Muti-Stage Hierarchical Food Classification

Xinyue Pan
pan161@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Jiangpeng He
he416@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Fengqing Zhu
zhu0@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

Food image classification serves as a fundamental and critical step in image-based dietary assessment, facilitating nutrient intake analysis from captured food images. However, existing works in food classification predominantly focuses on predicting 'food types', which do not contain direct nutritional composition information. This limitation arises from the inherent discrepancies in nutrition databases, which are tasked with associating each 'food item' with its respective information. Therefore, in this work we aim to classify food items to align with nutrition database. To this end, we first introduce VFN-nutrient dataset by annotating each food image in VFN with a food item that includes nutritional composition information. Such annotation of food items, being more discriminative than food types, creates a hierarchical structure within the dataset. However, since the food item annotations are solely based on nutritional composition information, they do not always show visual relations with each other, which poses significant challenges when applying deep learning-based techniques for classification. To address this issue, we then propose a multi-stage hierarchical framework for food item classification by iteratively clustering and merging food items during the training process, which allows the deep model to extract image features that are discriminative across labels. Our method is evaluated on VFN-nutrient dataset and achieve promising results compared with existing work in terms of both food type and food item classification.

CCS CONCEPTS

• **Applied computing** → Health informatics; • **Computing methodologies** → **Object recognition**; *Neural networks.*

KEYWORDS

datasets, hierarchical structure, clustering, transfer learning

ACM Reference Format:
Xinyue Pan, Jiangpeng He, and Fengqing Zhu. 2023. Muti-Stage Hierarchical Food Classification. In *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management (MADiMa '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3607828.3617798

1 INTRODUCTION

Image-based dietary assessment, which involves analyzing nutrient and energy intake from food images captured by an individual, is becoming increasingly prevalent[44]. With the ubiquity of mobile devices, many people find it useful to snap pictures of their meals to track dietary intake and monitor adherence to a healthy eating regimen [7, 41]. In addition, image-based dietary assessment is crucial for healthcare applications[2].

A vital component of image-based dietary assessment is food image classification, which aims to predict the food consumed in an eating occasion image[6, 15]. In this paper, we establish that a food can be annotated using both its food type and specific food items, as illustrated in Figure 1. Food types, which are typical classes (*i.e.,* Apple, Bagels, Burgers, etc.) found in datasets such as Food-101 [5] and UEC-256 [25], do not carry any associated nutritional composition information. Food items (*i.e.,* Cheeseburger from fast food, Cheeseburger on a white bun, Beef sandwich, etc.) do possess linked nutritional composition information. Because the annotations for food items are more discriminative compared to those for

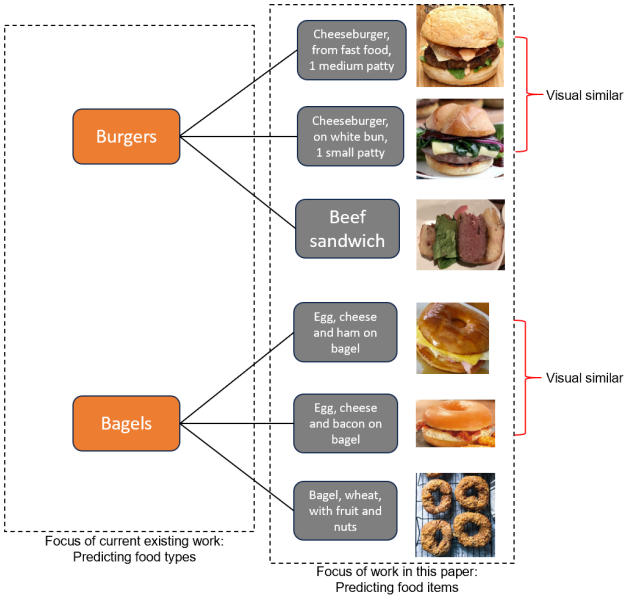

Figure 1: The hierarchical structure in food categorization is depicted in two parts. The left part illustrates the task of food classification in existing work, which focuses on predicting food types. The right side shows the focus of our work: predicting specific food items, each of which is uniquely associated with a particular nutritional composition. A major challenge in classifying food items is that they do not always exhibit visual relations; that is, visually similar images could belong to different food items.

This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

MADiMa '23, October 29, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0284-6/23/10.
https://doi.org/10.1145/3607828.3617798

food types, as shown in Figure 1, we consider food items as subcategories of food types, thereby forming a hierarchical structure within the dataset. While extensive research has been conducted to enhance the accuracy of food image classification based on food types [17, 18, 32, 33], such approaches cannot make classifications based on food items. Therefore, the current food classification results are not directly applicable for dietary assessment, even when the volume of food is estimated [43, 45]. This is because the primary objective of dietary assessment is to perform a nutritional analysis, and the existing classification data lacks corresponding nutritional composition information such as the energy per 100 grams of food. The inability to predict food items stems from the absence of datasets annotated with food items. The lack of such datasets, which would pair each food image with a food item, can be attributed to the domain-specific expertise required for such annotation.

In this paper, we bridge this gap by linking the VFN dataset [33] to nutritional composition information from the USDA Food and Nutrient Database for Dietary Studies (FNDDS) using a USDA food code identifier, as in [29]. In this way, each food image is matched with a food item in FNDDS, which corresponds to a specific USDA food code and is in turn linked to its respective nutritional composition information. Hence, the dataset, named **VFN-nutrient**, presents a two-level hierarchical structure with food types at the top level and food items at the bottom level. Ultimately, our goal is to conduct food image classification based on food items in the dataset.

