IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641426

COMPARATIVE EVALUATION OF XAI METHODS FOR TRANSPARENT CROP YIELD ESTIMATION USING CNN

Arun Venugopal, Mahdi Farnaghi, Raúl Zurita-Milla

Faculty of Geoinformation Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands

ABSTRACT

While Deep Learning (DL) has significantly improved the estimation of crop yields from satellite imagery, the intricacies of the decision-making processes within the utilized DL models remain obscure. To ensure transparency and dependability in crop yield predictions, unravelling the complex mechanisms of DL models is imperative. This task is challenging due to the varying outcomes presented by different eXplainable Artificial Intelligence (XAI) methods. This study examines a range of eXplainable Artificial Intelligence (XAI) techniques applied to Convolutional Neural Networks (CNNs) for Soybean yield estimation from Sentinel-2 satellite imagery. Methods like Layerwise Relevance Propagation (LRP), SmoothGrad, Deep Taylor, and gradCAM have been employed to elucidate the CNN model, producing saliency maps that are subsequently evaluated through a perturbation analysis. The study also examines how each XAI method identifies the model's focus on Soybean fields compared to other crop fields or land uses. Our findings indicate that LRP outperforms other methods, offering more accurate saliency maps and highlighting critical spatial information for crop yield estimation. The knowledge gained about the XAI methods facilitates understanding the behaviour of complex CNN model architecture used for crop yield estimation in the future.

Index Terms—Crop yield estimation, eXplainable Artificial Intelligence (XAI), Perturbation analysis, saliency map, Deep Learning (DL), Remote Sensing (RS)

1. INTRODUCTION

To sustain the projected global population of 9.3 billion people, 60 percent more food must be grown by 2050. Methods that accurately and timely estimate crop yield are essential to achieve such an increase in food production [1]. Deep Learning (DL), a subdivision of Machine Learning, is prevalently utilized for crop yield estimation, especially from large-scale and high-resolution satellite imagery [2]. DL can automatically learn complex feature representations from raw data, making it more effective for handling high-dimensional data and capturing intricate patterns [3]. Even though DL techniques allow us to estimate the crop yield with acceptable accuracy, there needs to be more focus on explaining the model's behaviour and decision-making process when estimating the crop yield. Composed of several interweaved layers of computational units, it is hard to decipher how DL models work and explain how this collective computational unit arrived at its estimation [4]. Considering crop yield estimation, it is crucial to determine if the model effectively estimates the yield by utilizing the correct spatial information from the input data [5]. This ensures that the model is not biased and learns effectively from the provided inputs; thus, it can be trusted to be applied to future datasets and estimate crop yield in the coming years. EXplainable Artificial Intelligence (XAI) techniques can be employed to demystify the complex decision-making process of DL models and validate the model's usage of appropriate spatial information [5].

The primary goal of this research is to assess various XAI techniques, intending to decipher the mechanisms through which DL models process and leverage spatial information for accurate crop yield prediction. We focus only on the Convolutional Neural Networks (CNN) models, commonly used DL for image-based inputs like satellite imagery and popular for crop yield estimation [2]. We use *perturbation analysis* to quantify the feasibility of different XAI methods applicable to a base CNN model that estimates Soybean yield. Perturbation analysis involves systematically altering input features and observing corresponding changes in model predictions, allowing us to determine feature importance based on the magnitude of changes. In the case of CNN-based models, perturbating different locations in the input image allows us to understand how the model deals with spatial information, especially when the input data encompasses satellite imagery that covers both the crop type of interest (in this study, Soybean) and other areas. Each XAI method highlights different areas in the image, indicating that the model has used those areas to estimate the yield. By perturbing the input image according to the areas highlighted by each XAI method and measuring the changes in the model's accuracy, the perturbation analysis provides insights into the explanatory capability of each method [6]. To further assess the reliability of the XAI methods, a comparison of these methods is made to identify the spatial importance given by each XAI method to the specific crop type compared to other areas in the image.

