LIME EXPLAINABILITY REVEALS IMPLICIT ASSOCIATIONS IN DEEP CONVOLUTIONAL ARCHITECTURES

Bogdan Diaconu, *University "Constantin Brâncuşi" of Tg-Jiu, ROMANIA*
Lucica Anghelescu, *University "Constantin Brâncuşi" of Tg-Jiu, ROMANIA*
Mihai Cruceru, *University "Constantin Brâncuşi" of Tg-Jiu, ROMANIA*
Bâca Gheorghe, *University "Constantin Brâncuşi" of Tg-Jiu, ROMANIA(student)*

ABSTRACT: Deep learning models for image classification trained on low-quality data (insufficient dataset volume, unbalanced datasets, ambiguous im- ages) may exhibit surprisingly high values of the performance metrics. However, explainability may be an issue for such models since their classi- fication decisions may be completely different from human reasoning. In this paper, four architectures were considered as follows: a custom model with two convolutional layers and two pooling layers and three models based on EfficientNet B0, Inception V3 and ResNet V2.50. The models were trained on a publicly available (Kaggle) image dataset "Solar Panel Images Clean and Faulty Images" with per-class accuracy values exceed-ing 0.9. The explainability algorithm LIME was applied for predictions of all models and for each class. For some classes, the decisions of the models (and especially the custom model) cannot be explained in terms of human reasoning. Explainability was assessed for predictions made by Inception V3 architecture using the LIME algorithm with the same parameters. In case of two random images (from the ImageNet classes 486 -" cello" and 657 -" missile") the classification decision of Inception V3 is grounded on features that make perfect sense for human reasoning. How- ever, for ambiguous images, the inference of Inception V3 seems to rely on underlying associations - groups of object occurring frequently together with the target object in the same image.

Key-Words: models, algorithm, function, images.

1.INTRODUCTION

Although machine learning and deep learning models (as universal function approximations [1]) outperform many other algorithms in classification and regression tasks, in general, they act like black boxes, in the sense that it may not always be possible to understand, in a quantifiable and reproducible way, the prediction of such a model. Explainability is considered one of the four ethical principles for trustworthy AI [2]. Longo et al. [2] reviewed the advances in Explainable AI (XAI) and applications in the real world, discussing the current challenges.

The necessity for explainability and interoperability of AI systems results has been understood with the emergence and deployment of such systems, especially in decision problems [3]. Unlike, for example, decision trees, which are algorithms with explainable structures, deep neural networks with their various architectures (Convolutional Neural Networks - CNN, Recurrent Neural Networks - RNN, Long Short-Term Memory - LSTM, and several more) pro- duce outputs that cannot be explained, i.e. their internal inference processes are neither known to the observer nor interpretable by humans, Guidotti et al. [4]. This work will discuss the explainability of three CNN models for im- age classification, trained on a publicly available, low-volume, and imbalanced image dataset. The LIME (Local Interpretable Model-agnostic Explanation) algorithm will be applied to the trained models to gain insight into the pre- diction mechanisms. Explainability is a very useful instrument that can shed light on the black-box model mechanisms for some image classification

models. In particular, some CNN-based models, trained on low-volume (and sometimes unbalanced) datasets, display unusually high-performance metrics and no over- fitting during the training process. In such cases, explainability could reinforce trust in the predictions of such models, or, conversely, flag possible underlying issues with such models.

