# APS360 FINAL REPORT: USING DEEP LEARNING TO GENERATE INGREDIENTS FROM IMAGES

**Gal Cohen**
Student# 1007757768
gal.cohen@mail.utoronto.ca

**Hannah Kim**
Student# 1008062310
hannahh.kim@mail.utoronto.ca

**Terrence Zhang**
Student# 1008029802
terrencez.zhang@mail.utoronto.ca

## ABSTRACT

Our project introduces a convolutional neural network (CNN) model designed to automate ingredient recognition from food images, facilitating recipe discovery and dietary management. Our model exhibits a promising ability to discern and classify a wide array of ingredients. Through extensive experimentation with various dishes, we illustrate the model's strengths in identifying prominent ingredients and seasoning, while also acknowledging its limitations in distinguishing similar-looking items and processing complex dishes with numerous components. This report further discusses the results of the model's performance on new test data and the associated quantitative results, supplemented by the qualitative results. —-Total Pages: 9

## 1 INTRODUCTION

Our project aims to elevate the culinary experience by developing a deep learning model that can accurately identify and classify ingredients from images of food items. Motivated by the increasing importance of ingredient recognition in the digital age of online food ordering and social media, our goal is to provide a tool that enhances recipe discovery and improves dietary management Banerjee & Chaudhury (2021).

This project's importance lies in the fact that it addresses a real-world challenge that affects a wide range of users, from home cooks to professional chefs. By automating the process of ingredient identification, we aim to streamline culinary exploration and provide valuable information for individuals with dietary restrictions or allergies.

Deep learning, particularly CNNs, is well-suited for this task due to its ability to process and learn from large amounts of image data Rawat & Wang (2017). CNNs excel at recognizing complex patterns in images, making them ideal for identifying the varied and intricate appearances of ingredients in different dishes. Additionally, deep learning models can be trained to recognize a wide range of ingredients, making them more versatile and accurate than traditional image processing techniques. By leveraging the capabilities of deep learning, our project aims to provide a practical and innovative solution for ingredient identification in food images, thereby contributing to the advancement of culinary technology.

## 2 ILLUSTRATION

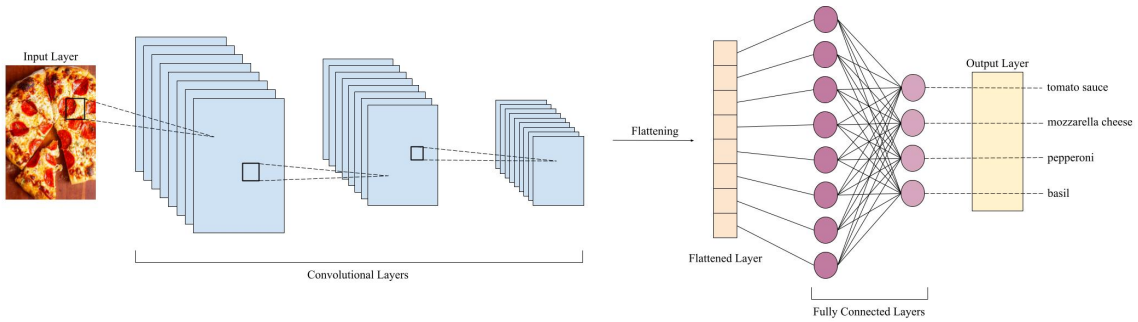


Figure 1: Proposed model architecture of the project.

## 3 BACKGROUND AND RELATED WORK

In the field of ingredient identification and food recognition, several studies have already laid the groundwork for our project. Here are five related works that provide context for our deep learning model:

1. **Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images** (Marin et al. (2019)): This paper introduces a large-scale dataset called Recipe1M+, which contains over one million cooking recipes and corresponding food images. The dataset is used to train deep learning models for various tasks, including ingredient prediction and recipe retrieval.

2. **Food Ingredients Recognition Through Multi-label Learning** (Bolaños et al. (2017)): This study proposes a multi-label learning approach for recognizing food ingredients in images. They present a method that combines CNNs with a specific loss function to enhance the accuracy of ingredient identification. The research highlights the potential of deep learning techniques in the field of culinary technology and dietary management.

3. **Food-101 – Mining Discriminative Components with Random Forests** (Bossard et al. (2014)): This study presents a food recognition dataset called Food-101, which consists of 101 food categories with 1000 images each. The authors use random forests to identify discriminative components in the images, which are then used for classification.

4. **Pic2Recipe** (Salvador et al. (2017)):This project introduces 'im2recipe', a dataset and model that pairs food images with their corresponding recipes. The model is designed to predict the ingredients and cooking instructions based on the input image, leveraging deep learning to bridge the gap between visual food recognition and recipe retrieval.

