Week 7 Progress Report

Presentor: Matt Behnke
Mentor: Dr. Guo

Table of Contents

Testing Lower Training Ratios

1. Manual Crop
2. PCA Crop
3. Hottest Pixel Crop

Lower Training Ratio Test Methodology

1. Testing each crop style
    a. Manual
    b. PCA
    c. Hottest Pixel Search
2. Every crop style will have the following train/test ratios
    a. 10/90 Train/Test
    b. 20/80 Train/Test
    c. 30/70 Train/Test
    d. 40/60 Train/Test
3. Each train/test ratio will be put through both the Random Forest and Neural Network Model
4. Every individual triall will receive a new random split

Manual Crop Results - Random Forest

```
10/90 Split
-------------------------------------------
Ave Accuracy = 0.9845383177570092
Ave CV Score = 0.983135632183908
Ave Time Taken = 2.491233940124512 seconds
-------------------------------------------


20/80 Split
-------------------------------------------
Ave Accuracy = 0.9925820016820854
Ave CV Score = 0.9892531073446328
Ave Time Taken = 4.298071670532226 seconds
-------------------------------------------


30/70 Split
-------------------------------------------
Ave Accuracy = 0.9966938971648246
Ave CV Score = 0.9915755305867665
Ave Time Taken = 6.1834619140625 seconds
-------------------------------------------


40/60 Split
-------------------------------------------
Ave Accuracy = 0.9968161434977576
Ave CV Score = 0.9924930921521151
Ave Time Taken = 7.822036104202271 seconds
-------------------------------------------
```

- Lowest test accuracy score was ~98.4 with the 10/90 train/test split
  - But very fast speed at ~2.49 seconds per split
- Highest test accuracy of ~99.68 with the 40/60 train test split
  - Not a huge overall improvement from from 30/70 or 20/80 train/test split
  - Only took ~7.82 seconds

Random Forest Comparison to Higher Ratios

```
10/90 Split
-------------------------------------------
Ave Accuracy = 0.9845383177570092
Ave CV Score = 0.983135632183908
Ave Time Taken = 2.491233940124512 seconds
-------------------------------------------

20/80 Split
-------------------------------------------
Ave Accuracy = 0.9925820016820854
Ave CV Score = 0.9892531073446328
Ave Time Taken = 4.298071670532226 seconds
-------------------------------------------

30/70 Split
-------------------------------------------
Ave Accuracy = 0.9966938971648246
Ave CV Score = 0.9915755305867665
Ave Time Taken = 6.1834619140625 seconds
-------------------------------------------

40/60 Split
-------------------------------------------
Ave Accuracy = 0.9968161434977576
Ave CV Score = 0.9924930921521151
Ave Time Taken = 7.822036104202271 seconds
-------------------------------------------

50/50 Split
-------------------------------------------
Ave Accuracy = 0.9960969044414537
Ave CV Score = 0.9950388173408308
Ave Time Taken = 9.556606903076172 seconds
-------------------------------------------

60/40 Split
-------------------------------------------
Ave Accuracy = 0.9984861227922623
Ave CV Score = 0.995405435942502
Ave Time Taken = 11.37139702796936 seconds
-------------------------------------------

70/30 Split
-------------------------------------------
Ave Accuracy = 0.9984753363228698
Ave CV Score = 0.9961346153846152
Ave Time Taken = 13.072708721160888 seconds
-------------------------------------------

80/20 Split
-------------------------------------------
Ave Accuracy = 0.9980504201680672
Ave CV Score = 0.9968844449172076
Ave Time Taken = 14.656173028945922 seconds
-------------------------------------------
```

- Lower splits had very similar accuracy results compared to higher train/test ratios
- Every split from 20/80 got test accuraies above 99%

Manual Crop Results - Neural Network

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.59825545946757
Ave Test Loss Score = 3.316163842873223
Ave Time Taken = 12.020630327860514 seconds
-----------------------------------------


20/80 Split
-----------------------------------------
Ave Accuracy = 0.9465937813123068
Ave Test Loss Score = 0.18459141140107616
Ave Time Taken = 31.58828632036845 seconds
-----------------------------------------


30/70 Split
-----------------------------------------
Ave Accuracy = 0.9818997263908387
Ave Test Loss Score = 0.11530724370229442
Ave Time Taken = 37.021276791890465 seconds
-----------------------------------------