While most recent works employ Convolutional Neural Networks (CNNs) [1, 21, 46, 48, 50] for hierarchical-based image classification [3, 12, 22, 27], they assume that labels are visually related to each other at each hierarchical level. However, the food item labels in the VFN-nutrient dataset, which are built based on nutritional composition, do not always exhibit visual correlations. Different food items could be visually similar, as shown in Figure 1. This makes it challenging to apply CNN-based models to learn image features directly using food items as labels, because a CNN model learns features visually, and in our case, there is a lack of visual correlation across food items. Additionally, there is a class imbalance issue among food items, causing classification results to be biased towards food items that contain more images. To address these issues, we propose creating visual relations among labels by merging food items iteratively during the training phase through a clustering method and updating image features extracted by the CNN model. We introduce multi-stage hierarchical transfer learning to train and update the model and mitigate the class imbalance issue.

The main contributions of our work can be summarized as follows:

- Unlike most existing works that focus on predicting food types, we aim to predict food items, which include nutritional composition information.
- We introduce the VFN-nutrient dataset, which contains nutritional composition information for each food image.
- We propose an end-to-end food item classification system that iteratively merges visually similar food items and employ multi-stage hierarchical transfer learning for improved classification.

2 RELATED WORK

2.1 Food image classification

Many contributions have been made in the field of food image classification from various perspectives.

Improving Food Image Classification Performance on Food Types: The major challenge of food image classification is the higher inter-class similarity and intra-class dissimilarity compared to other general objects [10, 33, 38]. S. Abdulkadir *et al.* enhanced classification performance by concatenating deep features from different models such as VGG [46], ResNet [21], Wide ResNet [50], and InceptionV3 [48] for food type classification. They observed a noticeable improvement over previous classification performance [52]. D. T. Nguyen *et al.* combined local appearance and structural features, specifically integrating non-redundant local binary pattern features and encoding the spatial relationships between interest points, to improve food image classification performance [36]. R. Mao *et al.* proposed a visual hierarchy structure and employed multitask learning to enhance classification accuracy on food types in the VFN dataset [33]. This approach was further refined by integrating nutritional information into the hierarchy [32]. However, these works still focus on food type classification while the nutritional composition information used is generalized for each food type and does not correspond to each individual food image.

Food Image Classification Under Special Problem Settings Existing work also attempt to approach the food classification issue through the lens of continual lifelong learning [14, 16, 18, 19, 39], while the work in [23, 37] aims to address the issue from the perspective of personalized food image classification, focusing on food images that appear sequentially over time. Additionally, few-shot learning [4, 30, 40], fine-grained classification [24, 33, 49] and long-tailed classification [11, 13, 20] have also been extensively explored in this area.

Datasets Proposed for Food Image Classification: Several works have focused on proposing food image datasets, such as Food-101 [5], Food-2K [35], ISIA-500 [34], and UEC-256 [25]. These datasets are collected through various methods and often focus on specific regions. The aim of introducing these datasets is to enhance the generalization ability of trained models.

However, all of these existing works fall short in linking classification results to nutritional composition information, a crucial requirement for achieving the objectives of image-based dietary assessment.

2.2 Hierarchical image classification

In various contexts, images are associated with different labels, ranging from coarse to fine categories. As such, hierarchical classification is particularly suitable for these scenarios, and numerous important contributions have been made in this field.

Methods Proposed in Hierarchical Image Classification: H. Long *et al.* proposed a hierarchical feature fusion method that classifies target labels by training each level of the hierarchy separately and then fusing the features from each level for final predictions [22]. G. An *et al.* introduced a hierarchical transfer learning method that initially learns the top level of the hierarchy and then transfers the model to the bottom level to make predictions on those

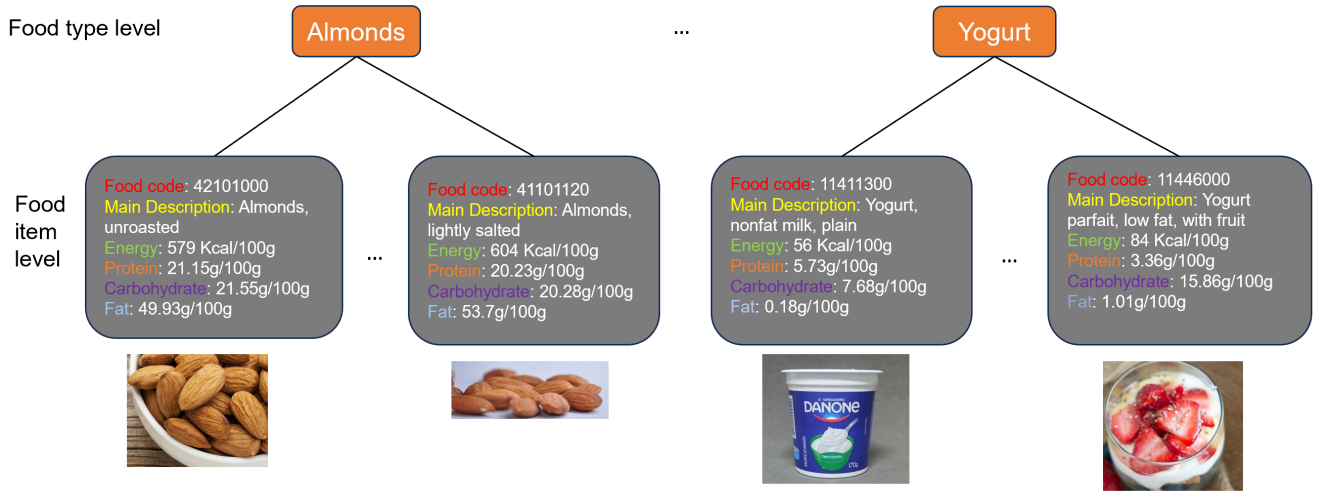

Figure 2: Structure of VFN-nutrient dataset

categories[3]. Additionally, [27] developed a conditional CNN network for multi-label image classification, which is also related to hierarchical classification. Furthermore, [12] presented a CNN-RNN model to classify leaf categories in a hierarchical structure, leveraging the CNN's capability for feature extraction and the RNN's strength in optimizing the classification of both coarse and fine labels. Y. Zhou *et al.* employed multi-task learning to simultaneously learn labels at different levels of the hierarchy[51].

