Development of a Machine Learning Algorithm for Classification of Red and Green Airglow Images

Matthew Burnes, Brian Harding
Space Sciences Lab, University of California, Berkeley

MANGO Network

Introduction & Motivation

The Midlatitude AllSky-Imaging Network for Geospace Observations (MANGO) uses wide-field imaging and high-resolution spectral interferometry to capture red and green airglow images[1]. Presently, analysis of these images is difficult due cloudy images needing to be manually sorted out of the dataset. However, machine learning classification algorithms can efficiently bin binary data using logistic regression. For this project, we propose the use of two supervised machine learning classification algorithms for the categorization of greenline and redline images into useable and unusable bins. This automated quality control will enable the long-term, reproducible analysis of the MANGO dataset.

Data Preparation

Image data was used from the MANGO dataset. The data used in this study spans from 2022-2024, encompassing twelve sites.
Redline images are filtered through a 2nm bandwidth centred at 630.3 nm, And the greenline images are filtered using a 2nm bandwidth centred at 557.7nm. To set up a base layer of images that can be used to train the model, 500 images were manually classified into clear (visible stars, no moon interference), cloudy (image dominantly covered by clouds) and contaminated images (images with some other major interference). These images were sampled randomly across available times into groups of 200 clear images, 200 cloudy images, and 100 contaminated images. In the classifier algorithm, the image data is separated into a training set (60% of the data used to train the model), and a testing set (40% of the data used to test the model). We also created a validation set of 50 images from a site that the model was not trained on to test how well it did in site-independent classification.
All inputs are in their raw form,
a 2D array of pixel brightnesses.

