Capstone 3: Final Project Report
Springboard Data Science Career Path
Kevin Wolfe

Identification of historic well pads based on Aerial imagery

Introduction:

The United States has nearly 5 million oil and gas wells that have been permitted and drilled in the 160 years of commercial development. While most of the data recorded with the regulatory agencies regarding these wells are based on the best information possible at the time, location information is unfortunately often misreported, particularly for older wells. Many of these errors are due to copying errors where a single digit is transposed or misidentified. These issues are regularly identified by oil and gas operators during their mapping and interpretation process. Specifically, these issues are often identified when data from a well looks "off" and then further investigation is merited.

The US National Agricultural Imagery Program (NAIP) collects areal imagery yearly of the acreage withing the United States to provide land use information. This dataset is freely available and collected yearly starting in 2003. The resolution of this data has historically been 1m resolution however the most recent data in many states has been collected at finer resolution (0.5m resolution). The goal of the project is to utilize the NAIP data to generate a computer vision model capable of validating the presence of an oil and gas pad at the location identified by the regulatory agency.

Databases:

The initial list of well header information, which includes a Latitude and Longitude for each conventional well in Westmorland Co. PA, will be generated from the PADEP's open data GIS portal (https://newdata-padep-1.opendata.arcgis.com/datasets/oil-gas-locations-conventional). Along with the well's location, this table also includes information about the well's vintage, current status, and API number (a unique identification number given to every well in the US).

The location information will be used to select from the satellite/areal imagery provided by National Agriculture Imagery Program (NAIP: National Agriculture Imagery Program | Earth Engine Data Catalog (google.com)). Google has provided a platform in which this data can be easily accessed called Earth Engine which provides access to a myriad of geospatial imagery information.

Potential Clients:

The primary client that would be able to make use of a dataset of verified well locations would be oil and gas operators. This is because the mapping done by oil and gas professionals is often hindered by incorrect location information. With a positive result of this project there would be a list of locations

to provide additional scrutiny were they included in mapping. A secondary client could be the regulatory agencies who are the primary source of the location data that will be evaluated.

Data Wrangling:

The initial database of potential well locations was collected from the PADEP's well header dataset for Westmorland Co. for conventional oil and gas wells. The well header has 4419 well locations with latitudes and longitudes. Utilizing this dataset in combination with Google's Earth Engine an 84mx84m image was cropped centered on the state well location from the NAIP dataset. This was done by taking the Latitude/Longitude applying a 250 ft buffer within geemap (a python library designed to interact with Google's Earth Engine) which was used to crop the NAIP imagery for each well location.

Once the initial imagery dataset was created using the above method, utilizing my industry knowledge of the signs of a well location to categorize the images in two groups, pad visible and pad not visible. There are several signs of a well location that are visible even with relatively low-resolution imagery. The first and most obvious is a visible clearing at the center of the image. A secondary feature that is almost always visible are linear features leading toward this clearing like lease roads and pipeline right-of-ways. Once the images were classified, they were stored as an array of arrays for analysis. With a corresponding array of for the classification (1 for visible, 0 for not visible).

Data Exploration:

Initially the group of images contained 3457 potential locations based on the coordinates from the well header data. This dataset is highly unbalanced with 2958 locations I was able to identify the well pad and 499 where the location was not visible (Fig. 1, left). To deal with this imbalance the not visible data was rotated 90, 180 and 270 degrees and then the original and rotated images were also flipped to bring the not visible image count to 2899 (Fig. 1, right).

