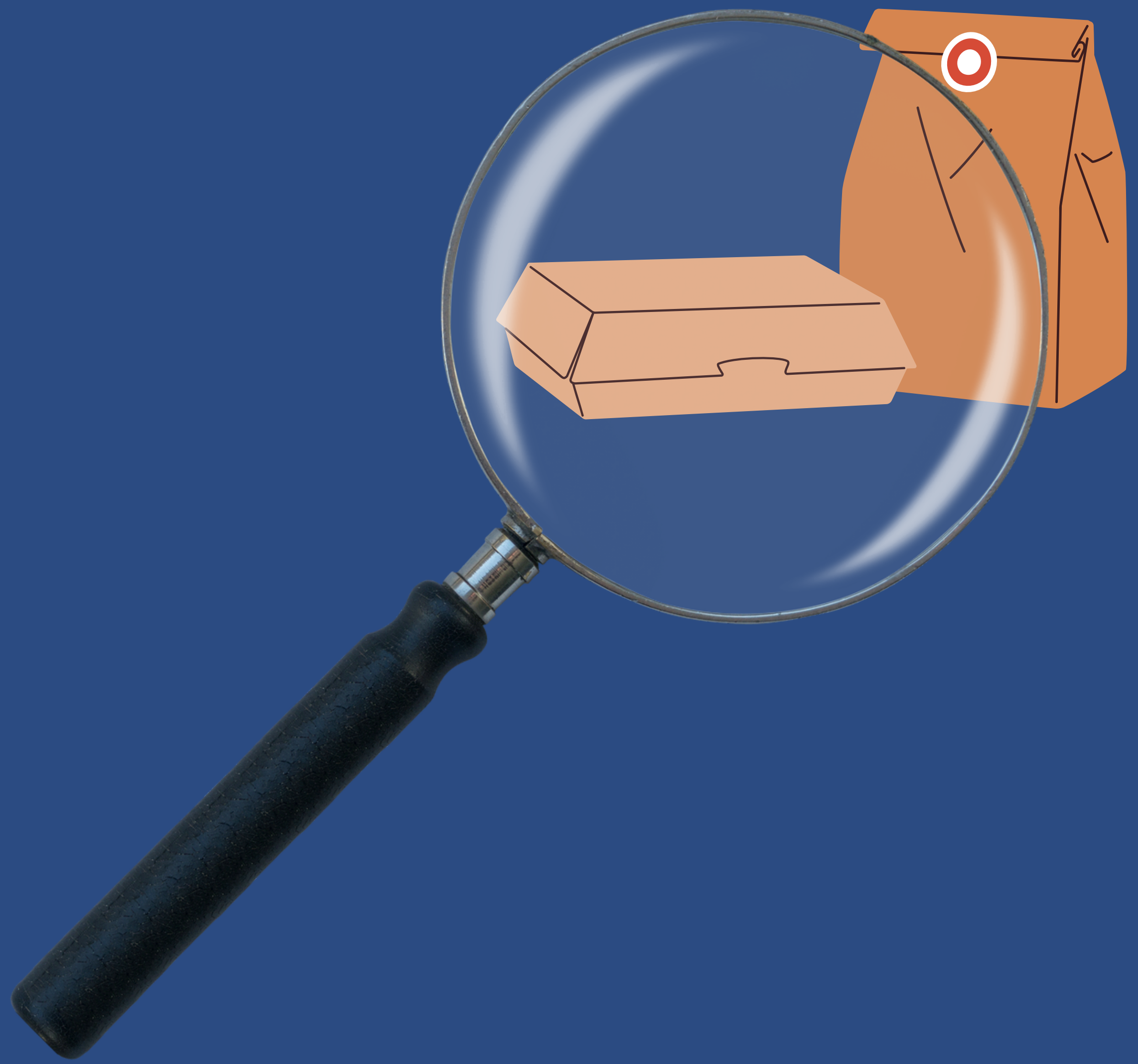
GROUP - 15



AI Powered Allergen Detection System

Custom OCR for food safety - Allergen Detection and Complex Food Ingredients simplifier

Team Members:
Kuldeep Choksi , Harshith Umesh
Ronit Naik , Kiran Deav

Introduction

Problem

- Approximately 32 million Americans have food allergies.
- Ingredient lists are hard to decipher quickly with confusing and hard to decipher ingredients names.

Solution

- Custom OCR for Food Ingredients Recognition using CNNs and Fuzzy Matching, to identify allergens and map complex ingredients to simpler synonyms.

