Food Desert CNN

Using Satellite Imagery & Machine Learning to Identify Food Deserts

Prerna Singh, Yu-Sheng Lee, Yiwen Hou & Kandace Webber
MIDS 207

UC Berkeley School of Information

Motivation & Problem Statement

🧩 The Problem

- Food deserts are geographic areas with limited access to affordable, nutritious food
- USDA classification rely on outdated Census data (once every ~10 years)

🚧 Limitations of Current Methods

- Traditional metrics don't capture real-time environmental change (recent changes in infrastructure, land use, and store access)
- Manual data collection is time-consuming and incomplete

🌍 Our Goal

Test if satellite imagery and ML Models can improve food desert classification

Research Question & Hypothesis

❓ Research Question

How can satellite imagery and ML models (CNNs) improve the classification of food deserts beyond the traditional metadata-based approaches?

🧪 Hypothesis

Overhead imagery of the environment contains latent features that correlate with food access (including road density, green space, commercial clusters, or housing patterns) that will help identify food deserts.

Related Work

📚 Jean et al. (2016) : Combining Satellite Imagery and Machine Learning to Predict Poverty

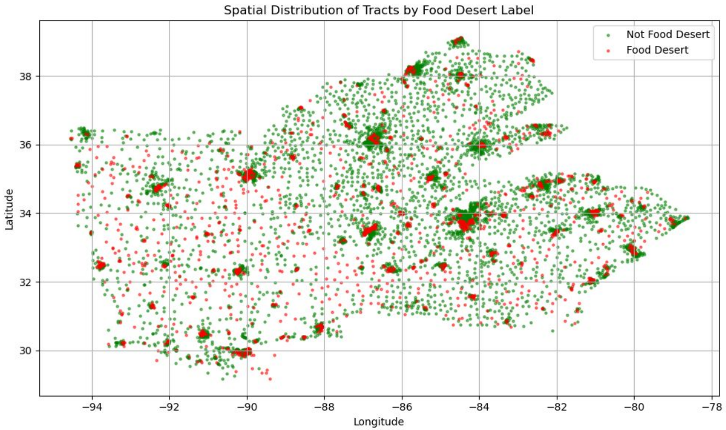
- Poverty Prediction with CNNs on satellite imagery
- Inspired our use of visual patterns and hybrid models

📚 George & Tomer (2021) :

- Critique of “food desert” concept
- Labels are imperfect; our model may capture what current definitions miss

Dataset (Pre-processing & EDA)

- **Input :** 400x400 RGB satellite images (Google Maps) + Metadata
  - **USDA Food Access Atlas** → Binary labels (LILA_Urban1_Rural10 = 1)
  - **Census TIGER shapefiles** → Tract boundaries & centroids
  - **Google Maps API** → 400x400 pixel satellite imagery
  - **Focus:** 8 Southeastern U.S. states
  - **Final dataset:** 7,121 tracts with image + metadata
- **Output:** IsFoodDesert (1 = food desert, 0 = not)

