Detect Aircraft in Satellite Imagery

Trent Dell’Oro, M.Sc.
March 2019

Business Case

Motivating Questions:

- *“Can learning algorithms be used identify planes in satellite imagery?”*
- *“How well can a learning model be tuned to identify planes?”*
- *“What learning algorithms are best suited for the problem?”*

The automation of satellite imagery analysis may provide unique insights into various markets such as agriculture, defense, intelligence, energy, and finance.

Dataset

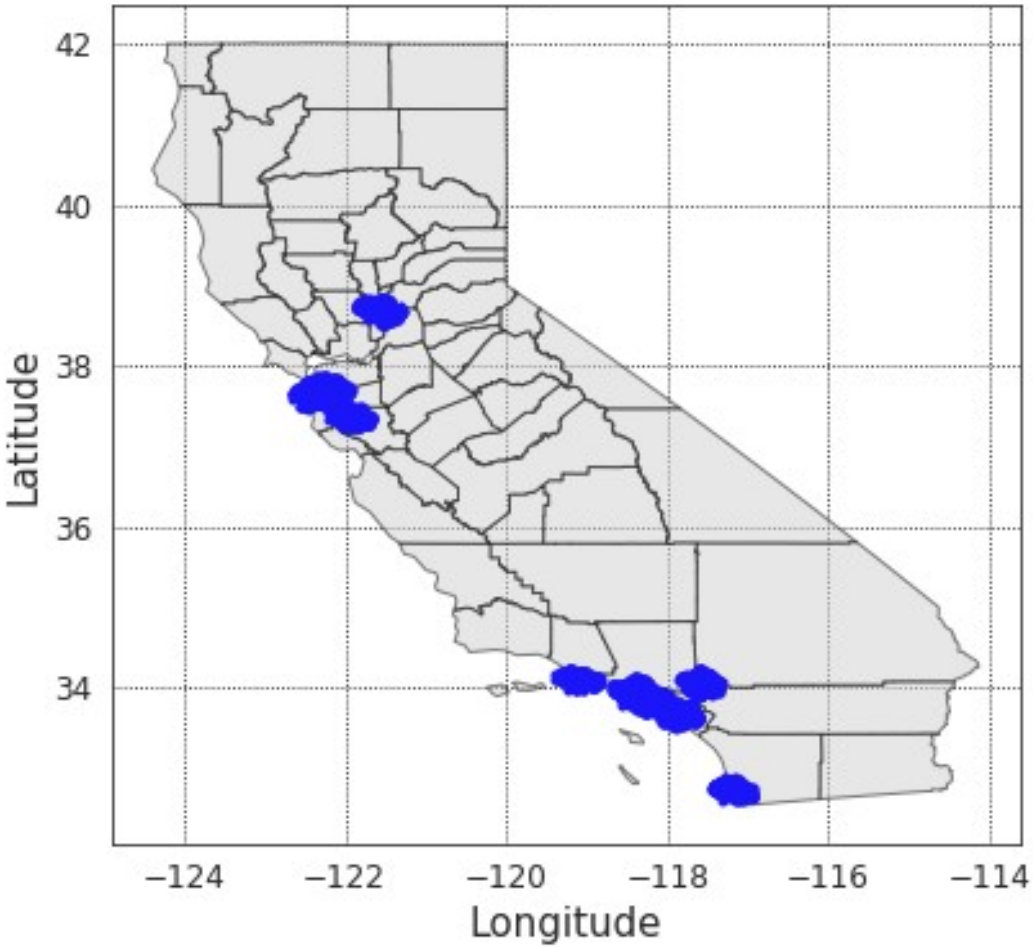
The PlanesNet satellite image chips used in this project are included in the Kaggle dataset “Planes in Satellite Imagery”, provided by Planet Labs Inc. The satellite imagery used to build the PlanesNet image chip dataset is made available through Planet Labs’ Open California dataset, which is openly licensed. A total of 32,000 images collected are provided, each image is 20x20 pixels in size, with three bands.

PlanesNet
www.kaggle.com/rhammell/planesnet#planesnet.json

	data	labels	locations	scene_ids
0	[206, 195, 187, 183, 177, 175, 174, 193, 198, ...	1	[-118.40497658522878, 33.940618514147936]	20170620_175442_0e30
1	[215, 209, 200, 196, 192, 197, 205, 168, 155, ...	1	[-122.392469714, 37.6176425378]	20161212_180859_0e30
2	[204, 214, 220, 219, 213, 205, 198, 193, 199, ...	1	[-122.397578597, 37.6209247852]	20170524_181349_0e2f
3	[179, 174, 179, 178, 173, 170, 168, 168, 168, ...	1	[-122.214849831, 37.7203378331]	20161110_180707_0e1f
4	[222, 222, 218, 214, 208, 205, 207, 206, 206, ...	1	[-117.862173435, 33.6796854072]	20160813_184932_0c64

