



Regional Food Recognition with Nutritional Analysis: Attention Mechanism Implementation for Indonesian Cuisine

Ahmad Fauzan* and Hedy Pamungkas

Cakrawala University, Jakarta, Indonesia
*ahmad.fauzan@cakrawala.ac.id

Abstract. Indonesia faces two major challenges: the preservation of its food heritage—at risk of digital extinction with 5,350 traditional recipes—and growing public health issues such as diabetes, which increased from 6.9% to 10.9% between 2013 and 2018. Tools that support culturally relevant food tracking are therefore urgently needed. This study has two main aims: (1) to evaluate whether a Convolutional Block Attention Module (CBAM) can improve food classification accuracy for Indonesian dishes, and (2) to assess whether a multi-task architecture can accurately predict nutritional categories using visual features alone. We trained and compared four EfficientNet-B0–based models (Vanilla, CBAM Single, Multi-Task, CBAM Multi-Task) using 815 images across 10 Indonesian food classes. Performance was evaluated through classification accuracy and nutritional prediction F1-score. Contrary to expectations, the CBAM Single-Task model (83.0% accuracy) did not outperform the baseline (85.0%). The best food classification result was achieved by the Multi-Task baseline (88.0%). For nutrition prediction, the CBAM Multi-Task model achieved the highest performance (91.1% F1-score). All models remained computationally efficient, with 3.6–4.0M parameters. The findings reveal a trade-off: multi-task learning provided strong regularization for improving food recognition, while CBAM was more effective when applied specifically to nutritional analysis. The high F1-score demonstrates the feasibility of visual-only nutrition prediction without relying on external databases. This study establishes competitive benchmarks for Indonesian food recognition (88.0%) and visual-based nutritional analysis (91.1%). The integrated framework shows strong potential for deployment in mobile health applications that support public health monitoring and cultural preservation in Indonesia.

Keywords: Attention Mechanism, Food Recognition, Indonesian Cuisine, Mobile Deployment, Multi-Task Learning, Nutritional Analysis.

1 Introduction

Indonesia's dilemma today, Indonesia battles with two main dilemmas – conserving its diverse cuisines (over 5,350 traditional dishes online are on the brink of digital extinction), and escalating public health challenges. Based on the Ministry of Health, between 2013-2018 diabetes prevalence climbed from 6.9% to 10.9%, and although there is a high level of awareness, balanced nutrition was not attained by 75.7 per cent among population [1]. This large knowledge-practice gap, suggests a pressing need for culturally-sensitive technological support; however, 85% of Indonesians have not used a food recording app (primarily due to poor coverage of indigenous foods).

Existing computer vision-based food recognition systems achieve reasonable performance (although they tend to overfit) on Western datasets, however their accuracy decreases by 15-20% when tested on culturally-specific Asian food. Moreover, the previous models for Indonesian food and dishes only achieved 81.51-85.52% accuracy which are not ideal for real-world implementation as well. The performance gap is due to the lack of large-scale and diversified Indonesian food datasets as well as unavailable comprehensive nutritional information, which can dissuade the development of dietary tracking tools.

Recent studies have been shown that the introduction of additional modalities to the model (such as depth) can significantly reduce estimation errors in terms of both calories and mass for food nutrition prediction [2]. such as employing a RGB-D fusion network with Convolutional Block Attention Module (CBAM). Similarly, Qi et al. [3], who introduced the Visual-Ingredient Feature Fusion (VIF2) approach and demonstrated that integrating visual features with ingredient-level features improves nutritional estimation in large datasets like FastFood and Nutrition5k [3]. More recent multimodal models, such as FoodSAM, [4] rely on foundation models for holistic food understanding and offer scalable and generalizable frameworks that may be extended to underrepresented cuisines like Indonesian food.

Attention model such as Convolutional Block Attention Module (CBAM), which focuses on texture have shown promising results in addressing challenging food recognition tasks especially the Indonesian cuisine as they have like very similar foods based on texture (i.e. rendang and semur). Nevertheless, to the best of our knowledge no attempt has been made to adopt such attention mechanisms systematically on Indonesian food. In addition, current systems disregard to include nutritional analysis as part of the process presenting it as a separate task that needs to be done in tandem using external databases which are usually not available for conventional foods. For instance, Multi-Task Image-Based Dietary Assessment on Food Recognition and Portion Size Estimation [5]. Introduced an end-to-end multi-task framework to further recognize food category and estimate portion size, showing that integrating these tasks can benefit more practical applications. Also, some recent works like Vision-Based Approach for Food Weight Estimation from 2D Images [6]. Provide good accuracy in food weight estimation, an important variable when estimating a nutrient content. These indicate that tasks combining classification with portion or nutrient estimation are viable, and inspire our work on directly predicting nutritional categories based on visual features.

