SCIENTIA AGROPECUARIA

Scientia Agropecuaria

Web page: http://revistas.unitru.edu.pe/index.php/scientiaagrop

Facultad de Ciencias Agropecuarias

Universidad Nacional de Trujillo

RESEARCH ARTICLE

Applying artificial intelligence in durian fertile lobe detection: Attention-Residual Unet and Test Time Augmentation algorithm

Thanh Tung Luu[1, 2]; Nhat Quang Cao[1, 2*]

[1] Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology, 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam.

[2] Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam.

* Corresponding author: quang.cao2705@hcmut.edu.vn (N. Q. Cao).

Received: 21 April 2025. Accepted: 18 July 2025. Published: 8 August 2025.

Abstract

The key factor in durian fruit trading is ripeness. Several studies have been conducted on non-destructive durian maturity classification using near-infrared (NIR) spectroscopy. However, most of these studies manually determined the most accurate measurement position, which was the durian's main fertile lobe center. This research aims to automate the stage of detecting this position of the durian by using UNet segmentation method, which leverages differences in rind texture between the center of the main fertile lobe and other areas (lobe grooves and stems), prior to conducting NIR measurements. The rough and non-uniform surface of the durian rind presents a significant challenge for segmentation. However, the large size of the durian spines in the main fertile lobe serves as an identification characteristic for the segmentation model. This study uses the Ri-6 durian in Vietnam as the samples for the experiment. The model was developed using three architectures: Unet, Attention-Unet and Attention-Residual Unet. According to the analysis results on test set, Unet, Attention-Unet and Attention-Residual Unet algorithms achieved %accuracy of 78.22%, 81.34%, 82.89% and %intersection over union of 79.49%, 80.47%, 80.72%, respectively. After that, the model was further enhanced using the test time augmentation algorithm, improving the %accuracy to 85.24%, 85.68%, 86.85% and %IoU to 81.65%, 82.03% and 83.12%. Among the three architectures, the Attention-Residual-Unet model demonstrated the highest efficiency in detecting the center of the durian's main fertile lobe for non-destructive durian maturity classification. This method can be applied to the development of an automatic durian's maturity classification machine, which would save time and improve economic efficiency.

Keywords: Durian; fertile locule's center; Unet; Att-Unet; Att-ResUnet; Test time augmentation.

DOI: https://doi.org/10.17268/sci.agropecu.2025.038

Cite this article:
Luu, T. T, & Cao, N. Q. (2025). Applying artificial intelligence in durian fertile lobe detection: Attention-Residual Unet and Test Time Augmentation algorithm. *Scientia Agropecuaria*, *16*(4), 499-511.

1. Introduction

The durian tree is a highly valuable fruit crop that plays a significant role in the agricultural economy of countries with climatic conditions suitable for its cultivation, particularly those in the ASEAN region. In Vietnam, the total area dedicated to durian cultivation reached 30,000 hectares (Ngoc et al., 2024), contributing to the country's global agricultural exports, which amounted to 21.82 billion US dollars (Likhitpichitchai et al., 2023), and this is estimated to increase rapidly in the future. Therefore, the demand for higher quality and productivity is expected to increase as well.

One method to increase the value of the durian market is to accurately classify durians according to their ripeness. Various approaches have been explored in this field, with non-destructive techniques using NIR spectroscopy, such as the study by Ditcharoen et al. (2023), demonstrating a high efficiency. Although NIR spectroscopy has proven to be effective, the measurement position is still selected manually. Although this approach is feasible for a small number of durians, it becomes impractical for large-scale applications due to its slow processing speed.

The NIR method has been applied to multiple individual positions on durian fruit, with the middle position of fertile locules and stems yielding the most accurate results among the single measurement locations (Puttipipatkajorn et al., 2023). However, some fruits lost their stems before classification. Therefore, the fertile locule center (FLC) is the most suitable location for classification. During the durian development process, the shape of the fruit is divided into two main parts (excluding the stem): fertile lobe and lobe groove (Figure 1). In this research, the special shape of the durian fertile lobe is used as the main factor to detect its center. A fertile lobe refers to a durian locule that appears visibly full and completely developed along the entire length of the fruit (Glinpratum, 2003). Between each pair of locules, a suture extends from the stem to the stylar end. This suture represents the dehiscence zone, where the ovary walls of individual carpels fuse during early flower development to form the locules. Ultimately, this is the region where the fruit naturally splits or dehisces (Siriphanich, 2011). As the fruit matures, the spines along the suture at the groove turn brown or even black (Siriphanich, 2011). In the fertile lobes, the grooves among the spine bases expand and also be darker (Pascua & Cantila., 1992).

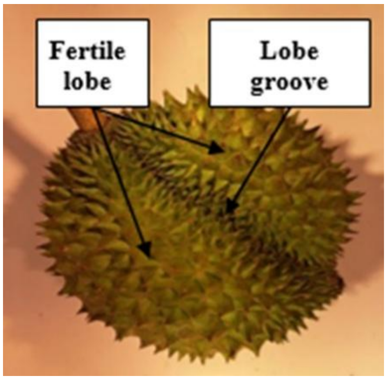


Figure 1. The shape of durian.

Based on the lobe morphological characteristics above, the segmentation algorithm is a potential method to determine the FLC point. The Unet model has been proven effective in various studies on image segmentation, with applications in fields such as healthcare, agriculture, and industrial products, among others. Specifically, in agriculture, this model has been studied and found to be well-suited for fruit image segmentation when usually achieve the accuracy over 90% (Mane et al., 2023). Moreover, The Unet model is notably fast and efficient to train, as it is less susceptible to overfitting compared to most other segmentation models. Additionally, it supports contextual learning, further enhancing its performance (Rehman & Rehman, 2022).

According to the literature review, Unet model and its variations have the potential to be applied to segment the center zone, which is used to measure the durian maturity. The variations in shape and color of the durian rind will undergo an image preprocessing stage before being fed into the model to enhance the Unet models feature extraction capability. This step is particularly important due to the rough texture of the durian rind, which poses challenges for segmentation. Although Unet is a promising model, it also has several shortcomings. The direct skip connections in Unet create a significant semantic disparity between the inputs of the two convolutional layers. Furthermore, these skip connections link feature maps at the same scale without accounting for the relationships between feature maps across different stages. As a result, the feature representations may lack consistency (Qurri & Almekkawy, 2023). Attention (Att) and Attention-Residual (Att-Res) mechanisms are utilized to enhance accuracy and address the limitations of the original Unet model. The Att model enhances critical features while suppressing less relevant ones through the use of channel attention and spatial attention modules (Li et al., 2025). In the research of Li et al. (2025) for sugar beet and weed segmentation, the intersection over union (IoU) ratio of Attention-Unet (Att-Unet) model is 98.76%, 89.69% and 60.63% compared to 98.64%, 89.47%, 55.49% of original Unet, respectively. The mean intersection over union (mIoU) index of Att-Unet (83.03%) is also higher than that of Unet model (81.20%). Additionally, the residual architecture, when integrated with the U-Net network, demonstrates an improved ability to enhance segmentation accuracy compared to the standalone U-Net model. When using Residual-Unet (ResUnet), the residual unit simplifies network training and the skip connections within a residual unit, as well as those linking low and high levels of the network, facilitate information propagation without degradation. This enables the design of a neural network with significantly fewer parameters while achieving comparable or even superior performance in semantic segmentation (Zhang et al., 2018). The ResUnet show the greater breakeven point than Unet model (0.9187 compared to 0.9053) so it has a better performance in precision and recall (Zhang et al., 2018). Due to Rehman et al. (2023), the segmentation experiment can achieve the best IoU index (81.16%) when using Att-ResUnet (between Unet, Att-Unet and Att-ResUnet). Although Att-Unet and Att-ResUnet demonstrate several advantages, they also introduce increased computational complexity and a higher risk of overfitting, which can result in lower accuracy compared to the traditional UNet model in certain cases. This was observed in the segmentation models for Colonoscopy, Liver, and Heart MRI datasets in the study by Weng et al. (2023), where the traditional UNet outperformed its variants, including Att-Unet, in terms of both accuracy and performance. Therefore, these variants were evaluated with different loss functions to ensure the selection of the model that achieves the highest possible accuracy. After training the model, to enhance accuracy, test time augmentation (TTA) method is applied before the model makes a segmentation decision by averaging the outputs of multiple images to produce the most precise prediction. Finally, the center determination algorithm is used to detect the FLC point.