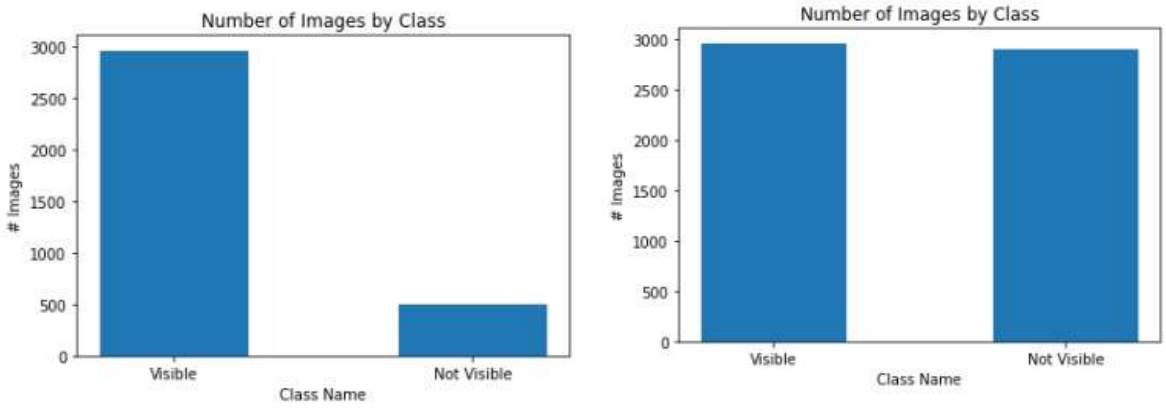


Figure 1: Left, distribution of the initial dataset of conventional well locations. Because of this imbalance the not visible images will need to be rotated/flipped to generate additional samples for model training. Right, after rotating and flipping Not Visible images the dataset is more even.

