EXPLAINABLE AI FOR SHORT-TERM LIGHTNING PREDICTION

Prahalath Anbu Bharathi
Tufts University, NESSI
Mentors: John Schreck, David John Gagne II

July 28, 2021

NCAR

NSF

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Motivation

https://pixabay.com/photos/lightning-storm-weather-sky-399853/

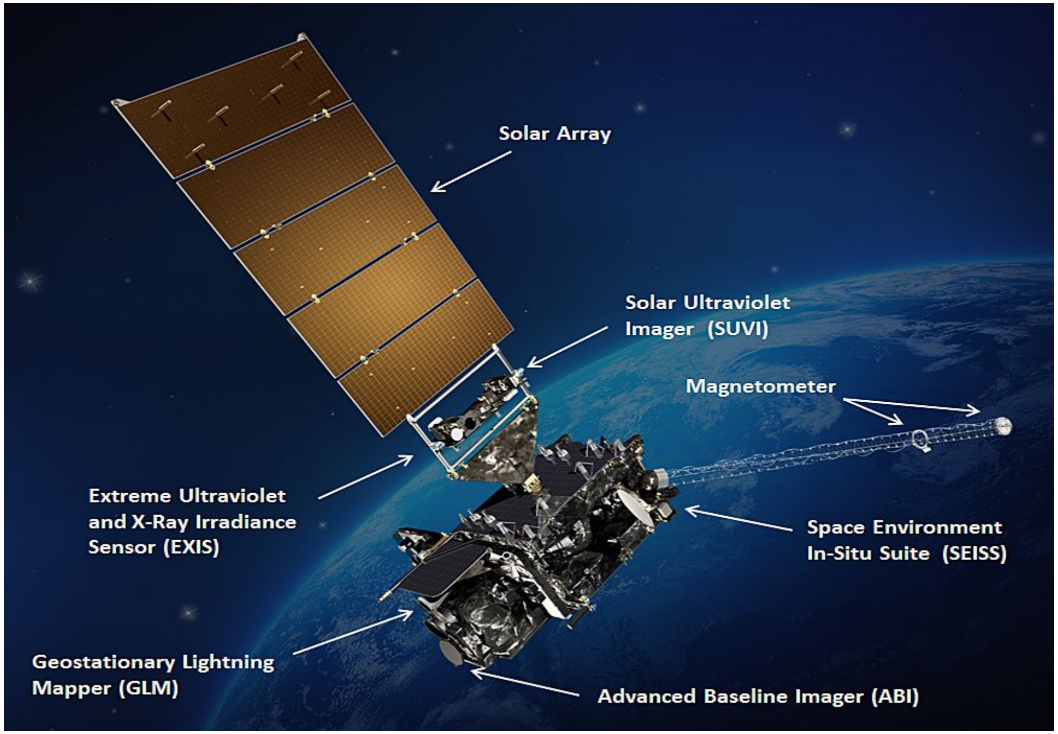
Severe weather kills on average **500** people and causes **billions of dollars** in property damage every year in the United States.

- Short-term prediction of Lightning can reduce some of these losses

Current methods of lightning forecasting do not provide sufficient lead times and lack in accuracy

- We can apply deep learning to this problem
- Leverage data from GOES-16 satellite

Data



https://commons.wikimedia.org/wiki/File:GOES-R_SPACECRAFT.jpg

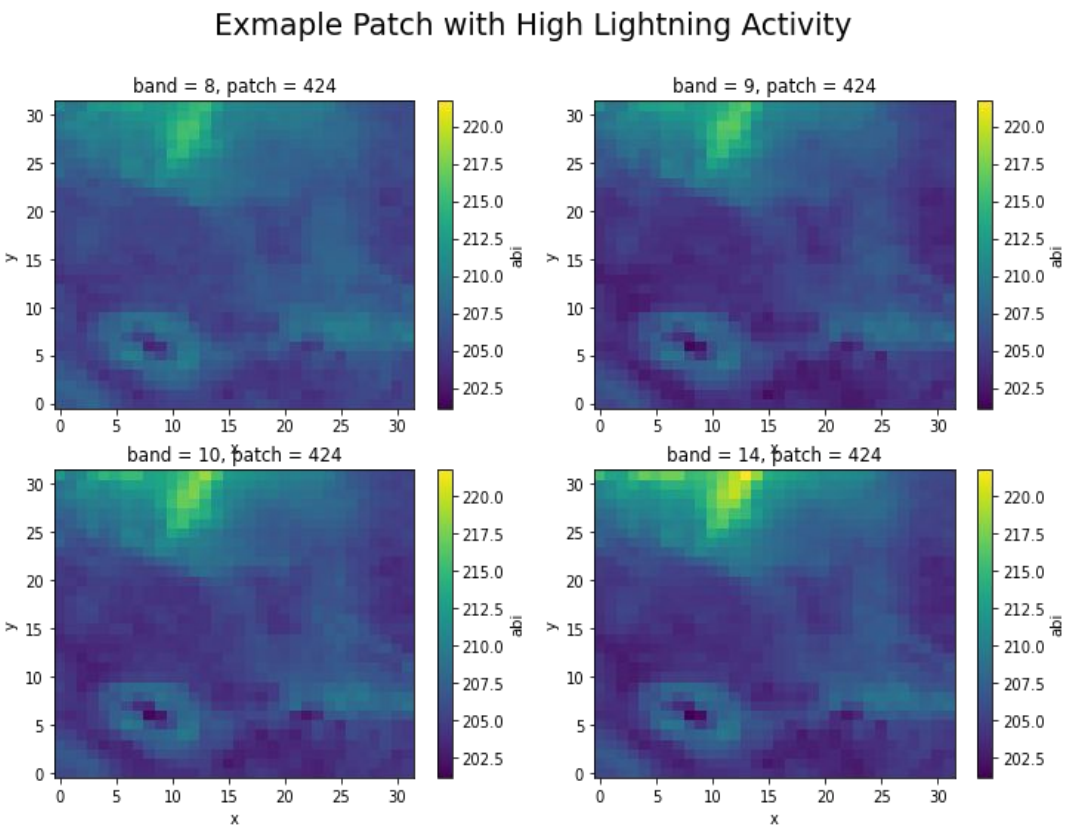
The Geostationary Operational Environmental Satellite (GOES-16), is a weather satellite centered on the Americas

- Provides high spatial and temporal satellite imagery through Advanced Baseline Imager (ABI)
- Houses the Geostationary Lightning Mapper (GLM)

ABI provides imagery through sixteen spectral bands across the visible, near-infrared, and infrared wavelengths