In the following sections, this study will explore the architectures of the models used, the sample collection methods for the dataset, and finally, an evaluation of the models' performance after training. This analysis will help assess the feasibility of identifying FLC using U-Net and its variations.

2. Model architecture

The working principle of the model follows Figure 2. The input of the model is the folder of durian images and the respective masks. The output is the coordinates of the FLC location. The training stage is experimented with three architectures: Unet, Attention-Unet and Attention-Residual Unet.

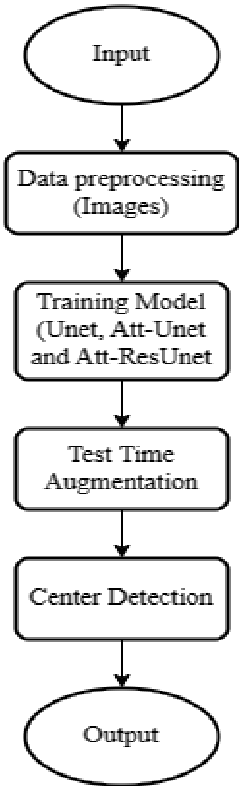

Input
Data preprocessing (Images)
Training Model (Unet, Att-Unet and Att-ResUnet
Test Time Augmentation
Center Detection
Output

Figure 2. Model architecture.

2.1 Data preprocessing

In this stage, the collected images will undergo a preprocessing algorithm in JPEG format to enhance the model's learning capability and improve accuracy. Subsequently, background removal is performed by extracting contours from the input image and preserving only the largest contour, which represents the durian fruit. All of the input data are scaled into 256×256 size. After that, all of the photos are moved to the contrast limited adaptive histogram equalization (CLAHE) stage. CLAHE applies a lower threshold when modifying the histogram, enhancing image quality by improving contrast while preventing noise over-enhancement and minimizing edge shadowing (Saifullah & Drezewski, 2023). The limit value is determined using the clip limit derived from the equation (1) below.

$$\beta = \frac{M}{n}\left(1 + \frac{a}{100}(s_{max} - 1)\right) \quad (1)$$

Where M denotes the region's area size, n represents the 8-bit grayscale value (0–255) and α defines the clip factor threshold applied to the histogram (0–100). After that, the dataset is divided into two main parts: 10% for the testing set and 90% for training and validation. Of this 90%, 90% is allocated for training set, while the remaining 10% is used for validation.

2.2 Unet

The network exhibits a symmetrical design, consisting of an Encoder that extracts spatial information from the input image and a Decoder that reconstructs the segmentation map from the encoded features (Figure 3). The Encoder adopts a conventional convolutional neural network structure, with two 3 × 3 convolutional layers followed by 2 × 2 max pooling with stride 2. This pattern repeats four times, doubling the number of filters at each stage. Two final 3 × 3 convolutional layers then serve as a bridge to the Decoder (Ibtehaz & Rahman, 2019).

Conversely, The Decoder reconstructs the output image using Encoder features. Each block mirrors the Encoder’s structure but replaces max pooling with a 2 × 2 transposed convolutions to upsample and halve the feature channels.

In detail, 3×3 filters enable the network to learn essential features from the input image. They enhance feature extraction, preserve image details, reduce the number of parameters compared to 5×5 and 7×7 filters, improve non-linearity, and facilitate effective connections between the Encoder and Decoder. Furthermore, for capturing the features non-linearity, Relu activation function is almost all blocks in the network (Enshaei et al., 2020) except the output layer at the end of the model, where a sigmoid activation function is utilized to assign a smooth weight to each channel and produce the corresponding weight value as output (Men et al., 2021). The formula of Relu and Sigmoid are presented by functions (2) and (3):

$$Relu\colon f_{(x)} = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases} \qquad (2)$$

$$Sigmoid\colon f_{(x)} = \frac{1}{1+e^{-x}} \qquad (3)$$

Besides, Max pooling in the encoder reduces spatial dimensions while preserving key features, improving efficiency. Transposed convolution in the decoder restores resolution by interpolating missing details for accurate segmentation. U-Net also uses concatenation to combine decoder context with encoder features, enhancing segmentation performance (Sulaiman et al., 2024).

Finally, each block incorporates a dropout function to prevent overfitting. The dropout rate is set at a low range of 0.2 to 0.3 to preserve essential features, considering the limited quantity and high complexity of the input data.

2.3 Attention-Unet

Expanding on the original Unet framework, the Attention-Unet architecture improves performance by integrating Attention Gate (AG) layers (Figure 4). It includes four encoder and four decoder blocks, with attention gates before each decoder block to emphasize relevant encoder features and suppress irrelevant ones. Although this refinement enhances segmentation accuracy, it also necessitates a careful balance between computational cost and the potential risk of overfitting (Sulaiman et al., 2024).

The Attention Gate (AG) uses coarse-scale information to suppress irrelevant signals and noise in skip connections before concatenation, retaining only meaningful activations. It also refines neural responses during training by downweighting gradients from background regions, focusing updates on task-relevant areas (Oktay et al., 2018).

2.4 Attention-Residual Unet

The Att-Unet architecture shows great potential in extracting complex features from datasets. However, the durian shell presents a significant challenge due to its rough and irregular surface, making feature extraction less stable and consistent. Although increasing network depth can enhance the performance of a deep neural network, it may hinder training and lead to degradation issues (Zhang et al., 2018). He et al. (2016) proposed the residual neural network to facilitate training and address the degradation problem.

The residual network consists of a series of stacked residual units. Each unit includes a convolutional block (F_{x1}) and an identity mapping (X_1) running in parallel, which are combined at the block's output (Figure 5a, 5b). Aside from the similarities in the convolutional blocks compared to Unet and Att-Unet, Batch Normalization in the Att-ResUnet architecture (Figure 6) plays a crucial role in stabilizing training, accelerating convergence, mitigating internal covariate shift, reducing the vanishing gradient problem, and minimizing overfitting, ultimately enhancing the model's overall performance.

