Open Access Library Journal
2025, Volume 12, e12846
ISSN Online: 2333-9721
ISSN Print: 2333-9705

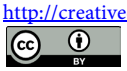
Predicting Categories and Ingredients of Traditional Dishes Using Deep Learning and Cross-Attention Mechanism

Ima Sokolo, Chidiebere Ugwu, Friday E. Onuodu

Department of Computer Science, University of Port Harcourt, Port Harcourt, Nigeria
Email: i.sokolo@yahoo.com, chidiebere.ugwu@uniport.edu.ng, friday.onuodu@uniport.edu.ng

How to cite this paper: Sokolo, I., Ugwu, C. and Onuodu, F.E. (2025) Predicting Categories and Ingredients of Traditional Dishes Using Deep Learning and Cross-Attention Mechanism. *Open Access Library Journal*, **12**: e12846.
https://doi.org/10.4236/oalib.1112846

Received: December 21, 2024
Accepted: March 16, 2025
Published: March 19, 2025



Abstract

The food recognition systems are available for foreign dishes but much work has not been done for our traditional dishes making it difficult to classify the traditional dishes and the ingredients they are made up of. From extensive literature reviews conducted, the existing models on food recognition are not robust enough to handle classification and identification of ingredients in traditional dishes. This study developed an improved food recognition system for the classification and identification of ingredients in traditional dishes. The food image dataset used to build the model was gotten from Kaggle, which was not standardized. It was preprocessed and standardized for consistency across datasets in eighteen different classes for model building. The standardized dataset was split into two; 80% for training and 20% for testing, convolutional neural network and cross attention mechanism were used to build the model. The cross-attention mechanism was used to selectively pick features across the multiple classes in the food dataset. ReLU was used as activation function and Adam optimizer was used as optimization function in building the model. The object oriented analysis methodology was used in the design, while python programing language was used in the development of the system. The result obtained shows an accuracy of 93.57% for training and 90.0% for validation and error loss of 0.062% and 0.001% respectively and interestingly, during testing the model gave 99% accuracies to traditional food images inputted on it. The results from application were able to detect and classify traditional dishes into different classes and outline the ingredients used to prepare them which shows tremendously performance of the system.

Subject Areas

Complex Network Models

Keywords

Traditional Dishes, Cross-Attention Mechanism, Convolutional Neural Network

1. Introduction

Image recognition and classification are becoming increasingly popular due to continuous and improved research in machine learning and deep learning models, with food images being no exception. These systems have become increasingly widespread with the development of online cooking platforms and mobile applications. Such systems aim to suggest recipes tailored to individual user preferences, dietary restrictions, and available ingredients. Traditional methods often rely on collaborative filtering, content-based filtering, or a combination of both. However, these methods can struggle with the complexities and degrees of food data. Food recognition technology offers the capability to automatically identify food items from images captured by a camera, image form, etc. making it easier for humans to discover the food types and ingredients in traditional food images [1]. Traditional food image recognition has become a favorable computer vision application [2] [3]. However, the diverse nature of traditional dishes makes predicting or classifying these food images challenging due to the many similarities and diversities between different types of foods, which can be difficult for humans to recognize correctly. Early recipe recommendation systems were primarily rule-based or employed simple collaborative filtering techniques. As the volume of available recipes grew, these systems faced challenges in scalability and accuracy. The introduction of machine learning techniques allowed for more sophisticated models that could better understand user preferences by analyzing past behaviors and recipe content [4] [5]. Deep learning, which is within the broader domain of machine learning, has transformed various disciplines, notably Natural Language Processing (NLP) and Computer Vision (CV) [6]. In the context of recipe recommendations, deep learning models can process and learn from large amounts of textual and visual data. Likewise, ingredients and presentation styles are identified from images of dishes utilizing Convolutional Neural Networks (CNN) [7], while Recurrent Neural Networks (RNNs) [8] and transformers [9] can understand the context and structure of recipe instructions.

Attention mechanisms, particularly in the form of transformers, have significantly improved the accuracy of deep learning models in NLP tasks [10]. Cross-attention extends this concept by allowing the model to concentrate on different features of input data simultaneously, making it particularly useful for tasks involving multiple data modalities. Specifically, when recommending a recipe, a model might need to consider both the textual description of a recipe and an image of the finished dish. Cross-attention can help the model align and integrate these different data sources effectively.

