Improving Fungi Classification with Metadata Integration

Rachel Kalafos, Ryan Farhat-Sabet, William Seward
DATASCI 207, Prof. Livinisky

Table of Contents

01 **Motivation**
\+ Research Question

02 **Data**
Summary Statistics, Preprocessing

03 **Modeling**
Naive → Enhanced

04 **Experiments**
CNN only, CNN + Metadata, FFNN on image data

05 **Conclusions**

1. Motivation

Research Question

*"Can the inclusion of metadata (e.g. elevation, habitat) enhance model performance in classifying fungal **taxonomic class** compared to a model using only image data?"*

Motivation

Fungi classification has applications in…

Biodiversity Tracking

Citizen Science

Conservation Efforts

An "automatic fungi recognition" mobile app is a fun challenge, as it would have to…

Work with ***limited resources***

Learn to classify ***extremely rare*** fungi

Visually ***distinguish similar species***

Initial Approach

1.
Build a baseline CNN model that classifies fungi into **classes** using only image data.

2.
Design a multi-input neural network that also includes spatial and environmental **metadata**.

3.
Use strategies to mitigate class imbalance (image augmentation, downsampling, filtering).

Summary of Results

- The addition of tabular metadata boosts the accuracy of the CNN

- FFNN performs better than a CNN on this particular problem

- Class imbalance proved to be a major challenge

2. Data

Data

Dataset:

- FungiCLEF 2025 (Kaggle Competition)
- 6,391 unique fungal observations, 12,015 total images
- Each observation = 1 fungus found in the wild, often with multiple photos
- Each observation includes **images** + **metadata**

Target Variable:

Taxonomic class (e.g., Agaricomycetes, Leotiomycetes)
- More common and stable across observations

Image Features:

- RGB photos with 300px width

Metadata Features:

- Numerical: latitude, longitude, **elevation**
- Categorical: **habitat**, land cover, substrate, region