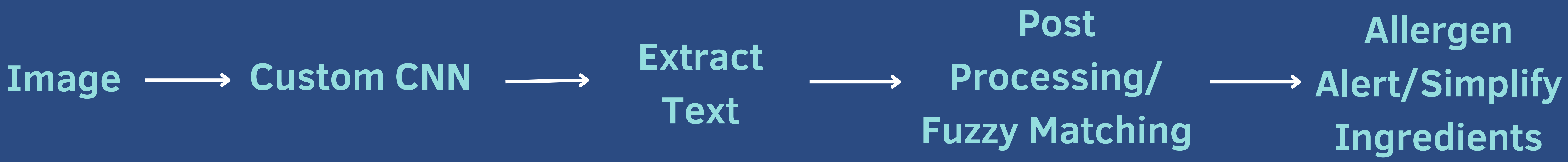


Image → Custom CNN → Extract Text → Post Processing/ Fuzzy Matching → Allergen Alert/Simplify Ingredients

Data Pipeline

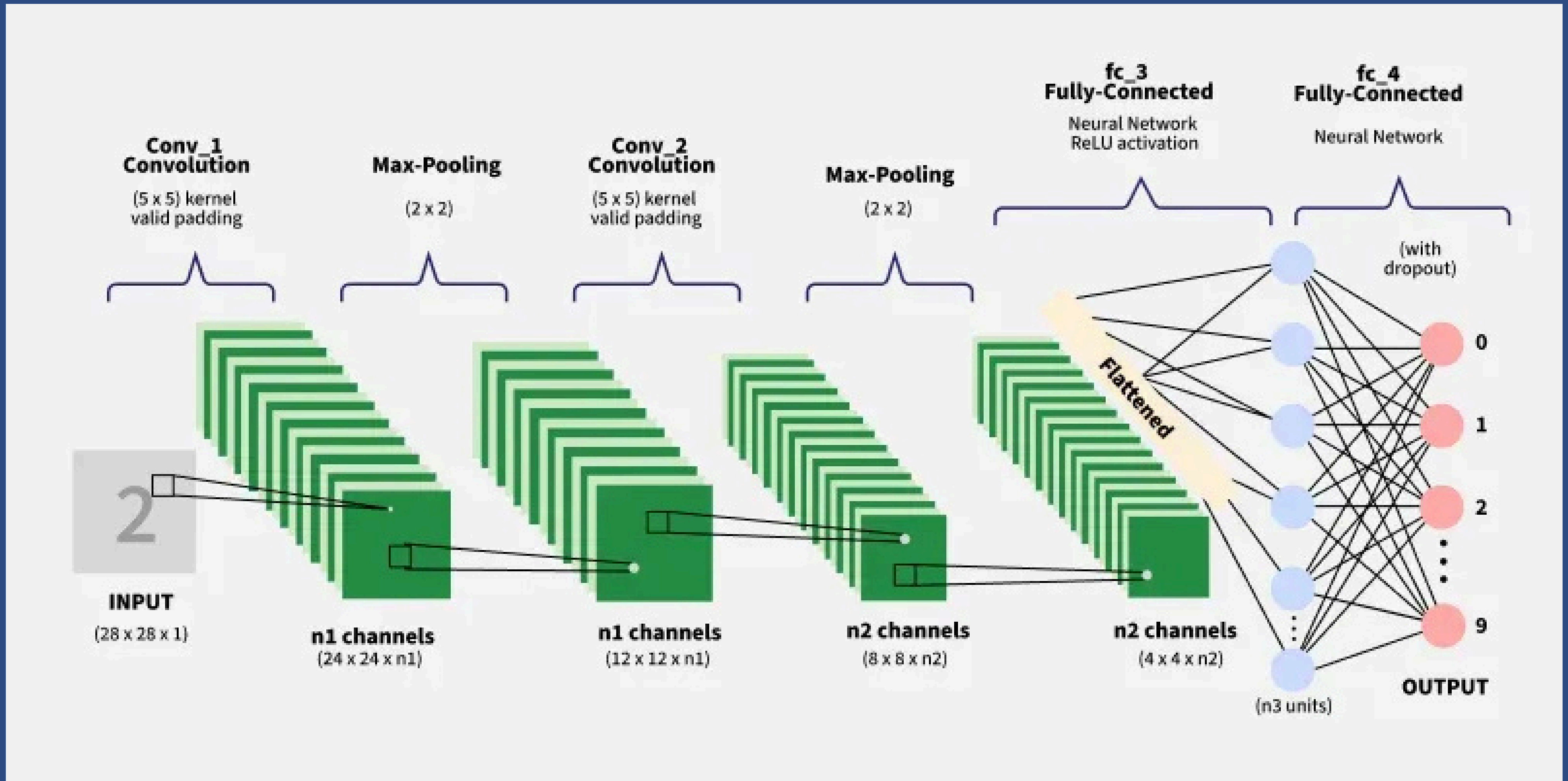
Dataset

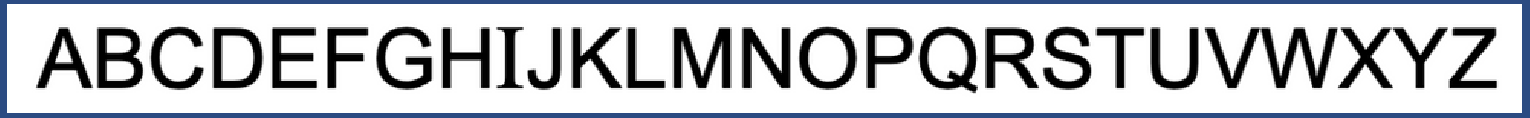
- 174,000+ character images (A-Z, a-z).
- Real-world images and variations with rotations and noise.

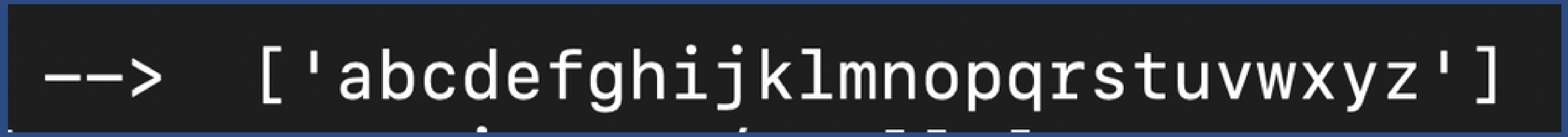
Preprocessing

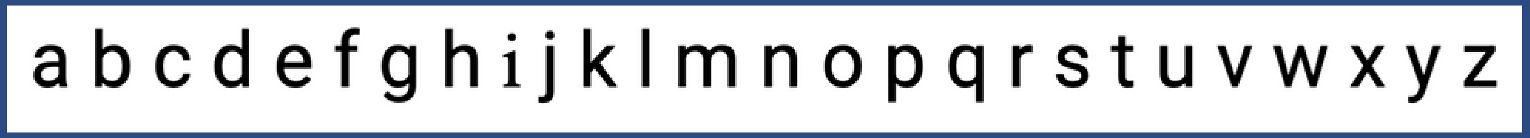
- Grayscaling the images for better detection.
- Binarization using Otsu’s thresholding.
- Edge detection: Dilating image regions to find word boundaries.

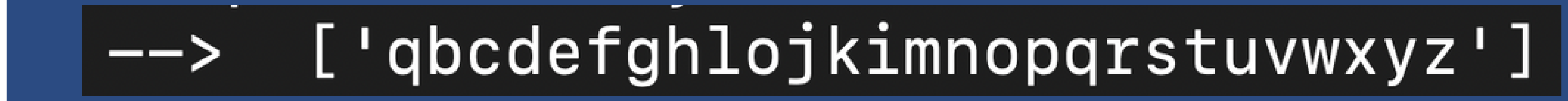
Model Architecture




ABCDEFGHIJKLMNOPQRSTUVWXYZ


```
-->  ['abcdefghijklmnopqrstuvwxyz']
```


a b c d e f g h i j k l m n o p q r s t u v w x y z


```
-->  ['qbcdefghlojkimnopqrstuvwxyz']
```

We achieved an accuracy of 95% for our Custom CNN.

Problems we faced

- Unable to identify punctuation characters.
- Images with coloured backgrounds.
- Initially trained on Macbook-only fonts which was impacting the accuracy.
- The similarity between “I” and “l” and “I”, “c” and “(”, “j” and “i”.
- Mapping OCR extracted words to our two datasets.

Post Processing

Candidate Generation:

- Trimming (removing extra characters).
- Ambiguous Character Replacement (i/l and r/f and i/j).

