XAI for Satellite Classifiers

Joshua Friesen, Andrew McMullin, Ziming Huang, Yutong Pan, Marcel Chlupsa
University of Michigan
{friesej, mcmullin, ziming, ytpan, mchlupsa} @umich.edu

Abstract

Explainable artificial intelligence (XAI) enables researchers to gain trust in their models and new intuition regarding their results that is otherwise unavailable when working with black-box models. Here, we seek to apply XAI to satellite imagery classifiers, a subset of ML models that could greatly benefit from cogent explanations. This is done through extending SHAPley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). These techniques are then employed in a testing environment: two satellite image classifying models, one with RGB bands and the other with RGB & near-infrared (NIR). The techniques output the desired results in this environment and show promise for future more complex applications. Code for this project can be found at our GitHub repository.

1 Introduction

In recent years, deep learning has made a considerable impact in many fields, including computer vision, robotics, and speech recognition. However, due to their black-box nature, it is difficult to discern the inner workings of deep learning models. Several state-of-the-art explainable artificial intelligence (XAI) methods have been developed to better understand the abstract logic in these trained models. XAI has taught researchers new ways to think about their problems and tune their models accordingly, all while increasing trust in deep learning.

We seek to extend existing XAI algorithms to analyze deep learning models dealing with satellite imagery. This is an important problem to deal with as satellite classifiers span several industries, including national defense, agriculture, weather forecasting, mobile navigation apps, etc. A tested XAI technique for these models would help improve network accuracy, increase trust in the system, and give these industries more insight as to why the models are behaving the way that they are.

Existing XAI techniques are not able to provide these explanations because they typically consist of 2D heatmaps of contribution levels over pixels or superpixels in images. For a trivial example of the drawbacks of these explanations, one might wonder if a color was important in a model's decision making process, (e.g. dart frog poison levels). This information is not captured by available XAI techniques. A more complex example is weather prediction, in which the meteorologist must know which satellite-band a model is basing its decision off of. What would be a useful explanation for these examples is the aforementioned heatmaps spread across individual channels. In this paper, we create these explanations through modifications to existing XAI techniques SHAP [5] and LIME [3].

2 Methodology

2.1 SHAP

To create an XAI technique for the models we seek to given explanations for, we modify SHAP [5]. SHAP is based on Shapely values, a concept from game theory now being applied to XAI. In this

Preprint. Under review.

context, the game is reproducing the outcome of the model, and the players are the features included in the model. A Shapley value is a numerical representation of how much each player affected the outcome of the game. The basic idea of how to calculate a Shapley value is to calculate a player's contribution to the model output for each subset of player groups and averaging over all the contributions. The mathematically correct way of showing this is in the equation below.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!}(v(S \cup \{i\}) - v(S)) \tag{1}$$

In this equation φ is the Shapely value for some feature i, v is the black-box model, S is some subset without feature i, and n is the total number of features. The difficulty in using this approach for XAI, is that with n features, and the cardinality of a power set equalling 2^n, the exponential increase in computing cost can be prohibitive for complex models.

To get around this problem, Lundberg and Lee [5] devised the Shapley kernel. The Shapley kernel is a method of approximating Shapley values through fewer subsets. SHAP samples feature subsets and fits a linear regression model with the variables being based on averaging the results of permuted features run through the model. As seen in their paper [5], the solution to the linear regression model guarantees that the calculated coefficients are equal to the Shapley values. This procedure forms the basis of the Shapley kernel.

A base partitioning scheme exists in SHAP that gives local explanations. How this algorithm works is that it recursively partitions along the largest axis of either rows or columns of the data. Then it will split along channels until a base case is reached (a single pixel). The problem is that the algorithm typically does not reach this step because a maximum evaluation limit is reached first. Additionally, even without this constraint, with too many splits along the horizontal and vertical axes the benefits of using partitioning (groups of data) to provide explanations is lost. Intuitively, switching to a partitioning scheme prioritizing only channels as seen in Figure 2 might seem ideal, but this leads to no explanations existing for correlated bands. Instead, we opt for recursively splitting into channels, and then superpixels until the maximum evaluation limit is reached (Figure 3).

Figure 1: Spatial Partitioning

Figure 2: Channel-Wise Partitioning

Figure 3: Combination Partitioning

The next significant change is the use of maskers. SHAP generates a tree of maskers to run the feature subsets through the model. As shown in Figure 4 row "PartitionSHAP Maskers", the default SHAP maskers are either solid or blurred with Inpaint. To match the previous step, before being broken spatially the maskers are split across channels and into superpixels. The result of this can be seen in 4 row "Modified Maskers". Additionally, the change in the type of masker plays a large role in determining the resulting explanations, so a random color per pixel method was also created to provide better results.