Accordingly, we have two goals in this paper: first, to examine whether an EfficientNet-B0 backbone model equipped with a CBAM attention mechanism can outperform the initial classification mode for texturally-complex Indonesian foods; and second, to investigate if multi-task learning architecture can correctly predict multi-label nutritional categories (highCarb-HighProtein- HighFat) from visual features alone. This paper contributes research to close the above-mentioned research gaps (section 2), by experimenting with and comparing four different model architectures – a vanilla baseline, a multi-task baseline, CBAM single-task model and a composite CBAM multi-task models – to determine optimal technique.

Two main priorities are addressed in this study. First, we carry out a systematic comparison of multi-task learning and CBAM attention at the Indonesian food recognition task, where it is found that multi-task baseline model successfully achieved higher accuracy (88.0%) in classifying the foods. Second, we show that visual-based nutrition analysis is feasible and accessible, achieving up to 91.1% F1-score in a nutshell for nutrition prediction without any dependence on external databases.

This paper consists of five sections. Section two is a literature review of food recognition systems, mechanisms of attention, and nutritional analysis. Section three talks about the research methodology including model architecture, dataset preparation, and metrics of evaluation. Section four presents experimental results and performance analysis. Section five presents conclusions and future directions of research for regional food recognition systems.

2 Literature Review

2.1 Limitations of Food Recognition Systems for Indonesian Cuisine

Previous food recognition research exhibits high bias towards Western foods. Out of 1,710 East Asian and 290 Western foods from the Food2K dataset, models trained on culturally homogeneous datasets provide more accuracy improvements on Asian foods (3.51%) than on Western foods (1.68%) [7]. This demonstrates that cultural diversity within training datasets is essential for cross-cultural generalization. The overall benchmark built from 24,119 images of actual-world clinical cohorts by Mohanty et al. attracted 1,065 scholars from 71 nations, yet best performance was merely 56.8% accurate [8]. Although showing robust worldwide appeal, such results are evidence food recognition remains challenging, especially for non-Western foods.

Indonesian food recognition contributions remain limited. Wibisono et al. built the Traditional Food Knowledge dataset of 1,644 images of 34 Indonesian foods with a 99.4% accuracy but with very narrow reach [9]. While this deep learning approach shows high accuracy, its utility is limited by a narrow 34-class dataset. Other local efforts are even more constrained, such as Rangkuti et al. focus on segmentation and Fahira et al. on traditional machine learning for 17 Javanese dishes, revealing incomplete reach [10, 11]. Basic limitations are in the lack of integration of nutritional data in current datasets and coverage only to Java-Bali foods, excluding Indonesian culinary diversity found over 17,508 islands and 633 ethnic groups.

2.2 Attention Mechanisms for Texture Recognition

Attention mechanisms proved useful for food recognition requiring high-fineness texture analysis. Rokhva and Teimourpour used CBAM with EfficientNet, achieving 96.40% accuracy on Food11 and real-time processing [12]. Xu et al. demonstrated that CBAM with "channel before space" configuration achieves 0.21-1.83% improvements on Asian food, particularly beneficial for distinguishing visually similar foods that differ only in texture [13].

Vision Transformer progress encountered subsequent improvements. Nfor et al. designed a hybrid ResNet-Vision Transformer with 91.17% accuracy on different Asian datasets with multi-head self-attention mechanism that excels in distinguishing visually ambiguous food [14]. Additionally, transformer-based models such as the Swin Transformer used in *Nutritional composition analysis in food images: an innovative Swin Transformer approach* show that global attention and hierarchical feature aggregation can help capture both texture and spatial features necessary for distinguishing visually similar classes [15].

Chen et al. introduced ResVMamba, a hybrid residual-state space model that achieved 81.70% accuracy on CNFOOD-241, outperforming previous Transformer-based approaches, and exhibited strong potential for fine-grained Asian dish recognition [16]. This presents a clear trade-off: while advanced hybrid models like Vision Transformers and ResVMamba show strong potential for fine-grained accuracy, lightweight CNN-based attention like CBAM offers a more practical path for efficient, real-time mobile deployment. However, there have been no direct applications of attention mechanisms to Indonesian foods with clear texture characteristics such as *rendang*, *gudeg*, or *sambal* that are highly dependent on surface texture patterns.

