Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP

Anastasios Temenos, *Graduate Student Member, IEEE*, Nikos Temenos, Maria Kaselimi, Anastasios Doulamis, and Nikolaos Doulamis

***Abstract*— An interpretable deep learning framework for land use and land cover (LULC) classification in remote sensing using Shapley additive explanations (SHAPs) is introduced. It utilizes a compact convolutional neural network (CNN) model for the classification of satellite images and then feeds the results to a SHAP deep explainer so as to strengthen the classification results. The proposed framework is applied to Sentinel-2 satellite images containing 27 000 images of pixel size 64 × 64 and operates on three-band combinations, reducing the model's input data by 77% considering that 13 channels are available, while at the same time investigating on how different spectrum bands affect predictions on the dataset's classes. Experimental results on the EuroSAT dataset demonstrate the CNN's accurate classification with an overall accuracy of 94.72%, whereas the classification accuracy on three-band combinations on each of the dataset's classes highlights its improvement when compared to standard approaches with larger number of trainable parameters. The SHAP explainable results of the proposed framework shield the network's predictions by showing correlation values that are relevant to the predicted class, thereby improving the classifications occurring in urban and rural areas with different land uses in the same scene.**

***Index Terms*— Convolutional neural network (CNN), EuroSAT, explainable AI (XAI), land cover, land use, remote sensing, Shapley additive explanation (SHAP).**

I. Introduction

KNOWLEDGE of land use and land cover (LULC) is important for the conceptual design of infrastructure projects in urban and rural areas [1]. The acquisition of such knowledge can be difficult, due to the complexity of urban/rural areas; in remote sensing and specifically in high-resolution Sentinel-2 satellite images, one pixel corresponds to 10 m on ground, meaning that a very small image, e.g., of size 64 × 64, covers a huge area, which is approximately 42 km^2. Therefore, the ground sampling distance in such images may contain many land uses, for instance, crops with roads and factories. Traditional methods, such as land surveying, provide accurate results, but are both time-consuming and cost-dependent. On the other hand, with Earth observation (EO) data, the task of classifying LULC is accelerated [2], as wide areas of interest are investigated and analyzed with a "birds-eye-view."

Geographic information systems (GISs) provide robust solutions for annotating fast and accurate LULC from EO data [3]. To improve their annotation, deep neural networks (DNNs) with an emphasis on convolutional neural networks (CNNs) for image classification are considered [4]. They are compelling for object detection in remote sensing data, covering several applications, including building extraction [5], deforestation [6], land cover change [7], and others.

In remote sensing, data can be very complex, as different objects belonging to the same category appear in the same scene [8], for instance, permanent crops, herbaceous vegetation, and forest areas, all belong to the vegetation category. DNNs, on the other hand, classify the output image, without further interpreting the results corresponding to a scene [9], [10]. Therefore, the establishment of techniques for black-box procedures to be more transparent and understandable has critical importance in remote sensing.

Explainable AI (XAI) is a technique for interpreting machine learning algorithms and DNNs models, making them more understandable to humans [10], [11]. Focusing on remote sensing data, XAI methods on images, DeepLIFT [12] and Grad-CAM [13], as well as on both images and features, local interpretable model-agnostic explanation (LIME) [14], have been used in the literature to provide insight into how a model is making decisions about LULC classification. In detail, in [9], a CNN is applied to SEN12MS [15] and to BigEarthNet [16] datasets along with selected XAI methods, such as LIME [14], DeepLIFT [12], Grad-CAM [13], Guided Grad-CAM [13], and others in [9] to show the correlations among the classes. In [17], a CNN is applied to EuroSAT [18], while LIME [14] is used to extract the correlations. However, the experimental results in [17] are limited to red green blue (RGB), and in addition, the CNN is trained using these channels only.