40/60 Split
-----------------------------------------
Ave Accuracy = 0.9918161511421204
Ave Test Loss Score = 0.057953184803527395
Ave Time Taken = 43.39260740280152 seconds
-----------------------------------------
```

- Lowest test accuracy score was ~59.8% on the 10/90 train/test split
  - Very low compared to other splits and random forest score at the same split (~98%)
- Highest test accuracy of ~99.18 with the 40/60 train test split

Neural Network Comparison to Higher Training Ratios

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.59825545946757
Ave Test Loss Score = 3.316163842873223
Ave Time Taken = 12.020630327860514 seconds
-----------------------------------------
```

```
50/50 Split
-----------------------------------------
Ave Accuracy = 0.990040385723114
Ave Test Loss Score = 0.07103357190469313
Ave Time Taken = 50.97666838169098 seconds
-----------------------------------------
```

```
20/80 Split
-----------------------------------------
Ave Accuracy = 0.9465937813123068
Ave Test Loss Score = 0.18459141140107616
Ave Time Taken = 31.58828632036845 seconds
-----------------------------------------
```

```
60/40 Split
-----------------------------------------
Ave Accuracy = 0.9968881487846375
Ave Test Loss Score = 0.034946644029723396
Ave Time Taken = 60.77804760932922 seconds
-----------------------------------------
```

```
30/70 Split
-----------------------------------------
Ave Accuracy = 0.9818997263908387
Ave Test Loss Score = 0.11530724370229442
Ave Time Taken = 37.021276791890465 seconds
-----------------------------------------
```

```
70/30 Split
-----------------------------------------
Ave Accuracy = 0.9949551463127136
Ave Test Loss Score = 0.03611448702022367
Ave Time Taken = 72.20257966518402 seconds
```

```
40/60 Split
-----------------------------------------
Ave Accuracy = 0.9918161511421204
Ave Test Loss Score = 0.057953184803527395
Ave Time Taken = 43.39260740280152 seconds
-----------------------------------------
```

```
80/20 Split
-----------------------------------------
Ave Accuracy = 0.998991596698761
Ave Test Loss Score = 0.026952999633181245
Ave Time Taken = 75.84459526538849 seconds
-----------------------------------------
```

- Starts to hit 99% average accuracy at 40/60 split and continues to have good accuracy after that
- 10/90 had very low accuracy
  - Implying that there is not enough training data

PCA Search Results - Random Forest

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.9546317757009344
Ave CV Score = 0.9535218390804597
Ave Time Taken = 0.8038173389434814 seconds
-----------------------------------------


20/80 Split
-----------------------------------------
Ave Accuracy = 0.9788393608074011
Ave CV Score = 0.9733073446327682
Ave Time Taken = 1.5749772357940675 seconds
-----------------------------------------


30/70 Split
-----------------------------------------
Ave Accuracy = 0.9880442095146563
Ave CV Score = 0.9820364544319605
Ave Time Taken = 2.250534381866455 seconds
-----------------------------------------


40/60 Split
-----------------------------------------
Ave Accuracy = 0.9903587443946185
Ave CV Score = 0.9866349522859994
Ave Time Taken = 2.9538190078735354 seconds
-----------------------------------------
```

- Lowest test accuracy of ~95.46% with 10/90 train/test split
  - Although less than a second per split!
- Highest test accuracy of ~99.03% with 40/60 split

PCA Random Forest Comparison to Higher Training Ratio

```
10/90 Split
----------------------------------------
Ave Accuracy = 0.9546317757009344
Ave CV Score = 0.9535218390804597
Ave Time Taken = 0.8038173389434814 secon
----------------------------------------
```

```
20/80 Split
----------------------------------------
Ave Accuracy = 0.9788393608074011
Ave CV Score = 0.9733073446327682
Ave Time Taken = 1.5749772357940675 secon
----------------------------------------
```

```
30/70 Split
----------------------------------------
Ave Accuracy = 0.9880442095146563
Ave CV Score = 0.9820364544319605
Ave Time Taken = 2.250534381866455 second
----------------------------------------
```

```
40/60 Split
----------------------------------------
Ave Accuracy = 0.9903587443946185
Ave CV Score = 0.9866349522859994
Ave Time Taken = 2.9538190078735354 secon
----------------------------------------
```

