Multimodal Deep Learning and IoT Sensor Fusion for Real-Time Beef Freshness Detection

Bambang Kurniawan[1,*], Refni Wahyuni[2], Yulanda[3], Yuda Irawan[4], Muhammad Habib Yuhandri[5]

[1,2,3,4]*Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[5]*Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia*

(Received: March 25, 2025; Revised: June 22, 2025; Accepted: October 01, 2025; Available online: October 22, 2025)

Abstract

Beef freshness quality is one of the important indicators in ensuring food safety and suitability. However, conventional methods such as manual visual inspection and laboratory testing cannot be widely applied in real-time and mass scale. To overcome these challenges, this study proposes a meat freshness detection system based on a multimodal approach that combines visual imagery and gas sensor data in a single IoT-based framework. This system is designed by utilizing the YOLOv11 architecture that has been optimized using the Adam optimizer. The dataset consisted of 540 original beef images, expanded into 1,296 images after augmentation. The model is trained on these augmented images and is able to achieve detection performance with a mAP@0.5 value of 99.4% and mAP@0.5:0.95 of 95.7%. As a further improvement, the cropped image features from the YOLOv11 model are processed through a combination of the ViT model and CNN to classify the level of meat freshness into three classes: Fresh, Medium, and Rotten with an accuracy of 99%. On the other hand, chemical data was obtained from the MQ136 and MQ137 gas sensors to detect H_2S and NH_3 levels which are indicators of meat spoilage. Data from visual and chemical data were then combined through a multimodal fusion method and classified using the Random Forest algorithm, producing a final prediction of Fit for Consumption, Need to Check, and Not Fit for Consumption. This multimodal model achieved a classification accuracy of 98% with a ROC-AUC score approaching 1.00 across all classes. While the proposed system achieved very high accuracy, further validation across diverse real-world environments is recommended to establish its generalizability.

Keywords: YOLOv11, Vision Transformer, CNN, IoT, Multimodal

1. Introduction

Beef consumption in Indonesia continues to increase, but domestic production is still inadequate, causing prices to soar and fraudulent practices such as mixing fresh meat with meat that is no longer fit for consumption are rampant [1]. The freshness of meat is difficult for the general public to determine because visual inspection is not always accurate, while laboratory methods are expensive and time-consuming [2]. According to the World Health Organization, foodborne illnesses affect around 600 million people annually, with contaminated meat being one of the leading sources of outbreaks. In addition, the Food and Agriculture Organization reports that economic losses due to food spoilage and waste are estimated at more than USD 400 billion per year, underlining the global urgency of improving meat freshness monitoring systems [3]. Therefore, an automated system based on artificial intelligence is needed that is able to detect the freshness of beef quickly and accurately to ensure meat quality.

Various studies have been developed to address the problem of meat quality detection. Among them is using a combination of CNN and YOLOv5 in a mobile application to detect meat quality based on visual images, but it is still limited because YOLOv5 only focuses on bounding box detection and is less than optimal in capturing texture details and subtle color variations that indicate early spoilage [4], [5]. Another study applied CNN to classify chicken freshness, but did not integrate odor data as an additional indicator [6]. Machine learning-based approaches such as K-NN, SVM, and Naïve Bayes have also been tested in meat freshness classification, but the results are still susceptible to dataset dependency and are less robust in handling complex features [7].

***Corresponding author: Bambang Kurniawan (ibenk.psht@gmail.com)**
DOI: https://doi.org/10.47738/jads.v5i2.XXX

The Internet of Things (IoT) integrates networked sensors, edge computing, and cloud services to enable continuous real-time data acquisition, analytics, and actuation enhancing monitoring, traceability, and decision-making across domains such as food safety, agriculture, healthcare, and environmental surveillance while remaining scalable and cost effective [8], [9], [10], [11], [12]. On the other hand, IoT-based research with gas sensors has been applied to detect NH_3 and H_2S levels, which are the main indicators of meat spoilage [13]. However, this approach is still limited because it has not optimized image processing. Meanwhile, other researchers use E-Nose which has high accuracy, but faces cost constraints, periodic calibration, and sensitivity that can be affected by environmental factors [14].

Various studies have developed deep learning and machine learning methods for meat freshness detection. Other researchers used MSI and CNN with 93% accuracy, but were dependent on lighting [9]. Other researchers applied ResNet-50 with 92.5% accuracy, but were only image-based [15]. The next researcher used AutoML with 84% accuracy but only relied on color and texture [16]. Then developed deep learning and OpenCV 75% but had difficulty distinguishing types of meat cuts [17]. Other researchers compared CNN, with Ayam6Net (92.9%), but did not consider odor [18]. Other researchers compared CNN, with Ayam6Net (92.9%), but did not take smell into account [6]. Researchers used Xception achieving 86.92% accuracy but were only limited to visual imagery [19].

The gas sensor and IoT approaches have also begun to be applied, including integrating CNN with gas sensors with an accuracy of 93% but have not been tested in complex environments [14]. Furthermore, applying K-NN with gas and color sensors with an accuracy of 93% but are vulnerable to external factors [20]. Other researchers use E-Nose based on metal oxide sensors, but are expensive and sensitive [21]. Other researchers developed fluorescence sensors and deep learning achieving 94.17% but difficult in complex implementation [22]. The researchers further applied gas and color sensor-based analysis achieving 93% accuracy [23].

Several studies comparing machine learning algorithms, among them researchers found K-NN superior with 95% accuracy, but dependent on large datasets [24]. Researchers further evaluated SVM, K-NN, and Naïve Bayes, with K-NN the best achieving 93% accuracy, but having difficulty handling small color changes [25]. Researchers further developed a mobile application based on ResNet-50, but requiring a larger dataset [26]. Previous research developed CNN and YOLOv9-based models with 95% accuracy but still susceptible to lighting [27].

Beyond deep learning on visual data, hyperspectral imaging has emerged as a powerful technique for food quality inspection. As highlighted by [28], hyperspectral imaging combines spectroscopic chemical information with high-resolution imaging, offering detailed insights into food safety, adulteration detection, and traceability. However, its widespread use is hindered by the high cost, bulky equipment, and difficulty in interpreting large, complex datasets. Another promising modality is the electronic nose, which mimics human olfaction by capturing gas fingerprint patterns. Other researcher reviewed recent advancements in this field, showing how electronic noses have evolved from bulky, expensive devices into portable and cost-effective systems that have been applied in food analysis, environmental monitoring, and even medical diagnostics [29]. While both hyperspectral imaging and electronic noses demonstrate strong potential, their practical deployment remains challenged by cost, calibration, and environmental sensitivity underscoring the practicality of our proposed multimodal approach integrating YOLOv11, ViT-CNN, and IoT gas sensors.

Although previous meat freshness detection models achieved high accuracy, they remain limited by unstable performance under varying environmental conditions, restricted dataset size and diversity, and poor generalization to unseen data. Image-based models cannot capture chemical indicators such as odor, while gas sensor systems are prone to disturbances from humidity, temperature, and environmental cross-contamination. To address these issues, this study develops a multimodal hybrid deep learning framework that integrates visual and odor-based analysis. The IoT device is designed as a closed box equipped with controlled lighting, built-in cameras, and MQ136–MQ137 sensors for detecting H_2S and NH_3 gases key indicators of meat spoilage. In the visual stream, YOLOv11 is applied for real-time meat detection, Vision Transformer (ViT) for spatial feature enhancement, and CNN for initial classification. Both modalities are fused through a multimodal feature table combining visual predictions and gas concentration values, which are retrained using Random Forest as the final classifier.

The proposed hybrid framework is expected to overcome the limitations of previous single-modality approaches by combining complementary strengths: visual analysis for surface and texture changes, and gas sensing for internal

 ISSN 2723-6471

biochemical spoilage processes. The integration of YOLOv11 with ViT-CNN enables more accurate feature extraction under diverse lighting conditions, while IoT-based MQ136 and MQ137 sensors provide real-time chemical indicators. By fusing these data streams into a multimodal learning pipeline and retraining with Random Forest, the system can achieve robust performance and higher generalization ability, even in non-ideal environments.