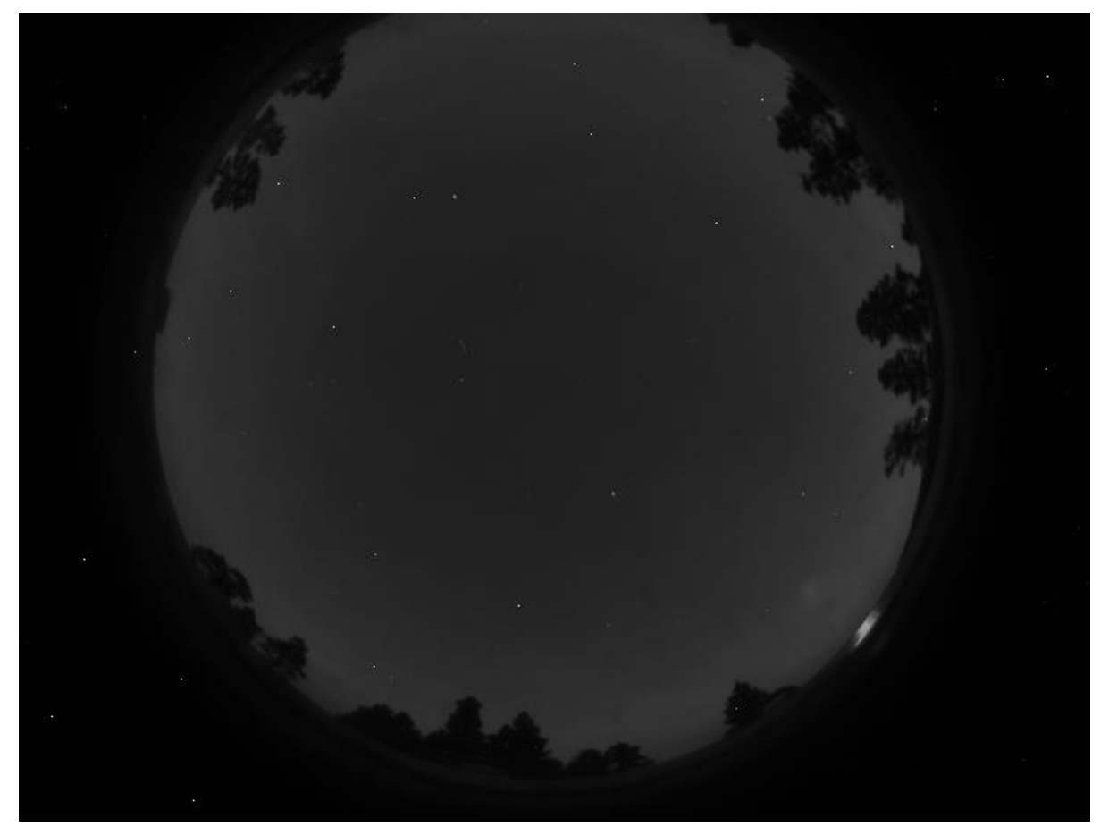

Fig 1: *An example of an image that is classified as 'clear'*

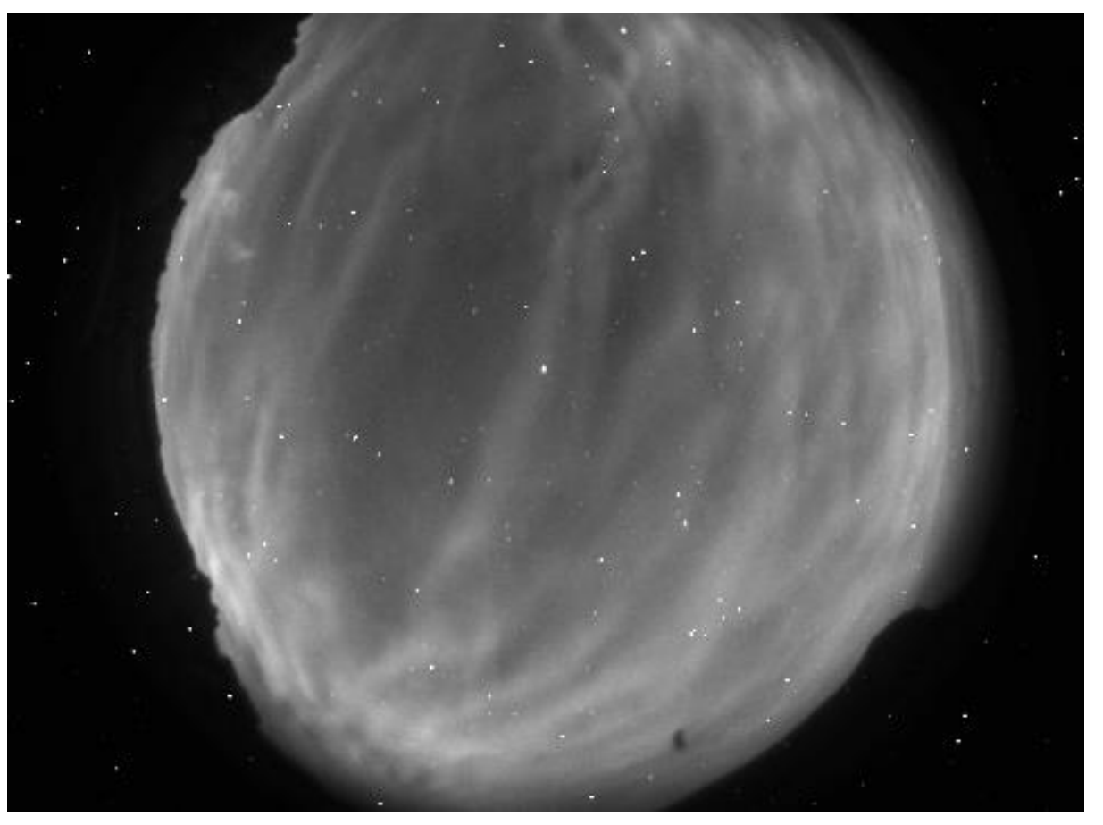

Fig 2: *An example of an image that is classified as 'cloudy'*

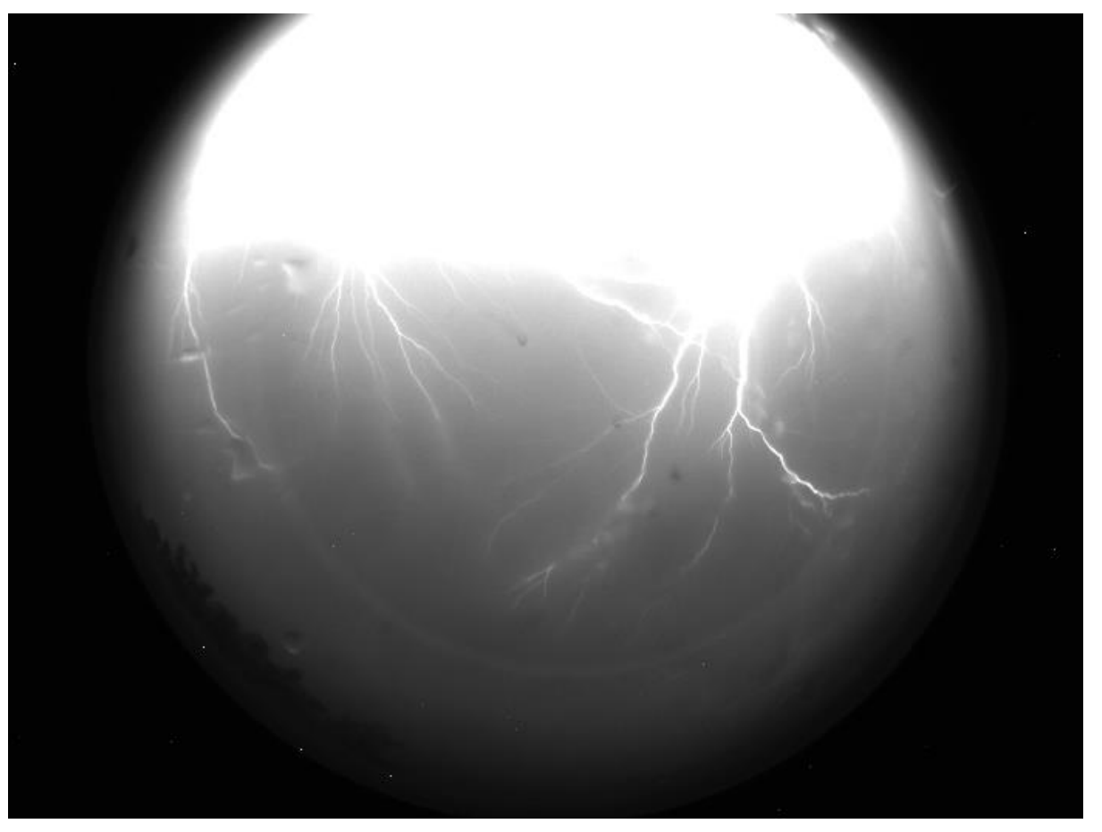

Fig 3: *An example of an image that is classified as 'contaminated'*

Algorithm Chaining

As logistic regression is limited to binary classification, we filter out contaminated data that would otherwise pollute the dataset using a K-nearest neighbour algorithm. The resulting data labelled as non-contaminated is then fed into the logistic regression classifier to sort into clear and cloudy data.

Classifier Algorithm Inputs

Input Pre-Processing

All images are stored as their raw form, a 2D integer array where each point represents a corresponding pixel's brightness. Each image is cropped to remove the interior of the camera lens from the photo, then normalized and run through a 2D Fourier transform. We found the classifier was significantly better at identifying clear images over cloudy images, likely due to the larger variance in appearance when an image is cloudy. To account for this, the algorithm is 1.5x weighted towards a classification of cloudy, so that when an image is difficult to identify, it is more likely to be set as cloudy to prevent pollution of the clear dataset.