Datasets Used in Hierarchical Image Classification: Publicly available hierarchical image datasets like CIFAR-10 [28] and ImageNet [8] provide training data for hierarchical image classification. In ImageNet [8], the presence of multiple objects in an image contributes to the hierarchical structure, as each image can be assigned multiple labels. For CIFAR-10 [28], certain specific categories are grouped into super-categories, thereby forming a hierarchical structure within the dataset.

However, existing works in hierarchical image classification often assume that classes at the bottom level have visual relationships with each other. Therefore, these methods are not directly applicable to our scenario, where there is limited visual correlation between classes.

3 DATASET

In this paper, we introduce the VFN-nutrient dataset, which is an extension of the original VFN dataset [33]. This new dataset comprises 74 food types selected based on the 'What We Eat In America' (WWEIA) survey. The dataset construction follows the methodology presented in [29]. To annotate food items in each image within the VFN dataset, four experts from the nutrition science team were divided into two independent groups, and they rigorously reviewed all the food images to assign specific food codes from the FNDDS 2017–2018 database to each one. The purpose is to cross-verify each other's work, thereby minimizing subjective errors and improving the reliability of the annotation. Any discrepancies in annotation from two groups were resolved through additional rounds of review involving additional two experts. The aim of the review process was to correct errors, improve categorization, and establish a reliable link to the FNDDS for nutritional analysis.

Each food code corresponds to nutritional composition information and can be considered a food item in the dataset. Consequently, the dataset encompasses 15K images and 945 food items, which belong to 74 different food types. Each image has two labels: one indicating the food type and another specifying the food item.

The structure of the VFN-nutrient dataset is illustrated in Figure 2. The dataset employs a 2-level hierarchical structure: 'food types' make up the top level, while 'food items' constitute the bottom level. The 'food types' are based on the original food classes introduced in the VFN dataset[33].Each food item in the VFN-nutrient dataset is associated with a specific food code from the FNDDS, as well as its corresponding nutritional composition information based on a 100g food sample. Because the nutritional composition data allow the food items to be more discriminative relative to each other, we consider food items as subcategories of food types.

In the dataset, each food item is associated with nutritional composition information based on a 100g food sample. This implies that if we can accurately estimate the weight of the food shown in the image, we can directly calculate the nutrient value for each composition. Given that the ultimate aim of image-based dietary assessment is nutrient analysis based on captured food images, this dataset significantly advances the field toward achieving its intended goal.

However, annotating food items based solely on nutritional composition presents a unique set of challenges for food image classification. Specifically, food items do not necessarily exhibit visual correlations with one another, complicating the task of feature learning for CNN models. In this scenario, the features learned by a CNN model may not be discriminative enough to differentiate between various labels. In this paper, we address this challenge as it specifically relates to food image classification within the dataset. For the experiments to be conducted in subsequent sections, the

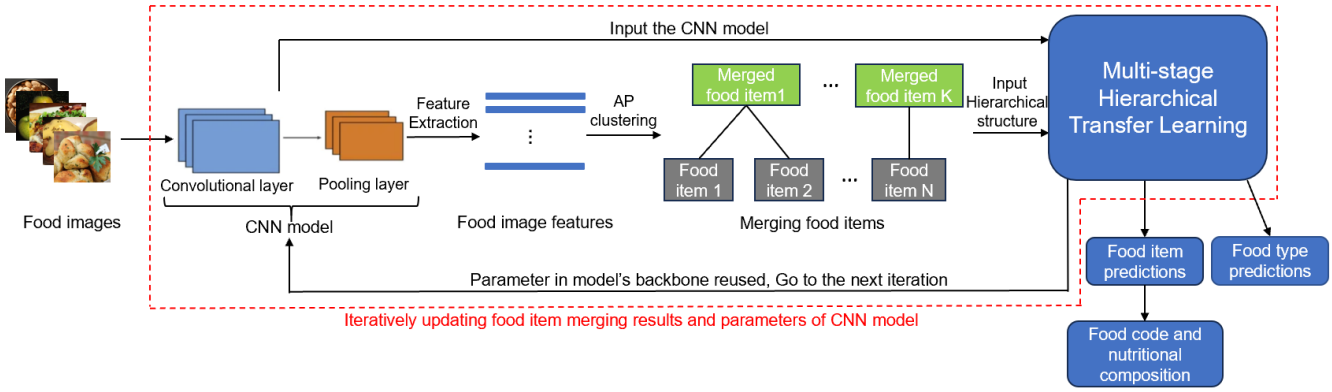

Figure 3: Overview of our proposed method: N represents the total number of food items available for classification. We first extract image features using a CNN model. Second, we merge food items iteratively via Affinity Propagation (AP) clustering, resulting in K merged food items. Third, we train the CNN model using a multi-stage hierarchical transfer learning approach. The trained CNN model is then fed into the next iteration for another round of food item merging, allowing us to update the merging results iteratively. Finally, the iterative process is halted if either the number of iterations reaches 5 or there is no further decrease in the validation loss for predicting food items. At this point, we make inferences on the food items and obtain their nutritional composition information.

VFN-nutrient dataset is divided into training, validation, and testing subsets using a 7:1:2 ratio.