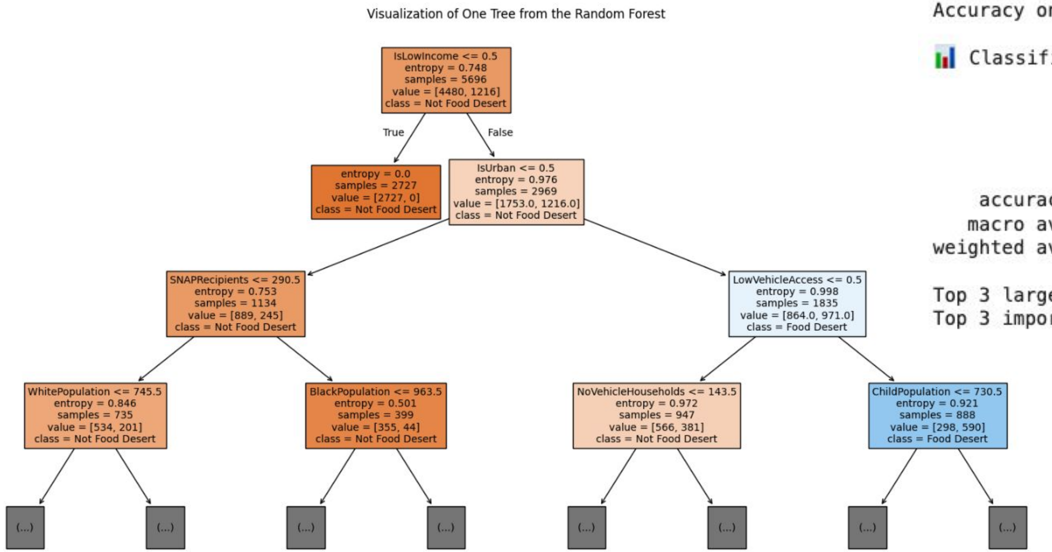


Key EDA findings:

- **Class imbalance:** urban ≠ not food desert
- Food deserts **cluster near city edges**
- Label correlates with **income & rural status**

Model 1: Random Forest

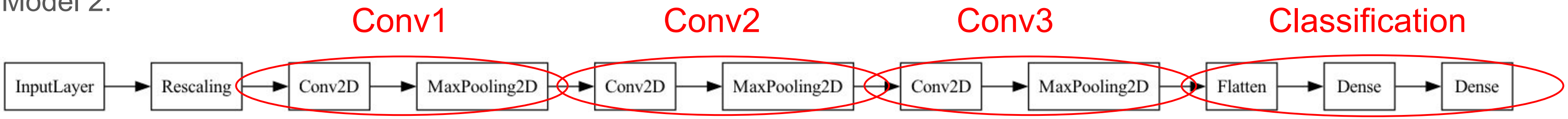
Model 1 : Random Forest (Metadata only)



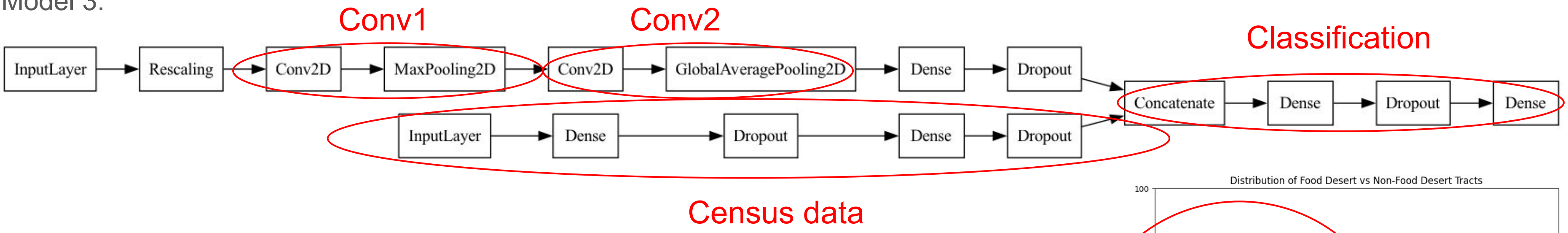

```
Accuracy on training data: 0.89
Accuracy on testing data: 0.82

📊 Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      1078
           1       0.64      0.60      0.62       347

    accuracy                           0.82      1425
   macro avg       0.76      0.75      0.75      1425
weighted avg       0.82      0.82      0.82      1425

Top 3 largest feature importance score: [0.55 0.09 0.08]
Top 3 important features: ['IsLowIncome' 'IsUrban' 'NoVehicleHouseholds']
```

