forest, support vector machine (SVM), back propagation neural network (BPN), and counter propagation networks (CPN). The authors used the same dataset of papaya images as their previous study [16] to train and test these classifiers. The results indicated that the SVM classifier yielded the highest accuracy among all the classifiers tested, with an accuracy of 95.2%.

Hossen [18] proposed a deep neural network (DNN) model for classifying papaya fruits as "diseased" or "healthy". The study employed a dataset of 234 images, with 184 images used for training, 28 for validation, and 22 for testing. The proposed network consisted of a basic convolutional neural network (CNN) with three convolutional layers followed by max pooling and two dense layers with a sigmoid function in the classification function. Although examples of both healthy and diseased papaya fruits and leaves were presented, the number of images used to form the training and test sets of leaves and fruits were not specified. The authors reported an average accuracy of 91% for the classification task, which is remarkable given the limited number of training examples in a CNN network. The source code and images utilized in the study were not made publicly available.

In a similar study, Hossen [19] also compared several algorithms for the classification of five diseases ("anthracnose", "black spot", "Phytophthora", "powdery mildew", and "ring spot") in papaya fruits. The study compared the performance of random forest, k-means clustering, support vector machine (SVM), and convolutional neural network (CNN) classifiers. The dataset used for the experiments consisted of 214 images, with 128 images utilized for training and 86 for testing. No validation set was used in the study. The CNN approach achieved the highest accuracy with 98%. The absence of a validation set, and the unavailability of source code and dataset, make reproducing the reported results difficult.

Moraes et al. [13] created a dataset of 15,179 images depicting 7 diseases/damages that affect papaya, including "anthracnose", "Phytophthora blight", "mechanical damage", "chocolate spot", "sticky disease", "physiological spot", and "black spot". The images were obtained in situ in a fruit packaging facility in a rural production environment and annotated to depict a single disease in each image. The authors divided the dataset into a training set (12,071 images), a validation set (1554 images), and a test set (1554 images) and employed the Yolov4 detector for the disease detection task. The study reported an f1-score of 80.1% for the disease detection task. However, the dataset's annotations were limited to depicting a single label in each image, even in cases where the fruit was affected by secondary diseases or multiple instances of the same disease, making it less useful for real-world applications.

2. Materials and Methods

The current study was conducted according to the following steps: (i) A dataset of images depicting the main diseases affecting papaya fruit was collected; (ii) The collected images were manually annotated by in field specialists who identified and classified all instances of disease present in the fruit; (iii) The resulting dataset was then partitioned into three subsets, including a training set, a validation set, and a test set, with proportions of 80%, 10%, and 10%, respectively; (iv) The dataset was employed on training and evaluating the performance of several implementations of convolutional neural networks (CNNs), as well as a novel CNN architecture proposed in the study. In the following sections, each of these processes is described in further detail.

2.1. Image Acquisition

Prior to the scope of this research, the sole publicly accessible dataset featuring a substantial number of papaya disease samples with precise annotations suitable for utilization in convolutional neural networks (CNNs) was *Sisfrutos Papaya* [13]. Nonetheless, this dataset's limitation stemmed from its labeling approach, which assigned only one disease per image even when the fruit displayed additional secondary diseases; this proved to be a limiting factor for practical applications.

Consequently, we developed an enhanced version of this dataset for our study, incorporating the following improvements: (i) Expansion of the dataset by incorporating over 8000 new instances; (ii) Introduction of a new class, "scar," to the dataset; (iii) Implementation of a model that facilitates the annotation of multiple diseases and/or several occurrences of the same disease within a single image; (iv) Provision of annotations in both TXT and COCO formats [20].

The supplementary images were obtained in a genuine production environment, in collaboration with fruit packaging companies over a period of six months. Throughout this interval, fruit samples that passed through the production line were photographed and evaluated by specialists in the quality control sector of the respective companies. All fruits used in this dataset were appraised in two phases by diverse experts following the evaluation protocol described below.

The first evaluation stage entailed a specialist randomly selecting fruit during its transport through the production conveyor, immediately after washing and before packaging. Subsequently, using a mobile device and an application designed explicitly for this task, the specialist captured images of the fruit and identified the regions of interest (ROI). These included the region of the fruit and the areas impacted by injuries, such as diseases, mechanical damage, and scars, which were then classified by name. Notably, a single image could display occurrences of several diverse diseases or several instances of the same disease. The captured image and annotation data were then stored in a database, and a unique identifier (ID) was generated to permit the unambiguous identification of the fruit.

In the second evaluation stage, another group of specialists with greater expertise than those in the first assessment reviewed the images that had undergone the first evaluation. In a blind assessment, without access to the results of the first evaluator, the second specialist analyzed the images, identified the regions of interest within the image, and categorized the injuries, where applicable. The results of the second evaluation were then stored in the respective fruits (using their IDs), along with the data from the first assessment. The second evaluation, which is considered the ground truth, supersedes the first evaluation in the event of discrepancies.

The existing images underwent a new annotation process through a two-stage evaluation process. In the first stage, the evaluator viewed the images using proprietary software and recorded the relevant information. After completing the first phase, the images were presented for a second evaluation, where, in a blind assessment, the evaluator recorded their own observations. The evaluators in the two stages were always distinct, and in line

with the aforementioned protocol, the second evaluator was considered the ground truth. Additional details on the image acquisition phase are accessible on the project's website.

2.2. Image Annotations

In an effort to broaden the accessibility of the new dataset to a wider range of researchers in the field, we have included annotations in two widely employed standards for state-of-the-art (SOTA) classifiers in the object detection task.

- *TXT Format:* (i) Each image (.jpg) has its respective .txt file. For example, the TR00001-4.jpg image is related to the TR00001-4.txt tag file; (ii) Each line of the .txt file describes an object (disease or healthy fruit) that appears in the respective image; (iii) The content of each line contains the following data:

  <class> <x_center> <y_center> <width> <height>
  Where:
  <class>—Id with the object's class;
  <x_center> <y_center>—central point of the object;
  <width> <height>—Object's width and height.

  These values are always given in relation to the image size, allowing the image to be resized without losses in relation to the regions of interest. This annotation standard is used, for example, by the detector of the Yolo family [21].