All the above methods discussed are constrained to local explanations, meaning that they may not be able to correctly capture information existing in the whole dataset used. Furthermore, the explanations in some cases are limited to specific

Manuscript received 16 November 2022; revised 30 January 2023 and 22 February 2023; accepted 22 February 2023. Date of publication 2 March 2023; date of current version 20 March 2023. This work was supported by the European Union Funded Project "Improved Resilience and Sustainable Reconstruction of Cultural Heritage Areas to cope with Climate Change and Other Hazards based on Innovative Algorithms and Modeling Tools" under the Horizon 2020 Program H2020-EU.3.5.6., under Grant 872931. *(Corresponding author: Anastasios Temenos.)*

The authors are with the School of Rural Surveying and Geoinformatics Engineering, National Technical University of Athens, 15780 Athens, Greece (e-mail: tasostemenos@mail.ntua.gr; ntemenos@gmail.com; mkaselimi@mail.ntua.gr; adoulam@cs.ntua.gr; ndoulam@cs.ntua.gr).

Digital Object Identifier 10.1109/LGRS.2023.3251652

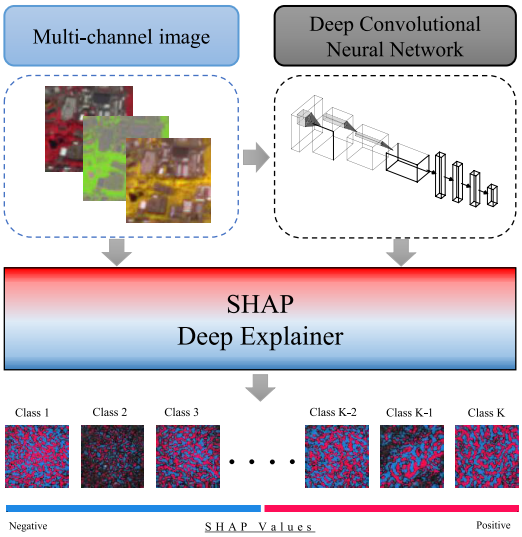

Fig. 1. Proposed explainable deep learning framework. Multichannel images and a trained deep CNN architecture are fed to an SHAP deep explainer.

channels, meaning that important information from other ones and their impact on the explanations are not investigated.

This work proposes an XAI framework for remote sensing data utilizing Shapley additive explanations (SHAPs) [19]. Compared with the existing XAI approaches, the use of SHAP enables both *local* and *global* explanations, allowing for information between different spectral bands in a dataset to contribute toward the explanations. Compared with the existing approaches being limited to RGB channels, the proposed approach considers different band combinations for the classification and the explanation of their results so as to highlight the interference of information from different wavelengths of the spectrum affecting the classes. This improves both the classification accuracy of each individual class and the explanations' interpretability, leading to a better estimation of the channels' contribution to the final prediction.

II. Proposed Deep SHAP Framework

The high-level model capturing the operation of the proposed deep SHAP framework is shown in Fig. 1. Initially, a deep CNN is trained using a dataset containing LULC images, and once trained, it classifies any multichannel image in one of its classes. The classification's result along with the image are then fed to the SHAP explainer, which outputs the pixels' positive or negative correlation for each one of the K existing classes. The positive or negative correlation corresponds to what extend each feature contributes to each class, allowing for a better interpretation and understanding of the following: 1) many different objects existing within an input image and 2) which image spectral band combinations responded better in the LULC classification. In Sections II-A and II-B, the CNN architecture and the deep SHAP model used are explained.

A. CNN Architecture

The CNN architecture utilized by the proposed deep SHAP framework is illustrated in Fig. 2. The input image is of size

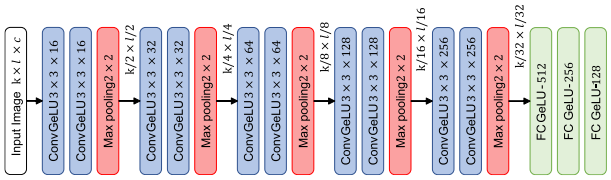

Fig. 2. CNN architecture used in the framework of Fig. 1.

