# FoodLens: Fine-grained and multi-label classification of Indian Food Images


## ABSTRACT

India has rich cultural diversity which reflected in its variety of food. In recent years, Computer vision has played a key role in classify food images for automated tagging, nutrition profiling and many other tasks. However, the existing state-of-the-art AI-based food classification models trained on global food images have subpar performance on Indian food images. This is due to the lack of representation of Indian food in existing food datasets and unique image classification challenges specific to Indian food, such as cuisines having multiple dishes within a single image and regional fine grained varieties of the dishes. To address these challenges, a dataset with 30K food images consisting of popular dishes from restaurant menus across India was curated and annotated with multi-label and fine-grained labels for each dish in the image. All the dishes were mapped onto a hierarchical tree which models a categorical breakdown of Indian food. Custom loss function was tuned to learn from hierarchical and multi-label information contained in the Indian food images. Augmenting our loss on existing methods gives 13% improvement on average AUPRC and shows better classification performance on Indian food dataset compared to state of art food classification models with comparable results for other food benchmark datasets.

More than 100k photos which are submitted each day on Google Maps on Indian restaurants and many more on social media channels were utilized for the project.



## KEYWORDS

Computer Vision, Machine Learning, Classification, Multi-label, Hierarchical

**ACM Reference Format:**

. 2024. FoodLens: Fine-grained and multi-label classification of Indian Food Images. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 18 pages.


## 1 INTRODUCTION

Food and nutrition have long played a crucial role in people's daily lives. In recent years, the pervasive nature of mobile devices has begun to redefine how we manage our food consumption facilitated by technology. With the advent of camera based smartphones, millions of pictures and videos of food are uploaded each day on numerous surfaces on the internet. These are often not tagged with the dish and relying on manual labelling is not scalable. Automated detection of food images provides scalable indexing and searching, making it easier to find relevant images, restaurants and blogs. In addition, many meal logging applications exist to help individuals monitor their eating patterns. These mobile applications elicit meal information including type and time of the meal. However, the primary method of this data collection is through text. Prior research has shown how this data collection can quickly turn tedious[28]. Identifying the food also helps in mapping the nutritional value of the food in many health applications. Automated food identification through images has thus emerged as an important domain for enabling healthy lifestyle changes.

Computer vision techniques have been used in food identification from images[20, 25, 32, 57] and videos, food volume estimation with sensors[32], segmentation of food images[32] and learning from recipes [31]. These techniques require supervised training data to learn and generalize to unseen images. Prior works have curated several datasets[16, 22, 31, 32] and developed approaches to tackle the problem of automated food identification. However, most existing datasets and techniques have focused exclusively on Western food, with exceptions of Banerjee, Rajayogi et al. and Nayak et al..

Indian food, in particular, consists of a variety of highly diverse regional cuisines under-represented in public food image datasets. Different states and regions within India have their unique styles of cooking, ingredients, and flavors, influenced by culinary traditions spanning centuries. Indian food draws on several staple ingredients such as rice, lentils, wheat, as well as spice combinations (e.g., cumin, turmeric, coriander). Furthermore, an Indian meal typically comprises of multiple items served simultaneously, such as lentils, rice, breads, and chutneys, in contrast to other cuisines that have a sequential meal structure starting from appetizer, to main course, to dessert.

To understand the performance of these models on Indian food images, we conducted preliminary annotation experiments to qualitatively assess the performance of Google Lens on Indian dishes. Google Lens is an image recognition software that was released in 2017 by Google and is now available in most smartphones. We collaborated with 10 annotators to label food in 100-150 images for 13 popular Indian dishes. Many of these images had multiple dishes in each image. For each displayed image, raters were asked mark the prominent and non-prominent (side dishes) along with confidence. Raters were asked to identify the most prominently present dish from the given options (coarse and then fine-grained label per dish) on Likert scale (min: 1, max: 5). Once a coarse label is added, options for the respective fine-grained labels were displayed in the next dropdown. Post data analysis of rated images, we found: (1) Overall ratings showed poor accuracy (less than 50%) when using a generic food classifier on Indian dishes across all 13 dishes. (2) Ratings based on granularity of the identified dish showed that most of the dishes were classified at coarser level than desired (40% of correctly classified dishes were identified at fine-grained level). (3) Ratings also showed that most of the identification focused on the main dish in the plate leaving many side dishes unidentified. (Less than 25% of multi dish plate classified correctly).


*Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.*

Identifying fine-grained labels and making better mistakes are important to improve the user experience in search and nutritional health applications. Detecting the non-prominent side dishes is important in nutrition mapping and the ability to search images with multiple dishes. The paper addresses the problems found in UXR study in Indian food image classification by developing hierarchy-aware multi-label classification algorithm. This improves the fine-grained classification while identifying multiple dishes in the plate. Overall, the key contributions contributions of this study are as follows:

- Creating an multi-rater annotated dataset of 26k images across 137 Indian food dishes, distilled from restaurant menus along with hierarchical relation among dish labels.
- Developing a methodology for fine-grained and multi-label classification using the hierarchy information and multi-label annotations.
- Showcase better classification accuracy over state-of-the-art food classification models trained on both Indian and non-Indian food image datasets.

## 2 RELATED WORK

Digital platforms like smartphones and wearables are having an increasing impact on lifestyle and well-being: physical activity, nutrition and sleep [53, 54]. Smartphones sensors like cameras are empowering users access searchable multimodal content across videos, images and text[23, 59]. Cameras have increasingly been used in food and nutrition science with advances in deep learning[57, 58]. Machine learning methods have been used for food identification from images[25, 26, 44], videos [8, 24], food volume estimation with sensors [26], segmentation of food images [20, 51], and learning from recipes [22, 31]. The performance of these algorithms on new examples depend on the training dataset and annotation[16, 33]. As a result, it is important to have sufficient representation of food images in training these systems to match use cases, region and culture. Food classification models have been adapted to specific cuisines by curating specialized datasets[10, 15] and developing algorithms to enhance localization in APAC regions[27, 37, 52, 56]. India has a rich cultural diversity of food with many centuries of influence. The Indian food image datasets[2, 36] and methods[40] are limited to a few dishes with no customization to the uniqueness of Indian food images.

Indian cuisine has multiple adaptations of popular dishes based on local regions which demand fine-grained classification [5]. Furthermore, most Indian food plates consist of many dishes. Our work brings the hierarchy and multi-label adaptation of deep learning based methods to Indian food images and develops comprehensive solutions for uniqueness of the Indian dishes. Previous work in multi-label classification has been around exploiting label correlation via graph neural networks [11, 12, 19] or word embeddings [13, 49]. Others are based on modeling image parts and attentional regions [21, 55], and using recurrent neural networks[34, 48], embedding space constraints, [39] region sampling [60]. Methods are proposed to incorporate hierarchical knowledge to single-label classifiers to add additional semantics to the models' learning capabilities such that when the model makes mistakes, it makes semantically better mistakes. Hierarchical information is important in many other applications such as food recognition [29, 50], protein function prediction [3, 4, 6], image annotation [18], text classification [30, 42, 43]. There has also been some recent work in hierarchical multi-label text classification [7, 9, 17, 30, 43].

## 3 METHODOLOGY

At present, there are no large scale open dataset dedicated for Indian images. Dataset consisting of images of Indian dishes is created for training and evaluation of Indian food classification models. Identifying the set of dish names and collecting images for the selected Indian dishes are the key steps in creating the dataset. Choices for selecting the dish types and the image source are aligned with identification of dishes from food images uploaded from restaurants in India on platforms like Google Maps restaurant reviews. The choice of the dish types and image source are for illustration of our methods on a generic application and not representative of the diversity or popularity of dishes in the country. The data collection strategy, number of dishes to be classified, number of images collected for each dish type can be swapped with any method tailored to the specific application of the food identification model. The following three subsections describe the label selection and curation process, also the food dataset creation. We then utilized hierarchy-aware multi-label classification algorithm in order for better Indian food classification.

### 3.1 Food Label Curation

Following methodology was adapted for the selection of top 200 dishes:

- 500 top restaurants were identified based on 4+ ratings on Google Maps with the highest number of reviews in popular cities across India. The restaurants were evenly distributed across 26 (out of 28) states in India with average number of reviews per restaurant 25000+ and 4.3+ rating out of 5. The restaurants ranged across franchises like McDonald's, local favorites and popular Indian chain restaurants.
- Menu Image was extracted from Google Maps for each of the 500 restaurants. Google OCR (Document AI) was used to digitize the menu image to extract the dishes served in the restaurant.
- Each dish name is preprocessed to correct for spelling correction and capitalization. N-gram phrase frequency is calculated for each dish type. We took the top 200 most frequent Indian food names as labels for creating the Image dataset.
- This method was aligned to identify most commonly occurring dishes among a diverse set of restaurants across various states in India. This serves as a proxy for commonly available dish types ordered in restaurants by Indians.

### 3.2 Distillation of Food labels

We collected a set of 235 unique food dishes labels from top-rated restaurant menus across India. However, the most common dishes from top-rated Indian restaurants contained several dishes that were clearly not of Indian origin, such as cheese dip, steak, and waffles. This added an additional challenge to the filtration process as many Indian restaurants have adopted dishes from other cultures, resulting in hybrid dishes that may not be easily classified as either

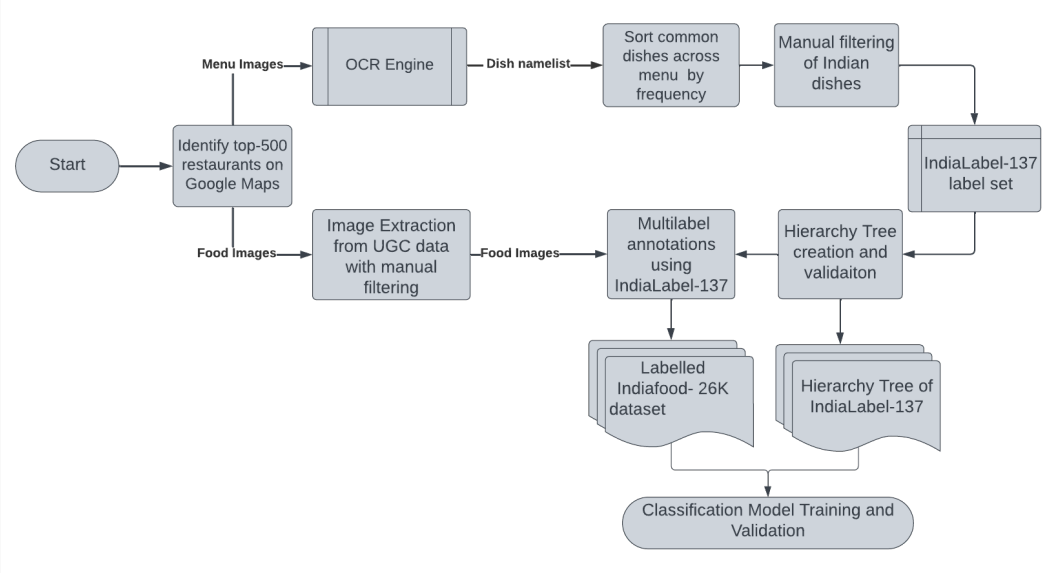


**Figure 1: A flowchart of the food label and food image dataset curation process**

Indian or foreign. As exemplified by the growing number of fast food chains to high end global cuisines appearing in the urban cities of India[14].

