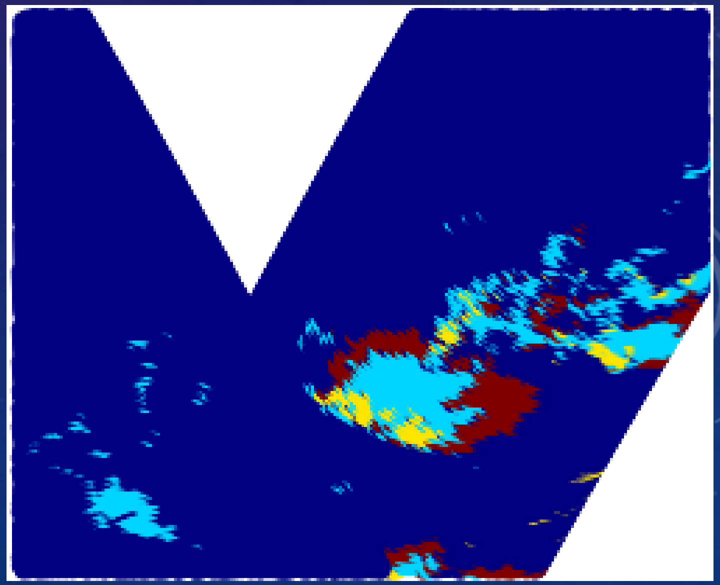
Machine Learning Approach to Classify Precipitation Type from a Passive Microwave Sensor

Spandan Das

Thomas Jefferson High School for Science and Technology

Mentor: Jie Gong



RATIONALE

- 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

To separate convective and stratiform precipitation using machine learning models trained on passive microwave data.

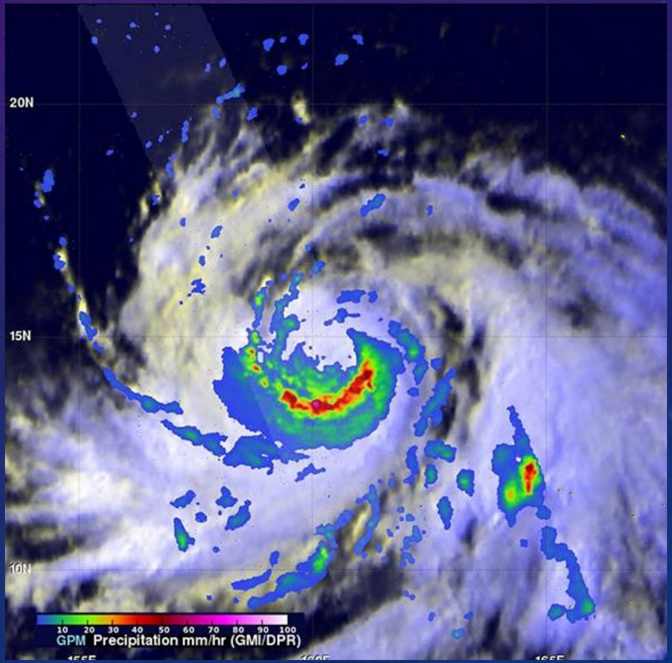
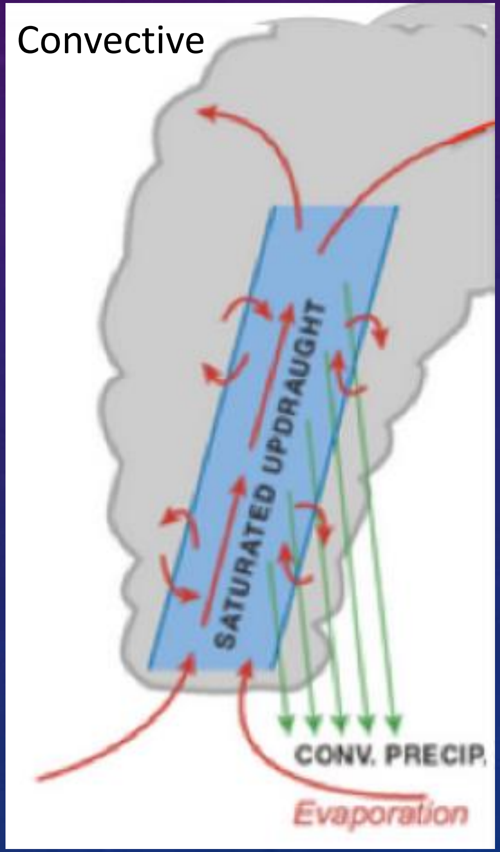


Figure 1. GPM Satellite Data Visualization

PRECIPITATION TYPES

- Convective
- Stratiform
- Mixture
- Other



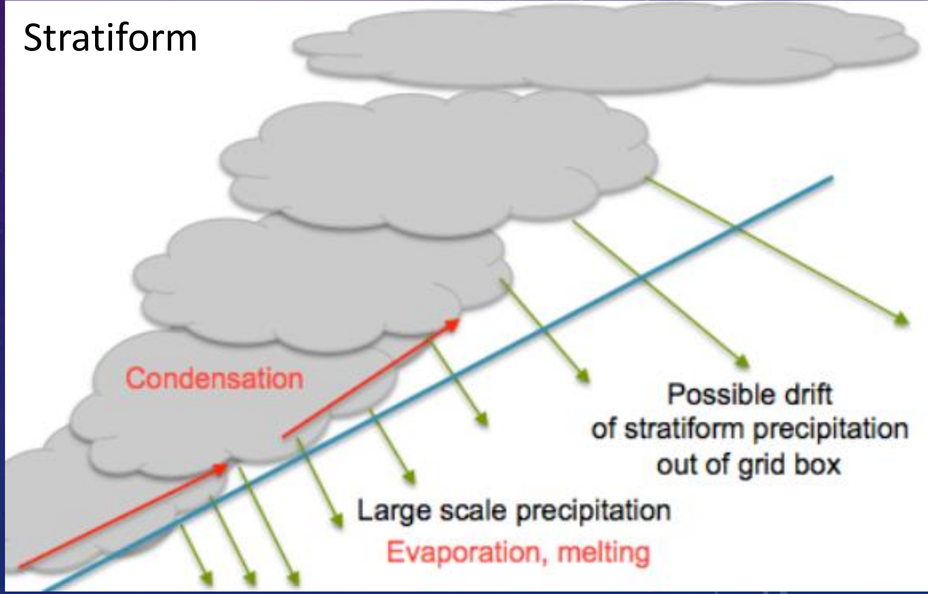


Figure 2. Diagrams of convective (left) and stratiform (above) precipitation

DATA

- GPM Core Observatory
  - Microwave Imager (GMI)
    - Passive Microwave
    - Features for training
  - Dual-Frequency Precipitation Radar (DPR)
    - Active Sensor
    - Precipitation Flag

