Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD26898

Multi-Scenario Evaluation of an Enhanced YOLOv8s-cls Model for Food Image Classification

Songyin Ran

School of Computer Science, Wuhan University, Wuhan, China
2022302111479@whu.edu.cn

Abstract. As the demand for intelligent dietary management continues to grow, food recognition technology shows great application prospects of areas like mobile platforms. In tackling the essential challenge of food image classification, the You Only Look Once (YOLO)v8s-cls architecture is augmented with the Focal Loss function in this paper, aiming to improve its discrimination capability for challenging or minority-class samples. The model is trained on a large food dataset containing 288 classes, and is optimized using data augmentation strategies such as MixUp and CutMix. It was evaluated in multiple application scenarios, including normal images, occluded images, blurred images, and high/low-frequency classes subsets. On normal images, the enhanced model attained 80.64% in Top-1 accuracy and 95.73% in Top-5 accuracy, outperforming the original model in robustness under challenging conditions. Overall, the results indicate that the Focal Loss function has improved the model's capability to classify diverse food images, supporting the practical implementation of intelligent dietary systems.

Keywords: Food image classification, YOLOv8s-cls, Focal Loss, Data augmentation, Multiple scenarios.

1. Introduction

Alongside the steady rise in living standards these years, there has been a growing societal emphasis on adopting scientifically informed dietary habits and maintaining nutritional well-being [1]. Intelligent dietary management has thus been widely developed in areas such as mobile platforms [1]. As a key application of computer vision in intelligent dietary management, food image classification techniques are receiving more attention [1]. In this context, it is of great practical significance and application value to construct a food image classification model with good robustness and easy deployment [1].

Convolutional neural networks (CNN) have been widely applied to food recognition tasks for over a decade, yielding consistently promising results [2]. Kagaya et al. applied CNN to a 10-classes food classification task and achieved an accuracy of 73.7% [3]; On the UEC-FOOD-100 dataset, Kawano et al. employed a CNN exclusively for feature extraction, which resulted in 72.3% classification accuracy [4,5]. Although CNN-based models have performed consistently in standard classification tasks and outperformed traditional machine learning methods [2], they may face significant degradation in accuracy and model inference speed when applied to real-world scenarios.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD26898

In recent years, food recognition techniques employing CNNs, instance-based learning such as k-nearest neighbors (k-NN), and margin-based classifiers like support vector machines (SVM) have demonstrated accuracy rates up to 86.85% in recent studies [6]. Elbassuoni et al. constructed a nutrition-oriented food image classifier based on CNN, achieving over 94% accuracy on a small-scale dataset [7]. Nevertheless, the model's excessive parameter counts pose challenges for deployment on mobile devices.

The You Only Look Once (YOLO) family represents a rapidly evolving line of end-to-end object detection frameworks. It is widely recognized in the computer vision community for combining high recognition accuracy with real-time inference [8]. Released by Ultralytics in 2023, YOLOv8 is the first generation in the YOLO series to adopt a unified architecture that supports multiple vision tasks —namely object detection, image classification, semantic segmentation, and pose estimation— within a single framework. It marks a significant advancement in both performance and flexibility, and has seen adoption across a wide range of application areas [9,10]. Compared with traditional CNN models, YOLOv8 offers advantages of faster inference speed, more compact model size, and greater deployment flexibility, while maintaining high accuracy and stronger scene adaptability. Additionally, the lightweight classification model YOLOv8s-cls can be directly applied to image recognition tasks, which provides a solid foundation for building accurate, efficient, and deployable food image classification systems [11].

This paper collected data from several food image datasets, including Food-101. And investigates food image classification tasks in complex backgrounds using YOLOv8s-cls as the base model. The generalization performance and robustness of the model were strengthened through the integration of data augmentation techniques such as MixUp and CutMix. In addition, Focal Loss was adopted to intensify the network’s emphasis on hard-to-classify samples. By integrating these techniques, a compact food image classification model was established, achieving a balanced compromise between predictive accuracy and computational efficiency.

2. Method

The food image classification method in this paper consists of three main stages: (1) data acquisition and processing; (2) model training and performance evaluation; (3) definition of multiple scenarios.