The shape of the images is very consistent. Both height and width range from 83-84 and 84-85 respectively (Fig. 2). In the modeling stage I will need to reshape my images to a consistent size (84x85 was used).

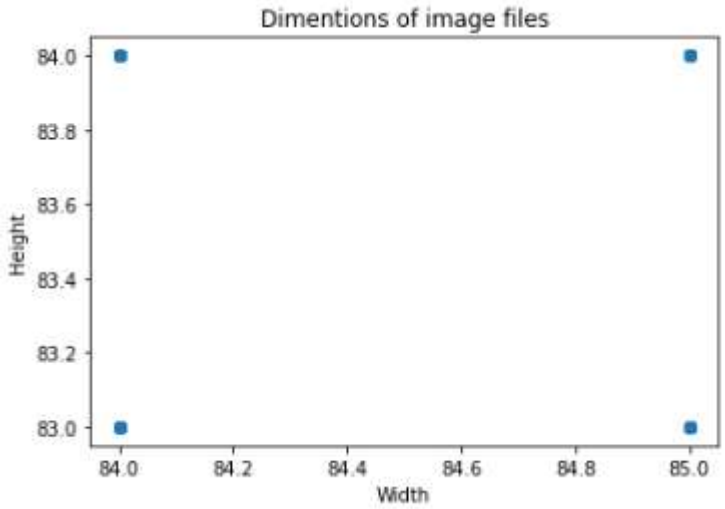
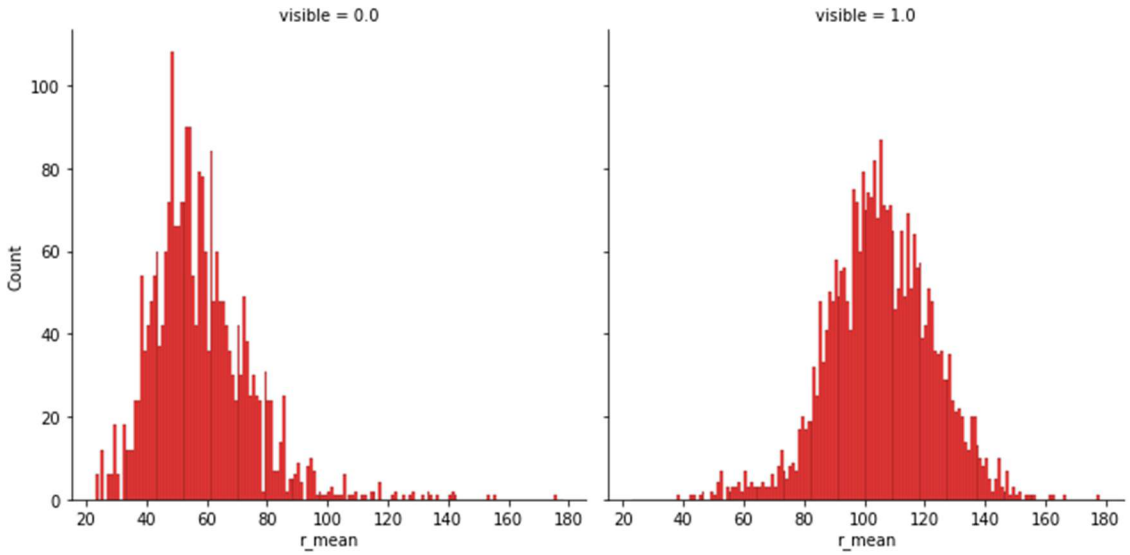


