ISSN 2007-9737

New Explainability Method based on the Classification of Useful Regions in an Image

Tonantzin Marcayda Guerrero Velázquez, Juan Humberto Sossa Azuela

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Laboratorio de Robótica y Mecatrónica,
México

{tmguerrerov, humbertososa}@gmail.com

Abstract. Machine learning is a necessary and widely used tool nowadays in Industry. Talking about evaluating its reliability, already known metrics are broadly used. These metrics are focused on how precise, accurate or sensitive the model is. Nevertheless, they do not offer an overview of the consistency or stability of the predictions, that is, how reliable the model is. This reliability can be deduced if the reasons behind the predictions are understood. In the present work, we propose a novel method that can be applied to image classifiers and allows the understanding, in a non-subjective visual manner, the background of how the model arrives at a prediction.

Keywords. Explainability, Classifier, XAI, Machine Learning.

1 Introduction

Nowadays, Explainable Artificial Intelligence (XAI) turns out to be an interesting area within the field of machine learning. Although it is a relatively new field, its attraction lies in the usability that it can be granted. XAI is about improving the human understanding of artificial models and trying to justify their decisions.

In the context of Artificial Intelligence, explainability refers to whatever action or process is carried out, intending to clarify the decision process.

Most of the time, the concept of explainability is used in the same manner as interpretability. However, interpretability refers to the level at which a model has a sense for a human being. This concept can also be expressed like the transparency of the model.

A model is considered transparent if, by itself, it is understandable such as a logistic regression model, a decision tree, or a classifier based on rules [1].

Some explanation methods and strategies have surfaced due to the need to analyze the decisions of machine learning models. In [2] the authors propose three primary classes of explanation methods. To the first class belong the rules-extraction methods. The goal here is to approximate the decision-making process for a model using its inputs and outputs. The second class corresponds to so-called attribution methods, which measure how much changing the inputs or internal components affects the model's performance.

The last class involves so-called intrinsic methods. Here, the goal is to enhance the interpretability of internal representations making methods derived from the model architecture. Among the different techniques that provide an explanation of a deep learning-based model are the explainability methods LIME (Local Interpretable Model-Agnostic Explanations), and RISE (Randomized input sampling for explanation of black-box models).

For image classifiers, LIME creates a set of images that result from perturbing the input image by dividing it into interpretable components (superpixels), to obtain a belonging probability for each of these perturbed instances. LIME generates a visual explanation based on the classification of this new perturbed data, resulting in an area of the input image that denotes what the model looked at to make a prediction [3].

On the other hand, RISE [4] produces a heat map or a saliency map showing those parts of an input image that are most important for the prediction made by a neural network. The heat map of an input image is obtained by generating random masks and superimposing them to the original image.

Afterward, those versions of the original image with the overlapping masks feed the neural network to observe the changes that happen at the network's output. When this process is repeated many times, it is possible to identify which image features are more important for the prediction made by the model.

Nevertheless, there is a major disadvantage in current methods for visual explanations, which refers to the subjectivity of the results. These results are subject to the interpretability of the user. In consequence, the method's reliability can be questioned. Another important disadvantage is that the results of the explanations turn out to be unstable. In [5], authors show that the explanations obtained for two very close points become highly variable with each other, which also makes these explanation tools unreliable [6].

In this work, we present a solution for overcoming one of the problems described above (the subjectivity of the results). The proposed solution consists in creating a visual explanation of the prediction based on the characterization of certain regions of the image according to its importance for the prediction.

The main contribution lies in the creation of a novel explainability method with non-subjective explanations, and this issue is tackled in two ways. Firstly, there is no configuration parameter for the algorithm, which ensures that it does not depend on the person who implements it; the same result is always achieved. Secondly, the resulting explanation is clear and easy to understand as well as intuitive due to the proposed categories for each useful region. In this way, anyone who knows the color code used will be able to give the same explanation of the prediction made by the model.