5. **Hyperspectral Fruit and Vegetable Classification Using Convolutional Neural Networks** (Steinbrener et al. (2019)): This study explores the use of hyperspectral imaging combined with CNNs for the classification of fruits and vegetables. It highlights the potential of integrating advanced imaging techniques with deep learning to improve the accuracy and efficiency of food item classification.

# 4 Data Processing

## 4.1 Data Collection and Initial Structure

Our dataset originates from the comprehensive "Recipe1M+" dataset, which comprises an extensive collection of culinary recipes and associated images. Initially, this dataset includes approximately 13 million images, presenting a significant challenge in terms of storage and computational resources. The volume of data, especially the high-resolution images, demands extensive storage capacity and considerable computational power for processing and model training, making it impractical for our resource-constrained environment.

## 4.2 Ingredient and Recipe Filtering

To manage this challenge, we implemented stringent filtering criteria to reduce the dataset to a manageable size while striving to maintain a balance and retain a diverse set of recipes. This process involved:

**Ingredient Validation:** Leveraging a curated list of 34 common ingredients, we filtered recipes to include only those containing these valid ingredients. This step ensures the relevance and consistency of our dataset, which corrected 611,690 invalid ingredients.

**Recipe Pruning:** We removed recipes that did not meet specific criteria, such as a minimum number of valid ingredients and associated images. This pruning helped us eliminate less informative recipes, contributing to a more focused and balanced dataset. This step led to the removal of 1,020,989 recipes, leaving 8,731 recipes for further processing. For instance, the top three ingredients by count were:

- Oil appeared 9,921 times.
- Water appeared 7,893 times.
- Butter appeared 7,373 times.

## 4.3 Image Handling

Given the initial challenge posed by the vast number of images, we adopted a strategy to substantially reduce the number of recipes (and consequently, images) by increasing the constraints on the recipes included in our dataset. Despite these constraints, we were left with over 30,000 images, offering a substantial volume for training and testing our models. This approach represents a compromise, allowing us to manage the dataset's size while retaining a rich variety of data for model development.

## 4.4 Dataset Preprocessing

We adapted the image normalization technique from the methodology outlined in (Marin et al. (2019)), ensuring uniformity across all images. To maintain consistency for convolutional processing, we resized each image to a standard dimension while preserving their original aspect ratios. Furthermore, we employed image augmentation strategies, which involved randomly altering the colors and rotations of the images to enhance model robustness. Additionally, we introduced a level of randomness by injecting some noise into the images, further diversifying the dataset and simulating real-world variations.

### 4.5 Dataset Summary

After preprocessing, the dataset was summarized as follows:

- Valid Ingredients: 34
- Recipes Retained: 8,731
- Invalid Ingredients Corrected: 611,690
- Invalid Ingredients Removed: 455,555

A cleaned training sample consists of a recipe ID, a list of valid ingredients represented by their indices in the valid ingredients list, and a path to the corresponding image, as seen in figure 2 and the relevant information.

- Recipe ID: ”01d2c5d...”
- Valid Ingredients: [0, 5, 12] (representing water, sugar, and vanilla)
- Image Path: ”/train/0/1/d/2/01d2c5d...”

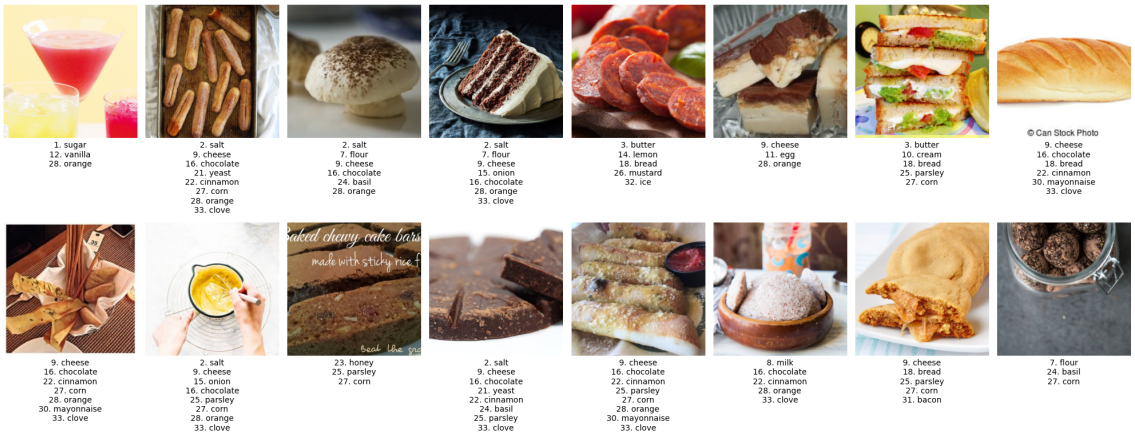


Figure 2: 16 images from the dataset with their corresponding ingredients and indices