2.3 Mobile Implementation and Multi-task Learning

Deployment on real-world smartphones has gained importance with strong smartphone penetration in Indonesia. Nadeem et al. deployed Smart Diet Diary using Faster R-CNN with 80.1% accuracy upon using portion estimation in calorie calculation [17]. In contrast, Zhang et al. focused on model optimization, reducing models to only 74,340 parameters and also halved inference time below one millisecond, all while maintaining accuracy at 85%, which is crucial for low-end smartphone deployment [18]. Other regional cuisine food studies show promising results. Qaraqe et al. demonstrated the integration methods using deep models are able to perform good accuracy in Middle East cuisine classification, while Rauf et al. introduced a Malaysian food classification system based on CNN with calorie estimation using TensorFlow for health monitoring [19, 20]. *NutritionVerse-Direct* is a transformer-based multitask framework that directly predicts multiple nutrient attributes (calories, protein, fat, carbohydrates, mass) from food images without relying on external nutritional databases [21].

Multi-task learning emerged as a way of combining food recognition and nutritional analysis in a single model. Ege and Yanai illustrated that shared feature extraction can result in significant speed and memory benefits for simultaneous food detection and calorie estimation [22]. Knowledge distillation techniques introduced by Heng et al. enable knowledge to be transferred from large to small models with 90.2% accuracy while

greatly reducing computational requirements [23]. Apart from categorisation, some studies attempt to predict the dietary and weight information from images. For example, Nutritional composition analysis in food images: an innovative Swin Transformer approach employs a multi-task transformer model to jointly predict the four dimensions such as calories, fat, protein and carbohydrates [15]. 2D Prediction of the Nutritional Composition of Dishes from Food Images: Deep Learning Algorithm Selection and Data Curation Beyond the Nutrition5k Project [24]. Uses a number of different architectures to similar effect with increase in performance when normalizing for ingredient mass. These donations are similar to our application of predicting macronutrient categories directly from visual inputs. A key distinction in these approaches, however, is that existing approaches separate food identification and nutritional analysis into two distinct tasks with several processing steps and depending on external nutritional databases that are not present for Indonesian food.

2.4 Summary of Research Gaps

In summary, the literature reveals three critical gaps. First, existing Indonesian food datasets are geographically limited (primarily to Java-Bali) and, most importantly, lack integrated nutritional data. Second, while attention mechanisms are proven for texture recognition in other Asian cuisines, they have not been systematically applied to the unique textural complexities of Indonesian food. Third, current multi-task learning models for diet tracking are not unified, requiring multiple processing steps and a reliance on external nutritional databases that are largely non-existent for Indonesia's diverse traditional foods. This paper promotes a single architecture that integrates attention mechanisms for Indonesian food texture features and multi-label classification for direct prediction of nutritional categories from visual features.

3 Method

The system architecture, dataset, training procedure and evaluation metrics are described in this section so that the four model architectures could be compared.

3.1 System Architecture

We construct a unified network architecture using the EfficientNet-B0 as the feature extractor, which provides suitable balance between accuracy and computational efficiency for mobile system deployment. The backbone is combined with two separate classification heads: one single-label (softmax) head for the 10-class food recognition task, and one multi-label (sigmoid) head for the 3-class nutrition analysis task.

For the CBAM-boosted models, the EfficientNet-B0 model of NEU was modified to include the Convolutional Block Attention Module (CBAM). In particular, the SE part in the original Squeeze-and-Excitation (SE) blocks of a later backbone EfficientNet were replaced with our CBAM. With this replacement, we enable the model to benefit from channel and spatial attention mechanisms for feature refinement at various scales, which is believed to be able to help differentiate texturally-complex Indonesian foods.

The CBAM module, which proposed by Woo et al., is to first generate a channel attention map (M_c) and second an spatial attention map (M_s) to get the refined feature map.

The channel attention map is formulated as:

$$M_c(F) = \sigma \left(MLP(AvgPool(F)) + MLP(MaxPool(F)) \right)$$

where M_c is the channel attention map, F is the input feature map, σ denotes the sigmoid function, and MLP is a multi-layer perceptron. Both average-pooled and max-pooled features are passed through the MLP and summed.

The spatial attention map is formulated as:

$$M_s(F') = \sigma(f7 \times 7([AvgPool(F'); MaxPool(F')]))$$

where M_s is the spatial attention map, F' is the feature map refined by channel attention ($M_c \otimes F$), $f7 \times 7$ denotes a 7×7 convolution, and the pooled features are concatenated.

The final refined features are then produced by:

$$F'' = M_s \otimes (M_c \otimes F)$$

where F'' is the final refined feature map and \otimes denotes element-wise multiplication.