In addition, the design of an IoT-based closed box ensures standardized data acquisition with controlled lighting and environmental stability, thereby reducing external noise that often hampers sensor reliability. This study not only contributes to advancing research in multimodal deep learning for food quality detection but also offers a practical solution for the food industry and regulatory agencies in Indonesia. The resulting system has the potential to strengthen consumer protection, reduce economic losses from meat spoilage, and support food safety policies at both national and global levels.

2. Research Methodology

To explain the model design process in its entirety, this section describes the main stages in the research from data processing to model performance evaluation. The model development flow can be seen in figure 1:

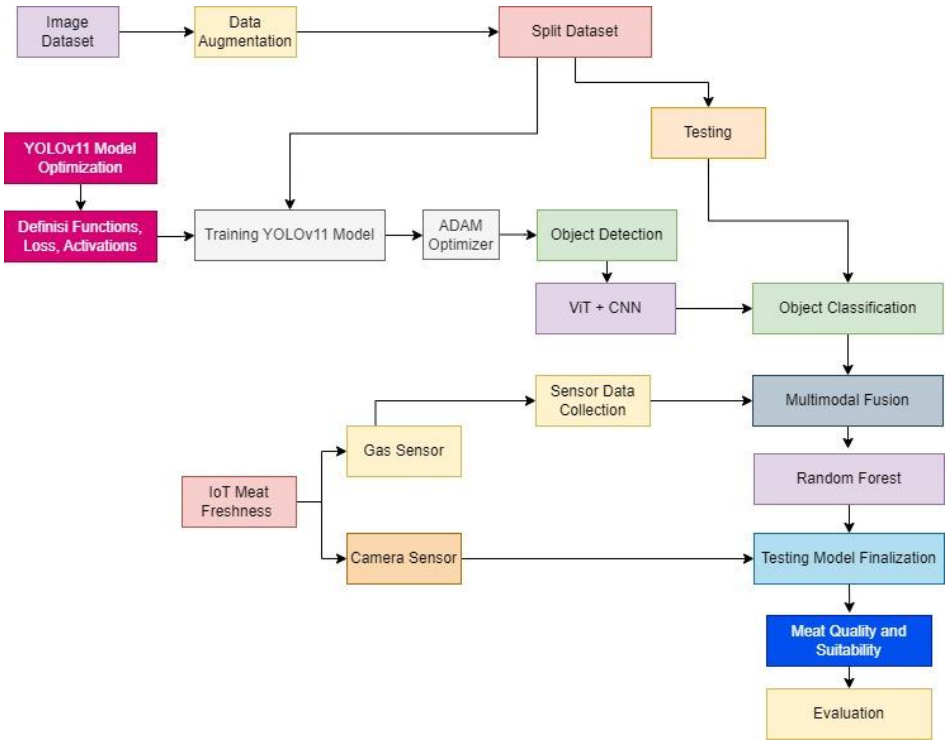

Figure 1. Flow of Model Development

2.1. Dataset Collection and Preprocessing

The image dataset used in this study consists of 540 images of beef with varying levels of freshness, which were collected as initial data before model training. The following is a view of the dataset in the form of images that have been entered into the Roboflow platform for efficient annotation and dataset management processes. This dataset visualization is presented in figure 2:

ISSN 2723-6471

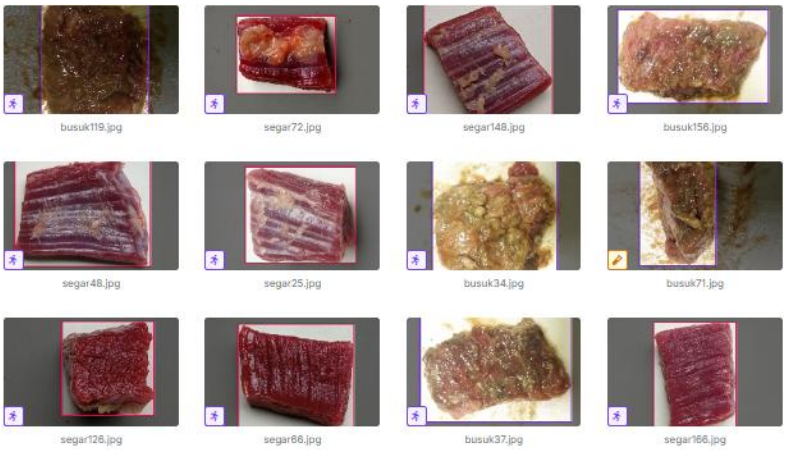

Figure 2. Beef Image Dataset

To ensure the consistency of image dimensions and orientation during training, a preprocessing stage is performed which includes two main processes. First, auto-orient is applied to automatically adjust the image orientation to be uniform horizontally or vertically. Second, all images are resized to 640×640 pixels using the resize-stretch method, so that all images have a uniform resolution that matches the YOLOv11 input architecture.

2.2. Data Augmentation

To increase the diversity of the training data and prevent overfitting, a data augmentation process was performed on the entire image dataset using automated features from Roboflow [30], [31]. Each original image was generated into three new variations through several transformations. The augmentation techniques applied include horizontal and vertical flip, crop with a maximum zoom limit of up to 20%, and random rotation between -14° to +14°. In addition, the image also undergoes horizontal and vertical shear up to ±10°, blur up to 2.6 pixels, and noise addition up to 0.69% of the total pixels.

2.3. Split Dataset

After the augmentation process is complete, the dataset is divided into three subsets for model training and evaluation purposes. A total of 88% of the data is used as training data, 8% as validation data, and 4% as test data. This proportion is designed to ensure that the model gets enough data to learn, while still being able to be evaluated objectively using data that has never been seen before. The division is done automatically through the Roboflow platform.

2.4. YOLOv11 Training Model

The detection model training is performed using the YOLOv11 architecture that has undergone an improvised stage to adapt to the characteristics of meat images. YOLOv11 optimization includes adjusting the model structure, redefining the activation function, loss function, and hyperparameter settings to improve the ability to detect subtle objects such as meat texture that changes in freshness. At this stage, YOLOv11 acts as the main detector to identify the meat object area in the image and automatically performs cropping process on the detected bounding box. The cropping results are then used as further input to the freshness classification process. After the definition and optimization process is complete, the model is trained with an image resolution of 640×640 pixels, an epoch of 50, and a batch size of 8, which is adjusted to the capabilities of the hardware to remain efficient but produce optimal performance. The Adam optimizer was chosen for its ability to maintain gradient stability on complex visual data [32], [33], and the cache=True feature was used to speed up data processing during training. Compared with YOLOv10, YOLOv11 incorporates a redesigned backbone with enhanced depth-wise convolutions, an optimized PAN-FPN neck for improved multi-scale feature fusion, and a refined loss function (Distribution Focal Loss) that increases sensitivity to subtle visual cues. These improvements are particularly beneficial for beef freshness detection, where minor changes in meat texture and color indicate spoilage.

2.5. ViT + CNN Model Training

After the meat object detection process using the YOLOv11 model, the system automatically crops the bounding box area of the detection results to produce a new dataset that is more focused on the meat part. The cropped dataset is then used as training data for a freshness classification model consisting of a combination of ViT and CNN. ViT is used to extract and enhance global spatial features of the meat image, such as discoloration and subtle textures that mark freshness degradation, while CNN is tasked with the final classification into categories such as Fresh, Medium, or Rotten. This approach combines the advantages of ViT in understanding the overall image structure with the strengths of CNN in detecting local patterns [34], resulting in a classification model that is more accurate and sensitive to visual changes. To minimize error propagation from object detection to classification, YOLOv11 outputs with confidence values below 0.5 were excluded from the cropping process. In addition, a subset of bounding boxes was manually verified to ensure consistency and correctness. Although these steps reduced the risk of mis-detections entering the classification pipeline, this remains a limitation of the approach. Future work will explore multi-stage verification strategies and ensemble detection models to further improve robustness.

2.6. IoT Sensor Data Acquisition

The IoT system in this study is designed to obtain odor data that is a chemical indicator of the level of meat freshness in real-time. The following is a series of IoT devices for detecting meat quality. The overall IoT system design can be seen in figure 3.

