



Universität St.Gallen



**Team undefined**

# **Machine Learning - Coding Challenge - Spring 2022**

Ruben Octavio Gonzalez Avilés, Alexander Lontke, Nora Choukrani



**Nora Choukrani**

Strong Background in Finance



**Ruben Gonzalez**

Working in Quantitative Risk Modelling



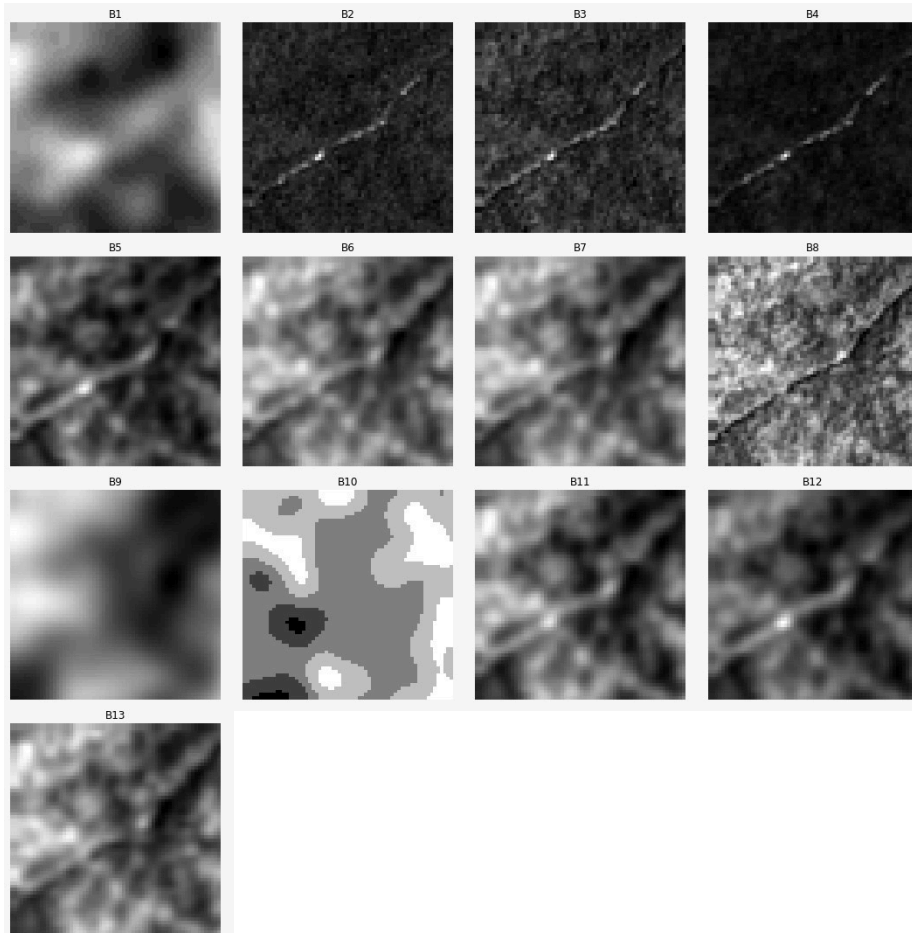
**Alexander Lontke**

Worked as a Machine Learning Engineer