Fuzzy Matching:

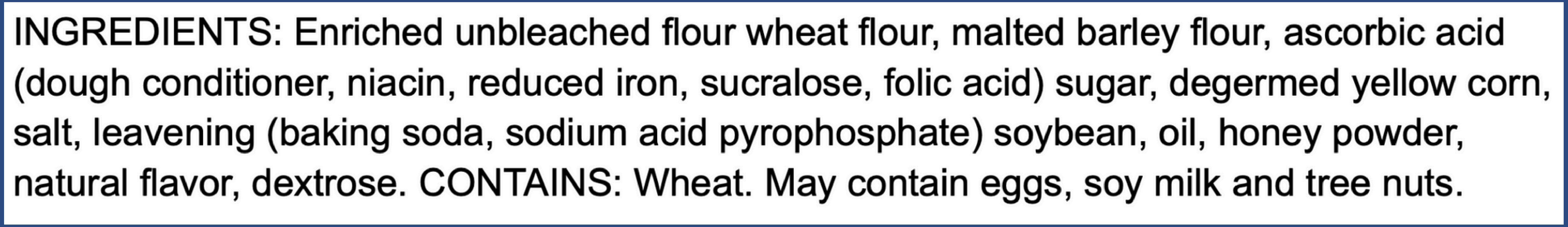
- Using Levenshtein distance for candidate validation.

Allergen and Complex Ingredients Dataset Integration:

- Mapping candidates to a CSV of food ingredients and their simpler synonyms.

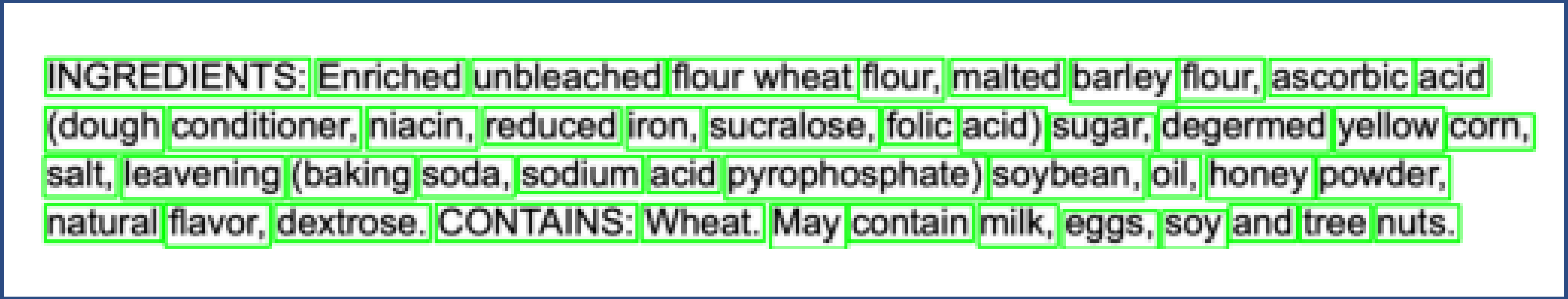
Results

Input Text:



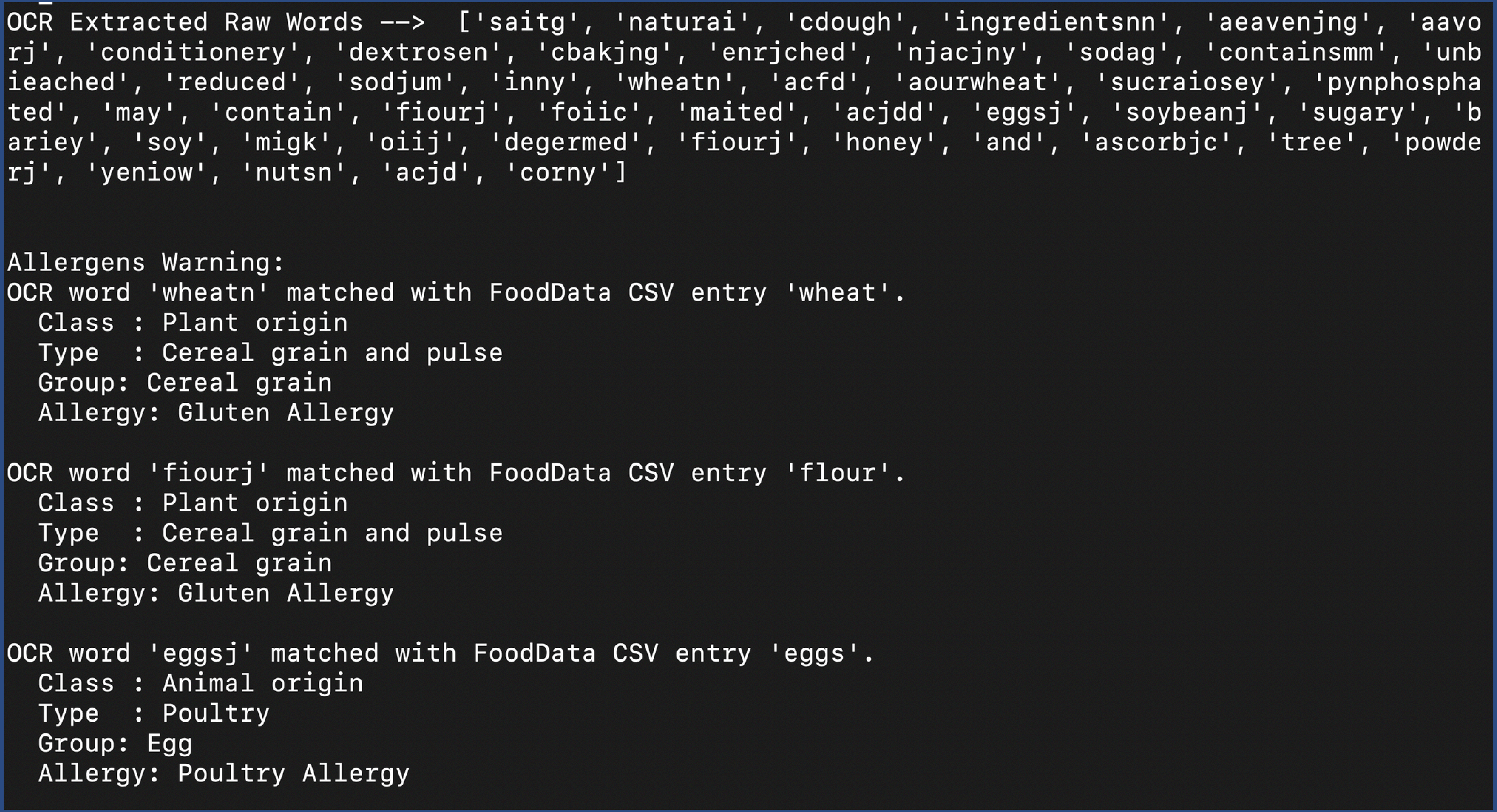
INGREDIENTS: Enriched unbleached flour wheat flour, malted barley flour, ascorbic acid (dough conditioner, niacin, reduced iron, sucralose, folic acid) sugar, degermed yellow corn, salt, leavening (baking soda, sodium acid pyrophosphate) soybean, oil, honey powder, natural flavor, dextrose. CONTAINS: Wheat. May contain eggs, soy milk and tree nuts.

Extract Words using computer vision:



INGREDIENTS: Enriched unbleached flour wheat flour, malted barley flour, ascorbic acid (dough conditioner, niacin, reduced iron, sucralose, folic acid) sugar, degermed yellow corn, salt, leavening (baking soda, sodium acid pyrophosphate) soybean, oil, honey powder, natural flavor, dextrose. CONTAINS: Wheat. May contain milk, eggs, soy and tree nuts.

