JIEET (Journal of Information Engineering and Educational Technology)

Volume x, Number x, 2025 pp. xx-xx
E-ISSN : 2549-869X
Open Access: https://journal.unesa.ac.id/index.php/jieet

Pclassification Of Tobacco Leaf Quality Using The Yolov8, Yolov9, Yolov11, and Yolov12 Algorithm

Suwarno Arieska[1*], I Gusti Putu Asto Buditjahjanto [2], Wiyli Yustanti [3]
[1,2,3] Department of Informatics, Universitas Negeri Surabaya, Surabaya, Indonesia

Corresponding author: 2405190503@mhs.unesa.ac.id

ARTICLE INFO

Article history:
Submitted 23-6-2026
Accepted 4-8-2026 Available online 5-8-2026

Keywords:
Tobacco leaf quality detection, YOLOv8, YOLOv9, YOLOv11, YOLOv12, object detection, computer vision

DOI:
https://doi.org/10.26740/jieet.v10n1.p26-38

ABSTRACT

Tobacco leaf quality assessment requires an accurate and efficient detection system because the grading process directly affects quality control and the economic value of tobacco production. However, automatic classification remains challenging due to the high visual similarity between quality classes, substantial intra-class variations, and diverse image backgrounds. This study evaluates the performance of YOLOv8, YOLOv9, YOLOv11, and YOLOv12 for tobacco leaf quality detection. The research uses a dataset consisting of 839 tobacco leaf images representing eight quality classes, namely L1L, L2L, L3L, L3R, L4R, L10, L10F, and LND. The dataset was collected from tobacco-producing regions in Tanzania and processed for object detection using an image size of 640 × 640 pixels. Model performance was evaluated using Precision, Recall, F1-Score, mAP@0.5, and mAP@0.5:0.95. The experimental results demonstrate that YOLOv11 provides the most balanced overall performance among the evaluated architectures, achieving a Precision of 0.913, Recall of 0.915, F1-Score of 0.910, mAP@0.5 of 0.973, and mAP@0.5:0.95 of 0.776. Statistical analysis using the Chi-Square test further indicates significant differences among the models in Precision and mAP@0.5, while no significant differences are observed in Recall and mAP@0.5:0.95. These findings demonstrate that YOLOv11 is an effective architecture for tobacco leaf quality detection and has strong potential for supporting accurate and efficient automated tobacco grading systems.

INTRODUCTION

The importance of tobacco quality in global trade and the domestic cigarette industry is the main factor in influencing competitiveness, market absorption, and selling prices. The quality of tobacco leaves that meet industry standards is an absolute requirement for cigarette factories. Therefore, accurate and efficient tobacco quality detection is required. Grading is the process of classifying or classifying tobacco leaves based on their quality. The purpose of grading is to separate tobacco leaves into several classes so that each class has appropriate quality characteristics. This quality conformity is very important to ensure the consistency of the final product (cigarettes) and becomes the basis for determining the selling price.

The determination of tobacco quality is based on a series of physical and chemical factors. In general, the assessment parameters include the position of the leaves on the stem, color, leaf size, texture, and the presence or absence of defects. Several types of tobacco, aroma, and chemical content are also important criteria. The tobacco leaf datasets used to classify tobacco quality will be helpful in this research activity, especially in Africa, as the datasets will be accessible in open-source dataset repositories.

Conventional approaches for detecting tobacco leaf quality continue to depend heavily on manual assessment, where evaluation outcomes are largely influenced by the experience and subjective judgment of the assessor. This condition not only limits the consistency of quality evaluation but also demands substantial human effort and time. Furthermore, manual inspection is prone to inaccuracies that may lead to misclassification, ultimately affecting the overall quality control process of tobacco products. Advancements in computer technology have encouraged the adoption of image processing techniques for object detection and classification tasks. Through the utilization of manually engineered image features, these methods have been able to reduce the workload and processing time required in tobacco quality assessment. Nevertheless, traditional image processing approaches are generally designed for specific tobacco leaf characteristics, restricting their adaptability and effectiveness when applied to diverse tobacco varieties and conditions. The advancement of deep learning has significantly revolutionized the field of image analysis by facilitating automatic feature extraction through the use of Convolutional Neural Networks (CNN). In contrast to traditional approaches that depend on manually engineered features, deep learning models are capable of learning intricate visual representations directly from raw image data. This capability enables the extraction of more discriminative and robust features, leading to improved performance in image-based classification tasks. In recent years, a variety of machine learning and deep learning techniques, such as K-Means, Support Vector Machine (SVM), and Convolutional Neural Network (CNN), have been widely adopted for tobacco leaf quality classification. These methods have demonstrated considerable effectiveness in achieving rapid and accurate classification results, thereby enhancing the efficiency, consistency, and reliability of tobacco quality assessment systems.