$k \times l \times c$, where k and l are the number of rows and columns, respectively, and c is the number of the image's channels. It consists of five convolution–convolution–max-pooling layers connected sequentially, where each one downsamples the input image by increasing powers of 2, having the number of filters doubled in each layer. The convolution–convolution–max-pooling layers are then followed by three fully connected (FC) layers, where each one has 512, 256, 128 neurons, respectively. The activation function used here in all layers is the Gaussian error linear units (GeLUs) [20], defined as follows:

$$\text{GeLU}(y) = yP(Y \leq y) = y\Phi(y) \tag{1}$$

where $\Phi(y)$ is the standard normal cumulative distribution function, $Y \sim \mathcal{N}(0, 1)$, and y is the input to the activation function. The difference of GeLU over the rectifier linear unit (ReLU) originates from the stochasticity of the former; GeLU samples from the standard normal distribution according to (1), thus introducing a natural dropout regularization, which is not feasible with ReLU [20].

B. Deep SHAPs

The SHAP is a method for interpreting machine learning models, introduced in [19]. It maps inputs, x', to the original ones, x, through a mapping function $x = h_x(x')$ so as to explain a prediction $f(x)$. The simplified inputs allow for the interpretable model to ensure that for any feature $z' \in \mathbb{R}$ and whenever $z' \approx x'$, then $g(z') \approx f(h_x(z'))$.

The SHAP's additive feature attribution method is based on a linear function of binary values as follows:

$$g(z') = \phi_0 + \sum_{n=1}^{N} \phi_n z'_n \tag{2}$$

where N is the number of input features, $n = 1, 2, \ldots, N$ is the feature index, the values of $\phi_n \in \mathbb{R}$ are the model's coefficients and the values of $z'_n \in \{0, 1\}^N$ denote the observation of a feature. Note that each z'_n refers to a feature of z'.

To explain the derivation of the coefficients ϕ_n, we proceed with some definitions. Let N_S be a feature subset, such that $N_S \subseteq \mathcal{N}$, where $\mathcal{N}$ is the set of all features with cardinality $|\mathcal{N}|$. Assuming that x_{N_S} represents the values of the input features existing in N_S and n is a feature, the model $f_x(N_S)$ used for the calculation of the SHAP values is defined as $f_x(N_S) = \mathbb{E}[f(x)|x_{N_S}]$. The model is trained two times; one including the feature n, i.e., $f_{N_S \cup \{n\}}(x_{N_S \cup \{n\}})$, and one excluding it, i.e., $f_{N_S}(x_{N_S})$. Predictions are then derived from their comparison $f_{N_S \cup \{n\}}(x_{N_S \cup \{n\}}) - f_{N_S}(x_{N_S})$, while the procedure is repeated for every possible subset, such that $N_S \subseteq \mathcal{N}$. Combining the above equations, the coefficients ϕ_n from (2)

are derived as follows:

$$\phi_n = \sum_{N_S \subseteq \mathcal{N}\setminus\{n\}} \frac{|N_S|!(|N| - |N_S| - 1)!}{|N|!} \times \left(f_{N_S\cup\{n\}}(x_{N_S\cup\{n\}}) - f_{N_S}(x_{N_S})\right). \quad (3)$$

Note that unique solutions within the class of additive feature attribution methods exist if and only if the following three key properties are satisfied [19]: 1) local accuracy; 2) missingness; and 3) consistency.

The deep SHAP framework of Fig. 1 combines the Shapley values calculated using (2), (3), and the DeepLIFT method. DeepLIFT is a compositional approximation of the Shapley values under the assumptions that the following hold: 1) the deep model is linear and 2) the input features are uncorrelated to one another. It is an additive feature attribution method that satisfies local accuracy and missingness, two of the three key properties for additive feature importance. Therefore, with the inclusion of the Shapley values, the consistency of the model is achieved [19].

III. LULC on Three-Band Combinations: Classification and XAI Results

A. Experimental Setup

The performance of the proposed framework is evaluated using the EuroSAT dataset proposed in [18]. EuroSAT contains LULC images taken from the Sentinel-2 satellite, covering 13 spectral bands and consisting of ten classes in total with 27 000 labeled and geo-referenced images. Out of the 13 spectral bands, we consider only the use of red, green, near infrared (NIR-Band 8), short-wave infrared (SWIR-Band 11), and the remote sensing normalized difference indexes stemming from them, including vegetation index (NDVI), buildup index (NDBI), and water index (NDWI).