The rest of the paper is organized as follows. In section 2, we describe the methods we use throughout the paper. In section 3 we present our results and provide a discussion on these results. In section 4, we finally conclude.

2 Methods

The main goal of the proposed method is to identify the regions of the input image that are most relevant for the prediction of the classifier and categorize them as *significant, relevant, and futile*. Then, those regions are highlighted as a visual explanation with a color code defined by the colors green, yellow, and red, respectively. This goal is achieved first by doing a selective search that will result in a set of candidate regions of the image, so named because they might be relevant; however, it is not yet known whether these regions are relevant to the classifier. Therefore, these candidate regions are evaluated using the same classifier and go through statistical analysis so that the most relevant regions can be chosen and now considered useful. Finally, these useful regions can be categorized and colorized as *significant, relevant, and futile.*

2.1 Searching Useful Regions

To search for the set of proposed regions in the image from which the most useful regions for the classifier will be obtained, we decided to use the selective search algorithm [7], where a graph-based segmentation method is used to carry out the search for regions in the image [7, 8]. In this algorithm, the input is considered as a graph $G = (V, E)$, where n represents the number of vertices and m the number of edges of G. Similarity between regions is hierarchically propagated, for which the following equations are used.

$$S_{colour}(r_i, r_j) = \sum_{k=1}^{n} \min(c_i^k, c_j^k). \quad (1)$$

Equation (1) is about the color similarity for each pair of regions r_i, r_j using the intersection histogram, where c_i and c_j refer to the histograms of these regions. Additionally, the texture histogram is obtained for each region. Then the texture similarity measure can be calculated as follows:

$$S_{texture}(r_i, r_j) = \sum_{k=1}^{n} \min(t_i^k, t_j^k). \quad (2)$$

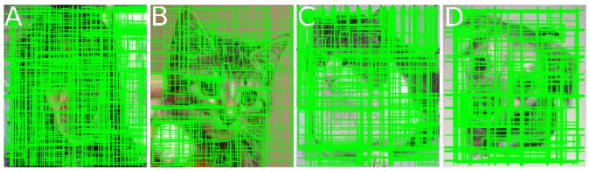

Fig. 1. Regions resulting from applying the selective search algorithm for four different images of the *Dogs vs Cats* dataset.

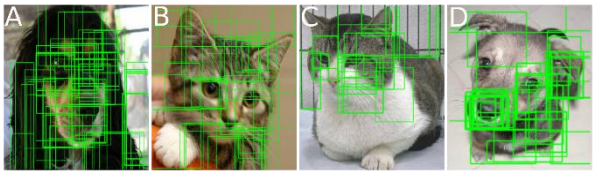

Fig. 2. Set of proposed regions found after applying the class membership filter for each region of Fig. 1.

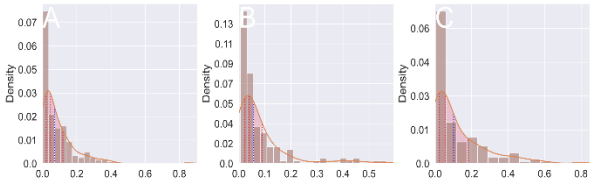

Fig. 3. Probability distributions obtained statistically with different pre-trained models. (A) *Inception model*, (B) *Resnet model*, (C) *Inception-Resnet* model.

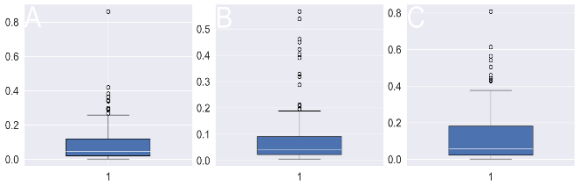

Fig. 4. Boxplots obtained statistically with different pre-trained models. (A) *Inception* model, (B) *Resnet* model, (C) *Inception-Resnet* model.

