


Machine Learning Approach to Classify Precipitation Type from A Passive Microwave Sensor



Machine Learning Approach to Classify Precipitation Type from A Passive Microwave Sensor

Spandan Das, Jie Gong, Chenxi Wang, Dong Liang Wu, Stephen Joseph Munchak, and William S. Olson
NASA Goddard Space Flight Center

Background

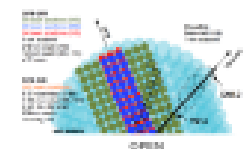


Figure 1: CMSR observing Earth from space.

Purpose

- Convective vs. Stratiform:
 - More accurate precipitation measurements/estimates
 - Operational cycles of convective and stratiform rain
- Machine Learning:
 - Understanding of complex precipitation mechanisms

Purpose of this project is to separate convective and stratiform precipitation using machine learning models trained on passive microwave data. Previous ML solutions were based on a heavily biased training samples (convective precipitation occurs more often than stratiform precipitation), so a goal of this project is to use a more balanced dataset.

Methods

FEATURES

Feature	Units	# of Channels
Black Cap (Ice Detection)	deg	1
Surface Temperature	°K	14
Surface Humidity	g/kg	14
Surface Wind Speed	m/s	14
Surface Wind Direction	deg	14
Surface Ice	deg	14
Surface Snow	deg	14
Surface Rain	deg	14
Surface Cloud	deg	14
Surface Wind	deg	14

Figure 2: Features (CMSR) 17 separate band, improved features: grey features removed

- Wide variety of atmospheric information from GMI data
- Surface temperature and humidity (convective rain) 14 channels (see Figure 3 below)
- Each channel is its own feature in the ML model (trained on 10,000 features)
- Frequency and persistence of events (shown in Figure 4 on the right)
- Many channels of identical frequencies have different polarizations (vertical and horizontal)
- Use data advantage of these polarizations by adding a band-improved feature: Polarization Difference

Analysis

CASE STUDY: SQUALL LINE

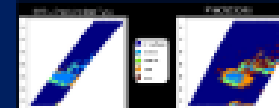



Figure 3: SQUALL LINE. Left: Convective rain (red) and stratiform rain (blue) regions. Right: Convective rain (red) and stratiform rain (blue) regions.

To better understand the project results and its significance, we examine a case study. On the left is the CMSR precipitation flag data for a squall line, which is a line of adjacent low-level clouds. Because the satellite's CMSR swath is limited, we can't see the entire squall line from the entire sensor's data.


The right figure ("Plot the different rain rates") shows the precipitation flag predictions made by the machine learning model for the entire squall line. These predictions are based on all of 2018 data, not just a single scan of the convective system with its various precipitation flags.

FEATURE IMPORTANCE



Results

Category	Convective Rain	Stratiform Rain	Machine Learning
Convective Rain	10.0	10.0	10.0
Stratiform Rain	10.0	10.0	10.0
Machine Learning	10.0	10.0	10.0
Convective Rain	10.0	10.0	10.0
Stratiform Rain	10.0	10.0	10.0



Conclusion

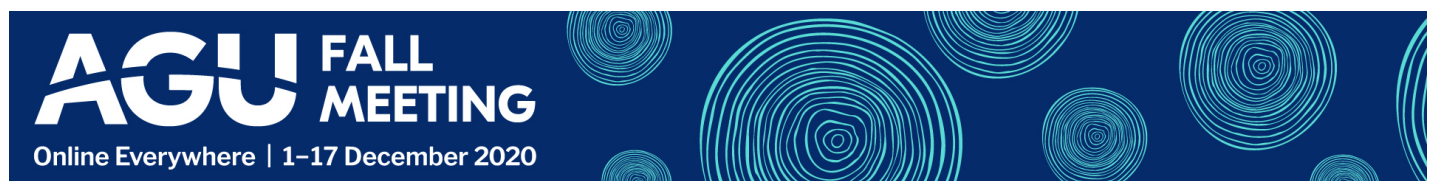
- Machine Learning models with good performance:
 - ~95% accuracy
 - ~90% accuracy
 - Convective rain data inclusion
- Convective rain data inclusion:
 - Higher frequencies were important
 - Polarization Difference very helpful
 - Future instrument channel selection
- Future Work:
 - Focus on improving (or removing) "Mixed"