In the experiments, the following different three-band combinations of the above selected bands are used, which are the following: 1) SWIR-NIR-RED; 2) NIR-RED-GREEN; and 3) NDBI-NDVI-NDWI. These combinations capture information existing in wavelengths that are able to identify vegetation, water bodies, soil, and man-made constructions, as their percentage reflectance is higher compared with other bands [18].

All the experiments are conducted using Google Colab Pro, Python 3, and TensorFlow. With respect to the training phase, the dataset is split into 70/10/20 train/validation/test sets, while the network is trained for approximately 70 epochs, using an early stopping criterion with patience of ten epochs, monitored using the validation loss. The batch size used is 64, the learning rate is 10^{-3} and the seed value is 42. Moreover, the selected optimizer is layer-wise adaptive moments optimizer for batch training (LAMB) proposed in [27]. Compared with adaptive moment estimation (ADAM), it applies adaptive elementwise updating and layerwise learning rates on large batches of input data, hence speeding-up the training when large datasets are used.

B. Experimental Results

To compare the performance of the proposed framework, we consider several deep NN architectures applied on the EuroSAT dataset, including the following: 1) a Shallow CNN [21]; 2) GoogleNet [22]; 3) DenseNet121 [23]; 4) Inception V3 [24]; 5) ResNet50 [25]; 6) ResNet101 [25]; 7) VGG16 [26]; and 8) GeoSystemNet [21]. Note that in [21], a fusion of the initial EuroSAT dataset along with different-scaled imaged derived from MapBox application programming interface (API) [21] is used. In the comparisons, we use the following standard classification metrics:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1 Score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \quad (4)$$

where TP, TN, FP, and FN denote, respectively, the true positive, true negative, false positive, and false negative values. The results of the precision, recall, and $F1$ score are cited in Table I, whereas the accuracy is cited in Table II. Note that the accuracy reported in Table II is calculated using all 13 bands so as to have a fair comparison between the works.

It should be noted that among the classification metrics in (4), precision and recall are the most important ones in the LULC classification; precision reflects a model's ability in identifying correctly a single object among many, thereby strengthening its reliability, whereas recall measures a model's ability in identifying correctly a single object regardless of the rest, thereby strengthening its effectiveness. On the other hand, accuracy measures the model's overall performance without considering that many classes have similar spectral signatures; for instance, permanent crop and forests are different, but belong to the vegetation category.

According to Table I, the proposed framework yields the highest precision value in all classes, except from pasture in which GeoSystemNet is better. The recall values follow similar behavior to those of the precision, with the main difference being the forest class in which GeoSystemNet has lower value. With respect to the $F1$ score, it is observed that our framework results in the highest values except from the herbaceous and permanent crop, which is expected, since it is the harmonic mean of the precision and recall, hence affected by them. Yet, considering all the 13 bands, the classification accuracy can be improved. From the results cited in Table I, one can conclude that reducing the number of channel bands improves the classification accuracy of each class separately.

From Table I, it can be seen that the use of three channel combinations greatly improves the classification of each class separately. However, the classification accuracy reported in Table II for the proposed framework is reduced compared with the other models, which is reasonable given the number of its trainable parameters. It should be mentioned though that the SHAP by itself is the computationally expensive XAI technique, as it calculates the Shapley values for various features in a prediction instance [28]. Therefore, despite the smaller number of trainable parameters of the CNN used by the proposed framework, it results in faster (in time) extraction of the explanations, as less features are fed to the SHAP explainer.