To create a more focused dataset of Indian food, we assembled a team of four native Indians to review the dish names and remove any that were not of Indian origin. With these adaptations to global cuisine, there have been more hybrid dishes that are mixture of authentic Indian food and strong external influences such as cheese pav, veg sandwiches and many more.

The team of 4 did their best to filter out any dishes that would not be considered Indian dishes (i.e., chocolate brownie, honey chili fries, etc.). In the cases of dishes with uncertain origin or broad classification, such as "corn chilly" and "curd", we consulted culinary experts. This resulted in a further removal of 19 labels.

At last, The final dataset of 137 unique Indian food labels was curated. These labels represent a wide range of Indian cuisine, from traditional dishes such as tandoori chicken and idli to more modern fusion dishes such as cheese pav and veg sandwiches. The label set will be referred to as IndiaLabel-137 in the paper. Figure 1 summarizes the steps involved in dataset and label set curation.

## 3.3 Food images collection and annotation

The following methodology was adopted to collect the images of identified 137 list:

Images are sourced from the user uploaded food images on Google Maps for the 500 preselected restaurants, which were used to gather the food labels. Which resulted in a total of roughly 53,400 images, with machine labels. In which, we discarded the dataset of all the labels that weren't on the final IndiaLabel-137. This resulted in 35,500 images that were applicable, which were sent to human annotation. Then, raters manually labeled food images from the image dataset and skipped images which were blurry, had PII (Personal Identifiable Information) or wasn't able to be identified with a fine-grained label.

This process resulted around 26,500 images with good image quality from the IndiaLabel-137 and around 8,200 images were skipped.

The image dataset will henceforth be referred to as IndiaFood26K in the paper. Dataset reflects images taken in restaurant settings which are well decorated and arranged compared to home cooked food. We expect models trained on IndiaFood26K to generalize to restaurant uploaded images which would aid in automatically tagging images to support better user experience in identifying relevant restaurants on the web. They may also generalize to aid digital apps for online food delivery, identifying new dishes served in Indian restaurants but may not work well for home cooked meals.

*3.3.1 Annotation Process.* The annotation process for our Indian food image dataset was designed to be robust, reliable, efficient, and scalable. To address the challenge of multiple food items being present in a single image, a multi-label and multi-rater approach was used. The dataset contains 137 food labels, with approximately 250 images per label, for a total of approximately 35,000 images. Two rounds of human annotation were conducted, with each image being annotated by a single annotator in each round.

*3.3.2 Labeling tool description and rater profiles and matching.* The image annotation user interface (UI) is a critical tool for ensuring the quality and accuracy of image annotation. The UI provides annotators with a convenient and efficient way to label images, and it also collects valuable feedback on the quality of the image data.

The UI consists of three main components: the image, two drop-down menus for labeling, and two multiple choice questions on image quality and image label completeness.

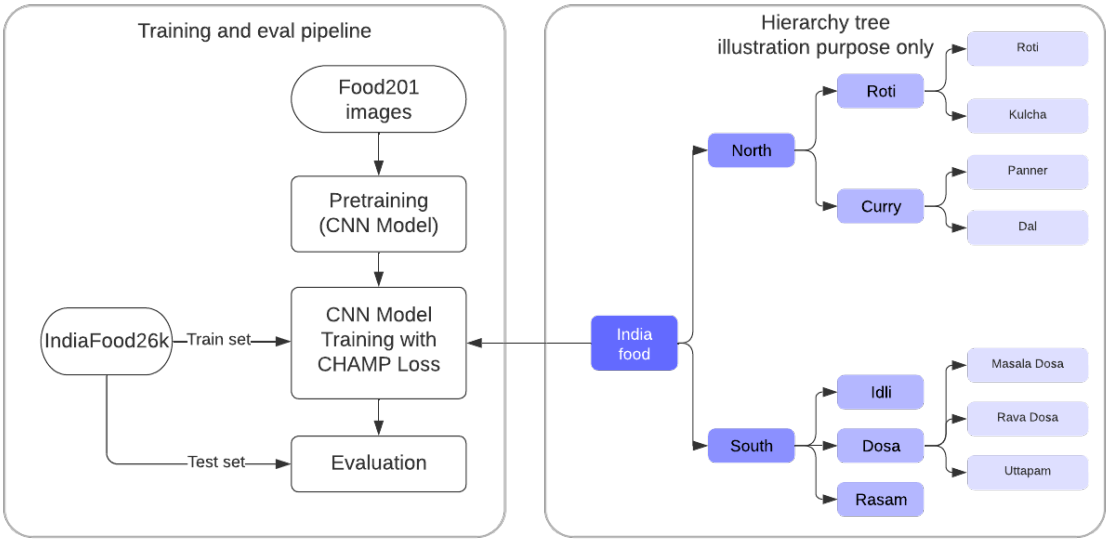


**Figure 2: India food classification pipeline**

The image 15 is displayed on the right-hand side of the UI. It is large enough to be easily viewed, and it is accompanied by a skip button that allows annotators to skip the image if they are unable to label it. The skip button also includes a dropdown menu with reasons for skipping: "Unable to identify the exact fine-grained label," "Unable to identify the dish at all," "Image contains text / caption / watermark of the label in it," "Image contains packaged food /canned food / food mixture," or "Image does not contain a food dish."

The two dropdown menus on the left-hand side of the UI allow annotators to label the image with precision and accuracy. The first menu is for fine-grained labeling, and it contains a dropdown section of the 137 food labels. An auto-complete feature helps to streamline the annotating experience, as annotators are not required to type out the entire name of the food to annotate the image with the dish name. The second dropdown menu records the annotator's confidence level between low or high depending on how well the label matches the image.

The multiple choice questions at the bottom of the UI allow annotators to provide feedback on the quality of the image and the label set. The first question, "Quality of Image?", assesses the image quality with the following options: "Good," "Low resolution," "Unclear / blurry," or "Noisy with perturbations." The second question, "How many dishes were labeled in this image?", measures the competence of the label set with the following options: "Are all dishes in the image tagged," "More than half which is 50% of the dishes are tagged," or "Less than half which is 50% of the items are tagged in the image."

*3.3.3 Hierarchical tree creation.* To improve the accuracy of AI-based Indian food classification, we sought to create a label set of fine-grained classifications structured within a hierarchical tree. The goal behind this creation is to assist AI systems in making better mistakes by having inherent semantic knowledge through the hierarchy [50]. Leveraging this inherent trait within hierarchies, we hope to close the gap in the accuracy of identifying Indian food images.

Due to the nuance of creating a hierarchy to assist AI systems with the goal of providing semantic information, we had to start from the ground to develop the bin. [38, 41] To gain a deeper understanding of how to create appropriate categories within Indian food, we first consulted with a culinary expert with over 20 years of experience in the field. We wanted to understand Indian food from their perspective and glean insights into how to generally structure the hierarchy.

The culinary expert informed us that Indian food can be broadly divided into two regional categories: North Indian and South Indian. This distinction is based on the distinct regional meal staples and corresponding visual differences between dishes from those regions. Although there are finer distinctions between East and West Indian cuisine, we found that the meals that we had narrowed down were better categorized with North and South, and that there would be no added benefit in distinguishing further with our current data [47].

At the second level of distinction, we made the separation between vegetarian and non-vegetarian. This distinction is important to ensure that vegetarian food does not get confused with non-vegetarian food. In India, there is a cultural significance to this distinction, as vegetarians tend to dislike their food being mistaken for non-vegetarian food. Therefore, we prioritized the distinction between veg and non-veg [35].

To implement these first two insights, we conducted a multi-rater process of manually labeling 137 food dishes as either North or South Indian, and as either vegetarian or non-vegetarian. We assembled a group of five non-expert Indians, two of whom were

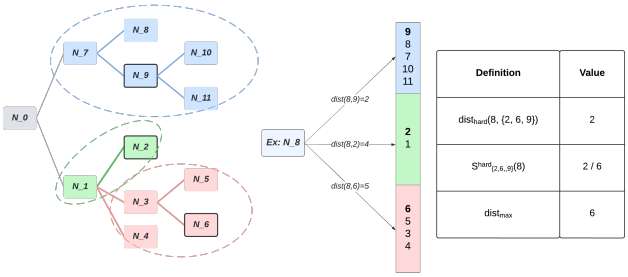


**Figure 3: HMC Loss function example**

North Indian, two of whom were South Indian, and one of whom was East Indian. We then classified each food label through a majority vote. Although this process was not perfect, it gave us an indication of whether a dish was North or South, and whether it was vegetarian or non-vegetarian.

After the multi-rater process, we further broke down the four high-level labels (North Indian, South Indian, vegetarian, and non-vegetarian) into more specific multi-labels. At this third level, we identified the specific type of dish the food would be categorized in, such as curries, sides, and drinks for vegetarian dishes. This distinction only affected vegetarian dishes, as non-vegetarian dishes fell into either starters or main dishes. The distinction between curries and sides was key to understanding the Indian staple of eating curry with roti or rice. This applies to non-vegetarian curries as well, which have an accompaniment of rice or roti.