2.PREVIOUS WORKS

Soiling of photovoltaic panels (PV) causes significant degradation of the efficiency. Adekanbi et al. [5] presented a comprehensive review of soiling caused by dust on the operation of PVs. The review focused on the effect of (1) various types of dust occurring naturally and (2) environmental conditions that could favor dust accumulation. Ballestrin et al. [6] reported daily electrical losses as high as 9% between February and November 2020 for a PV system installed in Madrid. The effect of soiling. The estimation reported in [6] was based on a database of 238 measurements of average daily losses of a photovoltaic module compared to an identical one that was maintained clean. Soiling of PV panels is a complex process with a significant random component and several environmental factors influencing it in ways not fully understood. Santhakumari and Sagar [7] conducting a literature review on studies investigating the effect of a wide range of environmental factors (dust, ambient temperature, wind velocity, humidity, snowfall, hail, and sandstorms) and the typical defects these factors incur on PV panels. A quantitative estimation of the effect of soiling was presented by Bessa et al. [8], reporting losses equivalent to 3%–4% of the global energy yield (2018), with total missed revenues of at least 3 to 5 billion euros. Cordero et al. [9] conducted a study on the PV panel economic losses caused by soiling in various regions in the Atacama Desert area. Based on the annual soiling rate, the amount and frequency of the rainfall, and the cost of cleaning operations, an optimum cleaning frequency was recommended for each site. Ilse et al. [10] reviewed the mechanisms of PV panel soiling, breaking up the soiling process into several sub-processes that favor particle deposition and adhesion, as follows: dew formation, cementation, particle caking, and capillary aging. Given the impact of PV faults, various methods have been developed to identify and classify abnormal operating conditions. Two main groups of approaches were identified in the literature [11]: (i) monitoring of electrical parameters and (ii) use of thermography images and computer vision techniques. The development of artificial intelligence both in terms of fundamental research, software libraries, and hardware, deep learning algorithms for the classification of anomalies from RGB and thermographic images become increasingly accurate and efficient. Ettaleby et al. [12] proposed a hybrid model combining Support Vector Machine (SVM) and Convolutional Neural Networks (CNN) to classify PV faults based on electroluminescence images from three classes (normal, cracked, and corroded). The architecture of the model was based on VGG16 (the feature extraction component). The actual classification was implemented by means of SVM. Two datasets were used: D1 with 2624 (300x300) electroluminescence images, manually labeled by a human expert; D2 with 1028 (250x250) images. The authors reported an accuracy of 99.49%, however, some important details such as class imbalance, the volume of the training/test sets and learning curves were not presented. A comprehensive review on photo- voltaic systems fault detection was conducted by Hong et al. [13]. The review discusses a significant number of studies employing deep learning algorithms, however, only one study - Sairam et al. [14] - integrated explainability into a model that could be deployed on edge devices to assist the field personnel in understanding the PV fault causes. The structure of the model consisted of a base model built on an Irradiance-based three-diode model (an electrical model of the photovoltaic cell) and XGBoost. The explainability component was

implemented using LIME - Local Interpretable Model-agnostic Explanations.. Harikumar et al. [15] used a dataset consisting of fetal screening ultrasound images (9308 RGB and gray scale images, five classes) and a custom CNN architecture to develop a model to predict the presence of some maternal and fetal anatomical parts. LIME algorithm was used to explain the outputs of the CNN model. In essence, LIME pointed out the region of the image contributing to the highest extent to a particular predicted class. The idea put forward in this work is that an explainability study is highly required for some deep learning models, especially when trained on low-volume datasets. Although such models can exhibit high values of the performance metrics, their explainability in terms of human understanding is low.

3.MATERIALS AND METHODS

A publicly available (Kaggle) image dataset "Solar Panel Images Clean and Faulty Images" with the classes and number of instances in each class presented in Table 1 was used to train the models. The dataset consists of images scrapped from the internet with various resolutions and file formats.

A custom CNN model was considered as a baseline with the architecture presented in Figure 1. Three architectures, EfficientNet B0, ResNet v2.50 and

Clean	Dusty	Bird drop	Electrical fault	Physical fault	Snow
192	194	191	104	70	124

Table 1: Content of the dataset.

Inception v3 (for all architectures, model variant feature-vector was used). All images were resized to 224x224 and the dataset was randomly broken down into train and test subsets (80/20). The test set was used for validation during the training process.