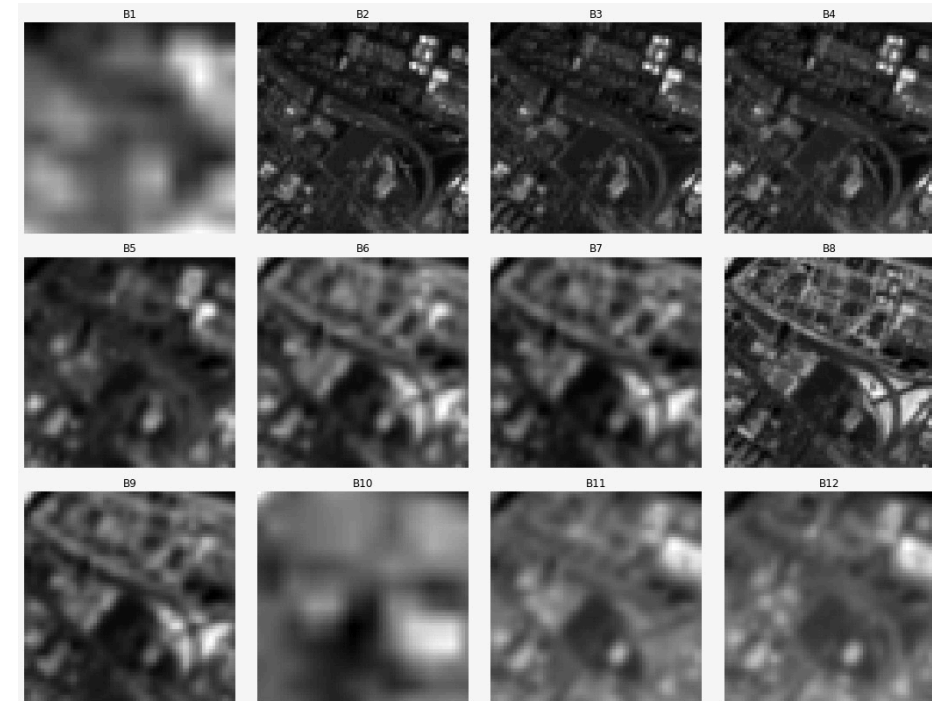


# Pre-processing and normalizing the data

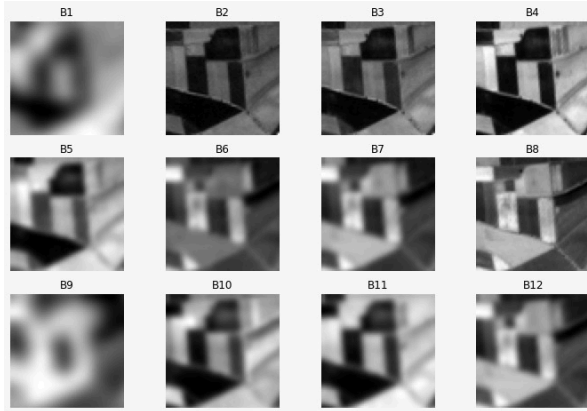
Training set sample



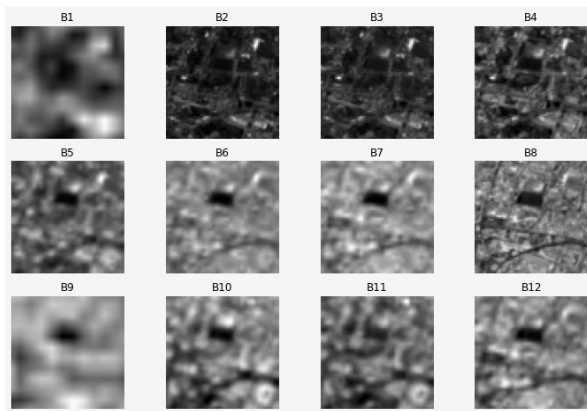
Submission set sample



# Pre-processing and normalizing the data



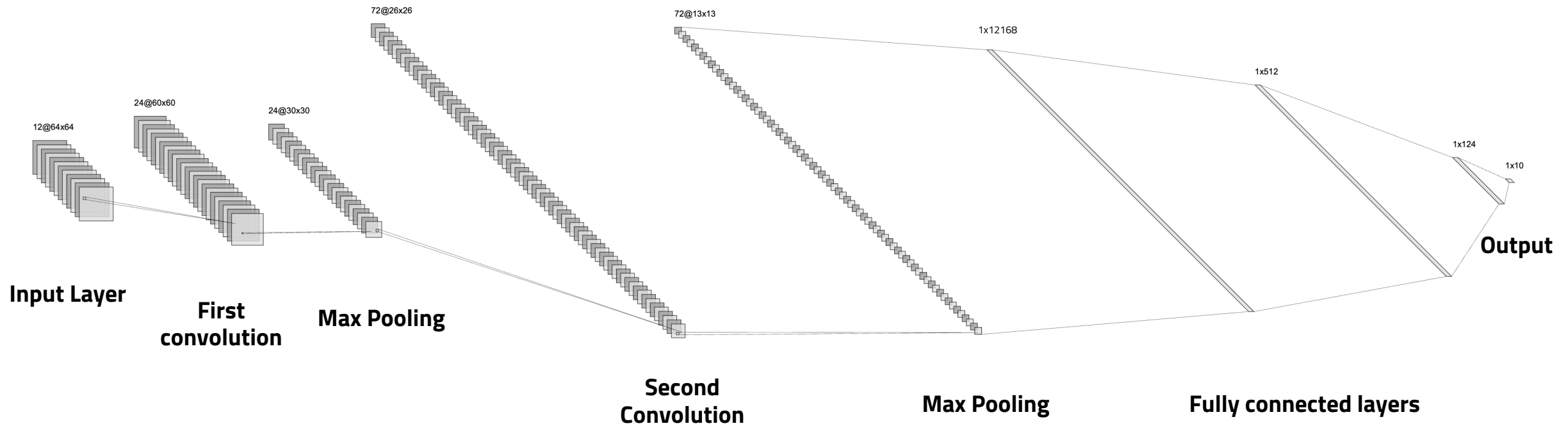
Training set sample



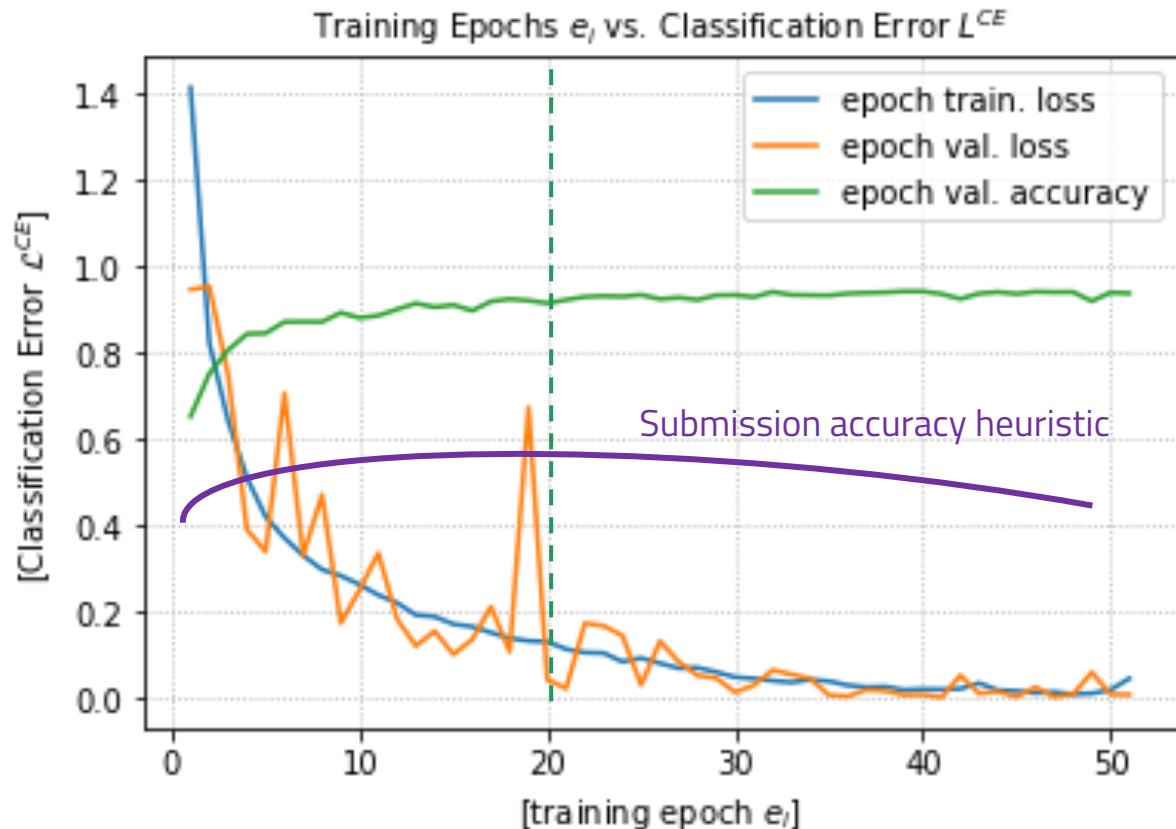
Submission set sample

- **Step 1:** exclude the 10<sup>th</sup> band of the training samples to only keep 12 bands as in the testing sample
- **Step 2:** move band 9 of submission data to last index to conform to training set
- **Step 3:** normalize all bands of the training data set to a mean of 0 and a standard deviation of 1
- **Step 4:** normalize all bands of the submission data set to a mean of 0 and a standard deviation of 1