The DMSO standard is used to identify the quality of tobacco leaves through visual image analysis of tobacco leaves (Novita Kurnia et al., 2019). The Support Vector Machine method in classifying tobacco quality through MATLAB software (Alfian Danu et al., 2023). The classification uses a naïve Bayes model in assessing tobacco quality (Dwi Sri et al., 2022). The GLCM and SVM methods are used at the digital image processing stage to classify the feasibility of tobacco leaf quality (Sriani et al., 2024). CNN's modeling technique in classifying leaf quality is by conducting the tobacco leaf sorting stage first, before modeling in the tobacco assessment process. Using the transfer learning strategy, they improved the classification of tobacco leaf quality by sorting tobacco leaves first to quickly collect tobacco leaf quality information (Nunik et al., 2021). The algorithms discussed previously belong to the category of two-stage object detection methods. These approaches operate by first generating candidate regions within an image that are likely to contain objects of interest, followed by a second stage that performs object localization and classification based on the corresponding tobacco quality categories. Although two-stage detection frameworks generally achieve superior detection accuracy due to their refined object proposal mechanism, they often require greater computational resources and exhibit slower inference speeds when compared with single-stage detection algorithms.

The enhanced YOLO-TS architecture, developed from the YOLOv8n model, has been utilized for tobacco stem detection tasks. Before model training, a collection of tobacco leaf images is gathered and processed through an annotation stage, where tobacco stems are labeled to establish a dedicated tobacco stem dataset (Chunjie Zhand et al., 2024). This approach enables the model to effectively learn the visual characteristics of tobacco stems and improve detection performance in real-world applications. Both YOLOv5 and YOLOv8 models have also been employed to classify fresh tobacco leaves according to their maturity levels, categorizing them into underripe, ripe, and overripe classes. To further enhance feature extraction capabilities, the proposed framework integrates the Squeeze-and-Excitation mechanism with ConvNet modules and constructs a small-sample learning model based on the stacked residual architecture of the standard module (David F Putradi et al., 2024). These improvements contribute to greater robustness and higher detection accuracy across varying environmental conditions. In addition, the YOLO-TobaccoStem-based detection system has been introduced to evaluate the loosening rate of tobacco leaves. To ensure reliable and consistent monitoring, a dedicated image acquisition system is designed to provide stable image capture throughout the observation process (Yansong Wang et al., 2025). Furthermore, several modifications have been incorporated into the YOLOv8 architecture to strengthen its object detection performance. To improve detection capability, the architecture is enhanced through the inclusion of dedicated feature extraction layers for small objects and a weighted bidirectional feature pyramid mechanism that facilitates more effective multi-scale feature fusion. As a one-stage detection model, YOLO streamlines the detection pipeline by generating object locations and class predictions directly from the input image without relying on an additional proposal generation stage. This integrated design reduces processing complexity, accelerates detection time, and enables efficient object recognition while maintaining reliable accuracy.

Previous studies have employed single-stage object detection models for tobacco leaf quality assessment; however, these approaches still exhibit limitations in terms of efficiency and overall performance. Consequently, there remains a need to develop more accurate and computationally efficient algorithms capable of detecting and classifying tobacco leaf quality. The YOLO family represents a widely adopted single-stage object detection framework due to its balance between speed and accuracy. In this research, the YOLOv8, YOLOv9, YOLOv11, and YOLOv12 architecture was utilized as the primary detection model for tobacco leaf quality evaluation. The key contributions of this study are summarized as follows:

(1) **Lightweight Detection Head:** Considering the diversity and complexity of tobacco leaf quality characteristics, a lightweight detection head is proposed to improve the model's detection capability while simultaneously reducing computational complexity and hardware resource consumption. This design enables the model to operate more efficiently without compromising detection accuracy.
(2) **Deep Supervision Technique:** A deep supervision strategy is introduced by integrating multiple detection heads for collaborative prediction. This approach enhances the model’s ability to capture and interpret complex quality-related features of tobacco leaves while avoiding additional inference costs, thereby improving overall detection effectiveness.
(3) **Dynamic Upsampling Module:** To further strengthen feature representation and multi-scale object detection performance, a dynamic upsampling module is incorporated into the framework. This module increases sampling adaptability and enhances feature reconstruction capabilities by leveraging the advantages of the YOLOv8, YOLOv9, YOLOv11, and YOLOv12 architectures, resulting in improved detection performance across varying tobacco leaf characteristics and image conditions.