TABLE I

MODEL PERFORMANCE IN THE CLASSIFICATION METRICS PRECISION, RECALL, AND $F1$ SCORE EVALUATED USING THE EUROSAT DATASET

Metric	Model	Channel Bands	Annual crop	Forest	Herbaceous	Highway	Industrial	Pasture	Permanent crop	Residential	River	SeaLake
Precision	Shallow CNN [21]	13	85.40	90.28	54.99	68.02	82.63	73.04	67.45	90.46	67.15	93.36
	GoogleNet [22]	13	85.90	81.17	57.22	56.15	91.03	81.75	69.50	83.93	57.12	96.91
	DenseNet121 [23]	13	71.12	75.39	52.39	51.54	94.52	63.11	57.23	79.74	78.34	91.82
	InceptionV3 [24]	13	86.80	80.27	56.32	57.15	91.03	82.71	68.51	84.96	56.18	97.90
	ResNet50 [25]	13	82.87	92.79	67.11	53.29	71.54	81.98	65.25	85.88	67.51	93.41
	ResNet101 [25]	13	83.61	91.62	72.54	57.70	73.07	61.95	59.18	87.08	53.55	92.04
	VGG16 [26]	13	79.03	80.69	50.27	66.38	86.64	58.16	58.03	93.77	66.82	89.74
	GeoSystemNet [21]	13	93.30	90.40	89.29	70.38	91.74	**89.85**	84.47	92.52	67.30	98.32
	Proposed	SWIR - NIR - RED	91.06	96.75	86.89	93.78	**96.72**	84.34	87.23	98.46	96.44	98.07
	Proposed	NIR - RED - GREEN	96.20	**98.94**	80.66	92.05	86.54	86.62	89.95	97.87	**98.23**	95.73
	Proposed	NDBI - NDVI - NDWI	**96.34**	98.38	81.80	91.40	86.96	87.82	87.97	**99.60**	95.63	99.04
	Proposed	13	92.94	94.24	**94.58**	**96.08**	92.91	81.64	**94.05**	97.40	93.65	**99.82**
Recall	Shallow CNN [21]	13	83.41	93.44	65.90	50.50	91.24	68.78	63.26	96.03	65.73	90.06
	GoogleNet [22]	13	76.45	98.03	71.80	46.21	70.04	55.20	55.68	95.06	79.91	82.41
	DenseNet121 [23]	13	85.67	92.40	64.38	53.81	40.39	74.97	53.33	95.66	53.88	72.18
	InceptionV3 [24]	13	77.58	97.54	72.94	44.00	71.30	56.84	56.76	96.47	78.79	83.40
	ResNet50 [25]	13	83.01	90.01	48.88	62.61	93.34	60.27	73.63	84.03	66.21	94.07
	ResNet101 [25]	13	81.32	93.88	45.90	53.24	86.07	68.28	44.73	94.62	59.21	85.18
	VGG16 [26]	13	80.45	92.33	56.33	49.97	83.01	63.84	54.09	92.16	66.91	86.66
	GeoSystemNet [21]	13	89.21	98.18	85.44	58.05	90.59	84.29	87.84	94.82	83.43	92.62
	Proposed	SWIR - NIR - RED	**93.96**	97.42	85.79	93.98	92.29	92.11	81.09	**98.29**	**97.87**	97.88
	Proposed	NIR - RED - GREEN	89.09	96.21	87.77	**94.62**	96.46	**93.68**	77.46	94.20	94.68	99.42
	Proposed	NDBI - NDVI - NDWI	92.62	94.15	**91.37**	93.76	97.29	89.21	83.90	85.67	97.66	99.61
	Proposed	13	91.88	**99.11**	85.90	90.31	**98.61**	93.56	**88.25**	95.25	97.52	**99.82**
F1 Score	Shallow CNN [21]	13	84.39	91.84	59.95	57.97	86.72	70.85	65.29	93.16	66.43	91.68
	GoogleNet [22]	13	80.39	89.25	62.52	48.69	82.71	66.35	64.18	92.01	64.34	89.41
	DenseNet121 [23]	13	77.72	83.03	57.77	52.65	57.29	68.53	55.21	86.98	63.85	80.82
	InceptionV3 [24]	13	81.93	88.07	63.56	49.72	79.97	67.38	62.08	90.35	65.59	90.07
	ResNet50 [25]	13	82.94	91.38	56.56	57.58	81.00	69.47	69.19	84.94	66.85	93.74
	ResNet101 [25]	13	82.44	92.24	56.22	55.38	79.04	64.96	50.95	88.95	56.24	88.47
	VGG16 [26]	13	79.73	86.12	53.12	57.02	84.78	60.87	55.99	92.96	66.87	88.18
	GeoSystemNet [21]	13	91.21	94.13	87.32	63.62	91.16	86.98	86.12	93.65	74.50	95.38
	Proposed	SWIR - NIR - RED	92.49	97.08	86.33	**93.88**	94.46	88.05	84.05	**98.38**	**97.15**	97.97
	Proposed	NIR - RED - GREEN	92.51	**97.56**	84.07	93.32	91.23	**90.01**	83.24	96.00	96.42	97.54
	Proposed	NDBI - NDVI - NDWI	**94.44**	96.22	86.32	92.57	91.84	88.51	85.89	92.11	96.63	99.33
	Proposed	13	92.41	96.61	**90.03**	93.11	**95.68**	87.20	**91.05**	96.31	95.55	**99.81**