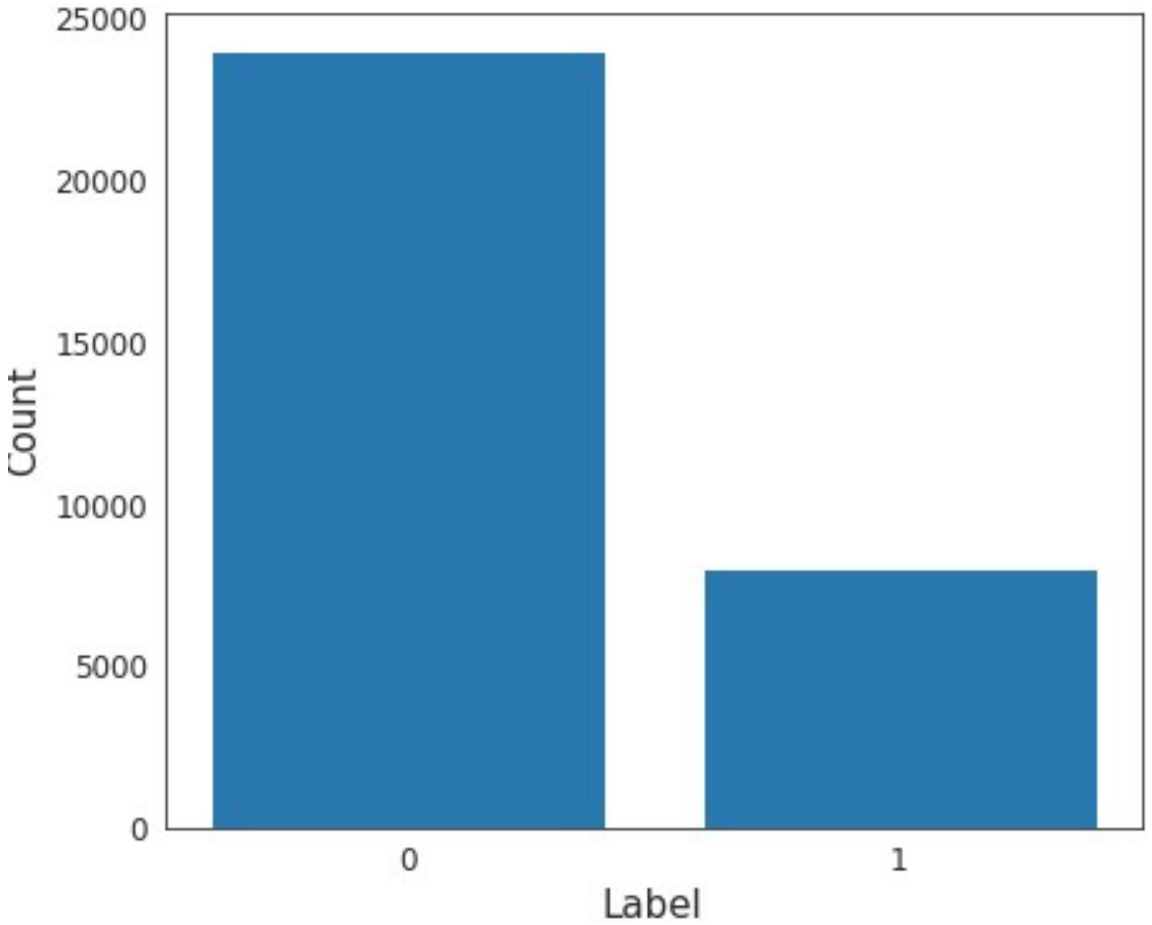
Location of Images Over California

- A map is displayed to show the locations of the satellite images taken over California, the map includes all 'plane' and all 'not-plane' image locations.
- The majority of images come from the Los Angeles area, the San Francisco area, and the San Diego area.

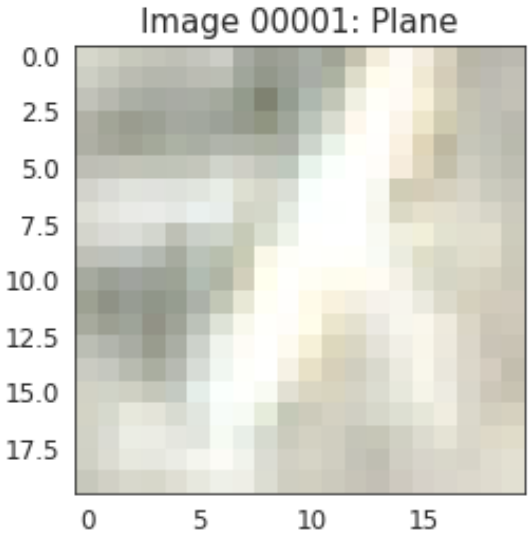


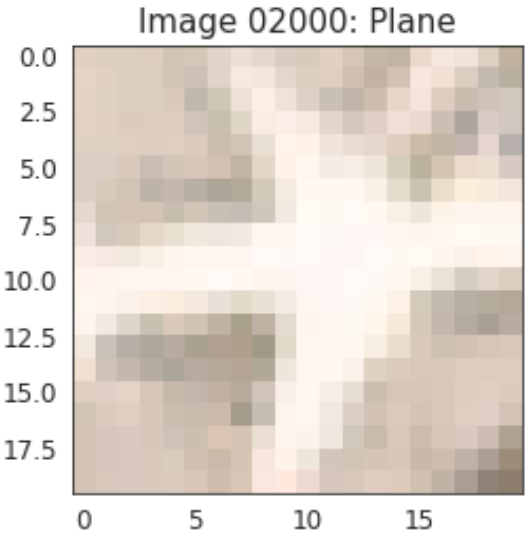
Plane and Not-Plane Class

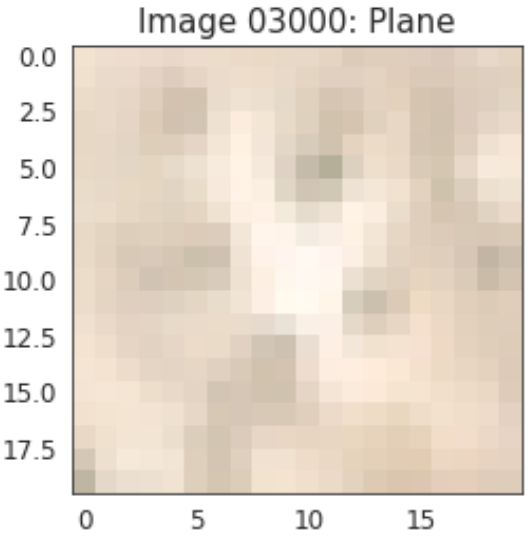
- The 32,000 images are split into the ‘plane’ class and ‘not-plane’ class
- 8,000 are of the ‘plane’ class.
- 8,000 and are a random sampling of different land-cover features, without a plane.
- 8,000 images contain only a portion of a plane.
- 8,000 images are confusers – image chips with bright objects or strong linear features that resemble a plane.