# The Model



# Training and Evaluation



## Setup:

- Cross Entropy Loss
- Stochastic gradient descent (LR=0.001)

## For experimental research:

- Training: 20'000 samples
- Validation: 7'000 samples
- Up to 200 epochs

## For submission:

- Training 27'000 samples
- No Validation set
- 20 epochs

# Validation Set - Report

Class	Precision	Recall	F1-Score
AnnualCrop	0.92	0.95	0.94
Forest	0.98	0.99	0.99
HerbaceousVegetation	0.96	0.91	0.94
Highway	0.82	0.86	0.84
Industrial	0.91	0.95	0.93
Pasture	0.92	0.93	0.93
PermanentCrop	0.92	0.89	0.91
Residential	0.98	0.92	0.95
River	0.94	0.96	0.95
SeaLake	1.00	0.99	0.99
Weighted avg	0.94	0.94	0.94

# Validation Set - Confusion Matrix



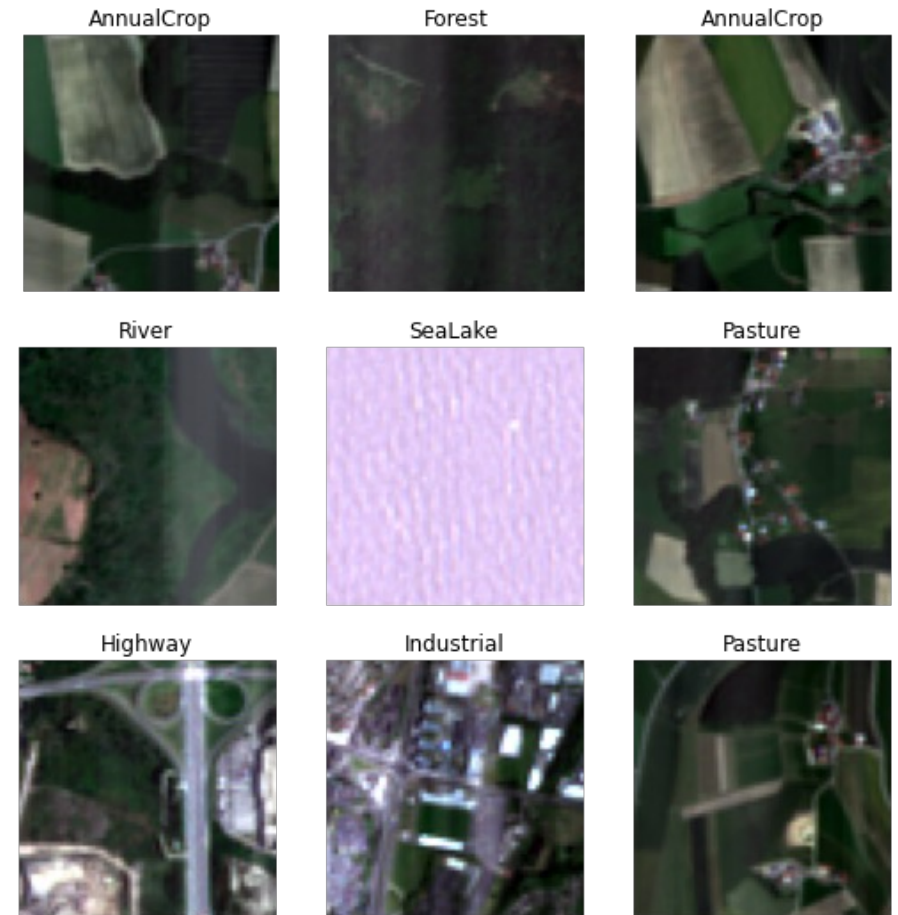