Figure 2: The dimensions of the images within the dataset. The process by which the image data was generated there was very little variance in the height and width in the image sizes.

The images from Geemap are contain three color channels. Looking at the mean value for each color channel within the images green tends to be the most dominant color in the images, followed by red, and blue tends to be the lowest valued channel in the images (Fig. 3). A second observation is that the distribution for the images categorized as not visible are shifted to generally lower values than what was observed in the images categorized as visible. Figure 4 presents the distribution of the standard deviations for these color channels. The general shape seems to be very similar between the visible and not visible categories, however the peak is shifted to slightly higher (<10) standard deviations for the visible images.



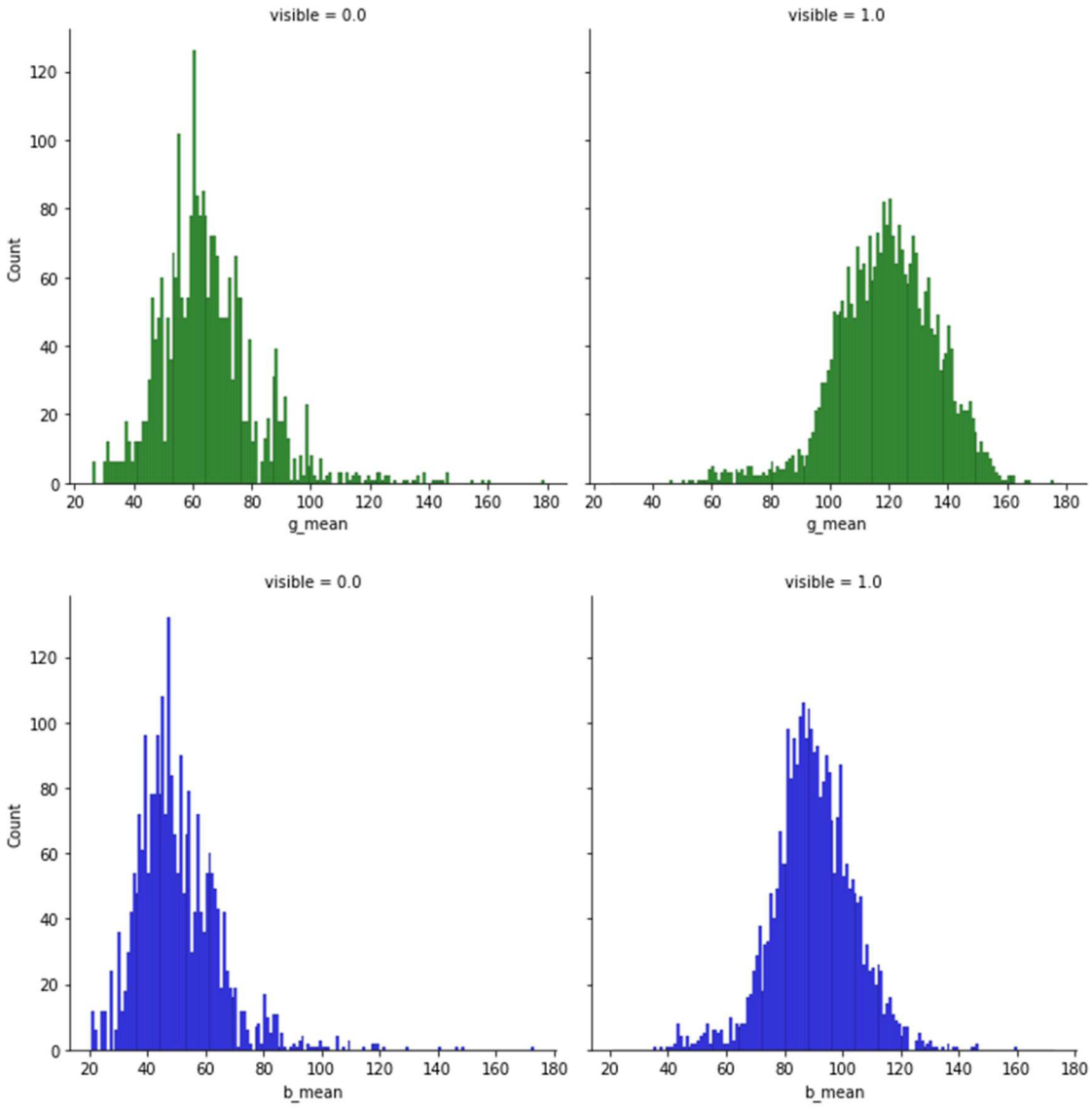


Figure 3: Histograms of the mean value for each color band separated into visible and not visible charts. The distribution of means is visibly lesser in values for blue color bands than green and red. All three color bands have near normal distributions of means.