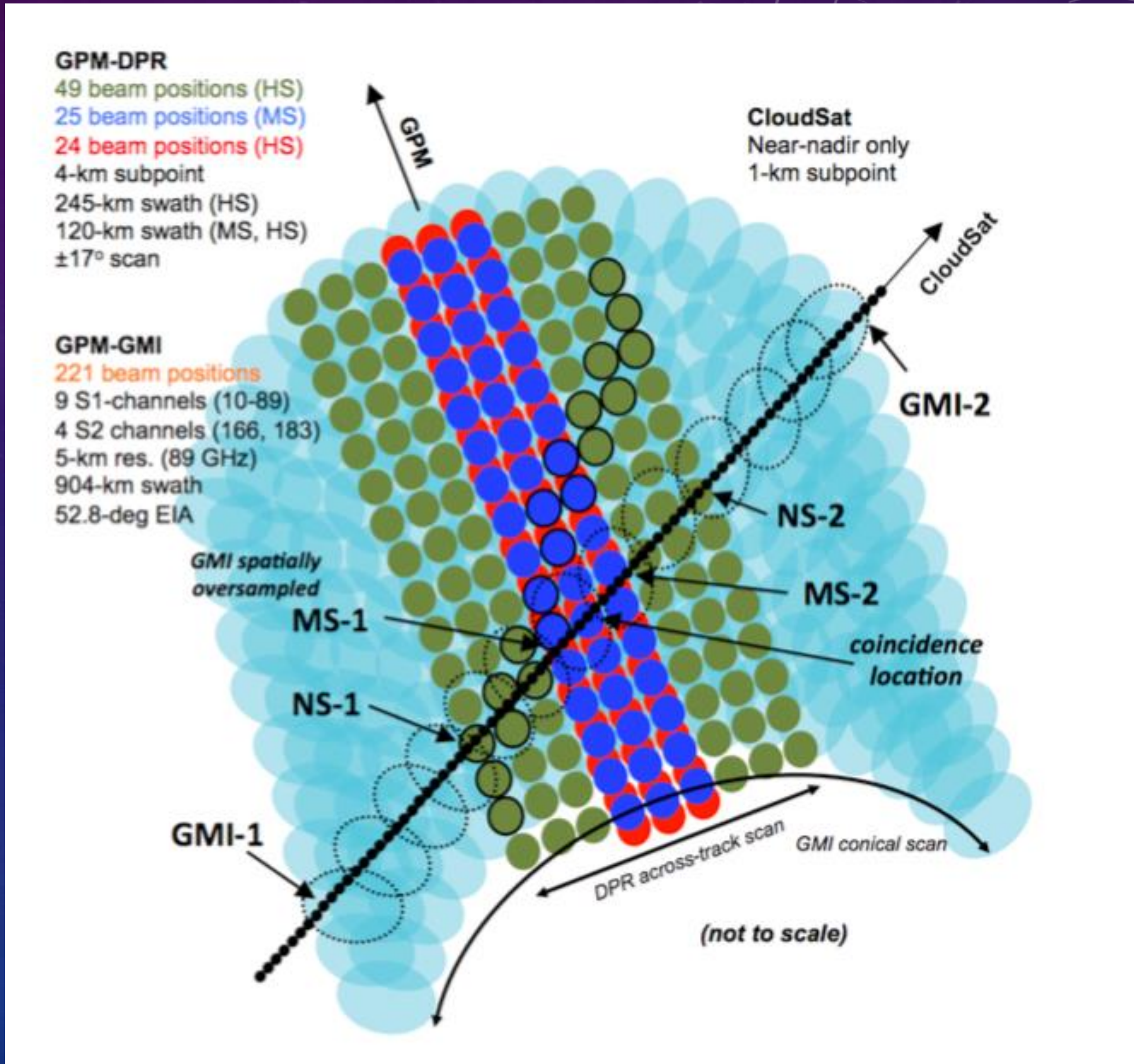


Figure 3. Details of GPM Core Observatory Scan

FEATURES (GMI-ONLY)

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	-
Surface Type	tysfc	1
Convergence Robustness Factor	chi	1
Universal Time	utc	1

** = Hand-engineered feature

Channel No	Central Frequency (Ghz)	Central Frequency Stabilization (±MHz)	Bandwidth (Mhz)	Polarization
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±3	200	3500	V
13	183.31±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 4. Channel Frequencies and Polarization

POLARIZATION DIFFERENCES

- Difference between brightness temperature values of vertical and horizontal polarizations: tc[V] – tc[H]
- For frequencies: 10.65, 89.00, 166.0 GHz