4 METHOD

In this section, we introduce a framework designed to classify food items, as illustrated in Figure 3. The framework aims to tackle two primary issues: the lack of clear visual correlations among different food items and the issue of class imbalance in the dataset. To address these challenges, we introduce an additional level of hierarchy between 'food types' and 'food items'. This is accomplished by iteratively clustering visually similar food items and updating the Convolutional Neural Network (CNN) model through multi-stage transfer learning during the training phase. During the validation and testing phases, the trained model in multi-stage transfer learning stage is applied directly for food image classification.

The method is composed of two main components during the training phase: (1) Employing clustering method to merge food items, (2) Utilizing multi-stage hierarchical transfer learning method to update CNN model and make classification on food items.

4.1 Merging Visual Similar Food Items

To merge food items visually, we first use a CNN model to extract image features. To capture the distribution of image features within each food item, we concatenate the mean and variance feature vectors and feed them as input into a clustering method to merge food items. This input can be represented as $f = [m_n, v_n]$, where m_n denotes the mean feature vector for food item n, and v_n represents the variance feature vector for food item n.

While our approach can accommodate various clustering methods, we specifically utilize Affinity Propagation (AP) [9] in this paper because it does not require a predefined number of clusters. The method operates by transmitting 'messages' among food items; these messages reflect the suitability of one food item to serve as a representative for another. This process updates iteratively across other pairs of food items until convergence. The input to Affinity Propagation consists of the mean and variance features of each food item, concatenated and denoted as f. f_x and f_y represent the input features for food items x and y, respectively. $s(x, y)$ represents the similarity between food items x and y, calculated based on the negative Euclidean distance between their input features. The clustering method utilizes the following equations:

$$s(x, y) = -||f_x - f_y|| \tag{1}$$

$$r(x, y) = s(x, y) - max_{y' \neq y}[a(x, y') + s(x, y')] \tag{2}$$

$$a(x, y) = min[0, r(y, y) + \sum_{x' \in \{x, y\}} r(x', y)] \tag{3}$$

where $r(x, y)$ indicates the suitability of food item y to be the exemplar for food item x, and $a(x, y)$ represents the accumulated evidence that food item x should select food item y as its exemplar. The method eventually outputs the food items that are clustered together. We then merge those that are in the same cluster under the same food type to maintain the hierarchical structure. The merged food items are denoted as $[m_1, m_2, ..., m_K]$, where K denotes the number of merged food items and $N > K$, given N food items.

The merging process is iterative and continues as the image features undergo refinement due to updates in the CNN model. This iteration persists until one of two conditions is met: either the number of iterations reaches five or the validation loss (i.e., loss on the validation set within the dataset) for food item classification ceases to decrease. This methodology confers several key benefits to our approach. First, the image features learned across different labels become more discriminative owing to the visual correlations among the merged food items. Second, the end-to-end system enables the CNN model to adapt its learning from different sets of

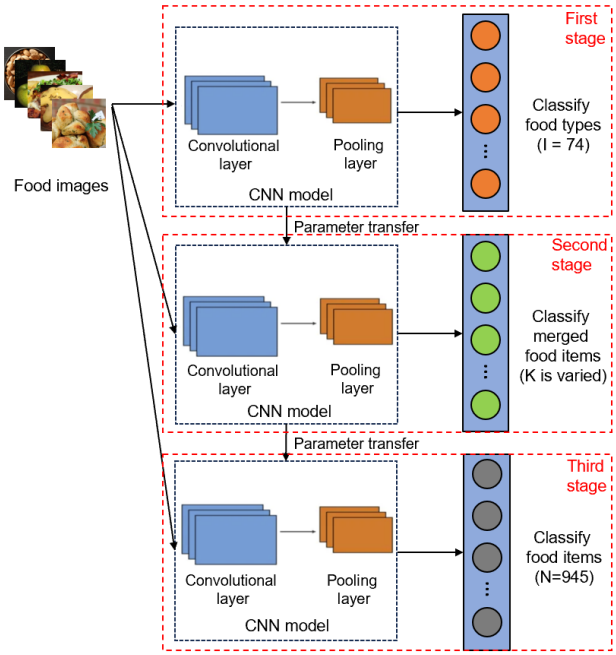

Figure 4: In the multi-stage hierarchical transfer learning approach, the initial CNN model is trained from scratch using food types as labels and is linked to a layer of dimension $I = 74$. The parameters from this model's backbone are then reused in the second stage to build a new CNN model that is connected to a layer of dimension K (variable) and trained with merged food item labels. Subsequently, these parameters are utilized once again in the third stage to create another CNN model that is connected to a layer of dimension $N = 945$ and trained using individual food item labels. The backbone parameters from this final model are then carried forward to the next iteration, continuing another cycle of hierarchical transfer learning. *Note: Labels for merged food items can change with each iteration.*

merged food items at different iterations, due to the variability in merging outcomes in each iteration. This means that distinct image features can be learned and refined during each iteration. Lastly, since many food items are represented by a limited number of training images, the merging process serves to mitigate the issue of class imbalance during the training phase.

4.2 Multi-stage Hierarchical Transfer Learning

To exploit the hierarchical structure created and to address the class imbalance issue in food items, we adopt multi-stage hierarchical transfer learning. This approach allows us to transfer the knowledge acquired from the top level (food types) to the bottom level (food items). Figure 4 outlines the pipeline of our multi-stage hierarchical transfer learning approach. Our methodology encompasses a three-stage, iterative training process.