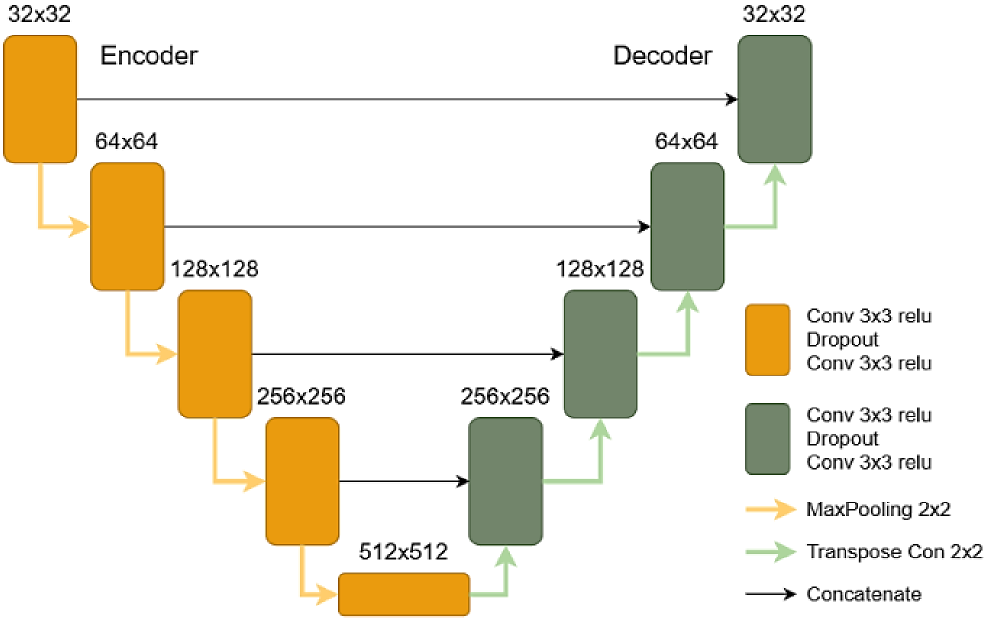


Figure 3. Unet architecture.

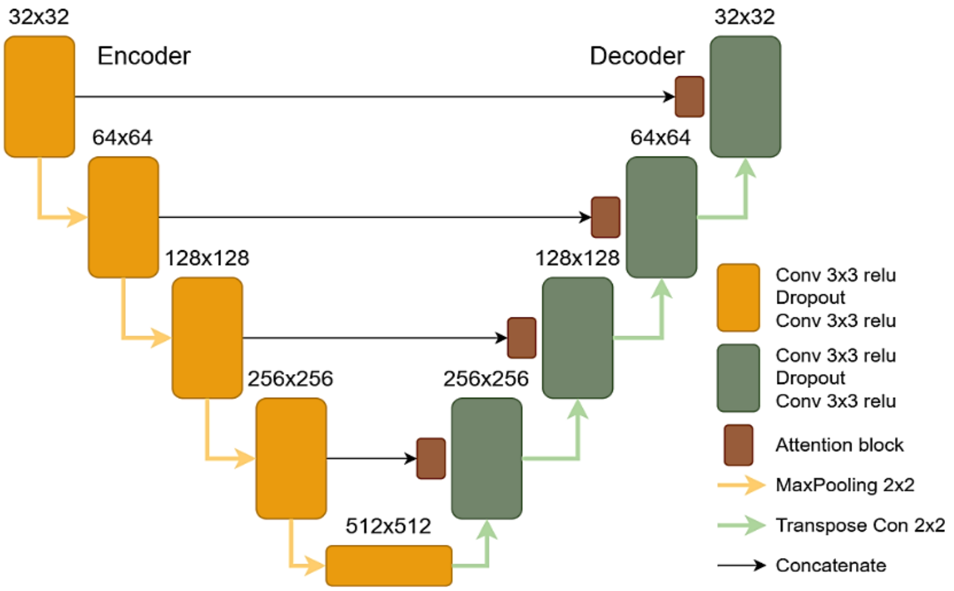


Figure 4. Attention-Unet architecture.

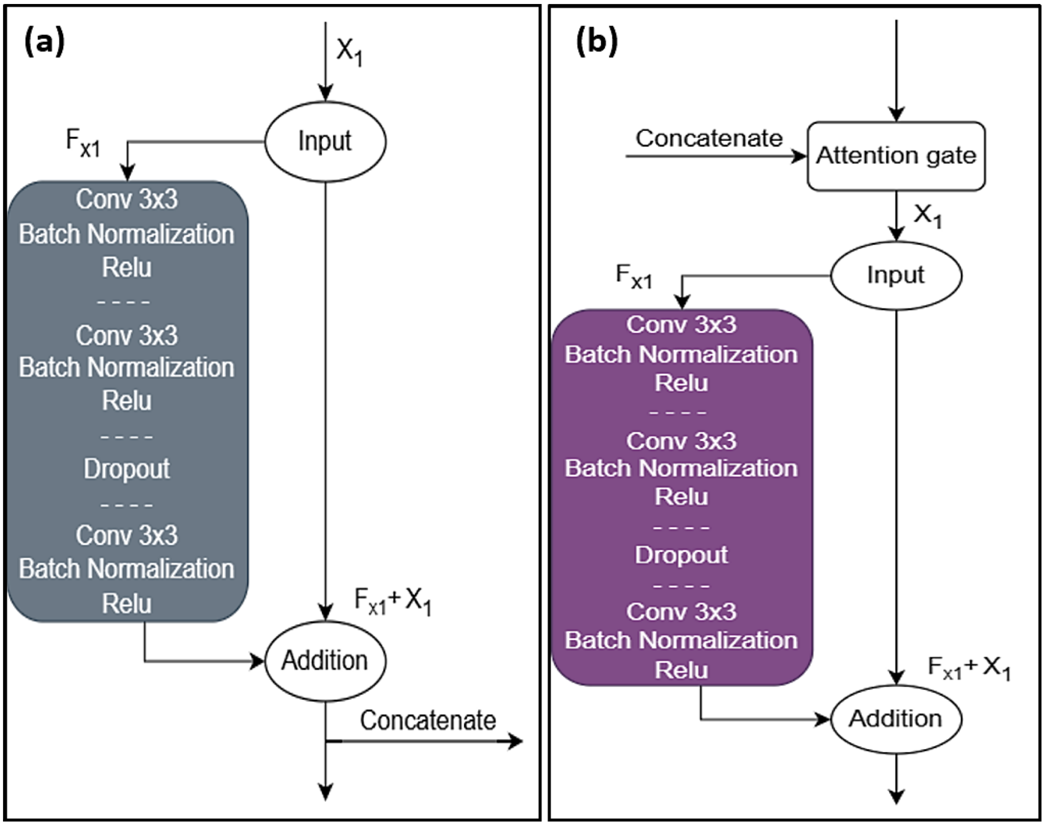


Figure 5. a) Encoder Residual Block, b) Decoder Residual Block.

2.5 Test Time Augmentation and Center Detection

After the training process, the output results still have noise. It can have negative affect to the prediction. Therefore, the test time augmentation (TTA) algorithm is applied to denoise for the final output. TTA includes four procedures: augmentation, prediction, dis-augmentation and merging (Moshkov et al., 2020). During the augmentation phase, the input images are transformed into multiple orientations (four times in this study). The augmented images are fed into the trained model for prediction. Afterward, a de-augmentation process restores them to their original orientations, and the results are merged to generate the final output.

In U-Net, the merging is straightforward since individual instances are not distinguished. De-augmented probability maps are averaged, yielding a floating-point image that is then converted into a binary mask (Moshkov et al., 2020). Applying element-wise thresholding at 0.47 effectively transforms the soft masks into binary masks

After the TTA stage, OpenCV is employed as a detection algorithm to extract the contour of the fertile lobe and determine its center coordinate, which represents the FLC position.

2.6 Loss Functions

Another crucial component of a neural network is the loss function. In the study of Neyestanak, et al., (2022), Binary Cross Entropy (BCE) and Binary Focal Loss (BFL) demonstrated high performance, consistently achieving an accuracy of over 98% for image segmentation. Meanwhile, in the segmentation research of Montazerolghaem, et al. (2023), Dice Loss (DL) proved to be more effective than BCE and BFL. In this study, all three loss functions are evaluated to identify the one that achieves the highest accuracy for FLC identification. Their general formulas are presented below, with specific parameters detailed separately:

g_i , g_i : voxels i in ground truth and segmentation output, respectively.