[NARRATION](#)
[ABSTRACT](#)
[CONTACT AUTHOR](#)
[PRINT](#)
[GET POSTER](#)

Spandan Das, Jie Gong, Chenxi Wang, Dong Liang Wu, Stephen Joseph Munchak, and William S. Olson

NASA Goddard Space Flight Center

PRESENTED AT:



BACKGROUND

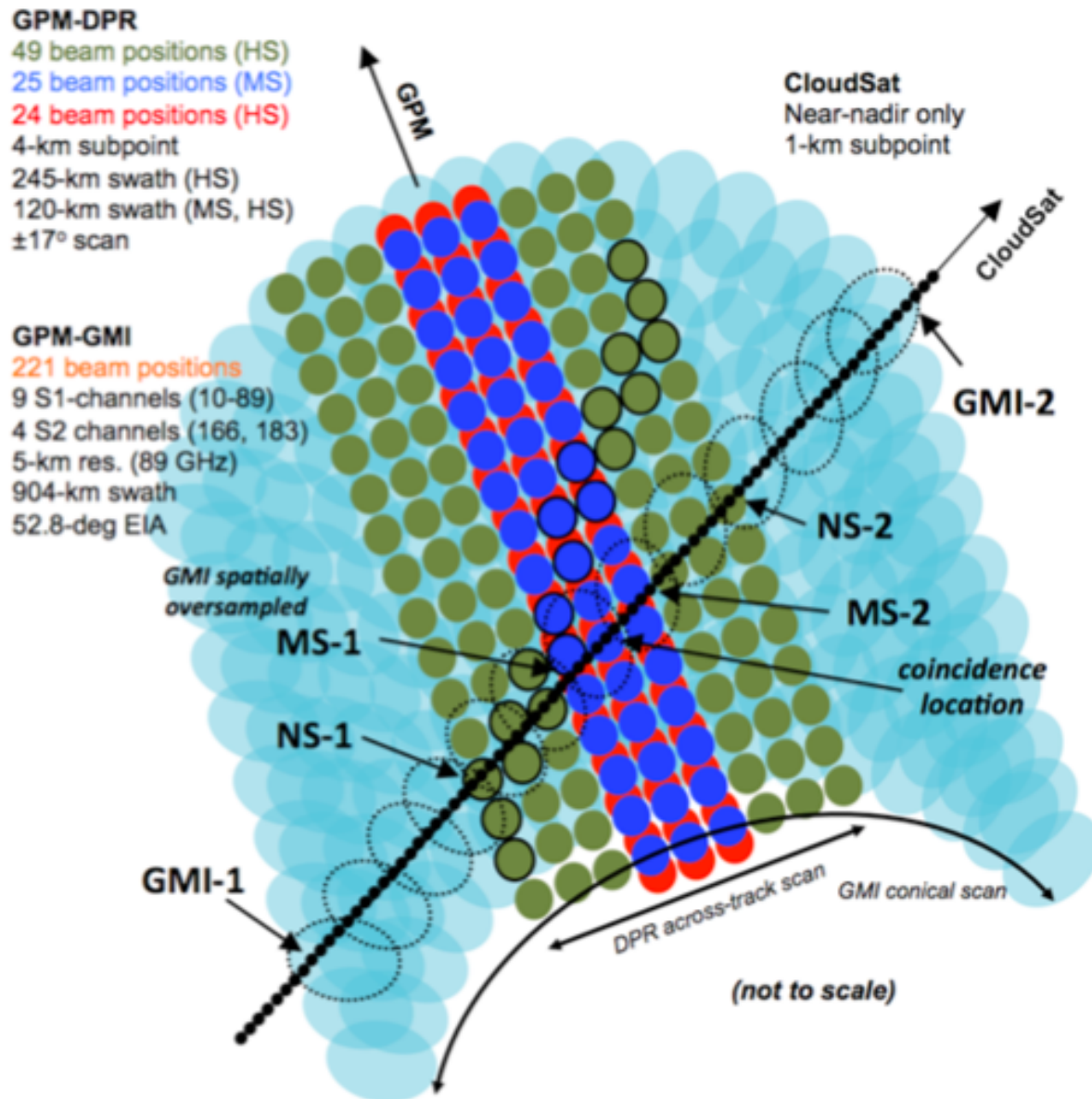


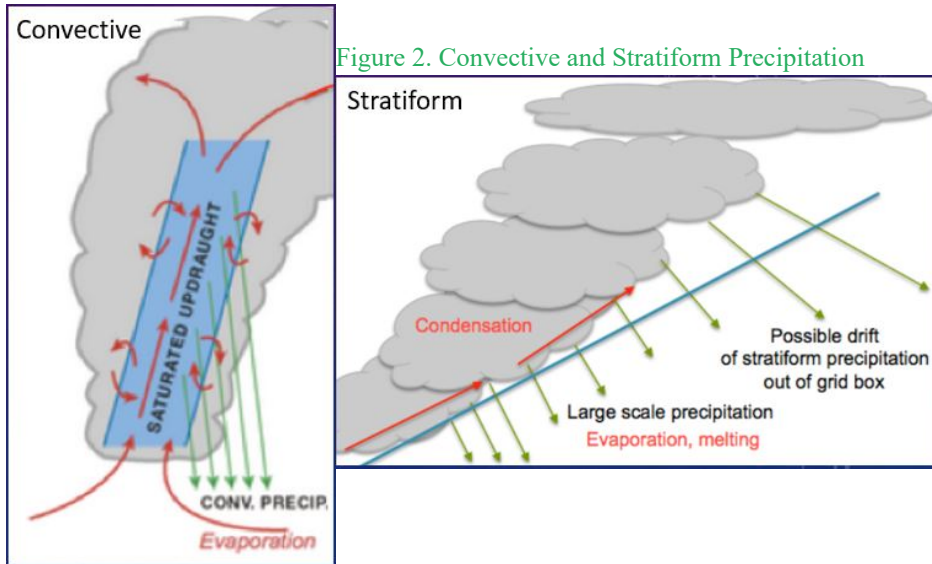
Figure 1. Details of GPM Core Observatory Scan; light-blue is GMI scan and green is DPR scan

NASA's Global Precipitation Measurement (GPM) mission's Core Observatory Satellite has two sensors:

- Microwave Imager (GMI)
 - Features for training
 - Passive Microwave Sensor
- Dual-Frequency Precipitation Radar (DPR)

- Active Sensor
- Precipitation Flag

Note the difference in scan-width of the two sensors. Goal of the project is to use ONLY passive microwave data to determine precipitation type.



In this project, we differentiate between 5 different types of precipitation:

- No Precipitation
- Convective
- Stratiform
- Mixture
- Other

PURPOSE