2. METHODS

To achieve our objective, a dataset is prepared by generating image patches with a target crop yield for each patch. Next, a base CNN regressor model is trained to receive the input image patches and predict the yield per patch. Since this study focuses on evaluating the explainable methods rather than building an efficient CNN architecture, we use the base architecture represented in Figure 1. Different XAI methods are implemented afterwards to explain the CNN model using saliency maps. Subsequently, a perturbation analysis is performed to evaluate the sensitivity of the various XAI methods and identify which method provides an optimal explanation. Finally, we look at how each XAI method prioritizes the crop type of interest in relation to other regions.

2.1 Data preparation and model tuning

The crop chosen for analysis is Soybean, and our study area corresponds to the five leading Soybean states in the US, namely Illinois, Indiana, Iowa, Minnesota, and Nebraska. Soybeans were chosen due to the U.S.'s role as a leading global producer due to the rising demand in recent years [7].

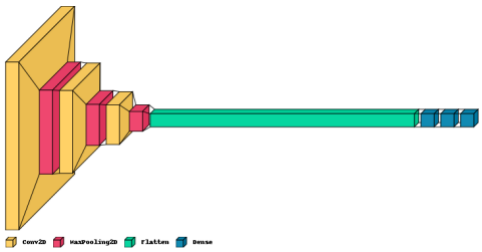

Figure 1: CNN model architecture

The Crop Land Data Layer, provided by USDA (United States Department of Agriculture) and accessed from GEE (Google Earth Engine), was used from 2017 to 2021 to identify Soybean fields. To simplify the process, we used the average of the Sentinel-2 images of July to predict yearly yield. July was selected because soybean typically reaches its mid-growth stage during this month, as indicated by the crop calendar [7]. The Sentinel-2 images were extracted from GEE, and crop yield data were downloaded from USDA (https://www.nass.usda.gov/). The pixel values of the Sentinel-2 bands were normalized to a range between 0 and 1. Afterwards, 256×256 image patches were created from the normalized dataset (12 bands of Sentinel-2) and a mask layer delineating the Soybean fields. These patches served as the input for the CNN model. For each patch, the target yield was calculated through a down-sampling of the county-level yield based on the area of the fields covered by each patch. Consequently, a dataset of 18816 records was constructed. The CNN model has three convolutional layers (3x3 kernel) and three max-pooling layers (2x2 kernel). It then flattens and connects to a dense network with 64 and 32 nodes. All layers use ReLU activations except the output layer, which uses a linear activation. Since the output of the CNN model is the yield value in Kg, the loss function used to train the model was MSE (Mean Squared Error). A learning rate of 10^{-4} with a batch size 64 and the Adam optimizer was fixed. The model was trained for 50 epochs. 16225 patches from four states, Iowa, Illinois, Minnesota, and Nebraska, were taken to train the CNN model. 25% of the patches from the four states were randomly chosen to improve the validation loss curve. The remaining 2591 patches from Indiana were kept aside for testing and evaluating the model.

2.2 Explainable methods

Saliency maps are a visualization technique used in XAI to interpret and understand the decisions made by Deep Learning models, especially for CNNs [8]. After training the CNN model, saliency maps are generated for the test dataset. Saliency maps visually highlight the most critical features in an input image that contribute to the network's prediction [9]. Different algorithms can be used to generate saliency maps. Below is a brief explanation of the algorithms used in this study.

SmoothGrad is a method that perturbs the input data by adding random noise and averaging the gradients obtained from multiple noisy samples. This process helps to reduce the influence of random noise on the gradient estimates and provides a smooth and stable interpretation [10].

Deep Taylor is a method based on the Taylor decomposition, used to simplify a complex formula by expanding the function as a series of higher-order partial derivatives [11]. When implemented for CNN, it considers the relationship between input features and the estimated class/value by decomposing the predictions based on the contributions from each input feature.