# Validation Set - Visualization

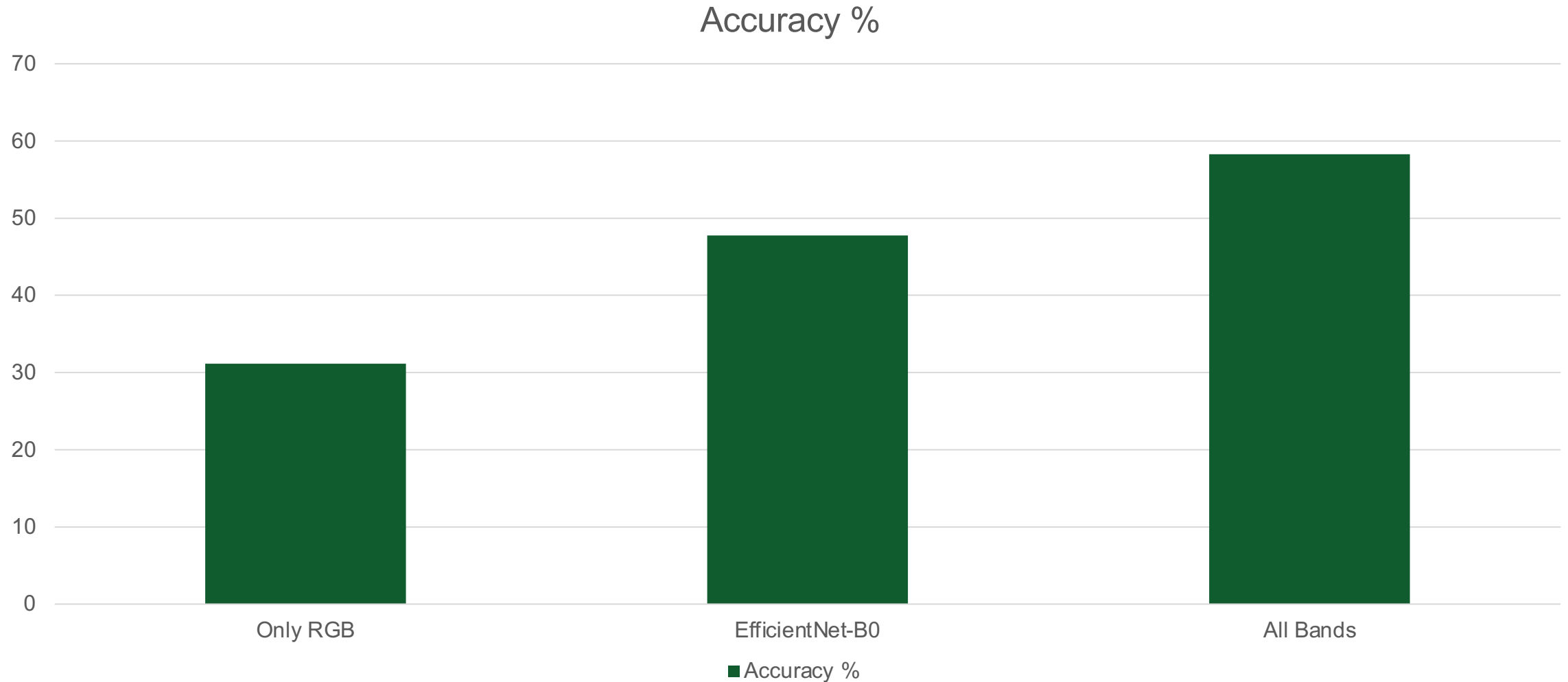


# Submission Predictions

Class	Predictions Count
SeaLake	1'012
PermanentCrop	714
Highway	572
River	408
AnnualCrop	393
HerbaceousVegetation	365
Pasture	335
Industrial	180
Forest	149
Residential	104
Total	4'232



# Additional Approaches



- Pre-processing of Sentinel Level-1C submission samples to Sentinel Level-2A
  - Would require original Sentinel data products
- Submission set seems unbalanced
  - Different normalization method might be more suitable
- Normalization approach considers bands independent from each other
  - Different normalization method might be more suitable
- We could try more variations of our CNN model architecture





# Questions?