- Convective vs. Stratiform
 - More accurate precipitation measurements/forecasts
 - Diurnal cycles of convective and stratiform rain
- Machine Learning
 - Understanding of complex precipitation mechanisms not required

Purpose of this project is to separate convective and stratiform precipitation using machine learning models trained on passive microwave data. Previous ML solutions were based on a heavily biased training samples (since no-precipitation scenes are much more frequent than precipitation scenes), so one aim of the project was to overcome the inherent data bias.

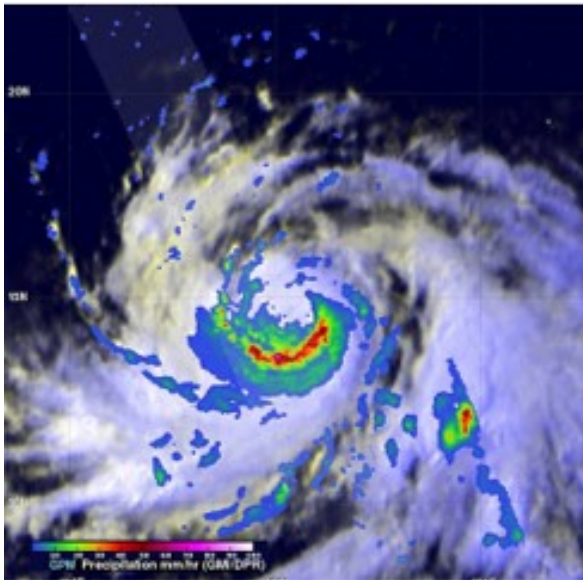


Figure 3. GPM Satellite Data Visualization

METHODS

FEATURES

Feature	Name	# of Channels
Cloud Liquid Water Path	clwp	1
Surface emissivity	emis	13
Latitude/Longitude	lat/lon	1
Brightness Temperature	tc	13
Surface Skin Temperature	ts	1
Total Column Water Vapor	twv	1
**Polarization Difference	PD	5
Surface Type	tysfc	1
Convergence Robustness Factor	chi	1
Universal Time	utc	1