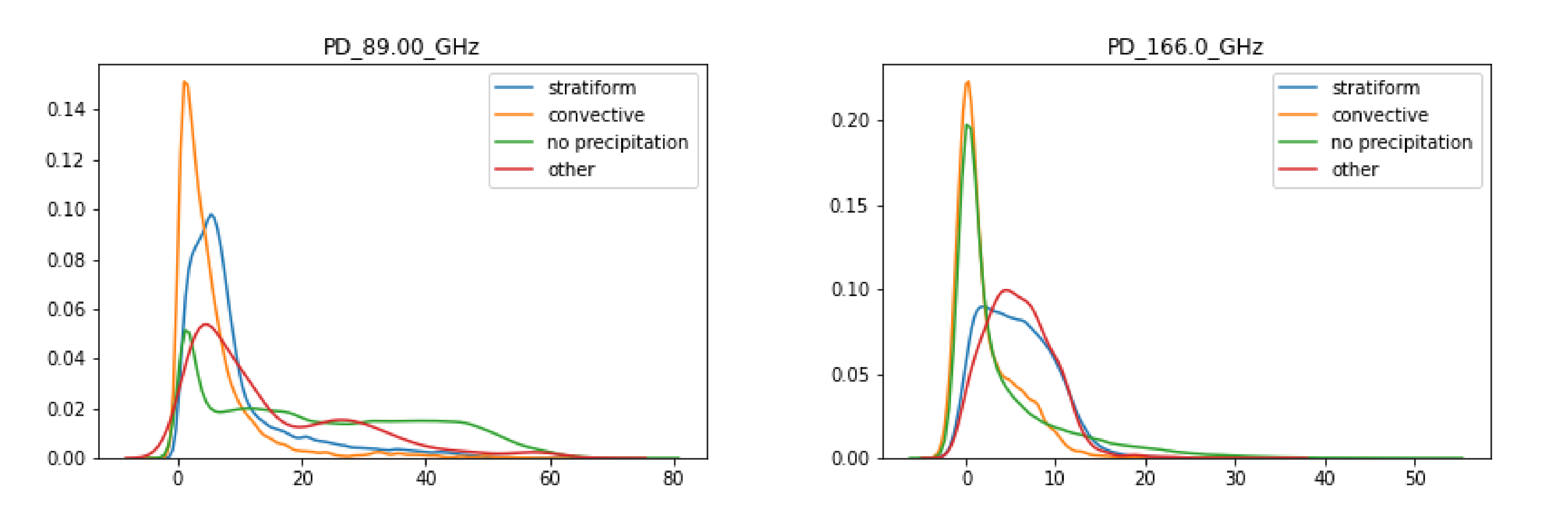
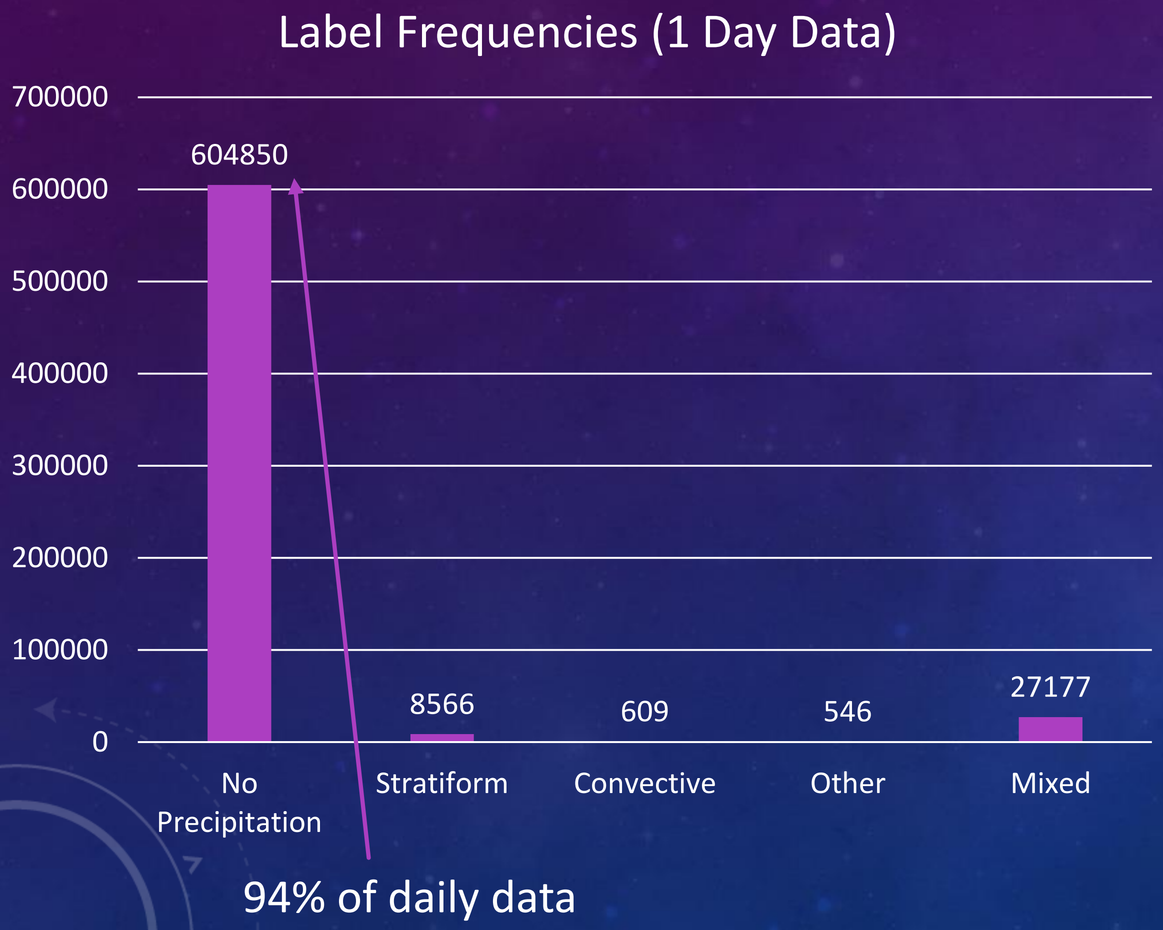
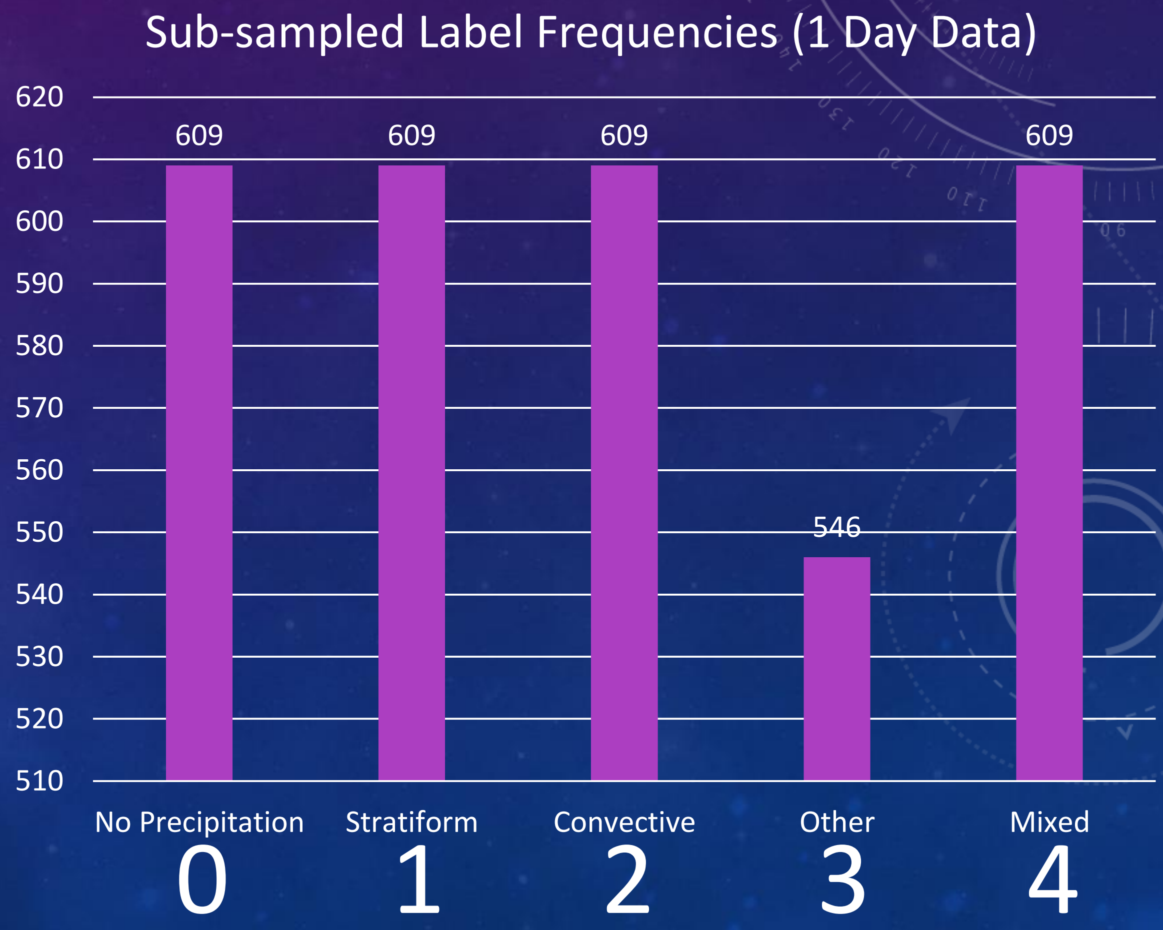