- Leveraged to train machine learning models to detect Convection (Lee et al., 2020)

Data



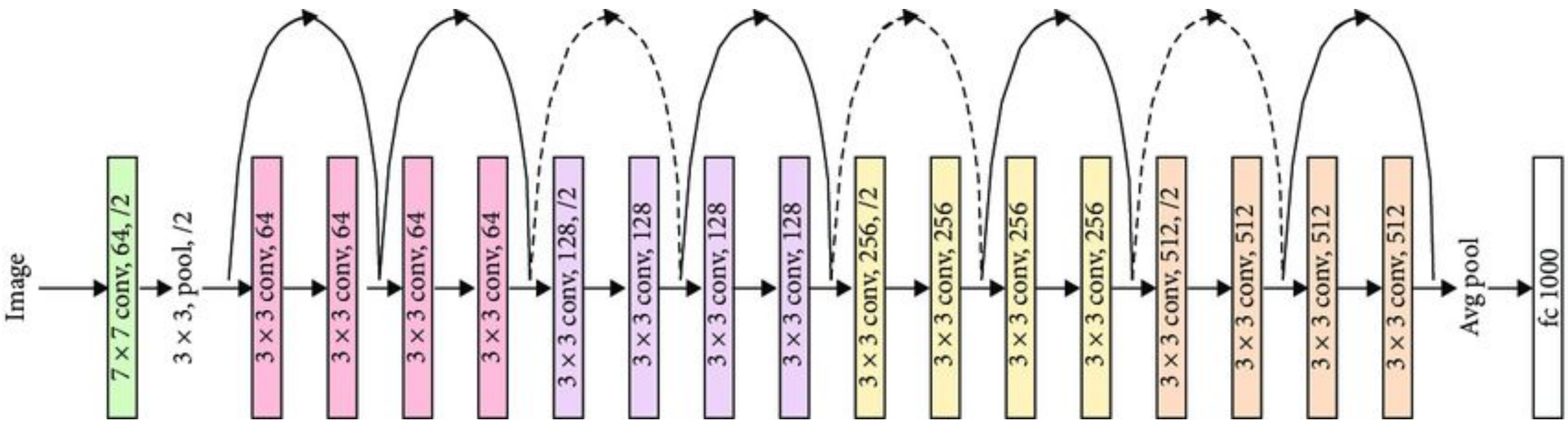
https://colab.research.google.com/github/NCAR/ai4ess-hackathon-2020/blob/master/notebooks/goes16.ipynb#scrollTo=Pcon8KbrnPNt

Continental United States GOES-16 satellite data (Patches of 32 X 32 pixels with a spatial resolution of 2 km) from 4 infrared bands

- Band 8 - Upper-level Water Vapor
- Band 9 - Mid-level Water Vapor
- Band 10 - Lower-level Water Vapor
- Band 14 - Longwave Window

Lightning count from GLM for the satellite data after 20 minutes

Machine Learning Model

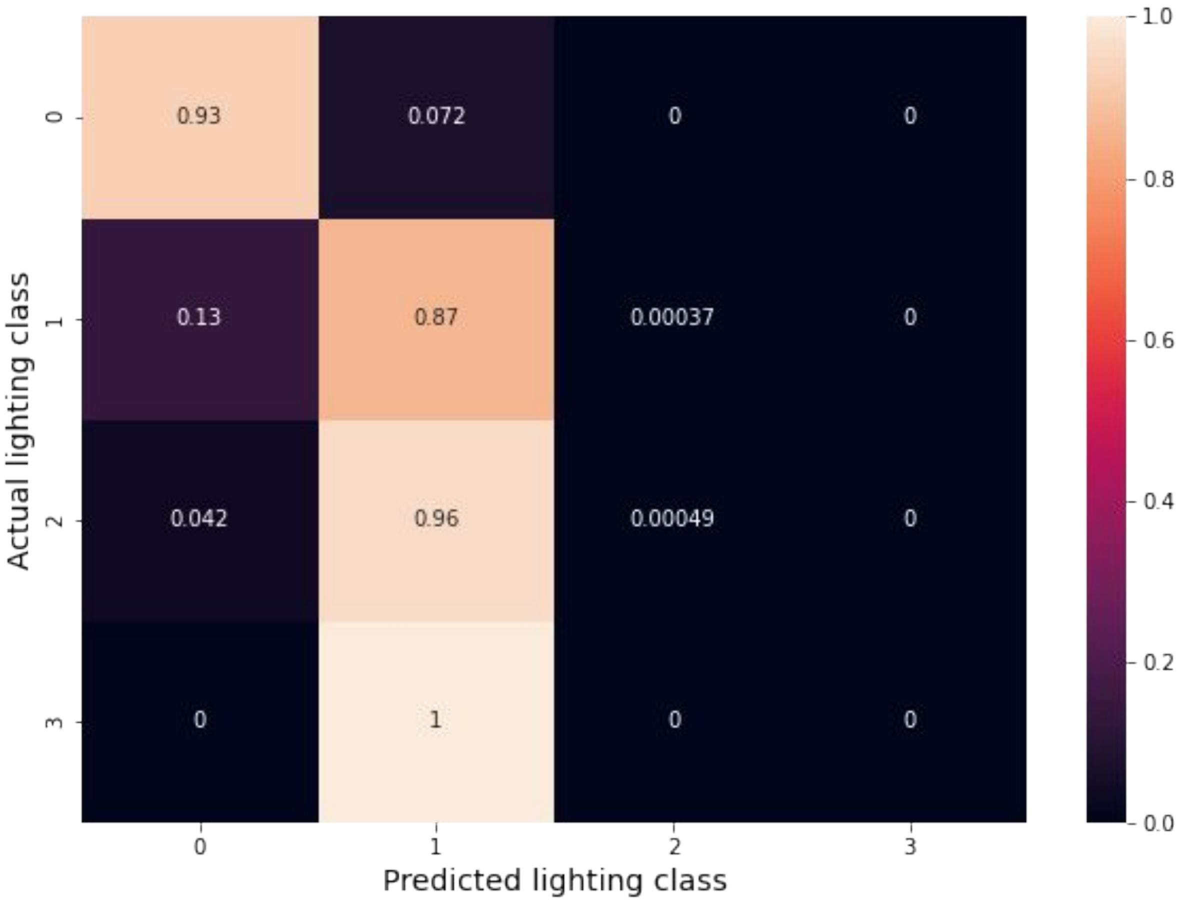


Architecture of ResNet18 borrowed from He et al., 2015

A pre-trained deep neural network with ResNet18 architecture was chosen

- Model Input: A 4-channel image with normalized bands 8,9,10,14 from GOES-16's ABI
- Model Output: Binned Lightning count (0 lightning, 1-10 lightning strikes, 10-100 lightning strikes, 100+ lightning strikes)