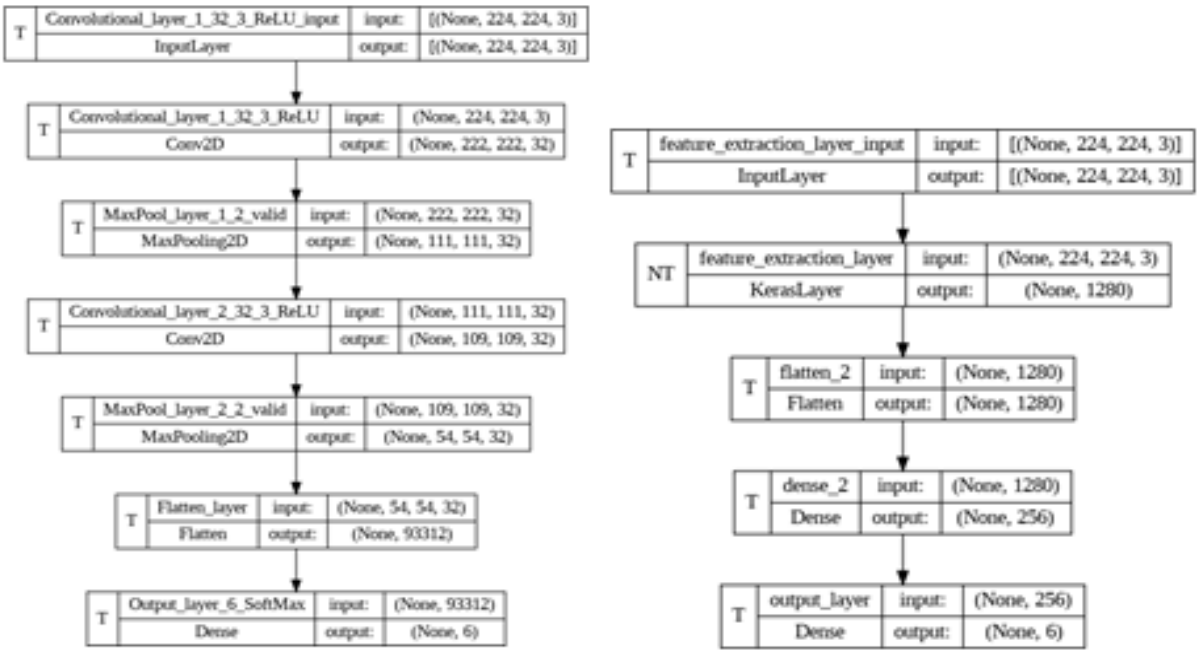

Figure 1: The custom model (left) and EfficienNet, Inception and ResNet architectures (right)

All models were compiled using Adam optimizer with the default parameters and categorical cross entropy as loss function. The training curves for the EfficienNet B0 architecture are presented in Figure 2. The confusion matrices (using the predictions on the test set) are presented in Figure 3. All models predict correctly 100% of the instances from the classes "Electrical damage" and "Snow-Covered". For the models based on Inception and ResNet models, the most frequent confusion occurs for the class "Dusty" (predicted as "Clean"). Another frequently occurring confusion is for the class "Physical-Damage", for which Inception and ResNet based models falsely predict "Clean" and "Bird- drop" classes.

The LIME algorithm will be employed to understand how the four models generate predictions. LIME has different variants depending on the data type. For image data, the first step of the algorithm is to segment the image into super pixels (defined as several adjacent pixels having RGB values close to each other). The super pixel regions can be turned on or off, setting the RGB values to a specific color [16]. LIME interpretation starts with a correct prediction of the trained model. Then a new data set is created by randomly perturbing the original image (for example, by turning on and off random superpixel regions). A local (in the sense that it is created based on the initial image considered) model is fitted taking as input the computed superpixels. Weights of the im- ages, close to the original image are applied to quantify the importance of the perturbed images. The importance of each perturbed image is determined by means of a distance metric that assesses "how far" each perturbed image is from the original image (the one for which all superpixel regions are turned on). The distance metric is calculated and assigned to each image in the perturbed dataset. The local model is then trained with the perturbed images, predictions and weights. A factor for each superpixel is calculated describing the effect of the superpixel on the right class prediction. The values of these factors are ordered to sort out the superpixel regions that contribute the most significantly to the model prediction. The approach used in this study is the following: first, the four trained models (as described in the previous section) will be used to generate a prediction. Then LIME algorithm will be used with each of the four models to generate the interpretable maps of the image. In the end, Inception V3 trained with Imagenet was be used with random images to demonstrate differences in the interpretability maps. The LIME algorithm was implemented as follows: the quickshift algorithm from the scikit-image library was used to segment the original image. The quickshift parameters kernel size, maximum (bottom right) confusion matrix. True class on the vertical axis, predicted class on the horizontal axis distance, and ratio were 4, N_{pert}=200, and 0.2 respectively. Then 200 perturbed images were generated randomly (drawing samples from a binomial distribution with a probability 0.5) by selecting superpixel patches. A set of predictions were generated using each trained model and the set of perturbed images as input. The distance values d_i between the original image and the perturbed images was determined by using the scikitlearn metric pairwise distance with the metric parameter cosine. The weight assigned to each perturbed image were calculated using a kernel function as follows:

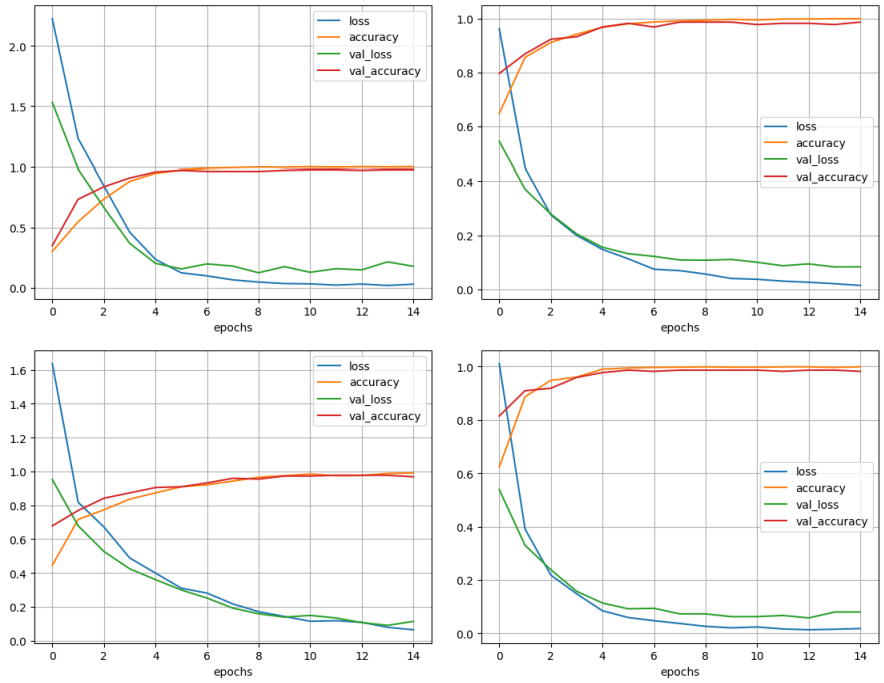

Figure 2: Custom model (top left), EfficientNet B0 (top right), Inception V3 (bottom left) and ResNet V2.50 (bottom right) training curves

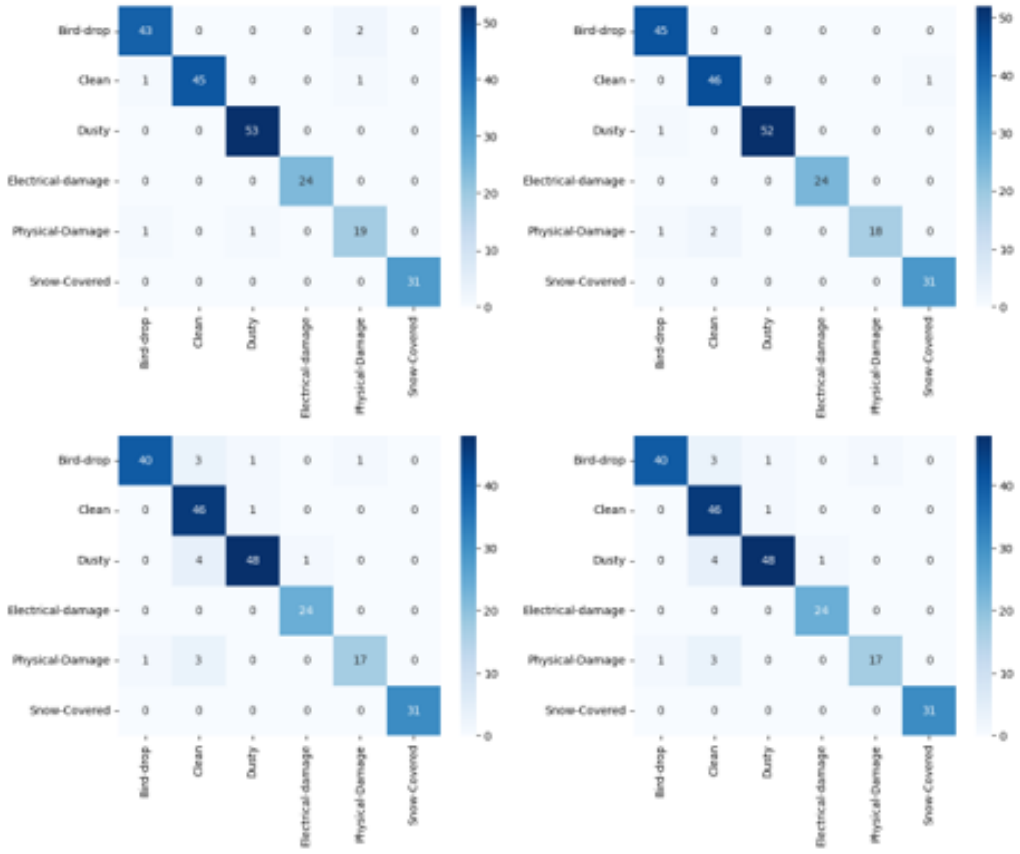

Figure 3: Custom model (top left), EfficientNet (top right), Inception (bottom left) and ResNet B0

$$w_i = \sqrt{exp - \left(\frac{d_i^2}{k^2}\right)}$$

in which k is the kernel size, for which the value 0.25 was used. A surrogate model (linear regression) was fitted using the perturbations and the corresponding predictions maximum probability (the maximum value was selected from the six-element prediction vector). The sample weight parameter of the Linear Regression model was set to w_i. The first four features will be used to create the features map.