Figure 5. Distribution of Polarization Differences by flag (89.00, 166.0 GHz)

LABELS (DPR) & SUB-SAMPLING





TRAINING AND VALIDATION DATASETS

Training

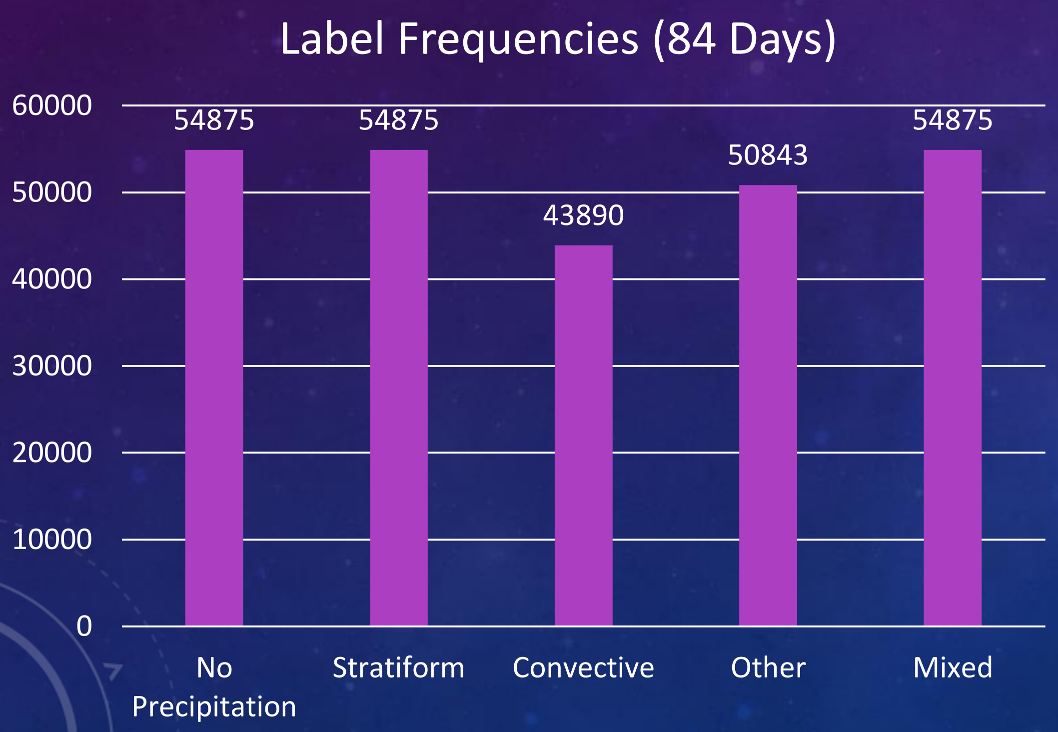
- 2017 data
- 84 days of data
  - 7 days/month
  - Randomly selected days
- Sub-sample daily data
  - Avoid bias caused by data imbalance

Validation

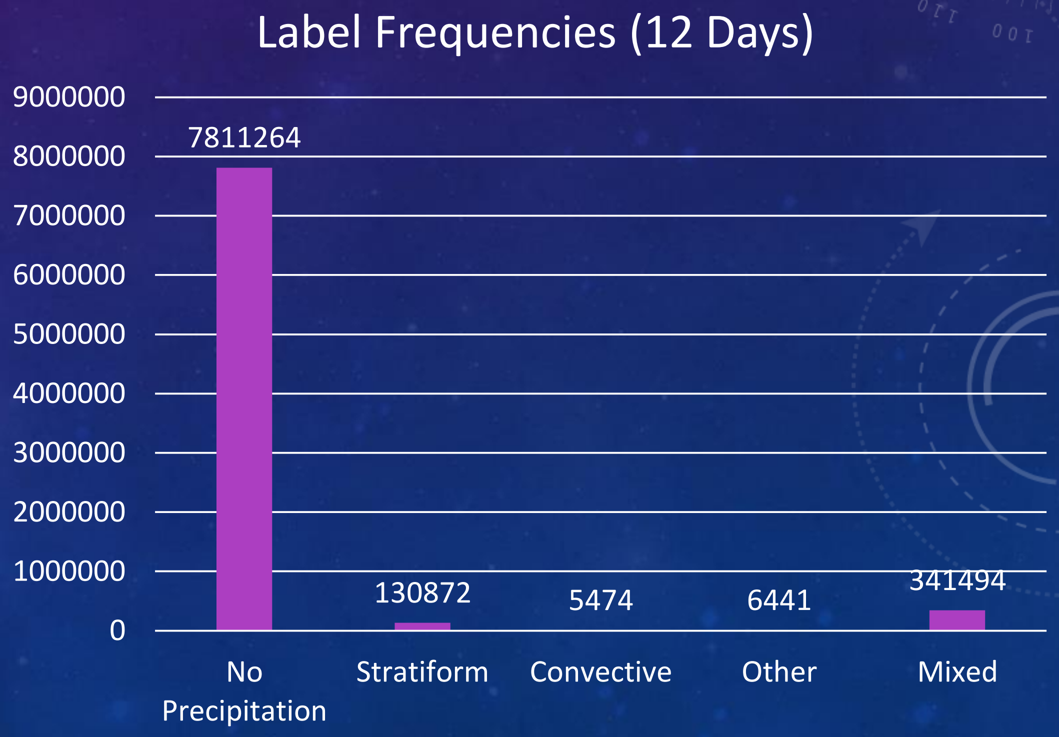
- 2017 data
- 12 days of data
  - 1 day/month
  - Randomly selected days
- No sub-sampling
  - Resemble distribution of real data