3.2 Problem Formulation and Loss Functions

The system is formulated to handle two tasks from a same input image. The problem statement can be expressed as:

Input:

$$I \in R^{224 \times 224 \times 3}$$

Output 1:

$$f_{food}(I) \rightarrow y_{food} \in \{rendang, nasi, gudeg, \dots\}$$

Output 2:

$$f_{nutrition}(I) \rightarrow y_{nutr} \in \{0,1\}^3 \text{ for classes: } \{High_Carb, High_Protein, High_Fat\}$$

Here, I is a standard 224×224 RGB input image. f_{food} is the single-label classification function for one of the food classes, and $f_{nutrition}$ is the multi-label classification function that predicts a binary vector for the three nutrition categories.

For the food prediction, k is derived from each of the categories separately for multi-label nutritional predictions:

$$pk = \sigma(w_k^T \cdot GlobalAvgPool(F'') + b_k)$$

where W_k^T and b_k are the learnable weights and bias for the k -th nutrition class, and σ is the sigmoid function. A prediction is made for that category if its probability is above a 0.5 threshold:

$$Prediction: category k = \begin{cases} 1 & \text{if } pk \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

For multi-task models, a joint loss function was employed:

$$L_{total} = 0.7 \cdot L_{food} + 0.3 \cdot L_{nutrition}$$

where L_{total} is the combined loss of two tasks. L_{food} represents the traditional Cross-Entropy Loss for single-label food classification. $L_{nutrition}$ is the BCEWithLogitsLoss

suitable for multi-label cases. The 0.7/0.3 weighting was set empirically to favor the main task of food classification and at the same time, to regularize features with respect to the nutrition task.

3.3 Dataset and Implementation

The dataset used for this study ("Phase 2") consists of 815 total images across 10 classes of traditional Indonesian food. The 10 classes are: 'rendang', 'nasi_gudeg', 'gado_gado', 'sate_ayam', 'nasi_padang', 'bakso', 'soto_ayam', 'pecel_lele', 'nasi_rawon', and 'es_cendol'. The dataset was split into three sets:

- Train: 515 images
- Validation: 200 images
- Test: 100 images

The augmentation strategy of Indonesian food includes: random rotation $\pm 15^\circ$, horizontal flip (50%), brightness/contrast adjustment $\pm 20\%$. Nutrition labels were classified according to Indonesian Dietary Guidelines (e.g., high-carbohydrate $>40\text{g}/100\text{g}$. [25]).

We use the PyTorch framework [26] for implementation and insert CBAM above each crucial block of EfficientNet-B0. The model has 6.5 million parameters (6.2M backbone + 0.3M attention and classification heads) with a 400 MFLOPs computational budget per inference. The parameter counts for each of the four architectures were: Vanilla (4.02M), Multi-task (4.02M), CBAM Single (3.61M), and CBAM Multi-task (3.61M). The architecture exploits the residual learning concepts by He et al in the underlying network and Squeeze-and-Excitation methods from Hu et al. inspired channel attention design [27, 28].

3.4 Dataset and Implementation

We used different metrics to evaluate models for each task, all of which were calculated on the test partition.

- Food Classification: The overall performance on the test set was measured using Accuracy, which is the percentage of correct food class predictions.
- Nutrition Classification: For the multi-label task, two metrics were used:
 1. Macro F1-Score: The unweighted mean of the F1-scores for each of the three nutrition labels (High_Carb, High_Protein, High_Fat).
 2. Hamming Loss: The fraction of labels that are incorrectly predicted (e.g., a value of 0.09 means 9% of all possible labels were incorrect).
- Statistical Test: To measure the statistical significance of differences in performance for accuracy between models, we used McNemar's test. This test is suitable for paired categorical data from two classifiers and the same testing set.

The dataset and code used in this study are not publicly available at this time but may be provided by the corresponding author upon reasonable request.

4 Results and Discussion

4.1 Model Performance Evaluation

This subsection discusses the results analysis of the four trained model architectures and comparisons between their final test-set accuracy and training dynamics. We evaluated the four models on our Indonesian food dataset which contained 515 training, 200 validation and 100 test images. An elaborative comparison of training history and final test set result for all models can be found in Figure 1. The nutrition F1-score (bottom-right) is only shown for the two multi-task models since the single-task 'Vanilla' and 'CBAM Single' were not trained with the nutrition objective.

