Vision Outlooker-based Hierarchical Food Classification

Pranav Kathar, Rajshree Khandare, Manisha Das, Deep Gupta,
Department of Electronics and Communication
Visvesvaraya National Institute of Technology Nagpur, India
katharbpranav@gmail.com, rajshreekhandare2810@gmail.com, das.manisha1989@gmail.com, deepgupta@ece.vnit.ac.in
Sneha Singh
School of Computing and Electrical Engineering
Indian Institute of Technology Mandi, India
sneha@iitmandi.ac.in

***Abstract*—In the modern world, where health concerns necessitate continual diet monitoring, the challenge of food image identification is crucial. Many machine learning models are available to automate the identification procedure. This is done predominantly with Convolutional Neural Networks (CNN) that help extract features for food images with different textures. But this comes with certain limitations such as diversity in food items, variation in the appearance of images, overfitting, and the inability to capture long-distance connections, which can result in inadequate feature representations. This paper attempts to explore Vision Transformers (ViTs) in an effort to overcome these limitations. ViTs are known for their attention mechanism, increased interpretability, better generalization, and robustness to adversarial cases. In this study, VOLO (Vision Outlooker for Visual Recognition), a contemporary vision transformer, improves learning by encoding fine-level information into the token representations. Also, a traditional flat classifier ceases to perform well because there are so many different cuisines and unique food items. Prediction systems with hierarchical classifiers were also developed to address this. Thus, the proposed method uses VOLO to accomplish hierarchical food classification. The experimental results support the proposed method's performance and contribution to an overall improvement in prediction accuracy.**

***Index Terms*—Vision outlooker, Food images, Convolutional Neural Network, Classification**

I. Introduction

Food Classification is a significant task with several applications in the fields of nutrition, healthcare, and the food industry. Creating individualized nutrition programs, keeping track of dietary intake, and detecting allergens all depend on being able to correctly recognize food items in images. Due to the variety in the appearance, shape, texture, and color of food products as well as the existence of occlusions, lighting, and camera angles, it can be difficult to identify food items from images. This can be seen from Fig 1 which shows food samples from the public food dataset MAFood121 [1]. Moreover, food Identification systems need to be highly accurate and efficient while handling complex and huge datasets. As a result, creating reliable and accurate food identification systems is a difficult problem that calls for innovative solutions [2].

In order to handle image input and perform classification tasks, convolutional neural networks (CNNs) have become a powerful technique. Convolutional filters are used in these networks to extract features from images by identifying patterns and edges in the image data. CNNs can learn

Fig. 1. Few sample images from MAFood121 dataset

increasingly more complicated representations of image features by stacking multiple convolutional layers and pooling layers. The spatial relationships between pixels in the image data may also be handled by CNNs, which is useful for identifying objects and patterns in images. However, they have disadvantages when working with huge datasets or complex foods with considerable intra-class variability and visual complexity [3]. CNNs have demonstrated promising results in the context of food classification when obtaining high accuracy in classification tests.

While CNNs have been successful in extracting features from images for food classification, they have certain limitations, such as the inability to capture long-distance connections and inadequate feature representations. The proposed work aims to overcome the drawbacks of current methods for food image classification using machine learning models, focusing specifically on the constraints associated with CNNs. The interpretability of the learned representations in CNNs is still an active area of research and remains a limitation in understanding the decision-making process of these models. In contrast, ViTs are known for their attention mechanism, which allows them to better capture long-distance connections and encode fine-level information into token representations. This makes them well-suited for the task. The novelty of the presented work lies in the use of ViTs for hierarchical food classification. This allows the system to first classify the cuisine and then the specific dish within that cuisine, resulting in improved accuracy.