## 3.4 ML Methodology

We propose to use the CHAMP[45] loss function that leverages the hierarchical relationships in the dataset and can easily be plugged into existing models and backbones. The loss function is described in Section 3.4.1 and evaluation methods are discussed in Section 3.4.2

*3.4.1 Loss Function.* Let $\mathcal{D}$ represent the training data distribution. We draw samples $(x, y)$ from $\mathcal{D}$ where $x \in \mathbf{R}^d$ denotes the input and $y \in \mathbf{R}^L$ denotes a binary vector of labels where one (zero) at position j indicates the presence (absence) of the class j. Let $\theta$ denote the set of parameters. Let $\hat{y} \in \mathbf{R}^L$ denote the vector of predicted class probabilities in the same order as $y$. The standard baseline for training multi-label classifiers is to predict probabilities for each class independently and minimize the binary cross entropy of each predictor. The instance-level loss for the standard baseline would be:

$$J(x, y; \theta) = -\sum_{j=1}^{L} y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)$$

We incorporate hierarchy $T$ into the method by expanding the label set to include non-leaf nodes of the hierarchy. To add positive examples for the internal nodes, we alter the ground-truth sets of each sample to include all the ancestors of every leaf label in the set. This reduces the hierarchical classification problem to a standard multi-label classification over an expanded label set. This serves as our hierarchical baseline.

To make more efficient use of the hierarchical information, we weigh the $L$ loss terms differently for each instance. Given an instance, the weight for the term corresponding to class j is proportional to the path distance between that class and the closest ground-truth to it in the hierarchy. The modified loss function is:

$$J(x, y; \theta) = -\sum_{j=1}^{L} y_j \log \hat{y}_j + (1 - y_j)(1 + s(j, y)) \log(1 - \hat{y}_j)$$

where $s$ is the instance-specific class weight. It is defined as:

$$s(j, y) = \beta \min_{i \in \{l \in L \mid y_l = 1\}} \frac{\text{distance}(j, i; T)}{\text{diameter}(T)}$$

where $\beta$ is a tunable hyperparameter.

*3.4.2 Evaluation Metrics.* Let $(x, y)$ denote a sample from the data distribution $\mathcal{D}$ where $x \in \mathbf{R}^d$ represents an image and $y \in \mathbf{R}^L$ is a one-hot binary vector of class ground-truths as described earlier. Given a model that outputs class probabilities, we assume the prediction to be positive for class j if the probability exceeds a class-specific threshold $T_j$.

We calculate per-class performance metrics. The overall performance is reported as the average per-class performance. Hence, for rest of the section, assume that we work with a binary classifier. Given a test distribution $(x, y) \sim \mathcal{D}$ and a model, we obtain the presence of the class, denoted by binary variable $\hat{y}$, by thresholding the predicted probability with $T$. We define precision and recall as:

$$\text{Precision} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[y = 1 | \hat{y} = 1\right]$$
$$\text{Recall} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\hat{y} = 1 | y = 1\right]$$

The choice of the threshold $T$ controls the precision-recall trade-off and is typically selected based on practical requirements. A higher threshold generally leads to higher precision and lower recall, and a lower threshold leads to lower precision and higher recall.

To gauge the general performance without committing to a specific threshold, we use the Area Under Precision-Recall Curve (AUPRC) to summarise the classification performance of a given class. The precision-recall curve is traced by computing the (precision, recall) for different values of thresholds. The average AUPRC over all classes or nodes in the tree to used summarise the overall performance.

# 4 RESULTS

## 4.1 Qualitative Results

*4.1.1 Breakdown of the Food Label Hierarchy.*

**Region:** The dishes are primarily divided into North Indian and South Indian cuisine, with 105 and 32 dishes, respectively. There are also a small number of dishes that are pan-Indian or specific to other regions of India.(See Figure 6 in Appendix)
**Dietary restrictions:** The dishes are also divided into vegetarian and meat dishes, with 102 and 35 dishes, respectively. (See Figure 6 in Appendix)
**Subcategories:** The dishes are further divided into subcategories at the third level of the hierarchy. For example, the North Indian dish "Chicken Tikka Masala" is classified as a "Curry" dish, while

**Figure 4: Images of chicken biryani from the IndiaFood26K**

the South Indian dish "Idli" is classified as a "Dosa" dish due to the ingredients used for both dishes. (see Figure 8 in Appendix)

*4.1.2 Datacard.* The IndiaFood26K dataset is a large-scale, multi-label dataset of Indian food images. The pre-annotated dataset contains 35,148 images across 137 data labels, and is 5-6 levels deep in terms of label hierarchy. The dataset was created by collecting non-sensitive static image data about food from a user-generated database within Google Maps. Labels were derived from restaurant menus. Refer to figure 3 for sample of the type of images within our dataset.

The IndiaFood26K dataset is motivated by the need for a large-scale Indian food classification dataset for computer vision projects. The dataset is not publicly available due to the fact that it is user-generated and we want to maintain user data privacy.

The IndiaFood26K dataset contains a number of biases, including:

- **North Indian vs. South Indian bias:** The separation of North Indian and South Indian food was not done with culinary expert feedback, and was reliant on 5 non-expert native Indians who are from similar backgrounds. This could have created a potential bias in annotation, which doesn't encapsulate the perception of the Indian populace.
- **Label hierarchy bias:** The label hierarchy was not verified and validated by experts, and we only consulted one expert. This may have induced selection bias, where our hierarchy would have catered to one perspective on Indian food. This effect will affect the model's semantic knowledge of Indian food, and may have left out important classifications.
- **Label selection bias:** The process in which we refined the labels from 235 to 137 labels could have resulted in a selection bias. This is because the selection of which labels to keep or throw out was done by a small group of people from the lab. This could have resulted in a dataset that does not represent the general perception of Indian food.

*4.1.3 Data trends post-annotation.* Annotating multi-image datasets is challenging because it can be difficult to identify the exact fine-grain label for a dish, especially if the image is blurry or if there are text/watermarks obscuring the food. In our study, we found that the most common reason for annotators to skip an image was that they were unable to identify the exact fine-grain label. Text/watermarks obscuring the food in the image was a close second. This led to a loss of 4-5k images per annotation round.

Of the 22k images that were labeled, annotator confidence and completion of labeling was high. In the first round of annotations (v1), 68% of images had all dishes labeled, 27% had over half labeled, and 4% had less than half labeled. The second round of annotations (v2) showed a similar trend. (see Figure 9 and 10 in Appendix)

The quality of images in the dataset was good, with annotators rating them as good 99% of the time. Annotator confidence was also high, with 93% of annotators rating their confidence as high.(see Figure 11, 12, 13 and 14 in Appendix for more detailed distribution)

The results from the annotation process strongly support the need for multi-label dataset as most Indian food images contain more than one dish. In both rounds of annotations, around 30% of images (7k images) received more than one label. This suggests that Indian food images are often complex and can be difficult to classify under a single label.

**Table 1: Performance on IndiaFood26K test set**

| Method | Backbone | AUPRC |
|---|---|---|
| Food201[32] | GoogleNet | 0.264 ± 0.005 |
| IFC[40] | ResNet50 | 0.276 ± 0.002 |
| SAM | EffNetV2s | 0.293 ± 0.009 |

## 4.2 Indian dish classification model results

*4.2.1 IndiaFood26K Train Test split.* The dataset was preprocessed to resize images and fix orientation before creating train-test splits. Labels that had fewer than twenty examples were removed (8 out of 137 labels). Examples that had no labels left after the previous step were removed (43 out of 23918 examples). Starting with the rarest label, a random example that contained the label was added to the test set. Examples were sampled without replacement until there were at least ten examples for that label. This processes is iterated by moving on to the next rarest label with fewer than ten labels and the process is repeated until every label had at least ten examples in the test set. The remaining examples formed the train dataset. At the end, there were 4775 test examples and 19100 train examples. Figure 19 shows the final distribution of labels in the test and train dataset.

Given our dataset is relatively small compared to the sizes required to train deep neural networks, we used pre-trained convolution neural network (CNN) backbone and fine-tuned on our target dataset. Three popular pretrained backbones were used to fine-tune the models for food classification: GoogleNet[1], ResNet[2] and EfficientNetV2s[3]. For each of these CNN architecture, ImageNet-pretrained weights were used and a linear layer of size 137 was added at the top. We use the standard binary cross-entropy loss as outlined in Section 3.4.1. We do a 80-20 train-validation split on the aforementioned train dataset and perform a grid-search on the hyperparameters based on the validation loss. More details on the exact training and grid-search procedure can be found in Appendix.

Table 1 presents the models' performance on IndiaFood26K test set using only multi label annotation for training. Macro average

**Table 2: Performance on Food201 test set**

| Method | Backbone | AUPRC |
|---|---|---|
| Food201 | GoogleNet | 0.489 ± 0.001 |
| IFC | ResNet50 | 0.519 ± 0.002 |
| SAM | EffNetV2s | 0.543 ± 0.005 |

**Table 3: Mean IoU between labels from annotation 1 and 2**

| Depth | Leaf | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| IndiaTree | 0.59 | 0.74 | 0.81 | 0.89 | 0.93 |
| Randomized Leaves | 0.59 | 0.62 | 0.66 | 0.79 | 0.90 |
| Height | Leaf | 1 | 2 | 3 | 4 |
| IndiaTree | 0.59 | 0.76 | 0.81 | 0.87 | 0.92 |
| Randomized Leaves | 0.59 | 0.61 | 0.66 | 0.75 | 0.8 |

AUPRC is used to evaluate the model. Performance is not satisfactory in all three methods. There is an increasing trend in performance with more recent backbones. To gauge a better understanding of the relative performance, we train models on Food201. Food201 is a multilabel dataset which is similar to our IndiaFood26K dataset in characteristics. Table 2 presents the results on the official Food201 test set. We note that the performance on Food201 is higher than IndiaFood26K.

We hypothesise that the poor performance on IndiaFood26K may be due to the intrinsic difficulty of classifying Indian food at a fine-grained level. We qualitatively describe two cases below highlighting the general difficulty in classifying the images:

- Many Indian dishes look visually similar but belong to different classes of food. Figure 18 shows similar looking images that were unanimously labelled differently.
- Many Indian dishes have large visual variations. Figure 17 shows images that look different but were unanimously labelled identically.