The following similarity measure is used to make small regions join with the larger regions:

$$S_{size}(r_i, r_j) = 1 - \frac{size(r_i) + size(r_j)}{size(im)}, \qquad (3)$$

where, $size(im)$ is the size of the image in pixels. With this process, we find the set of proposed regions= $\{r_1, \ldots, r_n\}$, where $r_i = \{x, y, w, h\}$ i.e., that each region r_i represents a bounding box with the pair (x, y) representing its position and (w, h) its size. Fig. 1 shows examples of proposed regions found by the selective search algorithm for some images from the *Dogs vs Cats* dataset taken from *Kaggle* [9]. The dataset includes12,500 images that correspond to images of dogs and cats.

Now, it will be necessary to find which of these resulting regions have the greatest influence on the model prediction. Given a classifier $C(x)$, of which the visual explanation is required, $C(x)$ is applied to each of the previously obtained regions, that is $C(R)$, thus generating a new set $P = \{p_i, \ldots, p_n\}$ where:

$$p_i = \begin{cases} prob(cls) & if \quad prob(cls) = \max(C(r_i)) \\ 0 & if \quad prob(cls) \neq \max(C(r_i)) \end{cases} \qquad (4)$$

and $prob(cls)$ is the belonging probability to the class of the result that is being explained. Now, if all the regions r_i for which $p_i = 0$ are removed from R, a new smaller and precise set R is obtained as illustrated in Fig. 2.

It is known that $\forall\, p_i \in P \rightarrow 0 \leq p_i \leq 1$. Therefore, a comparative study of the sets of probabilities, P, was carried performed on, for the 12,500 images from the *Dogs vs Cats* dataset with different classifiers (*Inception, Resnet, and Inception-Resnet*), resulting in a probability distribution skewed to right, shown in Fig. 3 and Fig. 4.

In the charts shown in Fig. 3, it is easy to observe that most of the proposed regions have a low probability. Also, given the bias present in their probability distribution, it is possible to think about utilizing the values of quartile ranges to differentiate and categorize these regions.

This can be best observed in the boxplots of Fig. 4 where the white line represents the median value of probability data, and the bounds of the box show the upper and lower quartiles, this is Q_1 and Q_3. The extreme upper and lower represent the highest and the lowest value, respectively, leaving out the outliers. Then, it is important to highlight that the so-called outliers, in this case, turn out to be regions with the highest probability and these regions will be directly categorized as the most significant regions to make a prediction. Now, we proceed to obtain the useful regions using the values of the quartiles and define them as:

$$R_u \subset R \mid r_i \in R \wedge p_i > Q_2. \qquad (5)$$

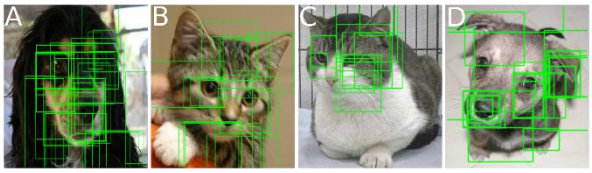

Fig. 5. Set of useful regions R_u for the images (A, B, C and D) in Fig. 1.

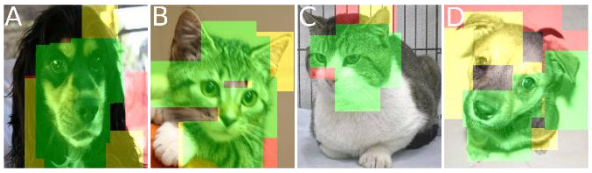

Fig. 6. Result of visual explanation of images in Fig. 1 (A) *Cocker_spaniel* class, (B) *Egyptian_cat* class, (C) *Egyptian_cat* class, (D) *Norwegian_elkhound* class, classified with the *Inception-Resnet* model, highlighting the significant, relevant, and futile regions with the colors green, yellow, and red, respectively.