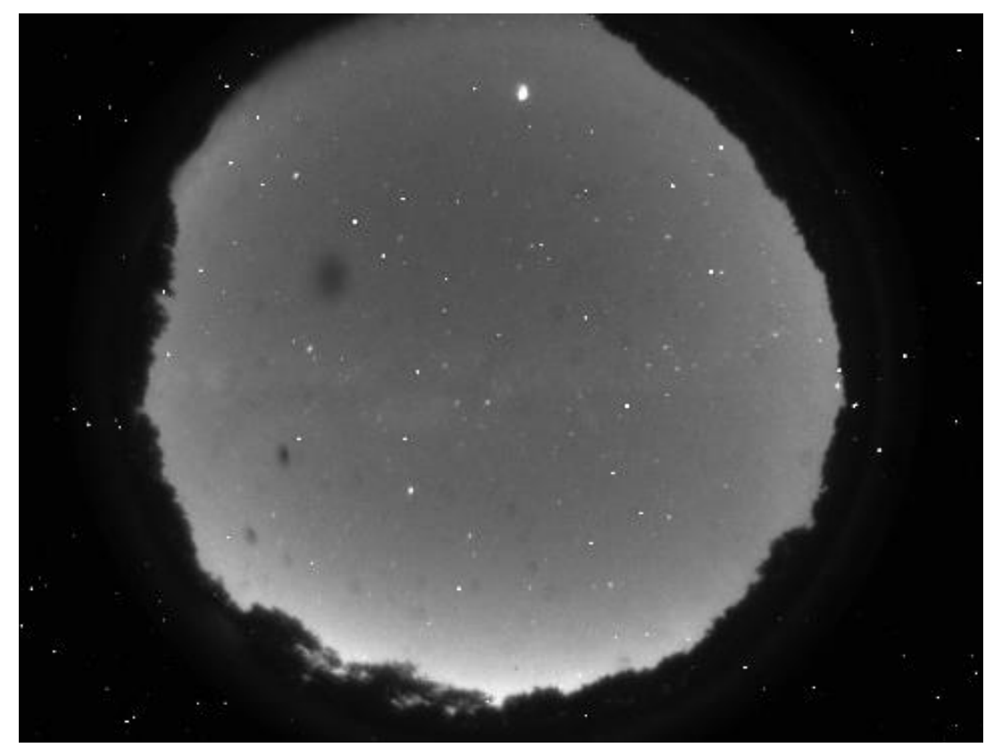

Fig 4: *A pre-crop image*

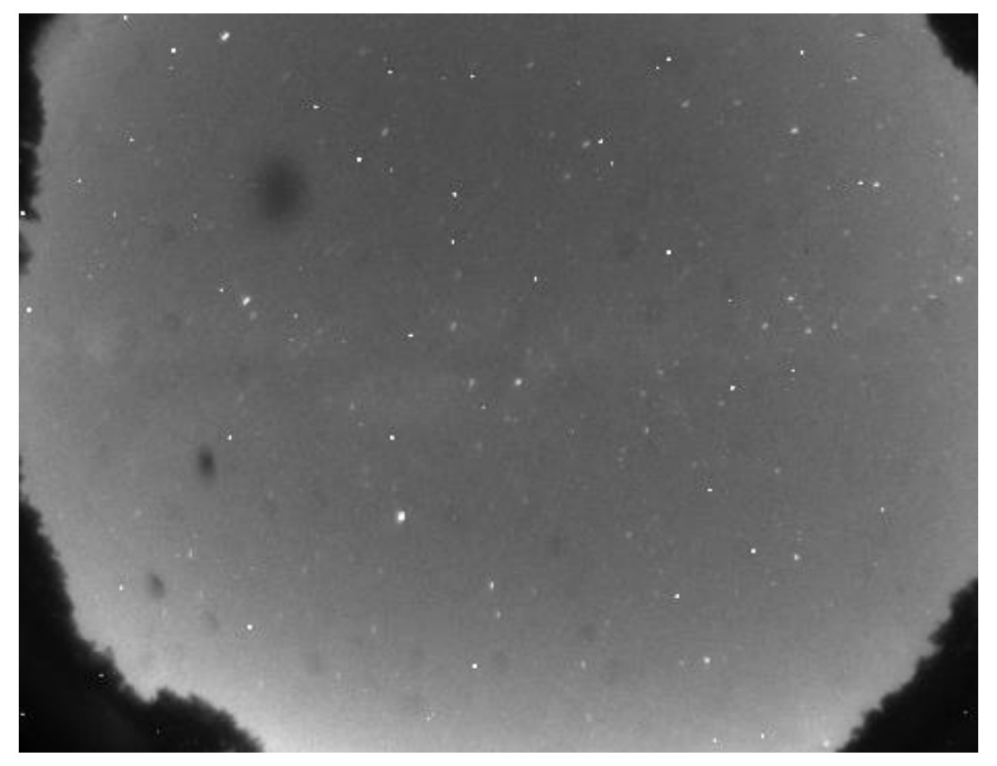

Fig 5: *A post-crop image*

Feature Selection

Mean Brightness: Mean brightness is extracted from each cropped image as cloudy images have a higher average brightness than clear images [2].
Fourier Features: When used for image analysis, Fourier Features can identify geometric texture features which work well with classification algorithms [3]. To extract the four 'best' Fourier Features, each training image was run through a fast 2D Fourier transform. Then, every matching point from each image was grouped together to find the correlation coefficient between each Fourier feature and the classified images. The Fourier features whose groups had the most significant correlation coefficients were used as features for the classification algorithm. The four features that showed the largest significance for classification were at points [1][3], [349][0], [28][449], and [349][3].

