Data in Brief 54 (2024) 110462

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Tea leaf age quality: Age-stratified tea leaf quality classification dataset

Md Mohsin Kabir[a], Md Sadman Hafiz[b], Shattik Bandyopadhyaa[b], Jamin Rahman Jim[c], M.F. Mridha[c,*]

[a] *Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka 1216, Bangladesh*
[b] *Institute of Information and Communication Technology, Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh*
[c] *Department of Computer Science, American International University-Bangladesh, Dhaka 1229, Bangladesh*

ARTICLE INFO

Article history:
Received 16 January 2024
Revised 21 March 2024
Accepted 16 April 2024
Available online 21 April 2024

Dataset link: Tea Leaf Age Quality: Age-Stratified Tea Leaf Quality Classification Dataset (Original data)

Keywords:
Tea leaf classification
Quality prediction
Deep learning
Image annotation
Machine Learning in Agriculture
Image Processing in Agriculture

ABSTRACT

The “Tea Leaf Age Quality” dataset represents a pioneering agricultural and machine-learning resource to enhance tea leaf classification, detection, and quality prediction based on leaf age. This comprehensive collection includes 2208 raw images from the historic Malnicherra Tea Garden in Sylhet and two other gardens from Sreemangal and Moulvibajar in Bangladesh. The dataset is systematically categorized into four distinct classes (T1: 1–2 days, T2: 3–4 days, T3: 5–7 days, and T4: 7+ days) according to age-based quality criteria. This dataset helps to determine how tea quality changes with age. The most recently harvested leaves (T1) exhibited superior quality, whereas the older leaves (T4) were suboptimal for brewing purposes. It includes raw, unannotated images that capture the natural diversity of tea leaves, precisely annotated versions for targeted analysis, and augmented data to facilitate advanced research. The compilation process involved extensive on-ground data collection and expert consultations to ensure the authenticity and applicability of the dataset. The “Tea Leaf Age Quality” dataset is a crucial tool for advancing deep learning models in tea leaf classification and quality assessment, ultimately contributing to the tech-

* corresponding author.
E-mail addresses: mdmkabi@gmail.com (M.M. Kabir), firoz.mridha@aiub.edu (M.F. Mridha).

https://doi.org/10.1016/j.dib.2024.110462

nological evolution of the agricultural sector by providing detailed age-stratified tea leaf categorization.

Specifications Table

Subject	Agricultural Sciences, Computer Science
Specific subject area	Computer Vision, Image processing, Deep learning, Agriculture
Data format	Raw, Annotated, Augmented
Type of data	Image
Data collection	The data collection for the "Tea Leaf Age Quality" dataset was conducted in three tea-producing zones in Bangladesh. An experienced and retired tea garden manager from Sreemangal, Moulvibazar, as well as consultations with a garden worker, provided initial insights into various tea leaf varieties. The tea leaves were then meticulously categorized into four classes (T1, T2, T3, and T4). Over three days, we captured approximately 2208 images across all categories. Then, these images were promptly annotated with their respective categories.
Data source location	The top 3 zones in Bangladesh that are famous for tea gardens and most of the tea gardens of Bangladesh are situated there: 1. Sylhet 2. Sreemangal 3. Moulvibajar
Data accessibility	• The dataset is published in Mendeley Data. • Dataset name: Tea Leaf Age Quality: Age-Stratified Tea Leaf Quality Classification Dataset • Data identification number(doi): 10.17632/7t964jmmy3.1 • Direct URL to data: https://data.mendeley.com/datasets/7t964jmmy3/1

1. Value of the Data

This dataset will be valuable for various fields, particularly agricultural research and machine-learning applications related to tea cultivation. The following are some insights and potential impacts of this dataset:

- The Tea Leaf Age Quality [1] dataset offers a meticulously categorized collection of tea leaf images based on age, providing unprecedented details. This precision facilitates more nuanced and accurate research on tea leaf quality, aiding scientists and agronomists in understanding and predicting quality based on leaf age.
- The categorization of the dataset into four distinct quality classes makes it an ideal resource for developing deep-learning-based detection algorithms. Researchers and technologists can leverage these data to design and train models capable of automatically detecting and classifying tea leaves into their respective categories.
- By understanding the relationship between leaf age and tea quality, cultivators and agricultural experts can optimize harvesting practices, leading to better quality control and potentially higher yields. Researchers can use these data to advise on the best harvesting time to ensure maximum quality.
- This dataset can act as a baseline for future research, allowing for comparative studies on tea leaf quality across different regions, climates, and cultivation methods. Researchers can use it to compare their findings, validate their methodologies, or explore regional variations in tea leaf quality.