- First stage: In the initial iteration, we employ a CNN model that is pretrained on ImageNet[8]. This model is then connected to a fully connected layer with a dimension of $I = 74$, utilizing food types as labels. In subsequent iterations, we retain the parameters in the backbone of the CNN model trained in the previous iteration. A new CNN model is constructed and connected to a fully connected layer with a dimension of $I = 74$ from scratch, and it is trained using food types as labels.
- Second stage: The parameters from the backbone of the CNN model trained in the first stage are reused to construct a new CNN model and make it connected to a fully connected layer with a dimension of K, where K varies in different iterations based on the updated food merging results. This model is then trained using merged food items as labels.
- Third stage: Leveraging the parameters from the backbone of the CNN model trained in the second stage, we construct a new CNN model. This model is connected to a fully connected layer with a dimension of $N = 945$ and is trained using food items as labels.

At each training phase, we utilize the cross-entropy loss as our objective function, formulated as follows:

$$L = -\sum_{i=1}^{N} y_{l,i} \log(p_{l,i}) \tag{4}$$

Here, l denotes the stage at which the model is being trained. $y_{l,i}$ represents the ground truth label i at stage l, and $p_{l,i}$ indicates the confidence score for predicting label i at stage l.

The CNN model, refined through this multi-stage hierarchical transfer learning process, is then employed in the food item merging process to generate new merging results, as described in section 4.1. Subsequently, the parameters from the model's backbone are transferred to initiate the next iteration of the multi-stage hierarchical transfer learning process. This cycle continues until either the validation loss on food item classification ceases to decrease in subsequent iterations, or until a maximum of 5 iterations is reached to mitigate the risk of overfitting. The CNN model trained in the final iteration of the third stage is ultimately used to make predictions on food items and to retrieve the corresponding nutritional composition information. For experimental comparison with related works, the saved model from the first stage of the last iteration is used to predict food types.

5 EXPERIMENT

We evaluate our proposed method based on average classification accuracy of predicting food items and mean absolute error in terms of nutritional composition on predicted food items and compare it with other related works on VFN-nutrient dataset.

5.1 Experimental setup

To train the CNN model, we partition the VFN-nutrient dataset into training, validation, and testing sets using a 7:1:2 ratio. The training set is utilized for model training, while the validation set is employed for determining whether the currently trained model should be saved for subsequent testing. The testing set is reserved for final model evaluation. In both our proposed method and related works,

Table 1: Average classification accuracy on predicting food items for different methods: The related work primarily is focusing on predicting food types. We re-implement these methods to fit into our secnario and make them predict food items

Methods	Average classification accuracy(%)
Flat-CNN[21]	46.35
VHML[33]	47.42
IHML[32]	47.59
HFF[22]	48.70
HTL[3]	49.55
Ours	**50.67**

the ResNet-50 model[21] serves as the backbone for the CNN architecture. For optimization, we use the Adam optimizer[26] with an initial learning rate of 0.0001, complemented by a cosine annealing scheduler[31]. Each stage within the multi-stage hierarchical transfer learning process is trained for 15 epochs. After each iteration, we decrease the initial learning rate by a factor of 0.8. The training process terminates either after 5 iterations or if the validation loss fails to decrease in a subsequent iteration.

Methods for comparison: Our method integrates an end-to-end food merging system to establish visual correlations and employs a multi-stage hierarchical transfer learning approach to continually update the CNN model. This facilitates learning across food types, merged food items, and individual food items. We evaluate our proposed framework by comparing with baseline and existing hierarchical based image classification work including:

- **Flat-CNN**[21]: Utilizes a CNN model pretrained on the ImageNet dataset[8] to train image features with target classes as labels. Inference is then performed on these target classes.
- **Visual Hierarchy with Multitask Learning (VHML)**[33]: Establishes a visual hierarchical structure by clustering target classes based on image features extracted using a pretrained Flat-CNN model. It then employs multitask learning to simultaneously classify both clustered and target classes.
- **Integrated Hierarchy with Multitask Learning (IHML)**[32]: Forms an integrated hierarchy by clustering target classes based on both nutritional composition information and visual features. It then applies multitask learning to concurrently classify both the clustered and individual target classes.
- **Hierarchical Feature Fusion (HFF)**[22]: Trains the CNN model based on labels from each hierarchical level separately. It then concatenates the image features extracted from each trained model for the final classification of target classes.
- **Hierarchical Transfer Learning (HTL)**[3]: Trains the CNN model first with the top-level classes in the hierarchy as labels and then retrains it using the bottom-level classes as labels for the final classification.

We adapt these methods to fit our specific use-case, enabling them to predict individual food items rather than just food types.

Table 2: Average classification accuracy on predicting food types for different methods

Methods	Average classification accuracy(%)
Flat-CNN[21]	74.66
VHML[33]	74.95
IHML[32]	75.13
HFF[22]	75.28
Ours	**75.83**

5.2 Experiment results

5.2.1 Average classification performance on predicting food items. Table 1 shows the average classification accuracy for predicting food items across different methods. Our method outperforms all other techniques due to several innovative strategies. Specifically, our method creates visual relations by merging food items, which allows the model to learn more distinct image features across labels compared to learning directly with food items as labels. Importantly, this merging process is updated iteratively as the image features are refined, enabling the model to learn different features in each iteration. This aspect distinguishes our work from **VHML**[33] and **IHML**[32], where the clustering results are static and do not adapt as the model improves.

In addition to this, our method employs an iterative, multi-stage transfer learning approach. After each iteration where the model finishes learning from food items as label, it reverts to learning from food types as label in the next iteration. Any performance gains achieved during this phase enhance the model's subsequent learning when it switches back to food items as labels again. This dynamic updating sets our method apart from **HTL**[3], which does not retrain the model at the top level of the hierarchy.Nevertheless, achieving improvements in food item classification remains challenging due to severe inter-class similarity, intra-class dissimilarity, and a low number of training images among various food items. Our proposed method partially mitigates this issue, but it remains a tough problem to fully resolve.