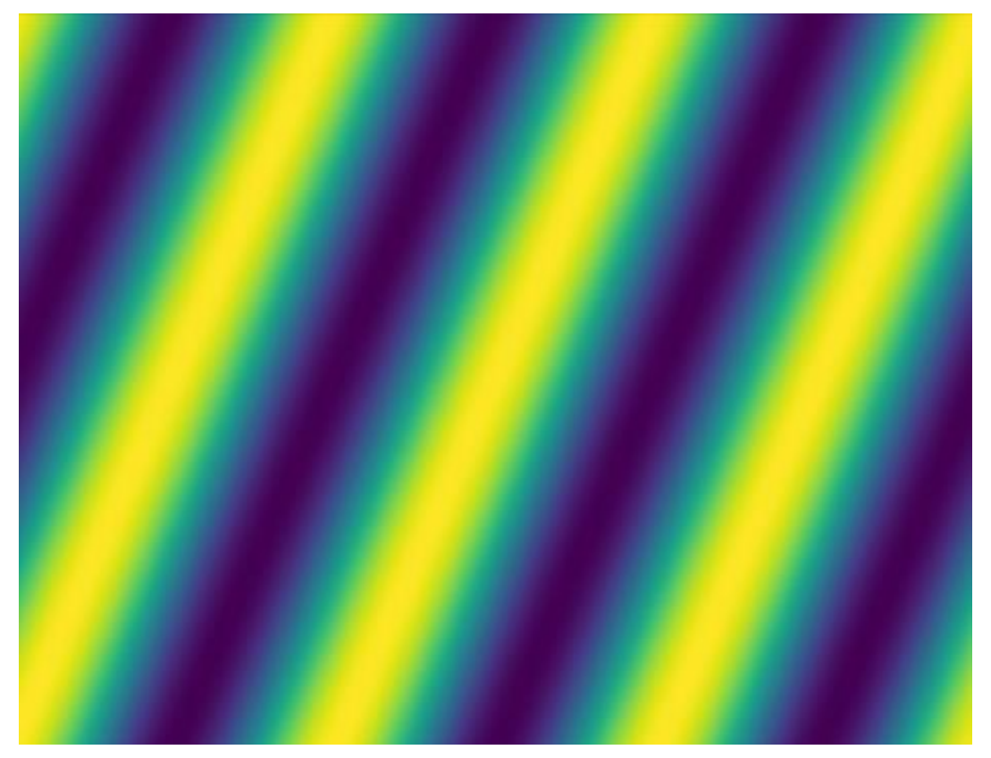

Fig 6: *Fourier Feature at point [1][3]*

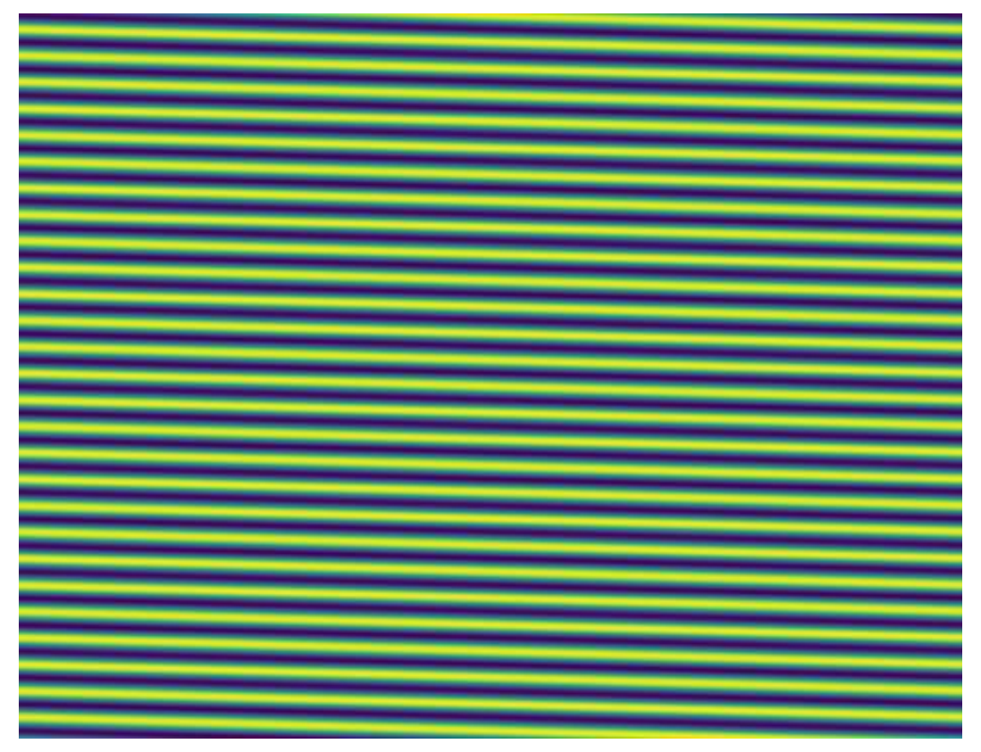

Fig 7: *Fourier Feature at point [28][449]*

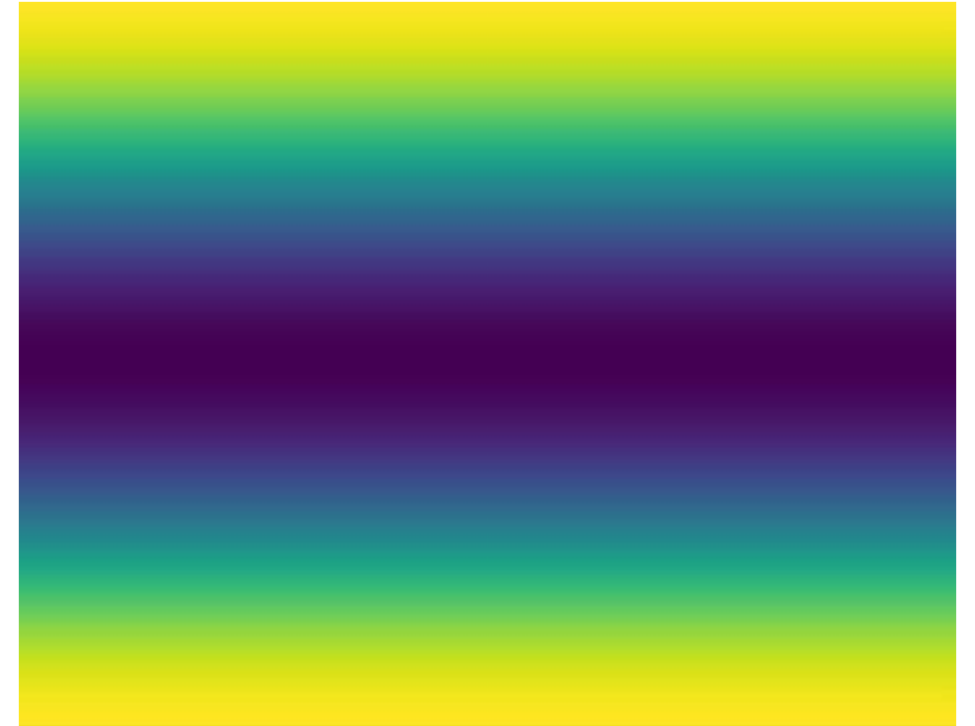

Fig 8: *Fourier Feature at point [349][0]*

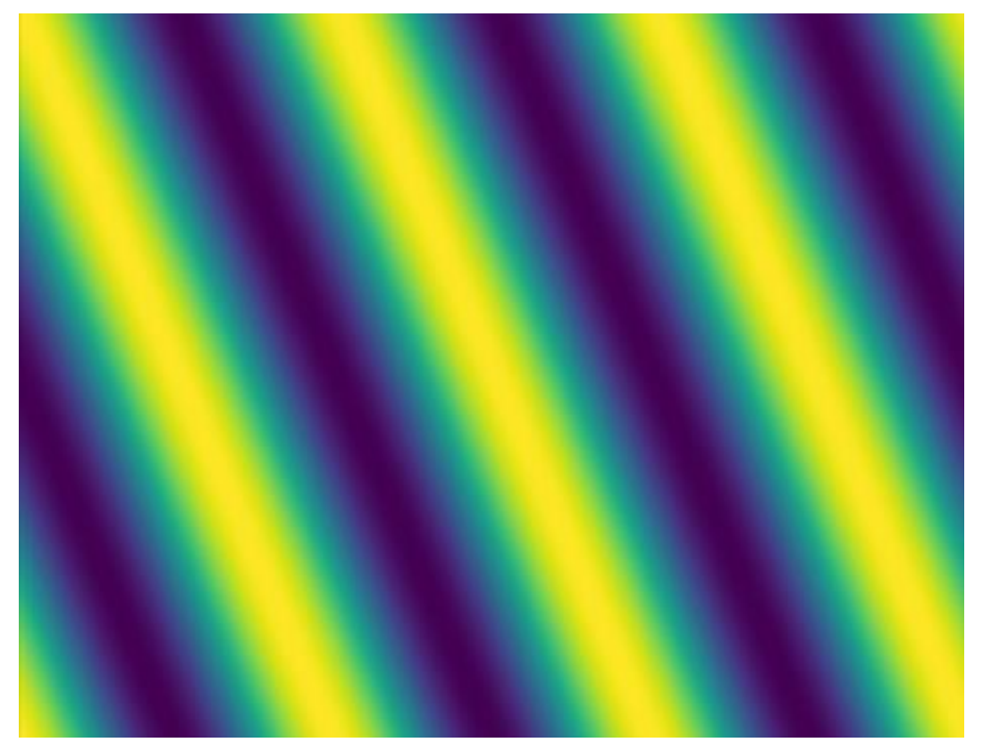

Fig 9: *Fourier Feature at point [349][3]*

Conclusions

After setting up the features and testing the algorithm on 200 red and green airglow images, the current accuracy of the classifier is at 87.05%. This is in line with other successful airglow image classifiers [4]. For the independent site analysis, the classifier was 92% accurate at categorizing images from a site it had not been trained on, demonstrating capability in real-time identification from independent sites.