C: the number of classes.

c: notation for an individual class. If class c is the correct classification for voxel i, g_i^c is equal to 1 and s_i^c is the corresponding predicted probability.

N: the total number of samples.

Initially, BCE loss is a widely used metric in deep learning, measuring the dissimilarity between probability distributions based on cross-entropy and the training data characteristics (Montazerolghaem et al., 2023). The mathematical formulation of the BCE loss function is:

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c}^{C} g_i^c \log(s_i^c) \qquad (4)$$

Secondly, BFL function addresses class imbalance by down-weighting well-classified samples and emphasizing harder examples through a scaling mechanism.

$$BFL = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c}^{C}(1 - s_i)^{\gamma} g_i^c \log(s_i^c) \qquad (5)$$

where hyperparameter γ is a focusing parameter.

The last loss function is DL. It is formulated to maximize the Dice coefficient directly. Models employing Dice loss have exhibited superior effectiveness in binary segmentation.

$$Dice = -\frac{2\sum_{i=1}^{N} s_i g_i}{\sum_{i=1}^{N} {s_i}^2 + \sum_{i=1}^{N} {g_i}^2 + \varepsilon} \qquad (6)$$

where ε is a small number to avoid division by zero. Those are all the necessary algorithms before proceeding to the experimentation phase.

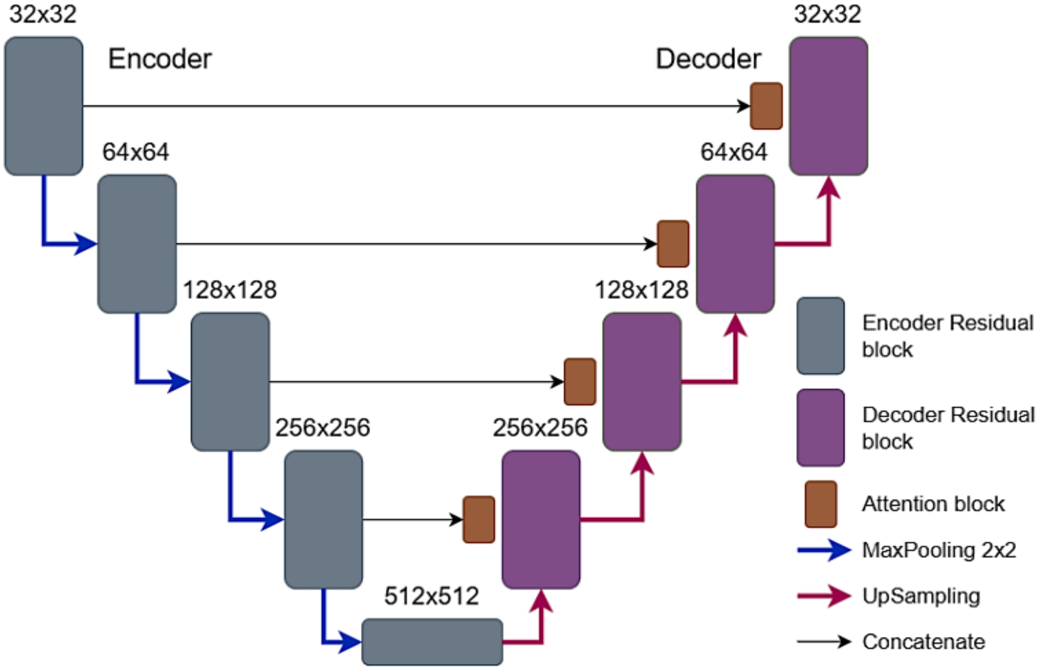


Figure 6. Attention-Residual architecture.

3. Methodology

3.1 Sample collection

The durian fruit sample for the experiment is the Ri-6 variety, which come from Tam Binh Commune, Cai Lay District, Tien Giang Province, Vietnam. Total of 80 fruit samples were collected from many different age periods (80 to 130 days), which were classified by the orchard's host, to ensure that the final model is suitable to all of these age periods (unripe and ripe). The samples were placed in the room temperature at 27 ± 2 °C for 1 day to ensure that all fruits are in the same condition before conducting the collection of image samples. After that, the fruits were moved to the sampling chamber for the experiment process.

3.2 Experiment

After 1 day, the durian samples will be put into the sampling box. The sampling chamber is a 50×50×50cm cube box which has rough white wall. The four upper corners of the box are arranged with the 50W halogen lights. The bulls are arranged at 45° angle from the vertical and pointing down toward the center of the chamber (Figure 7a, 7b).

The camera used in this research is a Samsung camera with 64MP resolution, and set to 1:1 zoom mode. The durian is placed at the center of the box's bottom, and the camera is positioned at the center of the box's top. During the sampling process, 80 durian fruits were photographed from various angles to enhance the training dataset for the model.

After the image sampling process, a set of images (Figure 8) was obtained for use as the training dataset for the deep learning model. The selected images meet the standard of sharpness, ensuring that the details on the durian's surface are clearly visible.

Based on the characteristics of fertile lobes in durian, as mentioned in the introduction, the collected image dataset will be annotated and masked using the LabelMe software tool (Figure 9). The generated masks will cover the most prominent and central part of the fertile lobe, including the largest, well-developed spikes that protrude higher than the rest of the fruit. These masks will then be added to the dataset along with the original images for training the model.

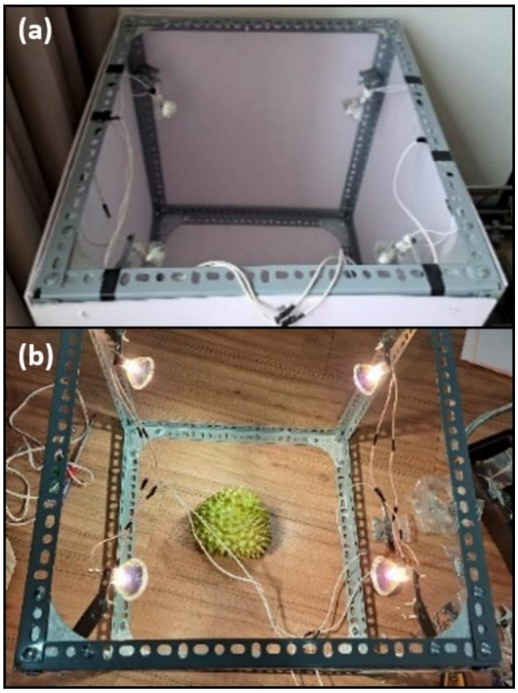


Figure 7. a) The sampling box, b) The sampling box (inside) during the sampling process.