Figure 4. Features (GMI-only); ** represent hand-engineered feature; gray features unused

- Wide variety of atmospheric information given from GMI data.
- Surface emissivity and brightness temperature have 13 channels (see Figure 5 below):
 - Each channel is its own feature in the ML model (instead of 8, ~30 features)
 - Frequencies and polarization of channels shown on figure on the right
 - Many channels of identical frequencies have different polarizations (vertical and horizontal).
 - We take advantage of these polarizations by adding a hand-engineered feature: Polarization Difference.

Channel No	Central Frequency (Ghz)	Central Frequency Stabilization (\pm MHz)	Bandwidth (Mhz)	<u>Polarization</u>
1	10.65	10	100	V
2	10.65	10	100	H
3	18.70	20	200	V
4	18.70	20	200	H
5	23.80	20	400	V
6	36.50	50	1000	V
7	36.5	50	1000	H
8	89.00	200	6000	V
9	89.00	200	6000	H
10	166.0	200	3000	V
11	166.0	200	3000	H
12	183.31 \pm 3	200	3500	V
13	183.31 \pm 7	200	4500	V

◆ **V:** Polarization vector is parallel to scan plane at nadir

◆ **H:** Polarization vector is perpendicular to scan plane at nadir

Figure 5. Channel Frequencies and Polarization

POLARIZATION DIFFERENCES

Polarization Differences (hand-engineered feature):

- Difference between brightness temperature values of vertical and horizontal polarizations: $tc[V] - tc[H]$
- Frequencies used: 10.65, 89.00, 166.0 GHz