Moreover, multi-stage hierarchical transfer learning in our method not only improves classification accuracy on food items but also on food types, as demonstrated in Table 2. Unlike **HTL**[3], which transfers the backbone parameters from the top-level hierarchy to the bottom level without further iteration (making its performance on food type classification equal to **Flat-CNN**[1]), our method continues to iterate, enhancing its performance. As a result, we do not include **HTL** in our comparison for food type classification.

5.2.2 Nutrition analysis. We utilize the Mean Absolute Error (MAE) as an evaluation metric to assess performance concerning the nutritional composition of food items. Introduced in [32], this metric is particularly useful for understanding the error in terms of nutritional composition on predicted food items. Specifically, a prediction is considered a 'better mistake' if the nutritional composition of the predicted food item closely resembles that of the target food item. The formula for calculating MAE is given by:

$$MAE = \frac{1}{N}\sum_{i=0}^{N-1} |A_i - \hat{A}_i| \tag{5}$$

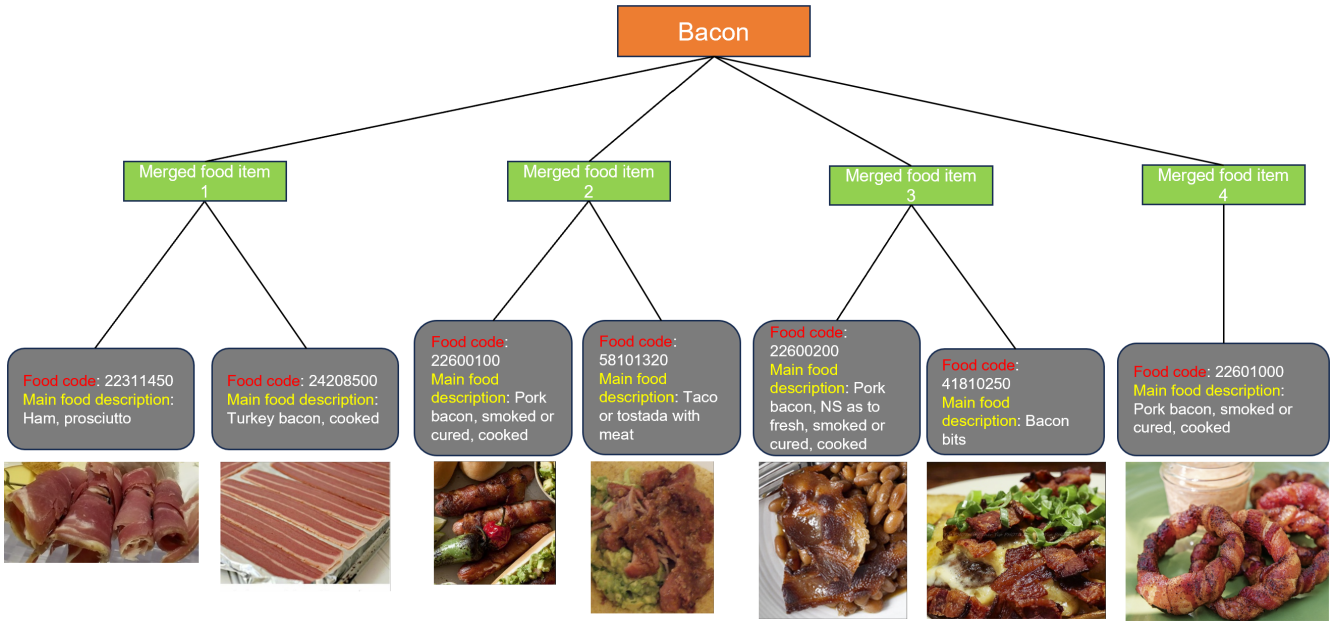

Figure 5: Example of food item merging results

Table 3: Mean absolute error in terms of nutritional composition per 100g food sample

Methods	Energy (Kcal)	Protein (g)	Carbohydrate (g)	Fat (g)
Flat-CNN[1]	39.36	2.39	5.31	3.20
VHML[33]	37.69	2.18	4.90	3.08
IHML[32]	37.74	2.18	4.88	3.05
HFF[22]	34.89	2.02	4.47	2.87
HTL[3]	34.98	2.06	4.48	2.85
Ours	**33.24**	**1.98**	**4.31**	**2.72**

Table 4: Clustering performance at in the first 5 iterations of our proposed method

Iteration Number	Silhouette score	Davies Bouldin index
1	0.036	2.059
2	0.177	1.685
3	0.216	1.585
4	0.236	1.449
5	0.258	1.374

In this equation, A_i represents the nutritional value per 100g of the predicted food item, and $\hat{A}_i$ represents the nutritional value per 100g of the target food item. N denotes the total number of testing images. We report the MAE in terms of macro nutritional components—energy, carbohydrates, fats, and proteins—as shown in Table 3.

Our method also demonstrates superior performance in terms of nutritional composition errors per 100g food sample compared to existing works. Specifically, our approach is designed to make "better mistakes," meaning that even if the model misclassifies a food item, the predicted item's nutritional profile closely resembles that of the target item. This is a significant advantage, as it ensures a higher degree of accuracy in the nutritional information provided, despite potential misclassifications.

5.2.3 Clustering performance at different iterations of proposed method. To substantiate the efficacy of our iterative method for updating both the merged food items and the backbone of the CNN model, as delineated in Section 4.1, we examine the clustering results obtained after each iteration using two established metrics: the Silhouette Score [42] and the Davies-Bouldin Index [47].

The Silhouette Score [42] serves as an indicator of the distinctness of the clusters relative to each other. Its value ranges from -1 to 1, with higher values representing more distinct and thus better clusters. Conversely, the Davies-Bouldin Index [47], ratio of intra-cluster to inter-cluster distance, offers an inverse interpretation. Lower values of this index imply better clustering performance, as they indicate clusters that are both tight and well-separated.