Finally, with this process, it has been reduced the set R to R_u that contains the most useful regions for the classifier to make their prediction. An example of the useful regions can be appreciated in Fig. 5.

2.2 Characterization of Regions and Visualization

Once the set of useful regions R_u has been found, as described in the previous section, each of these regions will have to be evaluated to classify them into the three possible categories, *significant, relevant,* and *futile*. As it is implicit in their names, each category refers to the level of importance that each has for the classifier's decision-making. One time the regions have been categorized, they can be easily color-coded by the colors green, yellow, and red, respectively, over the same image. Therefore, given the classifier $C(\mathrm{x})$ from which we want to obtain the visual explanation of the prediction, the category of each region R_u will be given by $F\,(C\,(R_u))$, where F is a function defined as:

$$F(x) = \begin{cases} significant, & p_i \geq Q_3 \\ relevant, & threshold \geq p_i < Q_3 \\ futile, & p_i < threshold \end{cases} \quad (6)$$

where the value of the $threshold$ is given by:

$$Q_1 + \frac{(Q_3 - Q_1)}{2}, \quad (7)$$

which is equal to the semi-interquartile range. This value was defined in this manner according to the statistical analysis discussed above, where it can be observed that this value ($threshold$) is always greater than quartile two Q_2.

Therefore, this range between the *threshold* and the quartile two Q_2 will serve to denote the regions that are within the *futile* range, i.e., those regions that have less relevance for the prediction made by the model.

Thus, those regions whose probability value are less than quartile three Q_3 and greater than the $threshold$ will be considered *relevant*, and those that are above quartile three Q_3 will be those that have a greater influence on the prediction of the model, therefore, these regions fall into the *significant* category.

Finally, the regions are colored according to their category with the colors green, yellow, and red to denote the *significant, relevant, and futil*e regions, respectively. These colored regions will be highlighted in the original image, where it is desired to obtain an explanation of the prediction made by the classifier model, as shown in Fig. 6.

3 Results and Discussion

Throughout this work, we present several examples and statistics resulting from the application of the algorithm proposed here, over three different datasets: (1) The dataset taken from *Kaggle* [9], that contains 12,500 images that correspond to images of dogs and cats, (2) the images from Microsoft COCO dataset [10], which contains more than 200,000 images with objects labeled and marked by human beings, and (3) the Places365 dataset [11] that contains more than 10 million images that comprise more than 400 categories of unique scenes. We also used the pre-trained models *InceptionResnet* [12], which is a convolutional neural network (CNN) that was trained with more than a million images from the ImageNet database, and the *Resnet50_places365,* which is also a

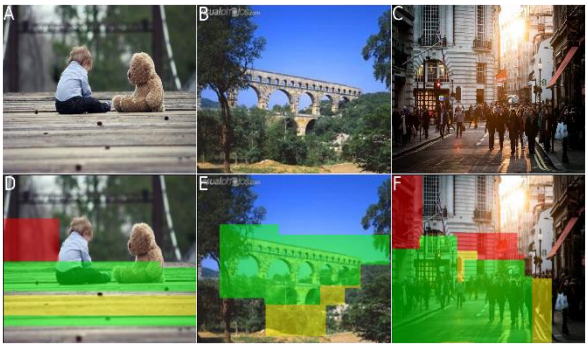

Fig. 7. Example of different images representing a scene. (A) Image with a probability of 0.728 for *pier* class, (B) Image with a probability of 0.935 for *aqueduct* class, (C) Image with a probability of 0.378 for *crosswalk* class (D), (E), and (F) represent the explanation obtained of each of them, respectively.

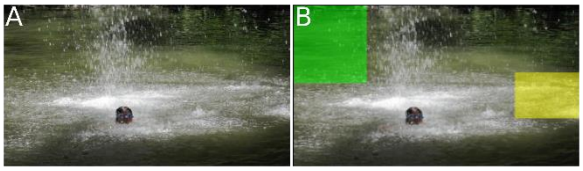

Fig. 8. (A) Image belonging to the *swimming_hole* class and the (B) explanation obtained by our proposed method.