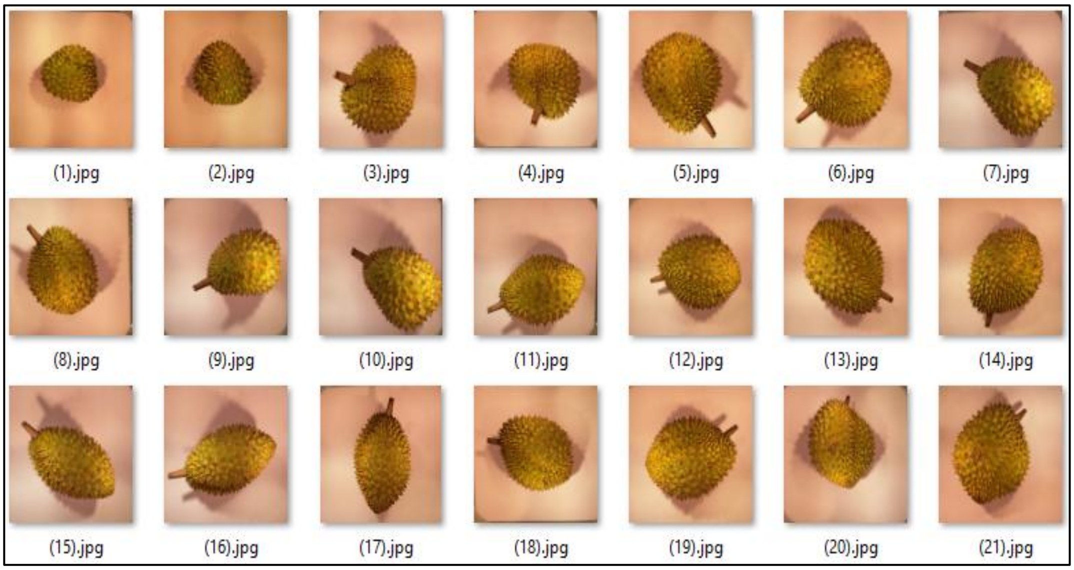
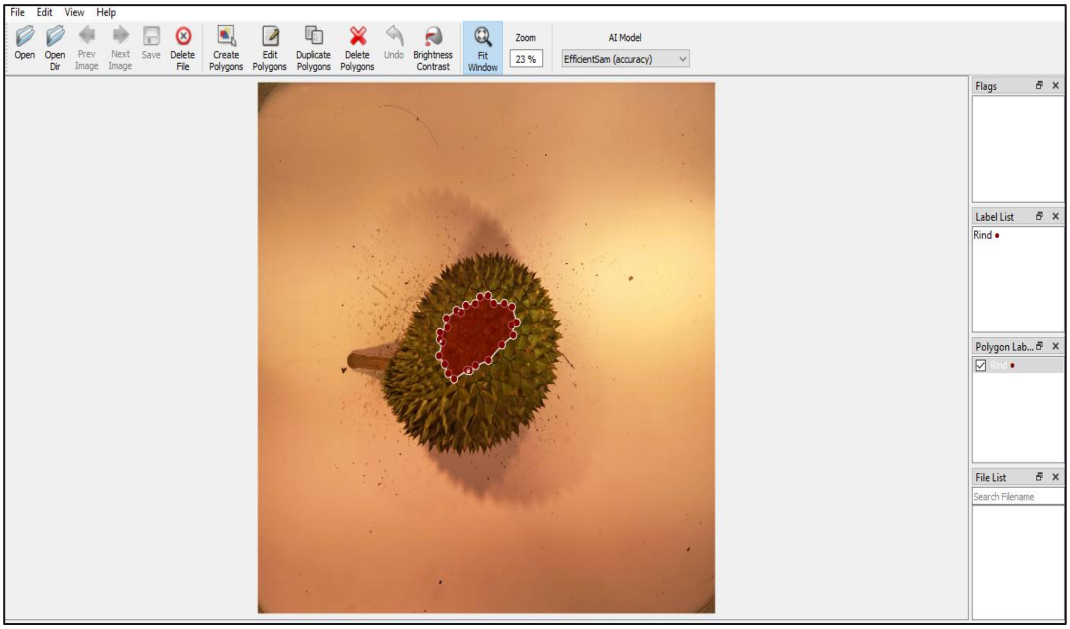


Figure 8. Raw durian dataset.



Zoom 23 %

AI Model EfficientSam (accuracy)

Label List: Rind ●

Polygon Lab...: ☑ Rind ●

Figure 9. Making mask with LabelMe software.

Then, the collected image dataset is divided into pairs of original images and corresponding masks before being fed into the neural network. This ensures alignment between the masks and the original images, preventing significant errors during the training process. After completing the above steps, the dataset is imported into the model for training. The original durian images are provided to the model as uint8 color images type with three channels, while the mask set is supplied as binary black-and-white images in Boolean format. This approach reduces computational resource consumption and enhances processing speed.

In the first stage of the model, the original data image set above will be taken through the preprocessing step. The contrast and color highlights of the durian spine and rind will clarify. In Figure 10, the spines of fertile lobe (yellow bounds) are large so this area reflects light better. The grooves at the base of the spines are clearly visible.

Next in the training process, three algorithms are Unet, Att-Unet and Att-ResUnet will be combined with the loss functions such as Binary Cross Entropy, Dice Loss, and Binary Focal Loss to determine which model yields the highest effectiveness. Additionally, techniques such as Dropout, Early Stopping (after 4 consecutive unchanged epochs) are employed to minimize errors and reduce overfitting in the model.

The training results are then evaluated and compared to identify the model that delivers the highest performance and accuracy. The final prediction will be finish by the center detection stage and get the coordinates of the FLC (Figure 11).

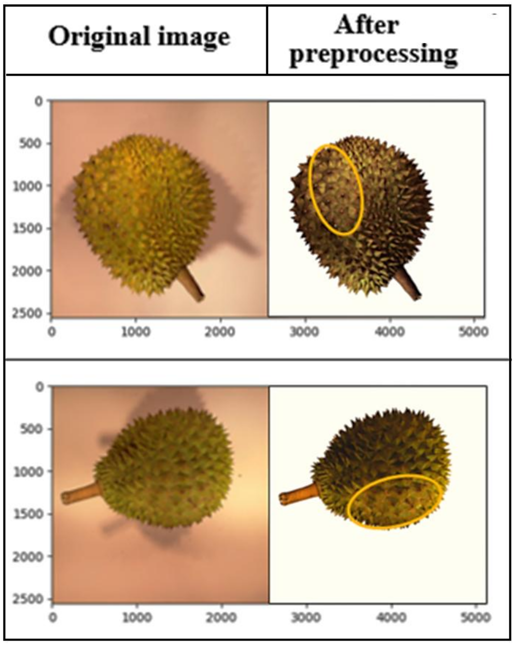


Figure 10. The contrast and color highlights of the durian spine and rind.

3.3 Evaluation metrics

The trained model will be evaluated and compared based on multiple criteria to ensure that the selected model achieves high accuracy and a short prediction time. The evaluation criteria include: recall (%), precision (%), accuracy (%), intersection over union (IoU%) and prediction time (seconds). The first three criteria are used to evaluate the model's performance on the validation set after training. Meanwhile, accuracy and IoU are specifically used to assess the model's effectiveness on the test set before and after applying TTA.

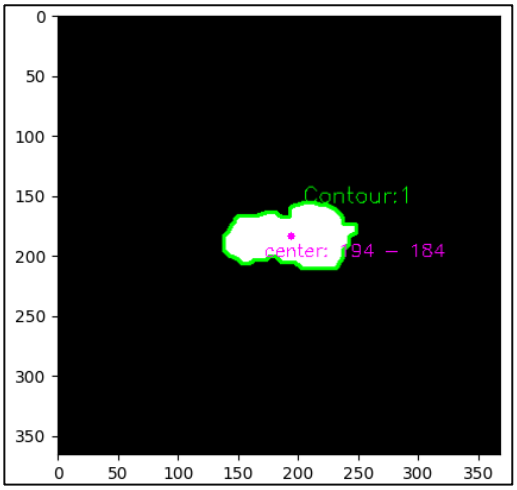


Figure 11. The coordinates of the FLC.