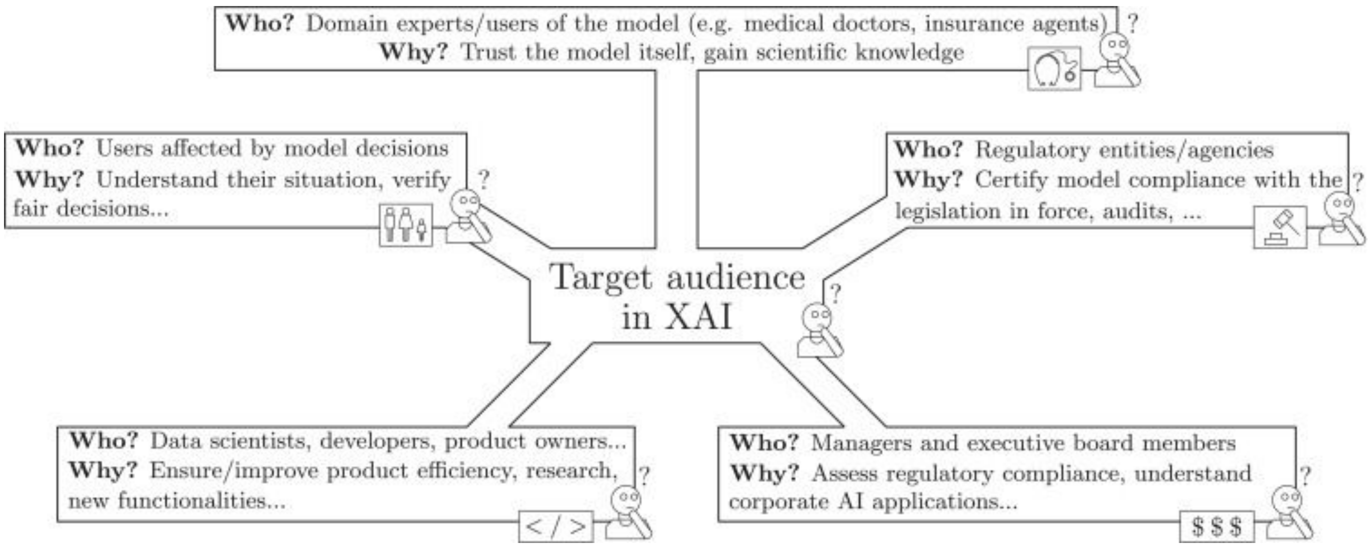
Machine Learning Results



- Model is 95% accurate at predicting lightning
- Performs poorly when predicting high lightning count images
- We can use XAI methods to interpret the model's predictions and reveal underlying phenomena

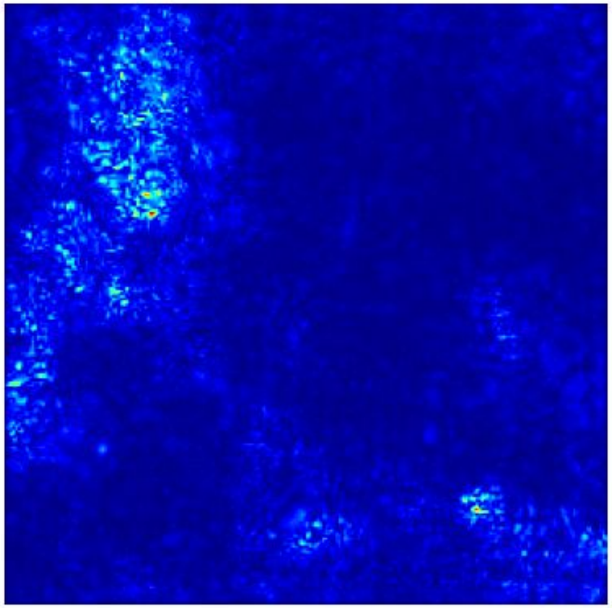
What is Explainable AI?

eXplainable AI (XAI) - A framework for machine learning that emphasizes the use of explainable models while maintaining key performance metrics and enables humans to understand, trust, and manage models



Target Audience of XAI borrowed from Arrieta et al., 2019

XAI for Neural Networks

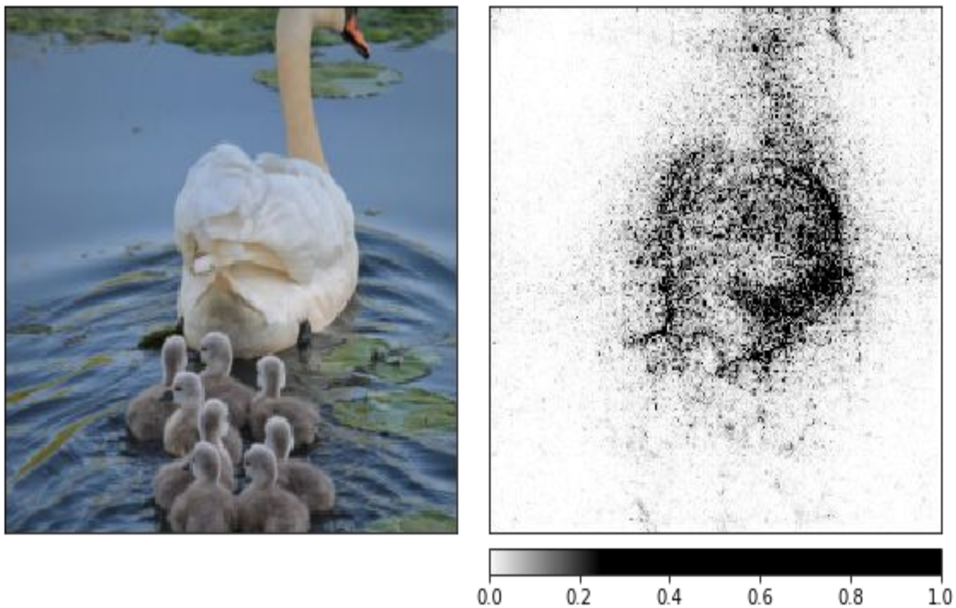


Example of saliency maps from Molnar 2020 (https://christophm.github.io/interpretable-ml-book/pixel-attribution.html)

Attribution-based methods (saliency maps) are designed for specifically for deep learning problems (i.e.image recognition) utilizing neural networks