Summary Statistics

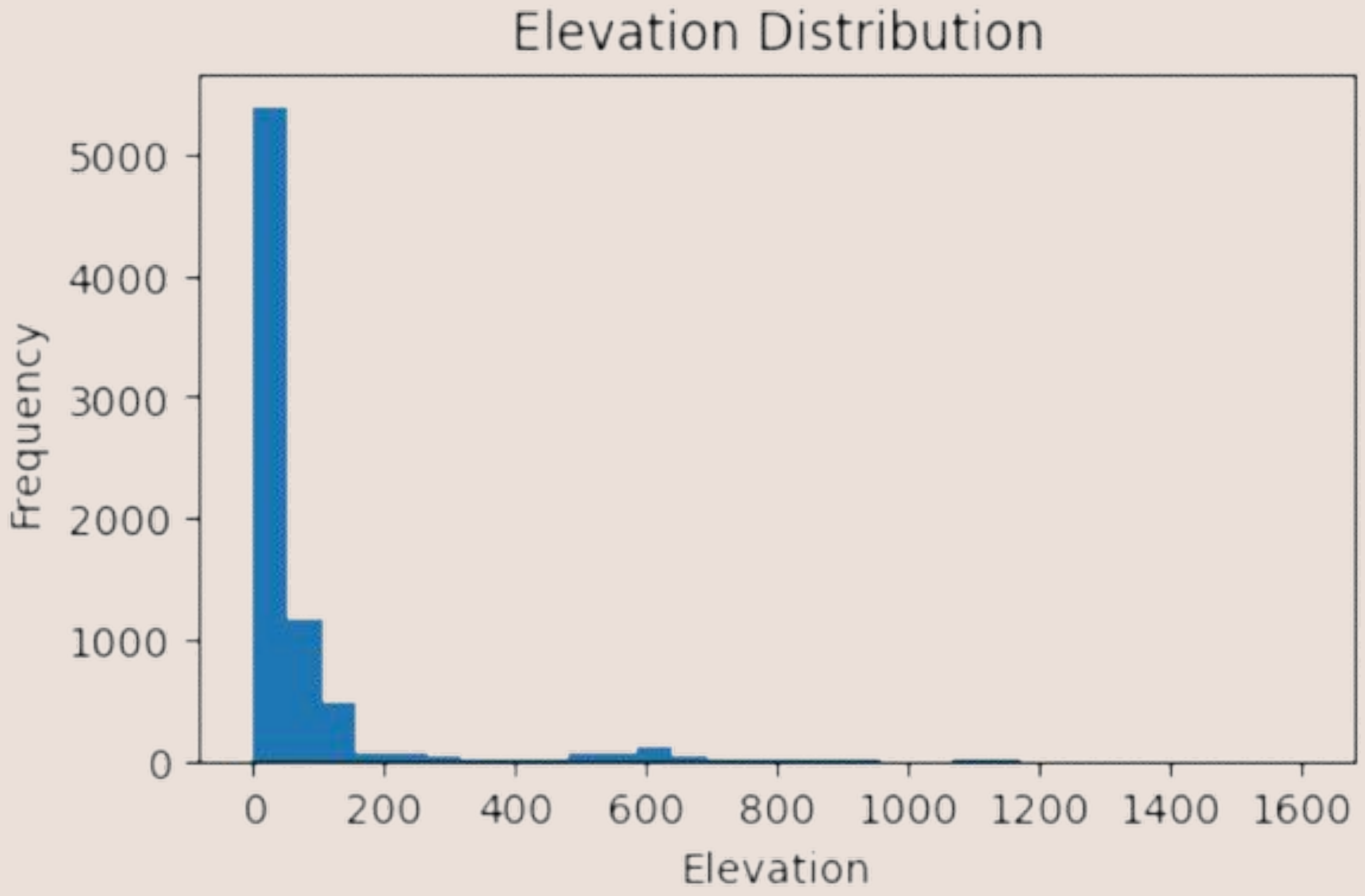
Image Counts

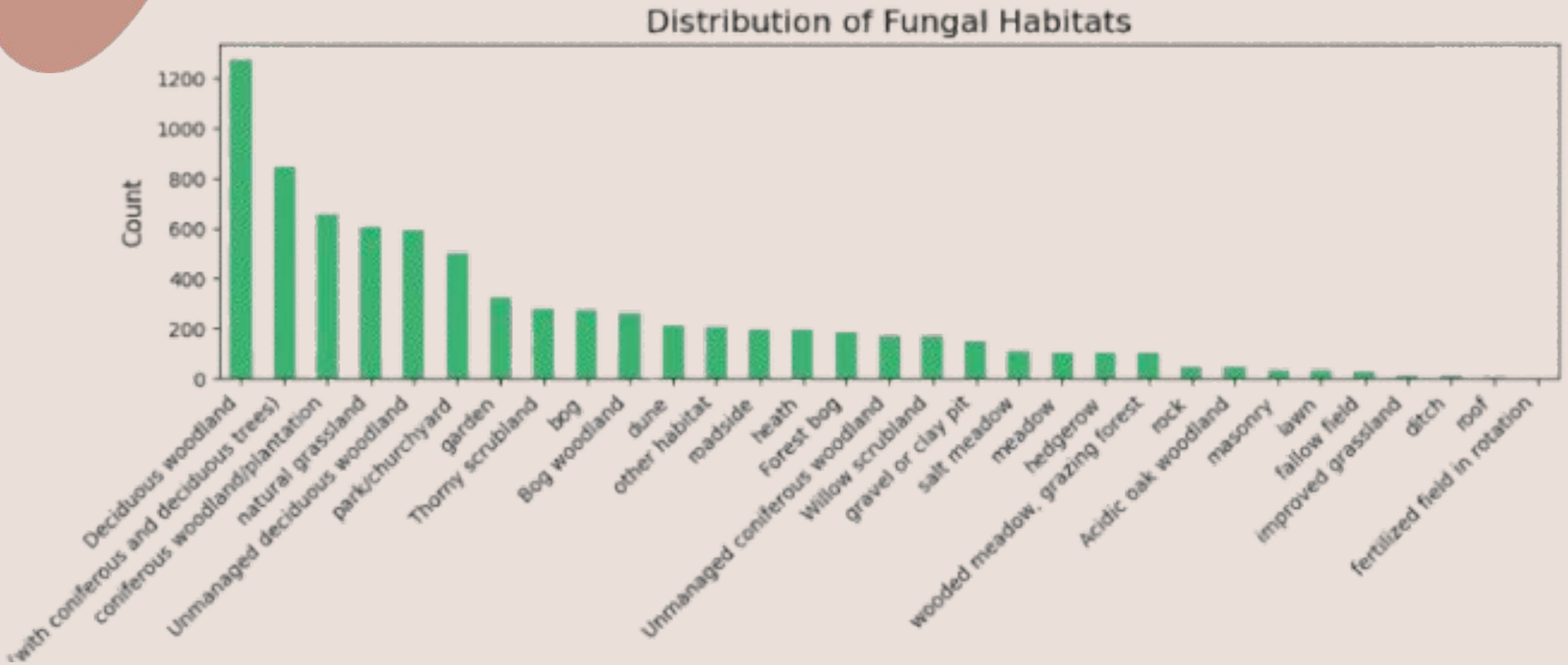
Train: 5,571 images
Val: 1,406 images
Test: 2,180 images