2. Background

Tea farming significantly boosts the economies of many countries, with the quality of tea leaves shaped by their age and other factors, directly influencing their taste, aroma, and value [2]. Traditionally, the assessment of tea leaf quality has relied on manual inspection, a method fraught with inefficiencies and human error. However, advancements in machine learning and computer vision are driving a shift towards automating this process to enhance the precision and efficiency of tea leaf classification [3].

However, there is still a lack of comprehensive datasets needed to develop advanced algorithms capable of accurately categorizing tea leaves based on quality parameters categorized by age. The “Tea Leaf Age Quality” dataset, collected from the famous Malnicherra Tea Garden in Sylhet and other well-known tea-producing regions in Bangladesh, fills this gap. The dataset offers a carefully organized compilation of 2208 unprocessed photos, divided into four unique categories, demonstrating a sophisticated understanding of the relationship between age and quality. This dataset is crucial for developing deep-learning models for tea leaf categorization and quality prediction. Additionally, it is an excellent resource for agricultural research, providing insights into the most favorable harvesting times to achieve the highest tea quality.

The “Tea Leaf Age Quality” dataset was meticulously developed through comprehensive on-site data gathering and expert consultation, guaranteeing its credibility and practicality. This dataset can potentially change tea production operations by enabling automated systems to detect the quality of the tea leaves. Consequently, it can enhance quality control measures and increase yields. Furthermore, its organized compilation of unprocessed, labeled, and enhanced photos establishes a novel benchmark for dataset curation in agricultural research, potentially motivating comparable endeavors for other crops.

3. Data Description

The “Tea Leaf Age Quality” dataset hosted by Mendeley is meticulously structured into three distinct directories to facilitate easy access and analysis [1]. The first directory, “Tea Leaf Age Quality (Raw Data),” encompasses a collection of raw, unprocessed images captured directly from various tea gardens. These images were categorized into four classes based on the age and quality of the tea leaves.

Category T1: Age 1 and 2 days, representing the highest quality tea leaves. (562 Raw Images) Category T2: Age 3 to 4 days, indicating good quality tea leaves. (615 Raw Images) Category T3: Age 5 to 7 days, suggesting average or below-average tea leaves. (508 Raw Images) Category T4: Age 7+ days, signifying tea leaves unsuitable for brewing drinkable tea. (523 Raw Images), respectively.

Each category is housed in its folder within the “Tea Leaf Age Quality (Raw Data)” directory, providing a structured and organized dataset for preliminary analysis. Table 1 presents the first directory of the dataset, and Fig. 1 visualizes the tea leaves of each of the four classes.

Table 1
Description of raw data by tea leaf age class.

Class	Age (days)	Description	No. of Raw Images
T1	1–2	Tea leaves picked within 48 h, highest freshness and aromatic quality	562
T2	3–4	Tea leaves plucked within 72–96 h, high quality with strong flavour retention	615
T3	5–7	Tea leaves plucked within 5–7 days, moderate deterioration in flavour and aroma	508
T4	7+	Tea leaves plucked beyond 7 days, significant loss in essential oils, not recommended for consumption	523
Total			2208

Fig. 1. Sample images of the 4 categories of the Tea Leaf Age Quality dataset.

The second directory, "Tea Leaf Age Quality (annotated)," contains the same images as the raw data, but with added annotations. The annotation process was conducted using Roboflow, which is a robust tool designed to enhance the precision and usability of the dataset. This annotated dataset is pivotal for researchers aiming to conduct detailed analyses and develop models based on the age and quality of the tea leaves.

Further enhancing the utility of the dataset, the third directory, "Tea Leaf Age Quality (Annotated and Augmented)," includes the annotated images and additional augmentations to simulate various conditions and scenarios. This augmented dataset is particularly valuable for developing robust and versatile machine-learning models capable of performing well under diverse conditions.