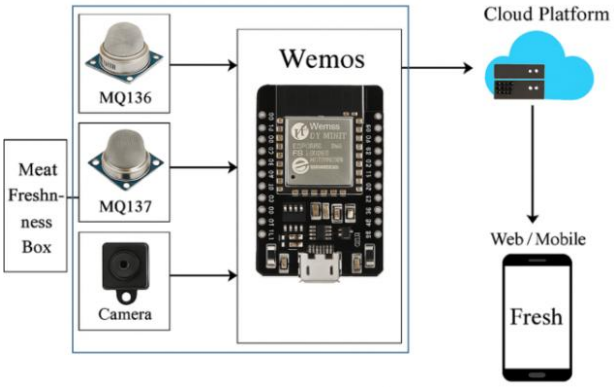

Figure 3. IoT System Design

The device consists of MQ136 and MQ137 gas sensors that detect the concentration of hydrogen sulfide (H_2S) and ammonia (NH_3) gas, respectively, which are two common volatile compounds produced during the meat spoilage process. These sensors are connected to an ESP8266-based Wemos D1 Mini microcontroller, which is in charge of reading the sensor values and sending them wirelessly via a Wi-Fi connection. The data obtained will periodically be sent and logged automatically to Google Sheets as a cloud-based recording platform, allowing real-time monitoring of sensor values and integrating with the built classification system.

2.7. Multimodal Integration and Random Forest Classification

After completing visual classification with ViT+CNN and gas sensor acquisition, both modalities were integrated into a multimodal feature table containing visual predictions (fresh, medium, rotten), H_2S concentrations (MQ136), and NH_3 concentrations (MQ137). These features were then retrained using Random Forest as the final classifier to determine meat fitness for consumption. This multimodal fusion ensures that the system leverages both visual and chemical cues, while Random Forest was selected for its robustness in handling mixed data and delivering stable, interpretable predictions. The final classification results are divided into three categories: Suitable for Consumption, Needs to be Checked, and Not Suitable for Consumption, which are the basis for the system's decision-making on the freshness of the meat in real time and accurately. Random Forest was selected as the fusion model because of its robustness in handling mixed feature types categorical predictions from the visual model and continuous gas sensor values while providing stable performance on moderately sized datasets. Compared to neural network–based fusion and gradient boosting, Random Forest required less hyperparameter tuning, reduced the risk of overfitting, and maintained interpretability of feature contributions. While this study demonstrated strong results with Random Forest,

future research will investigate gradient boosting and neural network based fusion strategies to further enhance multimodal integration.

2.8. Model Evaluation

The performance evaluation of the model in this study is done through various classification metrics that reflect the accuracy and reliability of the system in detecting meat freshness. Some of the main evaluation methods used include Confusion Matrix, Classification Report, and ROC Curve n35 [35]. Confusion Matrix is used to describe the distribution of model predictions against the correct class, visualizing the number of correct and incorrect predictions in each category [36], [37], [38]. Classification Report provides detailed information about the Precision, Recall, F1-Score, and Accuracy values for each class [39], [40], [41], [42]. While the ROC Curve is used to measure the model's ability to distinguish between classes, by examining the AUC value which is closer to 1 indicates better classification performance [43]. This evaluation was carried out for each model: CNN, ViT+CNN, Random Forest Sensor, and the final Multimodal model.

3. Methodology

3.1. Dataset Preprocessing and Augmentation Results

The preprocessing and augmentation steps enhanced the quality and diversity of beef images for YOLOv11 training. The dataset, initially 540 images, was expanded to 1,296 through Roboflow augmentation. Preprocessing included auto-orientation and resizing to 640×640 pixels to fit the model input. Augmentation techniques such as rotation, flip, shear, blur, crop, and noise addition improved robustness against variations in position, lighting, and shooting angle. The results of these preprocessing and augmentation steps are illustrated in figure 4.

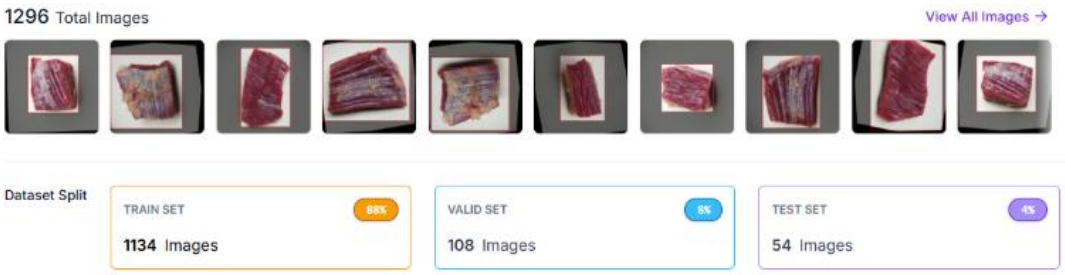

Figure 4. Dataset Preprocessing and Augmentation Results

The final result of the augmentation process produced 1134 images for training data (88%), 108 images for validation (8%), and 54 images for testing (4%). The various augmentation transformations applied allowed the model to learn from more visual variations, including subtle texture shifts, lighting distortions, and angular rotations. With proportional data distribution and sufficient augmentation variety, the YOLOv11 and ViT+CNN models are expected to be able to generalize better to new data, as well as reduce the risk of overfitting during training.

3.2. Performance of YOLO Model Training Results

The performance evaluation of the YOLOv11 model optimized using Adam was carried out after the training process was completed using the augmentation and preprocessing datasets. The model was trained to detect meat objects with three main classes, namely Fresh, Medium, and Rotten. The evaluation metrics used include Precision (P), Recall (R), mean Average Precision (mAP) at a threshold of 0.5 (mAP50) and mAP50–95. The evaluation results are shown in the following table 1.

Table 1. YOLOv11 Model Performance Evaluation Results

Class	Number of Images	Number of Objects	Precision (P)	Recall (R)	mAP50	mAP50–95
Rotten	35	35	0.997	1.000	0.995	0.925
Medium	39	39	0.995	1.000	0.995	0.964
Fresh	34	34	0.999	1.000	0.995	0.982

All	108	108	0.997	1.000	0.995	0.957

The YOLOv11 model showed very high performance in detecting meat objects in all three classes with near-perfect precision values (≥ 0.995) and 100% recall in all classes. The highest mAP50-95 value was achieved in the Fresh class at 0.982, indicating the model's ability to distinguish subtle visual characteristics. Overall, the mAP50 value of 0.995 and mAP50-95 value of 0.957 indicate that the YOLOv11 model used has been successfully trained very well and is able to perform precise object detection in the context of meat freshness classification.

3.3. Evaluation of the YOLO Model

Further evaluation of the YOLOv11 model is not only done through accuracy and mAP metrics, but also by analyzing the spatial distribution of bounding boxes generated during the labeling and detection process. This analysis is important to ensure that the model is not only numerically accurate, but also stable in recognizing and labeling objects at various positions and sizes in the image. The visualization of this analysis is presented in figure 5.

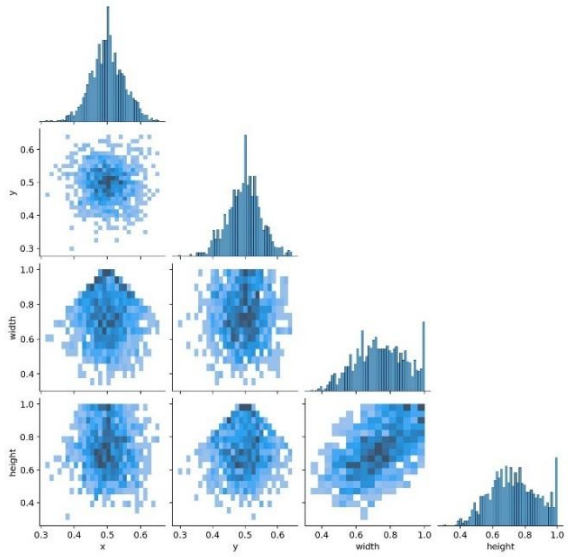

Figure 5. Spatial Distribution of Bounding Boxes on YOLOv11 Training Dataset

Figure 5 illustrates the heatmap pairplot of bounding box coordinates (x, y) and dimensions (width, height). The distribution of object positions appears symmetrical and centered, confirming consistent image capture and labeling. A strong correlation between width and height reflects the uniform shape ratio of meat samples, while the absence of extreme outliers indicates high annotation quality and evenly distributed data. These findings reinforce that the YOLOv11 model benefits from stable, representative training data in addition to strong predictive performance. The following is a visualization of the distribution of labels, bounding boxes, and object positions, as shown in figure 6.