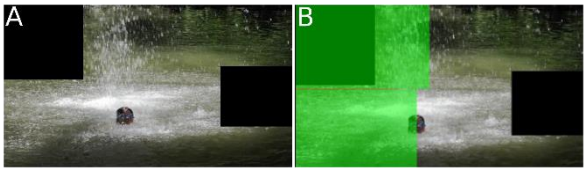

Fig. 9. Image belonging to the *fountain* class and its explanation after covering the *significant* and *relevant* areas marked in Fig. 8.

convolutional neural network (CNN), but this was trained with the Places365 dataset [11].

First, to show the performance of the algorithm proposed here, we used the classifier model *Resnet50_places365*. Fig. 7 shows three different scenes and the explanation of the prediction of the model in each case.

In Fig. 7, it is possible to see that the three explanations correspond to the region that the network should use to choose the label it predicts. As humans, it makes perfect sense since just by looking at these regions it is possible to say that the predicted class is correct.

However, the potential of this method goes beyond only explaining the correct classifications. It also works to understand the behavior of the network. For example, Fig. 8 shows an image that was predicted as *swimming_hole* class, which makes total sense to us.

As humans, we would think that one of the most important things for this prediction is the child swimming in the center of the image and we assume that for the classifier model as well. However, when doing the explanation, we can see that it shows how the model used totally other different regions of the image than we think; this makes us doubt the method's efficiency.

To verify that the explanation algorithm is correct, we cover those parts in the image marked as *significant* (green area) and *relevant* (yellow area) to check how the model prediction is affected. Then, when we classify this new marked image, the predicted class changes to *fountain* that is a different class, and therefore a different explanation as depicted in Fig. 9.

As we know, the classification models do not always respond as we would like, and according to the previous example, we can say that the use of the method proposed in this work helps to explain the decision-making of the model, as well as to the model improvement.

3.1 Impact of the Proposed Method

The relevance of the work presented here lies in the importance of knowing the reliability of the predictions given by a model, because, by definition, no model is perfect not even *InceptionResnet*, so when an incorrect prediction is made, it would be very helpful to know the reason and, thus, be able to improve the model.

If we ask a person to observe and identify the class to which the images in Fig. 10 belong, surely, they will answer to the *cat's* class. However, the *InceptionResnet* model classifies these images at different classes: *shopping_basket*, *quilt*, and *shoji*, respectively.

Thanks to the visual explanation method proposed here, it is possible to know how exactly the model makes its classification decision as can be appreciated in Fig. 11. Then, the impact of the method is demonstrated. Depending on the problem and the implementation of the model, this

explanatory factor will be decisive in the improvement and appropriate uses of the model, in addition to contributing to its reliability.

3.2 Advantages of the Proposed Method

The proposed method solves one of the problems strongly present in other methods such as LIME [3] and RISE [4]: the subjectivity of its results. As depicted in Fig. 12, our method overcomes this problem by providing well-defined results according to the different proposed categories (significant, relevant, and futile).

For example, in the explanation obtained by LIME, it can be observed that regions that are not within the region marked in black are marked as important. It is also observed that this explanation is like the one obtained with our method, however, a difference in the results is that LIME does not mark well-delimited regions that are considered important, that is, those regions in which a person could surely look at to determine that this image belongs to the *patio* class.

It can also be perceived that the region resulting from the explanation with LIME is not well-delimited, since it shows a non-uniform region that includes certain confusing parts of the image which, as human beings, it is difficult for us to identify what they are, and consequently to obtain different conclusions for each observer who analyzes the results.

On the other hand, RISE generates an explanation very different from the explanations obtained with our method and with LIME. This could be somewhat confusing, because in this explanation the regions of the greatest interest are highlighted as those marked in red and yellow, which as can be seen in Fig. 12 (D) are scattered throughout the image.