```
50/50 Split
----------------------------------------
Ave Accuracy = 0.9945625841184387
Ave CV Score = 0.9901948122619261
Ave Time Taken = 3.532757863998413 seconds
----------------------------------------
```

```
60/40 Split
----------------------------------------
Ave Accuracy = 0.9964339781328845
Ave CV Score = 0.9921026928629714
Ave Time Taken = 4.1870765876770015 seconds
----------------------------------------
```

```
70/30 Split
----------------------------------------
Ave Accuracy = 0.9978026905829596
Ave CV Score = 0.993
Ave Time Taken = 4.745021009445191 seconds
----------------------------------------
```

```
80/20 Split
----------------------------------------
Ave Accuracy = 0.9990588235294118
Ave CV Score = 0.9941430344289612
Ave Time Taken = 5.4596040821075436 seconds
----------------------------------------
```

- 40/60 train/test split had above 99% accuracy along with the higher training ratios

PCA Neural Network Results

```
10/90 Split
-------------------------------------------
Ave Accuracy = 0.4962990721066793
Ave Test Loss Score = 1.5333475124843399
Ave Time Taken = 4.131581735610962 seconds
-------------------------------------------


20/80 Split
-------------------------------------------
Ave Accuracy = 0.6200448671976725
Ave Test Loss Score = 1.148665061830338
Ave Time Taken = 6.857193692525228 seconds
-------------------------------------------


30/70 Split
-------------------------------------------
Ave Accuracy = 0.9363767385482789
Ave Test Loss Score = 0.20659438016302742
Ave Time Taken = 11.981612618764242 seconds
-------------------------------------------


40/60 Split
-------------------------------------------
Ave Accuracy = 0.9943198800086975
Ave Test Loss Score = 0.060495707358725045
Ave Time Taken = 23.986290804545085 seconds
-------------------------------------------
```

- Lowest test accuracy of ~49.62% with the 10/90 train/test split ratio
  - Implying not enough training data for the neural network
- Highest test accuracy of 99.43% with the 40/60 train/test split ratio
- Biggest drop off from 30/70 split (~93.63%) to 20/80 split (~62.00%)

PCA Neural Network Comparison to Higher Training Ratio

```
10/90 Split
-------------------------------------------
Ave Accuracy = 0.4962990721066793
Ave Test Loss Score = 1.5333475124843399
Ave Time Taken = 4.131581735610962 seconds
-------------------------------------------

20/80 Split
-------------------------------------------
Ave Accuracy = 0.6200448671976725
Ave Test Loss Score = 1.148665061830338
Ave Time Taken = 6.857193692525228 seconds
-------------------------------------------

30/70 Split
-------------------------------------------
Ave Accuracy = 0.9363767385482789
Ave Test Loss Score = 0.20659438016302742
Ave Time Taken = 11.981612618764242 seconds
-------------------------------------------

40/60 Split
-------------------------------------------
Ave Accuracy = 0.9943198800086975
Ave Test Loss Score = 0.060495707358725045
Ave Time Taken = 23.986290804545085 seconds
-------------------------------------------
```

```
50/50 Split
-------------------------------------------
Ave Accuracy = 0.9825033605098724
Ave Test Loss Score = 0.07390304885323323
Ave Time Taken = 34.462878465652466 seconds
-------------------------------------------

60/40 Split
-------------------------------------------
Ave Accuracy = 0.9969722509384156
Ave Test Loss Score = 0.030337238279241785
Ave Time Taken = 43.7130684375763 seconds
-------------------------------------------

70/30 Split
-------------------------------------------
Ave Accuracy = 0.9980941653251648
Ave Test Loss Score = 0.02437507739925398
Ave Time Taken = 48.024141263961795 seconds
-------------------------------------------