- *COCO Format*: The Coco data structure format is widely utilized by several state-of-the-art (SOTA) detectors, including EfficientDet [22] and YoloR [23]. The annotations of instances of each object consist of a basic data structure that includes several fields with information about the image, annotations, and classes. Structured records are used for storage, enabling a single JSON file to store annotations for an entire dataset.

To reduce the possibility of human error in identifying points of interest, all annotations were subject to automated verification at the conclusion of the process. Images that met the following criteria were excluded from the dataset: (i) The fruit area represented less than 10% of the total image area; (ii) The disease coordinates were not contained within the fruit area; (iii) The disease area was larger than the fruit area.

2.3. The New Extended Sisfrutos Papaya Dataset

The new dataset comprises 23,158 samples of eight distinct diseases, as well as samples of healthy papaya fruit. The dataset includes multi-class samples, wherein the same image may contain examples of several diseases or even multiple instances of the same disease. This characteristic of the samples brings the set of images closer to real-world scenarios, where it is common for fruit to be impacted by more than one disease. The images are presented in the RGB standard, with a size of 503×672 pixels, a complex background, and significant variations in luminosity, pose (rotation and translation), and focal length.

The selection of monitored diseases was based on those that caused the greatest financial losses, produced the most substantial damage, and were the most prevalent in the observed crop [4]. These diseases include anthracnose, Phytophthora blight, mechanical damage, chocolate spot, sticky disease, physiological spot, black spot, and scar. The dataset is unbalanced, with varying numbers of examples for each class. Diseases such as sticky disease and Phytophthora blight have fewer samples due to their lower occurrence rates in comparison to the other diseases.

The curated dataset in our study was compiled from samples gathered in an in-field fruit processing and packaging facility, where it is common for real-world datasets to display an unequal distribution among classes. To address this issue, we implemented several precautionary measures.

Initially, we employed the mAP (mean average precision) metric to assess our model's overall performance. This metric calculates the average precision of each class (AP), independent of the number of examples within that class. This approach also enables a dependable overall result, even when dealing with imbalanced datasets. Furthermore,

we examined the individual performance of each class, offering a more comprehensive understanding of the model's performance concerning specific classes.

As our dataset's objective is to set a robust benchmark for subsequent research, we refrained from using techniques such as oversampling, undersampling, and synthetic image generation to artificially balance the dataset. These approaches can be investigated in future studies that use our work as a reference point for comparison.

Table 1 displays the sample distribution for each class, while Figure 1 shows case examples of each disease.

Table 1. Distribution of classes in the dataset.

Class	Sample Quantity	(%)
Healthy papaya	7345	31.72%
Anthracnose	1104	4.77%
Phytophthora blight	207	0.89%
Mechanical damage	1072	4.63%
Chocolate spot	2622	11.32%
Sticky disease	208	0.90%
Physiological spot	1509	6.52%
Black spot	4661	20.13%
Scar	4430	19.13%
Total	23,158	100.00%

Figure 1. Examples of each dataset class.

Upon completing the creation of an image base capable of meeting all necessary requirements for the project, such as variability, quantity, and the presence of multiple diseases, we proceeded to tackle the second part of the problem: obtaining an object detector capable of adapting to the particularities of the domain.

Most of the conventional structures of cutting-edge detection algorithm models, such as those from the Yolo or EfficientDet [22] series, are customized with the objective of detecting objects in image databases, such as ImageNet [24] and COCO [20]. These datasets consist of hundreds of classes of objects, with well-defined object formats and significant intraclass similarity, but significantly differ from the scenario encountered in the task of

detecting diseases in fruits. In the fruit disease detection and classification domains, the number of classes is small, the objects sought are highly misshapen [6], and there are significant dissimilarities within the same class. As a result, the performance of these detectors falls short of expectations when applied in this specific domain.

The studied domain presents significant challenges for computer vision tasks, as the same disease (class) can exhibit drastically different size, shape, color, and texture. This necessitates a detector model with a level of generalization beyond that of conventional detectors to abstract these particularities. Figure 2 provides an example of this point, showcasing three examples of the "mechanical damage" class with highly distinct characteristics. To address these challenges, we first trained and tested the image base using state-of-the-art detectors commonly employed in the object detection task. Subsequently, we developed the Yolo-Papaya model: an adapted version of Yolov7 that employs additional layers of convolutional block attention modules (CBAM) to facilitate superior generalization of the classes without increasing inference time. The details of our model's implementation and the results we obtained are described in the subsequent sections.

Figure 2. Examples of the mechanical damage class.

2.4. Convolutional Block Attention Module (CBAM)

It is now widely accepted in the scientific community that attention mechanisms play a crucial role in human vision. One of the critical properties of the human visual system is that it does not process the entire scene image at once; instead, it selectively focuses on the salient parts to better capture the visual structure [25]. In recent years, researchers have been inspired by this concept to create selective attention mechanisms that allow neural networks to learn "where" or "what" to pay more attention to, building explicit dependencies between channels or weighted spatial regions.

In this context, Woo [26] proposed the convolutional block attention module (CBAM), which is an effective attention mechanism for convolutional neural networks. CBAM combines the concepts of channel attention module (CAM) and spatial attention module (SAM) and has demonstrated good results in residual networks.

In brief, CAM explores the inter-channel relationship of features, where each channel of a feature map is considered a feature detector. This module focuses on "what" is significant in the input image and generates two spatial context descriptors using average pooling and max pooling operations. These descriptors are then sent to a shared network to produce the channel attention map. SAM, on the other hand, focuses on "where" the most descriptive part of the input image is located. To calculate spatial attention, average pooling and max pooling operations are applied along the channel axis, concatenated and sent to a standard convolutional layer to generate an efficient feature descriptor.

CBAM combines both the above concepts: given an input image, two attention modules (CAM and SAM) calculate complementary attention, focusing on "what" is important and "where" is the descriptive part of the image. Figure 3 provides a diagram of the modules described. A detailed explanation of CBAM, CAM, and SAM can be found in [26].