Layerwise Relevance Propagation (LRP) is an algorithm that propagates the weights of each neuron in a layer from the last layer to the input to identify which pixel contributed the most to the output [12]. LRP's emphasis on the flow of relevance through the network layers makes it a suitable method for understanding the overall importance and information flow [13]. We used two variants of LRP, LRP-A and LRP-B, which differ in their propagation rules and methods of relevance assignment.

GradCAM extends the idea of CAM (Class Activation Maps) by incorporating gradient information from the target yield value [14]. Since our model has a regression output, GradCAM is adapted by attempting to highlight regions in an input image that have a higher influence on the model's regression output.

2.3 Perturbation Analysis

After implementing the XAI methods explained in section 3.2 and generating the saliency maps for the test data, perturbation analysis was performed to assess the reliability of each XAI method. Though the general purpose of the

saliency maps is to outline the important spatial information in the input images, the algorithm for each method listed in Section 3.2 generates distinct saliency maps. Therefore, perturbation analysis was utilized to examine the differences. Figure 2 depicts the implemented workflow for perturbation analysis applied to the saliency maps generated by each algorithm. The analysis starts by splitting an input patch into an 8×8 grid (64 cells of 32×32 pixels). These cells were ranked based on their mean pixel values. The cell with the lowest rank was perturbated with Gaussian noise, and the patch yield was estimated by passing it to the CNN model. The difference between the predicted and actual yield was saved to a list. This process was continued iteratively for the other cells based on the ranking in ascending order. The area under the curve for each explainable method was used to generalize the outcome of the perturbation analysis for the entire study region.

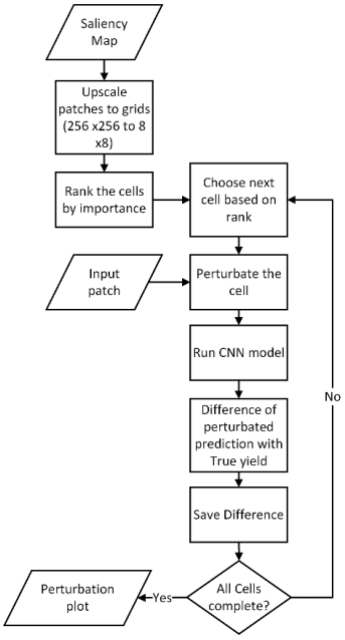

Figure 2: Perturbation Analysis Workflow

2.4 Focus on the Crop Type

In the last step, we wanted to identify the spatial importance given to the Soybean fields by each XAI method compared to other crops and land use types in the image. The pixel values of the saliency map, generated by each method, were categorized into two arrays by overlaying the saliency map with the Soybean field layer. The pixel values within the Soybean field polygons were added to the first array, and the pixel values of other regions were added to the second array. The mean of these two arrays was calculated, and the ratio of pixels in Soybean fields to pixels in non-soybean fields was plotted for all saliency maps. This ratio indicated the level of significance given to Soybean fields with respect to other areas in the image. This offered insight into the priority given to Soybean fields by each explainable method.

3. RESULTS

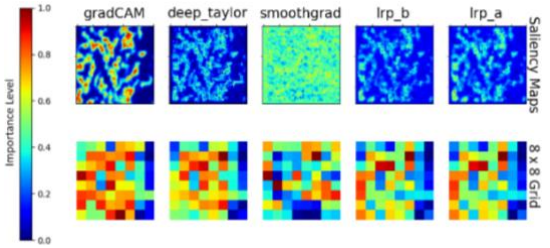

Figure 3: Perturbation Analysis based on the ranking of saliency

Figure 3 shows the saliency maps of one representative patch from Indiana and the ranking of cells on which the perturbation occurs. Blue cells have low importance, while red cells indicate that they constitute features of higher significance. We can see that the order in which the cells are ranked differs for each explainable method.