To verify the above hypothesis, second round of data annotation was performed on the IndiaFood26K using the same annotation protocol but with different annotators. Intersection over Union (IoU) was used on the two label sets to measure the consistency across annotations. Table 3 presents the results for IoU across the two annotation rounds. The IoU is around 0.59 at leaf nodes which indicate that human raters do not agree on the fine grained labels in more that 40% of cases. However, we observe better agreements on labels per image on common coarser grained labels. This motivates incorporating hierarchy in our model to make better mistakes and fail gracefully. Adding internal nodes may provide coarser labels of the fine grained labels to be represented in the classification. The hierarchy relations among these nodes also provides a way to leverage sibling relationships or patterns while learning. To train our hierarchical solution, we expanded the ground-truth set to include the internal nodes of the hierarchy as described in Section 3.4.1. The table 4 shows the results on three methods with non-leaf nodes on both IndiaFood26K and Food201 dataset.

We note the degradation in AUPRC after incorporating hierarchy. This may simply be due to addition of extra labels. The addition of hierarchy degrades performance of rare classes much more than common classes as shown in Table 4. We hypothesize that this may be due to classes recieving different effective weights; the more common internal coarse-grained labels might aid in learning representations of the children and thereby increase the effective weight of the children.

CHAMP loss function described in the section x3.4.1 (methods) is used to incorporate the hierarchical relation among the Indian food dishes more effectively. CHAMP uses the hierarchy as tree structure to compute the loss function. Table 5 shows the improvement in classification performance of CHAMP loss function in comparison to plain binary cross-entropy loss across all three backbones used in the experiments. CHAMP owing to hierarchy-aware loss may make better mistakes which is reflected in better performance of non-leaf node classification in the tree structure. The CHAMP shows the largest increase in performance relative to baseline loss function at leaf level. This may be attributed to larger gain in performance of nodes which have lesser training examples in the dataset leading CHAMP to perform well at fine-grained classification. However, the classification AUPRC attains good performance at higher levels indicating practical utility of the method for coarse grained classification.

The gain in CHAMP classification performance can be attributed to the hierarchy tree and multi label annotation. We conduct ablation experiments to show the value of the hierarchy tree construction on classification performance in table **??**. We note considerable degradation in the performance of internal nodes with the random hierarchy, but with a marginal improvement in the performance at leaf nodes. Random hierarchy at each level is created by randomly shuffling the leaf nodes while keeping rest of the structure intact. As seen, the structure of the hierarchy tree is an important design aspect of the algorithm and needs to be constructed according to use case.

## 4.3 Indian dish identification using Multimodal Large Language model

Recent advances in Multimodal Large Language Models (MLLMs) research is revolutionizing how we interact with technology. These models extend beyond the conventional text-based interfaces, to understand and generate content across a spectrum of formats including text, images, audio, and video. Gemini 1.0 Pro[1] is one of the popular MLLMs built on top of Transformer decoders [46] that are enhanced with improvements in architecture and optimization to enable training at scale as well as for efficient inference on Google's Tensor Processing Units.

Gemini was prompted with "classify the image with Indian dishes. Label all if the image has multiple dishes. Output label dish names only." to classify images form IndiaFood26K dataset. Both fine grained and multi label classification capability were analysed by comparing the output of Gemini model with annotator label. 6 shows the results summary for multilabel classification on test set of IndiaFood26K consisting of 4780 images with 6850 human labels. With just prompt engineering, MLLM models can identify India dishes at coarse granularity. The models also do not identify all the dishes present in the image. Further model finetuning is needed to

**Table 4: Performance on Indian food classification using fine and coarse grained labels.**

| Method | Backbone | Micro AUPRC | Macro AUPRC | Micro Top 25 | Micro Bottom 25 |
|---|---|---|---|---|---|
| BCE (no internal) | GoogleNet | 0.358 ± 0.006 | 0.264 ± 0.005 | 0.433 | 0.091 |
| Food201 | GoogleNet | 0.319 ± 0.011 | 0.221 ± 0.011 | 0.406 | 0.071 |
| BCE (no internal) | ResNet50 | 0.362 ± 0.002 | 0.276 ± 0.002 | 0.423 | 0.102 |
| IFC | ResNet50 | 0.351 ± 0.015 | 0.256 ± 0.011 | 0.426 | 0.088 |
| BCE (no internal) | EfficientNetV2S | 0.391 ± 0.009 | 0.293 ± 0.009 | 0.466 | 0.108 |
| SAM | EfficientNetV2S | 0.357 ± 0.06 | 0.256 ± 0.008 | 0.443 | 0.083 |

**Table 5: Classification performance comparison of CHAMP with BCE loss for multi label classification of IndiaFood26K dataset across depths of the hierarchy.**

| Method | Backbone | AUPRC@leaf | AUPRC@1 | AUPRC@2 | AUPRC@3 | AUPRC@4 | AUPRC@5 |
|---|---|---|---|---|---|---|---|
| Food201 + ERM (with internal) | GoogleNet | 0.221 ± 0.011 | 0.867 ± 0.007 | 0.667 ± 0.01 | 0.497 ± 0.013 | 0.393 ± 0.017 | 0.194 ± 0.001 |
| Food201 + CHAMP | GoogleNet | 0.253 ± 0.003 | 0.884 ± 0.004 | 0.71 ± 0.005 | 0.532 ± 0.006 | 0.422 ± 0.006 | 0.225 ± 0.003 |
| ERM (with internal) | ResNet50 | 0.256 ± 0.011 | 0.885 ± 0.005 | 0.697 ± 0.012 | 0.528 ± 0.015 | 0.422 ± 0.014 | 0.229 ± 0.011 |
| CHAMP | ResNet50 | 0.258 ± 0.011 | 0.889 ± 0.006 | 0.708 ± 0.013 | 0.531 ± 0.014 | 0.428 ± 0.016 | 0.23 ± 0.01 |
| ERM (with internal) | EffNetV2 | 0.256 ± 0.008 | 0.891 ± 0.003 | 0.717 ± 0.007 | 0.54 ± 0.005 | 0.431 ± 0.007 | 0.227 ± 0.008 |
| CHAMP | EffNetV2 | 0.269 ± 0.01 | 0.886 ± 0.007 | 0.714 ± 0.011 | 0.549 ± 0.013 | 0.442 ± 0.013 | 0.239 ± 0.009 |

**Table 6: Performance of MLLM on IndiaFood26K**

| Evaluation Method | Labels matched | Coverage |
|---|---|---|
| Fine grained prediction | 866 | 12% |
| Coarse grain prediction | 3300 | 48% |
| No human label match images | 2630 | 54% |
| Partial human label match images | 2070 | 44% |
| Complete human label match images | 80 | 2% |

increase the accuracy of the Indian food classification on GenAI tools.

### 4.4 Result summary

Summary of the experiments and ML model results are as follows

- Multi label classification of Indian food images have low accuracy due to high variations in images as seen with higher inter grader variability among human annotators.
- Incorporating hierarchical relationships among Indian dishes with CHAMP loss yields better classification accuracy for both fine and coarse grained labels.
- Ablation experiments attribute the gain in classification performance to the hierarchy tree information.

## 5 DISCUSSION

Wide-ranging geographical differences in Indian food, including ingredients, cooking techniques, and presentation styles, make complete picture classification extremely difficult. This intricacy is emphasized by our dataset, which is generated from user-uploaded photos of Indian restaurant meals on Google Maps. We do accept, however, that the dataset may not be fully representative of the entire range of Indian cuisine offers due to inherent biases in both dish selection and visual depiction. In this work, we investigate the challenges associated with vision-based categorization in Indian food and suggest possible ways forward. Large-scale, publicly-available statistics are crucial for promoting research and innovation in the Indian food and nutrition industry. Furthermore, considering the variety of possible uses (such as nutrition analysis, recipe development, and food discovery), equitable methods to dataset construction that prioritize fairness and representation are crucial. Adding more information than just picture hierarchies and co-occurrence patterns will greatly improve the categorization of Indian cuisine. Research on nutrition and cooking might be revolutionized by adding comprehensive metadata to datasets from professionals in these domains. The additional visual variety seen in photos of homemade Indian cuisine must also be expressly taken into consideration by machine learning techniques. These photos frequently feature overlapping parts, partly devoured portions, locally sourced food, and muted hues. Widespread image-based meal recording in social media and nutrition apps offers a great chance to address representational and data volume issues. Lastly, while maintaining user privacy, federated learning provides a way to improve food recognition algorithms. Our work on Indian food image classification is an excellent example to look for more accurate classification models. Addressing data bias, using multi-domain information, and using machine learning techniques can advance the field.

## 6 CONCLUSION

In this work, we consider the problem of multi-label India food classification. We created a new multi-label annotated IndiaFood26K dataset consisting of 139 Indian dishes. We improve the state of the art classification performance by utilizing the relationship among labels by creating the hierarchical tree by adding coarse grain dish categories as parent nodes. Our CHAMP loss uses both multi-label and hierarchy classification to improve both fine-grain and coarse grain classification of Indian food images.

# REFERENCES


[1] 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL]

[2] Sourav Banerjee. 2022. *Indian Food Images Dataset (Kaggle).*

[3] Rodrigo C. Barros, Ricardo Cerri, Alex A. Freitas, and André C. P. L. F. de Carvalho. 2013. Probabilistic Clustering for Hierarchical Multi-Label Classification of Protein Functions. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 385–400.

[4] Wei Bi and James Tin-Yau Kwok. 2011. MultiLabel Classification on Tree- and DAG-Structured Hierarchies. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:18111582

[5] Vishwanath C. Burkapalli and Priyadarshini C Patil. 2019. Food image segmentation using edge adaptive based deep-CNNs. *Int. J. Intell. Unmanned Syst.* 8, 4 (Dec. 2019), 243–252.

[6] Nicolò Cesa-Bianchi and Giorgio Valentini. 2009. Hierarchical Cost-Sensitive Algorithms for Genome-Wide Gene Function Prediction. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology (Proceedings of Machine Learning Research, Vol. 8)*, Sašo Džeroski, Pierre Guerts, and Juho Rousu (Eds.). PMLR, Ljubljana, Slovenia, 14–29.