TABLE II

MODEL CLASSIFICATION ACCURACY USING THE EUROSAT DATASET

Model	Classification Accuracy (%)	Trainable Parameters
Shallow CNN [21]	87.96	422,378
GoogleNet [22]	96.02	6,797,700
DenseNet121 [23]	96.64	8,062,504
Inception V3 [24]	96.86	23,851,784
ResNet50 [25]	96.43	25,636,712
ResNet101 [25]	95.57	44,707,176
VGG16 [26]	96.65	138,357,544
GeoSystemNet [21]	94.65	1,324,526
Proposed	94.72	1,869,082

C. XAI Results

According to Fig. 1, the CNN model and an image to be classified are fed to the SHAP deep explainer for the derivation of the SHAP values. Once derived, the SHAP image plot tool is used to visualize the positive (red) and negative (blue) correlations in each pixel of each of its classes. An example case of annual crop, river, and highway using the three different band combinations is illustrated in Fig. 3.

In the NDBI-NDVI-NDWI case of Fig. 3, as expected, positive correlations exist in the annual crop class given the CNN's correct classification. In the second case, SWIR-NIR-RED, positive correlations exist in the river class, while negative correlations are denser in the annual crop class, implying that it is less likely for annual crop to exist within the image. Of important interest is the final case, NIR-RED-GREEN, in which the positive correlations are intense in the area where the highway is present. Apart from the local explanations

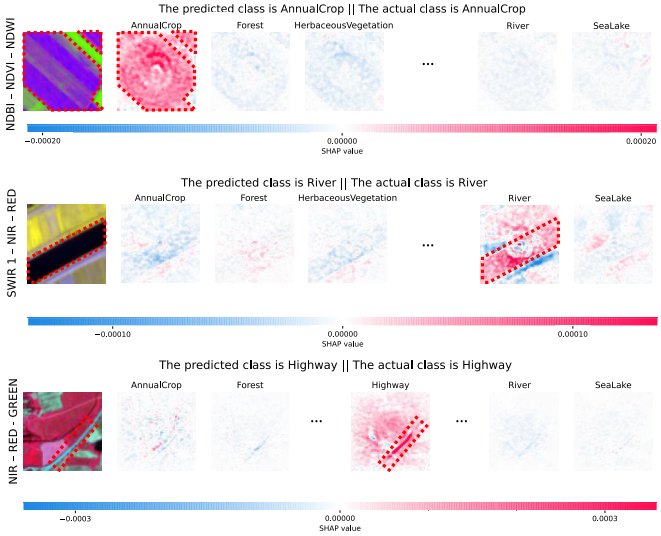

Fig. 3. SHAP image plot depicting the impact of each pixel on the models predictions. SHAP image plots and classification from top to bottom. First row: Annual crop using NDBI-NDVI-NDWI, Second row: river using SWIR-NIR-RED, and Third row: highway using NIR-RED-GREEN. The red pixels indicate strong correlation among the predicted classes, whereas the blue ones indicate weak correlation.

shown in Fig. 3, the proposed framework can be used to extract global explanations, as shown in Fig. 4. It can be seen that the selected bands red, green NIR, and SWIR 1 result in the highest average SHAP values for almost all classes, meaning that they are critical in the explainability of the