80/20 Split
-------------------------------------------
Ave Accuracy = 0.9912605047225952
Ave Test Loss Score = 0.040819265365271896
Ave Time Taken = 47.78992938995361 seconds
-------------------------------------------
```

- 40/60 train/test split had above 99% accuracy along with the higher training ratios
-

Hottest Pixel Random Forest Results

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.964844859813084
Ave CV Score = 0.9650850574712645
Ave Time Taken = 2.2022187328338623 seconds
-----------------------------------------


20/80 Split
-----------------------------------------
Ave Accuracy = 0.9914213624894868
Ave CV Score = 0.9839932203389833
Ave Time Taken = 3.950077323913574 seconds
-----------------------------------------


30/70 Split
-----------------------------------------
Ave Accuracy = 0.9958097068716961
Ave CV Score = 0.9891895131086145
Ave Time Taken = 5.705218553543091 seconds
-----------------------------------------


40/60 Split
-----------------------------------------
Ave Accuracy = 0.9983856502242152
Ave CV Score = 0.9954230769230767
Ave Time Taken = 12.167110567092896 seconds
-----------------------------------------
```

- Lowest test accuracy of ~96.48% with the 10/90 train/test split ratio
- Highest test accuracy of ~99.84% with the 40/60 train/test split ratio

Hottest Pixel Random Forest Comparison to Higher Ratios

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.964844859813084
Ave CV Score = 0.9650850574712645
Ave Time Taken = 2.2022187328338623 seconds
-----------------------------------------
```

```
50/50 Split
-----------------------------------------
Ave Accuracy = 0.9956931359353969
Ave CV Score = 0.9948781062942138
Ave Time Taken = 10.085623331069947 seconds
-----------------------------------------
```

```
20/80 Split
-----------------------------------------
Ave Accuracy = 0.9914213624894868
Ave CV Score = 0.9839932203389833
Ave Time Taken = 3.950077323913574 seconds
-----------------------------------------
```

```
60/40 Split
-----------------------------------------
Ave Accuracy = 0.9982169890664423
Ave CV Score = 0.9948219195279643
Ave Time Taken = 12.004673089981079 seconds
-----------------------------------------
```

```
30/70 Split
-----------------------------------------
Ave Accuracy = 0.9958097068716961
Ave CV Score = 0.9891895131086145
Ave Time Taken = 5.705218553543091 seconds
-----------------------------------------
```

```
70/30 Split
-----------------------------------------
Ave Accuracy = 0.997892376681614
Ave CV Score = 0.9959807692307691
Ave Time Taken = 13.799284038543702 seconds
-----------------------------------------
```

```
40/60 Split
-----------------------------------------
Ave Accuracy = 0.9983856502242152
Ave CV Score = 0.9954230769230767
Ave Time Taken = 12.167110567092896 seconds
-----------------------------------------
```

```
80/20 Split
-----------------------------------------
Ave Accuracy = 0.9975126050420168
Ave CV Score = 0.9963454242456479
Ave Time Taken = 15.548892192840576 seconds
-----------------------------------------
```

- Lowest test accuracy of ~96.48% with the 10/90 train/test split ratio
- Hits ~99% average test accuracy at 20/80 split

Hottest Pixel Neural Network Results

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.6623302181561788
Ave Test Loss Score = 2.379482292922115
Ave Time Taken = 10.123479127883911 seconds
-----------------------------------------


20/80 Split
-----------------------------------------
Ave Accuracy = 0.9855621019999187
Ave Test Loss Score = 0.10708985485309874
Ave Time Taken = 21.8948610941569 seconds
-----------------------------------------


30/70 Split
-----------------------------------------
Ave Accuracy = 0.9808425466219585
Ave Test Loss Score = 0.09497460637223104
Ave Time Taken = 29.13773142496745 seconds
-----------------------------------------


40/60 Split
-----------------------------------------
Ave Accuracy = 0.9901345332463583
Ave Test Loss Score = 0.06169282445246955
Ave Time Taken = 31.69415764808655 seconds
-----------------------------------------
```

- Lowest test accuracy of ~66.23% with the 10/90 train/test split ratio
    - Implying not enough training data for the neural network
- Highest test accuracy of ~99.01% with the 40/60 train/test split ratio