TRAINING AND VALIDATION DATASETS

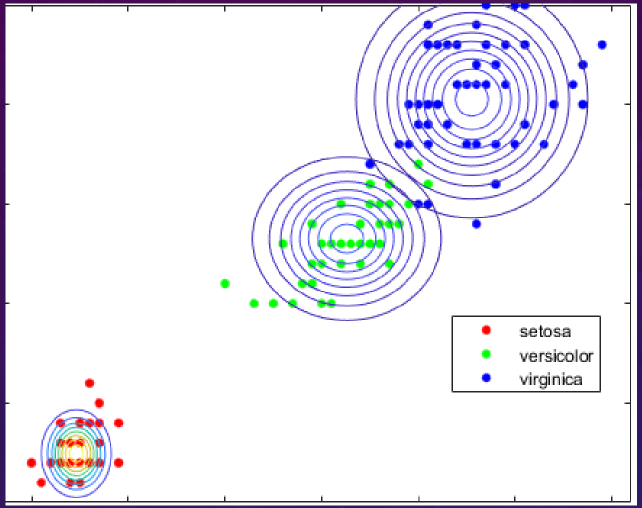
Training



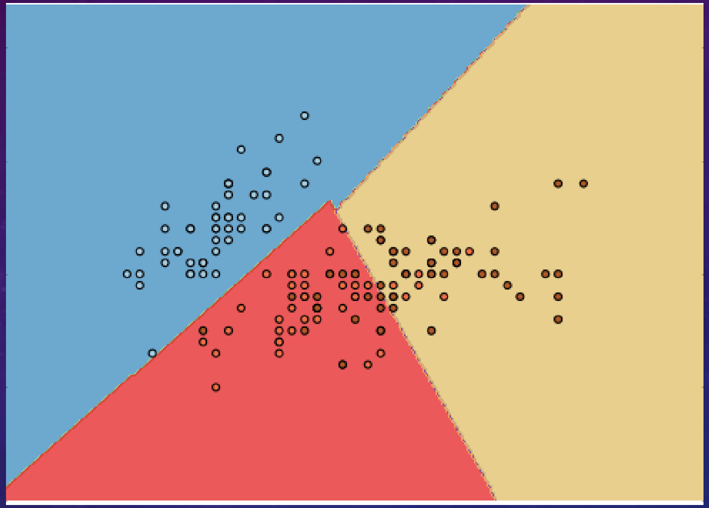
Validation



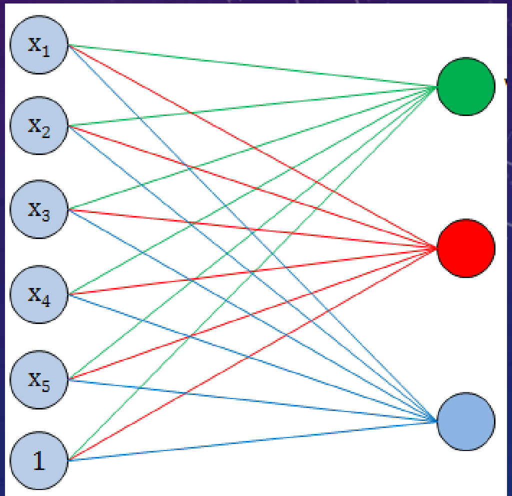
MACHINE LEARNING MODELS (scikit learn)



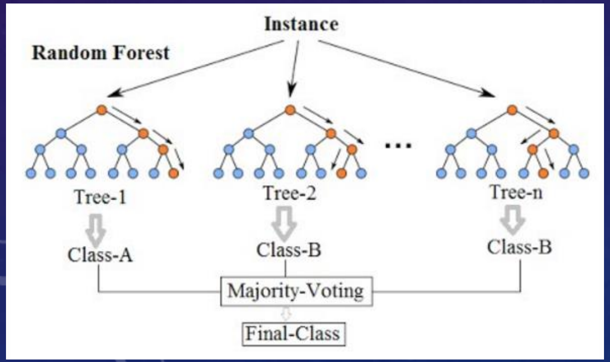
Naive Bayes Classifier



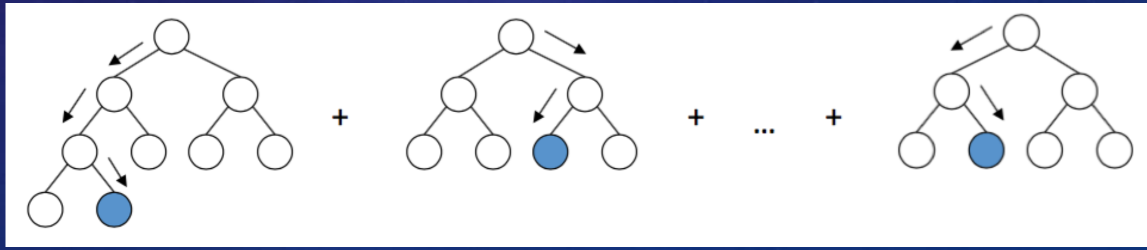
Support Vector Machine



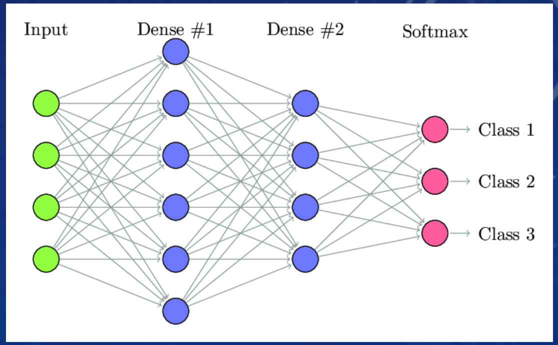
Softmax Regression



Random Forest



Gradient Boosting



Neural Network

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

CONFUSION MATRICES (TOP 3)