[7] Soumya Chatterjee, Ayush Maheshwari, Ganesh Ramakrishnan, and Saketha Nath Jagaralpudi. 2021. Joint Learning of Hyperbolic Label Embeddings for Hierarchical Multi-label Classification. arXiv:2101.04997 [cs.LG]

[8] Sachin Chaudhary and Subrahmanyam Murala. 2019. Deep network for human action recognition using Weber motion. *Neurocomputing* 367 (2019), 207–216. https://doi.org/10.1016/j.neucom.2019.08.031

[9] Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. 2019. Hyperbolic Interaction Model For Hierarchical Multi-Label Classification. *CoRR* abs/1905.10802 (2019). arXiv:1905.10802 http://arxiv.org/abs/1905.10802

[10] Xin Chen, Hua Zhou, and Liang Diao. 2017. ChineseFoodNet: A large-scale Image Dataset for Chinese Food Recognition. *CoRR* abs/1705.02743 (2017). arXiv:1705.02743 http://arxiv.org/abs/1705.02743

[11] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition with Graph Convolutional Networks. *CoRR* abs/1904.03582 (2019). arXiv:1904.03582 http://arxiv.org/abs/1904.03582

[12] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. 2019. Multi-Label Image Recognition with Joint Class-Aware Map Disentangling and Label Correlation Embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 622–627. https://doi.org/10.1109/ICME.2019.00113

[13] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition with Graph Convolutional Networks. arXiv:1904.03582 [cs.CV]

[14] Miss Chitnis. 2019. A Study on Scenario of Fast-Food Industry in India. *International Journal of Trend in Scientific Research and Development* Special Issue (03 2019), 88–90. https://doi.org/10.31142/ijtsrd23071

[15] Minki Chun, Hyeonhak Jeong, Hyunmin Lee, Taewon Yoo, and Hyunggu Jung. 2022. Development of Korean Food Image Classification Model Using Public Food Image Dataset and Deep Learning Methods. *IEEE Access* 10 (2022), 128732–128741.

[16] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2017. Food Recognition: A New Dataset, Experiments, and Results. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2017), 588–598. https://doi.org/10.1109/JBHI.2016.2636441

[17] Katie Daisey and Steven D. Brown. 2020. Effects of the hierarchy in hierarchical, multi-label classification. *Chemometrics and Intelligent Laboratory Systems* 207 (2020), 104177. https://doi.org/10.1016/j.chemolab.2020.104177

[18] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. 2011. Hierarchical annotation of medical images. *Pattern Recognition* 44, 10 (2011), 2436–2449. https://doi.org/10.1016/j.patcog.2011.03.026 Semi-Supervised Learning for Visual Content Analysis and Understanding.

[19] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. 2019. Learning a Deep ConvNet for Multi-label Classification with Partial Labels. *CoRR* abs/1902.09720 (2019). arXiv:1902.09720 http://arxiv.org/abs/1902.09720

[20] Charles N. C. Freitas, Filipe R. Cordeiro, and Valmir Macario. 2020. MyFood: A Food Segmentation and Classification System to Aid Nutritional Monitoring. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 234–239.

[21] Bin-Bin Gao and Hong-Yu Zhou. 2020. Multi-Label Image Recognition with Multi-Class Attentional Regions. *CoRR* abs/2007.01755 (2020). arXiv:2007.01755 https://arxiv.org/abs/2007.01755

[22] Jun Harashima, Yuichiro Someya, and Yohei Kikuta. 2017. Cookpad Image Dataset: An Image Collection as Infrastructure for Food Research *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 1229–1232. https://doi.org/10.1145/3077136.3080686

[23] Azra Ismail and Neha Kumar. 2019. Empowerment on the Margins: The Online Experiences of Community Health Workers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300329

[24] Hilde Kuehne, Juergen Gall, and Thomas Serre. 2016. An end-to-end generative framework for video segmentation and recognition. In *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*. Lake Placid.

[25] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. 2016. DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. *CoRR* abs/1606.05675 (2016).

[26] Frank Po Wen Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. 2020. Image-based food classification and volume estimation for dietary assessment: A review. *IEEE J. Biomed. Health Inform.* 24, 7 (July 2020), 1926–1939.

[27] Peihua Ma, Chun Pong Lau, Ning Yu, An Li, Ping Liu, Qin Wang, and Jiping Sheng. 2021. Image-based nutrient estimation for Chinese dishes using deep learning. *Food Research International* 147 (2021), 110437. https://doi.org/10.1016/j.foodres.2021.110437

[28] Megan M MacPherson, Kohle J Merry, Sean R Locke, and Mary E Jung. 2022. How Can We Keep People Engaged in the Behavior Change Process? An Exploratory Analysis of Two mHealth Applications. *Journal of Technology in Behavioral Science* 7, 3 (2022), 337–342.

[29] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. 2020. Visual Aware Hierarchy Based Food Recognition. *CoRR* abs/2012.03368 (2020). arXiv:2012.03368 https://arxiv.org/abs/2012.03368

[30] Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical Text Classification with Reinforced Label Assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1042

[31] Javier Marín, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2018. Recipe1M: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *CoRR* abs/1810.06553 (2018). arXiv:1810.06553 http://arxiv.org/abs/1810.06553

[32] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE international conference on computer vision*. 1233–1241.

[33] Bhalaji Nagarajan, Eduardo Aguilar, and Petia Radeva. 2021. S2ML-TL Framework for Multi-label Food Recognition. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 629–646.

[34] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing Subset Accuracy with Recurrent Neural Networks in Multi-label Classification. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/2eb5657d37f474e4c4cf01e4882b8962-Paper.pdf

[35] Sasmita Nayak, Mamata Beura, Mohammed Siddique, and Siba Mishra. 2021. Analysis of Indian Food Based on Machine learning Classification Models. *Journal of Scientific Research and Reports* 27 (07 2021), 1–7. https://doi.org/10.9734/JSRR/2021/v27i730407

[36] Paritosh Pandey, Akella Deepthi, Bappaditya Mandal, and Niladri B. Puhan. 2017. FoodNet: Recognizing Foods Using Ensemble of Deep Networks. *CoRR* abs/1709.09429 (2017). arXiv:1709.09429 http://arxiv.org/abs/1709.09429

[37] Lee Chang-Ho Jeong Nanoom Cho Young-Im Lee Hae-Jeung Park Seon-Joo, Palvanov Akmaljon. 2019. The development of food image detection and recognition model of Korean food for mobile dietary management. *nrp* 13, 6 (2019), 521–528. https://doi.org/10.4162/nrp.2019.13.6.521 arXiv:http://www.e-sciencecentral.org/articles/?scid=1138038

[38] Vishwesh Pillai, Pranav Mehar, Manisha Das, Deep Gupta, and Petia Radeva. 2022. Integrated Hierarchical and Flat Classifiers for Food Image Classification using Epistemic Uncertainty. In *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*. 1–5. https://doi.org/10.1109/SPCOM55316.2022.9840761

[39] Xiwen Qu, Hao Che, Jun Huang, Linchuan Xu, and Xiao Zheng. 2021. Multi-layered Semantic Representation Network for Multi-label Image Classification. arXiv:2106.11596 [cs.CV]

[40] J R Rajayogi, G Manjunath, and G Shobha. 2019. Indian Food Image Classification with Transfer Learning. In *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. 1–4. https://doi.org/10.1109/CSITSS47250.2019.9031051

[41] Brian H. Ross and Gregory L. Murphy. 1999. Food for Thought: Cross-Classification and Category Organization in a Complex Real-World Domain. *Cognitive Psychology* 38, 4 (1999), 495–553. https://doi.org/10.1006/cogp.1998.0712

[42] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. 2006. Kernel-Based Learning of Hierarchical Multilabel Classification Models. *Journal of Machine Learning Research* 7, 59 (2006), 1601–1626. http://jmlr.org/papers/v7/rousu06a.html

[43] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4239–4249. https://doi.org/10.18653/v1/2021.naacl-main.335

[44] Kyoko Sudo, Kazuhiko Murasaki, Tetsuya Kinebuchi, Shigeko Kimura, and Kayo Waki. 2020. Machine Learning–Based Screening of Healthy Meals From Image Analysis: System Development and Pilot Study. *JMIR Form Res* 4, 10 (26 Oct 2020), e18507.

[45] Ashwin Vaswani, Gaurav Aggarwal, Praneeth Netrapalli, and Narayan G Hegde. 2022. All Mistakes Are Not Equal: Comprehensive Hierarchy Aware Multi-label Predictions (CHAMP). arXiv:2206.08653 [cs.LG]

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[47] Mark L Wahlqvist and Meei-Shyuan Lee. 2007. Regional food culture and development. *Asia Pac. J. Clin. Nutr.* 16 Suppl 1 (2007), 2–7.

[48] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2285–2294. https://doi.org/10.1109/CVPR.2016.251

[49] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2019. Multi-Label Classification with Label Graph Superimposing. arXiv:1911.09243 [cs.CV]

[50] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R. Smith. 2016. Learning to Make Better Mistakes: Semantics-Aware Visual Food Recognition. In *Proceedings of the 24th ACM International Conference on Multimedia* (Amsterdam, The Netherlands) *(MM '16)*. Association for Computing Machinery, 172–176.

[51] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven C. H. Hoi, and Qianru Sun. 2021. A Large-Scale Benchmark for Food Image Segmentation. *CoRR* abs/2105.05409 (2021). arXiv:2105.05409 https://arxiv.org/abs/2105.05409

[52] Bing Xu, Xiaopei He, and Zhijian Qu. 2021. Asian food image classification based on deep learning. *J. Comput. Commun.* 09, 03 (2021), 10–28.

[53] Hongxuan Xu and Huanyu Long. 2020. The Effect of Smartphone App–Based Interventions for Patients With Hypertension: Systematic Review and Meta-Analysis. *JMIR Mhealth Uhealth* 8, 10 (19 Oct 2020), e21759.