According to the aforementioned criteria, this study will delve deeper into the formulas and mechanisms of each parameter to gain a better understanding of the evaluation results. Figure 12 provides a more intuitive view of how the following formulas work. Following to Figure 12, True Positive (TP) represents the overlapping region of the fertile lobe between the actual mask and the predicted result. Similarly, True Negative (TN) corresponds to the surrounding background. False Positive (FP) and False Negative (FN) refer to the non-overlapping regions of the fertile lobe in the actual mask and the predicted result, respectively.

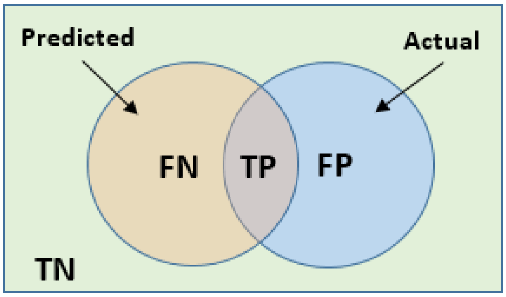


Figure 12. Distribution areas of actual and prediction.

From all of the area in Figure 12, the metric parameters are calculated by the equations. Firstly, recall is the ratio of the number of correctly classified pixels and the number of the actual target feature pixels. The formula of recall is:

$$\%Recall = \frac{TP}{TP+FN} \times 100 \quad (7)$$

Secondly, precision is the ratio of the number of correctly classified pixels to the number of mask pixels:

$$\%Precision = \frac{TP}{TP+FP} \times 100 \quad (8)$$

Thirdly, accuracy is a fundamental criterion in classification, providing a direct measure of how well a model performs its intended task. It represents the ratio of correctly predicted instances to the total number of instances in the dataset. It is the proportion of correctly classified instances, including both positive and negative cases. Mathematically, it is defined as:

$$\%Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (9)$$

Finally, IoU is a metric computed as the ratio of the overlapping area between the ground truth and the predicted segmentation to the combined area. It is mathematically defined as:

$$\%IoU = \frac{TP}{TP+FP+FN} \times 100 \quad (10)$$

4. Results and discussion

Based on the above criteria, the model will be evaluated for effectiveness. The model's performance on the validation set after training is evaluated based on the Table 1. It shows the percentages of recall, precision and accuracy of three algorithms, which are combined with three types of loss functions. From the data in Table 1, it is evident that the segmentation models are feasible for determining the location of fertile lobes, as most parameters exceed 80%. When monitoring the parameters between the loss function, it can be seen that the efficiency nearly increases in the order BCE, DL and BFL. However, BFL shows superior accuracy while BCE and DL do not seem to be too different from each other although DL mostly shows better performance than BCE (7 out of 9 cases).

From the results in Table 1, the fact that BFL is often 2% - 4% higher than the other two types can be explained by the significant variation in the shape and color of durian shells within the dataset, which leads to data imbalance. BFL demonstrates better adaptability in handling data imbalance by focusing more on hard-to-classify samples compared to BCE and DL. This has also been studied by Neyestanak, M. S. et al. (2022) (Neyestanak et al., 2022).

As for the algorithms, model Att-ResUnet achieves higher ratio of recall, precision and accuracy than the other two model, Unet and Att-Unet. Once again, the complexity of the durian shell is a great challenge for the segmentation model. Therefore, attention gates with focusing on important features contributes to improving the ability to recognize and set points for the model. This is also supported by the image preprocessing process which helps highlight the color points of the rind. Therefore, Residual Learning is responsible for helping to transfer information better through deep layers, minimizing the problem of disappearing gradients, thereby helping the model converge faster and more stably.

Table 1
Overall comparison the efficiency for validation set of different method

Classification algorithm	Loss Function	%Recall	%Precision	%Accuracy
Unet	Binary Cross Entropy	84.46	75.33	90.16
	Dice Loss	85.18	76.75	91.14
	Binary Focal Loss	91.77	80.24	92.57
Attention-Unet	Binary Cross Entropy	90.93	79.26	91.36
	Dice Loss	92.24	79.34	93.23
	Binary Focal Loss	93.05	82.80	94.31
Attention-Residual Unet	Binary Cross Entropy	92.56	80.67	94.25
	Dice Loss	94.06	83.43	94.80
	Binary Focal Loss	95.11	86.72	96.68

In Table 1, Attention-Residual Unet shows the most promising results in identifying fertile lobes based on the morphological characteristics of durian spines and peels, especially when combined with BFL loss function. This combination shows a rapid increase in efficiency during the first half of the training process, which gradually decreases but maintains an increasing trend in the second half of the process (Figure.13). It can be seen that there is a slight decrease in the period between epoch 40 and epoch 50, but this does not greatly affect the improvement of the model.

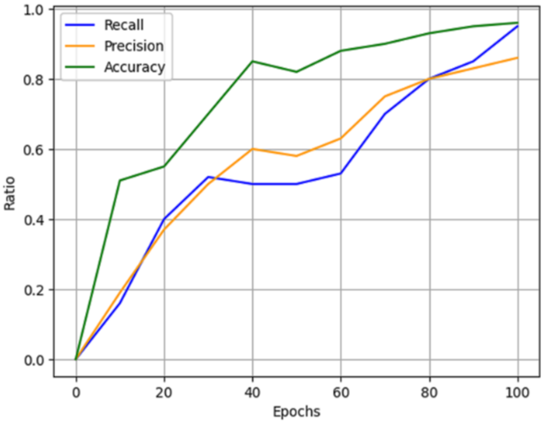


Figure 13. Training efficiency on the validation set of Att-ResUnet model.

After being evaluated on the validation set, each of the most effective models from the Unet, Att-Unet and Att-ResUnet architectures is continued to be tested on the test set. At this step, a test set accounting for 10% of the original dataset is used to test the accuracy of the prediction model. The metric used for testing will be %IoU and %accuracy, which can provide the objective evaluations. Table 2 shows the results of three models on the test set (one example is used for visualization). It can be easily seen that the predicted fertile lobe area has almost the same shape and position as the mask. This shows that the algorithm has effectively extracted the points of the durian peel. However, there are still noisy locations (red boxes) that the model did not recognize. %Accuracy and %IoU increase in the order Unet, Att-Unet and Att-ResUnet (78.26%, 81.34% and 82.89% for accuracy and 79.49%, 80.47% and 80.94% for IoU).

In the data from Table 2, noise artifacts appear at the edges of the FLC region due to reflection and diffusion on the rough surface, creating uneven bright spots. This can lead to confusion in the model's predictions. Model Attention-Residual Unet combined with binary focal loss method continues to outperform the rest by consistently outperforming them all. Although relatively good accuracy has been achieved for a complex classification task, noisy locations still need to be eliminated to avoid misidentifying the FLC location.

To achieve this, another technique that has been used to decrease noise levels is test time augmentation. The results after application on test set are presented again in Table. After the application of TTA, the neatness of the prediction compared to the true mask is significantly improved in Table 3. The noise area (yellow boxes) has disappeared so that the shape of the prediction matches the mask much better (Table 3). The two evaluation criteria accuracy and IoU improved by 3-5% compared to before combined with TTA. The %accuracy achieved the highest percentage with 86.85% and %IoU with 83.12%. From the above results, again in this study, the model with the combination of Att-ResUnet and test time augmentation algorithm achieved the best performance in determining the FL region achieved the best performance in determining the FL region. The four augmented predictions in the TTA algorithm eliminated unnecessary regions in the final result (Figure 14).