2.1. Data acquisition and processing

In this experiment, several public food image datasets (e.g., Food-101, ChineseFoodNet, etc.) were collected, filtered, and merged for subsequent use. The original datasets included some classes with very few samples. To ensure balanced training and enhance generalization, 22 classes with fewer than 100 samples were removed, and finally 288 classes are retained. The final dataset consisted of approximately 200,000 training samples, 20,000 validation samples, and 45,000 test samples. A partial class distribution is presented in Table 1.

DOI: 10.54254/2755-2721/2025.LD26898

Table 1. Number of images per class

Class name	Training set	Validation set	Test set	Total images
Abalone	664	80	104	848
apple_pie	670	80	250	1000
baby_back_ribs	670	80	250	1000
baklava	670	80	250	1000
Bean_curd	337	35	51	423
beef_carpaccio	670	80	250	1000
Beef_Seasoned_with_Soy_Sauce	856	80	131	1067

Before model training, the raw food images were first resized to a uniform resolution of 224 × 224. And the dataset directory structure was reconstructed based on class labels to meet the input data format requirements of YOLOv8s-cls.

2.2. Model training and performance evaluation

2.2.1. Experimental hardware and software environment

The experiments were performed on a laptop powered by an AMD Ryzen 9 6900HX processor with Radeon Graphics and 16 GB of memory. The system includes an NVIDIA GeForce RTX 3060 Laptop GPU with 6 GB of VRAM. The operating environment is Windows 11 64-bit, with CUDA version 12.2 and GPU driver version 552.12. The Python environment is managed by Miniconda, and development and debugging were performed using PyCharm. Model training and evaluation were implemented in PyTorch 2.4.1 (CUDA 12.1), incorporating TorchVision 0.19.1, Torchaudio 2.4.1, and Ultralytics YOLOv8 v8.3.161.

2.2.2. Model training configuration

This study employs the YOLOv8s-cls base model provided by Ultralytics, utilizing the official Python API for training and inference tasks. The model training configuration is presented in Table 2.

DOI: 10.54254/2755-2721/2025.LD26898

Table 2. Training configuration

Parameter	Value
Task	Classification
Mode	Train
Epochs	60
Batch Size	16
Image Size	224
Learning Rate	0.01
Momentum	0.937
MixUp	0.2
CutMix	0.2
Erasing	0.4
Fliplr	0.5

To strengthen the model's reliability when faced with complex and unpredictable real-world inputs, the training phase incorporated multiple data augmentation strategies, including MixUp (0.2) and CutMix (0.2). Furthermore, random erasing (0.4), and horizontal flipping (0.5) were applied to further enhance the diversity of the training samples.

2.2.3. Loss function

The dataset contains classes with large frequency differences. Traditional cross-entropy loss often focuses on high-frequency classes. This bias reduces recognition performance for low-frequency classes. To address this, Focal Loss is adopted as the primary loss function. It effectively mitigates class imbalance in the food image classification task. As detailed below:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

In Eq. (1), p_t denotes the probability that the model predicts a sample of class t correctly; The modulation factor γ reduces the relative loss for easily classified samples, and its value is fixed at 0.2; α_t is the weighting factor used to balance the importance of different classes. In this study, Inverse Probability Weighting (IPW) was applied to assign initial weights to each class, as detailed below:

$$\alpha_t = \frac{N}{C \cdot n_t} \tag{2}$$

In Equation (2), N represents the total number of samples in the training set; C represents the total number of classes; and n_t indicates the number of samples belonging to class t. Equation (2) calculates weights based on the frequency of each class in the training set, with weights inversely proportional to the class frequencies. This increases the loss contribution of rare classes, thereby encouraging the model to pay more attention to them while training. Since most class weights computed by Equation (2) exceed 1, each weight is divided by the sum of all initial weights to obtain the final class weights.

DOI: 10.54254/2755-2721/2025.LD26898

2.2.4. Evaluation metrics

For a thorough quantitative evaluation of the model, commonly used metrics such as training and validation losses, Top-1/Top-5 accuracies, and the weighted F1-score are adopted.

Training loss and validation loss are used to measure the difference between the model prediction and the true labels. Both are computed based on the Focal Loss defined in Equations (1). The training loss reflects how well the model fits on the training set, while validation loss is used to assess the model's generalization ability. Top-1 accuracy, widely adopted in image classification research, quantifies the proportion of test samples for which the model's most confident prediction corresponds correctly to the ground truth label. It is one of the key metrics for evaluating classification model performance. Top-5 accuracy quantifies the likelihood that the true class label is included within the five highest-ranked predictions of the model, offering a more tolerant performance indicator compared to the stricter Top-1 accuracy. Weighted F1-score considers class imbalance by computing the F1-score for every class and then taking a weighted average based on the number of true samples in each class.