The first step in evaluating the effectiveness of an XAI technique is to create a testing environment conducive towards it. In this case our main focus is to explain multi-spectral systems, so creating a model with the following characteristics is necessary: is a convolutional neural network, is easily explainable, makes predictable decisions based on channels other than RGB. A model such as this meets the requirements of a deep learning system in which we predict a channel-wise XAI technique will give explanations with better information than a traditional explainer.

EuroSAT [2] is one of the largest publicly accessible multi-spectral datasets, taken by the Sentinel-2 satellite. It has thousands of images divided into ten categories, each with 13 channels. By reducing the output categorization to two options (forest and river) and reducing those rasters to four bands (RGB and near infrared) it is possible to create a model that meets the requirements described

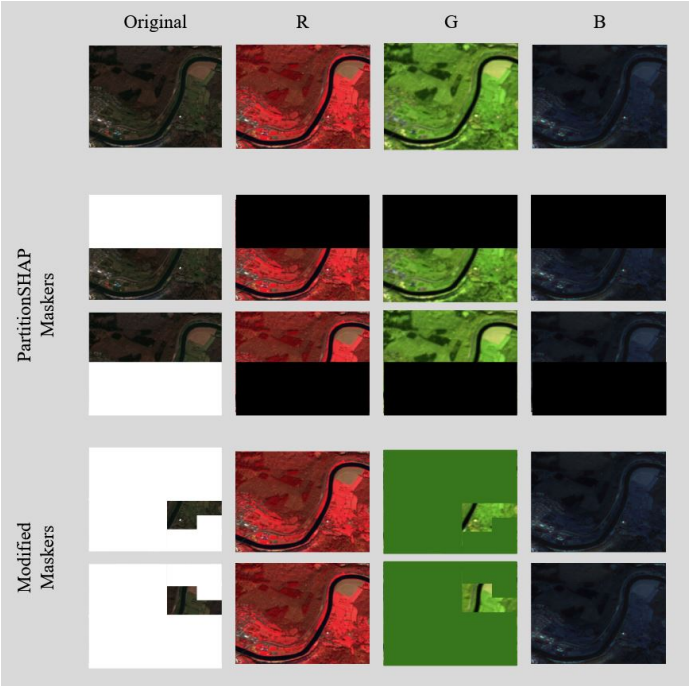
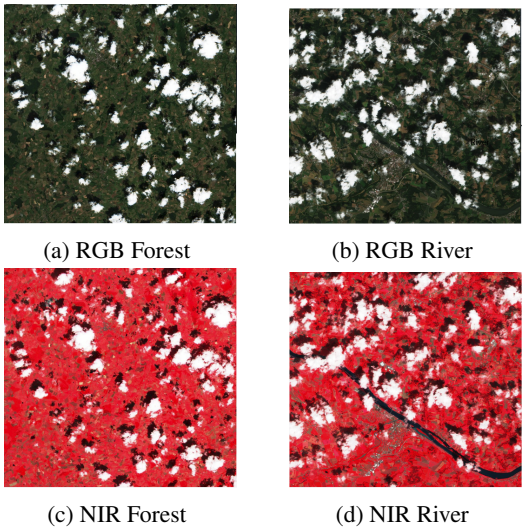


Figure 4: SHAP Maskers

above - one that we can easily predict the results of. The reason for this ease in predictability is that chlorophyll is present in plant life, which covers a substantial portion of both river banks and forests. It also reflects near infrared whereas water absorbs it. This creates a sharp contrast in the NIR spectrum which is not present in RGB channels. With this knowledge, we can predict that a model with RGB and NIR categorizing river and forest data would put significant weight on NIR. It would stay relatively simple with one extra band, contain the same base architecture as more complex geospatial models, and have the ability to host 3D XAI techniques. We can see an example of this data in Figure 5.



(a) RGB Forest (b) RGB River

(c) NIR Forest (d) NIR River

Figure 5: RGB & NIR Data Comparison

It is necessary for this model to show reliance on NIR instead of assuming it. One way of demonstrating this is to create another model identical to the first in training data, testing data,

and output categorization, but without the NIR band. Performance improvements in the NIR model in either training speed or validation accuracy are then indicative of the CNN putting emphasis on the NIR band to make predictions. This combined with sufficient model accuracy (as confirmed by a confusion matrix) completes a testing environment in which we can demonstrate the efficacy of XAI methods in multi-spectral systems. Both models were created off of a base ResNet-50 architecture. The hyper-parameters were batch size = 64, learning rate = 0.1, layers = 50, and number of epochs = 35.