Figure 6 shows that the label distribution of the three classes (Rotten, Medium, Fresh) is relatively balanced, with the number of instances close to 350-400 per class. Visualization of the bounding box accumulation shows that the majority of the objects are in the center of the image, indicating consistent camera framing quality during dataset acquisition. Heatmaps of the x-y position and width-height bounding box dimensions also show that the model was trained with realistic variations in position and size, while remaining centered and containing no extreme outliers.

ISSN 2723-6471

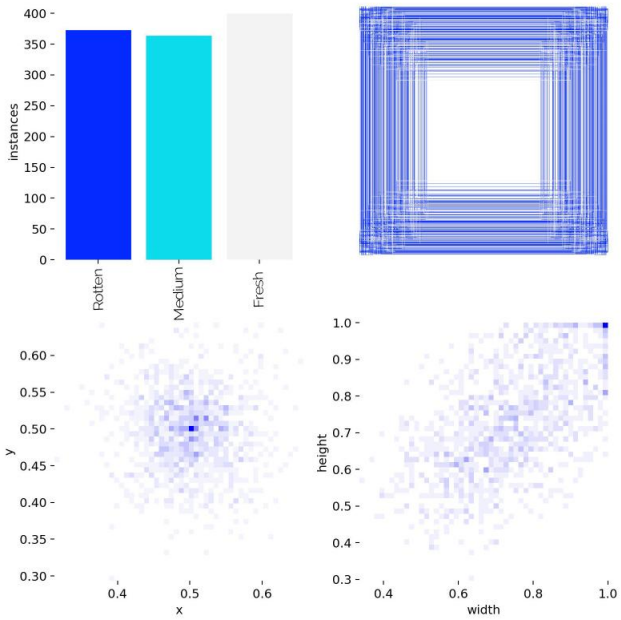

Figure 6. Visualization of Label Distribution, Bounding Box, and Object Position

The performance evaluation of the YOLOv11 model is also carried out by analyzing the loss trend and evaluation metrics during the training process until the 50th epoch. This graph includes the bounding box function loss, classification, distance distribution, as well as precision, recall, and mAP. The relationship between precision and recall is further depicted in figure 7.

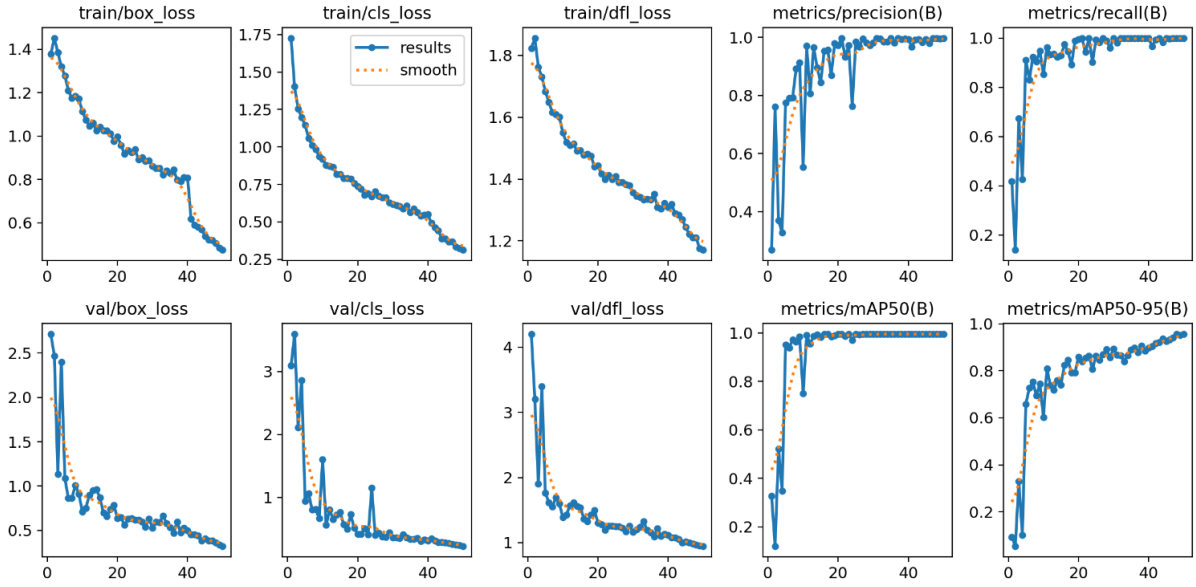

Figure 7. F1 Curve against Confidence Score for Each Class

Figure 7 shows that the loss values for box, classification (cls), and distance-from-location (dfl) on both training and validation data experience a steady decrease as the number of epochs increases, indicating that the model learning process is effective without any indication of overfitting. On the other hand, the precision and recall values continue to increase and reach values close to 1.0 at the end of training, indicating the model's ability to detect objects accurately and comprehensively. In addition, the mAP50 and mAP50–95 metrics have increased significantly since the beginning of training, with a flat curve after the 30th epoch, indicating that the model has reached optimal convergence. These results confirm that the training parameters used (epoch, batch size, Adam optimizer) have produced very good and

ISSN 2723-6471

stable model performance on the meat freshness detection task. As a final validation, a confusion matrix visualization is performed to evaluate the accuracy of the model's predictions for each class. This matrix displays a comparison between the actual labels and the model's predicted labels, as shown in figure 8.

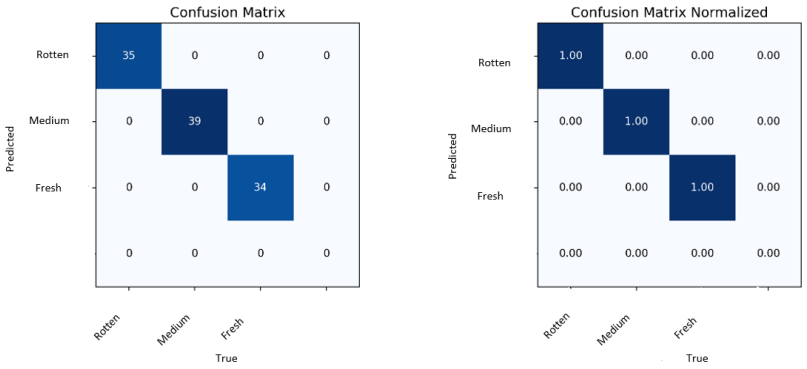

Figure 8. Confusion Matrix and Normalized Confusion Matrix

Figure 8 shows the confusion matrix results and its normalized version, where the YOLOv11 model successfully classified all samples in the Rotten, Medium, and Fresh classes 100% correctly without any misclassification. The numbers 35, 39, and 34 on the main diagonal reflect the number of correct predictions of each class in total, and this result is reinforced by the normalization which shows a score of 1.00 across all classes.

3.4. ViT + CNN Classification Results

Before being trained using the ViT + CNN model, the meat image dataset was first automatically cropped using the trained YOLOv11 model (best.pt). This process produces a set of cropped images that focus more on the meat object area without the background. To measure the performance of the visual classification model, an evaluation was conducted on the prediction results of the CNN and ViT models on the YOLO cropped image. Confusion matrix was used to identify the level of classification accuracy of each model in three main classes: Rotten, Medium, and Fresh. The comparison of these confusion matrices is presented in figure 9.