After an extensive literature review, we discovered that the existing models are not robust enough to detect and classify traditional dishes in addition to identifying their ingredients. There is limited research in the traditional dishes, making it difficult to classify traditional dishes and ingredients they are made up of. Whereas, food recognition systems are available for foreign dishes. Food image recognition contributes to health and nutrition monitoring by enabling users to track their dietary intake more accurately. This is particularly valuable for individuals with specific health goals, dietary restrictions, or medical conditions. This propelled us to restrict the dataset to traditional dishes so that the model will solve peculiar needs of our people based on the dishes that are common to them. Also, classifying traditional dishes preserves cultural heritage and enhances culinary knowledge. This paper seeks to develop a deep learning model that will have the capacity of identifying traditional dishes in an image form.

2. Related Works

The evolution of recipe analysis using machine learning has seen significant advancements over the past few years. One notable work in this area is the introduction of the RecipeQA dataset, designed for the multimodal comprehension of cooking recipes. This dataset facilitates visual question answering and recipe recommendation tasks, demonstrating the potential of combining textual and visual information for better recipe understanding. The paper "RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes" by [11] provides a comprehensive overview of this dataset and its applications. Additional crucial work is "Cross-modal Recipe Retrieval: How to Cook This Dish?" by [12], explores the learning of cross-modal embeddings for linking cooking recipes and food images. This study highlights the effectiveness of joint embedding models in processing and understanding recipe and image data, enabling tasks like recipe and image retrieval.

Ingredient recognition is a critical aspect of recipe analysis, and several studies have focused on this area using deep learning techniques. In the paper "Ingredient Recognition for Cooking Recipes Using Deep Learning" by [13] explored the automatic recognition of ingredients from textual recipes. They highlight the effectiveness of CNN and RNN in processing and understanding ingredient data. Furthermore, in their survey "Deep Learning for Food Ingredient Recognition and Recipe Inference" [14], adopted the use of deep learning models for recognizing ingredients in food images and inferring the corresponding recipes. Their work undermines deep learning potentials in enhancing the accuracy and efficiency of ingredient recognition as well as recipe recommendation systems. In addition, Attention Mechanisms have significantly enhanced the performance of various NLP and computer vision tasks. The foundational paper "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" by [15] has influenced many subsequent works, including those in the culinary domain. The concept of visual attention in image captioning has been introduced and is currently being adapted

for use in recipe analysis and recommendation. Additionally, [16] introduced the Recipe1M+, a large-scale dataset designed for learning cross-modal embeddings that link cooking recipes and food images, showcasing its potential to improve recipe and image retrieval, as well as recipe recommendation systems.

Merging textual and visual data for recipe analysis and recommendation has shown promising results. The paper "MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Cooking Recipe Retrieval and Exploration" by [17] proposed a model that fuses multimodal information using transformers. This approach enhances the retrieval and exploration of cooking recipes by leveraging the strengths of both text and image data. Building on these foundations, the hypothetical future work "Enhancing Recipe Recommendation with Deep Learning and Cross-Attention Mechanisms" would focus on integrating deep learning models with cross-attention mechanisms. This integration is aimed at improving multimodal data understanding and alignment, ensuing in more accurate and personalized recipe recommendations.

3. Material and Methods

The system architecture in **Figure 1** depicts the procedure of model building and application development which was strictly followed in developing the food recognition system. The dataset was downloaded from Kaggle repository, with other traditional food images crawled from food websites, amounting to 24,590 images divided into 18 categories or classes. To ensure that the classes are consistent and unambiguous, class definition was done to decide on the granularity of the classes. The classes are Abacha and Ugba (African salad), Akara and Eko, Amala and Gbegiri-Ewedu, Asaro, Boli(bole), Chin Chin, Egusi Soup, Ewa-Agoyin, Fried Plantains (dodo), Jollof Rice, Meat-pie, Moin-Moin, Nkwobi, Okro Soup, Pepper Soup, Puff-Puff, Suya and Vegetable Soup. Each of the classes contains various number of images that collectively formed the dataset that was used to build the model.