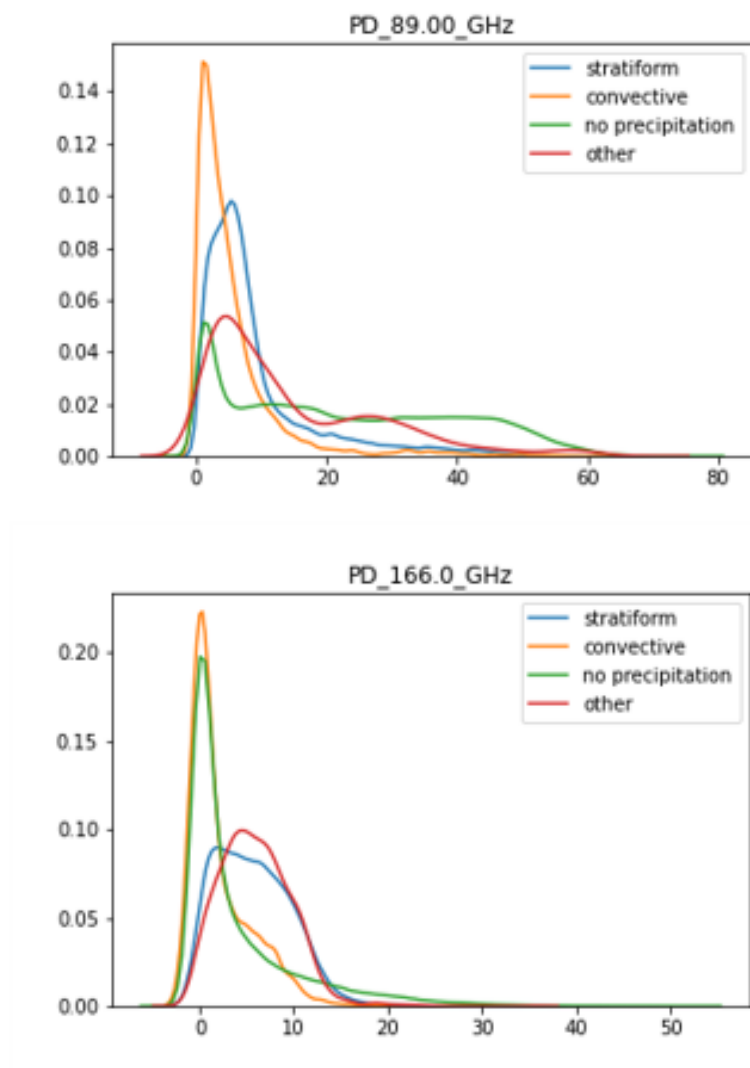


Figure 6. Distribution of Polarization Differences by flag (89.00, 166.0 GHz)

TRAINING AND VALIDATION DATASETS

Training Dataset (2017 data):

- 84 days of data
 - 7 days/month
 - Randomly selected days
- Sub-sample daily data
 - Avoid bias caused by data imbalance (since ~95% of daily data is non-precipitating)

Validation Dataset (2017 data):

- 12 days of data
 - 1 day/month

- Randomly selected days
- No sub-sampling
 - Resemble distribution of real-world data



Figure 7. Training and Validation Label Frequencies; all labels from DPR sensor

MACHINE LEARNING MODELS

We use train and test 6 different machine learning models:

1. Naive Bayes Classifier
2. Support Vector Machine
3. Softmax Regression
4. Random Forest
5. Gradient Boosting Classifier
6. Neural Network

All of the models are created, trained, and evaluated using Scikit-learn.

ANALYSIS

CASE STUDY: SQUALL LINE

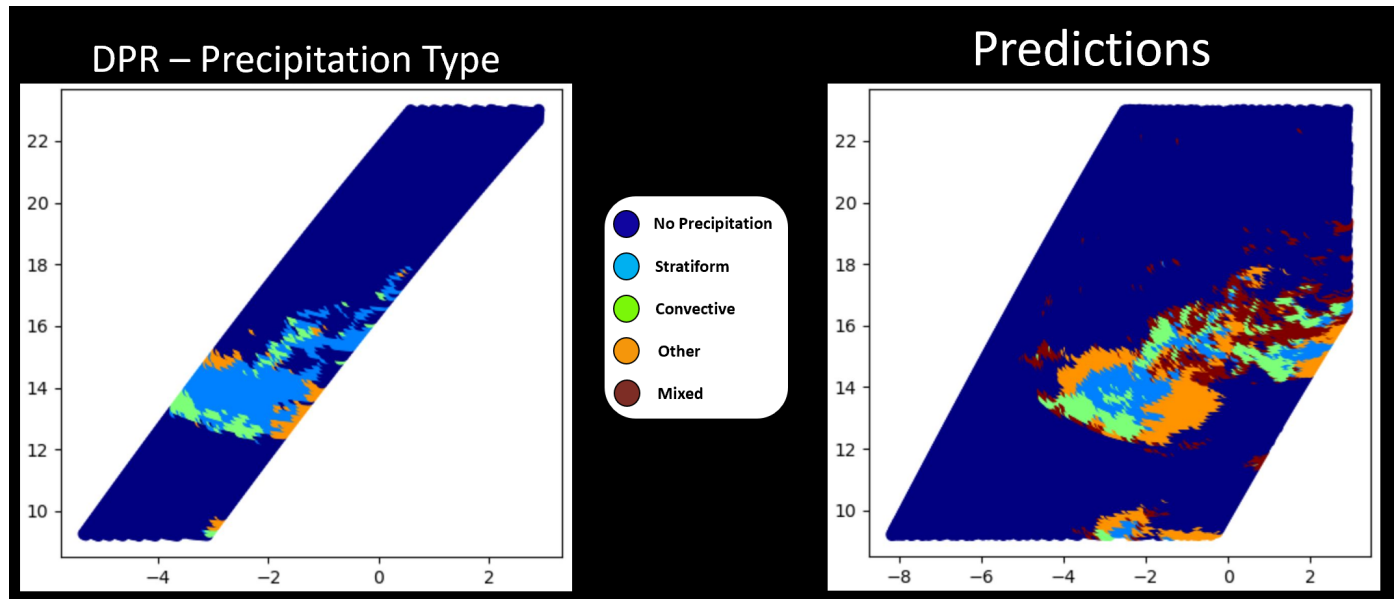


Figure 8. Squall line precipitation labels; DPR (given) labels on left, predictions on right

To better understand the project results and its significance, we examine a case study.