METHOD

1. Datasets

The dataset utilized in this research consists of tobacco leaf images categorized according to the leaf positions commonly used by tobacco farmers in Tanzania. In total, the complete dataset contains 49,779 labeled images, where each label represents a specific tobacco grade. Detailed information regarding the number of images and their specifications is presented in Table 2. The dataset is organized into 22 folders, with each folder corresponding to a particular class label. Compared to many previously published tobacco leaf datasets, this collection is considerably larger. For instance, the study conducted by (X. Xin et al., 2023) employed 21,113 images, while (M. Xu et al., 2023) gathered 11,849 images, and (M. Lu et al., 2023) utilized 22,322 images. These studies did not incorporate leaf-position information into their classification schemes. In contrast, the dataset used in the present study comprises 839 images distributed across the classes L1L, L2L, L3L, L3R, L4R, L10, L10F, and LND. A detailed distribution of the tobacco quality classes is provided in Table 1.

Although tobacco-producing countries generally adopt similar principles for grading tobacco leaves, each country applies its own classification standards, resulting in differences in the labels assigned to tobacco quality grades (M. Lu et al., 2023). Consequently, tobacco leaves with comparable characteristics may receive different grade labels depending on the country in which they are evaluated. This variation reflects the influence of regional grading systems and terminology on tobacco quality assessment.

One illustration of these differences can be observed in leaf-position labeling. In China, upper-position tobacco leaves are represented by the letter B. For example, the label B2F refers to an upper tobacco leaf with quality grade 2 and an orange coloration. In Indonesia, a different naming convention is used, where upper-position leaves are commonly identified using the code TNG. Meanwhile, in Tanzania, upper tobacco leaves are designated by the letter L. As an example, the label L2O indicates an upper tobacco leaf classified as quality grade 2 with an orange color characteristic.

The class labels contained in this dataset follow the tobacco grading standards established in Tanzania. Each label begins with a letter indicating the position of the leaf on the tobacco plant, followed by a number representing the quality grade, and ends with a letter denoting the leaf color. Only three color categories are considered in this grading system, namely Orange (O), Lemon (L), and Mahogany (R). Researchers or practitioners from other countries may adapt these labels by converting them into their respective grading nomenclatures. Furthermore, certain grade labels may include the additional suffix F, which signifies special tobacco characteristics such as mature spots or fully ripened coloration. Figure 1 presents several examples of tobacco leaf grade labels, including L1L, L5R, and L1OF. Specifically, L1L represents a leaf-position tobacco leaf with preferred quality grade 1 and lemon coloration; L5R refers to a leaf-position tobacco leaf with quality grade 5 and mahogany coloration; and L1OF denotes a leaf-position tobacco leaf with preferred quality grade 1, orange coloration, and a special tobacco factor indicated by the suffix F (Faith Nguleni et al., 2024). The naming structure of Tanzanian tobacco class labels is summarized in Table 1.

During the current phase of data collection, one class label, namely L1R, was not successfully obtained. Additional data acquisition activities are planned in subsequent collection stages to include this missing category. Nevertheless, the absence of this particular class does not compromise the integrity or representativeness of the remaining tobacco quality classes that have already been collected and incorporated into the dataset.

(a) (b) (c)

Figure 1. Sample picture (a) L1L (b) L5R (c) L10F.

Table 1. Tanzania class label naming structure

Example of value label name: L	1	O	F
Part 1 (position)	**Part 2 (quality)**	**Part 3 (Color)**	**Section 4 (Special factors)**
Meaning of symbols	Meaning of symbols	Meaning of symbols	Meaning of symbols
C Cutter position	1 Quality selection	L Lemon	F Special factors (spots of maturity or full maturity or full orange color)
L Leaf position	2 Good quality	O Orange	
M Position of thin leaves	3 Good quality	R Mahogany	
X Lug and priming position	4 Reasonable quality		
	5 Low quality		