In this study, each epoch comprises two main phases: the initial training is completed on the training dataset, followed by evaluation on the validation set for parameter tuning and accuracy monitoring (validation results are shown in Figure 1). At the end of each training epoch, performance indicators, including loss values for both training and validation, as well as classification accuracies under the Top-1 and Top-5 criteria, were recorded. Upon completion of training, the model achieved the performance presented in Table 3 on the validation set.



Figure 1. Val_batch1_labels(left) and Val_batch1_pred(right)

Table 3. Validation set performance

Model	Top-1 Accuracy	Top-5 Accuracy	Training loss	Validation loss
Focal Loss	77.43%	93.91%	0.7499	0.8933
Original	73.83%	92.52%	0.9066	1.1594

The results presented in Table 3 demonstrate that the enhanced model achieves superior outcomes in each assessed performance metric compared to the original one.

DOI: 10.54254/2755-2721/2025.LD26898

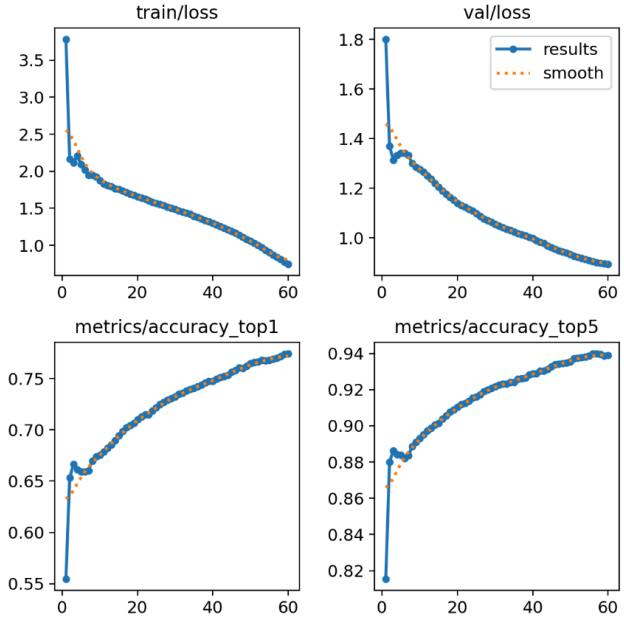

Figure 2. Trends of metrics over training epochs

To provide more intuitive views of the model's performance evolution during the training, the curves of various metrics for the Focal Loss model are plotted in Figure 2. From the curves, the model converges rapidly in the early stage and gradually stabilizes in the late stage. With no evident signs of overfitting or underfitting, the classification model achieved the desired training outcomes. It lays a solid foundation for subsequent generalization testing across multiple application scenarios.

2.3. Definition of multiple scenarios

In practical application scenarios, the uploaded food images may not be complete or clear. Additionally, some food classes may have limited training samples, resulting in low prediction accuracy. To make a comprehensive evaluation of the model's performance in realistic scenarios, five kinds of representative test sets were created:

Normal Images (original test set) — Images with clean backgrounds, uniform lighting, moderate shooting angles, and complete, clear food subjects. These are standard images without any occlusion or blur applied, encompassing all images in the test set as shown in Figure 3a.

Blurred Images — Samples affected by motion blur or compression artifacts causing blurriness or defocus. These are generated by applying mild Gaussian blur to all testing set images, forming a new test subset, as illustrated in Figure 3b.

Occluded Images — Images containing occluding objects such as utensils, hands, packaging, or plate edges, obscuring parts of the food area (simulated with black blocks). The occlusion ratio varies progressively from 0.2×0.2 to 0.8×0.8. These sets are created by applying varying degrees of occlusion to all test images, as depicted in Figure 3c1-4.

High-Frequency Classes Images — A subset of classes with abundant samples, specifically those appearing more than 670 times in the training set.

Low-Frequency Classes Images — A subset of classes with limited samples, defined as those appearing fewer than 300 times in the training set.