Table 4 shows the clustering performance for 5 iterations of our proposed method. The clustering performance is always improving and this can provide more visually correlated merged food items for CNN model to learn.

5.3 Qualitative results from merging food items

Figure 5 presents a the outcome of merging food items within the "bacon" food type category. Upon observation, it becomes evident that all food items with visual similarity have been cohesively

merged to form a newly merged food item. For instance, the food item identified by USDA food code 22311450, which originally has only three training images, is merged with the food item identified by USDA food code 24208500, which originally has one training image. As a result of this merging, we now have a training set of four images to train the model under a single merged food item. This is particularly beneficial in mitigating the class imbalance issue inherent in the data. Such a strategy highlights the efficiency of our approach in leveraging limited image data from specific classes for effective training. This, in turn, facilitates the CNN models in acquiring more discriminative image features.

6 CONCLUSION

In this paper, we exploited a VFN-nutrient dataset that labels each food image with a corresponding food item and its nutritional composition information. We consider food items as subcategories of food types, thereby forming a hierarchical structure within the dataset. Predicting food items enables us to retrieve the corresponding nutritional composition information, thus bringing us closer to the goal of image-based dietary assessment. To create visual relations among food items, we implemented an end-to-end food items merging method during training phase by updating CNN model for extracting image features iteratively through multi-stage hierarchical transfer learning, which can also address the class imbalance issue across food items.

Despite these contributions, there are other potential strategies that could further improve food item classification accuracy. For future work, we aim to incorporate ideas from multi-modal learning to enhance our classification of food items, and in turn, refine the results of food image classification on food items.

REFERENCES

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. *Proceedings of International Conference on Engineering and Technology* (2017), 1–6.

[2] Dario Allegra, Sebastiano Battiato, Alessandro Ortis, Salvatore Urso, and Riccardo Polosa. 2020. A review on food recognition technology for health applications. *Health psychology research* (2020).

[3] Guangzhou An, Masahiro Akiba, Kazuko Omodaka, Toru Nakazawa, and Hideo Yokota. 2021. Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. *Proceedings of the International Conference on Pattern Recognition Workshop* (2021), 571–598.

[4] Berker Arslan, Sefer Memiş, Elena Battini Sönmez, and Okan Zafer Batur. 2022. Fine-Grained Food Classification Methods on the UEC FOOD-100 Database. *IEEE Transactions on Artificial Intelligence* 3, 2 (2022), 238–243.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. *Proceedings of European Conference on Computer Vision* (2014).

[6] C. J. Boushey, M. Spoden, F. M. Zhu, E. J. Delp, and D. A. Kerr. 2017. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proceedings of the Nutrition Society* 76, 3 (2017), 283–294.

[7] Steven Coughlin, Mary Whitehead, Jeff Mastromonico, Dale Hardy, and Selina Smith. 2015. Smartphone Applications for Promoting Healthy Diet and Nutrition: A Literature Review. *Jacobs journal of food and nutrition* (2015).

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255.

[9] Brendan Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* (2007).

[10] Wenjin Fu, Yue Han, Jiangpeng He, Sriram Baireddy, Mridul Gupta, and Fengqing Zhu. 2023. Conditional synthetic food image generation. *Proceedings of the IS&T International Symposium on Electronic Imaging* 35 (January 2023), 1–6. San Francisco, CA.

[11] Jixiang Gao, Jingjing Chen, Huazhu Fu, and Yu-Gang Jiang. 2022. Dynamic Mixup for Multi-Label Long-Tailed Food Ingredient Recognition. *IEEE Transactions on Multimedia* (2022).

[12] Yanming Guo, Yu Liu, Erwin M. Bakker, Yuanhao Guo, and Michael S. Lew. 2018. CNN-RNN: a large-scale hierarchical image classification framework. *Multimedia Tools and Applications* (2018).

[13] Jiangpeng He, Luotao Lin, Heather A. Eicher-Miller, and Fengqing Zhu. 2023. Long-Tailed Food Classification. *Nutrients* 15, 12 (2023). https://doi.org/10.3390/nu15122751

[14] Jiangpeng He, Luotao Lin, Jack Ma, Heather A. Eicher-Miller, and Fengqing Zhu. 2023. Long-Tailed Continual Learning For Visual Food Recognition. *arXiv preprint arXiv:2307.00183* (2023).

[15] Jiangpeng He, Runyu Mao, Zeman Shao, Janine L Wright, Deborah A Kerr, Carol J Boushey, and Fengqing Zhu. 2021. An end-to-end food image analysis system. *Electronic Imaging* 2021, 8 (2021), 285–1.

[16] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. 2020. Incremental Learning In Online Scenario. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020), 13926–13935.

[17] Jiangpeng He, Zeman Shao, Janine Wright, Deborah Kerr, Carol Boushey, and Fengqing Zhu. 2020. Multi-task Image-Based Dietary Assessment for Food Recognition and Portion Size Estimation. *Proceedings of IEEE Conference on Multimedia Information Processing and Retrieval* (2020), 49–54.

[18] Jiangpeng He and Fengqing Zhu. 2021. Online Continual Learning For Visual Food Classification. *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops* (2021), 2337–2346.

[19] Jiangpeng He and Fengqing Zhu. 2022. Exemplar-Free Online Continual Learning. *2022 IEEE International Conference on Image Processing* (2022), 541–545. https://doi.org/10.1109/ICIP46576.2022.9897554

[20] Jiangpeng He and Fengqing Zhu. 2023. Single-Stage Heavy-Tailed Food Classification. *arXiv preprint arXiv:2307.00182* (2023).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016).

[22] Long He and Dandan Song. 2022. Hierarchical image classification with a literally toy dataset. *Proceedings of International Conference on Mechanisms and Robotics* 12331 (2022), 1101–1110.