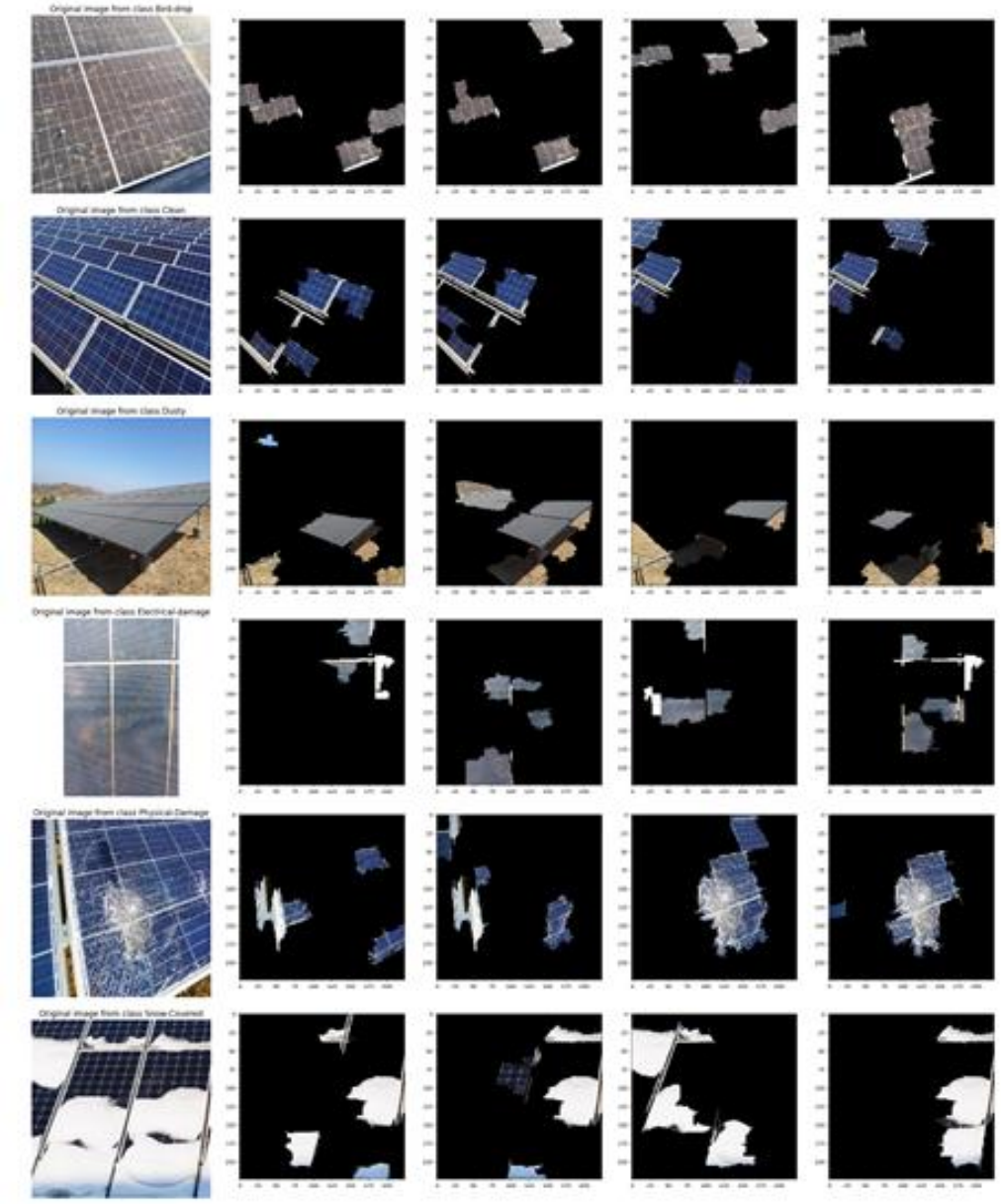

Figure 4: Interpretability maps: leftmost column: original random image from each class. Next columns: interpretability maps for the custom model, EfficientNet, Inception and ResNet-based models, respectively

4.DISCUSSION

In Figure 4 several random images from each class were selected and LIME algorithm was applied to identify the top four features that determine the models to predict a certain class. Before applying the LIME algorithm, it was ensured that each model predicts correctly the true class. For the classes "Bird-drop" and "Clean" the top four features that contribute to the model decision are similar in appearance for all models discussed in this paper. Although Inception and ResNet have a higher rate of failures for the class "Bird-drop", as shown in Figure 3, they still consider correctly the relevant features of the images. It can be argued though that the image area is almost fully occupied by the target object and the model has little room for misinterpreting the image content. This hypothesis seems to be confirmed for the class "Dusty". The image area contains other object than the target object (dust-covered PV panel), such as clear sky and dry, exposed soil patches. Although all models include the target object in the feature set used to decide the class, they also consider the exposed soil patches as elements that point to the "Dusty" class, which is rather counter intuitive for a human. However, although it is far fetched, one could argue that PV panels installed in areas with dry and exposed soil have more chances to get dusty. For the image in the class "Electrical damage", although all models predict correctly the class, only EffcientNet, Inception and ResNet consider correctly the specific feature for this class. The custom model prediction, al- though correct, has no intuitive grounds for a human. Inception and ResNet identify successfully the specific feature for the class "Physical Damage" while the custom model and EfficientNet fail to include their decision on the main feature that describes this particular class. For the class "Snow-Covered" all models identify correctly the defining features for this particular class. However, EfficienNet includes amongst its top features an image patch containing a clean area.