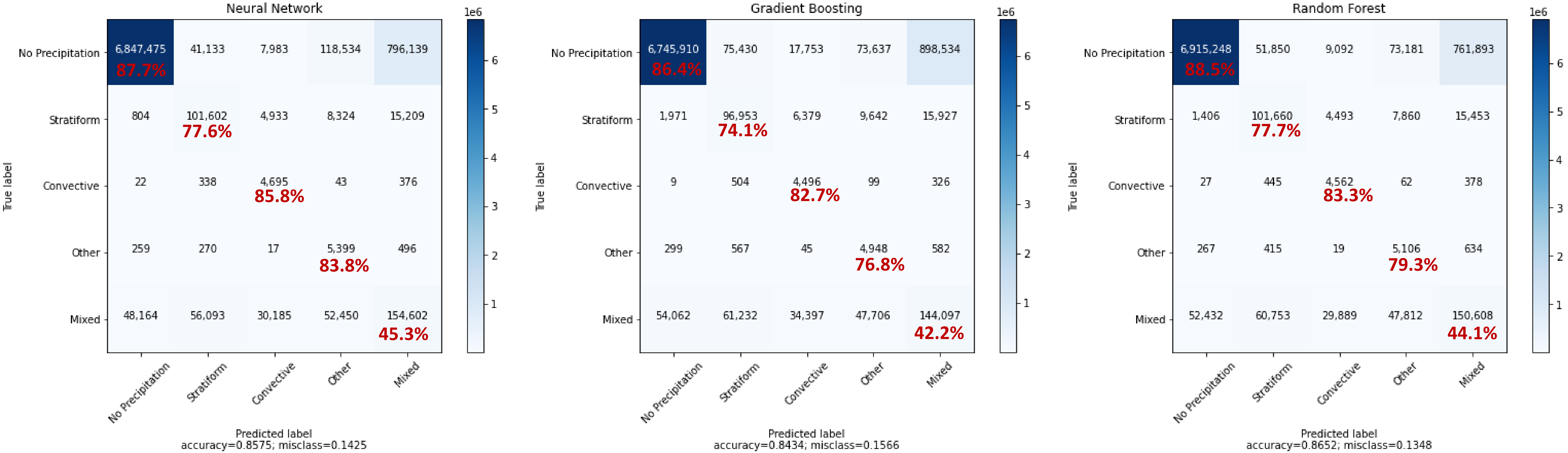
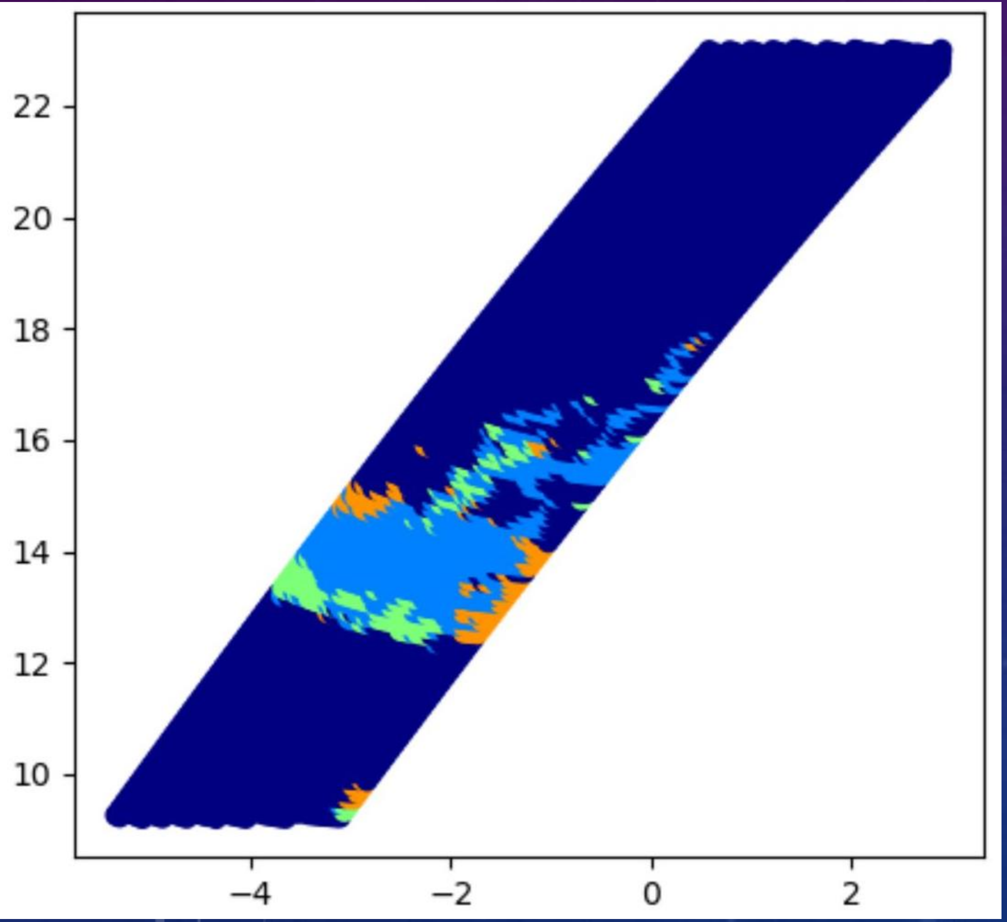


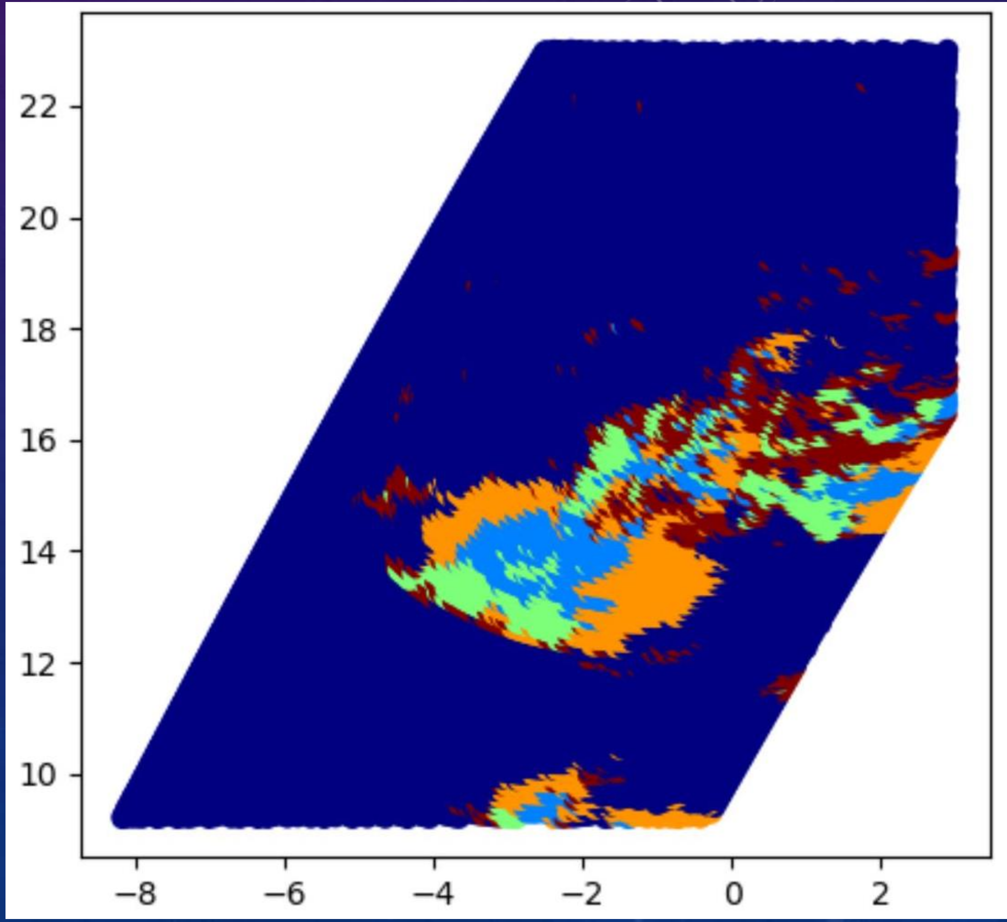
Figure 6. Confusion Matrices (top 3 models)

CASE STUDY: SQUALL LINE

DPR – Precipitation Type



Predictions



FEATURE IMPORTANCES

1. **89 GHz TB**
2. **89 GHz P.D.**
3. **89 GHz emis.**
4. **183.3 GHz TB**
5. **166 GHz emis.**
6. clwp
7. twv
8. ts
9. **166.0 GHz P.D.**
10. **183.3 GHz TB**