The rest of the paper consists of the following sections. Section 2 presents the work related to the proposed approach,

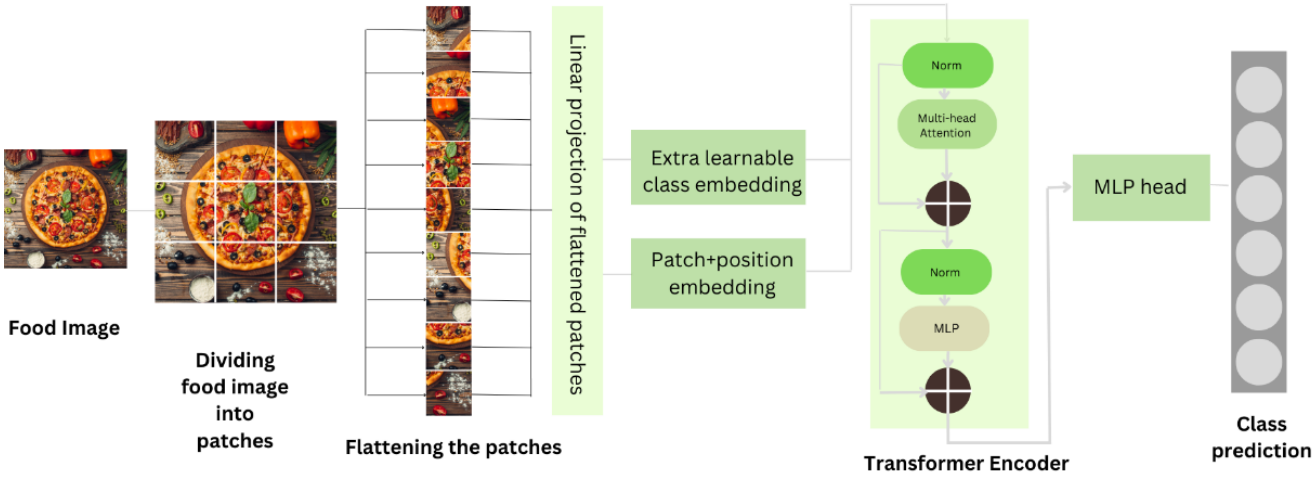

Fig. 2. Overview of Vision Transformer (ViT) model Architecture: An image is segmented into patches (here 9 by 9), after which position embedding is applied to the flattened patches, and then the patches are passed to a transformer encoder to make the final class.

Section 3 discusses the work with Methodology, and Section 4 and 5 presents the Experimental Details and the Results obtained respectively followed by the Conclusions in Section 6.

II. Related Works

Several research has already been conducted in diverse sectors such as healthcare, pattern recognition, classification, clustering, IoT technologies, etc based on the different neural models. This section examines the advances in food processing using neural networks, with a particular emphasis on the move from shallow to deep learning techniques. Collecting data and monitoring food processing processes has gotten easier because of industrial automation and the Internet of Things. ANNs have been critical in improving technology and attaining success in food grading, safety inspections, and quality assessment [4]. Accurate dietary assessment is critical for evaluating weight loss therapy. However, most current nutritional evaluation methodologies rely on recall, which has drawbacks. To address this issue, a modern computer-based food recognition system was created, allowing for reliable food evaluation on a variety of mobile devices and via Cloud services. The system tries to address the problem of detecting and distinguishing different types of food images. The great variety of food products, with both minor differences across categories and considerable variations within categories, complicates this work. The suggested technique involves the use of several fusion-trained classifiers which use features collected from various deep models and improves the system's ability to identify and recognize objects [5]. Deep learning is a powerful technique that is frequently used in image processing, speech recognition, and object detection. Author claims this study to be the first of its kind and investigates its recent application in the field of food science and engineering. The paper defines deep learning, covers prominent architectures and training approaches, and compiles a list of articles that use deep learning in food-related difficulties. Food recognition, calorie estimation, quality detection, food supply chain management, and contamination detection are among them. The publication delves into the specifics of each study, such as datasets, preprocessing methods, networks employed, performance, and comparisons to other methodologies [6].

There has been a surge in interest in using previously created ANNs to solve complicated real-world challenges in food processing. Deep learning techniques, on the other hand, have emerged as a substantial contributor to intelligent food processing. The study emphasizes the widespread application of deep learning, machine learning, and image processing in expanding the possibilities and growth potential in the field of food processing. Food recognition using Deep CNN integrated with the hand-crafted features is implemented in various tasks where they evaluated that network depth is an important aspect [7]. During training, CNN extracts feature and learns patterns in images which are important for classification and pattern recognition tasks. On successful training, the network can be used to classify new images using the learned parameters. The common architecture of CNN for this task is to use a series of pooling and convolutional layers to extract features from input images, then followed by fully connected layers to make the final classification. Convolutional layers are used to extract features from the input image filters, creating feature maps emphasizing different aspects of images and different levels of abstraction.