The tobacco leaf dataset utilized in this study was obtained through a collaborative research initiative involving NM-AIST in Arusha, Tanzania, TORITA located in Tabora, Tanzania, and the Tanzania Tobacco Board (TTB). The data collection process was conducted under the supervision and guidance of domain experts from TTB. In Tanzania, tobacco cultivation is concentrated within designated tobacco-growing zones and government-regulated agricultural areas (Faith Nguleni et al., 2024). Tobacco leaf images were gathered over a two-month period, from February 2024 to April 2024, involving four major tobacco-producing regions and a total of five farms within each region. The selected regions included Tabora, Uyui, Urambo, and Kaliua, which were chosen due to their significantly higher tobacco production levels compared with other tobacco-growing areas in Tanzania. Across these regions and participating farms, all tobacco grade classes were collected, although the number of samples varied among classes. As a result, the geographical variation within the dataset was relatively limited and was not considered to have a significant influence on the overall data distribution. For each class, the number of sample used the slovin method with an error of 10%. The research sample was calculated using equation 1 (Wahyudi et al., 2023).

$$s = \frac{N}{1+(N.e^2)} \quad (1)$$

where :
s = Number samples.
N = Total population
e = the percentage of chance due to sampling errors that are still tolerated is 10%

the number of detail of the research sample of tobacco quality class data is shown in table 2

Table 2. The amount of data used in the tobacco class study

Grade label	Research population	e =10%	Research sample
L1L	3570	97	105
L10	1413	93	104
L10F	2021	95	105
L2L	4914	99	105
L3L	7093	99	105
L3R	262	73	105
L4R	348	78	105
LND	187	66	105
whole	19808	700	839

2. **Performance evaluation methods**

In object detection studies involving tobacco leaf quality datasets, model performance is commonly evaluated using the mean Average Precision (mAP) metric, which has become one of the standard benchmarks for assessing detection effectiveness (Terven, J., 2023). This metric reflects a model's overall capability by simultaneously considering the accuracy of object localization and classification. Understanding mAP requires familiarity with three fundamental evaluation outcomes: True Positive (TP), False Positive (FP), and False Negative (FN), which are defined as follows:

- **True Positive (TP):** A detection outcome in which the model successfully identifies an object and accurately predicts its corresponding bounding box.
- **False Positive (FP):** An incorrect prediction generated when the model detects an object that does not actually exist or assigns an incorrect label to a detected object.
- **False Negative (FN):** A missed detection in which an existing object, represented by a ground-truth bounding box, is not recognized by the model.

To classify a prediction as either correct or incorrect, a quantitative evaluation criterion is required. One of the most frequently applied measures for this purpose is Intersection over Union (IoU). As described by (Terven, J., 2023), IoU evaluates the similarity between a predicted bounding box (Bp) and its corresponding ground-truth bounding box (Bgt) by comparing their overlapping region with the combined area covered by both boxes. A larger IoU value indicates a stronger correspondence between the predicted object location and the actual object position. The mathematical expression of IoU is provided in Equation (1), while Figure 5 presents a graphical illustration of the overlap concept used in its calculation.

$$\mathrm{IoU} = \frac{\mathrm{area(Bp \cap Bgt)}}{\mathrm{area(Bp \cup Bgt)}} \quad (2)$$

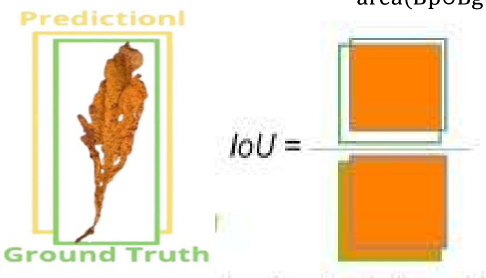

Figure 5. Illustration describing the calculation of the IoU

These metrics provide complementary insights into model performance by assessing the accuracy and completeness of object detection results. A detailed explanation of each evaluation metric is presented in the following sections.

- Precision: this shows how good the model is by looking at how many true positives it can find out of all the positive predictions (Padilla et al., 2020). Accuracy in math is given by Equation (3), where TP stands for true positives, and FP stands for false positives.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} = \frac{\mathrm{TP}}{\mathrm{all\ detections}} \quad (3)$$

- **Recall:** In object detection tasks, recall measures the ability of a model to capture and identify objects that truly exist within the dataset. According to (Padilla et al., 2020), this metric focuses on the extent to which positive samples are successfully detected.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP + FN}} = \frac{\mathrm{TP}}{\mathrm{all\ ground\ truths}} \quad (4)$$

- **mAP (mean Average Precision):** In object detection research, model performance across multiple classes is frequently assessed using the mAP metric. This metric is regarded as one of the most reliable indicators for comparing the effectiveness of detection algorithms and is extensively adopted in benchmark evaluations (Wang, B., 2022). The computation of mAP begins with the calculation of Average Precision (AP) for each class individually, after which the AP values are aggregated and averaged to obtain a single performance measure. AP corresponds to the area under the Precision–Recall relationship presented in Equation (5), providing an assessment of how well the model maintains both high precision and high recall. Within the AP computation framework, **M** denotes the total number of interpolation points used during the estimation process.