- Perturbation-Based Methods- computes the attribution of a feature by altering the feature and observing the changes in output
- Gradient and Backpropagation-Based Methods- computes the attribution of a feature utilizing gradients and a single pass through the Neural Network

Integrated Gradients (Sundararajan et al., 2017)



https://captum.ai/tutorials/Resnet_TorchVision_Interpret

A pixel wise gradient-based attribution method

- Computes the partial derivatives of the output w.r.t to each input feature
- Average gradient as the input varies linearly from a baseline (x’) to the actual input (x)

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Relies on two major Axioms

- Sensitivity: A feature change resulting in a change of output, should have a non-zero attribution
- Implementation Invariance: Identical attributions for functionally equivalent models

Deeplift (Shrikumar et al., 2017)

A Backpropagation-based approach

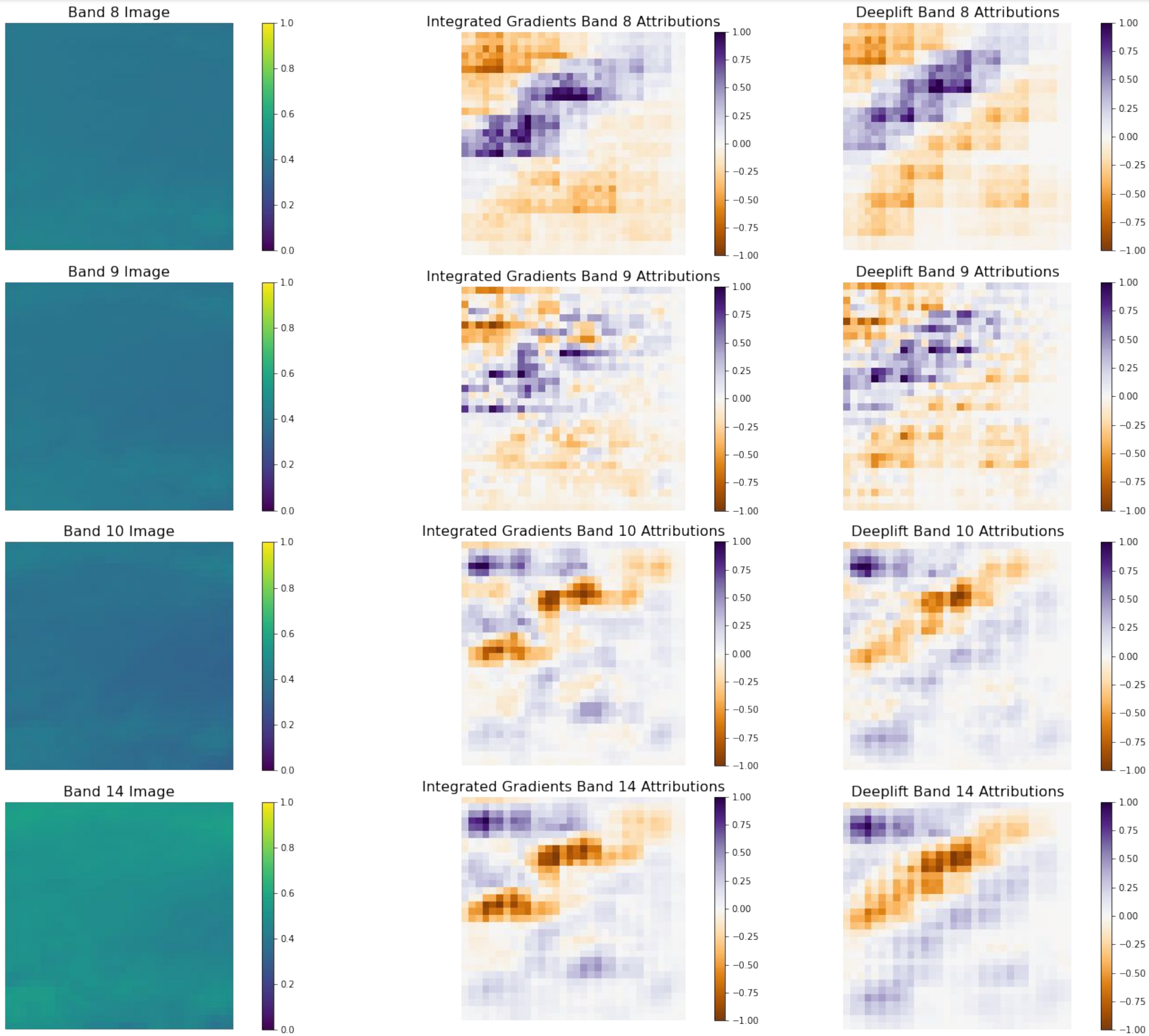
- Propagates change in output from reference back through the network
- Observes the activation of neuron and assigns contributions score