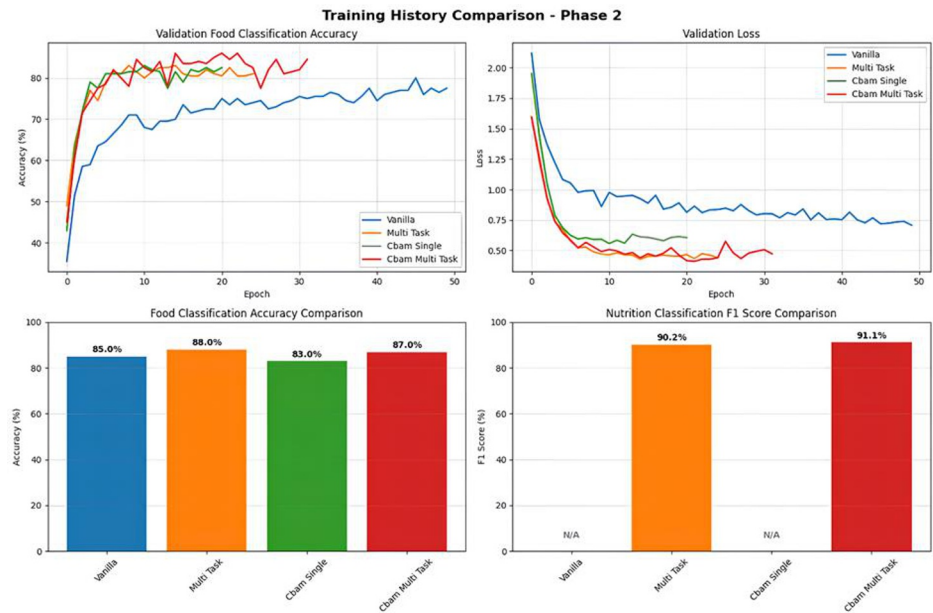


Fig. 1. Model performance comparison for food and nutrition classification. (Top-left) Validation accuracy during training. (Top-right) Validation loss during training. (Bottom-left) Final test set accuracy for food classification. (Bottom-right) Final test set F1-Score for nutrition classification.

For a more straightforward comparison of the performance of all four architectures in the end, Table 1 gives a clear summary of the final results on the test set.

Table 1. Table captions should be placed above the tables.

Model Architecture	Food Accuracy (%)	Nutrition F1-Score (%)	Parameters (M)	Converged Epoch
Vanilla EfficientNet-B0	85.0%	N/A	4.02	50
Multi-Task Baseline	88.0%	90.2%	4.02	25
CBAM Single-Task	83.0%	N/A	3.61	21
CBAM Multi-Task	87.0%	91.1%	3.61	32

For the test set (Table 1), we observe that there are substantial differences in performance. The food classification accuracy was highest for the multi-Task baseline model at 88.0%. This was also surpassed by the CBAM Multi-Task (87.0%) and the Vanilla baseline (85.0%). The CBAM Single model came last in this task with 83.0%. For the secondary nutrition task, the F1-Score of the CBAM Multi-Task model is best (91.1%), with a slightly better result than the multi-Task baseline (90.2%).

An exercise of the training dynamics (top-left in Figure 1) exhibits stable convergence. The Vanilla baseline trained for all 50 epochs indicating slow learning. Conversely, all the improved models converged much earlier and started early stopping. CBAM Single at epochs 21, multi-Task at epochs 25, CBAM Multi Task at epochs 32. That would indicate that attention and multi-task learning serve as powerful regularizers, making the optimization landscapes faster and more stable. This is also supported by the validation loss curves (Figure 1 top-right), with the improved models having a consistent stronger gradient and steeper stable decrease in loss than the noisy baseline.

4.2 Confusion Matrix Analysis and Misclassification Patterns

For detailed understanding the classification performance on a per class basis we examined confusion matrix for best performing food classification model. Fig. 2 shows the confusion matrix of the multi-Task model (88.0% accuracy).