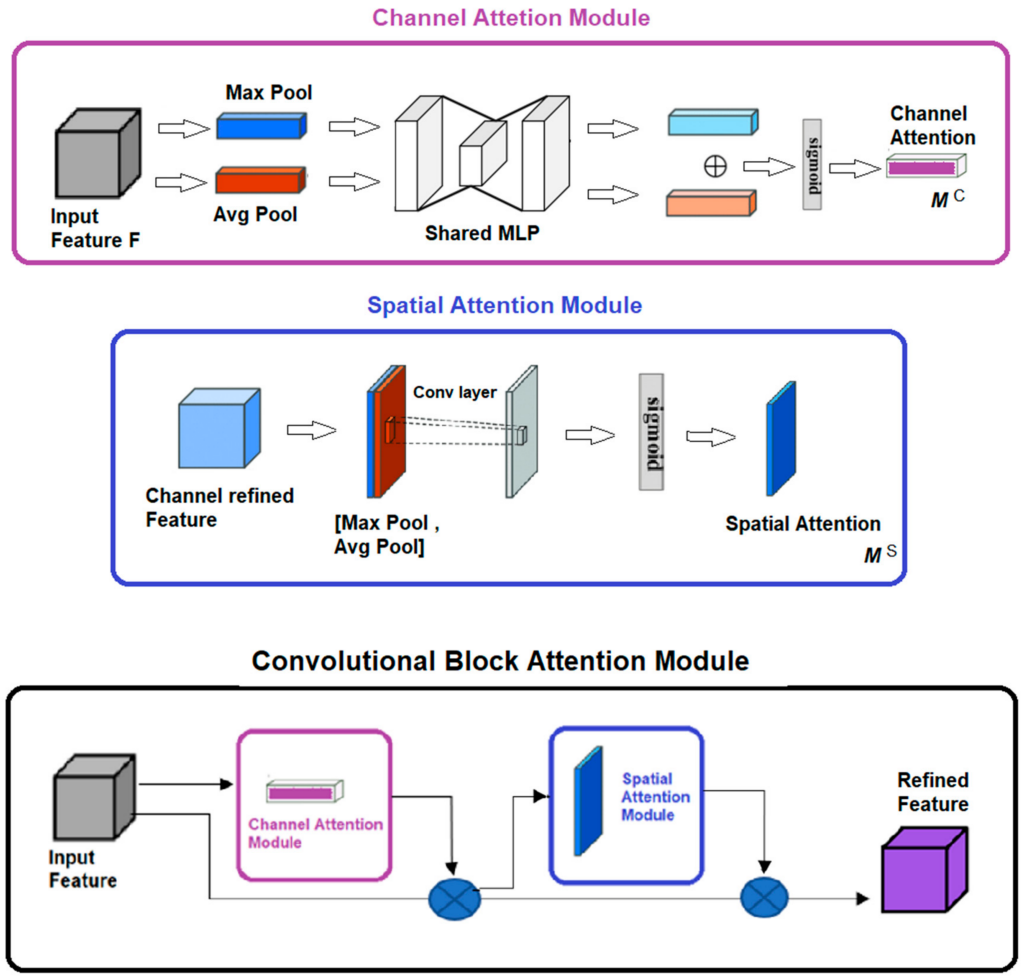

Figure 3. Description of the CAM, SAM, and CBAM Blocks. Adapted from [26].

Recent studies, including [27,28], have demonstrated successful combinations of convolutional block attention modules (CBAM) with convolutional neural networks in computer vision tasks. However, these studies have not provided the source code for the implementation of CBAM modules or the configuration of the network structure, making the reproducibility of reported results unfeasible.

Constructing a neural network is an intricate process that necessitates consideration of the dataset in use, the metrics to optimize, and the available computational resources. In our study, we determined the layers for incorporating the new CBAM layers based on our intuition and familiarity with the CNN literature. We opted for the last convolutional layers preceding the concatenation layer, as these layers are anticipated to encompass the most diverse and rich information with higher abstraction capabilities, including features such as shapes and textures. Additionally, the insertion of CBAM layers into the network's backbone and the number of CBAM blocks integrated should not significantly increase the network's overhead.

It is noteworthy that that our proposed integration of CBAM into YOLOv7 enables its insertion at various points within a CNN architecture, allowing researchers to develop a customized design for their specific problem. As such, the location, number of blocks, and their integration with other network components are all design decisions that may vary depending on the problem being addressed.

Figure 4 illustrates a simplified view of the backbone of the proposed structure and details the formation of the E-ELAN module where the CBAM modules were inserted. The CBS modules (k, s) represent various convolutional modules, where k denotes the convolution kernel size and s denotes the convolution step size. More information about

the complete structure of the Yolov7 network and the composition of each module can be found in [29].