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t$$

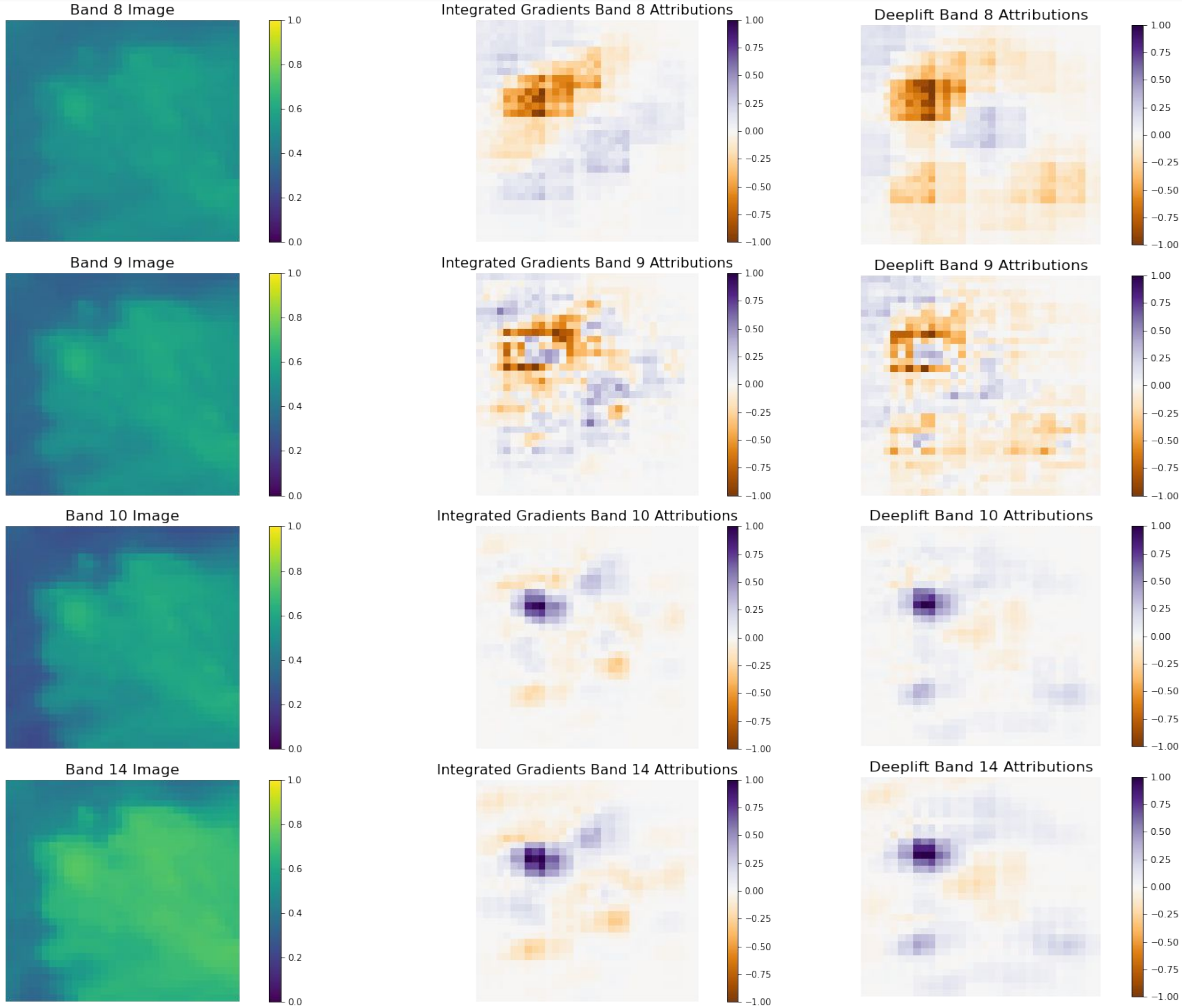
Reveals dependencies missed by gradient reliant methods

- Overcomes gradient saturation
- Treats positive and negative attributions separately