Both the "Tea Leaf Age Quality (Annotated)" and "Tea Leaf Age Quality (Annotated and Augmented)" directories are organized into three subfolders: 'train', 'test', and 'valid'. These subfolders contain images and their corresponding annotation labels, facilitating a structured approach for model training, validation, and testing. This comprehensive structure ensures that the "Tea Leaf Age Quality" dataset is a versatile and valuable resource for researchers and technologists aiming to advance tea leaf classification and quality prediction fields. The overall hierarchy of the dataset directories is illustrated in Fig. 2.

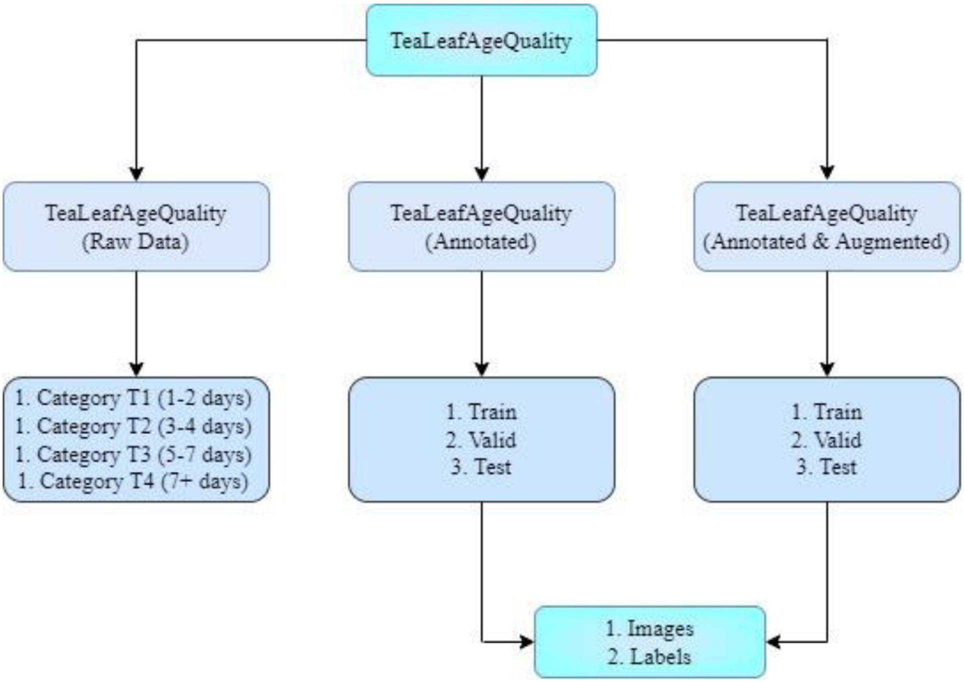

Fig. 2. Hierarchy of the Tea Leaf Age Quality dataset directories.

4. Experimental Design, Materials and Methods

This section discusses the overall experimental design and workflow of the dataset creation, which is presented in Fig. 3.

5. Experimental Design

5.1. Objective

To create a comprehensive dataset for tea leaf classification, detection, and quality prediction based on leaf age.

5.2. Location

Data acquisition was primarily conducted at the Malnicherra Tea Garden, Sylhet, with additional insights from a retired manager in Sreemangal, Moulvibazar, Bangladesh.

5.3. Sampling methodology

Tea leaves were sampled based on age and visual quality indicators categorized into four distinct classes (T1, T2, T3, and T4).