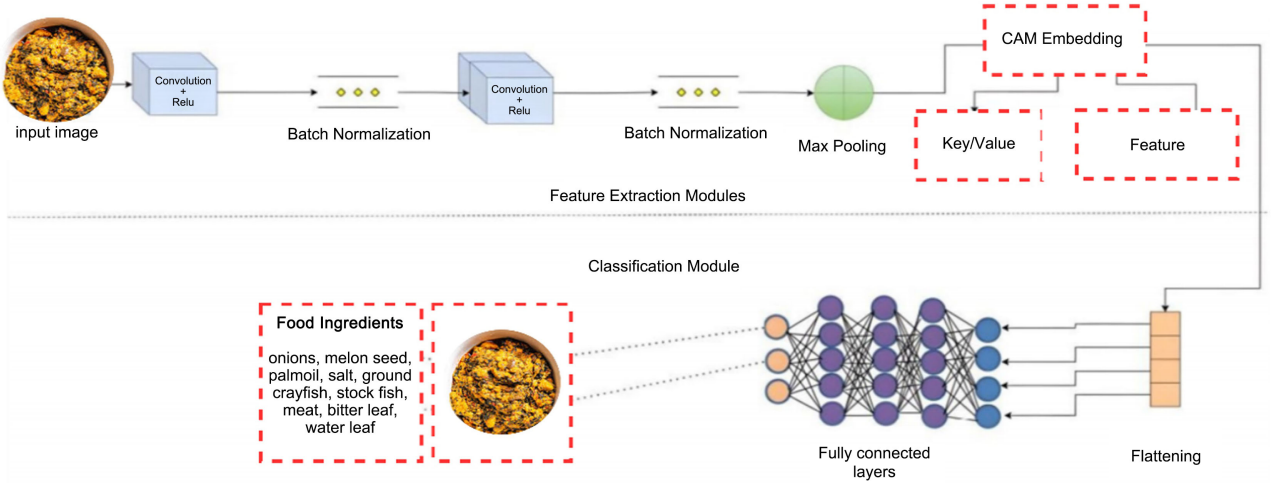


Figure 1. Architecture of the system.

We examined the dataset to identify any inconsistencies or anomalies. We noticed that some images were of low quality, exhibiting issues such as blurriness, poor lighting, or occlusions. See Figure 2 for sample food images of traditional dishes. These factors could potentially hinder the ability of the model in recognizing and classifying food items accurately if not properly handled at this stage. To address this, we filtered out the subpar images, retaining only those with clear and well-lit depictions of the food.

Figure 2. Sample images of traditional dishes.

3.1. Model Development

Several steps were initiated in the preprocessing stage to prepare the images for input into the convolutional neural network. First, all images were resized to uniform dimensions, in order to maintain consistency across the dataset. This was done using the random resizing approach, accomplished with cropping that helped in scaling images within defined range to randomly resizing images during training to increase the variety and robustness of the model. This was essential for maintaining a standardized input size, which facilitated the efficient training of the model. We selected a resolution that balanced detailed retention with computational efficiency, typically about (150, 150, 3) pixels. The robustness of the model was enhanced with rotation, flip, zoom, height and width using data augmentation techniques, which artificially increased the variability of the dataset. The pixel of the images values were normalized scaling the pixel intensities to a range of 0 to 1, which assisted in speeding up the convergence of the neural network (CNN) during training. Additionally, we ensured that the images were in the RGB colour space, as this is the standard input format for convolutional neural networks.

3.2. Training Stage

The data was split into two distinct set, the training and validation sets. This partitioning was done to evaluate performance of the model at different stages of development. In developing the model the training set was utilized, and the validation set was employed to optimize the model by fine-tuning and mitigate

overfitting. The test set was utilized in evaluating the accuracy and generalisation capabilities of the model. The CNN architecture, which was augmented with three dense layers, consisting of 1024, 512, and 18 units respectively is shown in Figure 3. EfficientNetB3 is a pre-trained model on large and diverse datasets and fine-tuned on food recognition dataset, which had already learned useful features assisted the model to generalize to new and unseen data by setting the base_model.trainable = False. The cross-attention was used to align and integrate these different data sources effectively in the model.