CNNs: Model 2 vs 3

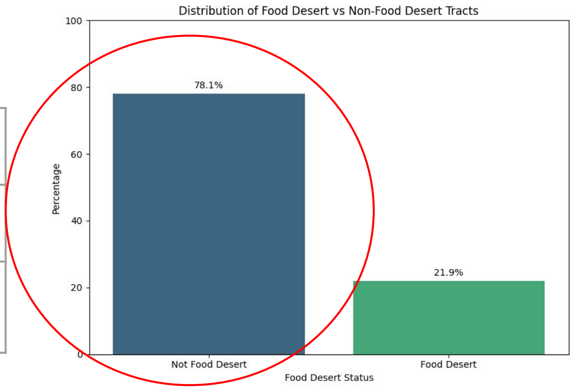
Model 2:



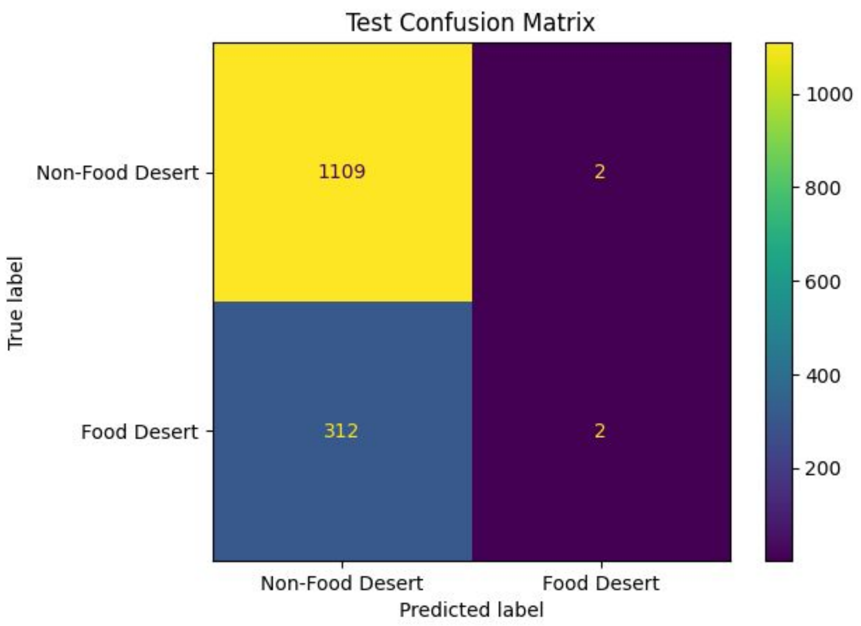
Model 3:

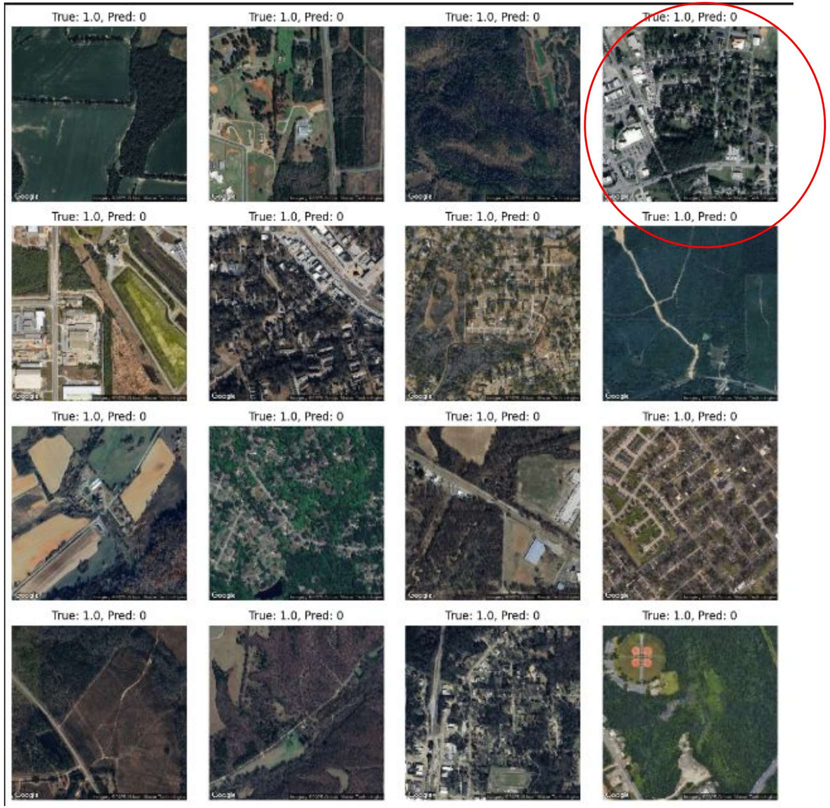


CNN Model	Accuracy	Precision	Recall	F1 Score
2. Image only	78%	0%	0%	0%
3. Image+meta	83%	58%	78%	67%

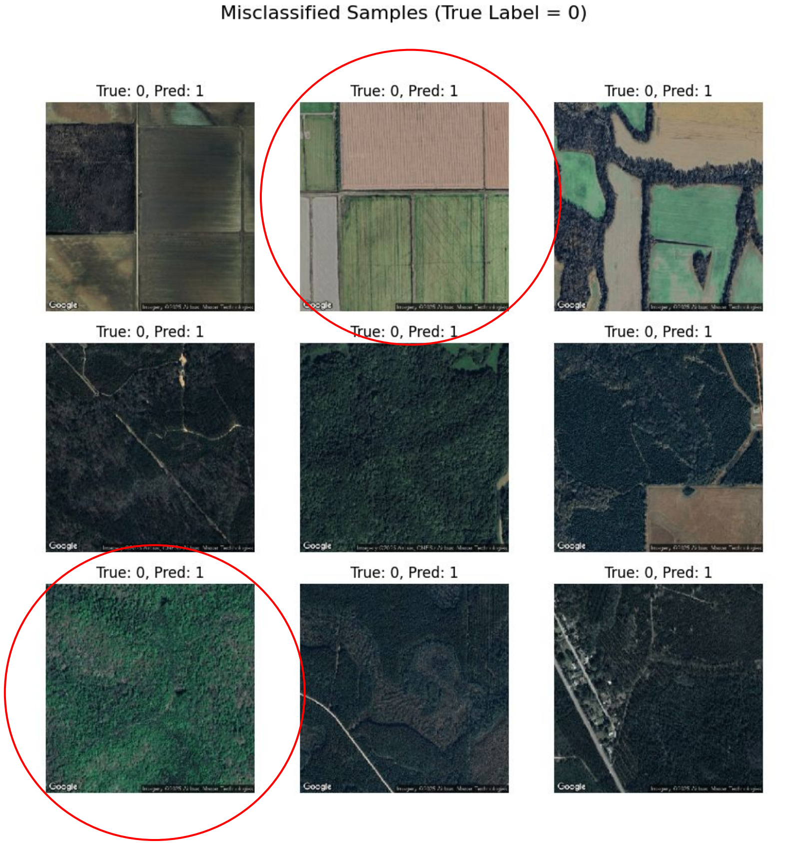


Confusion Matrix and Misclassified Images for Model 2





Misclassified Images for Model 3



Conclusions

✅ **Best performing Model :** CNN+Metadata

🔭 **Future Ideas/Learnings :**

- Did we have the right data?
  - Discrepancy in images + metadata
  - Real-time data sources of grocery store locations/business density/land use classification & zoning data
- Need for model that doesn't rely on government provided data
  - Support local policy and non-profits in food justice and urban planning
- Expand nationally

Table 1. Model Evaluation Metrics

Metrics \ Model	1. Random Forest	2. CNN Images Only	3. CNN Image + Metadata
Test Accuracy	82%	78%	83%
F1 Score	62%	0%	67%
Precision	64%	0%	58%
Recall	60%	0%	78%

Thank you!

MIDS-207

Prerna Singh
Yu-Sheng Lee
Yiwen Hou
Kandace Webber

UC Berkeley School of Information