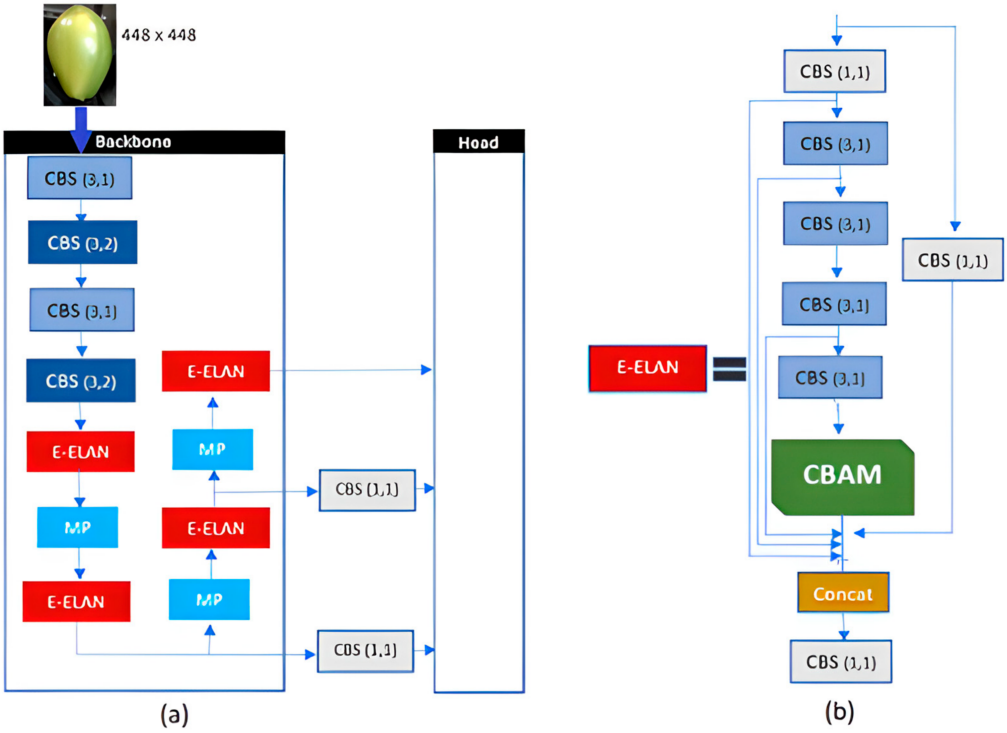

Figure 4. (**a**) Simplified structure of the backbone with the implementations of 4 CBAM modules (inserted in the E-ELAN blocks); (**b**) Detailing of the E-ELAN module. The CBS modules (*k*, *s*) represent various convolutional modules, where *k* denotes the convolution kernel size and *s* denotes the convolution step size.

3. Results

To conduct the experiments, the original dataset was partitioned into three non-overlapping datasets: (i) A training set with around 80% of the examples for training the model; (ii) A validation set with around 10% of the examples for tuning the model's hyperparameters; (iii) A test set with around 10% of the examples for evaluating the model's accuracy. The class distribution for each dataset is presented in Table 2.

To determine the optimal detector for the given task of detecting diseases in papaya fruits, a comparative analysis was carried out between several state-of-the-art (SoTA) detectors, including EfficientDet-B3 [22], YoloR [23], Yolov7 [29], and ResNet50 [30].

These detectors were trained and tested on the original dataset to ascertain the upper and lower limits of the dataset.

In order to evaluate the performance of each detector in our dataset and establish a robust baseline for comparison, we adhered to the following protocol:

i. All detector models employed were trained, validated, and tested exclusively using the datasets outlined in Table 2;
ii. The models were trained without utilizing any pre-trained weights, meaning they were consistently trained from scratch;
iii. Each model was trained and tested in accordance with the same settings (structure, hyperparameters, preprocessing, normalization, etc.) delineated in the respective publications of each model and for an identical number of epochs. Consequently, we are confident that we achieved the optimal performance for each detector given the dataset at hand;
iv. We compared the results of each detector, selecting YOLOv7 as our baseline due to its superior performance.

Table 2. Distribution of classes in the respective datasets.

Class	Dataset Train	Dataset Validation	Dataset Test	Total Number of Samples
Healthy papaya	5873	736	736	7345
Anthracnose	883	110	111	1104
Phytophthora blight	165	21	21	207
Mechanical damage	858	107	107	1072
Chocolate spot	2095	263	264	2622
Stick disease	166	21	21	208
Physiological spot	1207	151	151	1509
Black spot	3727	467	467	4661
Scar	3542	444	444	4430
Total	18,516	2320	2322	23,158

Subsequently, the proposed Yolo-Papaya model, which integrated convolutional block attention module (CBAM), was trained and tested on the dataset using the same parameters and number of epochs as the baseline detector. The performance of the proposed model was evaluated based on various metrics, such as mean average precision (mAP), F1-score, weight file size, and inference time. All experiments were carried out on a machine equipped with an Intel Xeon CPU E5606 processor (2.13 GHz), 24 GB RAM and NVIDIA TITAN XP GPU (12 GB). The results obtained from the experiments are presented in Table 3.

Table 3. Performance of tested detectors.

Detector	Size (Mb)	Inference Time (ms)	mAP (%)
Yolo-Papaya (our)	76.1	3.9	86.2
Yolov7	74.8	3.4	83.8
ResNet50	68.3	2.8	83.4
Yolo-R	105.3	3.2	82.3
EfficienDet-B3	48.6	37.4	67.1

The proposed Yolo-Papaya model demonstrated a mean average precision (mAP) of 86.2%, indicating a significant enhancement in the disease detection task compared to all tested models. Despite the considerable improvement in detection, the model maintains a stable number of parameters and inference time, establishing itself as the state-of-the-art for the task of detecting diseases in papaya fruits in large image databases and setting a robust benchmark for this specific task. Samples of correct and incorrect detections made by the model are illustrated in Figures 5 and 6. In Figure 7, we show examples of the detection of the best detectors for the same image.

To provide a more comprehensive evaluation of the proposed model, we conducted an analysis of its performance for each class. This is presented through a comparative class chart in Figure 8 and a confusion matrix in Figure 9.

The analysis of our proposed technique revealed that some diseases, such as black spot and physiological spot, are more challenging to detect due to the difficulty of specialists in accurately defining the boundaries of each instance of the disease in advanced stages. Consequently, the detector may yield false negatives and/or false positives, even when the disease is correctly detected. Figure 10 illustrates this situation. In image a1, the specialist marked two distinct instances of the "black spot" disease (ground truth). However, the detector identified it as a single continuous instance, resulting in a false negative (a2). In image b1, the specialist marked one instance of "black spot", while the detector identified it as two separate instances, leading to a false positive (b2). In image c1, the specialist identified two instances of the physiological spot disease, but the detector identified it as a single instance, resulting in a false negative (c2).