Origins of fungal samples

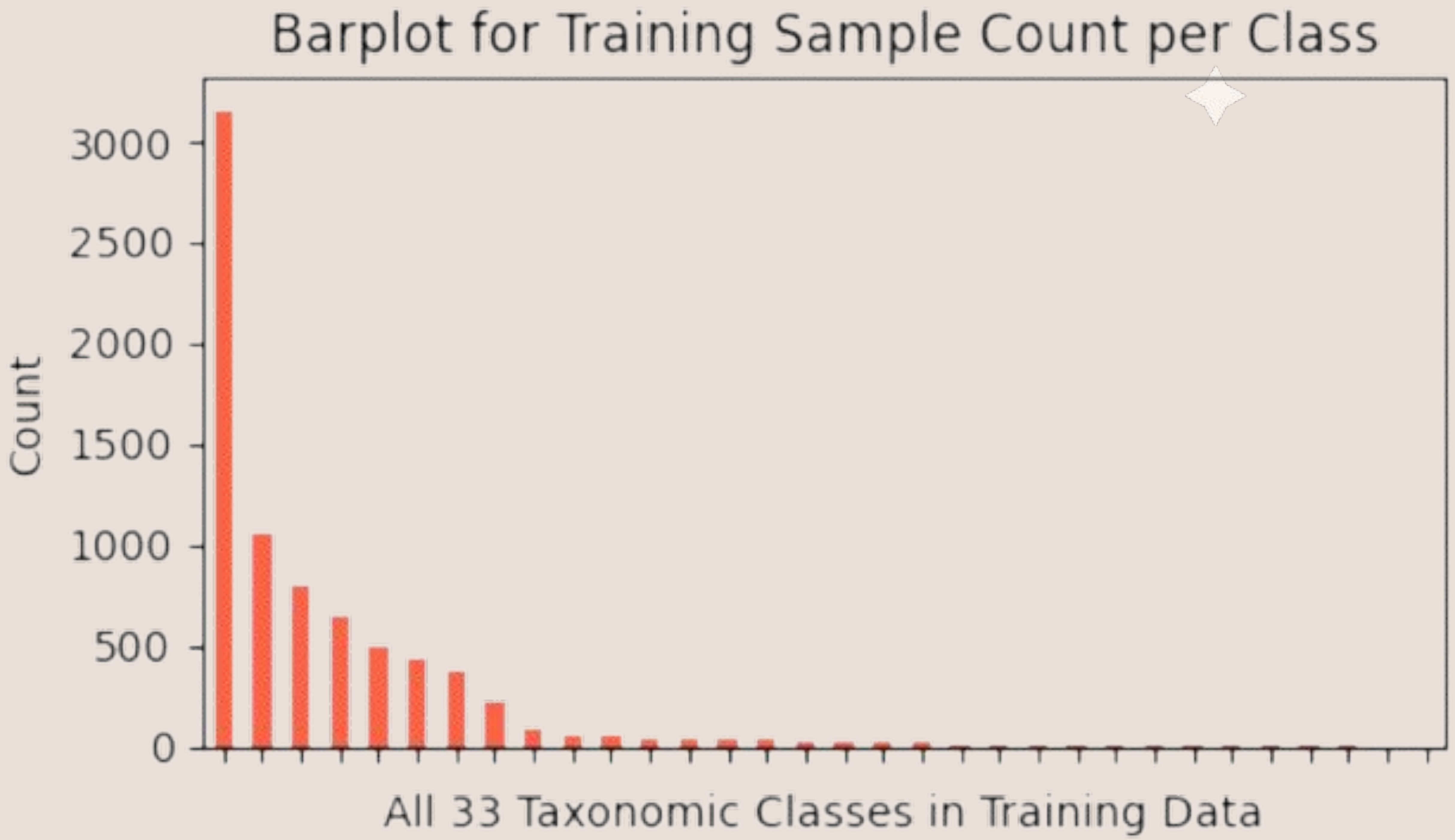
Metadata Skew

The metadata fields of interest (elevation and habitat) tend to skew





Class Imbalance Strategy



- Severe class imbalance, initial models learned how to predict majority
- Goal: aim for 600 images per class
  - Remove classes that have only 1 image
- How we did this:
  - Augment enough images in the minority classes to meet threshold
  - Also experimented with downsampling to meet threshold

```
class
Agaricomycetes          3156
Leotiomycetes           1052
Sordariomycetes          798
Lecanoromycetes          653
Dothideomycetes          503
Pezizomycetes            438
Myxomycetes              381
Pucciniomycetes          230
Eurotiomycetes            88
Ustilaginomycetes         52
Exobasidiomycetes         50
Tremellomycetes           47
Peronosporea              47
Orbiliomycetes            38
Microbotryomycetes        34
Dacrymycetes              33
Mucoromycetes             26
Arthoniomycetes           24
Taphrinomycetes           19
Coniocybomycetes          18
Entomophthoromycetes      16
Laboulbeniomycetes        15
Candelariomycetes         15
Geoglossomycetes          13
Zoopagomycetes             6
Lichinomycetes             6
Sareomycetes               6
Glomeromycetes             5
Cystobasidiomycetes        4
Atractiellomycetes         4
Chytridiomycetes           1
Blastocladiomycetes        1
Name: count, dtype: int64
```