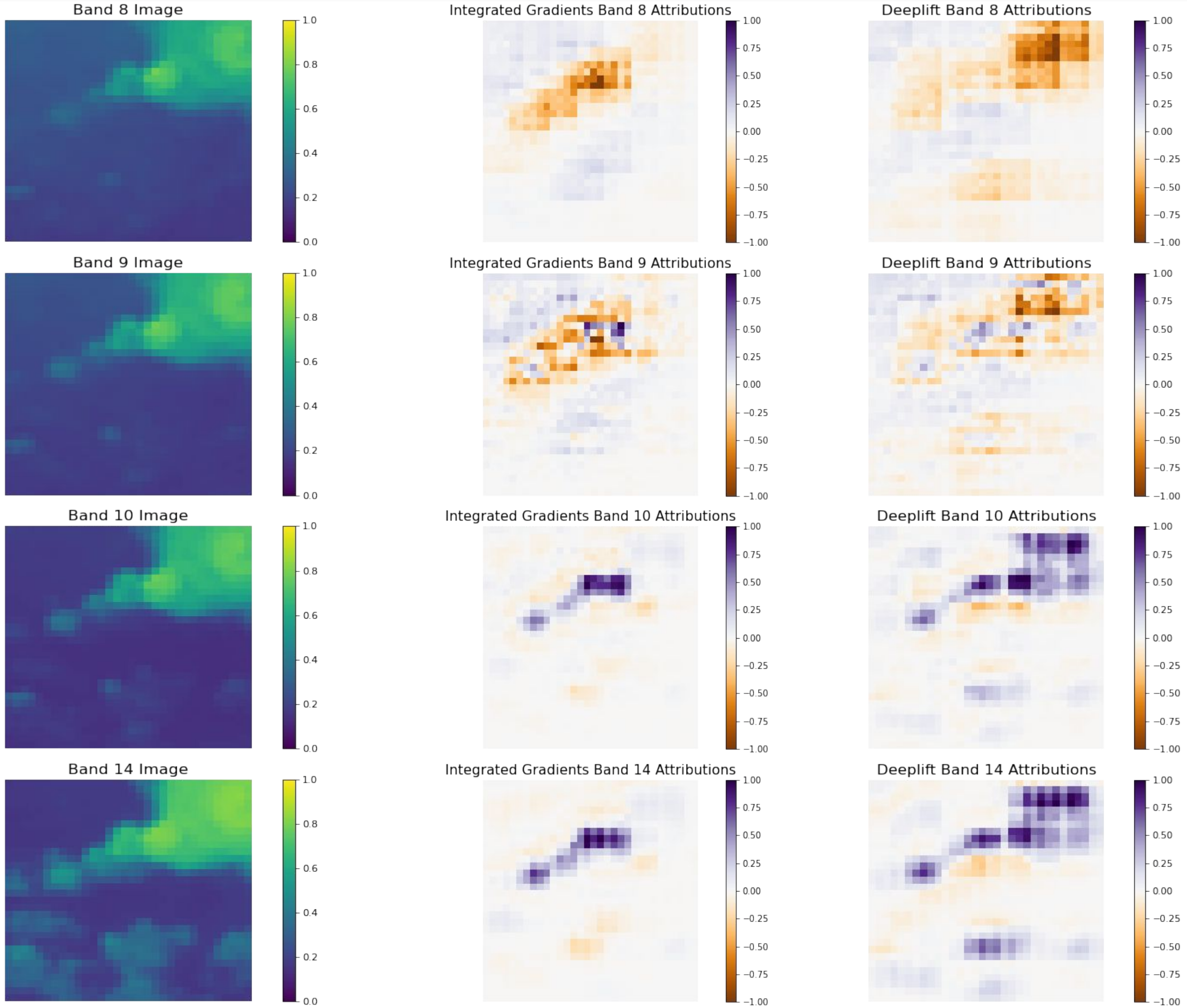
Example: Attributions for No Lightning



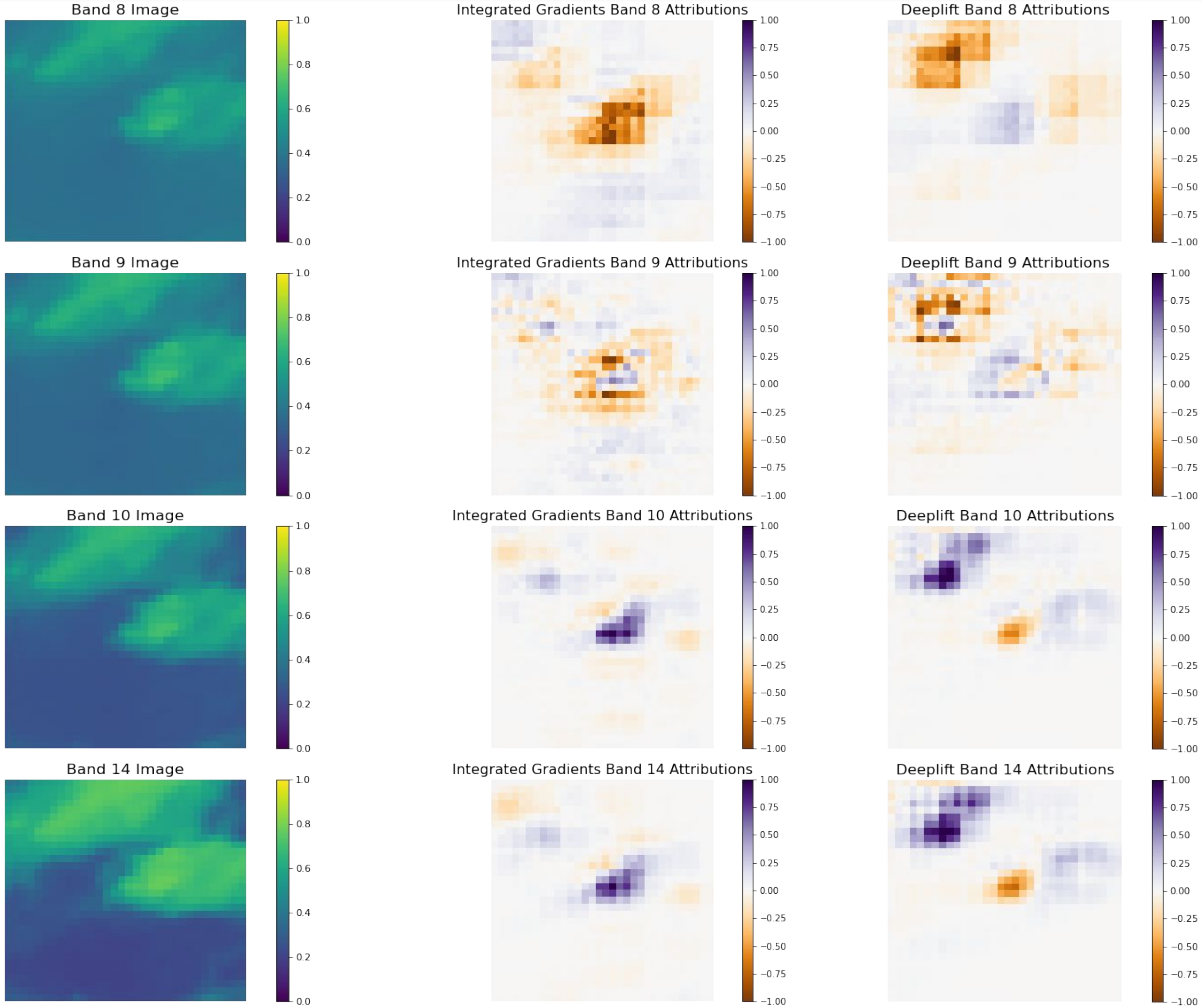
Example: Attributions for 1-10 Lightning Strikes



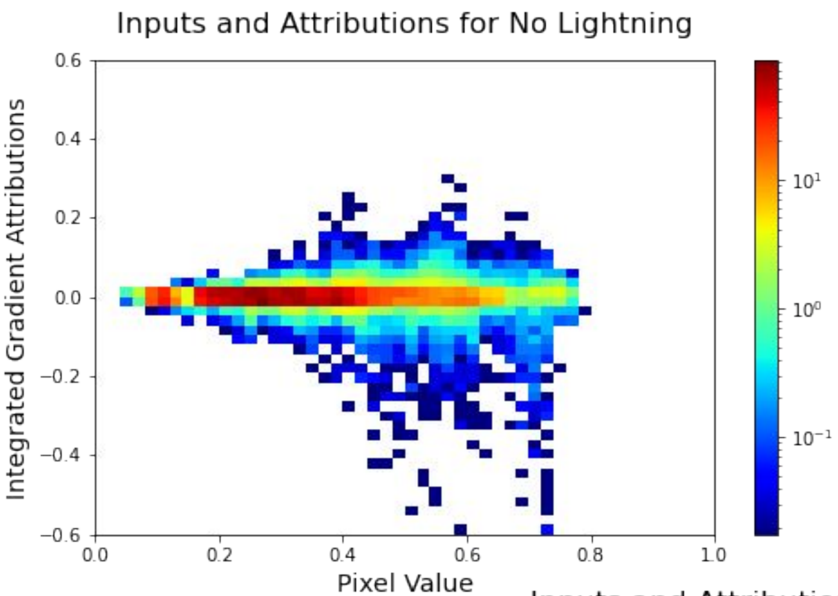
Example: Attributions for 10-100 Lightning Strikes