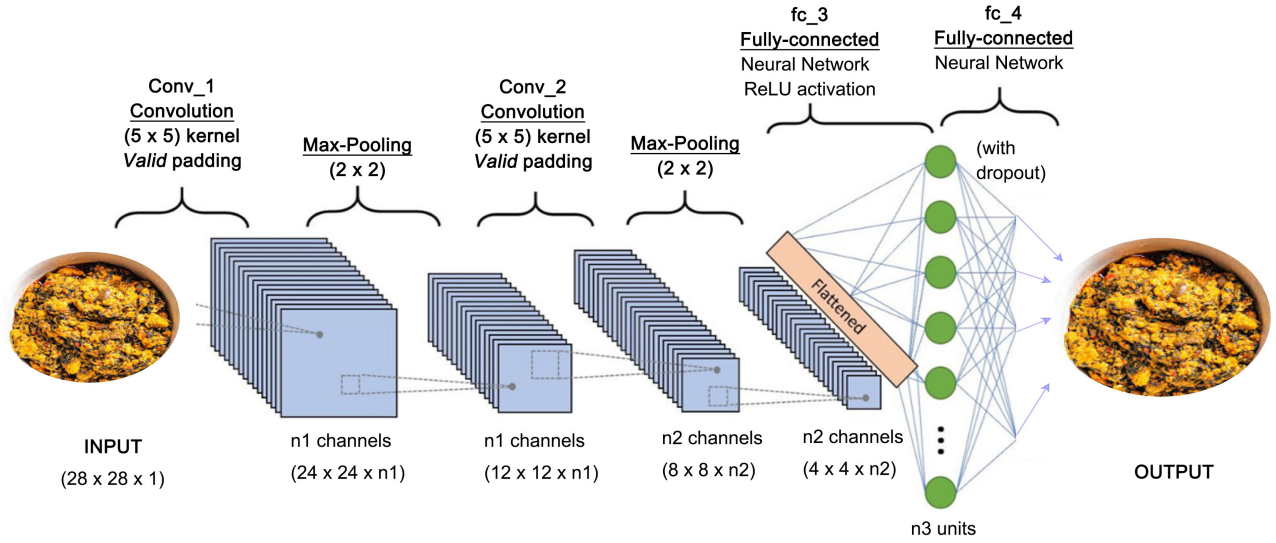


Figure 3. CNN Architecture of the system.

Convolution layer is the first layer in the CNN architecture, the kernel in the first layer was used to extract the features from the images by striding, the kernel was able to move through the entire images as shown in the equation, formula = $[i - k] + 1$ where i is Size of input, K is Size of kernel, the striding parameter was placed at 1 so that the images are picked one after the other as shown in equation, formula = $[i - k/s] + 1$ where i is Size of input, K is Size of kernel, S is Stride. At this point padding was used to prevent the loss of details of any image as a result of inability to capture features at the boarder. This is shown in the following equation, formula = $[i - k + 2p/s] + 1$ where i is Size of input, K is Size of kernel, S is Stride, p is Padding. The result was passed to the next layer after completion of the convolution operation in the input.

The Pooling Layer follows the Convolutional Layer. Primarily to decrease the size of the convolved feature map to reduce the computational costs. This is performed by decreasing the connections between layers and independently operates on each feature map. There are two types of pooling namely, average pooling and max pooling, for this paper Max Pooling was used, which took the largest element from feature map in a predined section. The total sum of the elements in the predefined section is computed in Sum Pooling. The Pooling Layer usually serves as a bridge between the Convolutional Layer and the Fully Connected Layer (FC). It

consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. These layers are usually placed before the output layer and form the last few layers of a CNN Architecture.

The output is flattened and inserted into the dense layers. The first dense layer, with 1024 units, acted as a significant transformation layer, capturing complex patterns and relationships within the extracted features. This was followed by a second dense layer with 512 units, which further distilled the information, reducing dimensionality while retaining critical data attributes. Finally, the third dense layer, containing 18 units, represented the output layer where each unit corresponds to a different food category.

There is possibility of overfitting occurring when all the features are connected to the FC layer, to overcome this problem, a dropout layer was used with a few neurons dropped from the neural network during training process resulting in reduced size of the model. A dropout of 0.3, 30% was passed and the nodes were dropped out randomly from the neural network.