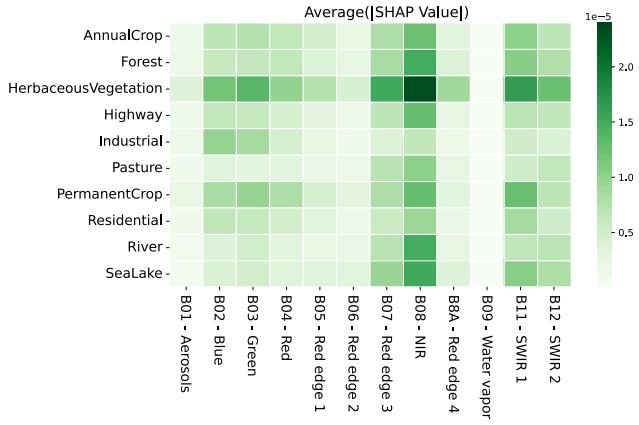

Fig. 4. SHAP global explanations on the spectral bands of Sentinel-2.

classification results. Note that the cirrus band (B10) is not included in Fig. 4, as it does not contain surface reflectance information [29].

IV. Conclusion

In this work, we presented a deep XAI framework based on SHAP applied on satellite images of Sentinel-2. Experimental results on different spectral band combinations of the EuroSAT dataset demonstrated that the proposed framework improves the classification accuracy of each individual class among the existing ones, also shown with comparisons to the existing CNN methods from the literature. The local explanation results verified that the model predictions were correctly derived, highlighting for each class which pixels had positive correlation, whereas the global explanation results showed the contribution of each individual band toward the explanations. Therefore, using the proposed framework, the end user can classify satellite images in an automatic and reliable way, as the introduced qualitative visual pattern assists the quantitative metric of the classification accuracy. This improves the concept of multilabel LULC classification, especially when multiple objects exist in the same scene.

References

[1] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[2] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[3] G. Kanagasundaram, K. Dissanayake, and C. Samarasuriya, "Remote sensing and GIS approach to monitor the land-use and land-cover change in kaduwela metropolitan area," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, Jul. 2022, pp. 1–6.

[4] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.

[5] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[6] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 2, 2022, doi: 10.1109/TNNLS.2022.3144791.

[7] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412712.

[8] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020.

[9] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102520.

[10] A. Temenos, I. N. Tzortzis, M. Kaselimi, I. Rallis, A. Doulamis, and N. Doulamis, "Novel insights in spatial epidemiology utilizing explainable AI (XAI) and remote sensing," *Remote Sens.*, vol. 14, no. 13, p. 3074, Jun. 2022.

[11] A. Temenos, M. Kaselimi, I. Tzortzis, I. Rallis, A. Doulamis, and N. Doulamis, "Spatio-temporal interpretation of the COVID-19 risk factors using explainable ai," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 7705–7708.

[12] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[13] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[15] M. Schmitt, L. Haydn Hughes, C. Qiu, and X. Xiang Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," 2019, *arXiv:1906.07789*.

[16] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5901–5904.

[17] M. Verma, N. Gupta, B. Tolani, and R. Kaushal, "Explainable custom CNN architecture for land use classification using satellite images," in *Proc. 6th Int. Conf. Image Inf. Process. (ICIIP)*, Nov. 2021, pp. 304–309.

[18] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[20] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[21] S. A. Yamashkin, A. A. Yamashkin, V. V. Zanozin, M. M. Radovanovic, and A. N. Barmin, "Improving the efficiency of deep learning methods in remote sensing data analysis: Geosystem approach," *IEEE Access*, vol. 8, pp. 179516–179529, 2020.

[22] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[27] Y. You et al., "Large batch optimization for deep learning: Training bert in 76 minutes," 2019, *arXiv:1904.00962*.

[28] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 740–770, Sep. 2021.

[29] Y. Li, J. Chen, Q. Ma, H. K. Zhang, and J. Liu, "Evaluation of Sentinel-2A surface reflectance derived using Sen2Cor in North America," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 6, pp. 1997–2021, Jun. 2018.