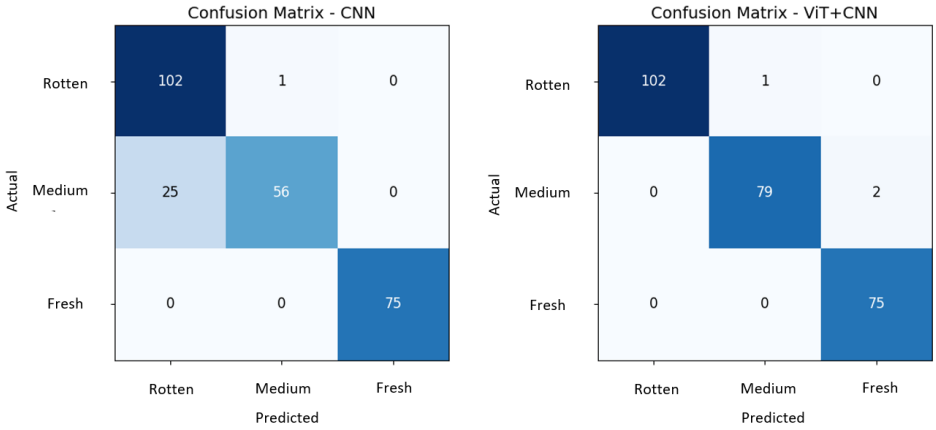

Figure 9. Comparison of Confusion Matrix Model CNN and ViT in Classification

Figure 9 shows that the CNN model produces several classification errors, especially in the medium class which tends to be predicted as Rotten (25 samples) and Fresh (8 samples). Meanwhile, the ViT+CNN model shows a much more stable classification performance with results that are close to perfect; all samples in the Rotten and Fresh classes are correctly classified, and only two samples in the Medium class are misclassified as Fresh. These results confirm that ViT+CNN excels in extracting visual features globally, producing more accurate and consistent classifications than conventional CNN in the context of meat freshness classification. As a complement to the confusion matrix evaluation, a classification report is used to compare classification metrics between models. The metrics displayed include precision, recall, f1-score, and total accuracy for each class. The results of this comparison are presented in figure 10.

 ISSN 2723-6471

Classification Report CNN

	precision	recall	f1-score	support
accuracy			0.87	
Rotten	0.80	0.99	0.89	103
Medium	0.86	0.69	0.77	81
Fresh	1.00	0.89	0.94	75
accuracy			0.87	259
macro avg	0.89	0.86	0.87	259
weighted avg	0.88	0.87	0.87	259

Classification Report ViT+CNN

	precision	recall	f1-score	support
accuracy			0.99	
Rotten	1.00	0.99	1.00	103
Medium	0.99	0.98	0.98	81
Fresh	0.97	1.00	0.99	75
accuracy			0.99	259
macro avg	0.99	0.99	0.99	259
weighted avg	0.99	0.99	0.99	259

Figure 10. Comparison of Classification Report Model CNN and ViT+CNN

Figure 10 shows that the CNN model has an accuracy of 0.87, with a notable drop in f1-score in the Medium class (0.77), indicating the model's weakness in distinguishing the medium class which has ambiguous visual features. In contrast, the ViT+CNN model achieved an accuracy of 0.99, with a high and balanced f1-score across all classes (≥ 0.98), including the Medium class which was previously a weak point of CNN. The superior performance of ViT+CNN shows that combining global (ViT) and local (CNN) feature extraction provides much more stable, accurate and even classification results across classes. To observe the stability of the training process, a comparison of the training loss between the pure CNN model and the combined ViT + CNN model was carried out. The following figure 11 shows the decrease in loss against the number of epochs.

Figure 11. Comparison of Training Loss between CNN and ViT + CNN Models

Figure 11 shows that the ViT+CNN model experiences a faster and more stable loss decrease than the CNN model alone. Since the beginning of training, the loss of the combined model drops drastically and remains low with minimal fluctuations, while the CNN model experiences a slower decrease and shows instability at several epoch points, including a significant spike at epoch 14. This indicates that the integration of ViT helps to absorb spatial information more efficiently, thereby accelerating convergence and reducing the risk of overfitting. This training loss performance strengthens the previous evaluation results that ViT + CNN is superior in terms of training accuracy and stability. To ensure the effectiveness of the model in a real context, direct testing was conducted on meat images from the three classes using the ViT+CNN model. The outcomes of these tests are presented in figure 12.

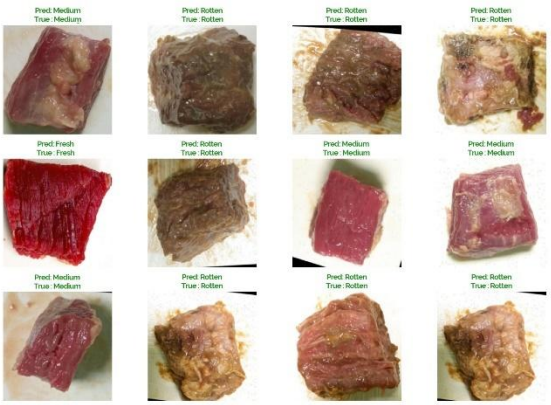

Figure 12. Visualization of ViT + CNN Model Prediction Results on Test Samples

ISSN 2723-6471

Figure 12 displays the prediction results of the ViT + CNN model showing that all test images were successfully classified according to their original labels. The model can distinguish the visual characteristics of Rotten, Medium, and Fresh meat consistently without any classification errors. This shows the generalization ability of the model to new data that has never been trained before. This performance reinforces that ViT+CNN not only excels in quantitative metrics such as accuracy and f1-score, but is also highly reliable when applied in real test conditions.

3.5. IoT Based Sensor Data Acquisition Performance

To support multimodal meat freshness detection, a prototype IoT device has been developed and tested directly. This device is designed using MQ136 and MQ137 gas sensors installed in a closed box with controlled lighting. The sample meat is placed in the box, and the results of the detection of H_2S and ammonia levels are displayed in real-time on the LCD screen. The following figure 13 is a display of the meat quality detection box that has been designed:

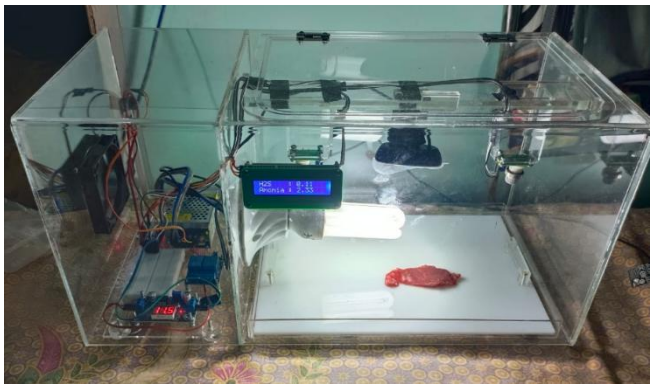

Figure 13. Prototype of IoT Device for Meat Quality Detection

Figure 13 shows the physical appearance of the IoT-based detection device that has successfully acquired and displayed H_2S and NH_3 gas levels from meat pieces in the test box. The MQ137 sensor detects ammonia, while the MQ136 reads H_2S levels. The distribution of gas concentration values was further analyzed per freshness class. Samples categorized as Fit for Consumption showed consistently low levels (MQ136 ≈ 0.97 ppm; MQ137 ≈ 0.87 ppm), while Need to Check cases exhibited intermediate and more variable ranges (MQ136 ≈ 1.19 ± 0.21 ppm; MQ137 ≈ 1.86 ± 0.87 ppm). In contrast, Not Fit for Consumption samples recorded distinctly higher concentrations (MQ136 ≈ 1.32 ppm; MQ137 ≈ 2.70 ppm). These distributions provide clearer separation among freshness categories and support the decision boundaries learned by the Random Forest classifier. The raw signals from MQ136 and MQ137 were preprocessed using a moving average filter (window size = 5) for noise reduction, baseline correction in clean-air conditions to address drift, and normalization (0–1) for comparability. The IoT detection box was equipped with internal LED lighting, ensuring consistent and uniform illumination across all samples. This design reduced variability due to external lighting conditions and supported stable data acquisition during testing.