In addition, regions that could be of greater weight to reach a prediction are left out, for example, all the chairs and tables that are in the image, as well as part of the floor. Or, on the contrary, regions that might not be so important are considered, such as the window. Although the window is part of the image, there is no clear relevance of it for the prediction since a window can appear in different types of images that may belong to a different class than the *patio* class. It is also noted that the effectiveness of RISE varies

Fig. 10. *Classifications made with InceptionResnet*. (A) as *shopping_basket*, (B) as *quilt,* and (C) as *shoji.*

Fig. 11. Visual explanations for the classes predicted by the *InceptionResnet* model for the images in Fig. 10.

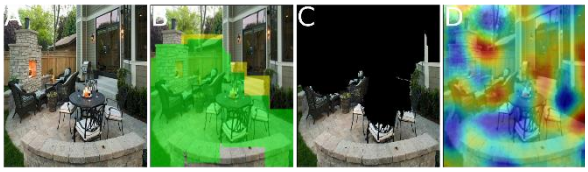

Fig. 12. Explanations resulting from the application of different methods to the same image, which the *Resnet50_places365* classifier model predicted as the main class *patio*, with a probability of 0.685. (A) Original image, (B) Explanation obtained by the proposed method, (C) explanation obtained by LIME, (D) Explanation obtained by RISE.

depending on the number of classes with which a model could classify an image and that the time to generate an explanation turns out to be somewhat high compared to that of our method.

In contrast to the explanations obtained with LIME and RISE, it is observed that the proposed method delimits, with known and well-defined patterns (rectangles) that are also easy to perceive and understand for humans, those regions of interest that are important to classify this image as a *patio*. In these regions of interest, we can observe the chairs, the tables, the fireplace, the fence, the stairs, and even the floor, which has a typical finish that a patio could have. In addition, the importance of these regions is clearly differentiated and denoted by the color code (green, yellow, and red) defined according to their relevance (significant, relevant, and futile) for the prediction made by the model.

With this, there is a clearer intuition as to what the model has given more weight to perform its classification task.

It can be seen, that, unlike the proposed method, the explanations obtained with LIME and RISE are not so clear and are also subjective, i.e., the interpretation may be different depending on the observer. These methods also require a previous configuration of parameters, on which the obtained result depends.

Table 1 shows a summary of the characteristics of LIME and RISE methods and the proposed technique. The time column is based on the explanation obtained by each method for Fig. 12 (A), classified by the model as *patio* class, with a size of 640 x 426 pixels.

3.3 Explanation by Class

Using the proposed method, it is possible to obtain not only the explanation of the class with greater probability but also of other classes with a slightly lower probability.

For example, in the case of the *InceptionResnet* model that was trained to predict 1000 different classes, it may be the case that given an image, this image may belong in different degrees to different classes, and this belonging can be explained by applying our method. We show this in Fig. 13.

3.4 Method Validity

In order to verify the validity of the algorithm proposed in this work, a comparison process is carried out between the useful region obtained by the proposed algorithm, and the region selected by a person within that image. This region represents the most important region to determine whether an image belongs to one class or another. This comparison is carried out using the COCO data set, consisting of images tagged according to objects within the image whose bounding box is marked by a human being, and it will be used as an indicator, since this region is what the network should ideally consider classifying the image.

Fig. 14 shows four examples of images belonging to the COCO dataset. These images have been marked and classified by human beings. These regions are compared against the

Fig. 13. Visual Explanation of the prediction for multiple classes present in an image. (A) *Weimaraner* (B) *shower_curtain* (C) *Siamese_cat.*

Fig. 14. Examples of images from the COCO dataset, marked and classified by a human.