The activation function used is ReLU, it has a derivative function and allows for backpropagation while simultaneously making it computationally efficient, it also accelerates the convergence of gradient descent towards the global minimum of the loss function due to its linear, non-saturating property. The main catch here is that the ReLU function does not activate all the neurons at the same time. The neurons will only be deactivated if the output of the linear transformation is less than 0. The formulae is $f(x) = \max(0, x)$. This is one of the most important parameters of the CNN model. They are used to learn and approximate any kind of continuous and complex relationship between variables of the network. In simple words, it decides which information of the model should fire in the forward direction and which ones should not at the end of the network.

For optimization Adam was used and batch processing was adopted to train the model in mini-batches to efficiently handle large datasets. The final dense layers culminated in an output layer, which varied depending on the exact application of the food image recognition system. Since the problem solved is a classification problem, we used softmax activation function to produce a probability distribution over different food categories, allowing the model to identify the most likely food item in the image.

3.3. Model Testing and Results

After training the CNN, we used a test dataset to verify its accuracy. The test dataset is a set of labelled images that were not being included in the training process. Each image is being fed to CNN, and the output is compared to the actual class label of the test image. The results from the model are listed as follows.

4. Discussions

Figure 4 depicts the model's accuracy progression over the epochs. The blue line represents the training accuracy, which improved consistently from an initial

value of 50% to 89% as the model learned to map the input features to their respective labels accurately. Conversely, the yellow line depicts the validation accuracy, which began at 90%, reflecting the model's strong initial generalization ability to unseen data. However, during the early training stages, the validation accuracy experienced a decline, reaching 81% at its lowest point. This decrease shows that the model temporarily overfitted the training data, thereafter prioritized learning specific patterns or noise present in the training set rather than generalized features applicable to the validation set. Subsequently, the validation accuracy recovered and stabilized at 84%, proving the model's ability to balance learning specific features while regaining its generalization capabilities.

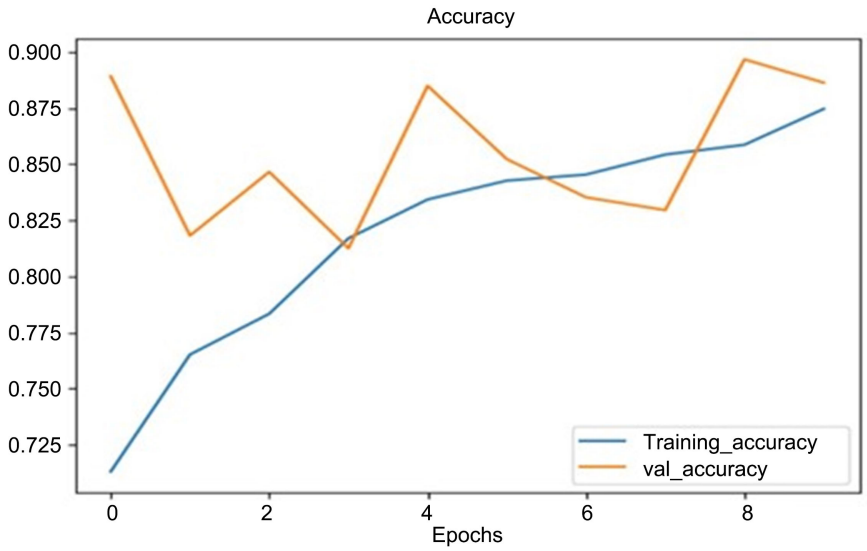


Figure 4. Accuracy of the model against the epoch.

Figure 5 is a plot of loss against the epoch, it is the different losses experienced at different epochs during the model building. **Figure 6** is the output of the test set, demonstrating the model's effectiveness on 20% of the dataset that was set aside specifically for testing purposes, it shows the percentage accuracies of the prediction and the images not predicted corrected. The images with red caption show that the predicted images are not seen as actual. The fourth image actual is Asaro (porridge Yam) but it was predicted as jollof rice. All the green captions show that predicted dishes are the same with the actual dishes. In such scenarios the model performed well with accuracies of almost 100%. **Figure 7** to **Figure 8** show the different classification results of the classifier application. After testing the model other food images that were used for training and testing were uploaded on the system. The system was able to classify them into one of the 18 classes identifying the ingredients they are made up of. The classifier was able to predict Okro Soup and the actual is Okro Soup with the ingredients listed, same with Moi Moi, Asaro (porridge Yam), Chin Chin, Jollof Rice. The classification ability of this model is very high so it can identify image of any traditional dishes and also list the content or ingredients used for the preparation.