Table 2
Comparison of %IoU and %Accuracy on the test set before applying TTA with different methods

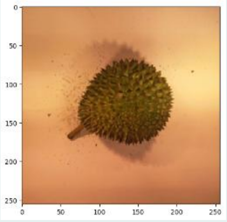
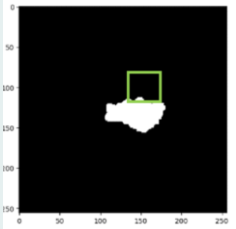
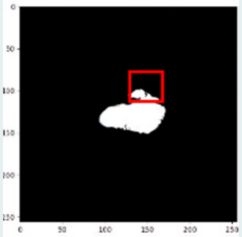
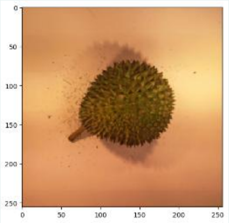
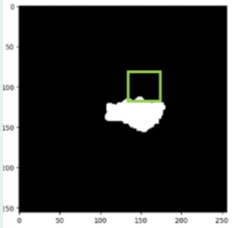
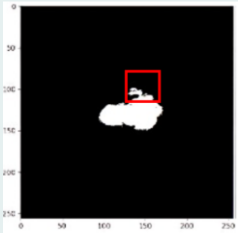
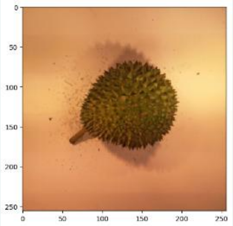
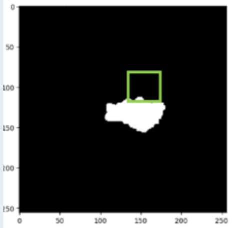
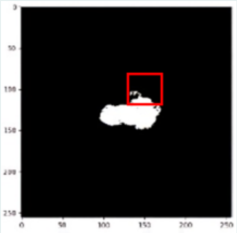
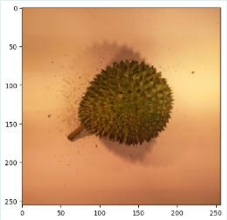
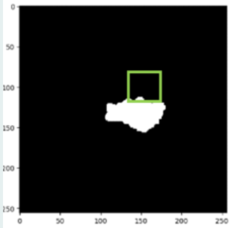
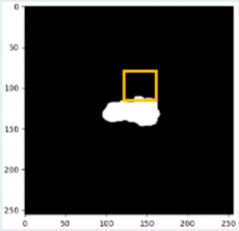
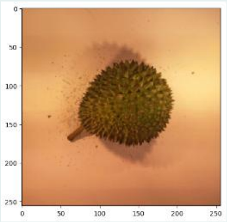
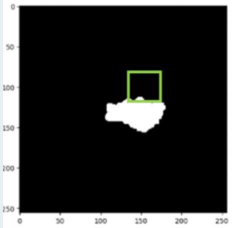
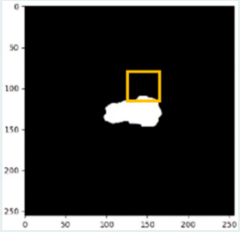
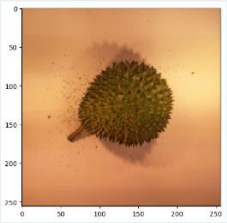
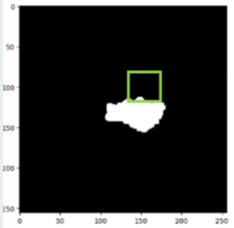
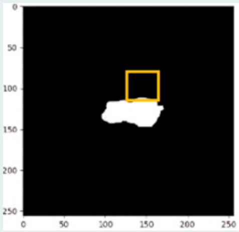
Classification Algorithm	Loss Function	Original Image (one example)	True Mask (one example)	Prediction (one example)	% IoU	% Accuracy
Unet	Binary Focal Loss				79.49	78.26
Attention-Unet	Binary Focal Loss				80.47	81.34
Attention-Residual Unet	Binary Focal Loss				80.94	82.89

Table 3
Comparison of %IoU and %Accuracy on the test set after applying TTA with different method

Classification Algorithm	Loss Function	Original Image (one example)	True Mask (one example)	Prediction (one example)	% IoU	% Accuracy
Unet	Binary Focal Loss				81.65	85.24
Attention-Unet	Binary Focal Loss				82.03	85.68
Attention-Residual Unet	Binary Focal Loss				83.12	86.85

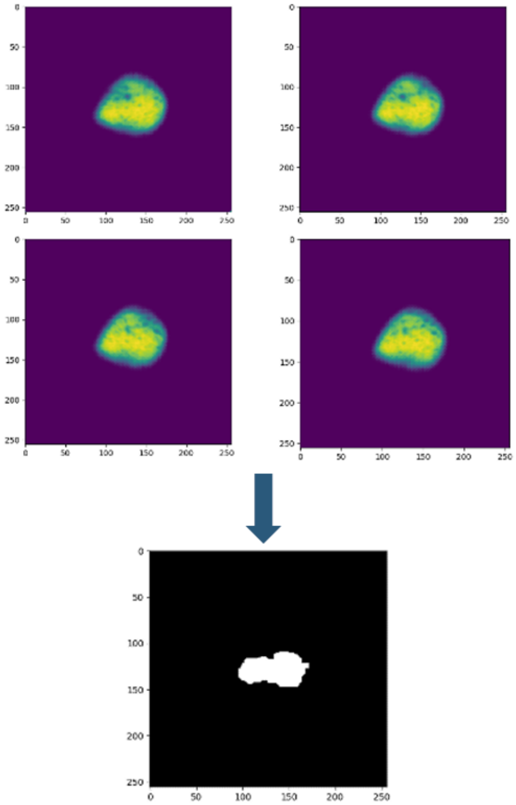


Figure 14. Result after TTA.

The results in Table 4, with 86.85% accuracy, demonstrate the high effectiveness of utilizing durian shell characteristics for classification or segmentation. This task is significantly more challenging compared to other fruits due to the irregularity and complexity of the shell and spikes. In Lim & Chuah's (2018) (Lim & Chuah, 2018) study, a CNN model was applied to images of durian surfaces, achieving high accuracy during training. However, when tested on the test set, accuracy dropped significantly to 82.5%. Similarly, in Hashim, et al.'s (2022) (Hashim et al., 2022) CNN-based durian classification model, accuracy reached 80%. Additionally, in the research of Mustaffa, et al. (2018), which employed Linear Discriminant Analysis (LDA) for durian recognition based on external features, the test dataset yielded significantly lower accuracy than the validation set, achieving only 72.38%.

Table 4
Comparison of different models in utilizing durian shell characteristics for classification or segmentation

Model architecture	%Accuracy
CNN (Lim & Chuah, 2018)	82.5
CNN (Hashim et al., 2022)	80
LDA (Mustaffa et al., 2018)	72.38
Att-ResUnet (this research)	86.85

Table 4 illustrates that the Att-ResUNet model in this study outperforms models *CNN* (Lim & Chuah, 2018), *CNN* (Hashim et al., 2022), and *LDA* (Mustaffa et al., 2018) in terms of accuracy.

5 Conclusions

From the results of the training and evaluation process in this study, it is evident that the combination of Attention-Residual Unet architecture and test time augmentation algorithm achieves the highest effectiveness in recognizing the characteristics of fertile lobes and determining their locations.

Additionally, when evaluating the model's performance on the test set using %IoU and %accuracy, the results are highly encouraging, reaching 83.12% and 86.85%.