$$\mathrm{AP} = \sum_{\mathrm{m=1}}^{\mathrm{M}} (\mathrm{Rm} - \mathrm{Rm} - 1).\mathrm{P(Rm)} \quad (5)$$

In multiclass object detection tasks, model performance is frequently measured using mean Average Precision (mAP), which summarizes detection effectiveness into a single evaluation score. The metric is derived by calculating the Average Precision (AP) for each object category separately and subsequently averaging these values across all classes. As a result, mAP reflects the overall capability of a model to accurately identify and localize objects belonging to different categories within the dataset (Terven, J., 2023). Equation (6) presents the formulation of mAP, where **N** denotes the number of classes used in the evaluation process.

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{APi} \quad (6)$$

The assessment of detection performance in this research was carried out using an IoU threshold of 50%. Under this evaluation protocol, the metric is denoted as mAP50**,** which signifies that a detection is classified as valid when the overlap between the predicted and actual bounding boxes is at least 0.5. This metric was adopted as the principal benchmark for measuring and comparing the effectiveness of the detection architectures investigated in this study.

3. Research hypothesis

A hypothesis is a temporary explanation of a certain behavior, phenomenon or circumstance that has occurred or will occur. The function of the hypothesis is as a guideline to be able to direct research to match what we expect (Mudrajad Kuncoro., 2009).

Statistical testing was conducted to determine if there were performance differences between YOLO architectures based on evaluation metrics during the training and validation process.

For each metric (Precision, Recall, mAP50, mAP50-95, Train Box Loss, Train Classification Loss, Train DFL Loss, Validation Box Loss, Validation Classification Loss, and Validation DFL Loss), the hypotheses tested are:

Hypothesis nol (H_0):

There was no significant difference in the distribution of evaluation metric values between YOLOv8, YOLOv9, YOLOv11, and YOLOv12 architectures during the training process. Mathematically shown in equation 7 :

$$H0: MYOLOv8 = MYOLOv9 = MYOLOv11 = MYOLOv12 \quad (7)$$

with (M) expressing the median (or distribution) of the metric being tested.

Alternative hypothesis (H_1):

There is at least one YOLO architecture that has a significantly different distribution of evaluation metric values than the other architectures. Mathematically shown in equation 8:

$$H1: \exists i \neq j,\ Mi \neq \mathrm{Mj} \quad (8)$$

Decision-Making Criteria

- If the p-value ≤ 0.05, then H_0 is rejected, so there is a significant difference between the YOLO architectures in the metric being tested.
- If the p-value > 0.05, then H_0 fails to be rejected, so there is not enough statistical evidence to show that there is a difference between YOLO architectures in the metric.

RESULT

The machine learning platform used for the model evaluation uses Google Colaboratory with a library of hardware specifications and training configurations shown in Tables 3 and 4.

Table 3. Hardware specifications

Parameter	Configuration
CPU	CPU Intel Xeon @2.20Ghz
GPU	Intel Core
GPU memory size	16 G
Operating system	Windows 10
Deep learning architecture	python-3.12.12 pytorch-2.9.0

Table 4. Training configuration

Parameter	Configuration
Batch size	5 groups
Image size	640 x 640
Optimizer	AdamW
Early learning level	0.05
Final learning	0.05
Momentum	0.9
Epoch	100

The larger the area, the better the performance of the PR curve comparison reveals that the classification of tobacco class using YOLOv11 achieves excellent detection efficiency compared to the YOLOv8, YOLOv9, and YOLOv12 versions. The results of the tobacco class dataset exercise using YOLOv8, YOLOv9, YOLov11, and YOLOv12 are shown in Table 5.

Table 5. Comparison of the performance of tobacco grade detection algorithms.

Model	*Layer*s	P	R	mAP50	mAP50:90	F1
YOLOv8n.pt	168	0,925	0,911	0,965	0,752	0,910
YOLOv9c.pt	168	0,961	0,888	0,955	0,765	0,920
YOLOv11n.pt	168	0,913	0,915	0,973	0,776	0,910
YOLOv12n.pt	168	0,868	0,954	0,971	0,778	0,910

DISCUSSION

1. Hypothesis results

   From the results of the research that has been carried out, the results of the hypothesis of the training process are obtained which are displayed in the form of hypothesis testing results and the results of the Post Hoc Dunn Test.