2.2 LIME

LIME is another XAI technique that returns a mask over the provided image, to show what parts of the image the model is using to make its decisions. We use LIME as a baseline explanation that gives us another point of comparison to our modified SHAP algorithm.

LIME is built to take in a three channel RGB image. First, a 2D greyscale image is converted to a 3D RGB image. Then LIME generates neighborhood data by randomly perturbing features from the instance. This is done using scikit_image.SegmentationAlgorithm with algo_type 'quickshift'. It then learns locally weighted linear models on this neighborhood data to explain each of the classes in an interpretable way. The linear models are returned using a lime.lime_image.ImageExplaination class. This class uses the model to create the mask that shows what parts of the image contribute the most. LIME is built to use all RGB values of the image. This image is of size, Height x Width x 3.

Because LIME already returns a mask over the image with respect the RGB values, we modified this to return a mask over the image and also over the different channels of the image. We modified LIME to not only return one mask but to return 4 or 5 masks depending on if an NIR channel was provided. These masks are the explanations of the original RGB values, the red channel, the green channel, the blue channel and the optional NIR channel, for the 3-4 single channels it is passed into the LIME explainer using 2 zeroed out channels to make it of size Height x Width x 3.

Breaking up the image and modifying LIME to take 4 channels is used to get an understanding of how the model performs on just the red, green, blue, and NIR channels separately. You can see in the results of LIME that we get a mask that in some cases prove that some of the channels do not make much of a contribution to the output/classification.

3 Related Works

While numerous papers have been published exploring XAI, there are few that provide methods of explanation for models dealing with multi-modal CNNs (convolutional neural networks), or any network trained on a 3D dataset. Among the few, two studies are listed that follow a similar approach to the work presented in this paper. The work in [4] modifies common 2D XAI methods to provide an explanation for a network analyzing 3D CAD figures. The explanation determines which geometric features were most utilized by the model to determine the cost of manufacturing and automation. The XAI techniques modified were LIME, Grad-CAM, Sensitivity Analysis (SA), and LRP. The techniques were evaluated and compared using a common 3D CAD model of a part and a 3D CNN that was pre-trained with an autoencoder, then trained again with the NeuroCAD trainset via transfer learning. While the results found that the new 3D methods were only able to highlight significant points of interest on a part that indicated an abstract feature, it did open up doors for more complex abstract-features to be examined than previously possible with 2D inputs.

The second work [6] adapted XAI methods to obtain visual explanations of Alzheimer's disease classifications made by 3D-CNNs. Within their work, three approaches were developed, each of which with their own perks and limitations. The Sensitivity Analysis by 3D Ultrametric Contour Map (SA-3DUCM) worked best on the homogeneous regions of the brain and was model-agnostic and more adaptable to accommodate other types of 3D images. The two other approaches using 3D-CAM and Grad-CAM [5] were more receptive and analytical of the 3D-CNN with their production of heatmaps. The main pitfall was that they would have lower resolutions than that of the original. While less adaptable than SA-3DUCM, these approaches could still be used in other medical contexts involving 3D images or videos.

While [1] did not deal with XAI techniques or explicitly with multi-channel networks. However, it does offer insight on how XAI techniques are capable of being extended and applied to yield to

extract visualizations and interpretations of various methods used in image search, image retrieval, and determining image similarity. Among the methods extended in this work were 3 that have been previously mentioned: CAM, SHAP, and LIME. The technique, integrated gradients, was found to produce the most unhelpful explanations because rather than focusing on important pixels or pixel groups as methods like SHAP does, minor pixel changes do not change the classifiers if they are well trained [1]. Additionally, it served to highlight the development of a new approach to obtain more consistent explanations of image search interpretability by means of projected Harsanyi Dividends.

4 Results

4.1 SHAP

The first results of this project are the demonstrations of an effective testing environment. As referenced in the methodology section, the first step in this is to show a performance improvement of the RGB & NIR model over the pure RGB model. After running these models we have confirmed that RGB & NIR outperforms RGB in both training speed and final accuracy.