There are several factors affecting the quality of the final results, such as the lighting angle in certain cases failing to create a clear distinction in the durian spines, surface damage on the shell, and so on. However, in the future, these limitations can be addressed by optimizing the illumination angle for the light source and applying more rigorous data selection before training.

These figures demonstrate the high feasibility of the proposed model for practical applications, aiming to improve the efficiency of FLC identification in durians. This research also contributes to the development of a durian ripeness classification system using the NIR method, previously studied by Ditcharoen et al. (2023). In addition, the model's ability to extract characteristics of durian peels can be a premise for improving durian variety classification models in the future.

Acknowledgments
We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

Authors' contribution

T. T. Luu: Methodology, Investigation, Writing – review & editing, Supervision, Project administration. N. Q. Cao: Methodology, Formal analysis, Investigation, Writing – original draft.

ORCID
T. T. Luu https://orcid.org/0000-0002-4042-6968
N. Q. Cao https://orcid.org/0009-0006-9482-1935

References

Ditcharoen, S., Sirisomboon, P., Saengprachatanarug, K., Phuphaphud, A., Rittiron, R., & Terdwongworakul, A. (2023). Improving the non-destructive maturity classification model for durian fruit using near-infrared spectroscopy. *Artificial Intelligence in Agriculture, 7,* 35–43. https://doi.org/10.1016/j.aiia.2023.02.002

Enshaei, N., Ahmad, S., & Naderkhani, F. (2020). Automated detection of textured-surface defects using UNet-based semantic segmentation network. *IEEE International Conference on Prognostics and Health Management (ICPHM).*

Detroit, MI, USA, pp. 1-5. https://doi.org/10.1109/ICPHM49022.2020.9187023

Glinpratum, S.-a. (2003). Thai Agricultural Commodity and Food Standard: Durian. *National Bureau of Agricultural Commodity and Food Standards* . TAS 3-2003.

Hashim, N. M., Bahri, M. H., Ghani, S. M., Sulistiyo, M. D., Kassim, K. A., & Zahri, N. A. (2022). An Introduction to A Smart Durian Musang King and Durian Kampung Classification. *IEEE Xplore*. https://doi.org/10.1109/CONIT55038.2022.9847909

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE* https://doi.org/10.48550/arXiv.1512.03385

Ibtehaz, N., & Rahman, M. S. (2019). MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Networks 121*, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025

Li, Y., Guo, R., Li, R., Ji, R., Wu, M., & Chen, D. (2025). An improved U-net and attention mechanism-based model for sugar beet and weed segmentation. *Front. Plant Sci.*, *15*, 1449514. https://doi.org/10.3389/fpls.2024.1449514

Likhitpichitchai, P., Khermkhan, J., & Kuhaswonvetch, S. (2023). The Competitiveness of Thai Durian Export to China's Market. *GMSARN International Journal*, *17*, 371-378.

Lim, M. G., & Chuah, J. H. (2018). Durian Types Recognition Using Deep Learning Techniques. *IEEE Control and System Graduate Research Colloquium (ICSGRC 2018)*. https://doi.org/10.1109/ICSGRC.2018.8657535

Mane, S., Bartakke, P., & Bastewad, T. (2023). Efficient Pomegranate Segmentation with UNet: A Comparative Analysis of Backbone Architectures and Knowledge Distillation. *ITM Web of Conferences*, *54*, 01001. https://doi.org/10.1051/itmconf/20235401001

Men, G., He, G., & Wang, G. (2021). Concatenated Residual Attention UNet for Semantic Segmentation of Urban Green Space. *MPDI Forests*, *12*. https://doi.org/10.3390/f12111441

Montazerolghaem, M., Sun, Y., Sasso, G., & Haworth, A. (2023). U-Net Architecture for Prostate Segmentation: The Impact of Loss Function on System Performance. *MDPI Bioengineering*, *10*, 412. https://doi.org/10.3390/bioengineering10040412

Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., & Horvath, P. (2020). Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports*, *10*, 5068. https://doi.org/10.1038/s41598-020-61808-3

Mustaffa, M. R., Yi, N. X., Abdullah, L. N., & Nasharuddin, N. A. (2018). Durian Recofnition Based on Multiple Features and Linear Discriminant Analysis. *Malaysian Journal of Computer Science. Information Retrieval And Knowledge Management Special Issue, 2018*. https://doi.org/10.22452/mjcs.sp2018no1.5

Neyestanak, M. S., Jahani, H., Khodarahmi, M., Zahiri, J., & Hosseini, M. (2022). A Quantitative Comparison between Focal Loss and Binary Cross-Entropy Loss in Brain Tumor Auto-Segmentation Using U-Net. *SSRN*. http://dx.doi.org/10.2139/ssrn.4142314

Ngoc, N., Dang, L., Ly, L., Thao, P., & Hung, N. (2024). Use of the Diagnosis and Recommendation Integrated System (DRIS) for Determining the Nutritional Balance of Durian Cultivated in the Vietnamese Mekong Delta. *Horticulturae*, *10*(6), 561. https://doi.org/10.3390/horticulturae10060561

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., & Heinrich, M. (2018). Attention U-Net: Learning Where to Look for the Pancreas. *Medical Imaging with Deep Learning*. https://doi.org/10.48550/arXiv.1804.03999

Pascua, O., & Cantila., M. (1992). Maturity indices of durian (*Durio zibethinus* Murray). *Philippine Journal of Crop Science*, *17*(3), 119-124.

Puttipipatkajorn, A., Terdwongworakul, A., Puttipipatkajorn, A., Kulmutiwat, S., & Sangwanangkul, P. (2023). Indirect prediction of dry matter in durian pulp with combined features using miniature NIR spectrophotometer. *IEEE Access*, *11*, 84810-84821. https://doi.org/10.1109/ACCESS.2023.3303020

Qurri, A. A., & Almekkawy, M. (2023). Improved UNet with Attention for Medical Image Segmentation. *Sensors*, *23*, 8589. https://doi.org/10.3390/s23208589

Rehman, A., Butt, M. A., & Zaman, M. (2023). Attention Res-UNet: Attention Residual UNet With Focal Tversky Loss for Skin Lesion Segmentation. *International Journal of Decision Support System Technology (IJDSST)*, *15*(1), 1-17. https://doi.org/10.4018/IJDSST.315756

Rehman, H. H., & Rehman, M. (2022). 2D UNET Segmentation of Mitochondria of Electron. *International Research Journal of Modernization in Engineering Technology and Science*. https://www.doi.org/10.56726/IRJMETS29962

Saifullah, S., & Drezewski, R. (2023). Modified Histogram Equalization for Improved CNN Medical Image Segmentation. *Procedia Computer Science*, *225*, 3021–3030. https://doi.org/10.1016/j.procs.2023.10.295

Siriphanich, J. (2011). Durian (*Durio zibethinus* Merr.). *Postharvest Biology and Technology of Tropical and Subtropical Fruits Cocona to Mango Woodhead Publishing Series in Food Science, Technology and Nutrition, 80-114, 115e-116e*. https://doi.org/10.1533/9780857092885.80

Sulaiman, A., VatsalaAnand, SheifaliGupta, A. R., & HaniAlshahrani. (2024). Attention based UNet model for breast cancer segmentation using BUSI dataset. *Scientific Reports*, *14*, 22422. https://doi.org/10.1038/s41598-024-72712-5

Weng, W., XinZhu, Jing, L., & MianxiongDong. (2023). Attention Mechanism Trained with Small Datasets for Biomedical Image Segmentation. *Electronics*, *12*(3), 682. https://doi.org/10.3390/electronics12030682

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749-753. https://doi.org/10.1109/LGRS.2018.2802944