Pooling layers are used to reduce spatial dimensions of data allowing the network to focus more on abstract features. In a fully connected layer, it attempts to generate class scores from the activations. Extracted features from the previous layer are taken as input and further used for prediction. The output layer consists of a probability distribution over the classes [8], [9].

Despite their effectiveness, the limited ability of CNNs to capture long-range connections between image parts can lead to inadequate feature representations, which is one of their key problems. To solve this problem, ViTs, a new kind of deep learning architecture, has recently been proposed by Alexey Dosovitskiy et al. in 2020 [10]. ViTs employ a self-attention method that dispenses with explicit modeling of the spatial structure of the image and enables the network to capture long-range relationships between image patches. This makes it possible for ViTs to represent the connections more accurately between various aspects of the image, which is crucial for identifying complex food items with substantial intra-class variability. Moreover, ViTs is a possible substitute for CNNs for large-scale image recognition applications since they are extremely scalable and have a high computational efficiency while training on big datasets. ViTs has recently been modified for image classification tasks, and the results are encouraging. The basis of ViTs is the idea of self-attention, which enables the model to pay attention to various input sequence parts when generating its output. ViTs use a patch-based strategy to represent the input image as a series of fixed-size image patches when it comes to image categorization. The final prediction is created by running the patches through a sequence of self-attention and completely linked layers. In order to recognize complex patterns and objects in images, ViTs must be able to record long-range dependencies between image patches [11], [12]. A detailed Overview of Vision Transformer Architecture is provided in Fig. 2. Despite their impressive representational capacity, current ViT models experience inconsistency and inaccurate dense predictions at local regions. But it is observed that the effectiveness of their self-attention mechanism is constrained to shallower and thinner networks. Hence, a recently advanced ViT named VOLO is explored in the proposed study [13].

Nowadays, where health concerns demand continuous diet monitoring, accurate food classification becomes crucial. The task is compounded by the variety of food products with low-inter and large intra-class differences, as well as the restricted information in a single image [14], [15]

- Inadequate accuracy of existing food classification methods: Current food classification methods are not able to achieve the desired level of accuracy, especially when dealing with diverse food items and variations in appearance. This limitation hinders the effectiveness of the task.
- Demands of modern health-conscious society: In today's world, where people are increasingly concerned about their health and dietary choices, accurate food classification is of paramount importance. However, existing methods fall short of meeting the expectations and requirements of continuous diet monitoring.

The conventional models' effectiveness declines as the number of classes and the number of images inside each class increases. It suggests that models that perform effectively across a small number of classes cannot be expanded by simply scaling them for more complex applications. Therefore, this paper proposes a hierarchical classification approach for food classification along with VOLO. The primary objective of this study is to enhance the accuracy of food classification, particularly when dealing with diverse food items and variations in appearance, which pose significant limitations for existing models. To address this problem, the proposed study presents the utilization of ViTs and a hierarchical classification approach for food classification along with VOLO.

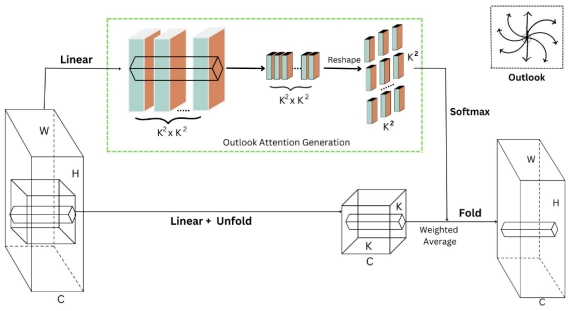

Fig. 3. Outlook Attention Mechanism illustrated with a window size of KxK using the center token, a linear layer followed by a reshape operation can easily construct the outlook attention Matrix