In order to understand the differences in terms of explainability between the four models considered in this paper (trained on a low-volume and unbalanced dataset), the Inception V3 architecture trained on ImageNet will be used to make predictions on a couple of random classes and the results of the LIME algorithm will be discussed. First, a random image from the class 657 ("missile") was collected from the public internet (search term "missile", only images). The architecture Inception V3 was used to make predictions, resulting the following classes and corresponding probabilities: missile: 0.645 projectile: 0.314 cannon: 0.00141 tank: 0.00054 The original image, the superpixel segmented and a random perturbed image are presented in Figure 5 Applying the LIME algorithm and selecting successively the first top feature only, then the first two and lastly the first three features, the feature maps presented in Figure 6 are obtained. The results from another class (486 - "cello, violoncello") are presented in Figure 7. The predicted classes for the target image are as follows: "cello": 0.871, "stage": 0.013, "drum": 0.002. The top features that determined the model prediction are presented in Figure 8

In both cases, the LIME algorithm demonstrates that Inception V3 with

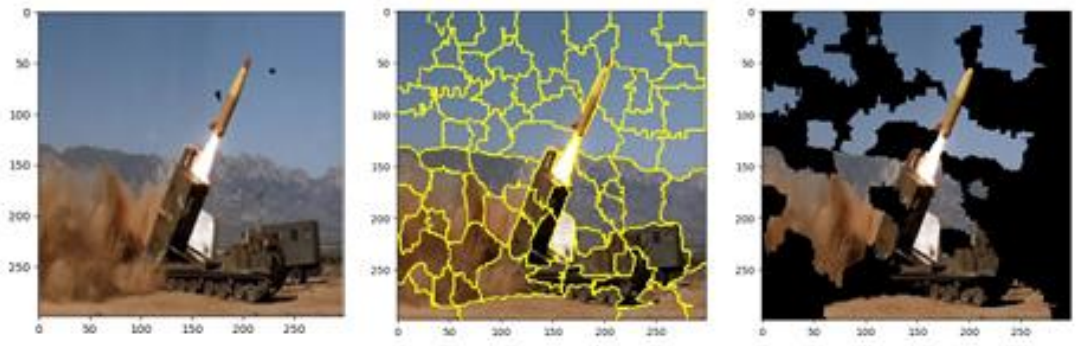

Figure 5: Target image from the class "missile", segmented image and one random perturbed image

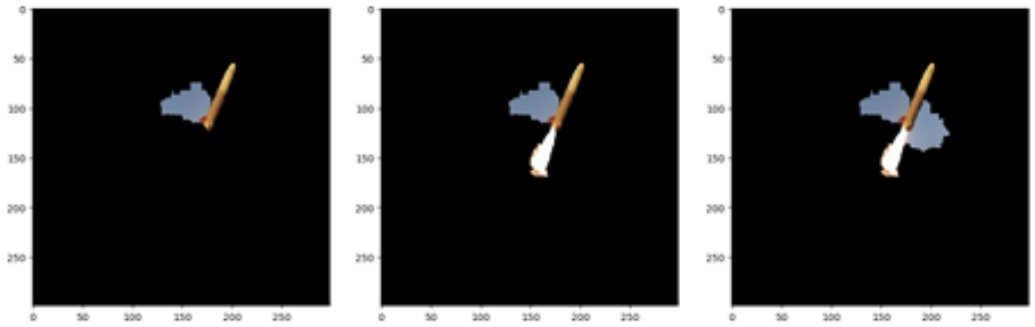

Figure 6: Target image from the class "missile", top three features used in the model decision

.

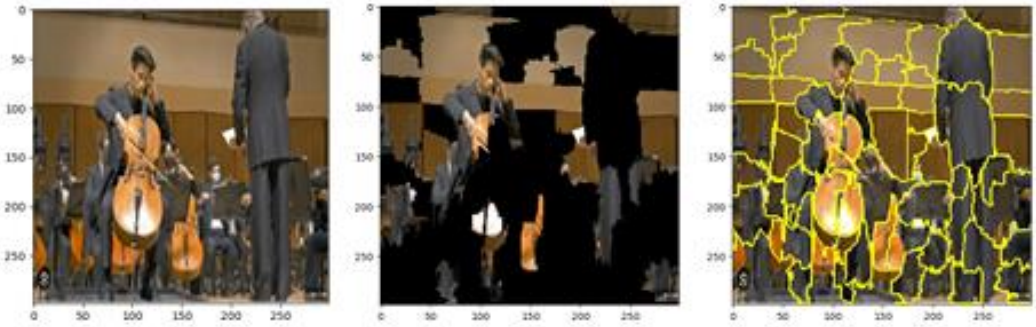

Figure 7: Target image from the class "cello", segmented image and one random perturbed image