### 4.6 Configurability and Balance

Our data processing pipeline is designed with flexibility in mind, allowing us to adjust parameters such as the minimum number of images per recipe or the minimum number of valid ingredients. This configurability enables us to fine-tune the balance between dataset size and diversity, ensuring we have a rich dataset that is both manageable and conducive to training effective models.

## 5 Baseline Model

To evaluate the performance of our model, we will compare the performance of our model against the results of human user studies as a baseline. This is done as the task itself a complex, multi-class image recognition task that is relatively difficult for deep learning models and easy for humans due to the ability to easily recognize context in the images easily and make inferences, in contrast to the deep learning model.

A sample of 10 3rd year Engineering Science students at the University of Toronto were selected for the user studies. The user studies consist of a subset of 10 images from the test set, selected at random. The user is then given a list of the 34 valid ingredients and asked to select any ingredient they believe is part of the recipe of the image. The user results are evaluated based on the accuracy, precision, recall, specificity, and F1 score. By default, an ingredient will not be selected and be counted as a negative label.

The results of our user studies are presented below.


Figure 3: Example of a prompt in the user studies form.

| | | True Class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Class | Positive | 453 | 189 |
| | Negative | 265 | 3012 |

Table 1: Confusion matrix from user study results.

| Metric | Result |
|---|---|
| Accuracy | 0.8842 |
| Precision | 0.7056 |
| Recall | 0.6309 |
| Specificity | 0.9410 |
| F1 | 0.6662 |

Table 2: Performance metrics from the user studies.

One issue with our methodology is that given there are 34 ingredients, most recipes do not use all 34 and thus there is a heavy slant towards the true negatives, which will skew the accuracy higher. Thus, the precision, recall, and F1 score are also being evaluted to get a better understanding of human and model performance when evaluating the results.

Challenges in performing the user studies include the avoidance of bias, as the use of humans in the trial means certain ingredients will tend to automatically be selected when shown in a list, such as salt and oil. The presence of a list also allows participants to guess or infer, which increases their accuracy as they will not be able to put any ingredients they believe may be in the food image but are not present in the list due to the constraints imposed. Nonetheless, these constraints are shared by the model itself, and the ability to guess based on contextual factors is not limited to human participants, but also the model itself, and thus we believe these are reasonable biases to have in our baseline evaluation methods.

To avoid bias in the data set, we selected a diverse group of images that reflect all the ingredients on the list equally to ensure that no ingredient is significantly more likely to appear than others. This is especially the case as oil and salt tend to be staple ingredients that can be guessed easily, and thus ensuring a representative sample test set allows to avoid the issue of certain ingredients being overrepresented and skewing the results.

Another bias is the sample group, which is skewed towards students and young people, who on average will have less experience cooking which may result in lower performance relative to older generations who have had more time to gain experience cooking and will be more familiar with the ingredients of dishes. However, we believe that as a baseline this is acceptable and thus benchmarking our model to these results will give us a reasonably good understanding of the capability of the model.

# 6 ARCHITECTURE

## 6.1 PRIMARY MODEL

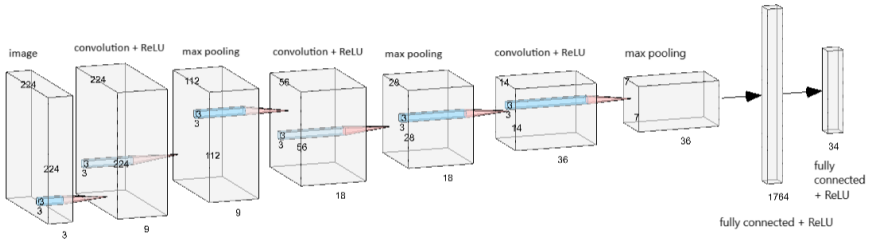


Figure 4: Diagram of our primary model.