3.6. Multimodal Visual Integration and IoT Sensors

The multimodal dataset in this study was developed through the integration of meat image classification results using the ViT+CNN model and readings of ammonia (NH_3) and hydrogen sulfide (H_2S) gas levels obtained from the MQ137 and MQ136 sensors. The final status labeling process was carried out by considering gas threshold values based on scientific references and expert validation in the fields of food technology and environmental health. There are a total of 2100 entries in the dataset, each consisting of three input features (visual image class, MQ136, and MQ137 levels) and one output label in the form of the final status: Appropriate, Not Appropriate, and Need to Check. The structure of this dataset is summarized in table 2.

Table 2. Multimodal Dataset for Random Forest Classification

Visual	MQ136	MQ137	Status
Medium	1.11	1.50	Need to Check
Fresh	0.99	0.88	Appropriate
Fresh	0.99	0.86	Appropriate

 ISSN 2723-6471

Rotten	1.32	2.70	Not Appropriate
Medium	1.13	1.12	Need to Check
Medium	1.17	1.63	Need to Check
Fresh	12.15	24.08	Need to Check
Rotten	9.26	9.133	Need to Check
Fresh	16.27	33.02	Need to Check
Fresh	1.00	0.88	Appropriate
Medium	1.05	0.73	Need to Check
Fresh	12.15	22.87	Need to Check
Fresh	0.99	0.87	Appropriate

Table 2 shows the combined data format between visual classification results and IoT sensors used to train the Random Forest model. The final status is classified into three categories: Appropriate (fresh visuals and low gas levels), Not Appropriate (rotten visuals and high gas levels), and Need to Check for ambiguous or borderline cases. This approach improves accuracy while adding an element of caution to decision making, in line with food safety principles.

To evaluate the classification performance on multimodal systems, the Random Forest model was trained using a dataset that combines visual and gas sensor data. The evaluation was performed on three final status categories: Appropriate, Not Appropriate, and Need to Check. The following figure 14 is the confusion matrix and classification report from Random Forest:

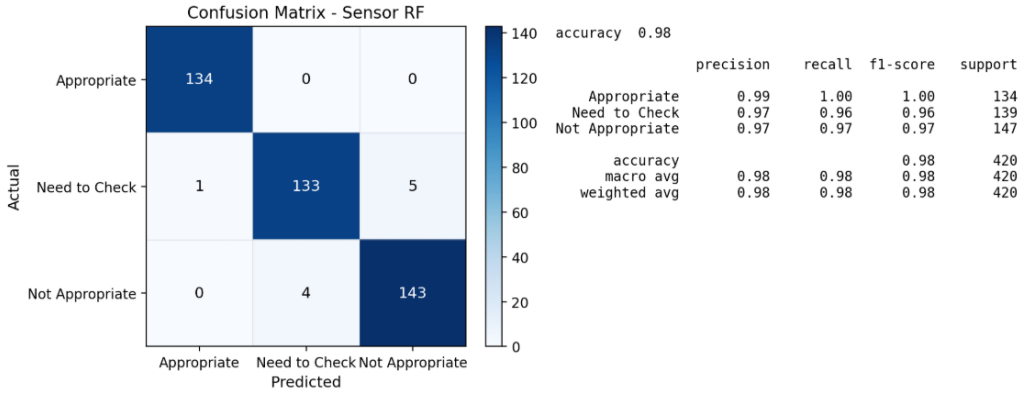

Figure 14. Confusion Matrix and Classification Report Model Random Forest

Figure 14 shows that the Random Forest model is able to classify the data very well, with an overall accuracy of 98%. The Appropriate class is perfectly recognized (recall and f1-score = 1.00), while the Need to Check and Not Appropriate classes also show high performance with f1-scores of 0.96 and 0.97, respectively. The number of misclassifications is very minimal, indicating that the combination of visual and sensor features provides complementary information. This confirms the reliability of the multimodal approach in providing accurate and careful predictions, especially for threshold categories such as Need to Check.

As part of the system implementation, a Graphical User Interface (GUI) application was designed using the Streamlit framework. This application allows users to detect meat freshness directly and interactively through a camera or image upload that can be accessed via mobile devices or personal computers. The interface and workflow of this application are presented in figure 15.

ISSN 2723-6471

Figure 15. Implementation of Streamlit-Based Meat Freshness Detection Model

Figure 15 shows the application interface built using Streamlit, where users can select the input source, either from the camera or image upload. The camera used comes from the Meat Quality Detector IoT Box device, which is integrated with a gas sensor. After the image is processed, the application displays the visual detection results, MQ136 and MQ137 gas level readings, and the final status such as Consumable. This display is designed to be easily accessible and used in the field, allowing the meat quality assessment process to be carried out in real-time, accurately, and multimodally. An ablation study was conducted to evaluate the contribution of each modality. The visual-only (ViT+CNN) model achieved 99% accuracy, while the chemical-only (MQ136, MQ137 + Random Forest) model reached 93% accuracy. In contrast, the multimodal fusion achieved 98% accuracy with a ROC-AUC close to 1.0, demonstrating that each modality contributes complementary information and that their integration improves reliability in ambiguous cases.

3.7. Comparative Analysis of Model Performance

To find out how much performance improvement is achieved by the YOLOv11 model proposed in this study, a comparison is made with several previous YOLO versions (YOLOv7 to YOLOv10). The evaluation is carried out on the same dataset to ensure consistency. The evaluation metrics used include precision, recall, mAP@0.5, and mAP@0.5:0.95. The comparison results are shown in table 3.

Table 3. Yolo Model Performance Comparison

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv7	0.75	0.66	0.705	0.412
YOLOv8	0.78	0.68	0.722	0.435
YOLOv9	0.76	0.67	0.710	0.426
YOLOv10	0.79	0.70	0.739	0.456
YOLOv11 (Research)	0.99	1.00	0.99	0.957

Table 3 shows that YOLOv11 consistently outperforms earlier versions, achieving precision 0.997, recall 1.000, mAP@0.5 0.995, and mAP@0.5:0.95 0.957 demonstrating stable and accurate performance through architecture optimization, Adam optimizer, data augmentation, and hyperparameter tuning. To validate the ViT-CNN approach, comparisons were also made with ViT-LSTM, ViT-RNN, ViT-GRU, and ViT-UNet, confirming the effectiveness of ViT-CNN for meat image classification. The performance comparison can be seen in table 4:

Table 4. Comparison of ViT Model Performance with Deep Learning

Researcher	Year	Model	Accuracy
[44]	2022	VIT-Bi-LSTM	86.67%
[45]	2024	ViT-LSTM	91.15%
[46]	2024	ViT-UNet	93.50%
[47]	2025	ViT-RNN	92.10%

 ISSN 2723-6471

[48]	2025	ViT-GRU	98.76%
(This Research)	2025	ViT+CNN	99.00%

Table 4 shows that the combination of ViT with CNN yields the highest accuracy of 99.00%, surpassing all comparison models. The ViT-GRU combination came in second with an accuracy of 98.76%, while ViT-UNet achieved 93.50% which excelled at segmentation but was less efficient in pure classification. ViT-RNN and ViT-LSTM models performed well but still below 93%. Meanwhile, ViT-BiLSTM which is the initial approach recorded the lowest accuracy of 86.67%. These results reinforce the position of ViT-CNN as the most optimal combination in visually detecting meat freshness, by utilizing the strength of ViT's spatial representation and CNN's superiority in local feature capture. To validate the significance of the observed improvements, paired t-tests were conducted on the accuracy and mAP values across 10-fold cross-validation. The results showed that YOLOv11 significantly outperformed YOLOv7–YOLOv10 ($p < 0.05$), and the ViT-CNN model achieved statistically significant improvements compared to ViT-LSTM, ViT-RNN, and ViT-UNet ($p < 0.05$). These findings confirm that the performance gains are not incidental but statistically meaningful.