[54] Hsin-Yen Yen, Grace Jin, and Huei-Ling Chiu. 2023. Smartphone app-based interventions targeting physical activity for weight management: A meta-analysis of randomized controlled trials. *International Journal of Nursing Studies* 137 (2023), 104384. https://doi.org/10.1016/j.ijnurstu.2022.104384

[55] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. 2020. Cross-Modality Attention with Semantic Graph Embedding for Multi-Label Classification. arXiv:1912.07872 [cs.CV]

[56] Peiran Yu and Min Fu. 2021. TPJF: Machine Learning Based Intelligent Prediction of Preference for Japanese Food. In *Proceedings of the 4th International Conference on Advances in Image Processing* (Chengdu, China) *(ICAIP '20)*. Association for Computing Machinery, New York, NY, USA, 158–162. https://doi.org/10.1145/3441250.3441273

[57] Weiyu Zhang, Qian Yu, Behjat Siddiquie, Ajay Divakaran, and Harpreet Sawhney. 2015. "Snap-n-Eat": Food Recognition and Nutrition Estimation on a Smartphone. *Journal of diabetes science and technology* 9, 3 (2015), 525–33.

[58] Lei Zhou, Chu Zhang, Fei Liu, Zhengjun Qiu, and Yong He. 2019. Application of Deep Learning in Food: A Review. *Comprehensive Reviews in Food Science and Food Safety* 18, 6 (2019), 1793–1811.

[59] Biao Zhu, Hongxin Zhang, Wei Chen, Feng Xia, and Ross Maciejewski. 2015. ShotVis: Smartphone-Based Visualization of OCR Information from Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 12, 1s, Article 12 (oct 2015), 17 pages.

[60] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning Spatial Regularization with Image-level Supervisions for Multi-label Image Classification. *CoRR* abs/1702.05891 (2017). arXiv:1702.05891 http://arxiv.org/abs/1702.05891