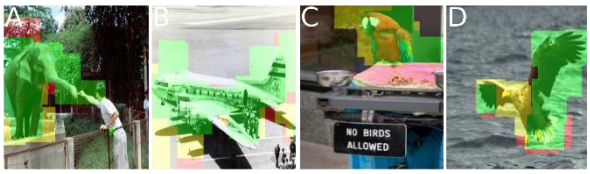

Fig. 15. Result of visual explanation for (A), (B), (C), and (D) from Fig. 10.

explanation obtained by the algorithm proposed here.

In Fig. 15, it is possible to observe that the useful region found by our explanation algorithm effectively surrounds the entire region marked by the human being.

These results coincide rather well with those that would be expected from a good visual explanation algorithm; this, of course, if the classifier model used is well trained. Otherwise, the utility of the explanation algorithm would change.

Keeping in mind the same logic, a subset of 200 different images of the COCO data set belonging to different classes were selected, and then a comparison with the proposed method explanation was made.

In order to carry out this comparison, the overlap of the two regions is measured as $O_v = A_o/A_c$, where A_o is the area of overlap of the area marked in the images belonging to COCO against the area of the useful regions found by our proposed explanation algorithm, and A_C is the area

Table 1. Characteristics of methods

Method	Year	Subjective	Time (sec)	Images	Text
LIME	2016	Yes	105	✓	✓
RISE	2018	Yes	480	✓	-
Proposed method	2021	No	25	✓	-

marked in COCO. Thus, we expect if the model used has been well trained, the visual explanation obtained should always cover what a human has determined as important for the classification of an image, such is the case of the *InceptionResnet* model, i.e., $O_v = 1$.

The results obtained were favorable for all cases. We observed that the regions marked in COCO are always within the useful regions founded by the proposed explanation algorithm or $O_v \cong 1$, demonstrating that the method is effective, and can be used successfully to find a visual explanation of model prediction.

In addition to this and very importantly, the explanation is given in a non-subjective way using three clear, color-coded categories *(significant, relevant, and futile*), according to the relevance of the regions.

4 Conclusions and Future Work

In the present work, a new method was proposed to try to give a simple and easy-to-understand explanation about the predictions of an image classifier model. It has been demonstrated its validity, relevance, and usefulness.

Furthermore, it has been shown clear advantages, such as solving the problem of subjectivity which is present in other explainability methods. This turns out to be very important since it is not subject to the interpretation of a particular person so that anyone will give the same interpretation to the explanation obtained and even it could be analyzed automatically, which has been left for future works.

It was also shown that the performance of the proposed method is useful not only for correctly trained models, but also helps in understanding the model prediction which sometimes goes against human intuition and thus be able to correct the model or data in a relevant way.

For future research, we propose working towards improving the time and performance of the proposed method and its generalization to other types of classifiers.

Acknowledgments

Tonantzin Guerrero thanks CONACYT for the scholarship to undertake her doctoral studies. The authors thank the Instituto Politécnico Nacional for the economic support under projects SIP 20200630 and 20200788, and CONACYT under projects 65 (Fronteras de la Ciencia) and 6005 (FORDECYT- PRONACES).

References

1. **Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Chatila, R. (2020).** Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, Vol. 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.
2. **Ras, G., van Gerven, M., Haselager, P. (2018).** Explanation methods in deep learning: Users, values, concerns and challenges. Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp. 19–36. DOI: 10.1007/978-3-319-98131-4_2.
3. **Ribeiro, M.T., Singh, S., Guestrin, C. (2016)**. Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