4. Results and Discussion

This study presents the development of a multimodal beef freshness detection system that integrates visual image processing and IoT-based odor sensing. The visual pipeline employs YOLOv11 for object detection, enhanced with Vision Transformer (ViT) and Convolutional Neural Network (CNN) for classification, while odor data is obtained through MQ136 and MQ137 gas sensors measuring H_2S and NH_3 concentrations. The YOLOv11 model optimized with the Adam optimizer achieved excellent performance (mAP@0.5 = 0.99, mAP50–95 = 0.957), demonstrating stability and fast convergence in handling complex visual data. The combination of ViT and CNN achieved the highest classification accuracy of 99%, outperforming other tested architectures. Multimodal feature fusion using Random Forest yielded a final classification accuracy of 98% with three decision categories: Fit for Consumption, Need to Check, and Not Fit for Consumption, effectively addressing ambiguity in visual and chemical signals. The system offers accurate, real-time detection supported by a Streamlit-based GUI for practical deployment in food safety, retail, and industry. Nevertheless, challenges remain in scalability, long-term sensor durability under environmental variations, and reliance on relatively limited datasets. Future work will focus on in-field testing, dataset expansion across diverse sources, Edge AI integration for autonomous processing, and exploration of temporal-based hybrid architectures such as ViT-LSTM or ViT-GRU to further enhance robustness. Although the experiments were conducted under controlled laboratory conditions, MQ136 and MQ137 are known to be cross-sensitive to other gases and influenced by environmental factors such as humidity and temperature. Addressing these issues through compensation strategies and sensor fusion will be the subject of future research to enhance robustness in real-world deployments. Future work will address this through automated calibration, sensor redundancy, and periodic recalibration to improve robustness in field deployment.

5. Declarations

5.1. Author Contributions

Conceptualization: B.K., R.W., Y., Y.I., and M.H.Y.; Methodology: R.W.; Software: B.K.; Validation: B.K., R.W., and M.H.Y.; Formal Analysis: B.K., R.W., and M.H.Y.; Investigation: B.K.; Resources: R.W.; Data Curation: R.W.; Writing Original Draft Preparation: B.K., R.W., and M.H.Y.; Writing Review and Editing: R.W., B.K., and M.H.Y.; Visualization: B.K. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

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

References

[1] S. N. Hadi and R. H. Chung, "Estimation of Demand for Beef Imports in Indonesia: An Autoregressive Distributed Lag (ARDL) Approach," *Agriculture*, vol. 12, no. 8, pp. 1–12, 2022, doi: 10.3390/agriculture12081212.

[2] S. A. Mehdizadeh, M. Noshad, M. Chaharlangi, and Y. Ampatzidis, "AI-driven non-destructive detection of meat freshness using a multi-indicator sensor array and smartphone technology," *Smart Agric. Technol.*, vol. 10, no. March, pp. 1–10, 2025, doi:10.1016/j.atech.2025.100822.

[3] M. Hashen3mi, M. Salayani, A. Afshari, H. S. Kafil, and S. M. A. Noori, "The Global Burden of Viral Food-borne Diseases: A Systematic Review," *Curr. Pharm. Biotechnol.*, vol. 24, no. 13, pp. 1657–1672, 2023, doi: 10.2174/1389201024666230221110313.

[4] D. L. Prado, E. J. T. Dajang, and I. K. Machica, "FRESHNet : A CNN and YOLO - based Mobile App for Meat Freshness Assessment," *Food Chem.*, vol. 463, no. January, pp. 1-12, 2024, doi: 10.13140/RG.2.2.29355.21289.

[5] S. Dalal, U. K. Lilhore, M. Radulescu, S. Simaiya, V. Jaglan, and A. Sharma, "A hybrid LBP-CNN with YOLO-v5-based fire and smoke detection model in various environmental conditions for environmental sustainability in smart city," *Environ. Sci. Pollut. Res.*, vol. 1, no. January, pp. 1-10, 2024, doi: 10.1007/s11356-024-32023-8.

[6] Calvin, G. B. Putra, and E. Prakasa, "Classification of Chicken Meat Freshness using Convolutional Neural Network Algorithms," *2020 Int. Conf. Innov. Intell. Informatics, Comput. Technol. 3ICT 2020*, vol. 02, no. January, pp. 3–8, 2020, doi: 10.1109/3ICT51146.2020.9312018.

[7] A. Arsalane, A. Klilou, and N. El Barbri, "Performance evaluation of machine learning algorithms for meat freshness assessment," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 5, pp. 5858–5865, 2024, doi: 10.11591/ijece.v14i5.pp5858-5865.

[8] R. Wahyuni, Herianto, Ikhtiyaruddin, and Y. Irawan, "IoT-Based Pulse Oximetry Design as Early Detection of Covid-19 Symptoms," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 3, pp. 177–187, 2023, doi: 10.3991/ijim.v17i03.35859.

[9] Y. Irawan, E. Sabna, A. F. Azim, R. Wahyuni, N. Belarbi, and M. M. Josephine, "Automatic Chili Plant Watering Based on Internet of Things (Iot)," *J. Appl. Eng. Technol. Sci.*, vol. 3, no. 2, pp. 77–83, 2022, doi: 10.37385/jaets.v3i2.532.

[10] Y. Irawan, A. W. Novrianto, and H. Sallam, "Cigarette Smoke Detection and Cleaner Based on Internet of Things (Iot) Using Arduino Microcontroller and Mq-2 Sensor," *J. Appl. Eng. Technol. Sci.*, vol. 2, no. 2, pp. 85–93, 2021, doi: 10.37385/jaets.v2i2.218.

[11] Y. Irawan, R. Wahyuni, and H. Fonda, "Folding Clothes Tool Using Arduino Uno Microcontroller And Gear Servo," *J. Robot. Control*, vol. 2, no. 3, pp. 170–174, 2021, doi: 10.18196/jrc.2373.

[12] A. R. Abidin, Y. Irawan, and Y. Devis, "Smart Trash Bin for Management of Garbage Problem in Society", JAETS, vol. 4, no. 1, pp. 202–208, Sep. 2022, doi: 10.37385/jaets.v4i1.1015.

[13] A. N. Damdam, L. O. Ozay, C. K. Ozcan, A. Alzahrani, R. Helabi, and K. N. Salama, "IoT-Enabled Electronic Nose System for Beef Quality Monitoring and Spoilage Detection," *Foods*, vol. 12, no. 11, pp. 1-10, 2023, doi: 10.3390/foods12112227.

[14] Z. W. Bhuiyan, S. A. R. Haider, A. Haque, M. R. Uddin, and M. Hasan, "IoT Based Meat Freshness Classification Using Deep Learning," *IEEE Access*, vol. 12, no. October, pp. 196047–196069, 2024, doi: 10.1109/ACCESS.2024.3520029.

[15] Z. Xun, X. Wang, Hao. X, "Deep machine learning identified fish flesh using multispectral imaging," *Curr. Res. Food Sci.*, vol. 9, no. June, pp. 1-12, 2024, doi: 10.1016/j.crfs.2024.100784.

[16] D. Setiawan, R. N. Putri, I. Fitri, A. N. Hidayanto, Y. Irawan, and N. Hohashi, "Improved Deep Learning Model for Prediction of Dermatitis in Infants," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 871–884, 2025, doi: 10.47738/jads.v6i2.542.

[17] T. Anwar and H. Anwar, "Beef quality assessment using AutoML," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, vol. 01, no. July, pp. 1–4, 2021, doi: 10.1109/MAJICC53071.2021.9526256.

[18] J. M. Lee, I. H. Jung, and K. Hwang, "Classification of Beef by Using Artificial Intelligence," *J. Logist. Informatics Serv. Sci.*, vol. 9, no. 1, pp. 1–10, 2022, doi: 10.33168/liss.2022.0101.

[19] E. N. Cahyo, E. Susanti, and R. Y. Ariyana, "Model Machine Learning Untuk Klasifikasi Kesegaran Daging Menggunakan Arsitektur Transfer Learning Xception," *J. Sist. dan Teknol. Inf.*, vol. 11, no. 2, pp. 371-284, 2023, doi: 10.26418/justin.v11i2.57517.

[20] A. Denih and I. Anggraeni, "Beef Freshness Detection Device Based on Gas and Color Sensors using the K- Nearest Neighbor Method," *Ind. Eng. Oper. Manag. Manila*, vol. 7, no. March, pp. 2539–2545, 2023, doi: 10.46254/an13.20230690.