On the left is the DPR precipitation flag data for a squall line, which is a line of adjacent thunderstorms. Because the width of the DPR swath is limited, we can't see the entire squall line from the active sensor's data.

The right figure (**Note the different x-axis scales**) shows the precipitation flag predictions made by the random forest model for the same squall line. Since predictions are based off of GMI data, we get a wider view of the convective system with its various precipitation flags.

FEATURE IMPORTANCES

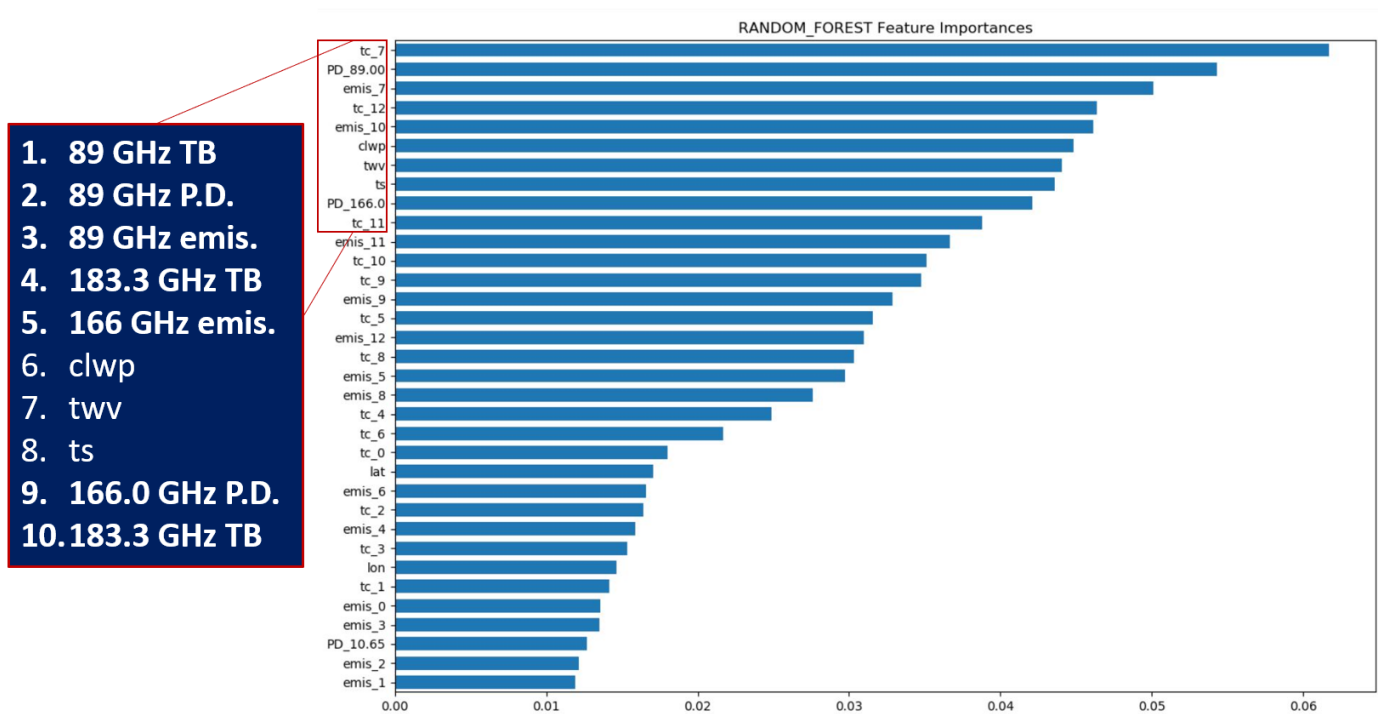


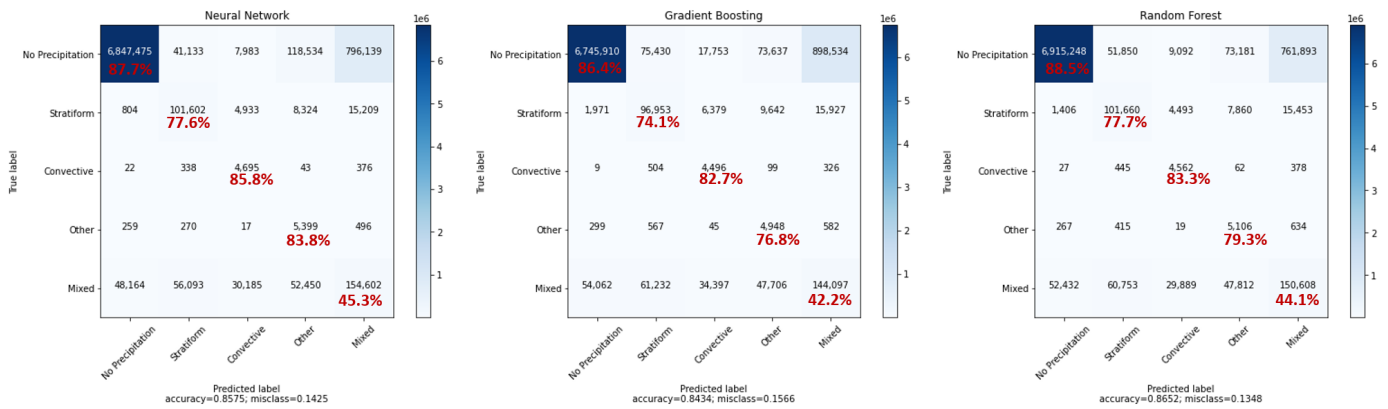
Figure 9. Feature importances for random forest model

The figure above shows a list of the feature importances for the random forest model.

From here, we see that the high frequency channels are the most important features for distinguishing between precipitation types. We also see that the polarization differences that we added to the data are very helpful for classification.

RESULTS

Classifier	Overall Accuracy (%)	Area Under ROC Curve (AUC ROC)
Naive Bayes	32.71	0.7312
Support Vector Machine	85.98	N/A
Softmax Regression	84.17	0.9077
Gradient Boosting	84.34	0.9360
Random Forest	86.52	0.9429
Neural Network	85.75	0.9432



CONCLUSION

- Multiple successful models with good performance
 - ~85% accuracy
 - >0.93 AUC score
 - Overcame inherent data imbalance
- Demonstrated relative significance of features
 - Higher frequencies more important
 - Polarization Differences very helpful
 - Future instrument channel selection
- Future Work
 - Focus on improving (or removing) “Mixed” class
 - Add nearby-pixel associations

Acknowledgements

- NASA Internship Program
- GSFC Climate and Radiation Lab