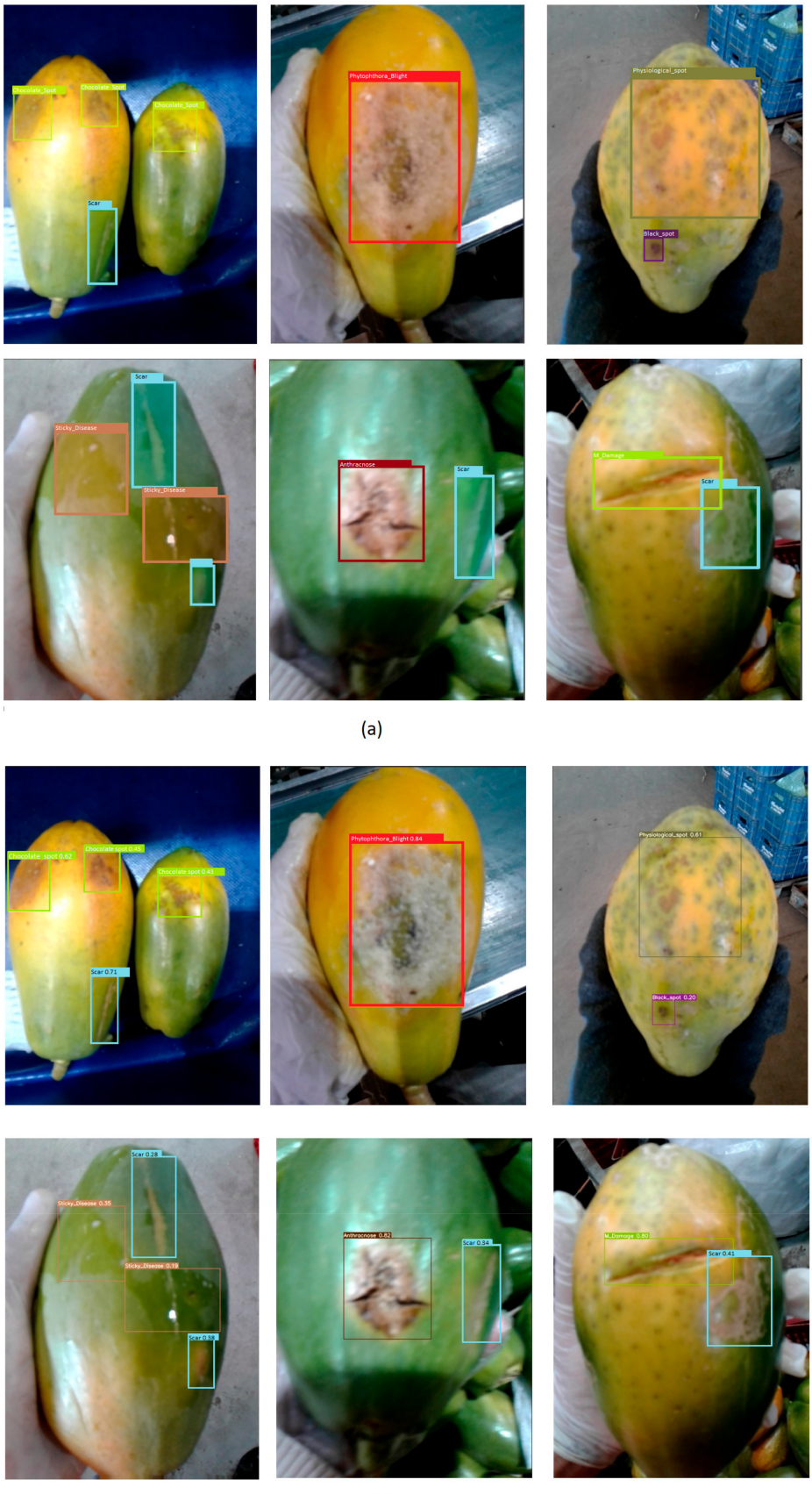

Figure 5. Examples of correct detections from the Yolo-Papaya model: (**a**) Shows ground truth; (**b**) Shows the detector predictions.

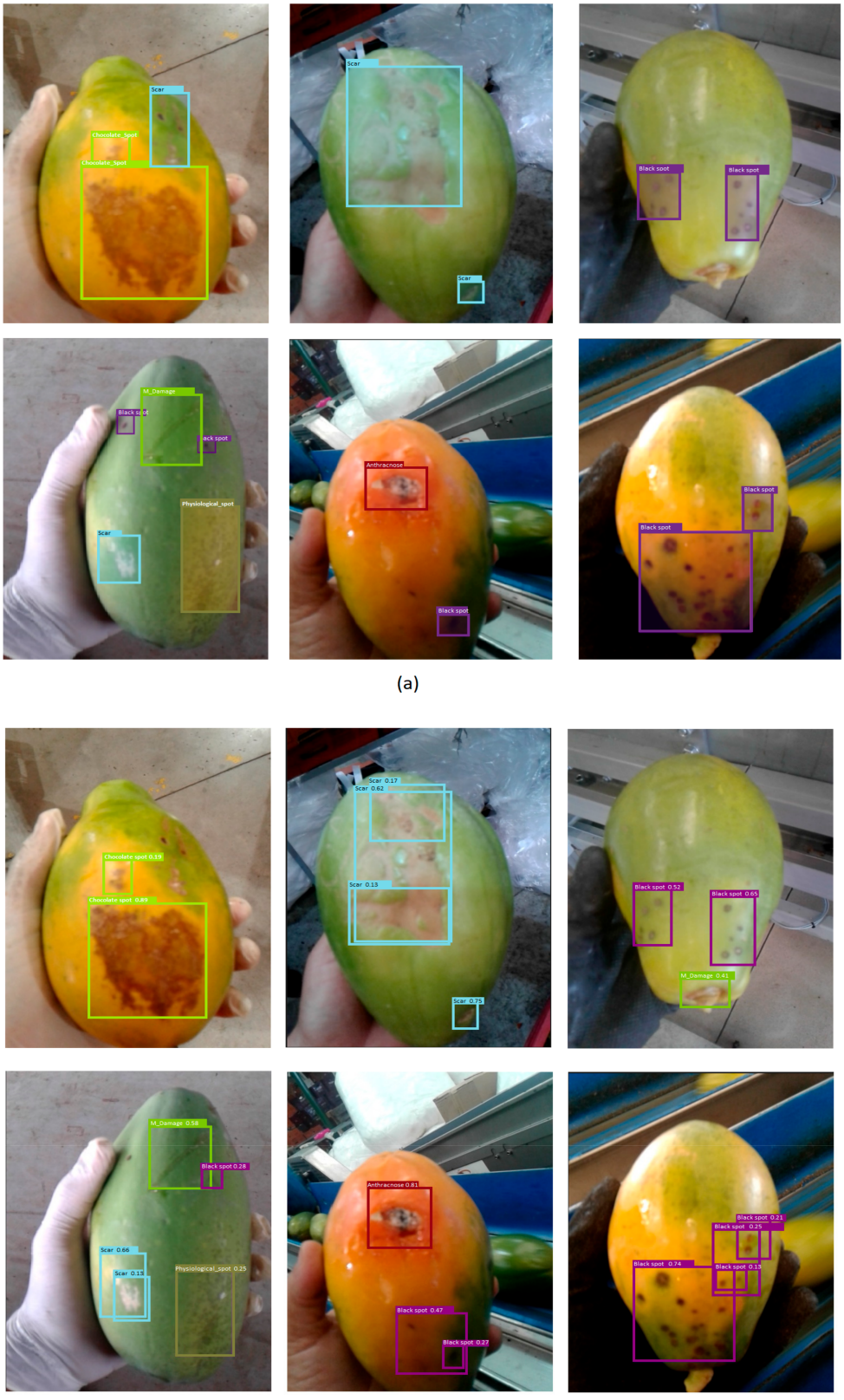

Figure 6. Examples of incorrect detections from the Yolo-Papaya model: (**a**) Shows ground truth; (**b**) Shows the detector predictions.