[21] O. B. Yurdakos and O. Cihanbegendi, "System Design Based on Biological Olfaction for Meat Analysis Using E-Nose Sensors," *ACS Omega*, vol. 9, no. 30, pp. 33183–33192, Jul. 2024, doi: 10.1021/acsomega.4c04791.

[22] Y. Lin, J. Ma, D. W. Sun, J. H. Cheng, and C. Zhou, "Fast real-time monitoring of meat freshness based on fluorescent sensing array and deep learning: From development to deployment," *Food Chem.*, vol. 448, no. March, pp. 1-11, 2024, doi: 10.1016/j.foodchem.2024.139078.

[23] J. Zhang, W. Jizhong, W. Wenya, "Olfactory imaging technology and detection platform for detecting pork meat freshness based on IoT," *Comput. Electron. Agric.*, vol. 215, no. December, pp. 1-10, 2023, doi: 10.1016/j.compag.2023.108384.

[24] A. Yudhana, R. Umar, and S. Saputra, "Fish Freshness Identification Using Machine Learning: Performance Comparison of k-NN and Naïve Bayes Classifier," *J. Comput. Sci. Eng.*, vol. 16, no. 3, pp. 153–164, 2022, doi: 10.5626/JCSE.2022.16.3.153.

[25] K. Kiswanto, H. Hadiyanto, and E. Sediyono, "Meat Texture Image Classification Using the Haar Wavelet Approach and a Gray-Level Co-Occurrence Matrix," *Appl. Syst. Innov.*, vol. 7, no. 3, pp. 1-10, 2024, doi: 10.3390/asi7030049.

[26] I. H. Kozan and H. A. Akyurek, "Development Of A Mobile Application For Rapid Detection Of Meat Freshness Using Deep Learning," *Theory Pract. Meat Process.*, vol. 9, no. 3, pp. 249–257, 2024.

[27] A. Febriani, R. Wahyuni, Y. Irawan, and R. Melyanti, "Improved Hybrid Machine and Deep Learning Model for Optimization of Smart Egg Incubator," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1052–1068, 2024, doi: 10.47738/jads.v5i3.304.

[28] D.-W. Sun, H. Pu, and J. Yu, "Applications of hyperspectral imaging technology in the food industry," *Nat. Rev. Electr. Eng.*, vol. 1, no. 4, pp. 251–263, 2024, doi: 10.1038/s44287-024-00033-w.

[29] A. Rabehi, H. Helal, D. Zappa, and E. Comini, "Advancements and Prospects of Electronic Nose in Various Applications: A Comprehensive Review," *Appl. Sci.*, vol. 14, no. 11, pp. 1-9, 2024, doi: 10.3390/app14114506.

[30] S. Tang and W. Yan, "Utilizing RT-DETR Model for Fruit Calorie Estimation from Digital Images," *Information*, vol. 15, no. 8, pp. 1-10, 2024, doi: 10.3390/info15080469.

[31] E. Hassan and H. Ghadiri, "Advancing brain tumor classification: A robust framework using EfficientNetV2 transfer learning and statistical analysis," *Comput. Biol. Med.*, vol. 185, no. February, pp. 1-12, 2025, doi: 10.1016/j.compbiomed.2024.109542.

[32] D. Rastogi, P. Johri, and V. Tiwari, "Augmentation based detection model for brain tumor using VGG 19," *Int. J. Comput. Digit. Syst.*, vol. 13, no. 1, pp. 1227–1237, 2023, doi: 10.12785/ijcds/1301100.

[33] H. A. Akbaci and E. Bayraktar, "Trajectory refinement in SLAM: the impact of Adam, AdamW, and SGD with momentum," in *Proc.SPIE*, vol. 13540, no. Feb, pp. 1-10, 2025, doi: 10.1117/12.3056417.

[34] N. S.T. and J. V Gorabal, "Design and Development of Multimodal Biometric System Using Finger Veins and Iris by CNN Integrated with Hybrid SIO and Whale Optimization Techniques," *Int. J. Interact. Mob. Technol.*, vol. 18, no. 22, pp. 97–114, 2024, doi: 10.3991/ijim.v18i22.50865.

[35] M. K. Anam, L. L. Van FC, H. Hamdani, R. Rahmaddeni, J. Junadhi, M. B. Firdaus, I. Syahputra, and Y. Irawan, "Sara Detection on Social Media Using Deep Learning Algorithm Development," *J. Appl. Eng. Technol. Sci.*, vol. 6, no. 1, pp. 225–237, Dec. 2024, doi: 10.37385/jaets.v6i1.5390.

[36] S. Tomar, D. Dembla, and Y. Chaba, "Analysis and Enhancement of Prediction of Cardiovascular Disease Diagnosis using Machine Learning Models SVM, SGD, and XGBoost," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, pp. 469–479, 2024, doi: 10.14569/IJACSA.2024.0150449.

[37] Y. Thanet, L. Potsirin, N. Wongpanya, and P. Nuankaew, "Information Systems for Cultural Tourism Management Using Text Analytics and Data Mining Techniques," *Int. J. Interact. Mob. Technol.*, vol. 16, no. 09, pp. 146–163, 2022, doi: 10.3991/ijim.v16i09.30439.

[38] H. Fonda, Y. Irawan, R. Melyanti, R. Wahyuni, and A. Muhaimin, "A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1701–1714, 2024, doi: 10.47738/jads.v5i4.388.

[39] K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education," *Comput. Educ. Artif. Intell.*, vol. 6, no. September, pp. 1-10, 2024, doi: 10.1016/j.caeai.2024.100205.

[40] P. Nasa-Ngium, W. S. Nuankaew, and P. Nuankaew, "Analyzing and Tracking Student Educational Program Interests on Social Media with Chatbots Platform and Text Analytics," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 05, pp. 4–21, 2023, doi: 10.3991/ijim.v17i05.31593.

[41] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1272–1285, 2024, doi: 10.47738/jads.v5i3.316

[42] A. Lubis, Y. Irawan, Junadhi, and S. Defit, "Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1625–1638, 2024, doi: 10.47738/jads.v5i4.343.

[43] K. A. Rashedi, M. T. Ismail, S. Al Wadi, A. Serroukh, T. S. Alshammari, and J. J. Jaber, "Multi-Layer Perceptron-Based Classification with Application to Outlier Detection in Saudi Arabia Stock Returns," *J. Risk Financ. Manag.*, vol. 17, no. 2, pp. 1-12, 2024, doi: 10.3390/jrfm17020069.

[44] H. Chen, J. Cui, Y. Zhang, and Y. Zhang, "VIT and Bi-LSTM for Micro-Expressions Recognition," in *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, vol. 9927522, no. September, pp. 946–951, 2022. doi: 10.1109/ICISCAE55891.2022.9927522.

[45] A. B. Nassif, I. Shahin, M. Bader, A. Ahmed, and N. Werghi, "ViT-LSTM synergy: a multi-feature approach for speaker identification and mask detection," *Neural Comput. Appl.*, vol. 36, no. 35, pp. 22569–22586, 2024, doi: 10.1007/s00521-024-10389-7.

[46] N. Zhou, M. Xu, "ViT-UNet: A Vision Transformer Based UNet Model for Coastal Wetland Classification Based on High Spatial Resolution Imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, no. October, pp. 19575–19587, 2024, doi: 10.1109/JSTARS.2024.3487250.

[47] A. R. Borah, A. A. Hameed, H. P. Thethi, J. L. Prasanna, A. Sangeetha, and D. D. Gautam, "ViT and RNN for Temporal and Spatial Analysis in Video Sequences," in *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*, vol. 2025, no. February, pp. 651–656, 2025, doi: 10.1109/IC363308.2025.10957553.

[48] M. J. Zobair, M. A. Rahman, M. S. Hossain, N. Khan, M. A. A. K. Akash, and M. H. I. Bijoy, "A Hybrid ViT-GRU Model for Breast Cancer Detection: Addressing Class Imbalance Challenges," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, vol. 2025, no. February, pp. 1–7, 2025, doi: 10.1109/ECCE64574.2025.11013095.