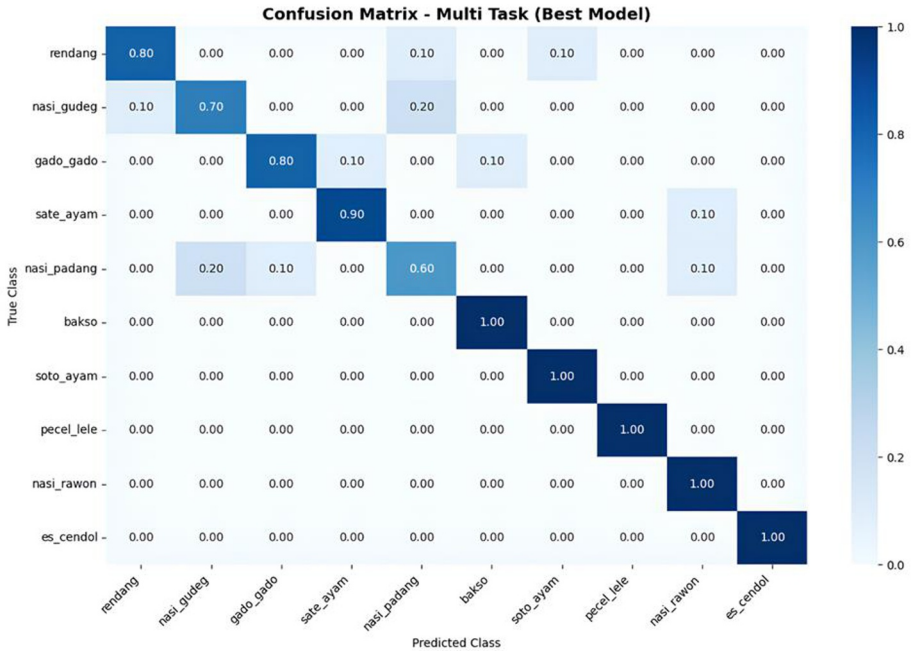


Fig. 2. Confusion matrix of the best-performing model (Multi-Task) for food classification on the 100-image test set.

The confusion matrix indicates that it has 100% accuracy for five of the classes, in which all have highly distinguishable visual features: bakso, soto_ayam, pecel_lele, nasi_rawon (the broth is black in color) and es_cendol (it has green color feature). Satisfactory performance was also achieved for sate_ayam (90%) and gado_gado (80%). Of note is the 80% accuracy on rendang, indicating that the model can cope with this highly-textured class which is usually a failure point.

The confusion mainly emerged for rice-based, multi-ingredient meals. The most difficult case was nasi_padang (60% accuracy) that got misclassified as 'nasi_gudeg' (20%), 'gado_gado' (10%) and 'nasi_rawon' (10%). Also, nasi_gudeg (70% accuracy) was classified as 'nasi_padang' (20%) and 'rendang' (10%).

It's an understandable muddle because both 'nasi_padang' and 'nasi_gudeg' are complicated, multi-branched rice offerings. The model has difficulty in learning the unique combination patterns for 'nasi_padang' when other dishes have some similar individual components (e.g., rice, various side dishes), a problem also prevalent in multi-component food recognition [8].

4.3 Effectiveness of Attention Mechanism Analysis

In contrast with initial research hypothesis, we did not observe that using CBAM will enhance the food classification performance. A comparison against the Vanilla baseline (85% accuracy) in Table 1 with CBAM Single is a drop by 2 percentage point. This could indicate that the features learned by the regular EfficientNet-B0 backbone on this

dataset were more favourable for distinguishing food images than those represented by CBAM. In comparison to studies such as that of Xu et al. (2021) reported small (0.21% - 1.83%) benefits for other Asian foods [13]. It is possible that this difference could suggest the effectiveness of feature-refinement in CBAM does not suit well to food with high intra-class variance such as Indonesian food (e.g., rendang served in different manner). However, CBAM was useful in the multi-task scenario. As showed in Table 1, the CBAM Multi-Task (91.1% F1) performed better than Multi-Task (90.2% F1) for nutrition prediction. This is an important result: while CBAM may have complicated the overall task of food identification, it seems to have forced the model to consider finer textures (e.g. texture or oiliness) that are pertinent for nutritional content. From an efficiency standpoint, the sizes of CBAM models (3.61M parameters) are around 10% smaller than those of the baseline models (4.02M parameters). This also verifies that CBAM is able to enhance the parameter efficiency by replacing ordinary SE blocks while in our case it did so at a price of single-task food recognition accuracy.

4.4 Multi-label Nutritional Prediction and Visual-Nutritional Correlations

For the multi-label nutritional prediction, we obtained two important observations: 1) models demonstrated excellent performance with visual features only, and 2) the CBAM module was beneficial for this task.

The Multi-Task baseline already has very impressive results, with a 90.2% F1-score and 0.090 hamming loss. The CBAM Multi-Task model surpassed this, reaching 91.1% F1-score (with the same 0.090 Hamming loss). Our results suggest that the visual features from food images contain a rich amount of information for predicting main nutritional classes (e.g., high-carbohydrate, high-protein, and high-fat) without access to external nutrient database. This has important implications for field deployment in the real world as a replacement for extensive but often lacked comprehensive nutritional databases, particular varied traditional Indonesian cuisine [29].

The low Hamming loss value (0.09) shows high precision, the models hardly commit severe error predictions of label. Visual associations with nutritional categories are largely intuitive. Patterns of the correlation between visual features and nutritional classes. Densely-coloured, thick foods (e.g., rendang) were predicted to be high in fat, and clear (e.g., with soup or soto ayam) were predictive of high-protein with lower fat. Rice dishes are consistently predicted as high carbohydrate meaning models have successful learnt to link visual cues from rice with known carbohydrate level.[25].