Our model is a condensed version of a VGG architecture for multi-class classification. It consists of 3 convolutional layers + ReLU, 3 max pooling layers, and 3 fully connected layers + ReLU. The input images are passed through the convolutional layers, which extract features from the images. Between the convolution layers are max pooling layers, which extract the useful features and reduce the amount of computation needed. The ReLU activation functions are added to the convolutional layers to introduce non-linearity and mitigate the effects of vanishing gradients. Following the the convolutional layers, the outputs are flattened and passed into the 2 fully connected layers with a rectified linear unit, which use the learned features to then classify ingredients and provide the probabilities of each of the 34 ingredients being in the image. The final output is then passed through the sigmoid function to transform the output into a probability.

An ingredient is considered in the image if the probability is greater than 0.5, and not in the image if the probability is less than 0.5.

To train our model, we use the ZLPR loss instead of cross entropy. The ZLPR is designed to handle multi-class tasks where the number of target labels is known, and considers the correlation between labels. According to Su et al. (2022), it has better performance compared to BCE with a similar computational complexity and thus we have selected this loss function for our training.

# 7 DISCUSSION OF QUANTITATIVE RESULTS

| | | True Class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Class | Positive | 167322 | 98340 |
| | Negative | 88976 | 704328 |

Table 3: Confusion matrix from model training.

| Metric | Result |
|---|---|
| Accuracy | 0.8329 |
| Precision | 0.6453 |
| Recall | 0.6647 |
| Specificity | 0.8856 |
| F1 | 0.6549 |

Table 4: Performance metrics from model training.

The results of our initial testing using the given setup and data processing strategy is an accuracy of 83%. The learning curve from the training is shown below. Our best parameters were achieved around epoch

45, with no significant improvement for the following epochs. The validation graph tracks the shape and accuracy of the training data, with a very low error after 30 epochs. This indicates our model is fitted well, and can identify ingredients from the images properly.

The performance of the model is comparable to that of the baseline, with very comparable F1 scores. This result indicates that our models robustness matches that of the average human. The significant number of true negatives does skew the accuracy, but this was expected and accounted for in the data processing, and is mitigated by the use of several performance metrics, like F1 and precision, among others.

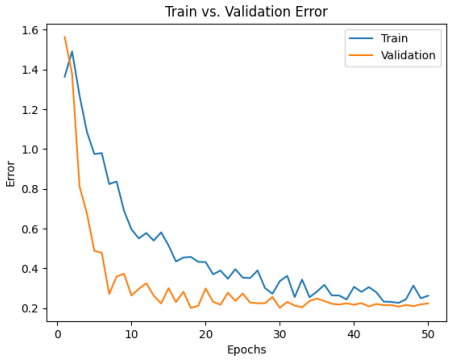


Figure 5: Learning curve for our food image recognition model.

## 8 Discussion of Qualitative Results

(a) Chopped salad
**True labels:** lettuce, tomato, cucumbers, white onion, feta cheese, dressing
**Predicted labels:** lettuce, tomato, cucumbers, dressing

(b) Chipotle chicken wrap
**True labels:** tortilla, chicken, corn, lettuce, tomato, white onion, salt, pepper
**Predicted labels:** tortilla, chicken, corn, lettuce, salt, pepper

(c) Gamjatang
**True labels:** pork, potatoes, green onion, napa cabbage, onions, ginger, salt, pepper
**Predicted labels:** pork, potatoes, green onion, salt, peppers

Figure 6: Sample outputs of the ingredient identification model.

Our deep learning model for ingredient identification has shown promising results in recognizing and classifying a wide range of ingredients from food images. To provide a qualitative analysis of our model's performance, sample outputs were analyzed to highlight its strengths and areas for improvement.

In many cases, our model was able to accurately identify and classify the ingredients in the dish. For example, in images of salads or simple dishes where all the ingredients are clearly visible, the model performs well as seen in Figure 6a. However, it fails to detect the white onion and feta cheese, which could be due to the less distinguishable characteristics of these ingredients.

One of the challenges of the model is when it tries to identify dishes where ingredients are overlapped, like in Figure 6b. The model is able to label most ingredients, but it misses the tomato and white onion. This is most likely because these ingredients are not clearly visible and hidden beneath the layers of the wrap. Yet our model was still able to predict things like salt and pepper even when it is not visible because it understands the connection between those seasonings and other elements of the dish.

When it comes to more complex dishes with multiple ingredients and varying textures, the model was able to identify most primary ingredients, but overlooked finer details. In Figure 6c, the model was able to accurately identify key ingredients of the Korean stew, but it missed the napa cabbage and onions, which are not visible in the broth. This highlights the challenge of identifying ingredients that do not stand out visually.