Image Augmentation

- <u>Problem</u>: *all images were of different sizes*
  - Shrinking images while keeping aspect ratio
  - Padding to fill out desired dimensions (224x224)
- <u>Problem</u>: *class imbalance, certain classes with too few images*
  - Random image augmentations
    - flip_left_right
    - flip_up_down
    - adjust_brightness
    - adjust_contrast

Overall Preprocessing Steps

In General:

- Remove classes with only one image so we can stratify
- Augment/downsample images due to class imbalance

For Images:

- Rescale
- Shrink according to aspect ratio
- Padding
- Normalize

For Metadata:

- Remove observations with missing data (no imputation)
- Normalize elevation
- Habitat embeddings

3. Modeling

Modeling

Naive Baseline: most common taxonomic class

- All class predictions are “Agaricomycetes”, being the majority class

Baseline Modeling: CNN (Image Only)

- Basic experimental architecture:
  - Layers: [(Convolution -> Pooling -> Dropout) repeated] -> Flatten -> Dense

Enhanced Modeling: Multi-Input Neural Network

- Architecture:
  - Three branches:
    1. **CNN branch** for image data
    2. **Fully connected branches** for metadata (elevation, habitat)
  - Both branches are concatenated before passing through a final dense layer for multi-class classification

Extra Modeling: Feed-Forward Neural Net

4. Experiments

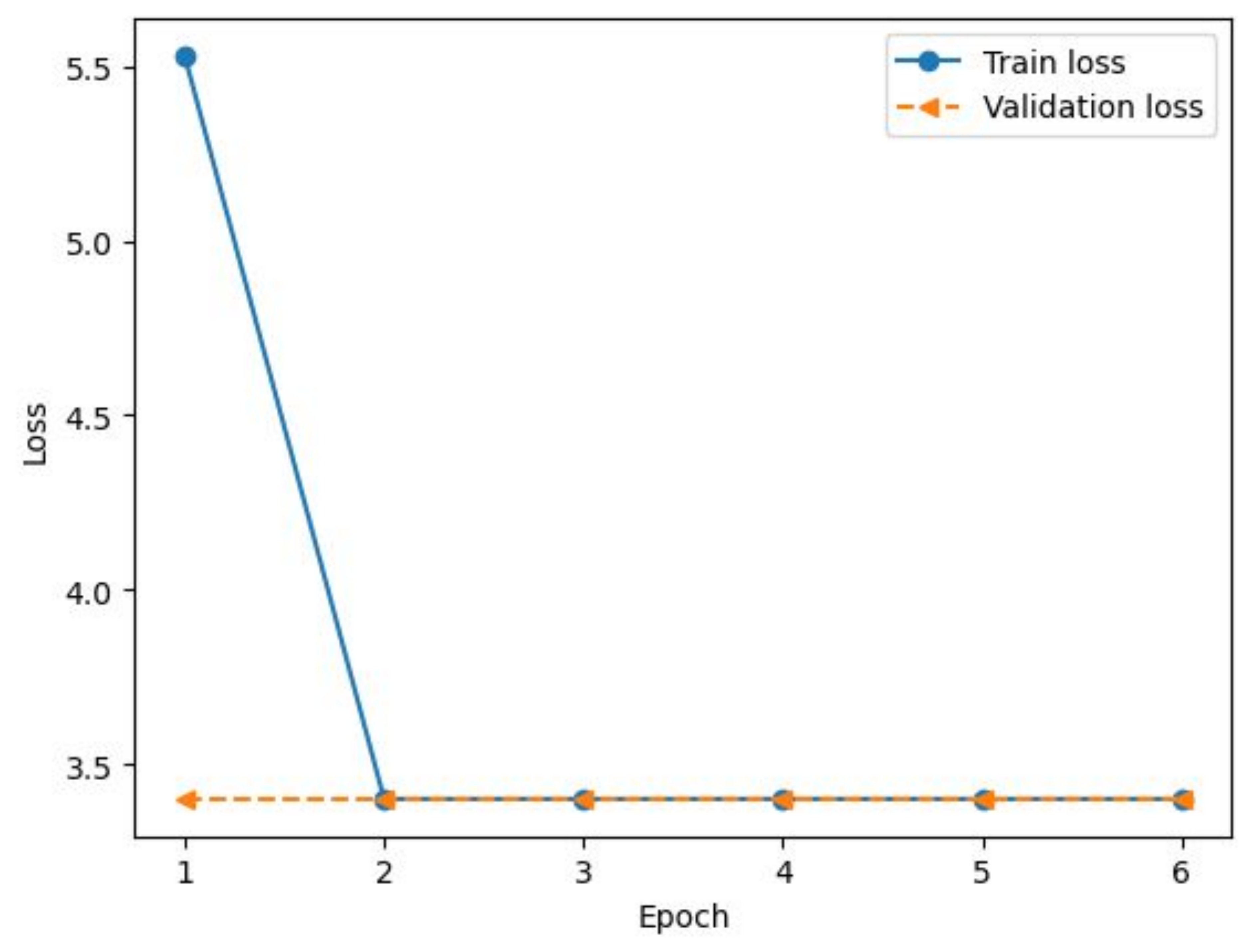
Experiments

Goal: Evaluate how different modeling choices affect classification accuracy for fungal **classes**, especially under class imbalance and with limited data.

Model Type	Test Accuracy
Naive Baseline (majority class)	3.3%
Image-Only CNN	2.4%
Multi-Input Network	8.2%
Feed-Forward Neural Network	11.1%

Image-only CNN

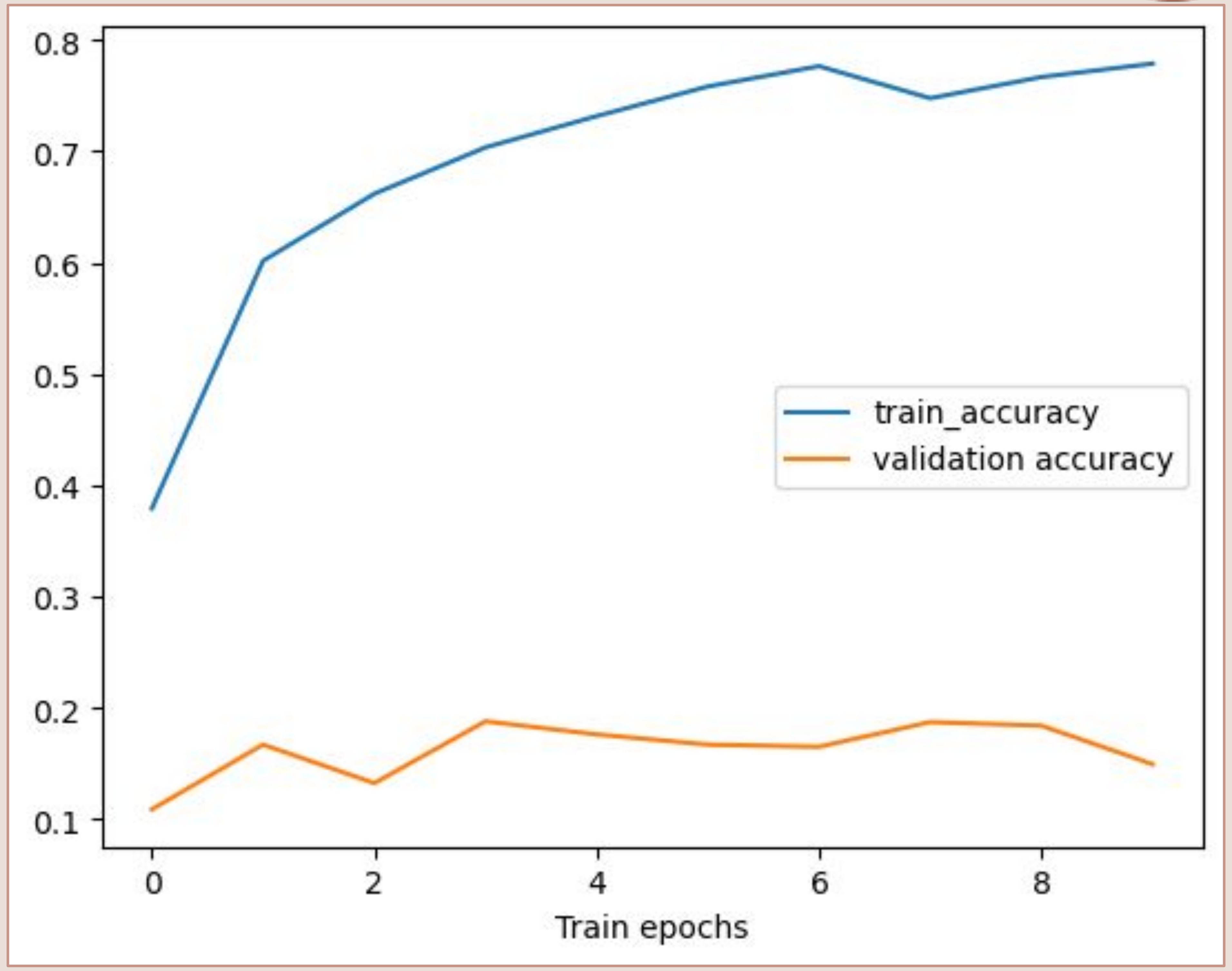
Layers	Params
Conv2D (filters=32, kernel_size=4, padding="same", activation="relu")	1,568
MaxPool2D	0
Dropout(0.25)	0
Flatten	0
Dense(activation=”softmax”)	12,042,270