Sample Images



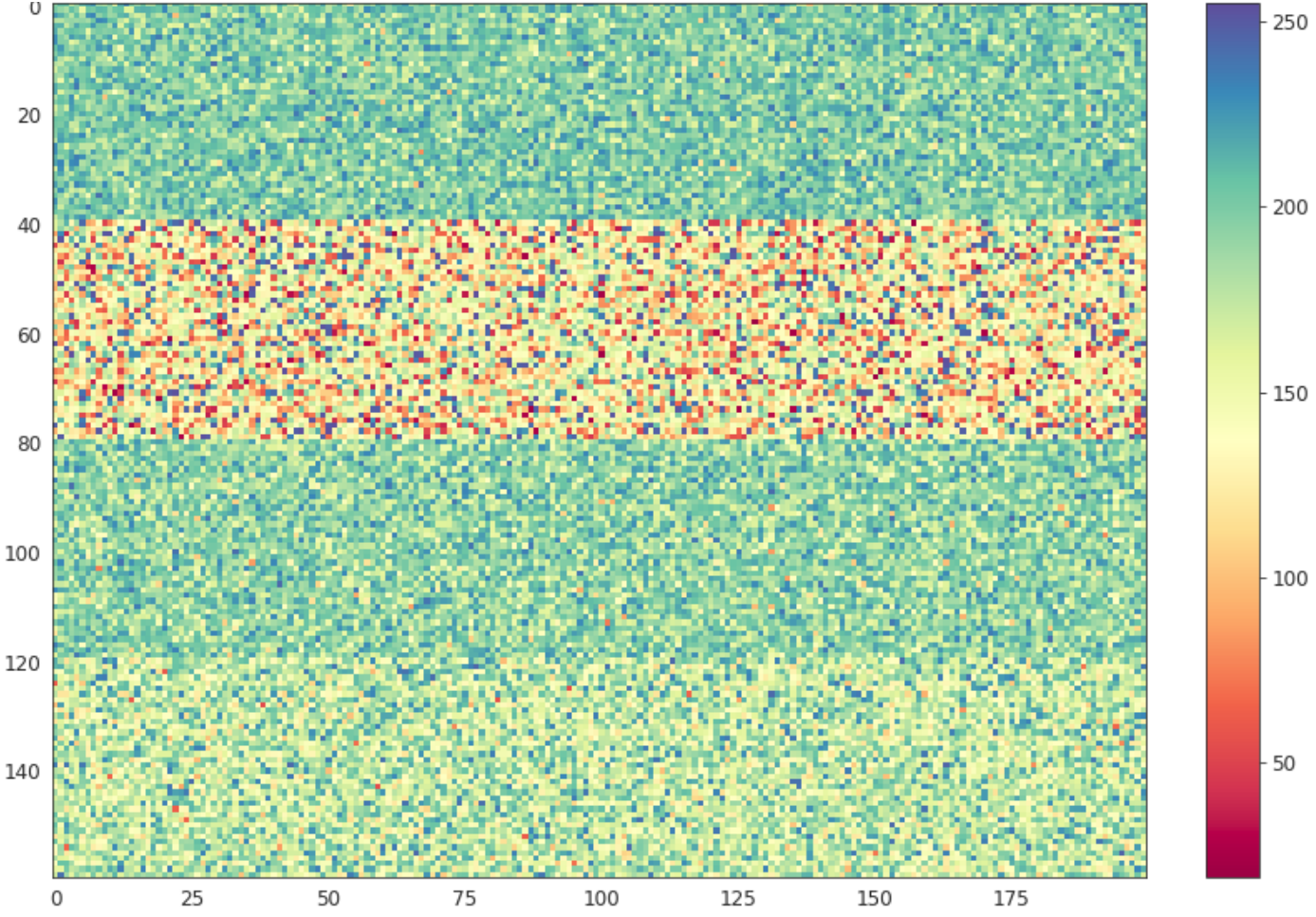




Three sample images from the ‘plane’ class

RMS Display of Satellite Images

- RMS pixel intensity is calculated for each satellite image and displayed for visual inspection.
- RMS displays can be used to visually identify pixel intensity spikes and outliers within the dataset, images that contain relatively high or low pixel intensity can be spotted easily.



RGB Bands

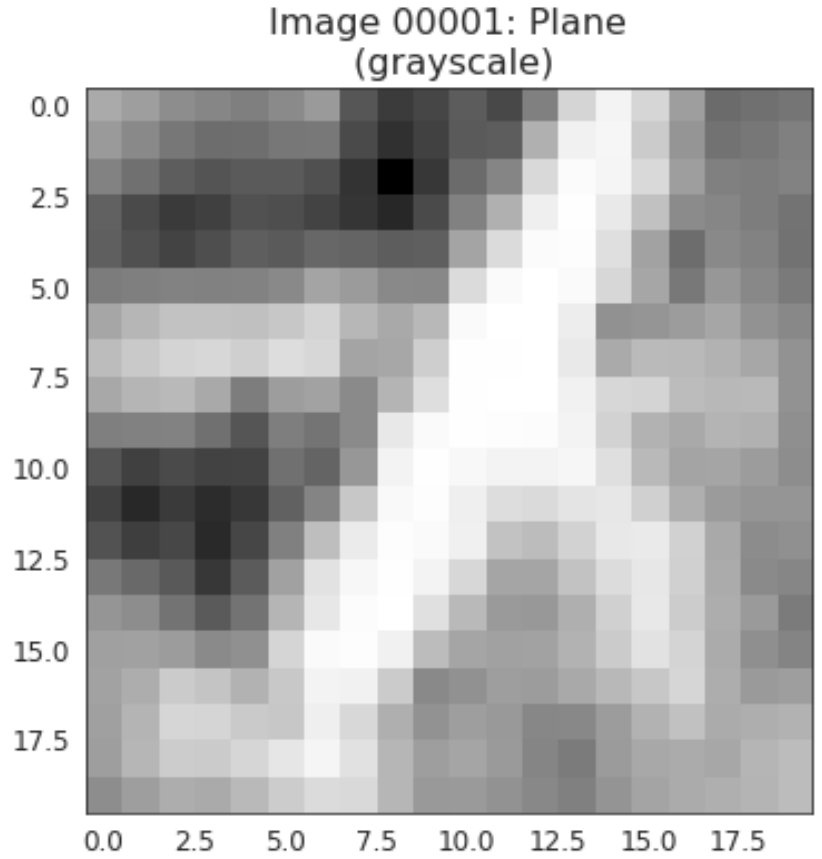
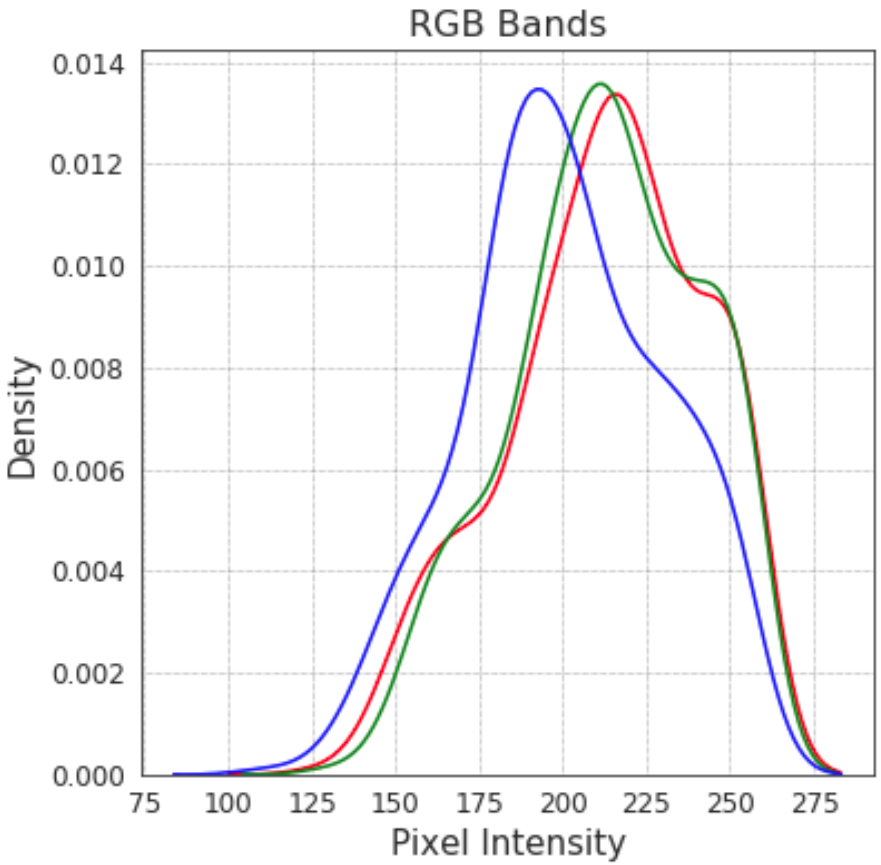