ABSTRACT

Precipitation flag (precipitating or not; stratiform or convective) is a key parameter for us to make better retrieval of precipitation characteristics as well as to understand the cloud-precipitation physical processes. The Global Precipitation Measurement (GPM) Core Observatory's Microwave Imager (GMI) and Dual-Frequency Precipitation Radar (DPR) together provide ample information on global precipitation characteristics. As an active sensor in particular, DPR provides an accurate precipitation flag assignment, while passive sensors like GMI were traditionally believed not to be able to tell apart precipitation types.

Using collocated precipitation flag assignment from DPR as the “truth”, this project employs machine learning models to train and test the predictability and accuracy of using passive GMI-only observations together with ancillary atmosphere information from reanalysis. Precipitation types are classified into the following classes: convective, stratiform, convective-stratiform mixed, no precipitation, and other precipitation. Sub-sampling with different probabilities is employed to construct a balanced training dataset. A variety of classification algorithms are tested, including Support Vector Machines, Naive Bayes, Random Forests, Gradient Boosting, and Neural Networks (Multilayer Perceptron Network), and their results are evaluated and compared. The trained model has $\sim 85\%$ of prediction accuracy for every type of precipitation. High-frequency channels (166 GHz and 183 GHz channels) and 166 GHz polarization difference are found among the most important factors that contribute to the model performance, which shed light on future instrument channel selection.