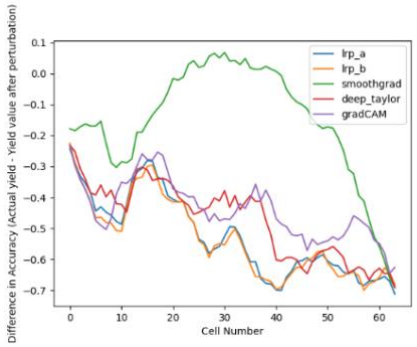

Figure 4: Perturbation plot – the difference in accuracy from the True value for each iteration

Figure 4 shows the perturbation plot where the x-axis represents the cell number (64, in this case), and the y-axis depicts the difference in accuracy between the perturbated patch and the actual value for each iteration. With each iteration, a cell is perturbated based on the ranking order, and the difference in accuracy is plotted. The lower and steeper the curve for each iteration, the better the explainable method since it indicates that the important regions are being perturbated in the correct order. Figure 4 shows that SmoothGrad is not a suitable explainable method for the crop yield model because of exceptionally high values in the y-axis. This means that the perturbated cells are not significant enough to result in a continuous drop in the y-axis values. LRPa and LRPb have the steepest decline in accuracy, signifying that their saliency maps are better at explaining the spatial information. GradCAM and Deep Taylor methods also have a steep decline in their curves.

Figure 5 depicts the Soybean fields for the patch on the left and the spatial importance of Soybeans given by each XAI

method on the right. The line plot shows that LRPb has the highest importance, followed by LRPa and Deep Taylor, while SmoothGrad has the lowest importance.

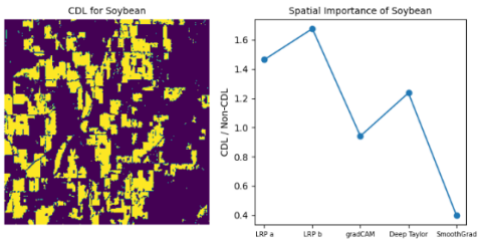

Figure 5: CDL for Soybean (left), Spatial Importance of Soybean by each eXplainable method (right)

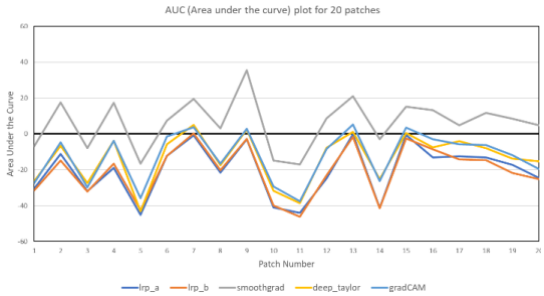

Figure 6: Area under the curve (AUC) plot for perturbation analysis

Figure 6 displays the plot of AUC (Area Under the Curve) for twenty patches from the test set. The higher the area under the curve, the lower the eXplainable method's quality. From the AUC plot, we can conclude that SmoothGrad (Grey) generally has a higher AUC compared to other XAI methods. The other explainable methods have lower AUC values, implying that they provide better levels of accurate information, with LRP being the best explainable method for these 20 patches.

4. DISCUSSION

From the results, it can be seen that running a perturbation analysis helps identify which explainable method is most reliable. In a former study, a perturbation analysis on explainable methods was performed for CNN to identify the best method for estimating fetal head circumference [15]. Their findings showed that LRP and $Input \times gradient$ were highly sensitive to perturbation and had a higher error rate [15]. Similarly, Kakogeorgiou and Karantzalos [16] assessed several explainable methods for multi-label classification in Earth Observation. They compared the saliency map methods using a sensitivity analysis and AUC-MoRF (Area Under the Curve Most Relevant First) [16]. They discovered that Occlusion, LIME (Local Interpretable Model-agnostic Explanations), and gradCAM were the most suitable explainable methods [16].