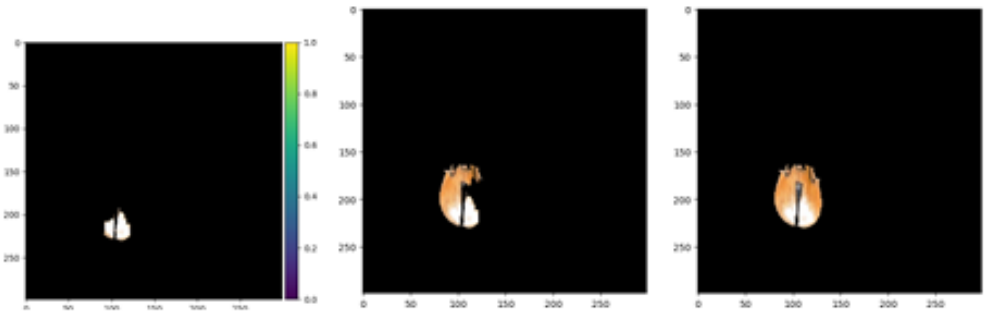

Figure 8: Target image from the class "cello", top three features used in the model decision

ImageNet weights spots correctly the top three defining features for the two images processed in this paper. An interesting occurrence can be spotted in the feature map for the class "missile". The third feature the model considers in its decision is a patch of clear sky, which, from human perspective, has nothing to do with the target object (a missile). This is possibly the result of an association the model infers during the training process, when some objects are frequently present in the image together with the target object (in this case, sky is often associated with missiles). To further investigate this phenomenon, a new image (more difficult to classify) from the class "cello, violoncello" is considered. For the original image presented in Figure 9, Inception V3 predicts the following classes: "cello": 0.518, "violin": 0.381, "microphone": 0.0245.

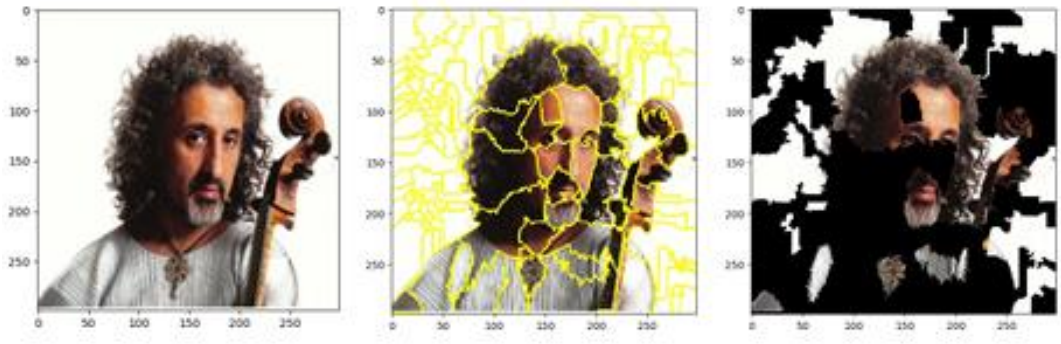

Figure 9: Target image from the class "cello", segmented image and one random perturbed image

The original, segmented and a random perturbed image are presented in Figure 9 The top features used in the model decision are presented in Figure 10. In the case of this particular image, it becomes clear that some sort of underlying association was generated during the training process since the features displayed in Figure 10 have nothing (from a human perspective) in common with the target object.

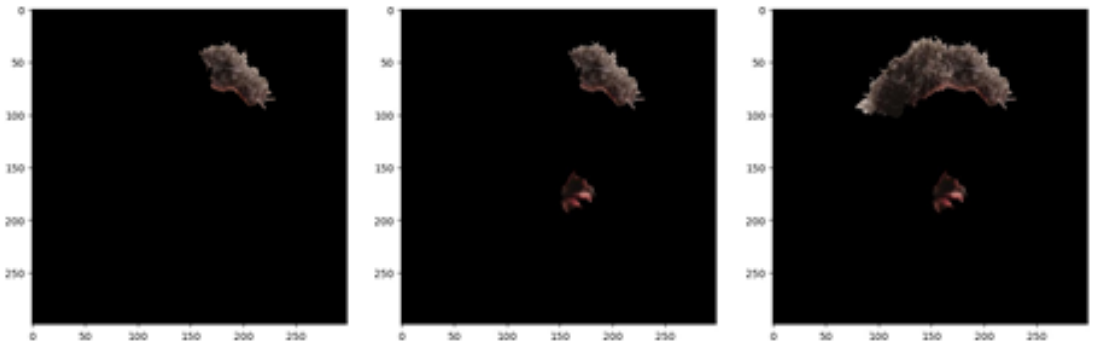

Figure 10: Target image from Figure 9, top three features (from left to right) used in the model decision