III. Methodology

In this section, the working of the proposed method using VOLO and its implementation details are discussed. Though ViTs have shown great potential, they have some limitations like low efficacy in encoding fine-level features. To solve this problem, a new architecture with outlook attention was introduced which was termed as VOLO. Different architectures of this model such as VOLO D1, VOLO D2, VOLO D3, VOLO D4, and VOLO D5 are presented with varying specifications as shown in Table I. VOLO-D5 is found to be the first model that exceeds 87% top-1 accuracy on ImageNet. The suggested outlooker creates a new focus for token aggregation and makes it possible for the model to effectively encode fine-level data. Outlook Attention, a layer that helps in focusing on a specific part of the input and represents the spatial relationship of the object in the image for further processing of information and generates attention weights while encoding fine-level information. A Multi-Layer Perceptron (MLP), is a feed-forward neural network used to model complex non-linear relationships of input and output. Let, Let $X \in R^{HXWXC}$,

$$\widetilde{X} = OutlookAtt(LN(X)) + X \tag{1}$$

$$Z = MLP(LN(\widetilde{X})) + \widetilde{X} \tag{2}$$

where LN is a Layer Norm, this is usually used in ViTs. Before the self-attention mechanism, in feed-forward layers, the LN is applied to activations. This improves the stability of the overall network during training and controls the scaling

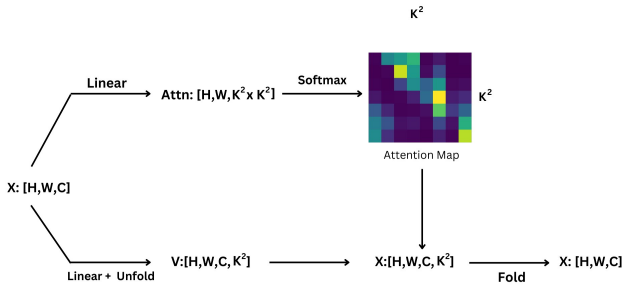

Fig. 4. Outlook attention flow with tensor shapes

of activations. Thus, resulting in improved performance and faster convergence on tasks.

Outlook attention calculates the resemblance of each spatial location (i, j) to all of its neighbors in a local window of size $K \times K$ centered at (i, j). Unlike self-attention, which must be calculated via a Query-Key matrix multiplication (i.e., Soft-max($Q^T K/\sqrt{d}$) outlook attention streamlines this process, by simply performing a reshaping operation as shown in Fig. 3. The weights from Outlook used for value aggregation like the attention weight by reshaping it to $\hat{A}_{i,j} \in R^{K^2 \times K^2}$, then a softmax function where the output of a preceding layer is converted into a vector of probabilities as shown in Fig. 4. With the given input vector and weight vector,

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3}$$

$$Y\triangle_{i,j} = MatMul(Softmax(\hat{A}_{i,j}), V\triangle_{i,j}) \tag{4}$$

After getting the projected values and adding up the weighted values from the different local windows at the same location yields the result. Fold and Unfold operations are performed followed by linear layer as in self-attention. Then multi-head outlook attention is implemented where N number of heads is taken into consideration, adjusting the weight shape. The value embeddings and outlook weight are uniformly split into N segments, which yield (A_n, V_n) for each such pair outlook attention is computed separately. Summation of which results in multi-head outlook attention. They have maintained an outlooker to transformer ratio of roughly 1:3, which has experimentally shown to be the most effective. To update the class embedding, also add two class attention layers. Outlookers have a hidden dimension that is set to half of the transformers. Fig. 5 depicts the precise workflow of the proposed method for flat dish classification, which is also applicable to cuisine classification. The outlook

TABLE I
MODEL SPECIFICATION DETAILS

Model	Dimension	Parameter count
VOLOD1	384	27
VOLOD2	512	59
VOLOD3	512	86
VOLOD4	768	193
VOLOD5	768	296