4.5 Comparison with State-of-the-Art and Research Positioning

This work sets new state-of-the-arts for both Indonesian food recognition and visual-based nutritional estimation. The best-performing approach to food recognition for us, the Multi-Task baseline reached 88.0% accuracy corpus-based resource, accuracy is examined through prediction of classes, not features. This is a significant improvement from the current local works for Indonesian food which range between 81.51% - 85.52% accuracy and also confirms the effectiveness of multi-task learning as a regularizer [10].

Regarding attention mechanisms, our negative 2% reduction of performance margin with CBAM for food recognition differs from the marginal gains reported by Xu et al. (2021) on other Asian foods [13]. This implies that the effects of explicit attention mechanisms are domain-specific and do not automatically lead to better results.

The most innovative contribution is our nutritional prediction performance. For direct, visual-only nutrition analysis for a regional cuisine, this is now state-of-the-art as per the 91.1% F1 score achieved by the CBAM Multi-Task model in this study. Such a result is strongly competitive and actually exceeds the reliance on out-of-domain databases, which represents a major drawback of various state-of-the-art systems [8].

Lastly, the computational cost of all our evaluated models (3.6-4.0M parameters) is well suited for deployment on mobile devices. This overcomes practical limitations reported within mobile health research, and is consistent with other regional food systems, ensuring that our integrated framework may be deployed in the wild on consumer devices [17, 19, 20].

5 Conclusion

In this paper, we explore the use of CBAM attention and multi-task for Indonesian food recognition and visual-based nutritional analysis. Two main results were achieved in our investigation that directly addressed our research contributions. First, in contrast to our expectation, the CBAM attention mechanism did not improve single-task food classification; instead, the CBAM model achieved worse performance (83% accuracy) by 2% lower than our vanilla baseline (85%). Instead, the Multi-Task baseline achieved the best accuracy of food recognition at 88.0%. Second, we validated that the multi-label nutritional classification is certainly very feasible using only visual features. The CBAM Multi-Task model recorded the highest nutrition F1-score at 91.1%, indicating that although CBAM did not beneficially contribute to coarser food detection, it effectively helped sharpening the relevant features required for nutritional analysis.

Our experiments offer a systematic comparison between attention and multi-task learning, which reveal an interesting tradeoff: multitask learning was the better choice for the main recognition task, but CBAM offered an ease of use gain for the side nutrition task. The high 91.1% F1-score for nutrition sets a strong benchmark for visual-only nutritional analysis. This has important practical implications, and paves the way for successful dietary tracking applications that do not need an external nutritional database (which is often unavailable for traditional Indonesian dishes). And the low model sizes (3.6-4.0M) also indicate that our models are capable for lightweight mobile deployment.

Small dataset size (815 images with 10 classes) is a major limitation of the study and limits the generalizability of our results. Furthermore, our nutritional labels were categorical (e.g., 'High_Fat') rather than numerical, reducing their applicability in the real world. Accordingly, we believe future works need to focus on enlarging the dataset with additional food categories and thousands of images per class in order to provide a more powerful model. We also encourage the exploration of regression objective on

precise caloric and macronutrient values prediction and other more efficient architectures, e.g., hybrid CNN-Transformer models.