The fact that the model closely matched the performance of the human baseline suggests that issues that humans face with identifying ingredients are not unique to humans, and is a challenge shared by the model, such as the aforementioned example of visually similar ingredients. Despite similar overall performance between the human baseline and model, we found that the baseline performance tended to a bimodal distribution. This suggests that when shown a dish that is familiar, humans tend to outperform the model. However, when presented a dish for which the participant was not familiar, they tended to perform worse than the model.

## 9 Model Evaluated on New Data

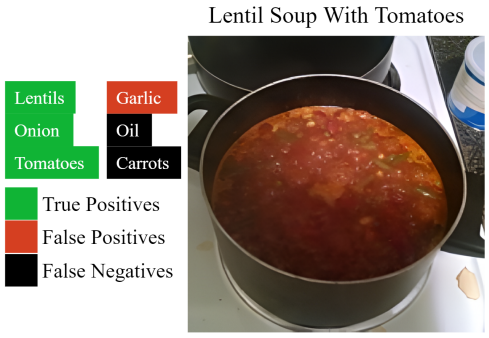


Figure 7: An image taken by one of our group members, including the true and predicted ingredients

First, using our testing set, consisting of approximately 6,000 images and corresponding labels, constituted 20% of our entire dataset. These images were exclusively reserved for the final evaluation phase and were not utilized during the model development stages, including hyperparameter tuning. This testing set was unique in that it comprised images from recipes that the model had never encountered before, thereby providing a fresh and unbiased dataset for evaluation.

Upon evaluating the model on this unseen testing set, we observed a high level of accuracy as seen above, with the model correctly classifying a significant majority of ingredients. The precision and recall metrics also indicated strong performance across the different categories of food items, suggesting that the model is not only accurate but also reliable in its predictions.

Secondly, to assess the model's generalization capabilities and its applicability to real-world scenarios, we independently sourced a new dataset by taking pictures of the food we consumed over the semester. This effort yielded 217 images, which were carefully labeled by our team. This dataset was entirely new to the model and represented a wide variety of real-world conditions under which the model might be expected to operate. While the model maintained a commendable level of accuracy of 81%, there were instances where the variability and complexity of real-world images posed challenges, leading to a slight dip in performance compared to the controlled testing set. However, after examining these images, the model still performed robustly, as some of the incorrectly classified ingredients weren't observable in the images.

## 10 Ethical Considerations

The development and deployment of our deep learning model to identify ingredients from images carry significant ethical considerations that we need to carefully examine. Firstly, the integrity and source of our training data are of paramount importance as discussed by Hussain & Abbas (2023). Our dataset comprises images and associated ingredient lists from diverse cuisines and cultures. It is crucial to ensure that this data is collected and used in a manner that respects the intellectual property rights of the creators of the dataset, or those who have a right to the pictures like chefs, food bloggers, and culinary experts. Moreover, the inclusivity and diversity of our dataset must be scrutinized to prevent cultural bias, ensuring that our model does not favor or under represent any particular cuisine or dietary preference.

## 11 Project Difficulty

The task of ingredient classification from images of food is linked to the diversity and complexity of visual data. The challenge increases when considering the array of possible ingredients that can vary based on factors like preparation, cooking methods, and presentation. This diversity adds a layer of complexity to the classification task, as the model must learn to recognize and differentiate between a wide variety of ingredients with unique visual characteristics.

Our deep learning model, based on CNNs, has been designed to tackle these challenges. Using preprocessing techniques such as resizing to maintain consistency and data augmentation to increase the variability of our training data. Our model has demonstrated its ability to accurately classify ingredients from images taken in real life, showing its ability to handle the diversity and complexity of the data. However, we acknowledge that the performance of our model may vary across different datasets and if there are new ingredients.

## 12 Link to Github Project

`https://github.com/Galc3882/Ingredient-Identifier-APS360`

## REFERENCES

Suprabhat Banerjee and Indrajit Chaudhury. An impact of social media as a promotional platform for food linked products in the new normal. *BULMIM Journal of Management and Research*, January 2021. doi: 10.5958/2455-3298.2021.00006.4.

Marc Bolaños, Aina Ferrà, and Petia Radeva. Food ingredients recognition through multi-label learning. pp. 394–402, 12 2017. doi: 10.1007/978-3-319-70742-6_37.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. pp. 446–461, 2014.

Nasir Hussain and Asad Abbas. Ethical considerations in artificial intelligence and machine learning. 11 2023.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, September 2017. doi: 10.1162/NECO_a_00990.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. 2017.

Jan Steinbrener, Konstantin Posch, and Raimund Leitner. Hyperspectral fruit and vegetable classification using convolutional neural networks. *Computers and Electronics in Agriculture*, 2019.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Zlpr: A novel loss for multi-label classification, 2022.