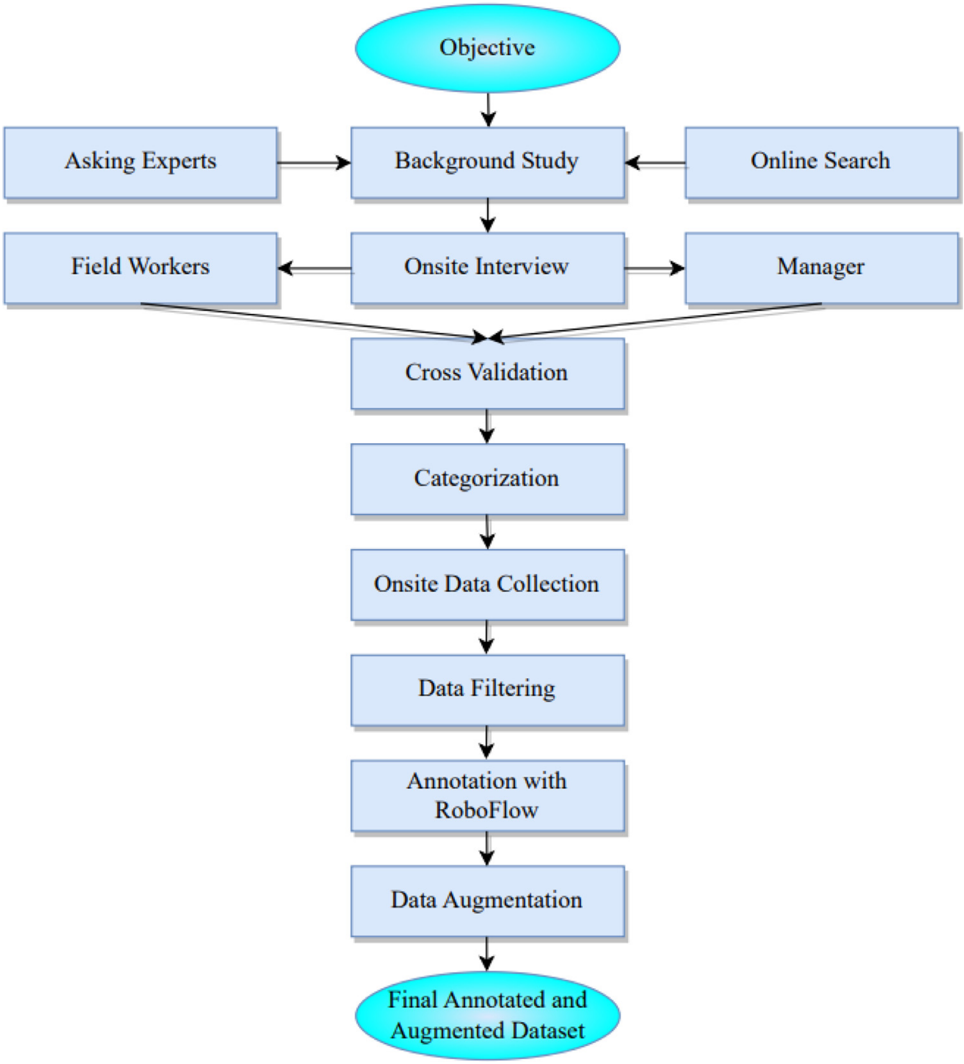

Fig. 3. Overall workflow of the Tea Leaf Age Quality dataset creation.

5.4. Data collection process

5.4.1. Camera device

A high-resolution camera on a smartphone was used to capture images of tea leaves. This ensures easy access and the ability to capture many images.

5.4.2. Engagement with experts

Initial discussions with garden workers and a retired tea garden manager provided foundational knowledge and validated tea leaf types and quality indicators.

5.4.3. On-site collection

We visited specific sections of tea gardens to capture images of the tea leaves from each category. The collection was spread over two days to ensure a diverse representation of the conditions and stages of leaf age.

Table 2
Annotation Excel file containing the image id and annotation classes.

Image_ID	Annotator_1	Annotator_2
20,231,224_160,737	1	1
20,231,224_160,738	2	1
20,231,224_160,740	1	1
20,231,224_160,743	2	1
20,231,224_160,747	1	1
20,231,224_160,752	1	1
20,231,224_160,758	1	1
20,231,224_160,765	1	1
20,231,224_160,773	1	1

5.4.4. Categorization

Leaves were categorized on-site into four classes (T1, T2, T3, and T4) based on age and perceived quality. Approximately 2208 images were captured, initially focusing on categories A and B (700 images).

5.5. Data processing and annotation

Software Tool: Roboflow, an image annotation tool, was used to annotate the collected images accurately. This involved labeling the images according to predetermined categories and ensuring that the annotations were consistent and precise.

The dataset was initially classified into four categories of tea leaves based on age, with the youngest leaves (1–2 days old) designated as the T-1 class. Images were systematically captured for each class, ensuring that all photographs for a given category were stored in a designated folder. Throughout the imaging process, tea garden owners and workers were present to guide the classification of images, thereby facilitating accurate categorization and minimizing ambiguity during annotation.

The annotation process was conducted by two experienced annotators. To assess the reliability of the annotations, we computed Cohen's Kappa score using data recorded throughout the annotation phase. The resultant Kappa score was 83.6 %, indicative of an 'Almost Perfect' level of agreement between the two annotators [4]. This high degree of concordance underscores the efficacy of our annotation methodology and the clarity of our class definitions.