# A APPENDIX

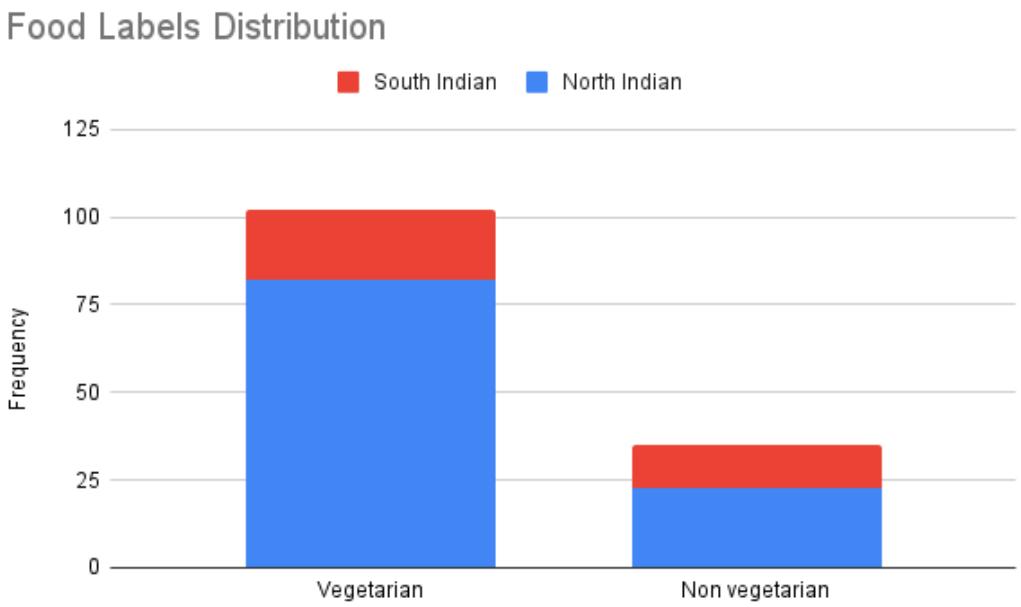


**Figure 5: High level food label distributions in hierarchy.**

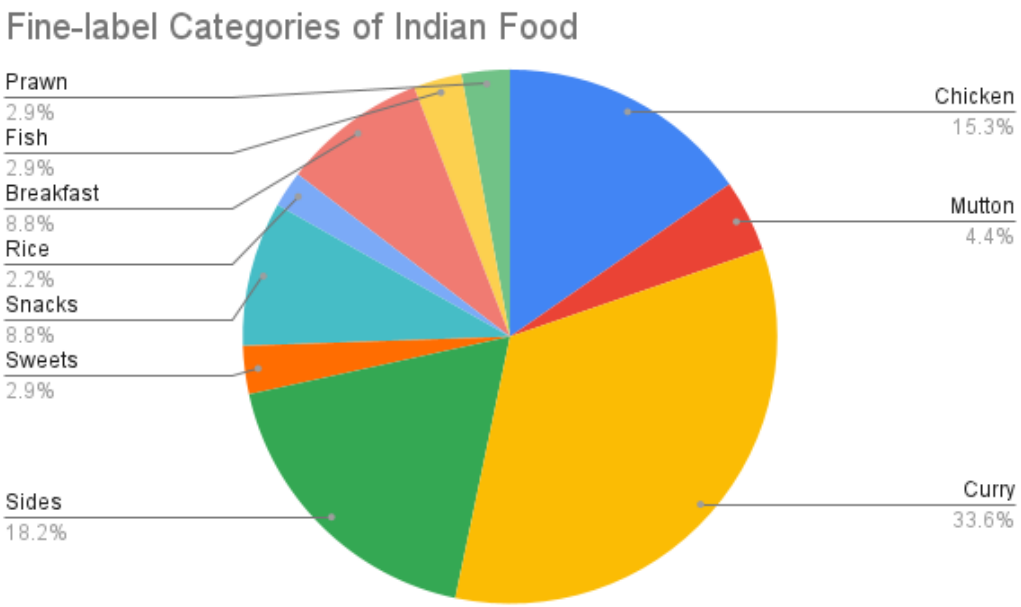


**Figure 6: Fine-level Food Label Distributions in Hierarchy.**

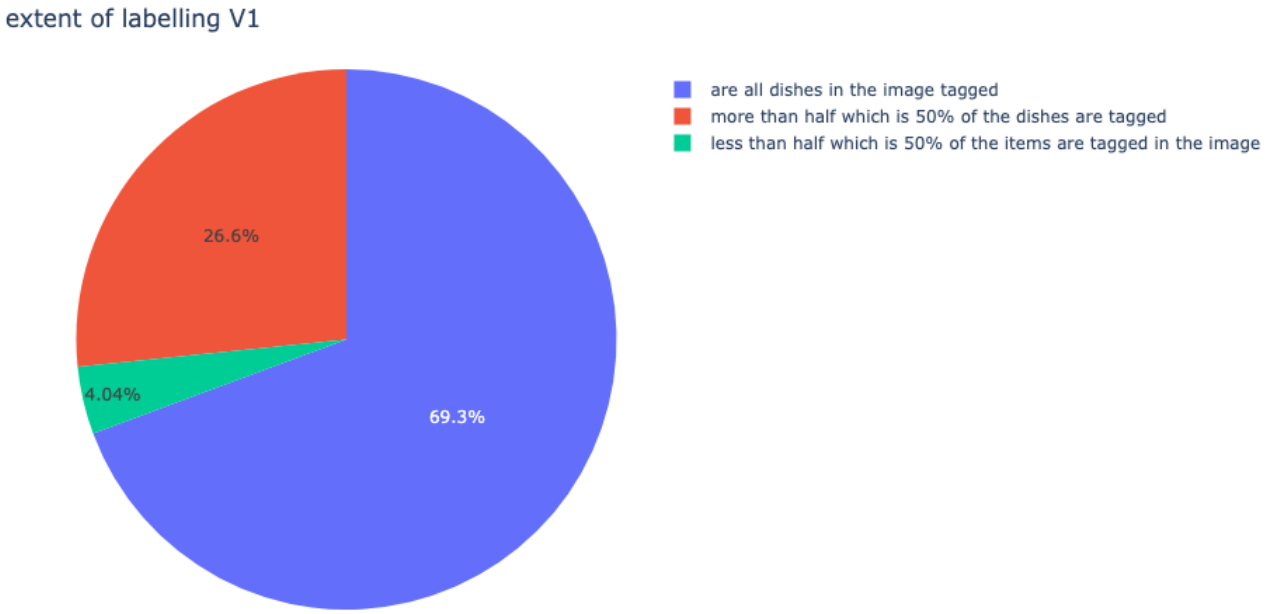


**Figure 7: Fine-level Food Label Distributions in Hierarchy.**

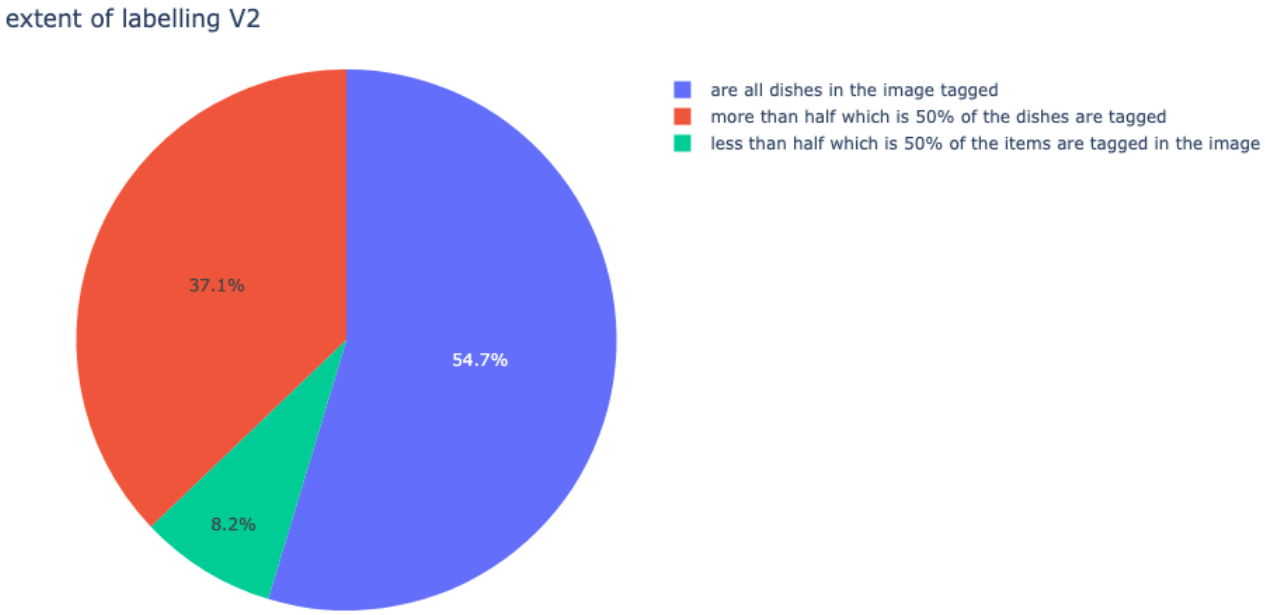


**Figure 8: Fine-level Food Label Distributions in Hierarchy.**

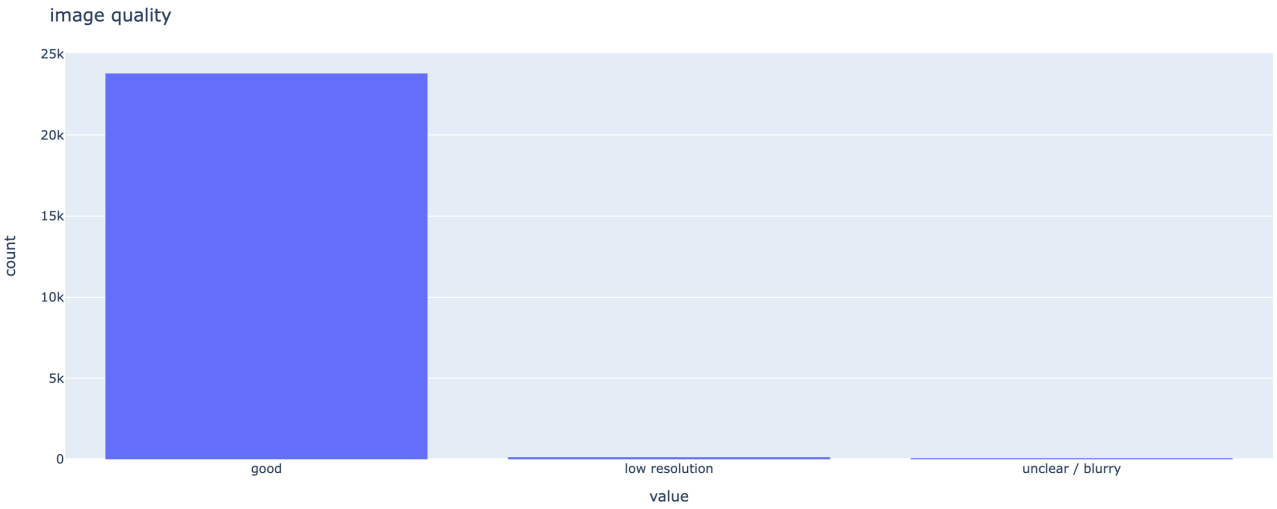


**Figure 9: Fine-level Food Label Distributions in Hierarchy.**

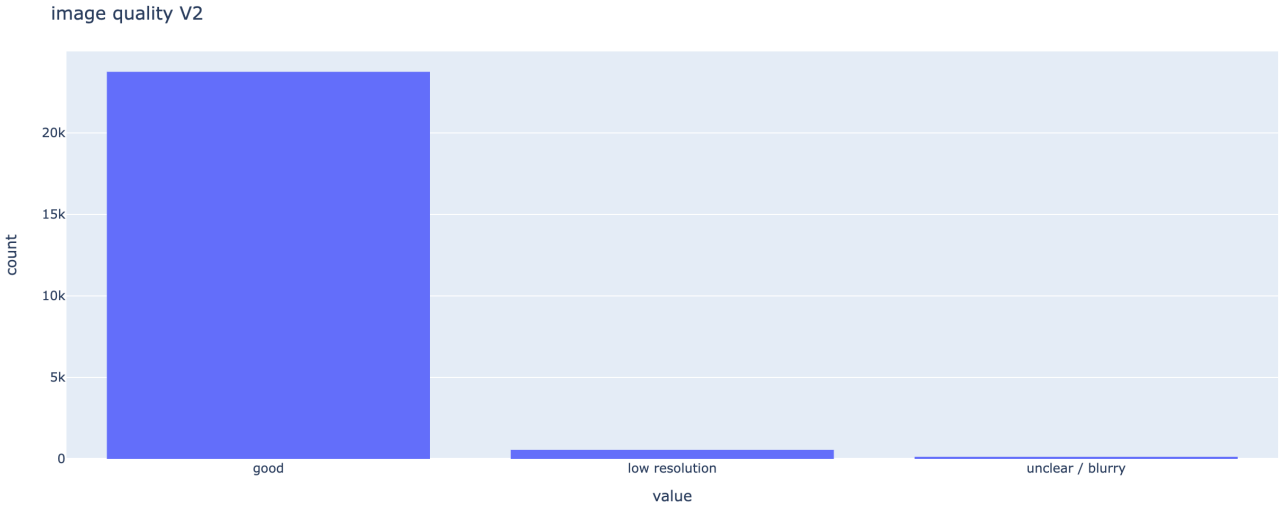


**Figure 10: Fine-level Food Label Distributions in Hierarchy.**

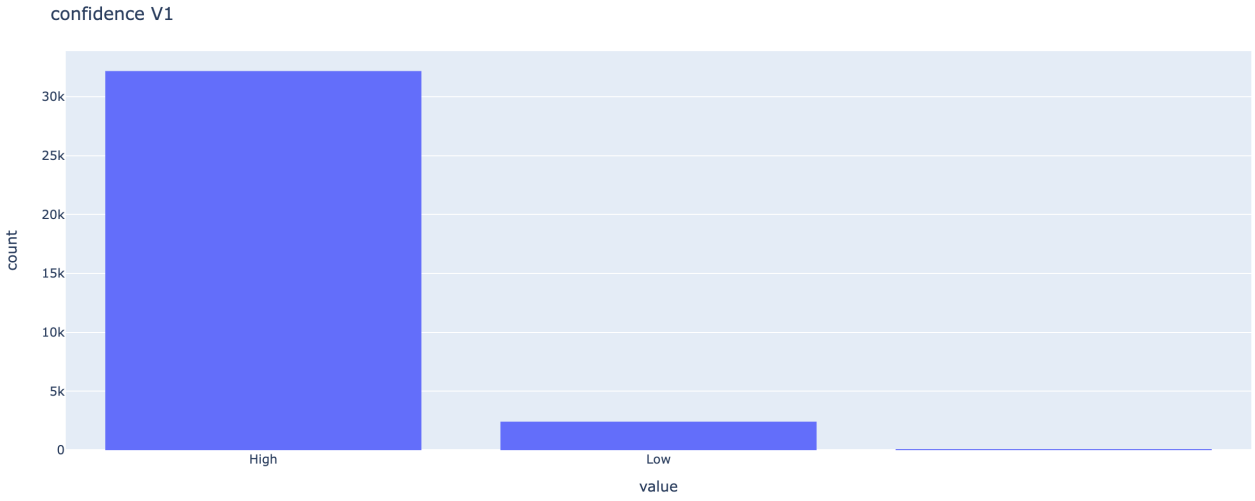


**Figure 11: Fine-level Food Label Distributions in Hierarchy.**

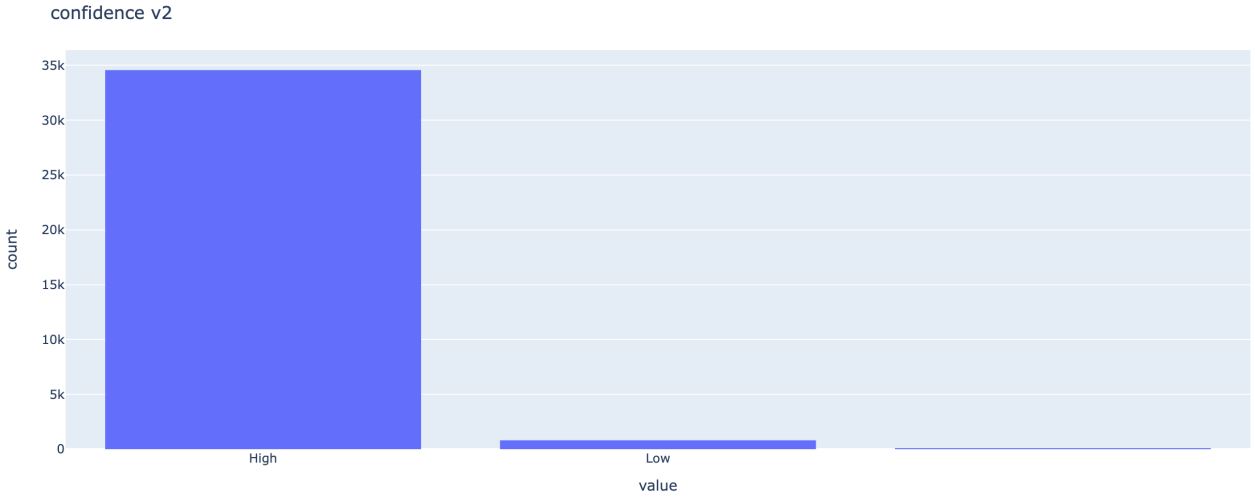