Image 00001: Plane
(grayscale)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


RGB Bands
Density
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
75
100
125
150
175
200
225
250
275
Pixel Intensity

RGB Bands

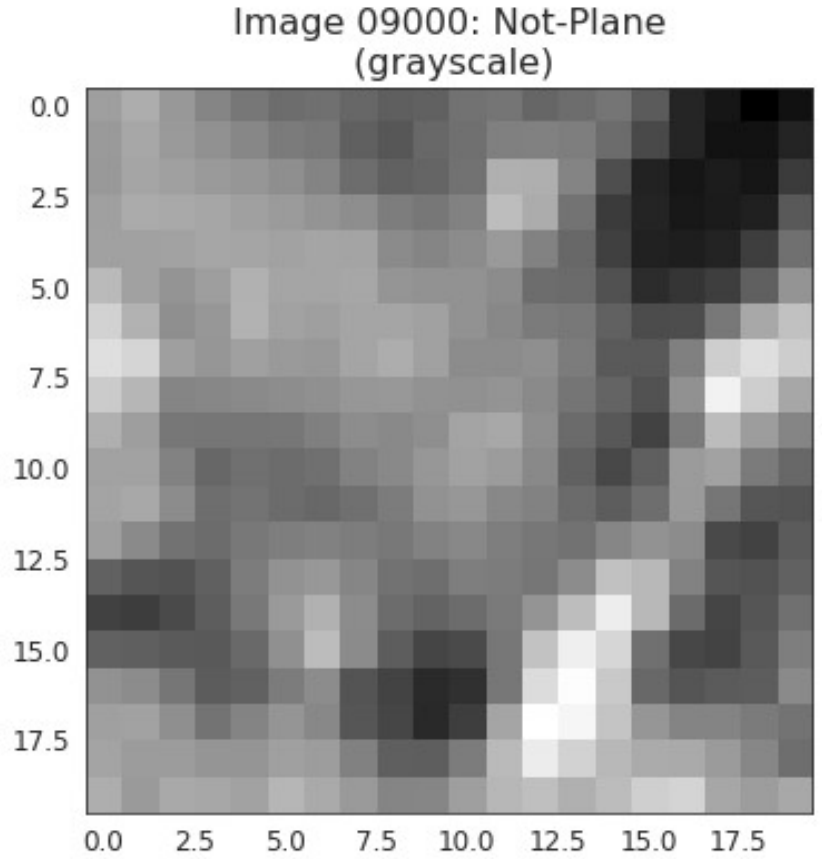
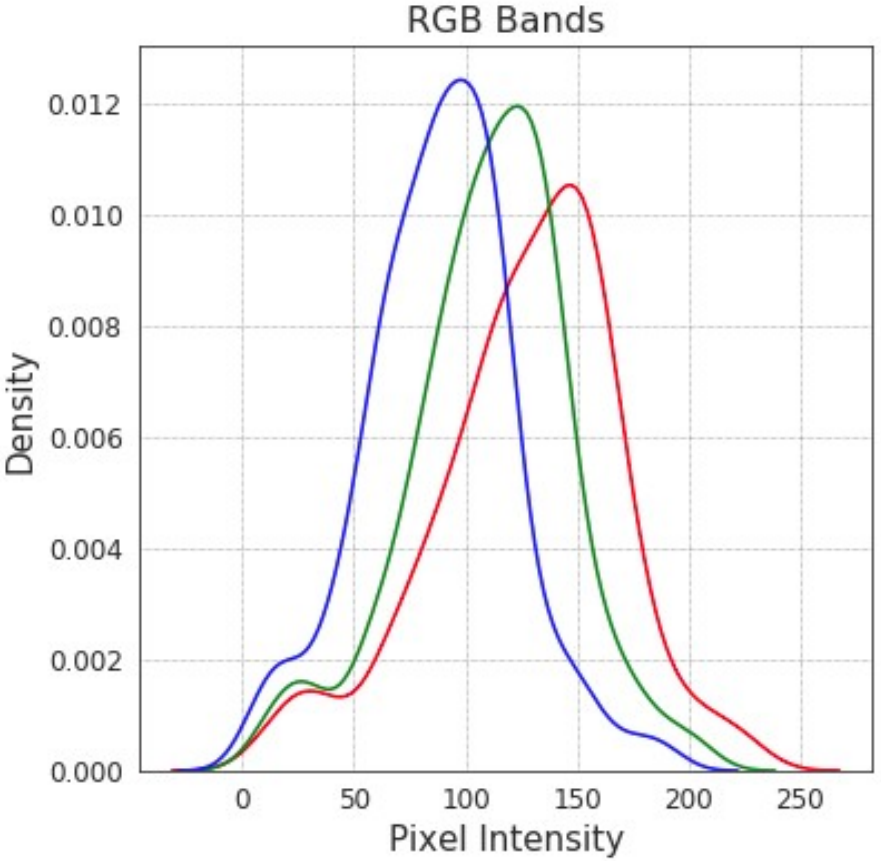
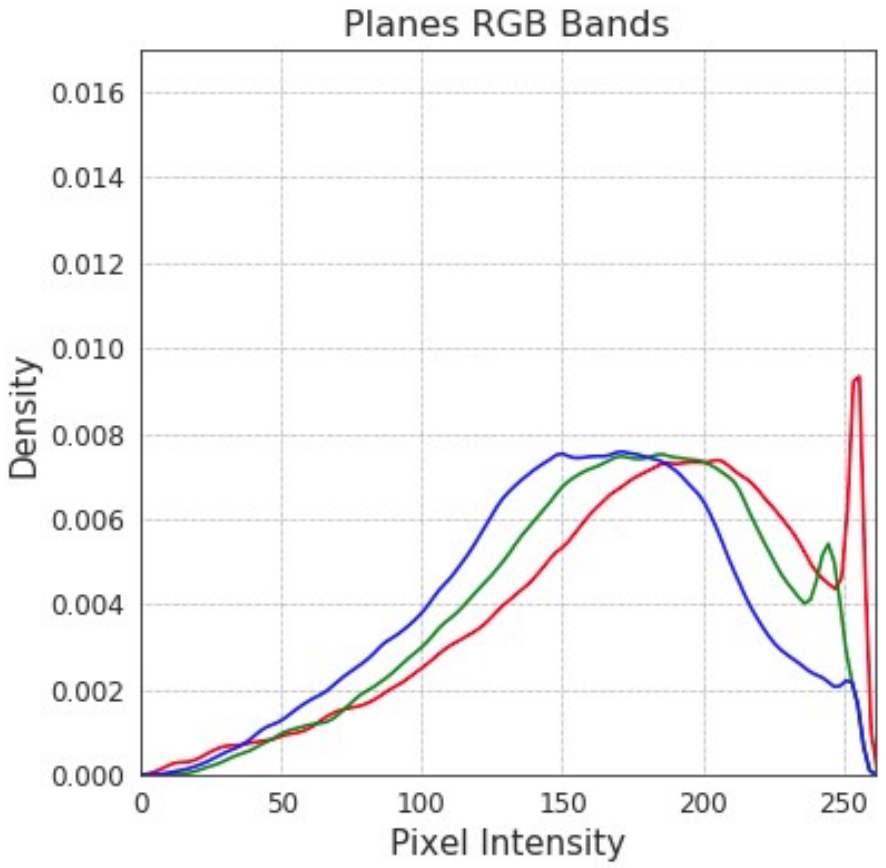
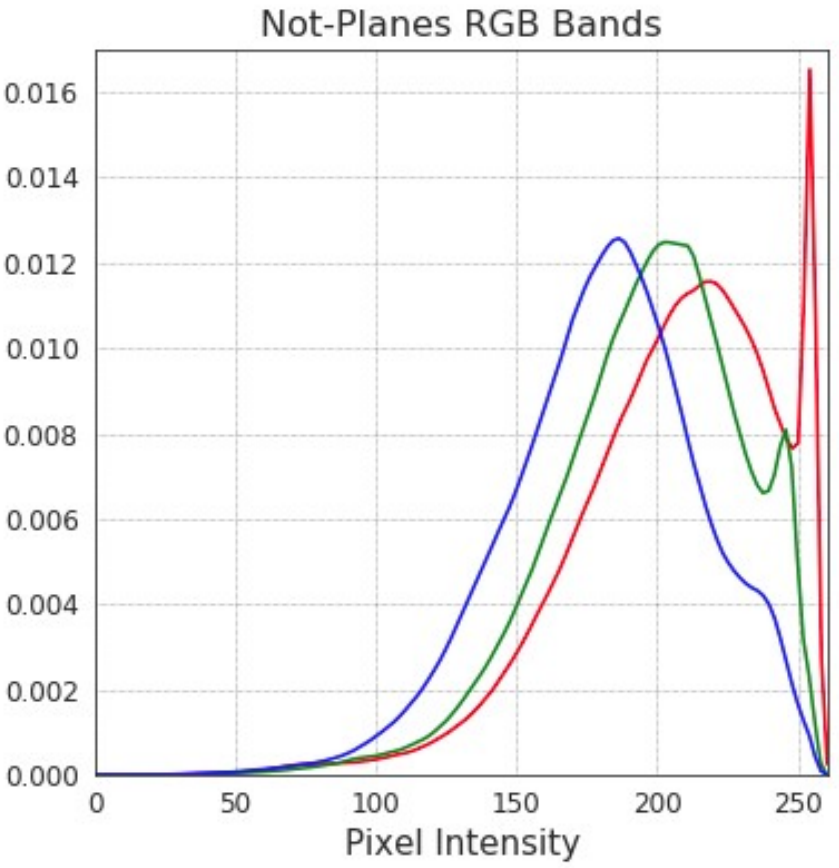