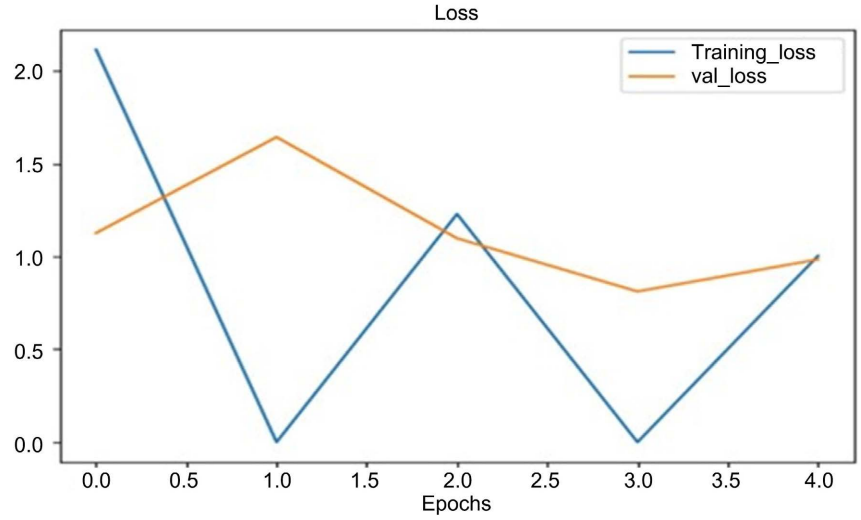


Figure 5. Loss against the epoch.

Figure 6. Classification results from testing.

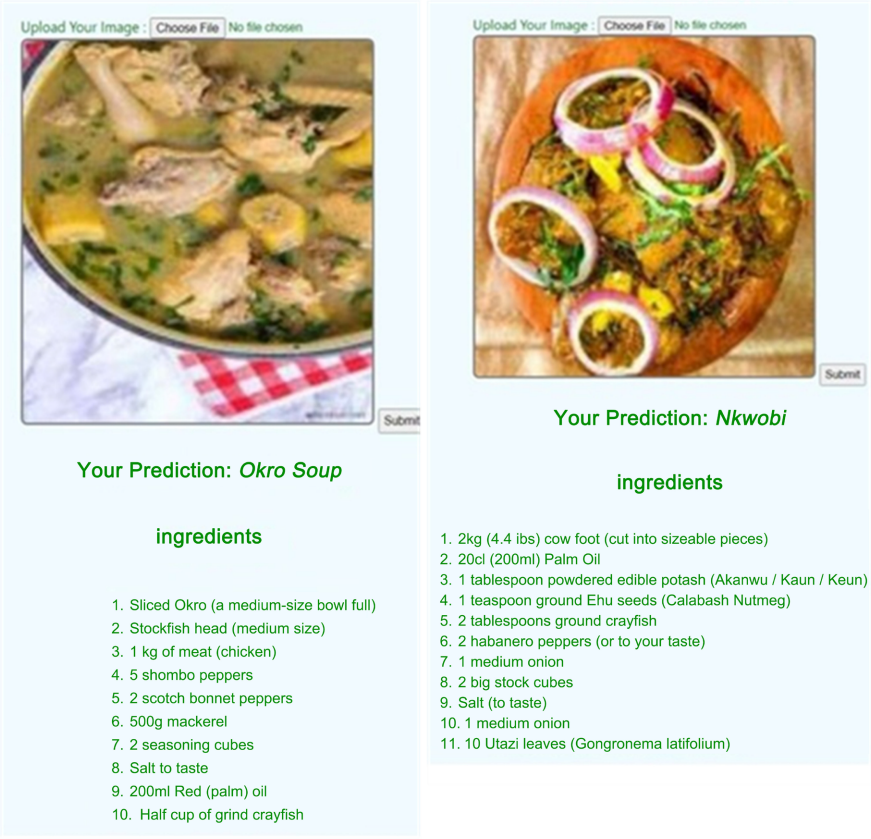


Figure 7. Classification results of Nkwobi and Okro Soup.

Figure 8. Classification results of Moi Moi and Yam Porridge.