1.1 Hypothesis test results

The results of the hypothesis test are shown in table 6 using the results of the Chi-Square (χ^2) test on various model evaluation metrics with the aim of finding out whether there are significant differences between the models or groups being compared. Decision-making is based on a p-value with a significant level of $\alpha = 0.05$. If the p-value < 0.05, then the null hypothesis (H0) is rejected showing a significant difference. On the other hand, if the p-value ≥ 0.05, then the HO fails to be rejected which identifies no significant difference

Table 6. Summary of Hypothesis Test Results

Evaluation Group	**Metric**	**x²**	**p-value**	**Verdict**	**Results**
Performance Metrics	Precision	12,244	0,0066	Minus H_0	Significant differences
	Recall	2,417	0,4910	Failed to reject H_0	No different
	mAP50	15,884	0,0012	Minus H_0	Significant differences
	mAP50-95	5,857	0,1190	Failed to reject H_0	No different
Training Metrics	Train Box Loss	3,147	0,3700	Failed to reject H_0	No different
	Train CLS Loss	1,278	0,7340	Failed to reject H_0	No different
	Train DFL Loss	2,771	0,4280	Failed to reject H_0	No different
Validation Metrics	Validation Box Loss	23,300	<0.001	Minus H_0	Significant differences
	Validation CLS Loss	5,205	0,1570	Failed to reject H_0	No different
	Validation DFL Loss	41,791	<0.001	Minus H_0	Significant differences

1.2 Post Hoc Dunn Test Results

The results of the Post Hoc Dunn Test conducted after conducting the Chi-Square test showed significant differences in several evaluation metrics. The purpose of Post Hoc analysis is to identify model pairs that have significant differences so that it can be known which model is the main cause of these differences. In table 4.2, the models compared include YOLOv8, YOLOv9, YOLOv11, and YOLOv12 while the interpretation is given based on the p-value of each model pair. The results of the Post Hoc Dunn Test are shown in table 7.

Table 7. Post Hoc Dunn Test Results

Metric	**Significantly Different Pairs**	**Interpretasi**
Precision	YOLOv12–YOLOv8 (p=0.0369); YOLOv8–YOLOv9 (p=0.0173)	YOLOv8 has different precision than YOLOv9 and YOLOv12.
mAP50	YOLOv11–YOLOv9 (p=0.0019); YOLOv8–YOLOv9 (p=0.0081)	YOLOv9 shows different mAP50 performance than YOLOv8 and YOLOv11.

Validation Box Loss	YOLOv12–YOLOv8 ($p<0.001$); YOLOv8–YOLOv9 ($p<0.001$)	YOLOv8 has different box loss validation characteristics than YOLOv9 and YOLOv12.
Validation DFL Loss	YOLOv11–YOLOv12; YOLOv11–YOLOv8; YOLOv11–YOLOv9; YOLOv12–YOLOv8; YOLOv8–YOLOv9	DFL Loss Validation is the metric that most distinguishes the four architectures.

CONCLUSION

Based on the results of the research that has been conducted, it can be concluded that

1. The YOLO model approach with YOLOv8, YOLOv9, YOLOv11 and YOLOv12 versions has proven to be very effective in classifying the quality of tobacco leaves. The application of the YOLOv11 algorithm in this study proved to be very effective and this model showed significant results in classifying the quality of tobacco leaves. Specifically, the YOLOv11 model tested on a *dataset* of 839 images managed to produce solid performance parameters with a *precision* value of 0.913, an F1 score of 0.915 and a value of mAP@0.5; 0.95 is 0.776. The main advantage of YOLOv11 compared to the YOLOv8, YOLOv9, and YOLOv12 versions lies in the accuracy metric in classifying tobacco leaves as a whole, reaching a value of mAP@0.5 with a figure of 0.973.
2. The results of the analysis of ten metrics grouped into performance metrics, Training metrics and validation metrics, obtained four metrics that showed significant differences and six metrics that did not show significant differences. In the performance metrics group, two of the four metrics with a 50% figure showed significant differences, especially in the aspects of precision and mAP50, while recall and mAP50-95 showed no significant differences. In the Training Metrics group, three metrics showed insignificant results with a percentage of 0%. So the learning process between models can be said to be relatively similar. Meanwhile, in the validation metrics group, two of the three metrics showed significant differences with a figure of 66.7% which identified variations in generalization capabilities between YOLOv8, YOLOv9, YOLOv11, and YOLOv12 architectures. Overall. The percentage of the metric shows a significant difference with the 40% figure. Thus, the hypothesis of this study can be declared partially accepted due to differences in model architecture that have been shown to have a significant influence on several performance and validation indicators but do not show a consistent influence on all evaluated metrics.