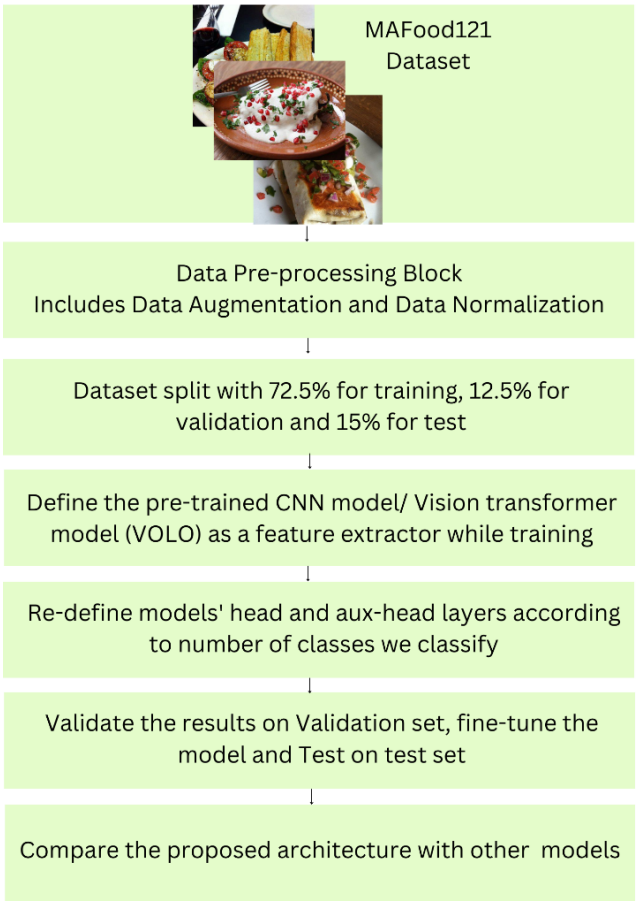

Fig. 5. Overview of proposed method

attention is simple, efficient, and easy to implement, and its main advantages are,

- The features present at each spatial location are representative enough to generate the attention weights for locally aggregating the neighboring features.
- Using dense and local spatial aggregation, it can encode fine-level information efficiently.

Hierarchical classification is employed while comparing it to a flat classification technique using pre-trained VOLO models with re-defined head and aux-head layers. The MAFood121 Dataset which we are working with consists of 11 cuisines and 11 dishes per cuisine. Fig. 6 and Fig. 7 show how hierarchical and flat classification will differ in the proposed scenario when dealing with cuisine and the corresponding dishes.

IV. EXPERIMENTAL DETAILS

A Multi-Attribute Food Dataset (MAFood121) with 11 most popular cuisines each containing 11 traditional dishes is used to train the classifiers. This dataset includes three different tasks: Dish, Cuisine, and Categories (food groups). There are a total of 21,175 images, of which 72.5% are used for training, 12.5% for validation, and 15% for testing. Dish and Cuisine support a single image-based value, whereas Categories support multiple label annotations. We performed data pre-processing which includes data augmentation and data normalization, details of which are provided in Table II. Then dataset is split into the train, test, and validation

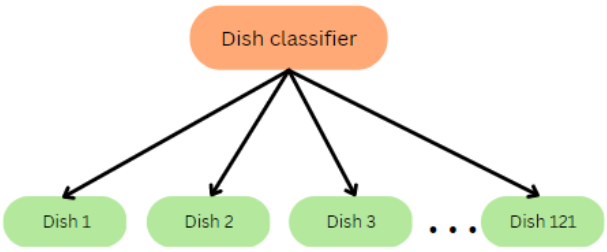

Fig. 6. Flat classifier

TABLE II
DATA AUGMENTATION DETAILS

Augmentation Technique	Value
Rotation	30
Horizontal Shear	0.2
Vertical Shear	0.2
Zoom	0.2
Horizontal flip	Yes

sets. We re-defined the model head and aux-head layers of VOLO's pre-trained models before using them for training. Then, we add the Adam optimizer to accelerate the gradient descent algorithm and Cross-Entropy-Loss is used to measure the performance of the classification model. The performance for VOLO D1 with the least (26.6M) parameters among all models and VOLO D5 with the highest (296M) of them are evaluated. Table I shows the parameters of VOLO models.

The relevance of layer activations at various 2-D spatial locations with respect to the objective task is represented by attention maps as a scalar matrix. Attention maps for VOLO D1 model can be seen in Fig.8. Then, its results were compared with pre-trained CNN models like EfficientNetV2, Inception-ResnetV2, and InceptionV3 on the same MAFood121 dataset. The hierarchical classification structure was trained on eleven cuisines - American, Japanese, Italian, Greek, Turkish, Chinese, Mexican, Indian, Thai, Vietnamese, and French with 11 dishes each in the respective cuisines, that is a total of 121 unique dishes. There are 13 classifiers in total 1 for flat classification, 1 for cuisine classification, and 11 for cuisine-specific dish classification. A learning rate of 1e-4 was used with a step decay of 0.5 applied after every 10 epochs. The model architecture used for the classifiers is the same as that used in the flat classification approach. The hyperparameters used are listed in Table III.