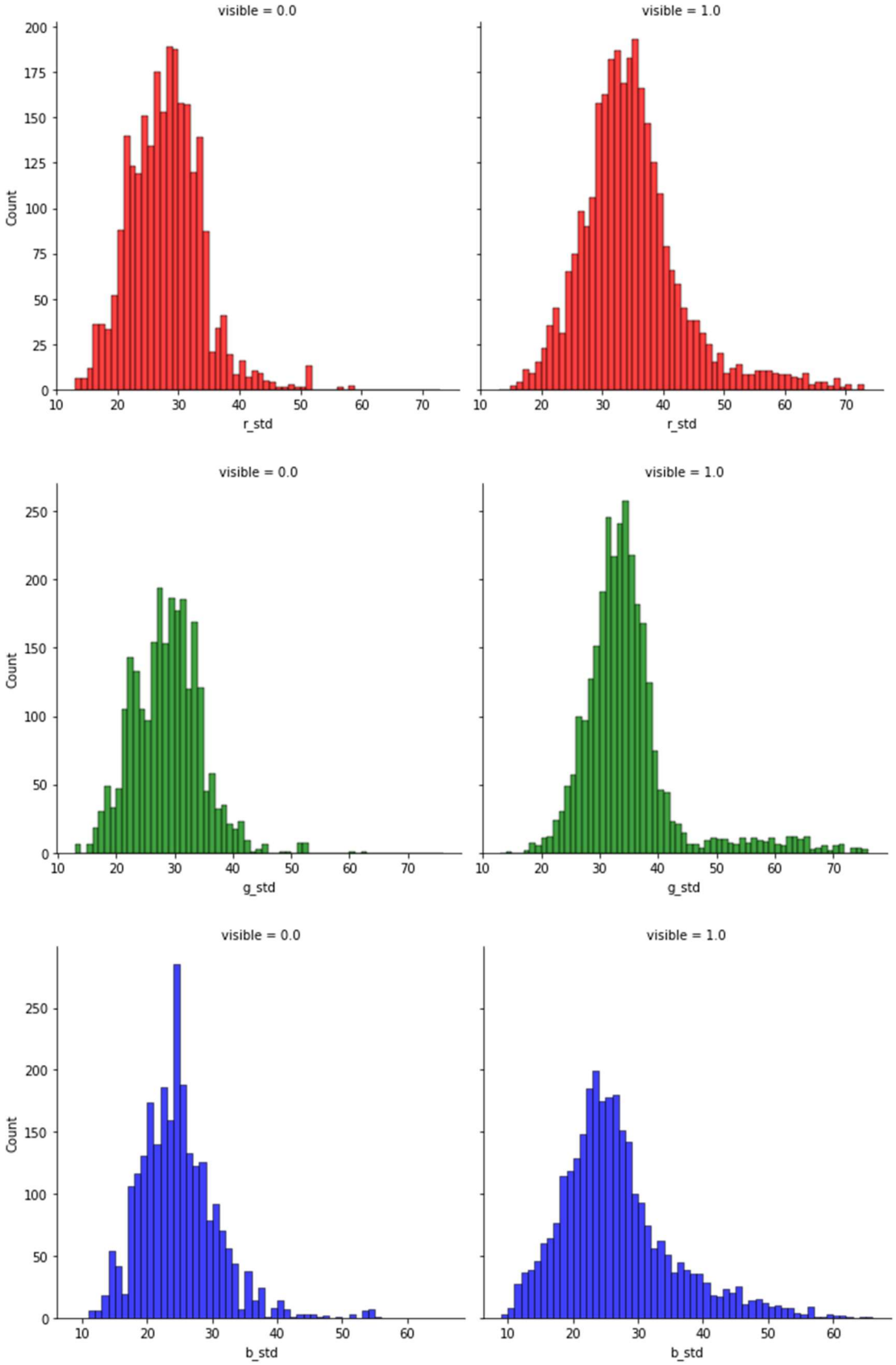

visible = 0.0
visible = 1.0
Count
r_std
visible = 0.0
visible = 1.0
Count
g_std
visible = 0.0
visible = 1.0
Count
b_std

Figure 4: Histogram of standard deviation of each color band by visible category. For all three color bands the standard deviations seem to peak between 25 and 35 for all color band and visible category options. One interesting trend in this data is that the images with a visible pad seem to have an increased spread in standard deviation when compared to the not visible pad images.

In addition to investigation of the coloration of the images I also utilized openCV to evaluate patterns within the images. Particularly of interest to me was the ability to inspect which parameters seemed most effective in highlighting the features within these images. For example, using canny edge detection aperture size of 3 was able to bring out the linear features visible within images (Fig. 5a). Interestingly, larger aperture sizes failed to produce similar results, which may suggest smaller window sizes are best for this dataset. Sobel edge detection which also utilizes a 3x3 matrix for edge detection yielded similar results (Fig. 5b)