REFERENCES

A. Harjoko, A. Prahara, TW. Supardi, I. Candradewi, R. Pulungan, S. Hartati, *image processing approach for grading tobacco leaf based on color and quality*, int, J, *Smart sens*, Intell syst.12 (1) (2019) 1-10, doi: 10.21307/ijsis-2019-010.

Alfian Danu Ismail, Danang E, Iska Y. 2023. Klasifikasi Kematangan Daun Tembakau Virginia Menggunakan Pengolahan Citra Digital. Journal Elektro Kontrol VOL. 3 NO.1 (2023). DOI: https://doi.org/10.24176/elkon.v3i1.10131.

AnonTTB, Tanzania *Tobacco Board*, 2023 https://www.tobaccoboard.go.tz/tobacco-growing-areas diakses pada tanggal 22 mei 2025.

Chunjie Zhang, Lijun Yun, Yansong Wang, Ruilin Luo, YOLO-TS: *Detection Loosening Rate of Tobacco Bundles Based on* YOLOv8, Januari 2024. *Applied Engginering in Agriculture* 40(6):623-634. Doi:10.13031/aea.16020.

D. F. Putradi, B. Setiawan and A. D. W. Sumari, "*Real time Tobacco Leaf Classification Machine Using the Computer vision with YOLO Approach on Jetson Nano," 2024 International Conference on Electrical and Information Technology (IEIT)*, Malang, Indonesia, 2024, pp. 19-24, doi: 10.1109/IEIT64341.2024.10763208.

Dewi, S. R. 2018. *Deep Learning Object Detection* Pada Video Menggunakan *Tensorflow* dan *Convolutiona*l *Neural Network*. Yogyakarta: Universitas Islam Indonesia.

Dwi Sri Anggun, *Classification of Harvestable Tobacco Quality with Naïve Bayes Classifier, September 2022. Journal of Information Technology and Computer Engineering* 3(2):46-52. doi: 10.37148/bios.v3i2.47.

Faith Nguleni, et al., *Dataset of* Virginia *Flue-cured Tobacco leaf images based on stalk leaf position for classification tasks: A case of* Tanzania. *Data in brief* 56 (2024) 110817 https://doi.org/10.1016/j.dib.2024.110817.

Iska Yanuartanti, *Classification of Maturity of Virginia Tobacco Leaves Using Digital Image Processor, Journal of Electro Control* (ELKON) https://doi.org/10.24176/ELKON.V3I1.10131.

J.L. Hintze and R.D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism," Am Stat, vol, 52, no.2,pp.181-184,1998, doi: 10.1080/00031305.1998.10480559.

Jalled, F., & Voronkov, I. 2016. *Object Detection Using Image processing*. Ithaca, New York: Cornell *University Library*.

K.Dormitius, *Impact of Tobacco Curing on the Environment Andsocio-Economic Aspects in the Urambo District*, 2021 [Online]. Available: https://dspace.nm-aist.ac.tz/handle/20.500.12479/1316.

L.Zhang, et al., *Metabolic profiling of tobacco leaves at different growth stages or different stalk positions by gas chromatography-mass spectrometry*. Ind. Crops Prod.116 (February) (2018) 46-55, doi:10.1016/j.indcrop.2018.02.041.

M. Lu, et al., *Intelligent grading of tobacco leaves using an improved bilinear convolutional Neural Network*, IEEE *Access* 11 (May) (2023) 68153-68170, doi: 10.1109/*ACCESS*.2023.3292340.

M. Xu, J. Gao, Z.Zhang, X. Guo, *Multi-channel and multi-scale separable dilated Convolutional Neural Network with attention mechanism for flue-cured tobacco classification*, *Neural comput*. *Appl*. 35 (21)(2023) 15511-15529, doi:10.1007/s00521-023-08544-7.

Mudrajad Kuncoro. 2009. Metode Riset Untuk Bisnis dan Ekonomi. Jakarta: Erlangga.

N. Mdurna, J. Leo, *Dataset of banana leaves and stem images for object detection, classification and segmentation: a case of* Tanzania, Data Br.49 (2023) 109322, doi:10.1016/j.dib.2023.109322.