The Cohen's Kappa score (**k**) was calculated using the formula:

$$\mathbf{k} = \frac{\mathbf{p\_o} - \mathbf{p\_e}}{1 - \mathbf{p\_e}}$$

where **p_o** represents the observed agreement between annotators, and **p_e** denotes the expected agreement by chance, derived from the probabilities of each annotator independently classifying each category. This metric provides a robust measure of inter-annotator reliability, adjusting for agreement that could occur by chance.

Presented below Table 2 is an excerpt from the annotation Excel file for the T-1 class, which showcases the concordance in annotations made by the two annotators. Here 1, 2 represents the classes (T1, T2).

Augmentation: The annotated images were subjected to augmentation techniques using standard image processing tools to enhance the dataset and simulate various conditions. This included transformations, such as scaling, rotation, and lighting adjustments.

5.6. Data organization

Directory Structure: The dataset was organized into three main directories: raw, annotated, annotated and augmented. Each directory was divided into subfolders for the training, testing, and validation sets to facilitate easy access and usage in machine learning workflows.

Table 3
Dataset evaluation on YOLOv8 model.

mAP	Precision	Recall
87.9 %	89.0 %	84.0 %

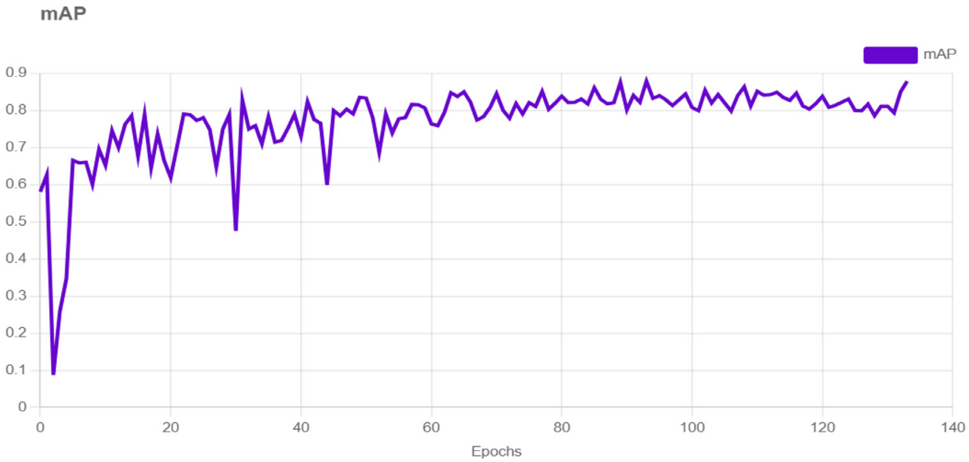

Fig. 4. Training graph for mAP.

File Naming Convention: Each image file was named systematically to reflect its category and other relevant details, ensuring easy identification and retrieval.

5.7. Quality assurance

Review and Validation: After annotation and augmentation, a subset of the dataset was reviewed for accuracy and consistency. Any errors identified during this review were corrected to maintain dataset integrity.

Adhering to this meticulous experimental design and methodological approach, the "Tea Leaf Age Quality" dataset was curated with the highest standards of quality and utility, serving as a valuable resource for the scientific community and industry practitioners.

5.8. Dataset evaluation

In introducing our new dataset, "Tea Leaf Age Quality," we thoroughly tested it using the YOLO (You Only Look Once) object detection model [5–7], specifically YOLOv8 [8]. This dataset is carefully created to evaluate both the age and the quality of tea leaves and has great potential for computer vision. Our detailed assessment with YOLOv8 shows how well the model detects and categorizes objects in this dataset, as well as highlights the importance of the dataset. It is designed for tasks such as assessing the quality of tea leaves in detail, making it a valuable resource for advancing computer vision research and applications. Table 3 shows the performance of the model on our dataset, and Figs. 4 and 5 display the training graph of the YOLO model on our dataset.