DOI: 10.54254/2755-2721/2025.LD26898

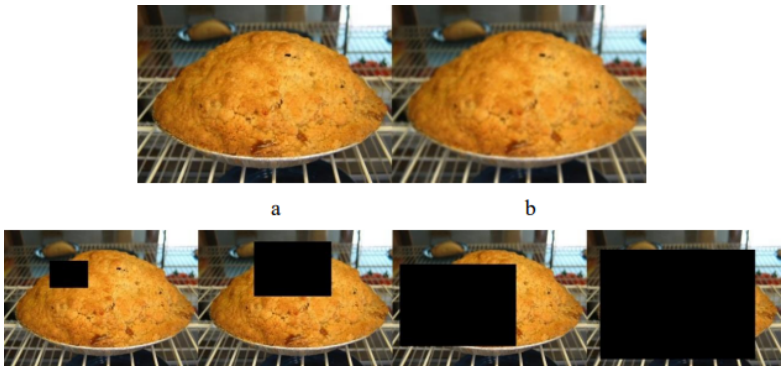

Figure 3. Example images (a: normal, b: blurred and c1-c4: occluded images)

3. Results and analysis

This experiment employs both the trained Focal Loss model and the original model to perform predictions on the five kinds of test sets. Top-1 accuracy, Top-5 accuracy, and weighted F1-score are adopted as evaluation metrics to measure the model's recognition accuracy and stability across different scenarios. Each metric was averaged over multiple repeated measurements to ensure the reliability and representativeness of the evaluation results. Detailed outcomes are presented in Table 4.

Table 4. Classification accuracy overview

Test set	Focal Loss			Original		
	Top-1 Accuracy	Top-5 Accuracy	F1-score(weighted)	Top-1 Accuracy	Top-5 Accuracy	F1-score(weighted)
Normal	80.64%	95.73%	0.805	77.06%	94.25%	0.768
Blurred	77.56%	94.44%	0.773	72.44%	91.55%	0.722
Occluded (0.2×0.2)	80.16%	95.60%	0.800	76.78%	94.11%	0.765
Occluded (0.4×0.4)	77.56%	94.29%	0.774	74.31%	92.79%	0.740
Occluded (0.6×0.6)	66.89%	87.59%	0.666	63.33%	85.86%	0.629
Occluded (0.8×0.8)	36.73%	59.20%	0.382	32.86%	55.15%	0.352
High-Frequency	82.63%	96.36%	0.833	78.80%	94.96%	0.798
Low-Frequency	62.96%	89.03%	0.753	58.98%	85.73%	0.716

DOI: 10.54254/2755-2721/2025.LD26898

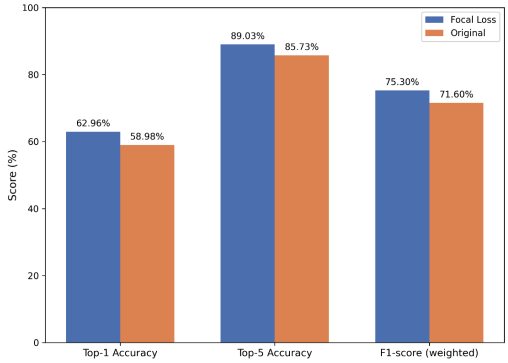

Figure 4. Performance on low-frequency subset

Overall, after the introduction of Focal Loss, the model achieves considerable improvement in all five kinds of application scenarios. Notably, for low-frequency classes (Figure 4), the Focal Loss model outperforms the original model in all three key metrics, demonstrating enhanced robustness under imbalanced data conditions. This indicates that incorporating Focal Loss effectively mitigates the model's tendency to overlook low-frequency classes. Such improvements in low-frequency classes are particularly valuable.

3.1. Analysis of blurred and occluded images

As can be seen from Table 4, the classification performance of Focal Loss model under different occlusion ratio and image blur conditions have some differences.

Figure 5 illustrates the changes in Top-1 accuracy, Top-5 accuracy, and weighted F1-score of the Focal Loss model under varying occlusion ratios. As shown, the model's Top-1 accuracy decreases from 80.64% to 36.73% as the occlusion ratio gradually increases from 0 to 0.8. Similarly, the Top-5 accuracy declines from 95.73% to 59.20%, and the F1-score drops from 0.805 to 0.382, which shows a significant decreasing trend overall. Notably, when the occlusion ratio exceeds 0.6, the model's recognition performance deteriorates substantially, likely due to difficulties in accurately extracting critical features from heavily occluded images. However, the model still maintains stable classification performance under mild occlusion conditions, indicating good adaptability to minor partial occlusion.