5. Conclusion

This paper has successfully hybridized a convolutional neural network and cross-attention mechanism in the development of traditional food recognition model that has the ability of classifying food images and identifying the ingredients they are made up of. The model was tested with real food images captured with cameras and images crawled from the internet and it was able to classify them and list the ingredients.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Zhou, L., Zhang, C., Liu, F., Qiu, Z. and He, Y. (2019) Application of Deep Learning in Food: A Review. *Comprehensive Reviews in Food Science and Food Safety*, **18**, 1793-1811. https://doi.org/10.1111/1541-4337.12492

[2] Mezgec, S. and Koroušić Seljak, B. (2017) Nutrinet: A Deep Learning Food and Drink Image Recognition System for Dietary Assessment. *Nutrients*, **9**, Article 657. https://doi.org/10.3390/nu9070657

[3] Prajena, G., Harefa, J., Alexander, Josephus, B.O. and Nawir, A.H. (2022) Indonesian Traditional Food Image Recognition Using Convolutional Neural Network. 2022 *International Conference on Informatics*, *Multimedia*, *Cyber and Information System* (*ICIMCIS*), Jakarta, 16-17 November 2022, 142-147. https://doi.org/10.1109/icimcis56303.2022.10017684

[4] Khan, M.A., Rushe, E., Smyth, B. and Coyle, D. (2019) Personalized, Health-Aware Recipe Recommendation: An Ensemble Topic Modeling Based Approach. *CEUR Workshop Proceedings*, Copenhagen, 20 September 2019, 2439.

[5] Yera Toledo, R., Alzahrani, A.A. and Martinez, L. (2019) A Food Recommender System Considering Nutritional Information and User Preferences. *IEEE Access*, **7**, 96695-96711. https://doi.org/10.1109/access.2019.2929413

[6] Chai, J., Zeng, H., Li, A. and Ngai, E.W.T. (2021) Deep Learning in Computer Vision: A Critical Review of Emerging Techniques and Application Scenarios. *Machine Learning with Applications*, **6**, Article ID: 100134. https://doi.org/10.1016/j.mlwa.2021.100134

[7] Liu, Y., Pu, H. and Sun, D. (2021) Efficient Extraction of Deep Image Features Using Convolutional Neural Network (CNN) for Applications in Detecting and Analysing Complex Food Matrices. *Trends in Food Science & Technology*, **113**, 193-204. https://doi.org/10.1016/j.tifs.2021.04.042

[8] Wang, H., Lin, G., Hoi, S.C.H. and Miao, C. (2022) Learning Structural Representations for Recipe Generation and Food Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 3363-3377. https://doi.org/10.1109/tpami.2022.3181294

[9] Li, D. and Zaki, M.J. (2020) RECIPTOR: An Effective Pretrained Model for Recipe Representation Learning. *Proceedings of the* 26*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 6-10 July 2020, 1719-1727. https://doi.org/10.1145/3394486.3403223

[10] Hu, D. (2019) An Introductory Survey on Attention Mechanisms in NLP Problems.

In: Bi, Y., Bhatia, R. and Kapoor, S., Eds., *Intelligent Systems and Applications*, Springer, 432-448. https://doi.org/10.1007/978-3-030-29513-4_31

[11] Yagcioglu, S., Erdem, A., Erdem, E. and Ikizler-Cinbis, N. (2018) RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, 31 October-4 November 2018., Brussels, 1358-1368. https://doi.org/10.18653/v1/d18-1166

[12] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., *et al.* (2017) Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 3068-3076. https://doi.org/10.1109/cvpr.2017.327

[13] Chen, J., Zheng, Y., Jiang, Z. and Lin, Z. (2020) Ingredient Recognition for Cooking Recipes Using Deep Learning. *Pattern Recognition Letters*, **131**, 194-200.

[14] Min, W., Jiang, S., Liu, L., Rui, Y. and Jain, R. (2019) A Survey on Food Computing. *ACM Computing Surveys*, **52**, 1-36. https://doi.org/10.1145/3329168

[15] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Bengio, Y., *et al.* (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 2048-2057.

[16] Marin, J., Escalante, H.J., Hernández, C.A., Gonzalez, J.A., Lopez-Lopez, A., Sucar, L.E., Guyon, I., *et al.* (2019) Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 1352-1362.

[17] Wang, J., Zhou, M., Chen, Q. and Jiang, X. (2020) MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Cooking Recipe Retrieval and Exploration. arXiv: 2007.16113.