Accuracy: 2.4%

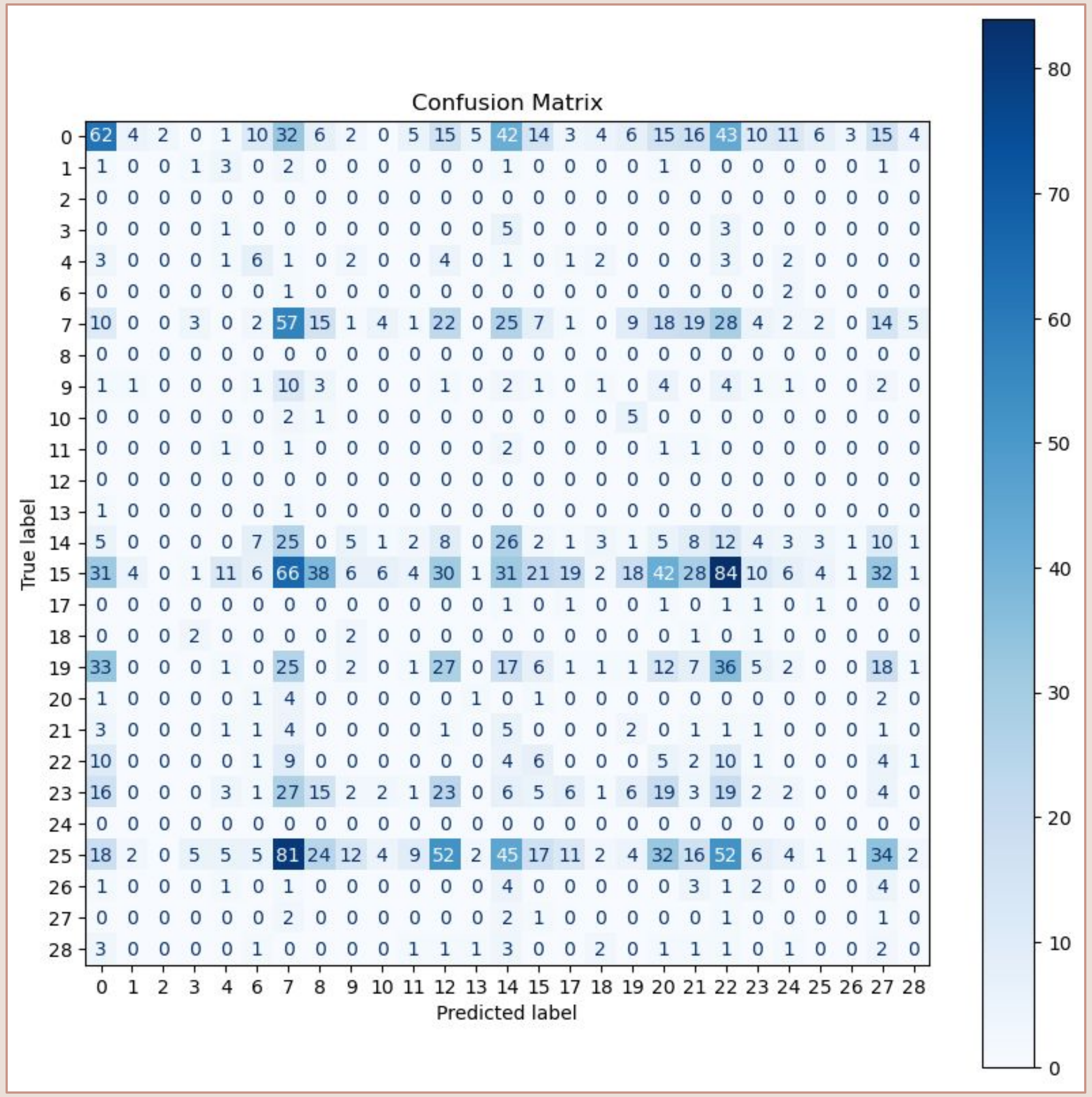
Multi-Layer with Metadata

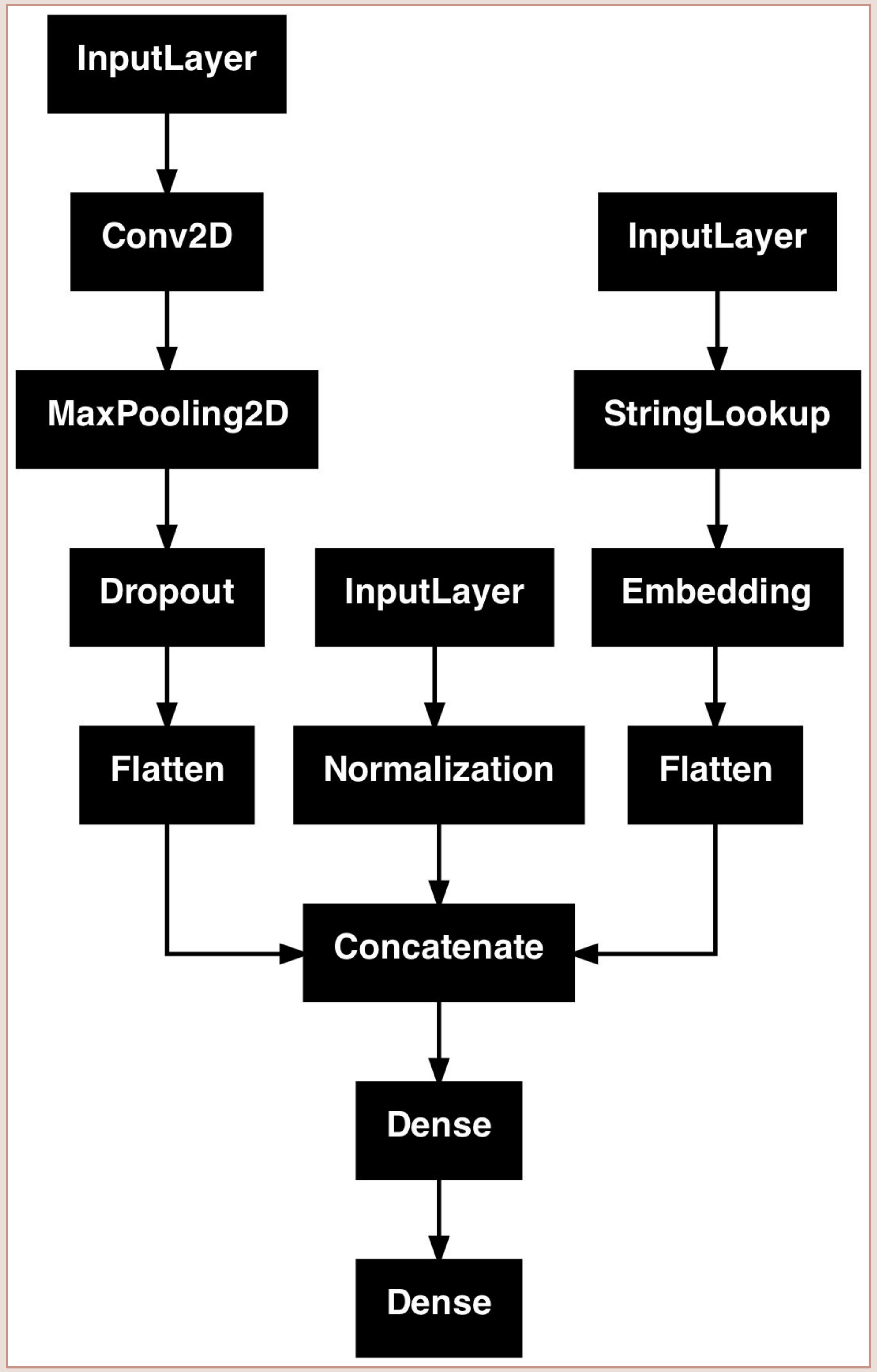
Image Layer	Elevation Layer	Habitat Layer
Conv2D (filters=32, kernel_size=4, padding="same", activation="relu")	Normalization	StringLookup
MaxPool2D		Embedding
Dropout(0.3)		Flatten
Flatten		
Dense(activation=" softmax")		
Concatenate		