Hottest Pixel Neural Network Comparison to Higher Ratios

```
10/90 Split
-----------------------------------------
Ave Accuracy = 0.6623302181561788
Ave Test Loss Score = 2.379482292922115
Ave Time Taken = 10.123479127883911 seconds
-----------------------------------------
```

```
50/50 Split
-----------------------------------------
Ave Accuracy = 0.9628532886505127
Ave Test Loss Score = 0.15698368040190488
Ave Time Taken = 15.208650159835816 seconds
-----------------------------------------
```

```
20/80 Split
-----------------------------------------
Ave Accuracy = 0.9855621019999187
Ave Test Loss Score = 0.10708985485309874
Ave Time Taken = 21.8948610941569 seconds
-----------------------------------------
```

```
60/40 Split
-----------------------------------------
Ave Accuracy = 0.9948696434497833
Ave Test Loss Score = 0.050269397379026824
Ave Time Taken = 22.180278539657593 seconds
-----------------------------------------
```

```
30/70 Split
-----------------------------------------
Ave Accuracy = 0.9808425466219585
Ave Test Loss Score = 0.09497460637223104
Ave Time Taken = 29.13773142496745 seconds
-----------------------------------------
```

```
70/30 Split
-----------------------------------------
Ave Accuracy = 0.997197300195694
Ave Test Loss Score = 0.03762968810099791
Ave Time Taken = 24.050864148139954 seconds
-----------------------------------------
```

```
40/60 Split
-----------------------------------------
Ave Accuracy = 0.9901345332463583
Ave Test Loss Score = 0.06169282445246955
Ave Time Taken = 31.69415764808655 seconds
-----------------------------------------
```

```
80/20 Split
-----------------------------------------
Ave Accuracy = 0.9952941179275513
Ave Test Loss Score = 0.03637705843864369
Ave Time Taken = 25.63216655254364 seconds
-----------------------------------------
```

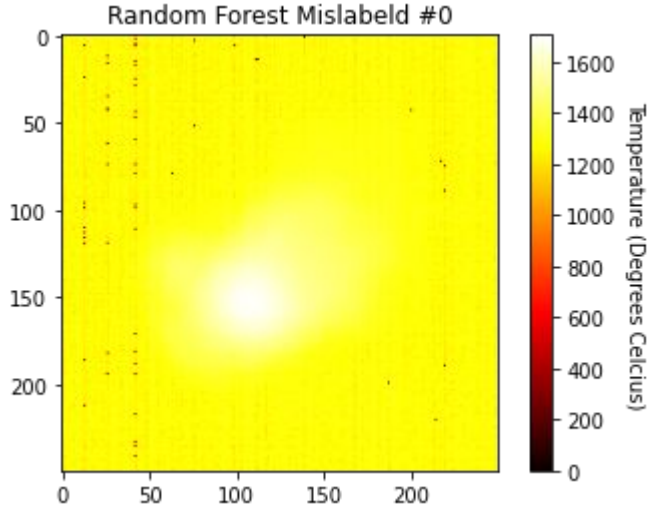
- Accuracy maintains at 99% at 60/40 split
- Time is inconsistent
    - I am currently attributing it to the patience value and how it can vary a lot