In our case, a regressor CNN model to estimate the Soybean yield is trained. From the results of the perturbation analysis, LRP gives the best results, while SmoothGrad performs poorly. The reason why LRP has better results might be due to the process in which each layer in the model has relevance. This is different from the other explainable methods like Deep Taylor and gradCAM. Though Deep Taylor provides an accumulated pixel score by propagating through the layers, LRP retains the importance level provided by each layer, which plays a role in getting the final score for the pixel. On the other hand, gradCAM takes the gradient of the weighted average of the feature maps, which makes it more suitable for visual interpretation. This also differs from LRP, where individual pixels are retained and do not lose their value by calculating the weighted average. SmoothGrad perturbates the input image multiple times and takes the average of the gradients from the perturbated images. This also differs from LRP, where random noise is not added to test the model's decision-making process. The additional assessment of the saliency maps by focusing on the Soybean field also gives similar results, where LRP has the highest importance, followed by Deep Taylor, gradCAM and SmoothGrad. This assessment assumes that the model estimates the yield by focusing on the pixels cultivated by Soybeans. This method can only be implemented based on this assumption, whereas the perturbation analysis is a generic method not restricted to this assumption.

5. CONCLUSION

Different XAI methods were compared by performing a perturbation analysis. In conclusion, the utilization of perturbation analysis proves instrumental in determining the reliability and suitability of various explainable methods. LRP and Deep Taylor were identified as the better explainable methods, with LRP giving the lowest AUC score. This was further supported by comparing the explainable methods with respect to the importance placed on the Soybean fields.

This study is significant in determining the ideal XAI method for crop yield estimation. Selecting the optimal XAI method will pave the way for future research, where complex models can be built and trained for multiple purposes. The development of a multi-model CNN that performs both crop classification and yield estimation and the implementation of a 3D CNN model to predict and forecast yield are some potential avenues for further research. The dependency and reliability of such complex models can be ascertained only with an optimal XAI method.

6. REFERENCES

[1] A. J. Challinor, J. Watson, D. B. Lobell, S. M. Howden, D. R. Smith, and N. Chhetri, "A meta-analysis of crop yield under climate change and adaptation," *Nat Clim Chang*, vol. 4, no. 4, pp. 287–291, Dec. 2014, doi: 10.1038/nclimate2153.

[2] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput Electron Agric*, vol. 177, p. 105709, Dec. 2020, doi: 10.1016/J.COMPAG.2020.105709.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5. Nature Research, pp. 206–215, May 01, 2019. doi: 10.1038/s42256-019-0048-x.

[5] A. Wolanin *et al.*, "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt," *Environmental Research Letters*, vol. 15, no. 2, p. 24019, Dec. 2020, doi: 10.1088/1748-9326/ab68ac.

[6] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (In)fidelity and Sensitivity for Explanations," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1901.09392

[7] USDA and N. A. S. Service, "Crop Production 2021 Summary." 2022. [Online]. Available: https://usda.library.cornell.edu/concern/publications/k3569432s

[8] G. Alicioglu and B. Sun, "A survey of visual analytics for Explainable Artificial Intelligence methods," *Computers and Graphics (Pergamon)*, vol. 102, pp. 502–520, Feb. 2022, doi: 10.1016/j.cag.2021.09.002.

[9] C. Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[10] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1706.03825

[11] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit*, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[12] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One*, vol. 10, no. 7, p. e0130140, Dec. 2015, doi: 10.1371/journal.pone.0130140.

[13] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-Wise Relevance Propagation: An Overview," 2019, pp. 193–209. doi: 10.1007/978-3-030-28954-6_10.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[15] J. Zhang, C. Petitjean, F. Yger, and S. Ainouz, "Explainability for Regression CNN in Fetal Head Circumference Estimation from Ultrasound Images," 2020, pp. 73–82. doi: 10.1007/978-3-030-61166-8_8.

[16] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, Dec. 2021, doi: 10.1016/j.jag.2021.102520.