Novita Kurnia, Septian Enggar Sukmana., *Classification of tobacco leaves according to the quality of single-leaf processing using backpropagation*, November 2019. *Symmetrical Journal of Electrical Mechanical Engineering and Computer Science* 10(2): 479-486. doi: 10.24176/simet.v10i2.3066.

Nunik eka diana, *Color-Based Visualization-Based Modeling Techniques for Virginia Tobacco Grading and Sorting Transparency*, November 2021. *Perspective* 20(1):50. doi:10.21082/psp.v20n1.2021.50-62.

Padilla, R., Netto, S.L., Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals, and

Image Processing (IWSSIP), Niterói, Brazil, 1–3 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 237–242. [CrossRef].

R. Megill, J.W.Turkey, and W.A. Larsen, "Variations of Box Plots," AmStat, vol,32, no.1, pp. 12-16, 1978, doi: 10.1080/00031305.1998.10480559.

Redmon, J., & Farhadi, A. (2016). *YOLO9000: Better, faster, stronger*. arXiv. https://arxiv.org/abs/1612.08242.

Redmon, J., & Farhadi, A. 2018. YOLOv3: *An Incremental Improvement.* arXiv. https://doi.org/10.48550/arXiv.1804.02767.

Redmon, J., Divvala, S., Girshick, R., & Farhadi Ali. 2016. *You Only Look Once: unified, Real-Time Object Detection.* arXiv. https://doi.org/10.48550/arXiv.1506.02640.

Roboflow (2025) https://roboflow.com/. Accessed on date 30 juli 2025.

Sapkota, R.; Meng, Z.; Churuvija, M.; Du, X.; Ma, Z.; Karkee, M. Comprehensive performance evaluation of YOLO11, YOLOv10, YOLOv9, and YOLOv8 on detecting and counting fruitlet in complex orchard environments. arXiv 2024, arXiv:2407.12040.

Sriani, Sullindawaty, Yulia Rizky, *Classification of tobacco leaf quality using GLCM (Gray Level Co-occurrence Matrix) and SVM (Support Vector Machine)*. Journal Informatika dan Teknik Elektro Terapan. https://doi.org/10.23960/jitet.v12i3.4599.

Sugiyono. 2019. Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta.

Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A.; Chavez-Urbiola, E.A. Loss functions and metrics in deep learning—a review. arXiv 2023, arXiv:2307.02694.

Ultralytics (2025) https:// github.com/ultralytics. Accessed on date 30 juni 2025.

Ultralytics (2025a) YOLOv8. https://docs.ultralytics.com/models/YOLOv8/ Accessed on date 22 juli 2025.

Ultralytics (2025b) YOLOv9. https://docs.ultralytics.com/models/YOLOv9/ Accessed on date 22 juli 2025.

Ultralytics (2025c) YOLOv11. https://docs.ultralytics.com/models/YOLOv11/ Accessed on date 22 juli 2025.

Ultralytics (2025d) YOLOv12. https://docs.ultralytics.com/models/YOLOv12/ Accessed on date 22 juli 2025.

Umar M, Altaf S, Ahmad S, mahmoud H, Mohamed ASN, Ayub R. 2024. *Precision agriculture through deep learning: tomato plant multi-disease recognition with CNN and improved YOLOv7*. IEEE Access.Digital Object Identifier 10.1109/ACCESS.2024.3383154.

Wahyudi, Widiya Avianti, Afrizal Martin, Jumali, Novita Andriyani, Diah Prihatiningsih, Dian Misesani, Fahrudin, Marianus Yufrinalis, Maria Angelina Fransiska Mbari, Arum Gati Ningsih, Aries Yulianto, M Taufiq Noor Rokhman, Aridhotul Haqiyah, Tatan Sukwika. 2023. Metode Penelitian. Sumatera Utara: PT. Mifandi Mandiri Digital.

Wang, B. A parallel implementation of computing mean average precision. arXiv 2022, arXiv:2206.09504.

X. Xin, H. Gong, R. Hu, X. Ding, S. Pang, Y.Che. *Intelligent large-scale flue-cured tobacco grading based on deep densely convolutional*,Sci, Rep. 13 (1)(2023)11119, doi: 10.1038/s41598-023-38334-z.

Yansong Wang, Chunjie Zhang, Mingjie Wu, Rullin Luo, Lin Lu, Zaiqing Chen and Lijun Yun, *A Method For Detecting the rate of Tobacco leaf Loosening in Tobacco leaf sorting scenarios, frontiers in plant science, June* 2025, doi: 10.3389/fpts.2025.1578317.