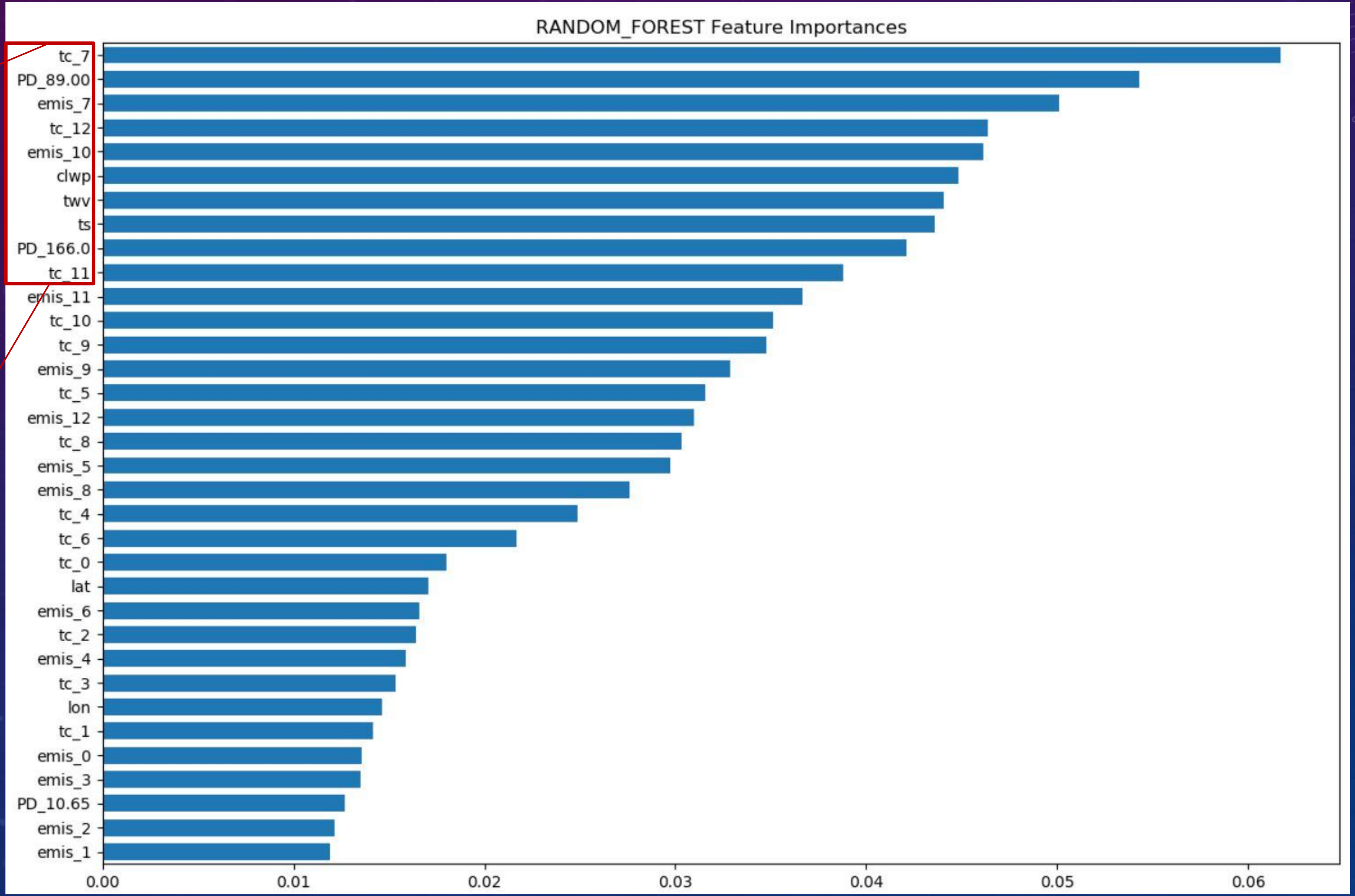


Figure 7. Feature Importance Rank for Random Forest classifier

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
  - Instrument channel selection
- Future Work
  - Focus on improving (or removing) “Mixed” class
  - Add nearby-pixel associations

ACKNOWLEDGEMENTS

- Jie Gong
- Chenxi Wang
- Dong L. Wu
- Stephen J. Munchak
- William S. Olson

Thank You!

IMAGE SOURCES

- https://www.clarksvilleonline.com/wp-content/uploads/2015/08/NASAs-Global-Precipitation-Measurement-satellite-observes-Typhoon-Atsani-building-1.jpg
- https://confluence.ecmwf.int/download/attachments/131397302/Fig2.1.4.4.3%20Conv%26Strat%20Cloud%26Precip%20Diag.png?version=1&modificationDate=1525963399687&api=v2
- https://www.star.nesdis.noaa.gov/mirs/gpmgmi.php
- https://www.star.nesdis.noaa.gov/star/documents/meetings/GPM2011/dayOne/Skofronick-JacksonG.pdf
- https://www.mathworks.com/help/examples/stats/win64/TrainANaiveBayesClassifierFitcnbExample_01.png (Naive Bayes)
- https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_111.png (Support Vector Machine)
- https://www.harrisgeospatial.com/docs/html/images/Classification/SoftmaxDiagram.gif (Softmax Regression)
- https://miro.medium.com/max/1042/1*6dGA56_UXJzuDAlAOd9Otw.png (Gradient Boosting)
- https://upload.wikimedia.org/wikipedia/commons/7/76/Random_forest_diagram_complete.png (Random Forest)
- https://www.researchgate.net/profile/Charlotte_Pelletier/publication/331525817/figure/fig2/AS:733072932745216@1551789615161/Example-of-fully-connected-neural-network.png (Neural Network)