Image 09000: Not-Plane
(grayscale)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


RGB Bands
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
Pixel Intensity

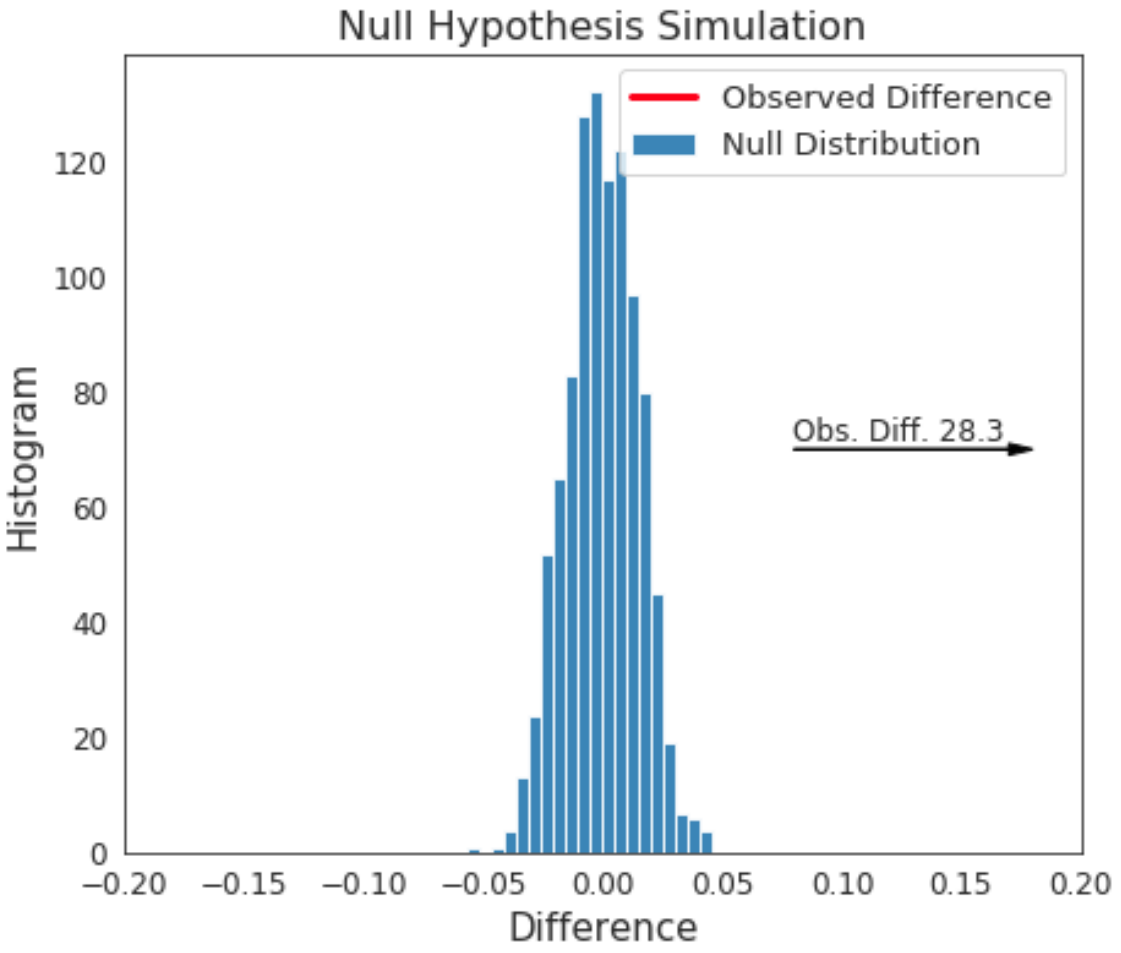
RGB Bands


Planes RGB Bands
Density
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
Pixel Intensity