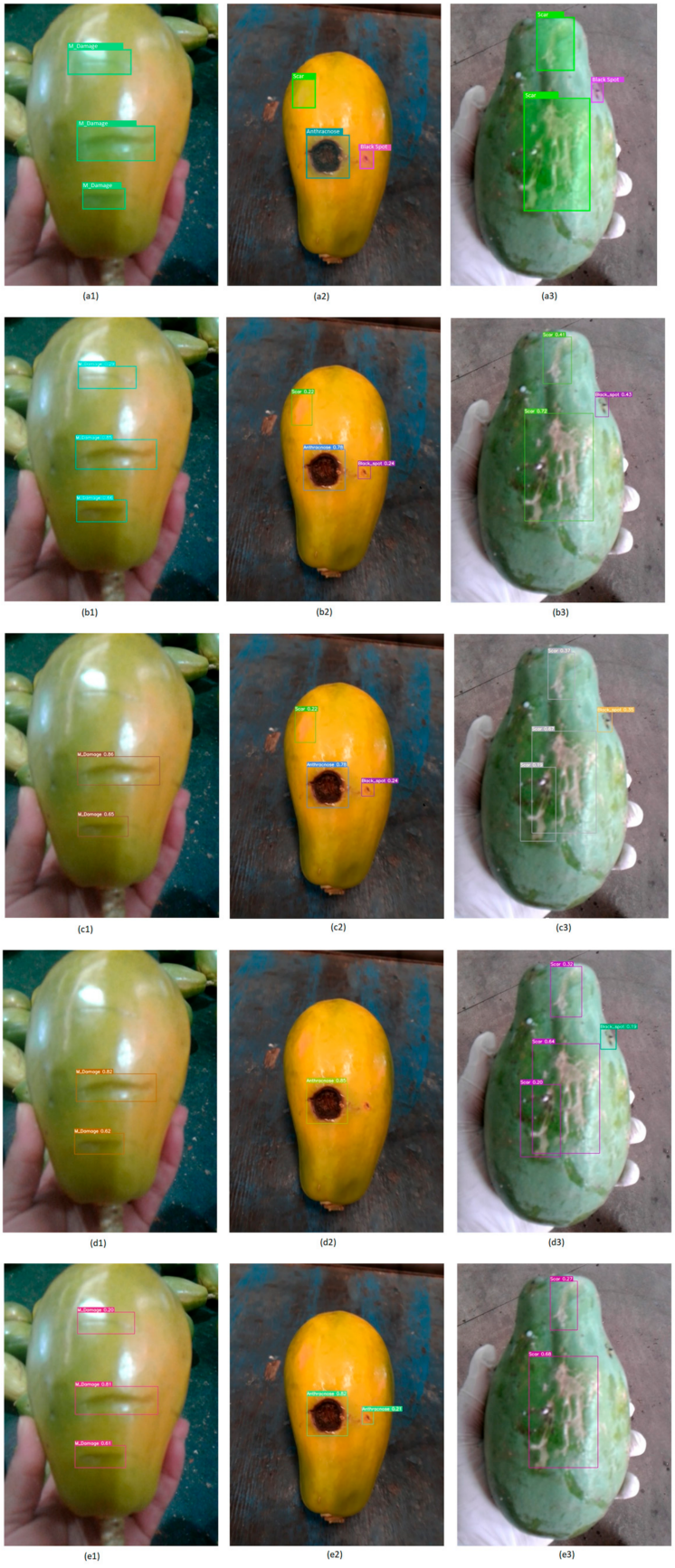

Figure 7. Example of the detections of the main models tested using a for the same sample: (**a1**,**a2**) and (**a3**) show the ground truth; (**b1**–**b3**) Yolo-papaya detections (our); (**c1**–**c3**) Yolov7 detections; (**d1**–**d3**) ResNet50 detections; (**e1**–**e3**) YoloR detections.