V. RESULTS AND DISCUSSIONS

The quantitative results obtained for the proposed method are presented in Table IV and Table V. Table V above summarizes the results obtained from evaluating the performance of various deep learning models on the MAFood-121 dataset. The models evaluated include EfficientNetV2, InceptionV3, Inception-ResNetV2, ViT-B-16, and VOLO D1 to D5. The overall accuracy of each model is reported in percentage, and the number of parameters for each model is also provided.

The results show that the VOLO D5 model achieved the highest overall accuracy of 84.71%, followed by VOLO D4 with an accuracy of 83.60%. Both VOLO D5 and D4 have significantly higher accuracy than the other models evaluated. The VOLO D3 model also performed well with an overall accuracy of 82.84%. Among the traditional CNN models, Inception-ResNetV2 achieved the highest accuracy of 80.16%, while EfficientNetV2 and InceptionV3 achieved 78.52% and 79.91% accuracy, respectively. Moreover, the performance of the proposed hierarchical classification method is also evaluated and the results are presented in Table IV. From the results mentioned in Table IV, it is observed that the proposed approach with a cuisine classifier achieved an accuracy of 84.63%, outperforming the flat classifier that achieved an accuracy of 82.95% for VOLO D5 backbone. Furthermore, it is also observed that the proposed approach with a dish classifier achieved the highest accuracy of 89.96%, highlighting the importance of hierarchical classification for food classification tasks.

The capacity of ViTs to efficiently capture long-range dependencies between picture components is one of the factors contributing to their high performance. ViTs, in contrast to traditional CNNs, process the entire image using the self-attention method, enabling them to gather global context data and dependencies between various parts of the image. Overall, the findings show that the suggested approach using VOLO models for hierarchical food categorization outperforms conventional CNN models, attaining excellent

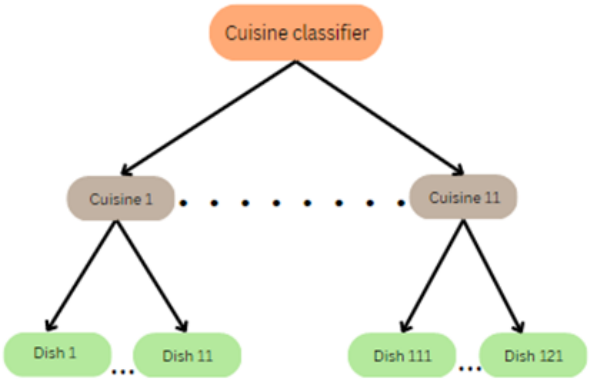

Fig. 7. Hierarchical classifier

TABLE III
HYPERPARAMETERS FOR EXPERIMENT

Parameters	Value
Batch size	16
Input Size Centre-crop	224x224
Initial Learning Rate	$1e^1$
Dropout	0.24
Optimizer	Adam

TABLE IV
MODELS TRAINED FOR HIERARCHICAL APPROACH DETAILS

Classifier	Backbone	Accuracy (%)
Flat classifier	VOLOD1	80.16
Cuisine classifier	VOLOD1	82.59
Dish classifier	VOLOD1	87.80
Flat classifier	VOLOD5	82.95
Cuisine classifier	VOLOD5	84.63
Dish classifier	VOLOD5	89.96