Not-Planes RGB Bands
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
Pixel Intensity

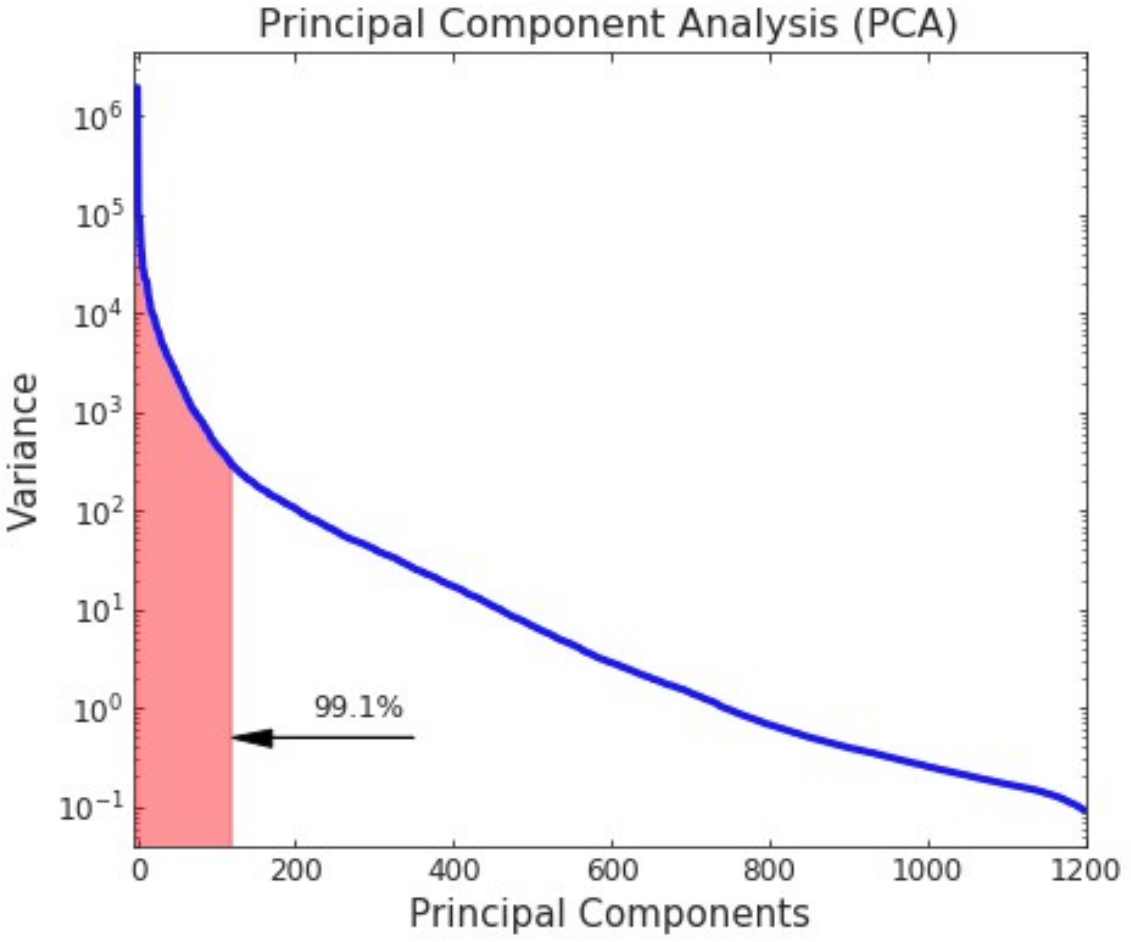
Hypothesis Test

- A hypothesis test is performed to determine whether there is a statistically significant difference in the mean pixel intensity between 'plane' and 'not-plane' classes.
- The null hypothesis is that there is no significant difference in the mean pixel intensity between 'plane' and 'not-plane' classes.
- The alternative hypothesis is that there is a significant difference between the two classes.