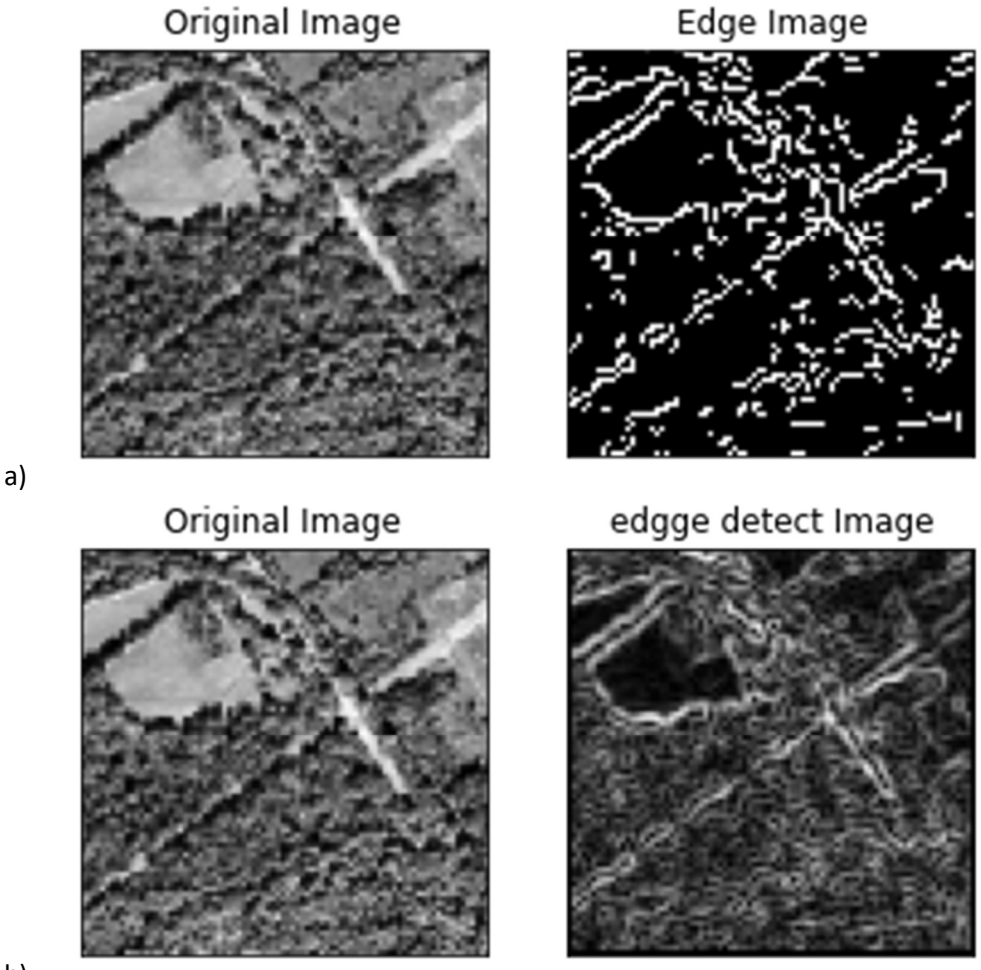


Figure 5: An example of the results of canny edge detection (a) and sobel edge detection (b). Linear features like the pipeline right-of-way running through this image were highlighted through these methods. The edges of fields like the one in the upper left also show up with this method (often somewhat more distinctly than the path running through the woods from a pipeline or road).

Data Modeling

Prior to modeling the balanced database of 5857 images (3457 original images plus 2400 created through rotating and flipping images without a visible pad) was split into 80/20 training and testing datasets. The selection was performed on the well header dataframe and then numpy arrays were built based on these selections to allow for Keras modeling. Three models with increasing complexity were then created in attempts to predict if a pad was visible at a well location (Table 1).

The first model was a dummy model that predicted based on the populations of the classifier. Due to the addition of the rotated/flipped versions of the not visible locations, 50.5% of the images were classified as having a location. The expectation is that a dummy model (one that picks 1 or 0 at random based on the population) should have an accuracy close to this percentage. As expected, the dummy model resulted in a 51% accuracy score which is not going to be sufficient for our needs but is an excellent baseline to compare subsequent models to.

The initial Keras model created contained a single convolution layer followed by a flattening and dropout layer prior to a sigmoid activated layer for prediction. This model does incredibly well for as simple as it is. After 5 epochs of training the modeled data was at near 100% accuracy however the test data yielded a 97% accuracy (both excellent). When evaluating this model against the test data both precision and recall of visible and not visible yield nearly 97% as well.