4. **Petsiuk, V., Das, A., Saenko, K. (2018).** Rise: Randomized input sampling for explanation of black-box models. arXiv:1806.07421.
5. **Alvarez-Melis, D., Jaakkola, T.S. (2018).** On the robustness of interpretability methods. arXiv:1806.08049.
6. **Molnar, C. (2020)**. Interpretable machine learning. A guide for making black box models explainable.
7. **Uijlings, J.R., van De Sande, K.E., Gevers, T., Smeulders, A.W. (2013).** Selective search for object recognition. International Journal of Computer Vision, Vol. 104, No. 2, pp. 154–171. DOI: 10.1007/s11263-013-0620-5.
8. **Felzenszwalb, P.F., Huttenlocher, D.P. (2004).** Efficient graph-based image segmentation. International Journal of Computer Vision, Vol. 59, No. 2, pp. 167–181. DOI: 10.1023/ B:VISI.0000022288.19776.77.
9. **Microsoft and PetFinder.com (2019)** Dogs vs. Cats dataset. https://www.kaggle.com/ c/dogs-vs-cats/data.
10. **Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L. (2014).** Microsoft coco: Common objects in context. European Conference on Computer Vision, Springer, pp. 740–755. DOI: 10.1007/ 978-3-319-10602-1_48.
11. **Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A. (2017).** Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, No. 6, pp. 1452–1464.
12. **Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. (2016).** Inception-v4, Inception-Resnet and the impact of residual connections on learning. arXiv:1602.07261.
13. **Amirata, G., Zou, J. (2019)**. Data Shapley: Equitable valuation of data for machine learning. arXiv:1904.02868.
14. **Robnik-Šikonja, M. (2018).** Explanation of prediction models with explain prediction. Informatica, Vol. 42, No. 1.
15. **Bansal, A., Farhadi, A., Parikh, D. (2014).** Towards transparent systems: Semantic characterization of failure modes. European Conference on Computer Vision, pp. 366–381. Springer. DOI: 10.1007/978-3-319-10599-4_24.
16. **Papernot, N., McDaniel, P. (2018).** Deep k-nearest neighbors: Towards confident, interpretable, and robust deep learning. arXiv:1803.04765.
17. **Ribeiro, M.T., Singh, S., Guestrin, C. (2016).** Model-agnostic interpretability of machine learning. arXiv:1606.05386.
18. **Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. (2015).** Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.
19. **Doshi-Velez, F., Kim, B. (2017).** Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.
20. **Simonyan, K., Vedaldi, A., Zisserman, A. (2013).** Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.
21. **Zhang, P., Wang, J., Farhadi, A., Hebert, M., Parikh, D. (2014).** Predicting failures of vision systems. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3566–3573.
22. **Adler, P., Falk, C., Friedler, S.A., Rybeck, G., Scheidegger, C., Smith, B., Venkatasubramanian, S. (2016).** Auditing black-box models for indirect influence. IEEE 16th International Conference on Data Mining (ICDM).
23. **LeCun, Y., Kavukcuoglu, K., Farabet, C. (2010).** Convolutional networks and applications in vision. Proceedings of IEEE International Symposium on Circuits and Systems, pp. 253–256.
24. **Gunning, D. (2017).** Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), Vol. 2, No. 2.
25. **Linardatos, P., Papastefanopoulos, V., Kotsiantis, S. (2021).** Explainable AI: A review of machine learning interpretability methods. Entropy, Vol. 23, No. 1, pp. 18. DOI: 10.3390/e23010018.
26. **Confalonieri, R., Coba, L., Wagner, B., Besold, T.R. (2021).** A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 11, No. 1. DOI: 10.1002/widm.1391.
27. **Islam, S.R., Eberle, W., Ghafoor, S.K., Ahmed, M. (2021).** Explainable Artificial Intelligence Approaches: A Survey. arXiv:2101.09429.
28. **Jiménez-Luna, J., Grisoni, F., Schneider, G. (2020).** Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, Vol. 2, No. 10, pp. 573–584. DOI: 10.1038/s42256-020-00236-4.
29. **Meske, C., Bunde, E. (2020).** Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. Proceedings of the International Conference of AI in HCI´20, pp. 54–69. DOI: 10.1007/978-3-030-50334-5_4.

30. Das, A., Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv:2006.11371.

Article received on 21/08/2020; accepted on 12/01/2021.
Corresponding author is Juan Humberto Sossa Azuela.