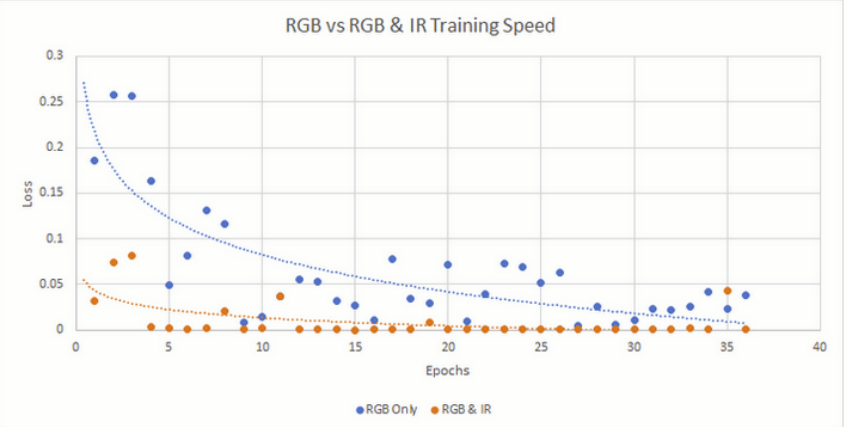


Figure 6: Model Performance

As shown in the Figure 6, both models achieve impressive accuracy ratings eventually, but the loss function in RGB and NIR decreases substantially faster and over fewer epochs than in the RGB only model. Both models showed minimal performance improvement after around thirty epochs, so training was capped at thirty-five.

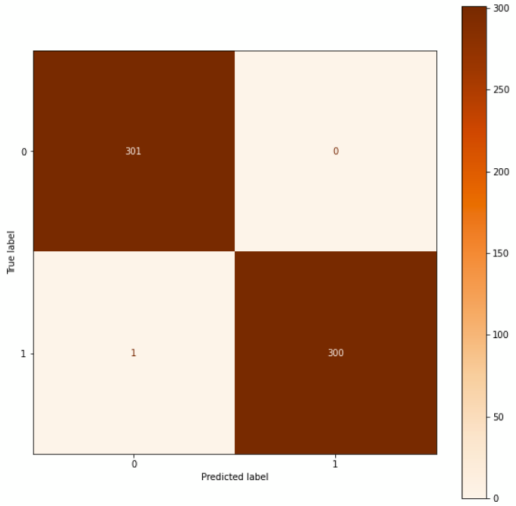


Figure 7: RGB & NIR Confusion Matrix

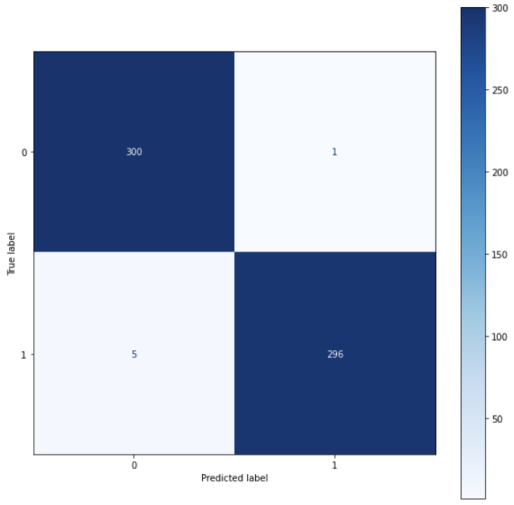


Figure 8: RGB Only Confusion Matrix

Figures 7 & 8 show the confusion matrices with the results of the RGB and NIR models run over the validation dataset, consisting of 602 images. 601/602 images were correctly classified with an average loss of 0.0037 in the NIR, compared to 596/602 images correctly classified in the RGB only model, with an average loss of 0.0277. These statistics fit well within our anticipated performance ranges and support the hypothesis of NIR playing an important role in model decision making.