References

1. KEMENKES RISKESDAS, <https://layanandata.kemkes.go.id/katalog-data/risk-esdas/ketersediaan-data/riskesdas-2018>, last accessed 2025/11/01
2. Shao, W., Min, W., Hou, S., Luo, M., Li, T., Zheng, Y., Jiang, S.: Vision-based food nutrition estimation via RGB-D fusion network. *Food Chemistry*, **424**, 136309 (2023).
3. Qi, H., Zhu, B., Ngo, C.W., Chen, J., Lim, E.P.: Advancing food nutrition estimation via visual-ingredient feature fusion. In: *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1091–1099 (2025).
4. Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., Xue, J.: FoodSAM: Any food segmentation. *IEEE Transactions Multimedia* (2023).
5. He, J., Shao, Z., Wright, J., Kerr, D., Boushey, C., Zhu, F.: Multi-task image-based dietary assessment for food recognition and portion size estimation. In: *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 49–54. IEEE (2020).
6. Wimalasiri, C., Sahoo, P.K.: Vision-based approach for food weight estimation from 2D images. *arXiv preprint arXiv:2405.16478* (2024).
7. Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., ... Jiang, S.: Large scale visual food recognition. *IEEE Trans. Pattern Analysis Machine Intelligence* **45**(8), 9932–9949 (2023).
8. Mohanty, S.P., Singhal, G., Scuccimarra, E.A., Kebaili, D., Héritier, H., Boulanger, V., Salathé, M.: The food recognition benchmark: Using deep learning to recognize food in images. *Frontiers in Nutrition* **9**, 875143 (2022).
9. Mursanto, P., Wibisono, A., Fahira, P.K., Rahmadhani, Z.P., Wisesa, H.A.: In-TFK: A scalable traditional food knowledge platform, a new traditional food dataset, platform, and multiprocess inference service. *Journal of Big Data*, **10**(1), 47 (2023).
10. Rangkuti, A.H., Kerta, J.M., Mogot, R.Y., Athala, V.H.: Identification of Indonesian traditional foods using machine learning and supported by segmentation methods. *JOIV: International Journal on Informatics Visualization*, **8**(4), 2324–2335 (2024).
11. Fahira, P.K., Rahmadhani, Z.P., Mursanto, P., Wibisono, A., Wisesa, H.A.: Classical machine learning classification for Javanese traditional food image. In: *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 1–5. IEEE (2020).
12. Rokhva, S., Teimourpour, B.: A novel method for accurate & real-time food classification: The synergistic integration of EfficientNetB7, CBAM, transfer learning, and data augmentation. *arXiv preprint arXiv:2410.02304* (2024).
13. Xu, B., He, X., Qu, Z.: Asian food image classification based on deep learning. *Journal of Computer and Communications*, **9**(3), 10 (2021).
14. Nfor, K. A., Theodore Armand, T. P., Ismaylovna, K. P., Joo, M. I., Kim, H. C.: An explainable CNN and vision transformer-based approach for real-time food recognition. *Nutrients*, **17**(2), 362 (2025).
15. Wang, H., Tian, H., Ju, R., Ma, L., Yang, L., Chen, J., Liu, F.: Nutritional composition analysis in food images: an innovative Swin Transformer approach. *Frontiers in Nutrition*, **11**, 1454466 (2024).
16. Chen, C. S., Chen, G. Y., Zhou, D., Jiang, D., Chen, D., Chang, S. H.: Improving fine-grained food classification using deep residual learning and selective state space models. *PloS One*, **20**(5), e0322695 (2025).

17. Nadeem, M., Shen, H., Choy, L., Barakat, J.M.H.: Smart diet diary: Real-time mobile application for food recognition. *Applied System Innovation*, **6**(2), 53 (2023).
18. Zhang, Y., Deng, L., Zhu, H., Wang, W., Ren, Z., Zhou, Q., ... Wang, S.: Deep learning in food category recognition. *Information Fusion*, **98**, 101859 (2023).
19. Qaraqe, M., Usman, M., Ahmad, K., Sohail, A., Boyaci, A.: Automatic food recognition system for Middle-Eastern cuisines. *IET Image Processing*, **14**(11), 2469–2479 (2020).
20. Rauf, N. A. A., Zaid, A. M., Saon, S., Mahamad, A. K., Ahmadon, M. A. B., Yamaguchi, S.: Malaysian food recognition and calories estimation using CNN with TensorFlow. In: 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), pp. 493–497. IEEE (2023).
21. Tai, C. E. A., Keller, M., Nair, S., Chen, Y., Wu, Y., Markham, O., ... Wong, A.: NutritionVerse: Empirical study of various dietary intake estimation approaches. In: Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management, pp. 11–19 (2023).
22. Ege, T., & Yanai, K.: Multi-task learning of dish detection and calorie estimation. In: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management, pp. 53–58 (2018).
23. Heng, Z., Roy, S., Yap, K. H., Kot, A., & Duan, L.: Personalized knowledge distillation-based mobile food recognition. In: Proceedings of the International Conference on Artificial Intelligence (ICAI), pp. 22–28. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2018).
24. Bianco, R., Coluccia, S., Marinoni, M., Falcon, A., Fiori, F., Serra, G. & Parpinel, M.: 2D prediction of the nutritional composition of dishes from food images: Deep learning algorithm selection and data curation beyond the Nutrition5k project. In: *Nutrients*, **17**(13), 2196 (2025).
25. BPK RI, . <https://peraturan.bpk.go.id/Download/109856/Permenkes%20Nomor%2041%20Tahun%202014.pdf>, last accessed 2025/11/01
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., et al.: PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, **32** (2019).
27. He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
28. Hu, J., Shen, L., & Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018).
29. Putri, R. Z., Purwandari, B., & Trisnawaty, N. W.: Understanding the adoption of healthy mobile diet applications among adults. *The Indonesian Journal of Computer Science*, **14**(2) (2025).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