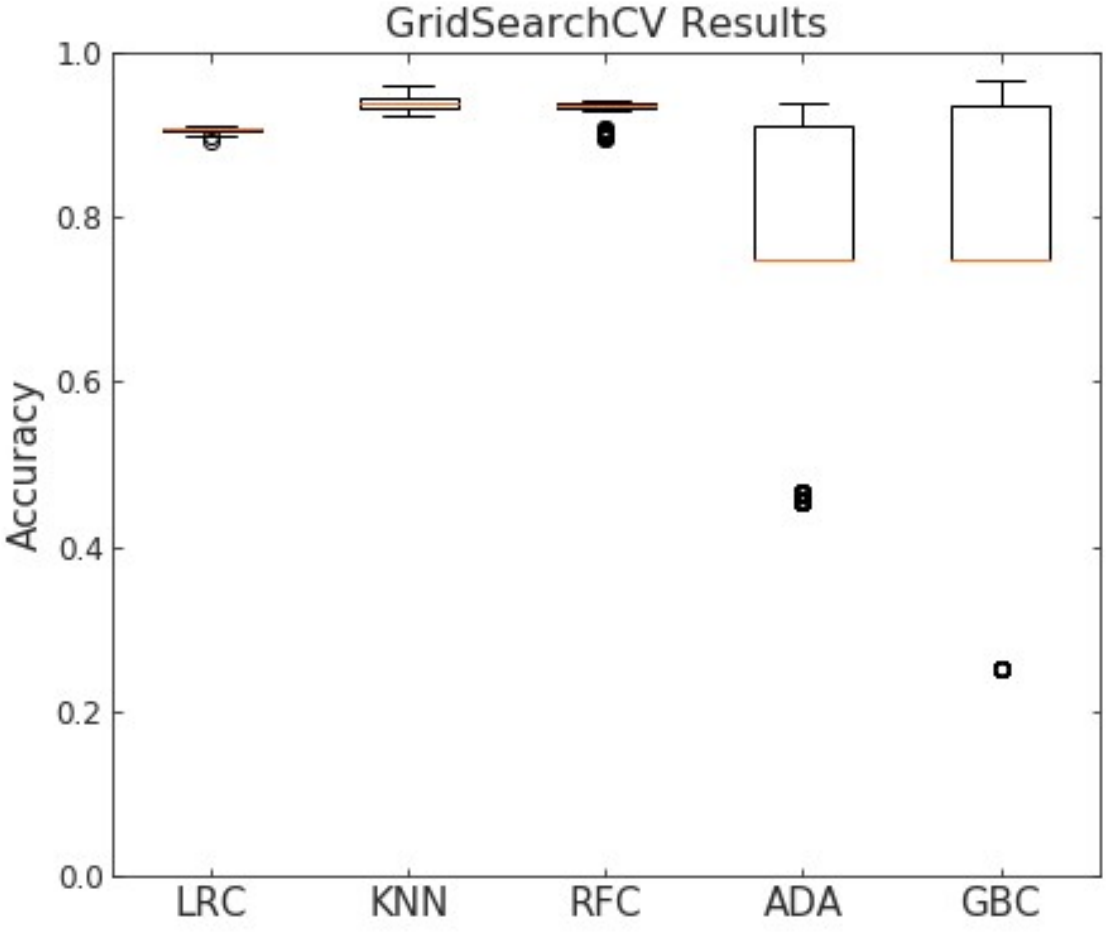
Feature Assessment with PCA

- Principal component analysis (PCA) is performed to assess the explained variance of 1200 principal components (PCs).
- The slope of the explained variance stabilizes around 120 PCs. In addition, 99.1% of the total explained variance is within the first 120 PCs. Therefore, 120 is the number of PCs chosen for input into predictive modeling.

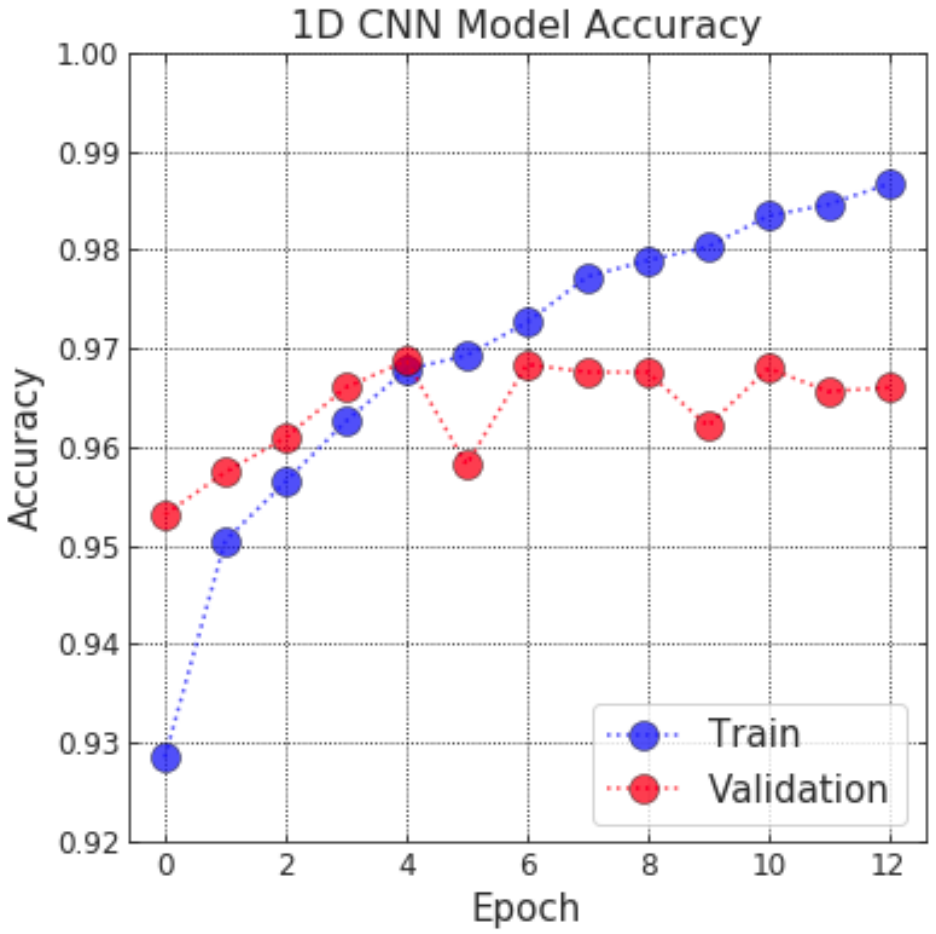


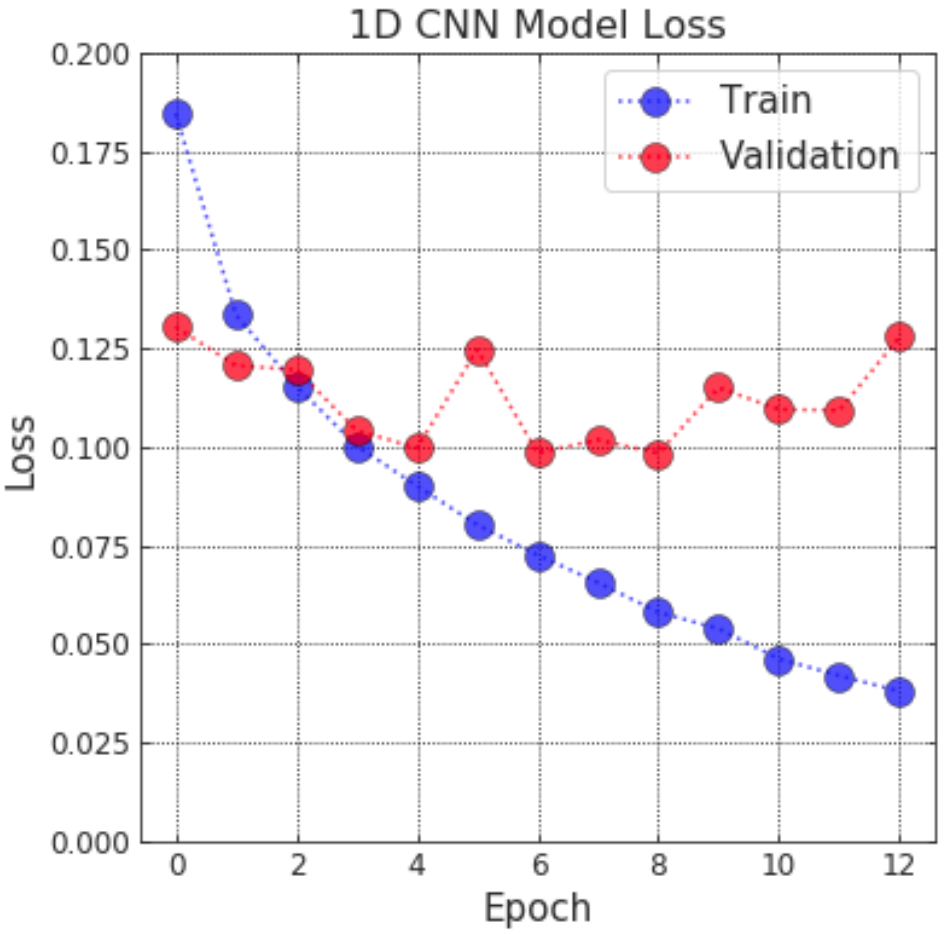
Machine Learning Optimization

- Logistic Regression (LRC), K-Nearest Neighbors (KNN), Random Forest (RFC), AdaBoost (ADA), and Gradient Boosting (GBC).
- LRC, KNN, and RFC result in relatively narrow accuracy ranges, however, both ADA and GBC have significantly low-accuracy outliers, with scores below 0.45 and 0.26 respectively.
- The low-accuracy outliers in the ADA and GBC GridSearchCV results correspond to test runs with the learning rate (i.e., learning_rate) set to 10.0.