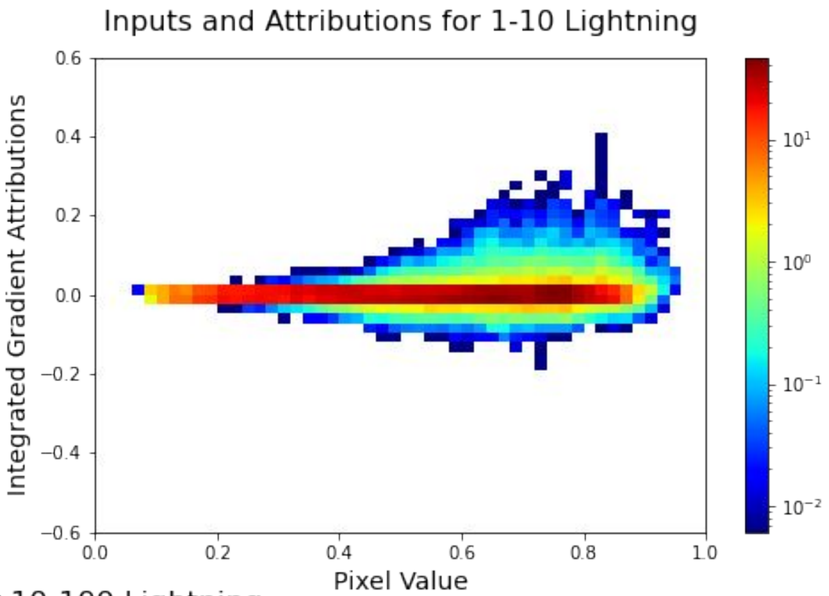


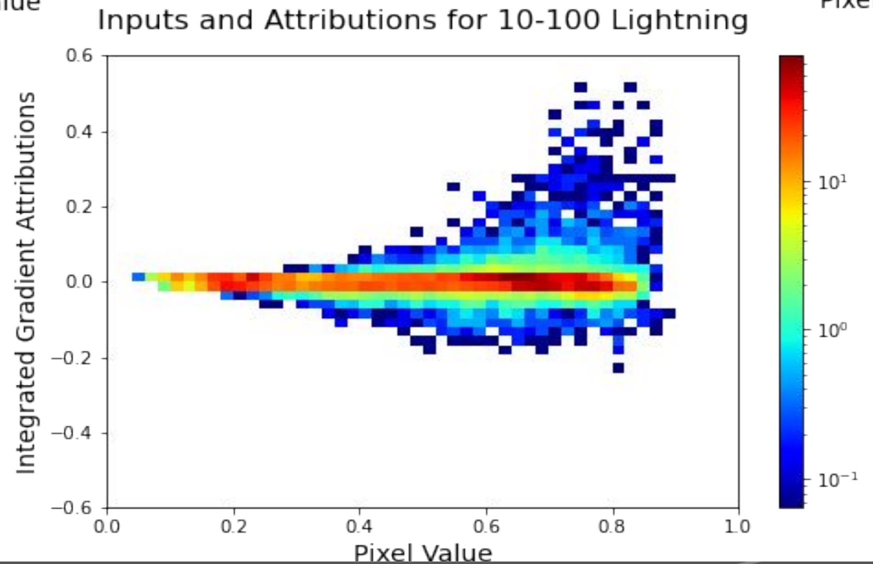
Example: Attributions for 100+ Lightning Strikes