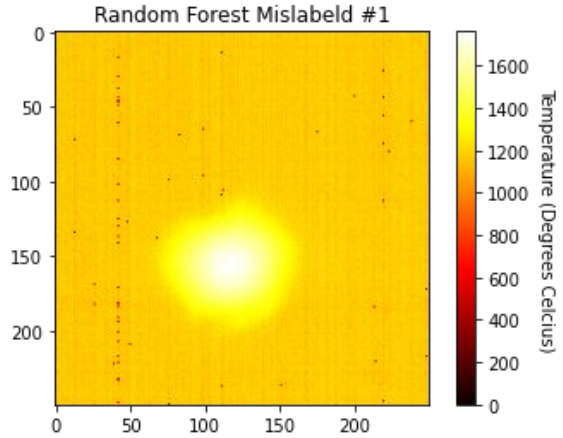
Visualizing Mislabeled Datasets

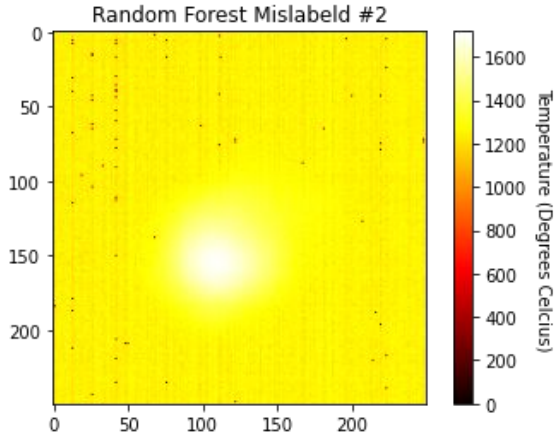
Problems Finding Mislabeled Titles

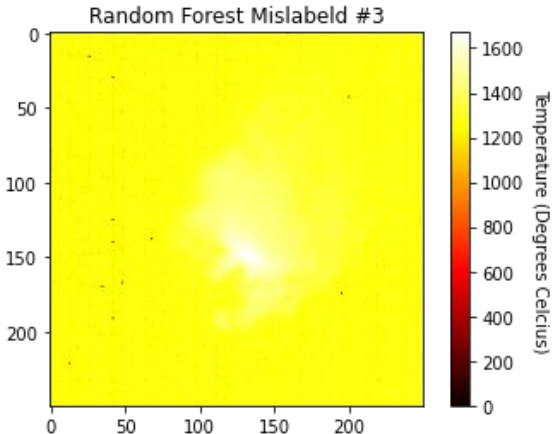
- Titles are mapped to Status (Title = Key, Status = Value)
- Then for each key, value pair
  - The key's respective data file is found and encoded then added to a list (inData)
  - The Status is then added to a seperate list (outData)
- Thus, the title is not considered in training
- Tried to use a dictionary to map data matrix to its title
  - Did not work, because a matrix is not a hashable type
- This would cause me to check if every single value matches it self, but it needs to be done as a search through all the datafiles
  - Essentially it is a very long operation and I am searching for a more elegant/working solution
- But, I could get the visualizations
  - Done on 50/50 splits for both Random Forest and Neural Net