5.CONCLUSIONS

Explainability is an extremely useful instrument in assessing the classification algorithms. In this paper, the explainability for several image classification mod- els (a custom model and feature vector for EfficientNet B0, Inception V3 and ResNet architectures) was assessed by means of LIME algorithm. A low-volume, unbalanced six-class dataset was used to train the models, which exhibited sat- isfactory values of the performance metrics. It was found that explainability varies, depending both on the class and model. For some classes, all models (especially the custom model) ground their decisions on features that are not understandable in terms of human reasoning. In order to get some insight into what causes this variability in explainability, the Inception V3 architecture with the weights for ImageNet was used to predict the class for some images. It was found that for sharply defined images, Inception V3 identifies correctly and grounds its decision on the most specific features of the target image (in Figure 6 the missile body and the jet and in Figure 8 the lower part of the cello body and the tailpiece). In the case of ambiguous images, the decision of the model (the prediction is still correct, with the highest probability for the true class) seems to be based on associations inferred during the training. An interpre- tation of the phenomenon observed in Figure 10 is proposed as follows: since the individual depicted in Figure 9 is the distinguished cellist Mischa Maisky (link to Wikipedia web page), it is plausible that images in the training dataset labeled as belonging to the "cello" class may have featured him. The Inception V3 model may, therefore, have inadvertently associated certain visual features of Maisky with the "cello" class, potentially due to his frequent representation in connection with the instrument. Another perspective considers that other notable cellists may also have been present in the training dataset, yet this specific image's defining characteristic may be the highly distinctive texture and arrangement of Maisky's hair, which could be a unique identifier. Future studies could explore this hypothesis by (i) experimenting with images of indi- viduals presented with the target object (the cello) and (ii) by testing a range of Maisky's images, particularly those where he appears alongside the instrument. This could help clarify the model's tendency to associate personal attributes with specific object classifications, thus contributing to understanding biases in image recognition systems

REFERENCES

[1] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward net- works are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359–366, 1989.

[2] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. P´aez, W. Samek, J. Schneider, T. Speith, and S. Stumpf, "Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions," Information Fusion, vol. 106, p. 102301, 2024.

[3] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in Natural Language Processing and Chinese Computing (J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, eds.), (Cham), pp. 563–574, Springer International Publishing, 2019.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," 2018.

[5] M. L. Adekanbi, E. S. Alaba, T. J. John, T. D. Tundealao, and T. I. Banji, "Soiling loss in solar systems: A review of its effect on solar energy efficiency and

mitigation techniques," Cleaner Energy Systems, vol. 7, p. 100094, 2024.

[6] J. Ballestr´ın, J. Polo, N. Mart´ın-Chivelet, J. Barbero, E. Carra, J. Alonso-Montesinos, and A. Marzo, "Soiling forecasting of solar plants: A combined heuristic approach and autoregressive model," Energy, vol. 239, p. 122442, 2022.

[7] M. Santhakumari and N. Sagar, "A review of the environmental factors degrading the performance of silicon wafer-based photovoltaic modules: Failure detection methods and essential mitigation techniques," Renewable and Sustainable Energy Reviews, vol. 110, pp. 83–100, 2019.

[8] J. G. Bessa, L. Micheli, F. Almonacid, and E. F. Fern´andez, "Monitoring photovoltaic soiling: assessment, challenges, and perspectives of current and potential strategies," iScience, vol. 24, no. 3, p. 102165, 2021.

[9] R. R. Cordero, A. Damiani, D. Laroze, S. MacDonell, J. Jorquera, E. Sepu´lveda, S. Feron, P. Llanillo, F. Labbe, J. Carrasco, J. Ferrer, and G. Torres, "Effects of soiling on photovoltaic (pv) modules in the atacama desert," Scientific Reports, vol. 8, p. 13943, Sep 2018.

[10] K. K. Ilse, B. W. Figgis, V. Naumann, C. Hagendorf, and J. Bagdahn, "Fundamentals of soiling processes on photovoltaic modules," Renewable and Sustainable Energy Reviews, vol. 98, pp. 239–254, 2018.

[11] D. Korkmaz and H. Acikgoz, "An efficient fault classification method in solar photovoltaic modules using transfer learning and multi-scale convolu- tional neural network," Engineering Applications of Artificial Intelligence, vol. 113, p. 104959, 2022.

[12] A. Ettaleby, Y. Chaibi, A. Allouhi, M. Boussetta, and M. Benslimane, "A combined convolutional neural network model and support vector machine technique for fault detection and classification based on electroluminescence images of photovoltaic modules," Sustainable Energy, Grids and Networks, vol. 32, p. 100946, 2022.