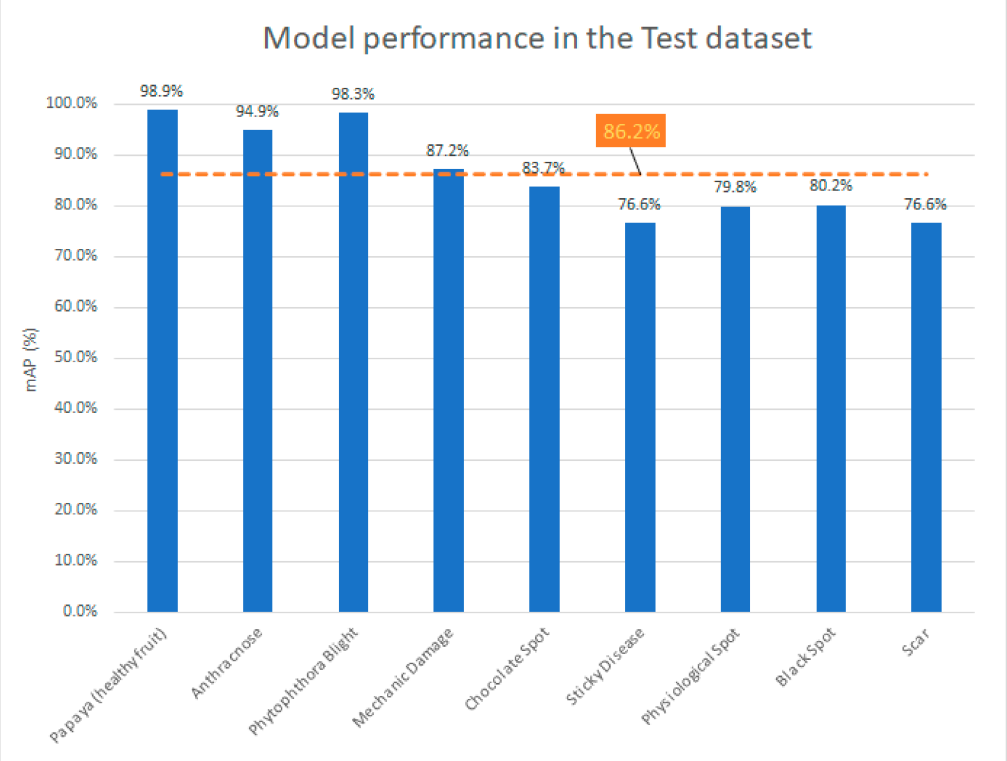

Figure 8. Graph with the performance of the model in each class considering the metric mAP. The dotted line indicates the average performance of the model.

		Ground Truth								
		Papaya (Healthy fruit)	Anthracnose	Phytophthora Blight	Mechanic Damage	Chocolate Spot	Sticky_Disease	Physiologicalspot	Black spot	Scar
Prediction	Papaya (Healthy fruit)	727								
	Anthracnose		105		2	2				
	Phytophthora Blight			20						1
	Mechanic Damage		2		93					20
	Chocolate Spot		4			220				
	Sticky_Disease						16			
	Physiological Spot							120	10	
	Black spot							15	374	
	Scar			1	12					340
	False Negative	9				42	5	16	83	83
	Total	736	111	21	107	264	21	151	467	444

Figure 9. Confusion matrix.

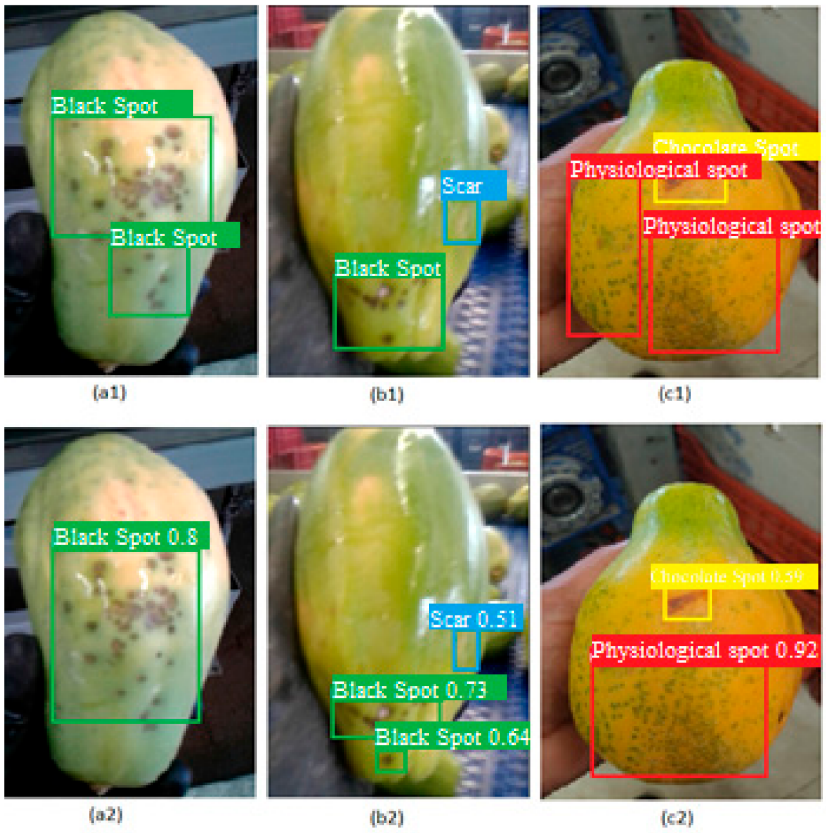

Figure 10. An illustration of the challenge in precisely delineating the boundaries for each instance of diseases that impact extensive regions of the fruit. Images (**a1**,**b1**,**c1**) show the expert's notes (ground truth). Images (**a2**,**b2**,**c2**) show the predictions of the detector.

The detection of diseases in Papaya fruits is an arduous task, even for experienced human specialists. According to Moraes et al. [13], the accuracy of a human expert, tested on an image base such as the one used in this study, was only 67.3%, and in some diseases, such as sticky disease, the performance was below 50%.

4. Discussion and Conclusions

Post-harvest losses in papaya cultivation are a significant challenge for the sector, necessitating efficient quality control to mitigate losses and ensure high-quality, healthy fruits for consumers. While computer vision techniques have advanced in recent years to address this issue, most works rely on small, self-created datasets. The datasets and codes of the proposed techniques are not publicly available, hindering reproducibility and making it difficult to establish a benchmark for this domain. This work proposes a publicly available image database with multi-class annotations, comprising 23,158 samples divided into 9 classes (8 diseases and 1 healthy fruit class), and implements a disease detector in papaya fruits based on convolutional block attention modules (CBAM). We compare the performance of our detector with several state-of-the-art detectors in the object detection task applied to the same image base. Our proposed detector demonstrates significant improvement in detecting diseases in papaya fruits when compared to other tested detectors, achieving an average mAP of 86.2% even in classes with high intra-class variation, such as "mechanical damage". Moreover, our detector maintains a stable number of parameters and inference times when compared to the other detectors, thus establishing itself as a benchmark for future work in this domain.

This work can mainly contribute to the several research domains: (i) More efficient network models: DNN network models are advancing rapidly in real-world applications, and with a comprehensive dataset, it is possible to test or develop new models to leverage

the state of the art in fruit quality control; (ii) Industrial automation: Research for the development of embedded systems aimed at the fruit processing industry, which would allow the autonomous or semi-autonomous detection of diseases in fruits, can benefit from the techniques proposed in this work; (iii) Precision agriculture: Harvesting robots, such as those proposed in [31,32], are already a reality in many cultures. A disease detection system embedded in a harvesting robot would enable selective fruit harvesting, bringing a huge cost reduction to this sector and improvements in the quality of the fruits offered to the end consumer; (iv) Small rural producers: A large portion of rural producers who are far from urban centers do not have access to specialized labor capable of correctly informing them of the diseases affecting their crops. A disease detector embedded in a simple smartphone or other mobile device could solve or alleviate this problem.