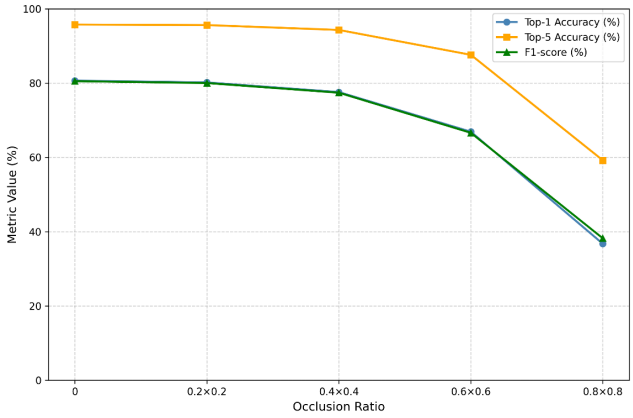

Figure 5. Performance of Focal Loss model under different occlusion levels

DOI: 10.54254/2755-2721/2025.LD26898

By contrast, evaluation on the blurred dataset yielded a Top-1 accuracy of 77.56%, along with a Top-5 accuracy of 94.44% and an F1-score of 0.773. Although these metrics are slightly lower than those for normal images—80.64%, 95.73%, and 0.805 respectively—the overall impact is less severe than under occlusion conditions, and notably better than the model's performance at moderate to high occlusion levels (0.6 and 0.8).

The comprehensive analysis shows that both image occlusion and blurring adversely interfere with the model performance, especially when the occlusion ratio is high. The model demonstrates strong adaptability to low-level occlusion and blurring, which may benefit from the multiple data augmentation and regularization strategies when training. To further enhance the model's stability in complex scenarios, future work may consider incorporating attention mechanisms or designing specific data augmentation strategies that simulate real-world occlusion and blurring conditions to boost model generalization. Meanwhile, in practical applications, it is necessary to minimize excessive occlusion of the main food subject, avoid too-far or too-close shooting distances, and prevent overly blurred images. Ensuring that the food image is complete and clearly visible from the outset, thereby improving recognition accuracy and stability.

3.2. Analysis of blurred and occluded images

When analyzing the performance of the Focal Loss model, it can be clearly found that its performance in the high-frequency subset is significantly better than the low-frequency subset. Specifically, the Top-1 and Top-5 accuracies for the high-frequency subset reach 82.63% and 96.36%, respectively, whereas these figures drop to 62.96% and 89.03% for the low-frequency subset. This disparity highlights the model's stronger discriminative capability and higher classification stability when dealing with categories that have sufficient training samples.

Although the Focal Loss model achieves a noticeable improvement over the original model in the low-frequency classes subset, its performance still exhibits a substantial decline compared with that on the high-frequency subset. These results suggest that Focal Loss cannot completely mitigate the performance disparity induced by data imbalance. From an application perspective, the Focal Loss model has achieved a substantial enhancement on the low-frequency classes, marking a transition from unusable to usable. However, there remains considerable room for improvement to achieve better accuracy and stability in these low-frequency images.

4. Discussion

This study systematically evaluated the performance of the improved YOLOv8s-cls model under different image qualities (e.g., occlusion, blur) and class frequency conditions (high-frequency vs. low-frequency classes). Compared to the original model, considerable improvements were observed. But there are still several issues of concern to be further explored in the subsequent research.

Firstly, class imbalance remains a critical challenge. After incorporating Focal Loss, metrics for low-frequency classes improved significantly by 3.98% (Top-1 accuracy) and 3.30% (Top-5 accuracy). This has mitigated the performance degradation caused by the long-tail distribution to some extent. Despite these improvements, the overall performance on low-frequency classes is still unsatisfactory, and the F1-score further indicates limited recall ability. These results suggest that the current strategy is not sufficient to fundamentally solve the degradation of the model's generalization ability caused by extreme class imbalance. Therefore, future research should consider integrating advanced approaches such as few-shot generation and data augmentation to systematically enhance performance on low-frequency classes.