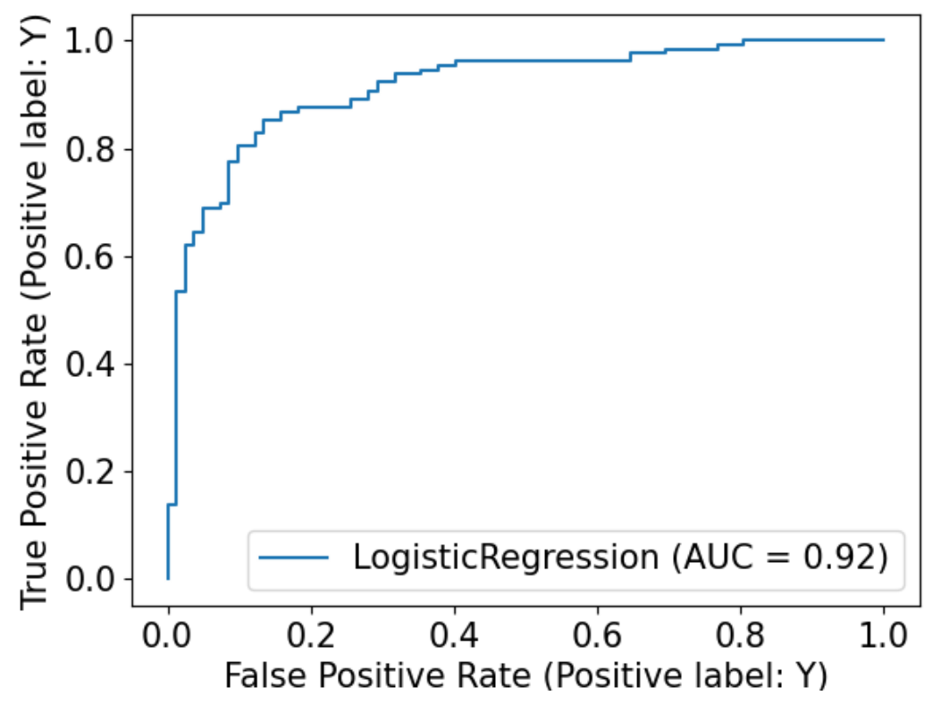

Fig 10: *Classifier ROC Curve*

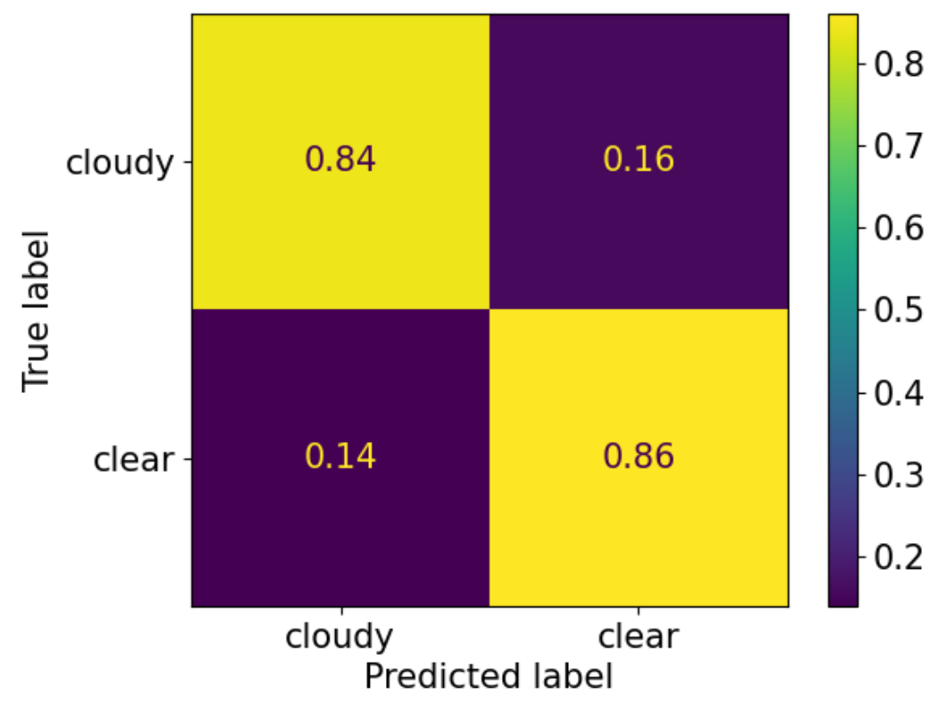

Fig 11: *Classifier Confusion Matrix (Decimals represent the percentage of y-axis images in that category)*

Future Work

One component of this project that could warrant future work is the difference observed when the algorithm is trained and tested on data of only greenline images instead of a dataset of both image types. The classifier performs much better in the greenline-only setting, and studying the reason for this disparity could improve the combined classification accuracy. This algorithm is planned to be included in the operational MANGO data processing pipeline, so that the dataset can be mainly clear images. With a clean dataset, future analysis of airglow images can be easily reproducible and have access to a much larger pool of data.

References and Acknowledgements

MANGO is supported by grant NSF AGS-1933077. We would like to thank Johnathan Makela, University of Illinois, and Asti Bhatt, Stanford Research Insitute The MANGO dataset is available at https://data.mangonetwork.org/data/transport/mango/archive/