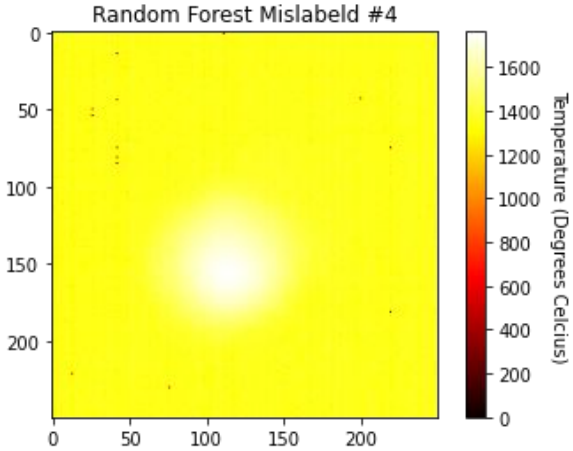
Random Forest Mislabeled Visualizations

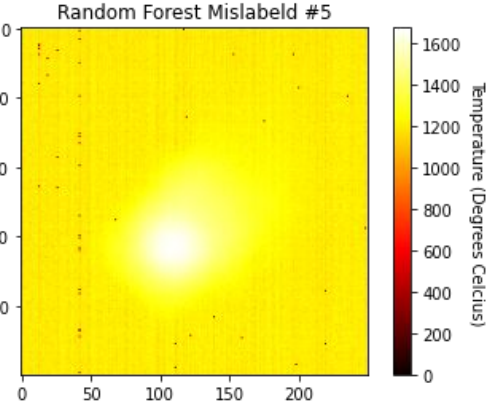




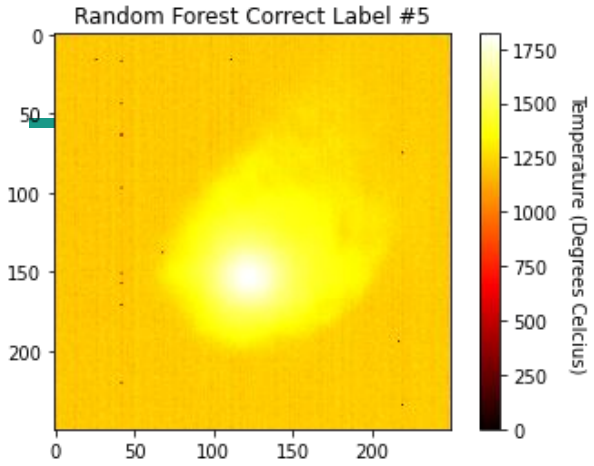


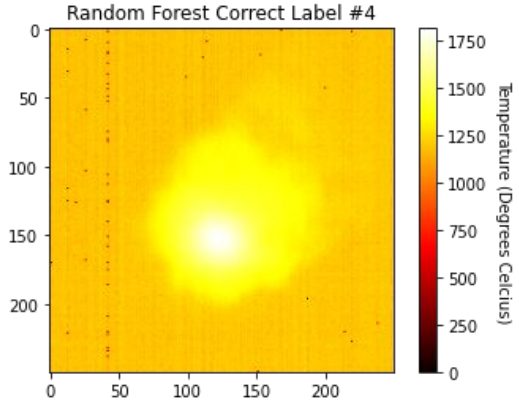


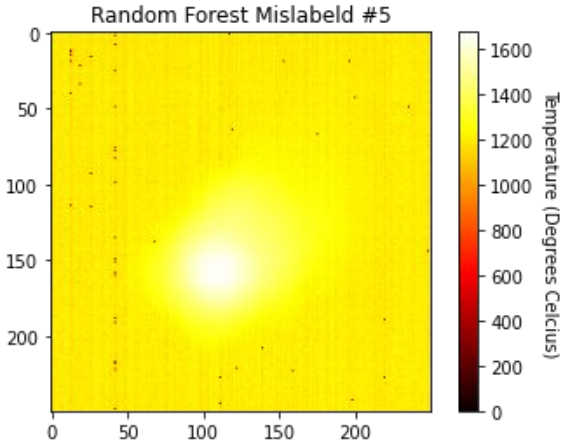


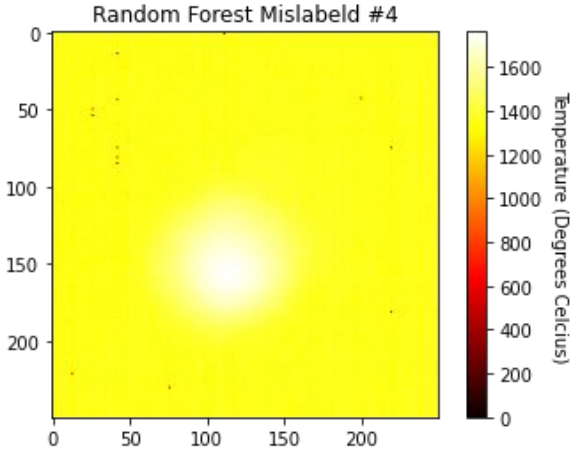


Random Forest Mislabeled Visualizations Comparison



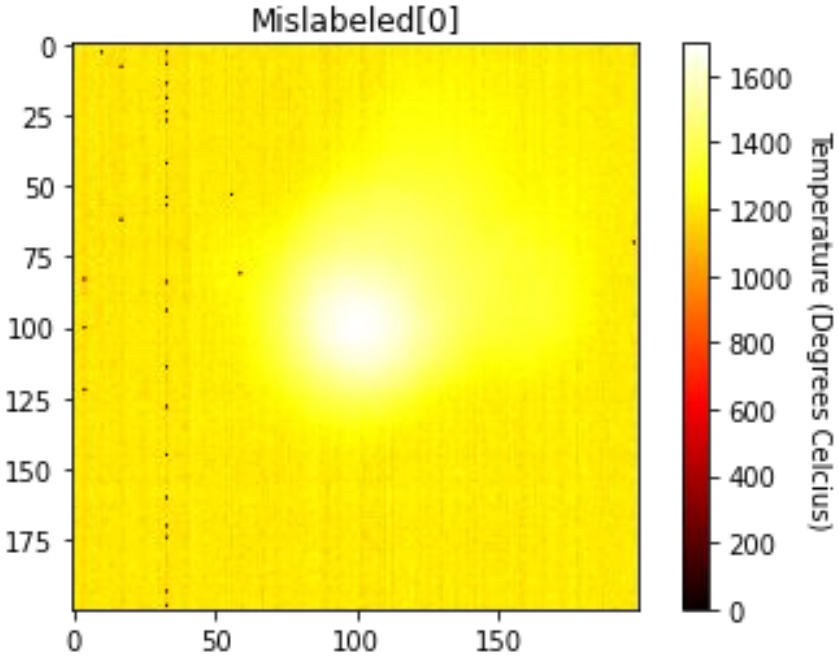




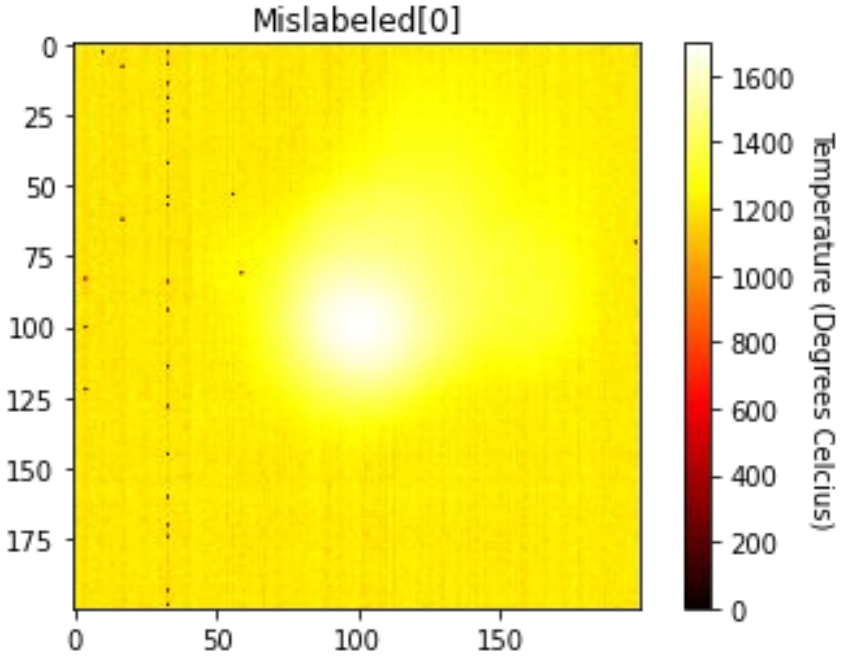


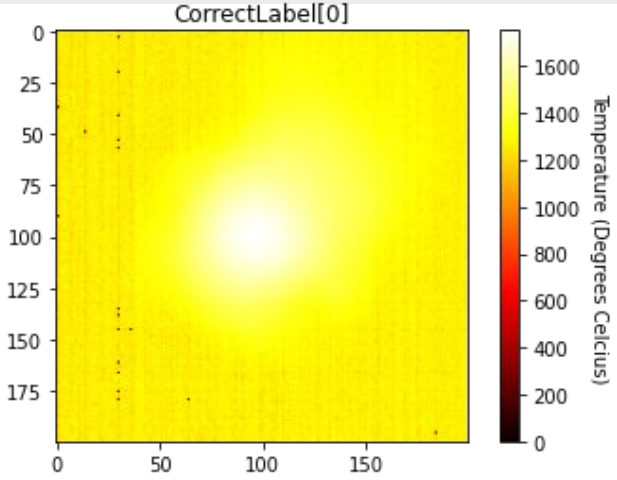