DOI: 10.54254/2755-2721/2025.LD26898

Secondly, there is room for improvement in the design of occlusion and blur experiments. Although this paper used block occlusion and Gaussian blur with a fixed standard deviation to simulate image degradation to some extent, real-world occlusions may arise from dynamic objects, scene overlap, or background interference. Similarly, blurs can be caused by factors such as motion, focus failure, or compression artifacts. Therefore, the current experiments cannot fully cover the model's robustness under complex visual disturbances. Future work should incorporate more realistic image degradation patterns to enhance the generalizability of the experimental results.

Finally, this study primarily evaluated performance based on the single YOLOv8s-cls architecture and has yet to verify the universality and effectiveness of the robustness strategies on other representative models such as ResNet50, EfficientNet, and MobileNetV2. Subsequent research will include comparative analyses across multiple architectures to explore the transferability and differences of the proposed improvements across various network scales and structures.

5. Conclusion

This work proposes a YOLOv8s-cls-based food classification framework incorporating Focal Loss to improve recognition accuracy for intelligent dietary applications. To thoroughly evaluate the model's performance under diverse conditions, comprehensive experiments were carried out on a large-scale dataset, revealing that the improved model surpasses the original one.

On the original test set, the refined model attained a Top-1 accuracy of 80.64%, a Top-5 accuracy of 95.73%, and a weighted F1-score of 0.805 after completing 60 training epochs, indicating robust classification capability. Under various scenarios such as high-frequency classes, mild occlusion, and slight blur, the model maintained high accuracy, showing a certain level of robustness across different scenarios. Additionally, the model features a compact parameter size and fast inference speed, offering good adaptability for deployment.

The proposed food classification model based on the improved YOLOv8s-cls achieves a balanced performance in terms of accuracy, model size, and recognition speed. It maintains relatively stable classification performance under various real-world interferences such as occlusion, blur, and lighting variations. Through the refinement of the loss function, the proposed model demonstrates enhanced proficiency in identifying insufficient training and visually diverse food classes, which substantially increases its potential for practical use within intelligent dietary management systems. However, the model still exhibits certain misclassifications when handling severely occluded images or low-frequency classes. Future work will focus on further improving the model's recognition and generalization capabilities for complex food images by employing techniques such as few-shot generation and augmentation, as well as increasing the diversity and quantity of training samples.

References

[1] VijayaKumari, G., Vutkur, P., & Vishwanath, P. (2022). Food classification using transfer learning technique. Global transitions proceedings, 3(1), 225-229.
[2] Shimoda, W., & Yanai, K. (2015). CNN-Based Food Image Segmentation Without Pixel-Wise Annotation. International Conference on Image Analysis and Processing. Springer International Publishing.
[3] Kagaya, H., Aizawa, K., & Ogawa, M. (2014, November). Food detection and recognition using convolutional neural network. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 1085-1088).
[4] Kawano, Y., & Yanai, K. (2014, September). Food image recognition with deep convolutional features. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (pp. 589-593).
[5] Matsuda, Y., Hoashi, H., & Yanai, K. (2012, July). Recognition of multiple-food images by detecting candidate regions. In 2012 IEEE international conference on multimedia and expo (pp. 25-30). IEEE.

DOI: 10.54254/2755-2721/2025.LD26898

[6] Srigurulekha, K., & Ramachandran, V. (2020, January). Food image recognition using CNN. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-7). IEEE.
[7] Elbassuoni, S., Ghattas, H., El Ati, J., Shmayssani, Z., Katerji, S., Zoughbi, Y., ... & Sassi, S. (2022). DeepNOVA: A Deep learning NOVA classifier for food images. IEEE Access, 10, 128523-128535.
[8] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. Procedia computer science, 199, 1066-1073.
[9] Saha, U., Ahamed, I. U., Imran, M. A., Ahamed, I. U., Hossain, A. A., & Gupta, U. D. (2024, November). Yolov8-based deep learning approach for real-time skin lesion classification using the ham10000 dataset. In 2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom) (pp. 1-4). IEEE.
[10] Dave, B., Mori, M., Bathani, A., & Goel, P. (2023). Wild animal detection using YOLOv8. Procedia Computer Science, 230, 100-111.
[11] Mo, H., Qu, Y., Li, X., & Wang, Z. (2024, December). Classification of Liver Tumors Based on YOLOv8s-cls. In 2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI) (pp. 1-5). IEEE.