[1] Bhatt, A. N., Harding, B. J., Makela, J.J., Navarro, L., Lamarche, L. J., Valentic, T., et al. (2023). MANGO: An optical network to study the dynamics of the Earth's upper atmosphere. Journal of Geophysical Research: Space Physics, 128, e2023JA031589. https://doi.org/10.1029/2023JA031589
[2] *Sedlak, R., Welscher, A., Hannawald, P., Wüst, S., Lienhart, R., and Bittner, M.: Analysis of 2D airglow imager data with respect to dynamics using machine learning, Atmos. Meas. Tech., 16, 3141–3153, https://doi.org/10.5194/amt-16-3141-2023, 2023.*
[3] *Martínez-Más J, Bueno-Crespo A, Khazendar S, Remezal-Solano M, Martínez-Cendán J-P, Jassim S, et al. (2019) Evaluation of machine learning methods with Fourier Transform features for classifying ovarian tumors based on ultrasound images. PLoS ONE 14(7): e0219388. https://doi.org/10.1371/journal.pone.0219388*
[4] *Lai, C.; Xu, J.; Yue, J.; Yuan, W.; Liu, X.; Li, W.; Li, Q. Automatic Extraction of Gravity Waves from All-Sky Airglow Image Based on Machine Learning. Remote Sens. 2019, 11, 1516. https://doi.org/10.3390/rs11131516*