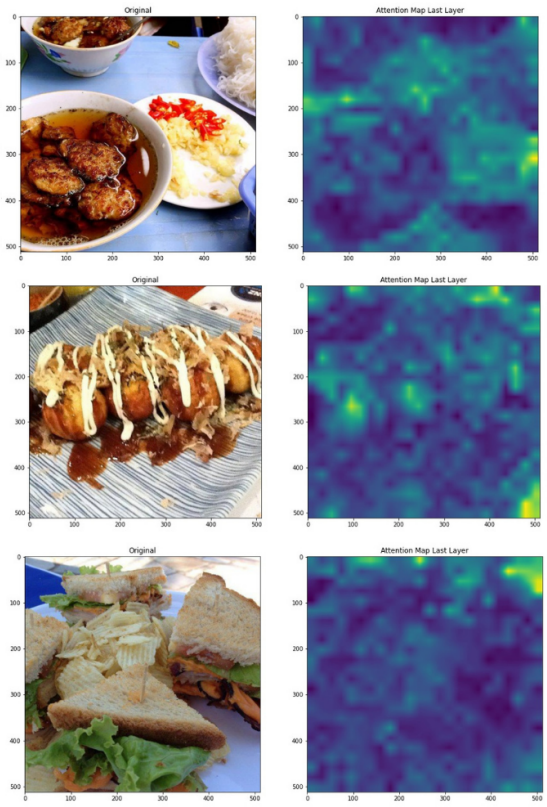

Fig. 8. Attention maps for food images (few samples) in MAFood121

TABLE V
PERFORMANCE COMPARISON OF VARIOUS MODELS ON MAFOOD121 DATASET

Model Architecture	Parameters count	Overall Accuracy (%)
EfficientNetV2	24	78.52
InceptionV3	22	79.91
Inception-ResNetV2	54	80.16
ViT – B -16	86	78.9
VOLO D1	27	81.16
VOLO D2	59	82.03
VOLO D3	86	82.84
VOLO D4	193	83.60
VOLO D5	296	84.71

accuracy even on a difficult dataset with a variety of food items. The potential of ViT in food image classification is shown by these results, which are encouraging.

VI. CONCLUSION

In contrast to convolutional neural networks (CNNs) having limitations when it comes to a variety of food items, variation in image appearance, overfitting, and capturing long-distance connections, this paper has explored the use of vision transformers (ViTs) for hierarchical food classification. VOLO, a modern vision transformer that enhances learning by embedding fine-level information into token representations, was used in the proposed study. The new outlook attention mechanism, which dynamically combines fine-level features in a dense manner, is what gives VOLO its performance advantage. The proposed method of using hierarchical classifiers successfully addressed the limitations of traditional flat classifiers, resulting in a general improvement in prediction accuracy. The proposed study does have some limitations, though. The hierarchical approach has issues with error propagation that needs to be addressed with uncertainty estimation techniques. Future studies could build on this by integrating a broader, more varied dataset because the dataset utilized was restricted to a specific number of foods and cuisines.

REFERENCES

[1] E. Aguilar, M. Bolaños, and P. Radeva, "Regularized uncertainty-based multi-task learning model for food analysis," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 360–370, 2019.

[2] W. Wang, W. Min, T. Li, X. Dong, H. Li, and S. Jiang, "A review on vision-based analysis for automatic dietary assessment," *Trends in Food Science & Technology*, 2022.

[3] A. A. Elngar, M. Arafa, A. Fathy, B. Moustafa, O. Mahmoud, M. Shaban, and N. Fawzy, "Image classification based on cnn: a survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, pp. 18–50, 2021.

[4] J. Nayak, K. Vakula, P. Dinesh, B. Naik, and D. Pelusi, "Intelligent food processing: Journey from artificial neural network to deep learning," *Computer Science Review*, vol. 38, p. 100297, 2020.

[5] N. O. Salim, S. R. Zeebaree, M. A. Sadeeq, A. Radie, H. M. Shukur, and Z. N. Rashid, "Study for food recognition system using deep learning," in *Journal of Physics: Conference Series*, vol. 1963, p. 012014, IOP Publishing, 2021.

[6] L. Zhou, C. Zhang, F. Liu, Z. Qiu, and Y. He, "Application of deep learning in food: a review," *Comprehensive reviews in food science and food safety*, vol. 18, no. 6, pp. 1793–1811, 2019.

[7] A. Fakhrou, J. Kunhoth, and S. Al Maadeed, "Smartphone-based food recognition system using multiple deep cnn models," *Multimedia Tools and Applications*, vol. 80, no. 21-23, pp. 33011–33032, 2021.

[8] Y. Liu, H. Pu, and D.-W. Sun, "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*, vol. 113, pp. 193–204, 2021.

[9] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 265–276, 2019.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229, Springer, 2020.

[12] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[13] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "Volo: Vision outlooker for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] E. Aguilar and P. Radeva, "Uncertainty-aware integration of local and flat classifiers for food recognition," *Pattern Recognition Letters*, vol. 136, pp. 237–243, 2020.

[15] V. Pillai, P. Mehar, M. Das, D. Gupta, and P. Radeva, "Integrated hierarchical and flat classifiers for food image classification using epistemic uncertainty," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, IEEE, 2022.