Accuracy: 8.2%

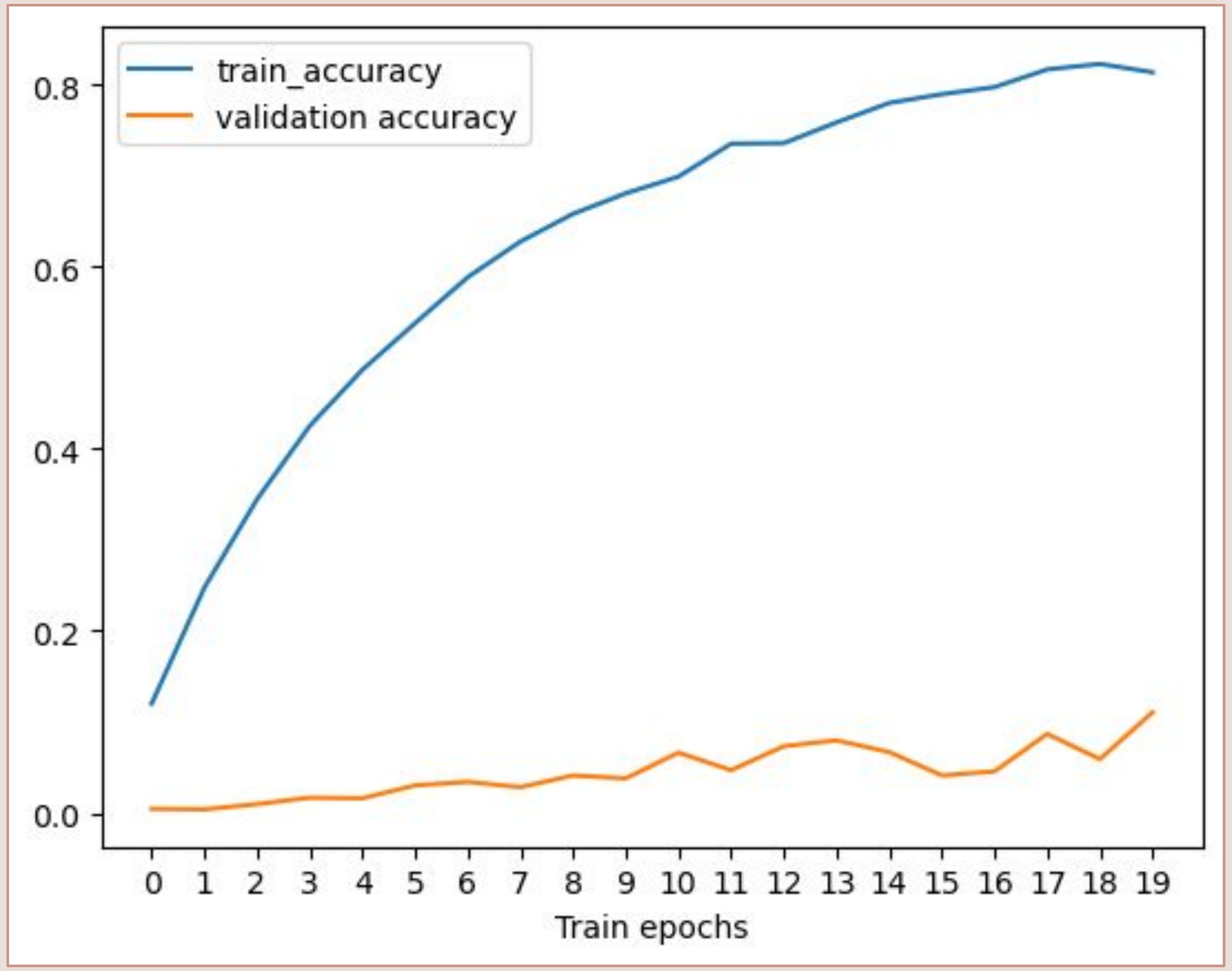
Multi-Layer with Metadata





Feed-Forward Neural Net

Hyperparams	
Hidden Layers	[256,128,64]
Activation	relu
Optimizer	SGD
Learning Rate	0.01
Epochs	20



Accuracy: 11.1%

5. Conclusion

Conclusions

Key Takeaways

- We showed that additional metadata can improve classification
- Simpler models can perform better than more complicated ones

What We Learned

- Class imbalance is a major challenge
- Takes a lot of compute (computer would crash)
- Rare fungi classes (<10 observations) are difficult to augment
- We could have started with the metadata only approach, then experimented combining it with a CNN to see if the accuracy would improve

Future Work

- Spend more time with CNN development
- Explore advanced architectures
  - Few-shot learning techniques
- Move towards species-level prediction

GitHub Repo Link

https://github.com/kalafosaurus/207-final-project

Contributions

Rachel

- EDA
- Experimented with different image preprocessing and data cleaning techniques
- Experimented with data augmentation
- Created baseline CNN
- Experimented with balancing classes
- Attempted to stratify splits
- Contributed to slide deck

Ryan

- Built and ran all our models: image-only CNN, CNN with metadata using functional API, FFNN
- Wrote lots of code:
  - image preprocessing, keeping aspect ratios when resizing, padding, etc
  - creating embeddings for habitat
  - class imbalance image augmentation and downsampling
- Also helped with slides

Will

- Contributed to data prep and preprocessing
- Designed and trained CNN for multi-class classification
- Addressed class imbalance through data augmentation (e.g., rotations, flips) and downsampling
- Evaluated in-memory vs. disk-based augmentation to balance training speed / storage constraints
- Gained practical experience in building and tuning CNNs
- Had fun with slides