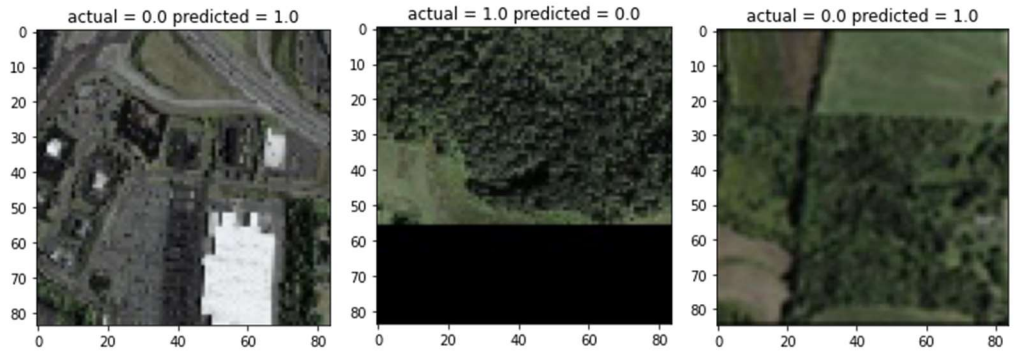


Figure 6: Examples of incorrectly predicted images. While the first is within an area that is now a shopping area and obviously no longer has a well location, I can see why the model struggled on the other two examples. One was missing a portion of the image and the pad location is in shadow, the other has linear features in the image that often would be associated with a well location, in this image they do not appear to be roads/pipelines.

The second Keras model that was created had an additional convolution and MaxPooling layer beyond what was used in the initial Keras model. After 8 epochs this model resulted in a 97% accuracy on the training data and a 95% accuracy for the testing data. This model is suffering compared to the less complex model with precision for not visible and recall for visible pad images (both ~91%). Overall, this model performs poorly enough that the previous model will be considered the best model.

Table 1: Results for well location identification models. Based on the better performance of the Keras 1x Conv model it is considered to be the final model moving forward. Were the initial Keras model not to have performed so well I would have been happy with the results of the Keras 2x Conv model.

Model	Training Score	Testing Score	Precision 'visible'	Precision 'not visible'	Recall 'visible'	Recall 'not visible'
Dummy	0.516	0.511	0.557	0.456	0.548	0.465
Keras 1x Conv	0.997	0.97	0.975	0.962	0.968	0.969
Keras 2x Conv	0.972	0.955	0.995	0.905	0.915	0.994

To evaluate the ability for this model to generalize to a different area I utilized Geemap to select the well locations in Sweetwater Co. WY and then a grid of locations within a non-oil and gas producing region of Idaho surrounding the town of Gooding. I selected Sweetwater Co. WY because this area does still possess similar amounts of oil and gas activity and infrastructure but has a significantly different landscape. Instead of the heavily forested areas of western PA, there is very little vegetation in Sweetwater Co. Gooding, ID was selected because I know there is not any oil and gas development there and based on aerial imagery looks more like Sweetwater Co. than western PA.

After evaluating the new image data with the final Keras model, I found that 25% of the locations within Sweetwater Co. were classified as not having a pad. Doing a quick visual inspection of the two classifications there many of the locations classified as not having a well definitely do have a pad (Fig. 7). An observation clear to me is that these images lack contrast when compared to the PA data.