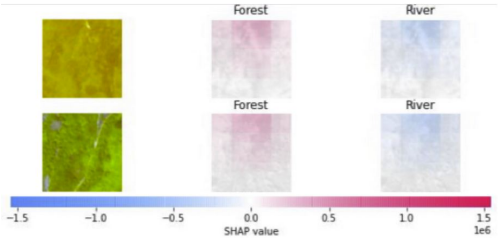


Figure 9: Standard SHAP Explanation

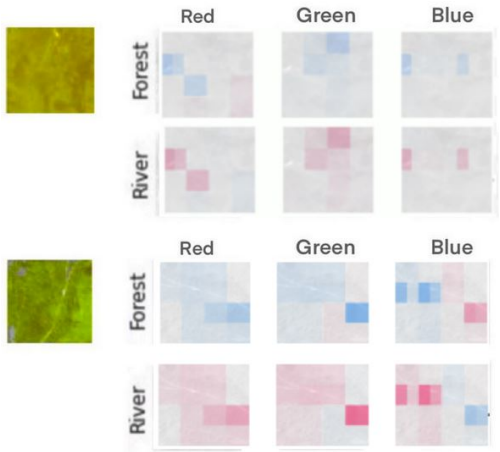


Figure 10: New SHAP Explanations - 3 band model

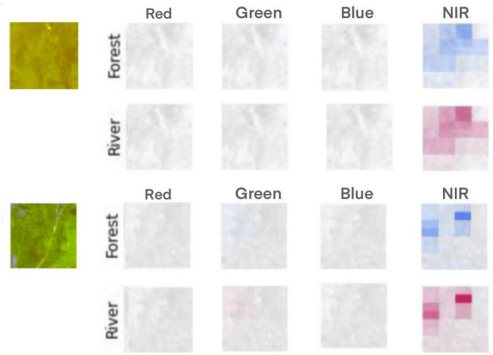


Figure 11: New SHAP Explanations - 4 band model

Finally, the results of our XAI technique can be seen in Figure 11 and compared against standard SHAP in Figure 9. We can also compare Figure 10 against Figure 11 to see the difference between a model where it makes sense to use this XAI (4 band model) and a model where it does not (3 band model). In the 4 band model we can see demonstrably high Shapley values in the NIR band and low values in RGB for images run through the NIR model, matching all previous referenced work. These results once again support our hypothesis of how this model makes decisions, and matches the performance comparison with the RGB only model. The explanations for the same images with standard SHAP provide minimal useful information, with vague image regions being highlighted and

no specific features. The absence of specific feature recognition reinforces the notion that a model with high channel-wise dependence looks less at individual image regions. With these results it is possible to expand testing to more complex multi-channel models in future works. We anticipate this XAI technique being able to provide consistent channel-wise explanations as complexity scales upwards.

4.2 LIME

LIME output looks different than SHAP but we can use the outputs of LIME to compare to the channel-wise outputs of SHAP. Here we get an explanation that displays the parts of the image that most contribute to the classification for each individual band.

As we can see in Figure 12, spatial explanations are dominant without significant changes to the base algorithm. LIME provides little to no channel-wise explanation as it stands, and can not express the same information that modified SHAP can. We can see in Figure 13 that the LIME identified features are evenly distributed throughout each band.

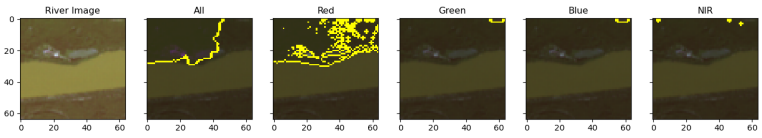


Figure 12: LIME Explanation on River Data

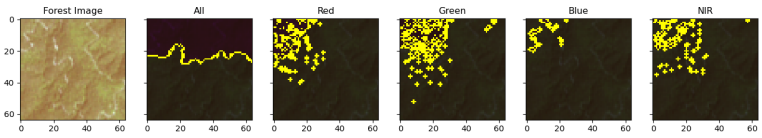


Figure 13: LIME Explanation on Forest Data

5 Conclusion & Discussion

After demonstrating the model's dependence on the NIR band without an XAI technique, we can informedly ascertain the utility of our XAI technique's explanations. The demonstration of significant Shapley values in the NIR band for the model's classifications embody the situations in which we predict it will be superior to standard XAI techniques. If standard SHAP is in turn not able to show the channel-wise dependencies suggested by the two model's performance differences, we have effectively demonstrated a situation in which this technique gives better explanations than existing XAI. From this point onwards our group can scale up this XAI to run on more complex models utilized by research groups today. Continual demonstration of this technique's superiority over other XAI techniques in multispectral models will validate its use for research and industry needs.

References

[1] Mark Hamilton, Scott Lundberg, Lei Zhang, Stephanie Fu, and William T Freeman. Model-agnostic explainability for visual search. *arXiv preprint arXiv:2103.00370*, 2021.

[2] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *CoRR*, abs/1709.00029, 2017.

[3] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[4] Raoul Schönhof, Artem Werner, Jannes Elstner, Boldizsar Zopcsak, Ramez Awad, and Marco Huber. Feature visualization within an automated design assessment leveraging explainable artificial intelligence methods. *Procedia CIRP*, 100:331–336, 2021.

[5] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[6] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification. In *AMIA annual symposium proceedings*, volume 2018, page 1571. American Medical Informatics Association, 2018.

Author Contributions

All authors contributed to writing this report and to team meetings.

- Joshua Friesen created the SHAP technique, datasets, and testing environment (ML models).
- Andrew McMullin created/programmed the LIME Implementation, currently in *lime/lime_image.py*, and created the *main_lime.py*.
- Yutong Pan helped with LIME and testing model for LIME.
- Ziming Huang helped with LIME and reviewed the report.