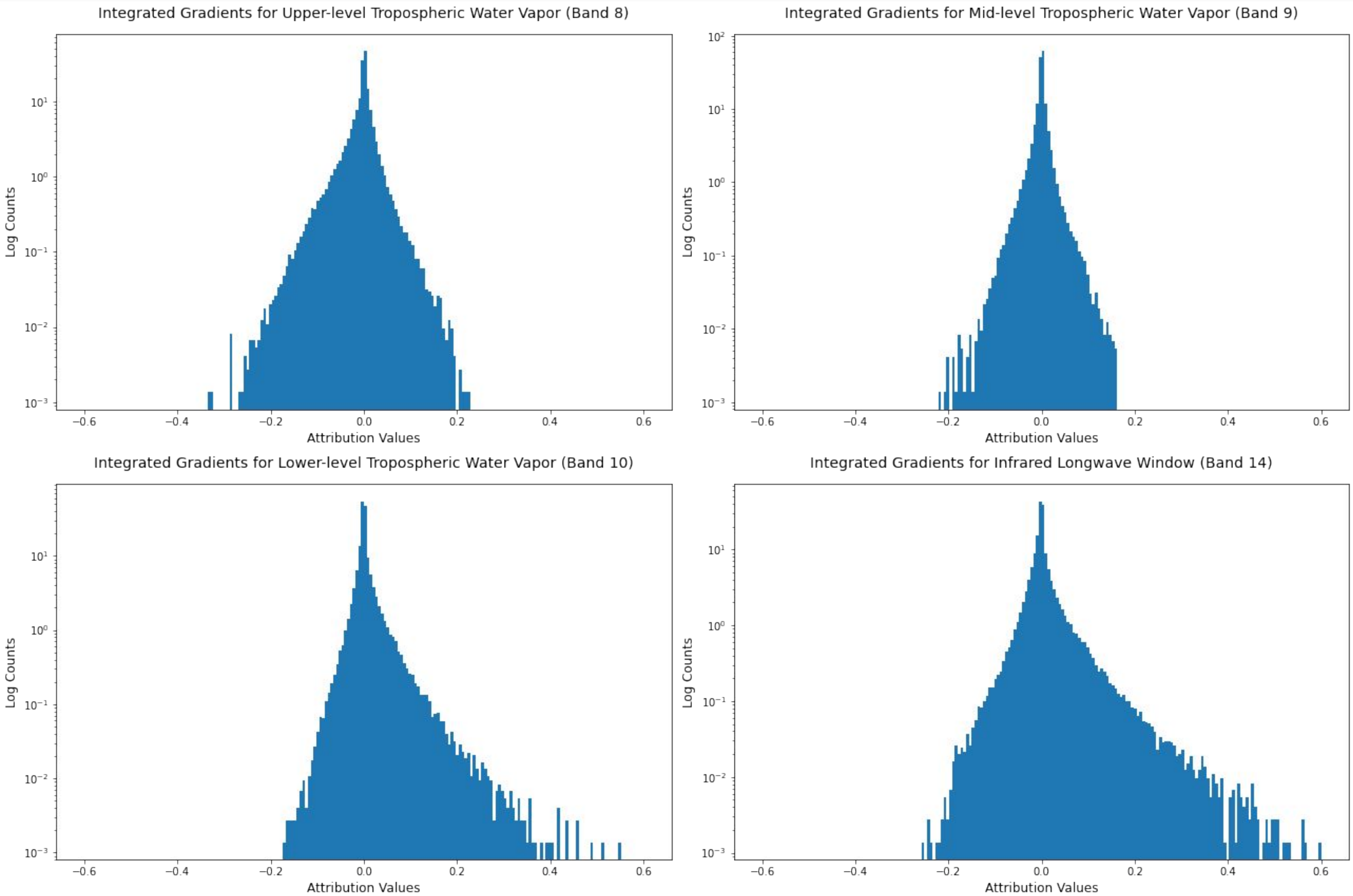
Attributions Across Classes







Attributions Across Channels



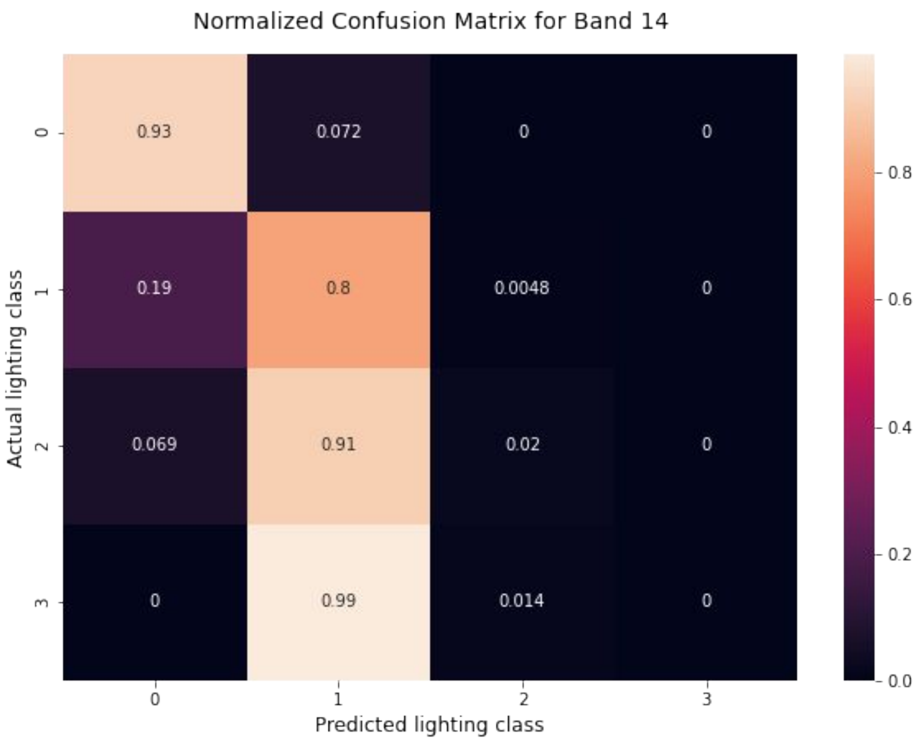
XAI Takeaways

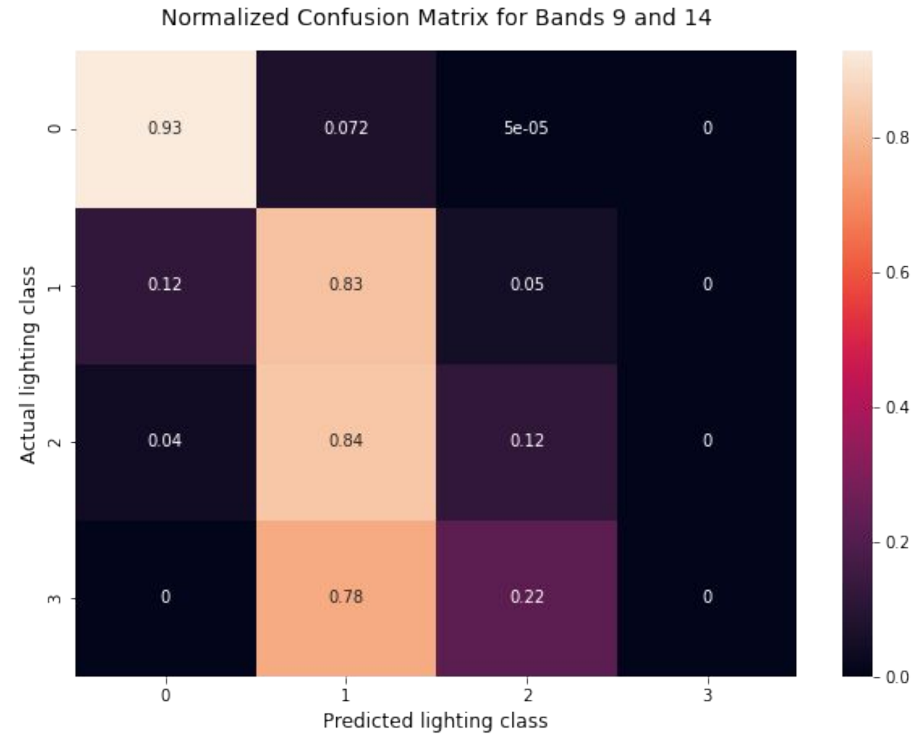
1. Integrated Gradients and Deeplift Attribution Visualizations
   - The two attribution maps are comparable, despite gradient saturation
   - Attribution maps of Band 14 reveals key features (overshooting cloud tops and lumpy cloud top surface) were learned from this band
   - Band 8 tended to have negative attributions for key features of lightning

2. Overall Attributions
   - Attributions for no lightning was almost symmetric across input values
   - Attributions for lightning had higher attributions for higher pixel values
   - Band 14 contained relatively "higher" attributions values

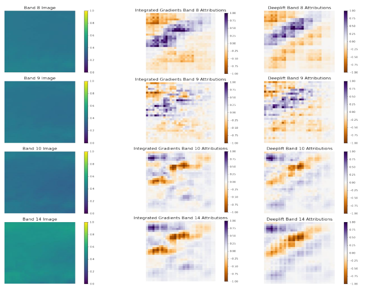
3. Model Retraining
   - Retrain model using only the "most significant band"

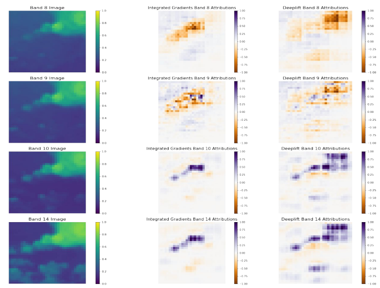
Retrained Model Results

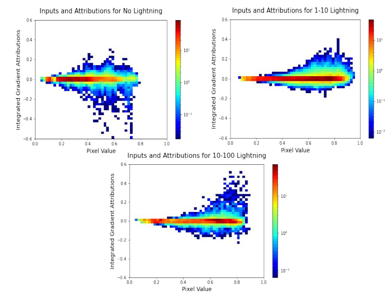


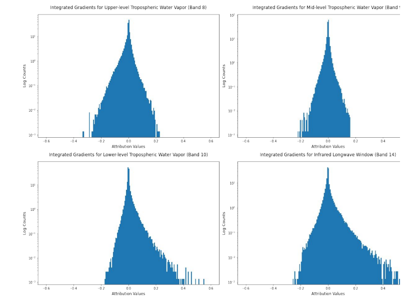


Findings









- ResNet trained on all 4 bands performs well at binary prediction
- XAI Methods indicate that images with lightning have more positive attributions
- Band 14 appears to be the most “significant” band
- Training using only Band 14 data gives similar results
- Adding band 9 achieves marginally better performance model begins to classify images as high lightning count

Questions?

prahalath.anbu_bharathi@tufts.edu