PREDICTION PROBABILITIES

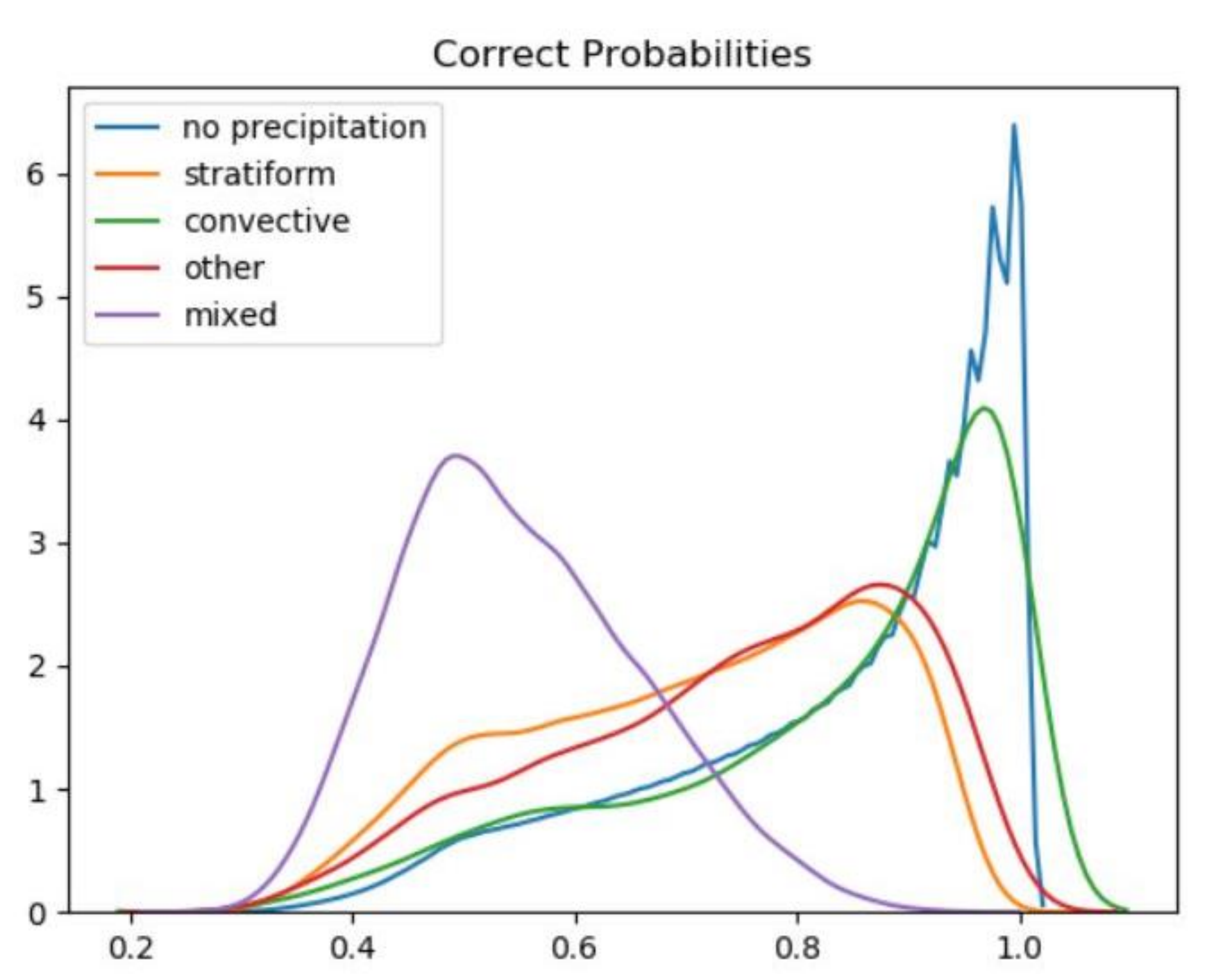


Figure _. Distribution of probabilities for correct instances
(Random Forest model)

FEATURE IMPORTANCES (CONT.)

1. **166 GHz emis.**
2. 36.5 GHz TB
3. 36.5 GHz emis.
4. twv
5. clwp
6. **89 GHz TB**
7. **89 GHz emis.**
8. **183.3 GHz emis.**
9. ts
10. **89.00 GHz P.D.**

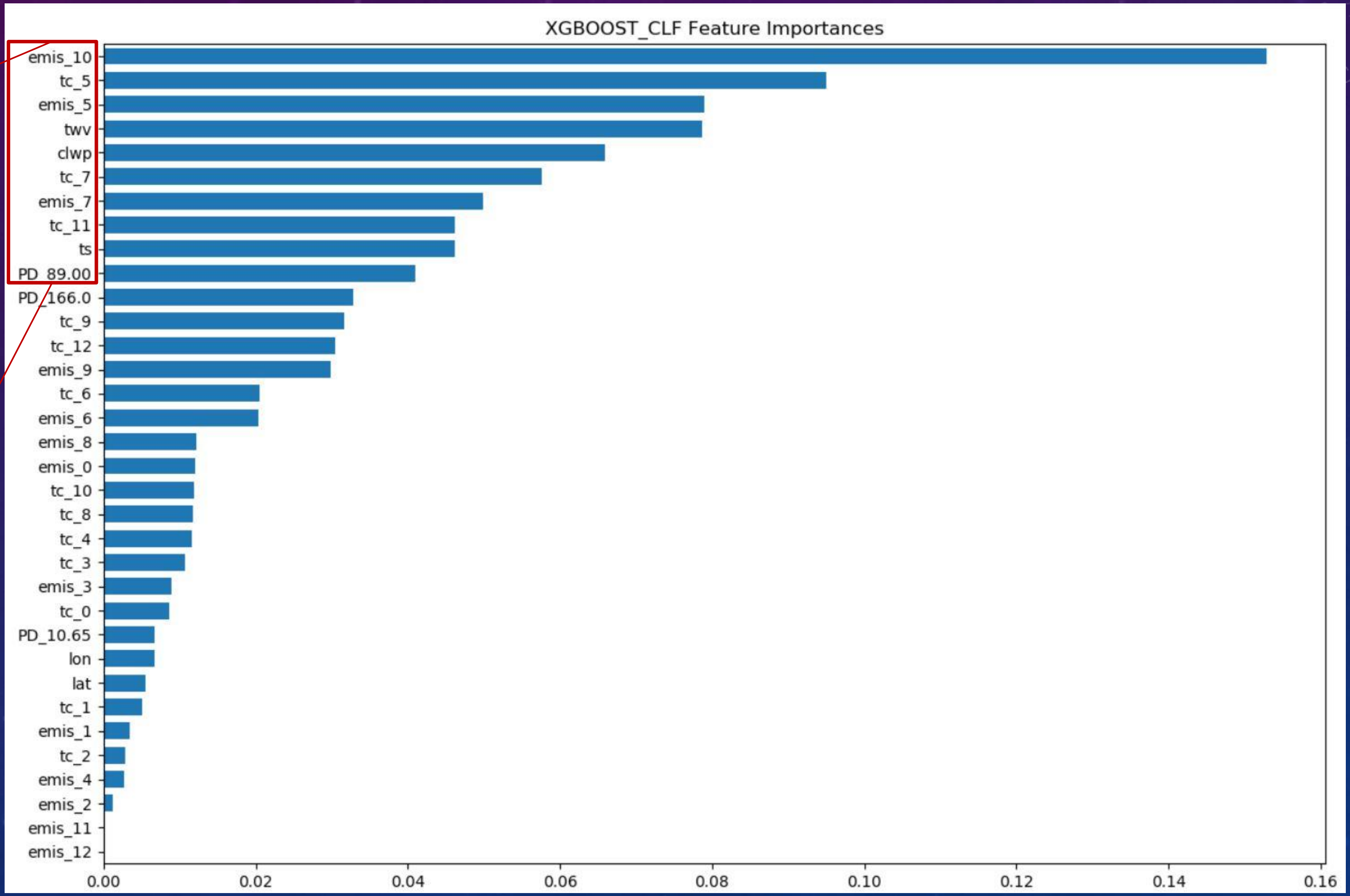


Figure _. Feature Importance Rank for Gradient Boosting classifier