Author Contributions: Data curation, J.L.d.M.; Formal analysis, J.L.d.M.; Investigation, J.L.d.M. and J.d.O.N.; Methodology, J.L.d.M.; Project administration, C.B., T.O.-S. and A.F.d.S.; Resources, C.B., T.O.-S. and A.F.d.S.; Software, J.L.d.M.; Supervision, J.L.d.M.; Validation, J.d.O.N., C.B., T.O.-S. and A.F.d.S.; Visualization, J.d.O.N.; Writing—original draft, J.L.d.M.; Writing—review and editing, J.d.O.N. and J.L.d.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset and source code employed in this study are available on the project page (https://github.com/jhony2507/Sisfrutos-Papaya, accessed on 1 May 2023) for free download, subject to the acceptance of the terms of use for academic research. Our team will continually maintain and update the dataset and codes, which may include the addition of new diseases, corrections, and other changes deemed significant by the authors for the benefit of the research community.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. IBGE-Instituto Brasileiro de Geografia e Estatística, "Produção Agrícola-Lavoura Permanente". Available online: https://cidades.ibge.gov.br/brasil/pesquisa/15/0 (accessed on 18 August 2022).
2. FAO. *"Fruit and Vegetables–Your Dietary Essentials", The International Year of Fruits and Vegetables, Background Paper*; FAO: Rome, Italy, 2021. [CrossRef]
3. Comex Stat-Portal de Comércio Exterior do Ministério do Desenvolvimento, Indústria e Comércio. Available online: http://comexstat.mdic.gov.br/pt/home (accessed on 2 October 2021).
4. Dantas, J.; Junghans, D.; Lima, J. *O Produtor Pergunta, a Embrapa Responde*, 2nd ed.; Embrapa: Brasilia, Brazil, 2013.
5. Barbedo, J.G.A. Factors influencing the use of deep learning for plant disease recognition. *Biosyst. Eng.* **2018**, *172*, 84–91. [CrossRef]
6. Arsenovic, M.; Karanovic, M.; Sladojevic, S.; Anderla, A.; Stefanovic, D. Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry* **2019**, *11*, 939. [CrossRef]
7. Barth, A.R.; Jsselmuidem, J.; Hemming, J.; Van Henten, E.J. Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset. *Comput. Electron. Agric.* **2018**, *144*, 284–296. [CrossRef]
8. Gongal, A.; Amatya, S.; Karkee, M.; Zhang, Q.; Lewis, K. Sensors and systems for fruit detection and localization: A review. *Comput. Electron. Agric.* **2015**, *116*, 8–19. [CrossRef]
9. Meshram, V.; Kailas, P. FruitNet: Indian fruits image dataset with quality for machine learning applications. *Data Brief* **2022**, *40*, 107686. [CrossRef] [PubMed]
10. Poulinakis, K. Are Transformers Replacing CNNs in Object Detection? Picsellia Blog. 2022. Available online: https://www.picsellia.com/post/are-transformers-replacing-cnns-in-object-detection (accessed on 2 March 2023).
11. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
12. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
13. de Moraes, J.L.; de Oliveira Neto, J.; Correia-Silva, J.R.; Paixão, T.M.; Badue, C.; Oliveira-Santos, T.; De Souza, A.F. Sisfrutos Papaya: A Dataset for Detection and Classification of Diseases in Papaya. In Proceedings of the 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, 14–17 September 2021; Springer: Cham, Switzerland, 2021.

14. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
15. Morshed, S.; Ahmed, S.; Ahmed, T.; Islam, M.U.; Rahman, A.B.M.A. Fruit quality assessment with densely connected convolutional neural network. *arXiv* **2022**, arXiv:2212.04255.
16. Habib, M.T.; Majumder, A.; Jakaria, A.Z.M.; Akter, M.; Uddin, M.S.; Ahmed, F. Machine vision based papaya disease recognition. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *32*, 300–309. [CrossRef]
17. Habib, M.T.; Majumder, A.; Nandi, R.N.; Ahmed, F.; Uddin, M.S. A Comparative Study of Classifiers in the Context of Papaya Disease Recognition. In *Proceedings of the International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems, Ashulia, Bangladesh, 14–15 December 2018*; Springer: Singapore, 2019.
18. Hossen, S.; Haque, I.; Islam, S.; Ahmed, T.; Nime, J.; Islam, A. Deep learning based classification of papaya disease recognition. In *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 3–5 December 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 945–951.
19. Islam, A.; Islam, S.; Hossen, S.; Emon, M.U.; Keya, M.S.; Habib, A. Machine learning based image classification of papaya disease recognition. In *Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 1353–1360.
20. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context, in European Conference on Computer Vision (ECCV). *arXiv* **2014**, arXiv:1405.0312.
21. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. Available online: https://arxiv.org/abs/2004.10934 (accessed on 30 April 2020).
22. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020), Seattle, WA, USA, 13–19 June 2020.
23. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. You only learn one representation: Unified network for multiple tasks. *arXiv* **2021**, arXiv:2105.04206.
24. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
25. Larochelle, H.; Hinton, G. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2010.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision–ECCV, Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Yan, J.; Zhou, Z.; Zhou, D.; Su, B.; Xuanyuan, Z.; Tang, J.; Lai, Y.; Chen, J.; Liang, W. Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Front. Mar. Sci.* **2022**, *9*, 2299. [CrossRef]
28. Fu, H.; Song, G.; Wang, Y. Improved YOLOv4 marine target detection combined with CBAM. *Symmetry* **2021**, *13*, 623. [CrossRef]
29. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Amsterdam, The Netherlands, 8–16 October 2016.
31. Kootstra, G.; Wang, X.; Blok, P.M.; Hemming, J.; Van Henten, E. Selective Harvesting Robotics: Current Research, Trends, and Future Directions. *Curr. Robot. Rep.* **2021**, *2*, 95–104. [CrossRef]
32. Zhou, H.; Wang, X.; Au, W.; Kang, H.; Chen, C. Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precis. Agric.* **2022**, *23*, 1856–1907. [CrossRef]