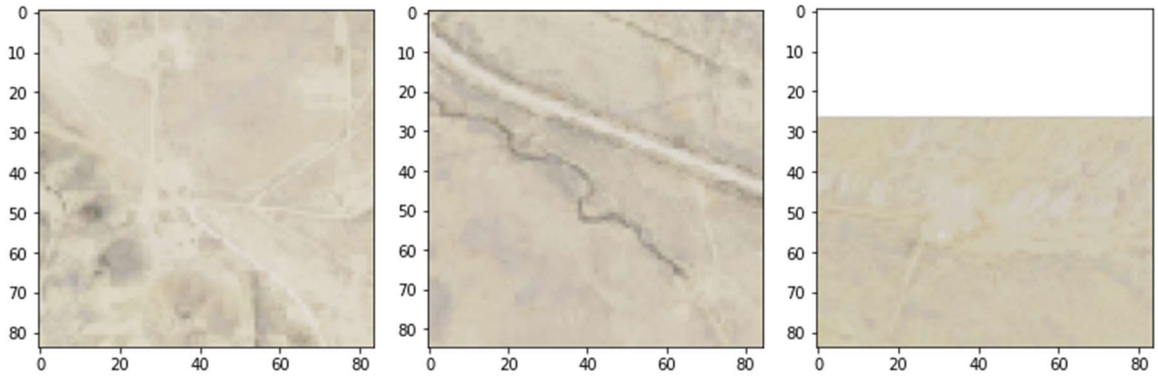


Figure 7: Example locations classified as not having a pad that have clear signs of lease roads providing access to the well locations. These 3 examples were within the first 4 images classified as not having a pad suggesting that the model is struggling to identify well locations in this area.

The model struggled even worse in the group of images from the area surrounding Gooding, ID where there is no oil and gas development. 75% of the random locations surrounding Gooding were classified as having a well. None of these locations have a well pad, so all the 'visible' classifications are incorrect. Like with the Sweetwater Co. WY locations there is little contrast between the features within the image which is likely hurting the model's ability to predict.

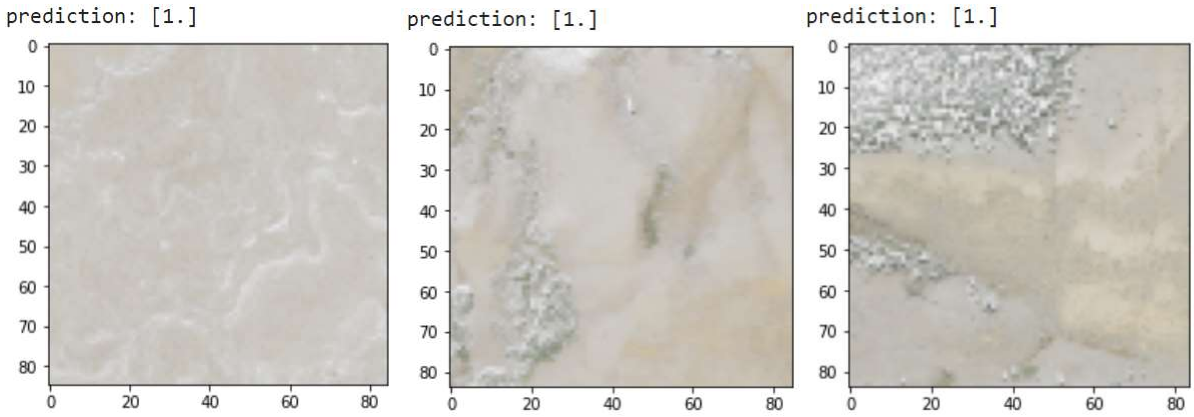


Figure 8: Locations from the area surrounding Gooding, ID that were classified as having a well pad. Although the color is significantly different than western PA the rocky landscape in areas looks to have similar texture as the forested areas of Westmorland Co.

Potential Utilization:

Based on the performance of the final model on the test data in Westmorland Co. PA it can be said that this model would be excellent in identifying suspect locations in this area. These suspect locations could be either excluded from mapping, provided lower weighting, or be flagged for further scrutiny after mapping. Utilizing this model in areas outside of western Pennsylvania was significantly less successful particularly in ID where there were no well locations this model massively overpredicted presence of well locations. The failing in this model was likely due to having training data from a single geographic area. Despite providing the model a balanced training set of locations with and without well locations, the model generalizes poorly to landscapes that look significantly different than western PA.

Further Work:

The main failing of this model is its poor generalization to areas outside of western PA. To improve the generalization of this model for areas outside of western Pennsylvania inclusion of locations from a wider variety of areas would be required. In the investigations I conducted post modeling I have provided one potential avenue for generating additional training data from areas outside of western PA. Additional work would need to be conducted to validate those additional visible locations were correctly categorized, but I would recommend including more locations that do not have oil and gas activity like Idaho to promote generalization.