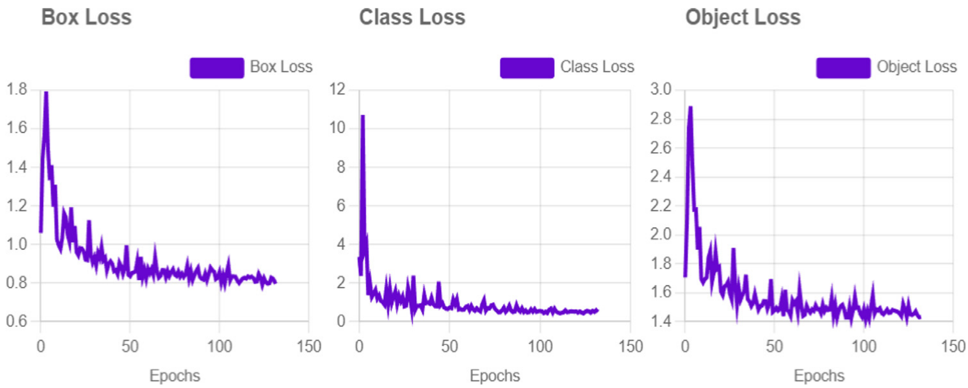

Fig. 5. Training graph for Box loss, Class loss and Object loss.

Limitations

The "Tea Leaf Age Quality" dataset may be limited in its representation of tea leaf diversity owing to geographical limitations and the possibility of human error in data collection and annotation, which could result in inconsistencies. Some of the limitations are as follows:

- The dataset may not capture the full spectrum of diversity within tea leaves. Tea cultivation is highly influenced by geographical factors, such as soil composition, altitude, season, and climate. Due to scope limitations, our dataset may be an incomplete representation of the diverse world of tea leaves affected by various environmental elements.
- Although we tried our best to marginalize the effects of human subjectivity in collecting images of different categories of tea leaves, the involvement of human beings might cause potential challenges and limitations in the exactitude of our dataset.
- Subjectivity in the annotation process may also cause inaccuracies in annotated images. In addition, a lack of training and experience among annotators, all of whom had varying backgrounds, might have led to uncertainties in the labeled dataset.

Ethics Statement

The 'Tea Leaf Age Quality' dataset was collected with the full consent and cooperation of the garden owner and managers. The collaborative initial classification process involved experts, garden workers, and managers to ensure accuracy. No harmful activities occurred during image collection, which was performed transparently in the presence of garden staff.

The purpose of the research was explained to the owners and managers beforehand. Each image was carefully captured and underwent a thorough manual review and filtering process before annotation, ensuring the quality and relevance of the dataset for tea leaf classification and research.

Data Availability

Tea Leaf Age Quality: Age-Stratified Tea Leaf Quality Classification Dataset (Original data) (Mendeley Data).

CRediT Author Statement

Md Mohsin Kabir: Conceptualization, Methodology, Software, Writing – original draft, Investigation; **Md Sadman Hafiz:** Data curation, Visualization, Formal analysis; **Shattik Bandyopadhyaa:** Data curation, Formal analysis; **Jamin Rahman Jim:** Writing – original draft, Validation; **M.F. Mridha:** Writing – review & editing, Supervision.

Acknowledgements

The authors would like to thank the Advanced Machine Intelligence Research Lab (AMIR Lab) for their supervision and resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that have influenced or could be perceived as influencing the work described in this paper.

References

[1] Kabir, Md Mohsin; Hafiz, Md Sadman; Bandyopadhyaa, Shattik; Jim, Jamin Rahman; Mridha, Firoz (2024), "Tea leaf age quality: age-stratified tea leaf quality classification dataset", mendeley data, V1, 10.17632/7t964jmmy3.1.
[2] A. Bhargava, A. Bansal, V. Goyal, P. Bansal, A review on tea quality and safety using emerging parameters, J. Food Measur. Characterization (2022) 1–21.
[3] P. Wang, E. Fan, P. Wang, Comparative analysis of image classification algorithms based on traditional machine learning and deep learning, Pattern. Recognit. Lett. 141 (2021) 61–67.
[4] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1) (1960) 37–46.
[5] A.R. Pathak, M. Pandey, S. Rautaray, Application of deep learning for object detection, Procedia Comput. Sci. 132 (2018) 1706–1717.
[6] P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of yolo algorithm developments, Procedia Comput. Sci. 199 (2022) 1066–1073.
[7] V. Ponnusamy, A. Coumaran, A.S. Shunmugam, K. Rajaram, S. Senthilvelavan, Smart glass: real-time leaf disease detection using YOLO transfer learning, in: 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE, 2020, pp. 1150–1154.
[8] F.M. Talaat, H. ZainEldin, An improved fire detection approach based on YOLO-v8 for smart cities, Neural Comput. Appl. 35 (28) (2023) 20939–20954.