OCR Output and Allergen Detection



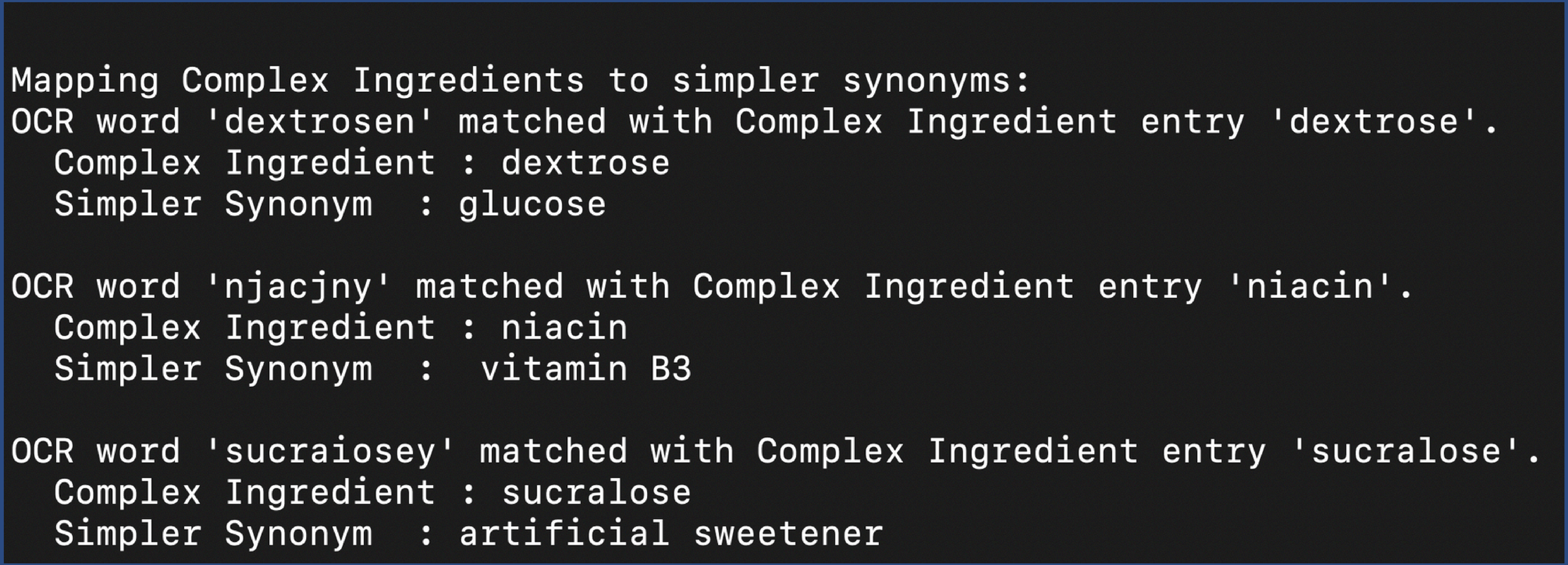
```
OCR Extracted Raw Words -->  ['saitg', 'naturai', 'cdough', 'ingredientsnn', 'aeavenjng', 'aavo
rj', 'conditionery', 'dextrosen', 'cbakjng', 'enrjched', 'njacjny', 'sodag', 'containsmm', 'unb
ieached', 'reduced', 'sodjum', 'inny', 'wheatn', 'acfd', 'aourwheat', 'sucraiosey', 'pynphospha
ted', 'may', 'contain', 'fiourj', 'foiic', 'maited', 'acjdd', 'eggsj', 'soybeanj', 'sugary', 'b
ariey', 'soy', 'migk', 'oiij', 'degermed', 'fiourj', 'honey', 'and', 'ascorbjc', 'tree', 'powde
rj', 'yeniow', 'nutsn', 'acjd', 'corny']


Allergens Warning:
OCR word 'wheatn' matched with FoodData CSV entry 'wheat'.
  Class : Plant origin
  Type  : Cereal grain and pulse
  Group: Cereal grain
  Allergy: Gluten Allergy

OCR word 'fiourj' matched with FoodData CSV entry 'flour'.
  Class : Plant origin
  Type  : Cereal grain and pulse
  Group: Cereal grain
  Allergy: Gluten Allergy

OCR word 'eggsj' matched with FoodData CSV entry 'eggs'.
  Class : Animal origin
  Type  : Poultry
  Group: Egg
  Allergy: Poultry Allergy
```

Complex ingredients list simplification



```
Mapping Complex Ingredients to simpler synonyms:
OCR word 'dextrosen' matched with Complex Ingredient entry 'dextrose'.
  Complex Ingredient : dextrose
  Simpler Synonym  : glucose

OCR word 'njacjny' matched with Complex Ingredient entry 'niacin'.
  Complex Ingredient : niacin
  Simpler Synonym  :  vitamin B3

OCR word 'sucraiosey' matched with Complex Ingredient entry 'sucralose'.
  Complex Ingredient : sucralose
  Simpler Synonym  : artificial sweetener
```

Demo Link:

https://drive.google.com/drive/folders/1JMAX1FDDiB_UeX_ssKL0udnuIF0AnaGb

Recap

- Developed a Custom OCR using CNN to achieve 95 % accuracy.
- 174,000 character images for training, testing and validation.
- Incorporated allergen and complex ingredients synonym datasets.
- We integrate Deep Learning with Heuristic Post-Processing.

Future Scope

- Build a Mobile Application with an easy to use user interface where a user can click a picture and do the food label analysis.
- Expand the allergen dataset.
- Expand the complex ingredient to simple synonym dataset.