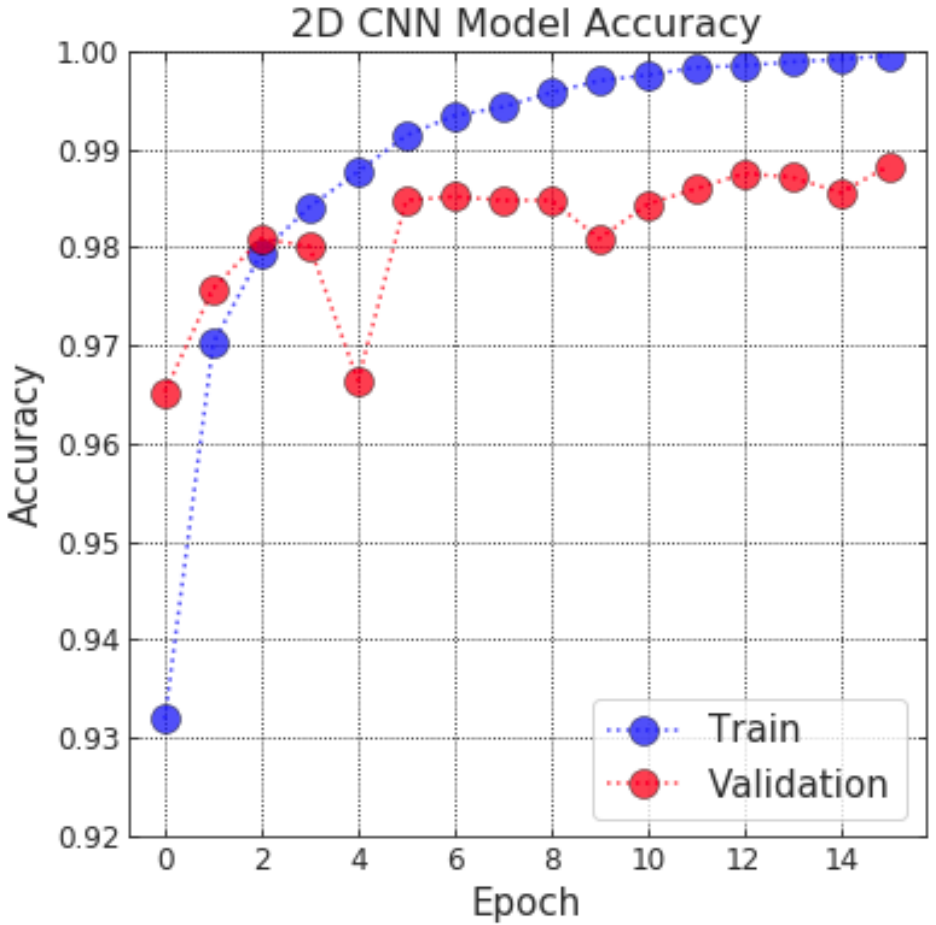


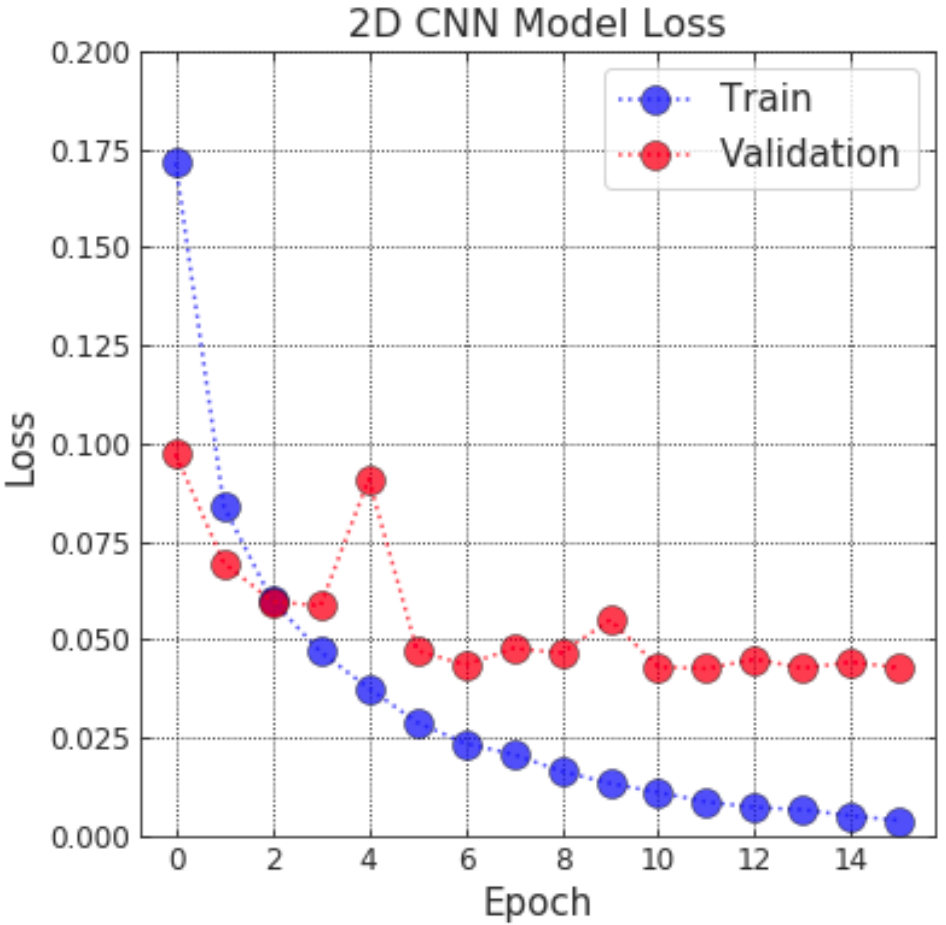
1D CNN Training





2D CNN Training





Conclusion and Recommendation

- The five machine learning models perform poorly. This could be due to poor feature selection. Better results might be obtained by using higher quality/resolution images, or using different image features such as Canny edges and/or band distribution features.
- The 1D CNN and 2D CNN models perform the image classification with test accuracy scores of 95.98% and 98.69% respectively.
- It is recommend that the client select either one of the two CNN models. If compute resources are limited and the ability to scale is important, then the 1D CNN is recommended. If maximizing accuracy is more important, then the 2D CNN is recommended.