Neural Network Mislabeled Visualization

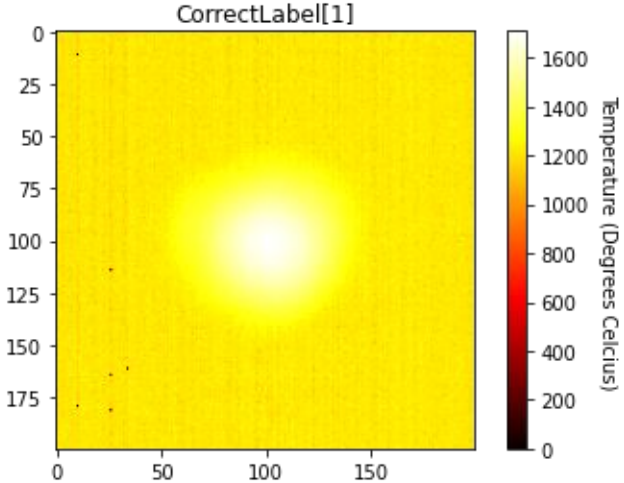
Only 1 file mislabeled in my 50/50 split



Neural Network Mislabeled Visualization







Working On Paper

1. Want to re-read and update figures
2. Want to put comments on my own writing where I know I need guidance
3. Share on google docs and receive feedback?

Future Steps / Goals

- Continue working on the paper
- Finding which titles are being misidentified