**Figure 12: Fine-level Food Label Distributions in Hierarchy.**

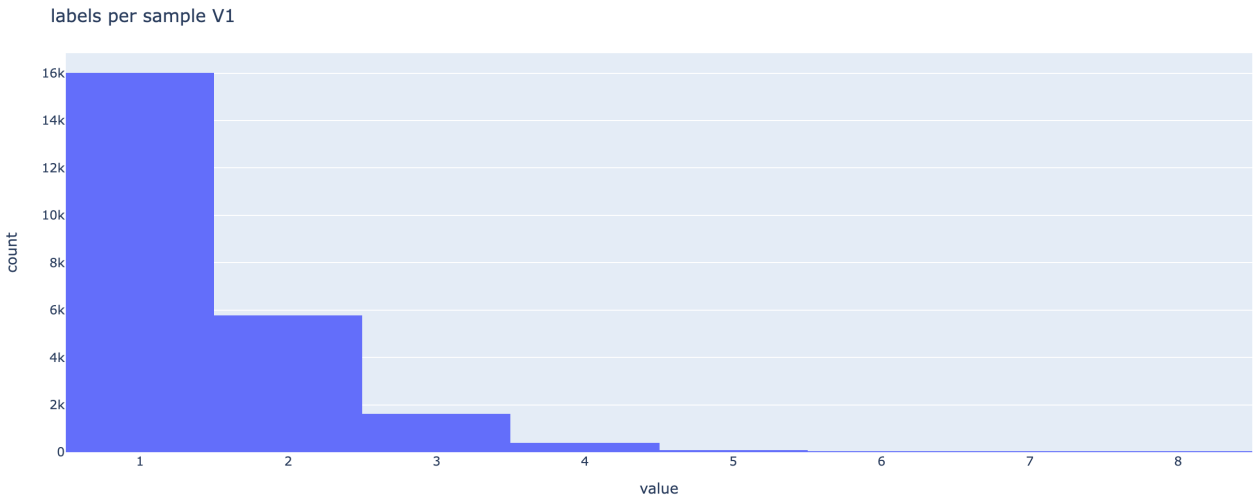


**Figure 13: Fine-level Food Label Distributions in Hierarchy.**

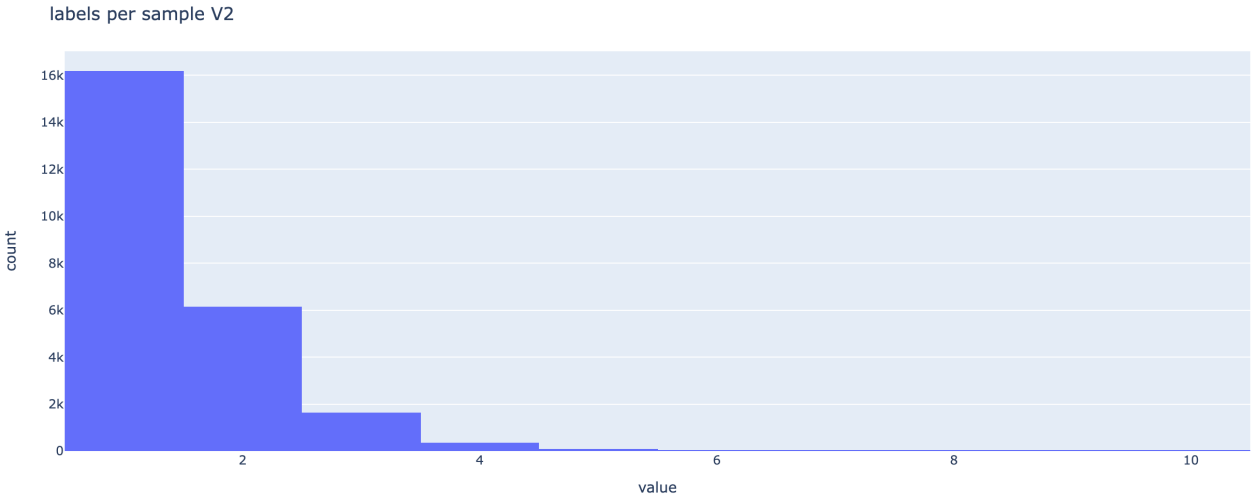


**Figure 14: Fine-level Food Label Distributions in Hierarchy.**

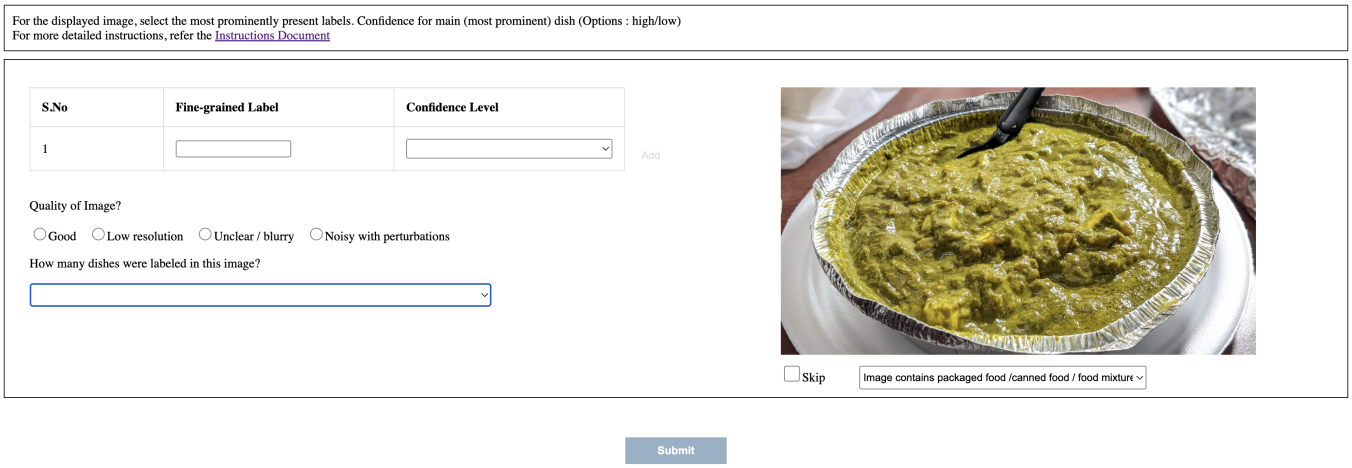


**Figure 15: Annotator UI**

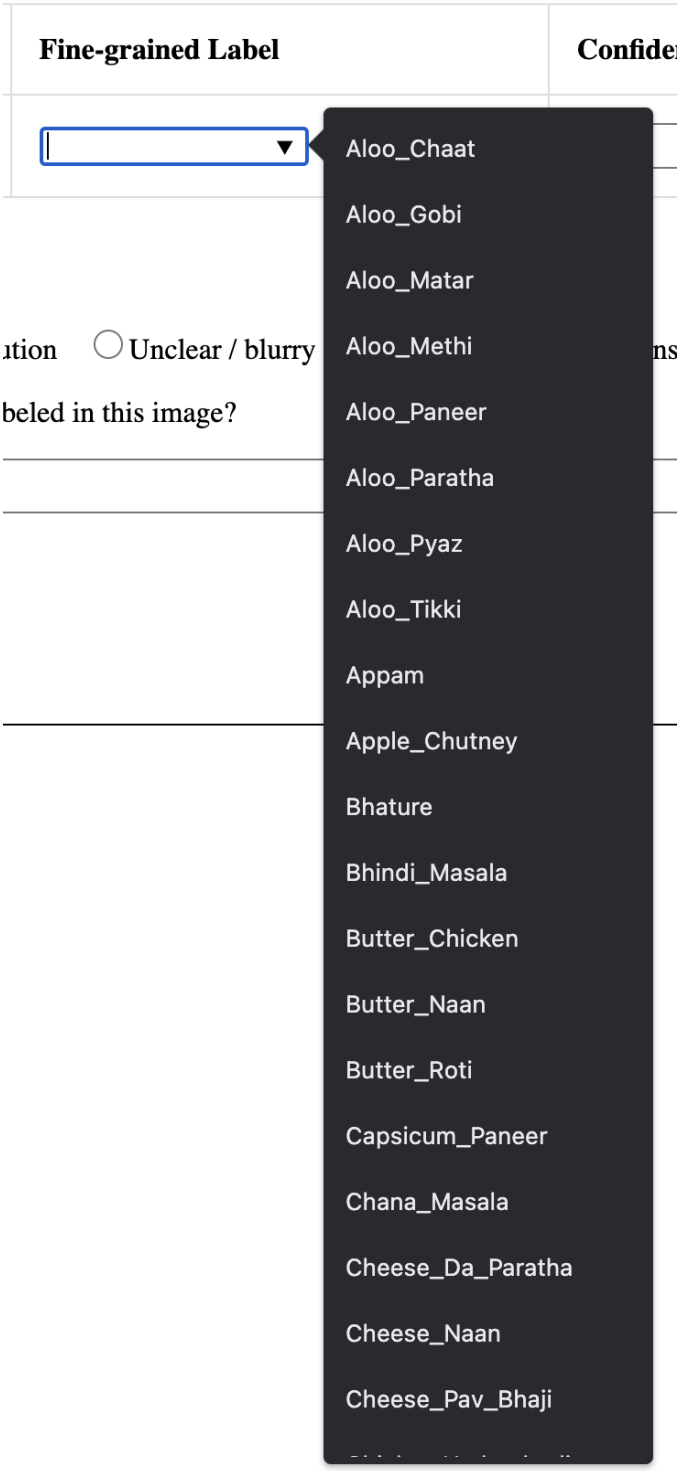


**Figure 16: Annotator UI Dropdown**

**Figure 17: Dishes of same kind (Biryani) has large visual variations among images.**

**Figure 18: Dishes look visually similar but belong to different classes like Biryani, Pulav, Fried rice.**

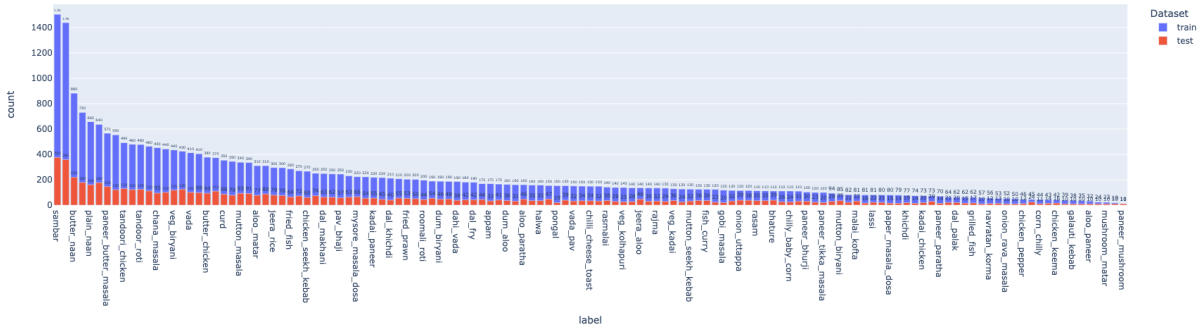


**Figure 19: Frequency of labels in the train and test datasets**