[23] Shota Horiguchi, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. 2018. Personalized Classifier for Food Image Recognition. *Proceedings of IEEE Transactions on Multimedia* 20 (2018), 2836–2848.

[24] Shuqiang Jiang, Weiqing Min, Yongqiang Lyu, and Linhu Liu. 2020. Few-Shot Food Recognition via Multi-View Representation Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 3 (2020).

[25] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. *Proceedings of European Conference on Computer Vision* (2014).

[26] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *Proceedings of International Conference on Learning Representations* (2017).

[27] Brendan Kolisnik, Isaac Hogan, and Farhana Zulkernine. 2021. Condition-CNN: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications* 182 (2021), 115195.

[28] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* (2009). Toronto, Ontario.

[29] Luotao Lin, Jiangpeng He, Fengqing Zhu, Edward J. Delp, and Heather A. Eicher-Miller. 2023. Integration of USDA Food Classification System and Food Composition Database for Image-Based Dietary Assessment among Individuals Using Insulin. *Nutrients* 15, 14 (2023).

[30] Chengxu Liu, Yuanzhi Liang, Yao Xue, Xueming Qian, and Jianlong Fu. 2021. Food and Ingredient Joint Learning for Fine-Grained Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 6 (2021), 2480–2493.

[31] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. *Proceedings of International Conference on Learning Representations* (2017).

[32] Runyu Mao, Jiangpeng He, Luotao Lin, Zeman Shao, Heather A. Eicher-Miller, and Fengqing Maggie Zhu. 2021. Improving Dietary Assessment Via Integrated Hierarchy Food Classification. *Proceedings of the IEEE International Workshop on Multimedia Signal Processing* (2021), 1–6.

[33] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. 2021. Visual Aware Hierarchy Based Food Recognition. *Proceedings of the International Conference on Pattern Recognition Workshop* (2021), 571–598.

[34] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network. *Proceedings of ACM International Conference on Multimedia* (2020).

[35] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2023. Large Scale Visual Food Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), 9932–9949.

[36] Duc Thanh Nguyen, Zhimin Zong, Philip O. Ogunbona, Yasmine Probst, and Wanqing Li. 2014. Food image classification using local appearance and global

structural information. *Neurocomputing* 140 (2014), 242–251.

[37] Xinyue Pan, Jiangpeng He, Andrew Peng, and Fengqing Zhu. 2022. Simulating Personal Food Consumption Patterns Using a Modified Markov Chain. *Proceedings of 7th International Workshop on Multimedia Assisted Dietary Management* (2022), 61–69. https://doi.org/10.1145/3552484.3555747

[38] Andrew Peng, Jiangpeng He, and Fengqing Zhu. 2023. Self-Supervised Visual Representation Learning on Food Images. *arXiv preprint arXiv:2303.09046* (2023).

[39] Siddeshwar Raghavan, Jiangpeng He, and Fengqing Zhu. 2023. Online Class-Incremental Learning For Real-World Food Classification. *arXiv preprint arXiv:2301.05246* (2023).

[40] Javier Ródenas, Bhalaji Nagarajan, Marc Bolaños, and Petia Radeva. 2022. Learning Multi-Subset of Classes for Fine-Grained Food Recognition. *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management* (2022), 17–26.

[41] Sabiha Samad, Fahmida Ahmed, Samsun Naher, Muhammad Ashad Kabir, Anik Das, Sumaiya Amin, and Sheikh Mohammed Shariful Islam. 2022. Smartphone apps for tracking food consumption and recommendations: Evaluating artificial intelligence-based functionalities, features and quality of current apps. *Intelligent Systems with Applications* 15 (2022), 200103.

[42] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics* (2020).

[43] Zeman Shao, Shaobo Fang, Runyu Mao, Jiangpeng He, Janine L. Wright, Deborah A. Kerr, Carol J. Boushey, and Fengqing Zhu. 2021. Towards Learning Food Portion From Monocular Images With Cross-Domain Feature Adaptation. *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)* (2021), 1–6. https://doi.org/10.1109/MMSP53017.2021.9733557

[44] Zeman Shao, Yue Han, Jiangpeng He, Runyu Mao, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu. 2021. An Integrated System for Mobile Image-Based Dietary Assessment. *Proceedings of the 3rd Workshop on AIxFood* (2021), 19–23. https://doi.org/10.1145/3475725.3483625

[45] Zeman Shao, Gautham Vinod, Jiangpeng He, and Fengqing Zhu. 2023. An End-to-end Food Portion Estimation Framework Based on Shape Reconstruction from Monocular Image. *arXiv preprint arXiv:2308.01810* (2023).

[46] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of International Conference on Learning Representations* (2015).

[47] Akhilesh Kumar Singh, Shantanu Mittal, Prashant Malhotra, and Yash Vardhan Srivastava. 2020. Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means. *2020 Fourth International Conference on Computing Methodologies and Communication* (2020).

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2015).

[49] Yanqi Wu, Xue Song, and Jingjing Chen. 2022. Few-Shot Food Recognition with Pre-Trained Model. *Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and Related APPlications*, 45–48.

[50] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. *Proceedings of the British Machine Vision Conference* (2016).

[51] Yu Zhou, Xiaoni Li, Yucan Zhou, Yu Wang, Qinghua Hu, and Weiping Wang. 2022. Deep collaborative multi-task network: A human decision process inspired model for hierarchical image classification. *Pattern Recognition* 124 (2022), 108449.

[52] Abdulkadir Şengür, Yaman Akbulut, and Ümit Budak. 2019. Food Image Classification with Deep Features. *2019 International Artificial Intelligence and Data Processing Symposium* (2019), 1–6.