CYBERNETICS and COMPUTER TECHNOLOGIES

УДК 519.67 DOI:10.34229/2707-451X.25.2.6

V. YAVTUKHOVSKYI, V. TRETYNYK

EVALUATION OF SIMILARITY OF IMAGE EXPLANATIONS PRODUCED BY SHAP, LIME AND GRAD-CAM

The level of differences between SHAP, LIME, and Grad-CAM on an image classification task is evaluated. The evaluation was performed on two datasets, with one fine tuned and one pre-trained model. The datasets were the CBIS-DDSM breast cancer dataset with fine tuned ResNet-18 model, and the Imagenet Object Classification Challenge (IOCC) with a VGG-16 pre-trained model.

***Keywords**: computer vision, convolutional neural network, Grad-CAM, LIME, SHAP, saliency maps, explainable AI, XAI.*

Introduction. Convolutional neural networks (CNNs) are a subtype of neural networks developed specifically to work with images [1]. They have achieved great success both in research and in practical applications in recent years, however, one of the major pain points when adopting them is the lack of ability to interpret what is the reasoning behind their conclusion. Because of this, various explainable artificial intelligence (XAI) methods have been developed; however, it is unclear if they show reasoning or the same aspects of reasoning of CNNs. In recent years some of the most popular methods, LIME [2], SHAP [3], and Grad-CAM [4], were evaluated using tabular data and it was showed how significantly different results are [5] or some were evaluated on a matter of trustworthiness with human evaluation on medical images [6], there is still a lack of measure of how different these methods are on image classification models. This study uses correlation and a popular segmentation measure, Intersection over Union (IoU)[7], to evaluate their differences.

Problem statement. The aim of this work is to evaluate the level of differences between SHAP, LIME, and Grad-CAM on an image classification task.

Description of methods. CNNs are based on a hierarchical representation of perceiving visible features, with each layer of the hierarchy representing higher and higher levels of features, from singular shapes on the lowest level to complex forms on higher levels. CNN is usually composed of two broad stages, the feature extraction stage and the classification stage. Feature extraction is done by using convolution filters, with possible hyperparameters such as a number of filters, filter size, stride, padding, type of pooling, and activation function. In the classification stage, the multidimensional results of feature extraction are first “flattened” into a 1-dimensional array with the previously mentioned pooling operation, after which classification is done as usual with artificial neural networks (ANNs) [1].

As was previously demonstrated, in practice, humans often trust more in systems that can provide explanations for their decisions [8], which made a significant impact and created a demand for algorithms that can produce explanations.

Pixel-attribution methods explain the predictions of a model (usually a convolutional neural network) in computer vision by attributing importance scores to individual pixels or regions in an input image. The goal is to highlight which parts of the input were most influential in the model's decision. Some of the widely used algorithms are LIME [2], SHAP [3] and Grad-CAM [4].

SHAP (SHapley Additive exPlanation) is a framework for using Shapley values [9] for explanation, which is a system-agnostic approach for evaluating the importance of features during classification. It's done in the following steps:

For each feature i in a set of all features F, which has all feature subsets S, calculate the difference in predictions of the model

$$d_i = f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S),\ S \subseteq F\backslash\{i\}.$$

Calculated weighted average of all such possible differences

$$\phi_i = \sum_{S\subseteq F\backslash\{i\}} |S|(|F| - |S| - 1)|F|! * d_i .$$

LIME (Local Interpretable Model-agnostic Explanations) is an approach based on creating perturbations in original data, where some features in the initial data are randomly modified, and the classification process is done again, and comparing result values to assign importance to variables.[2] The value of explanation can be defined as the following

$$\xi(x) = \arg\min L(f, g, \pi_x) + \Omega(g),\ g \in G.$$

Where G is a class of potentially interpretable models, such as linear models or decision trees, f is the initial model, πx is a measure of distance between original and perturbed data, and Ω(g) is the complexity of the selected model to be interpreted.

As the complexity is practically preselected when interpreting an algorithm, this task is transformed into the minimization of the following function

$$L(f, g, \pi_x) = \sum_{z, z'\in Z} \pi_x(z)\left(f(z) - g(z')\right)^2.$$

Grad-CAM is an algorithm specific to CNNs, which creates class-discriminative explanations. The core idea of the method is to compute gradients of the target class for the convolutional layers in the CNN. First, class activations y^c are calculated, where c denotes the selected class c. Then they are used to compute gradients with respect to the feature map activations A^k of the selected convolutional layer. The explanation weight for each feature map is computed by global-average pooling of these gradients:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}.$$

Where Z is the product of the dimensions of the feature map. Then the saliency map can be obtained as a weighted sum of feature maps. Optionally, as suggested by the authors of the method, the ReLU function is applied to retain only positive features, which is followed in this experiment.

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right).$$

This map can then be upsampled for alignment to the input image resolution by applying bilinear interpolation.

Evaluation methods. To assess the question about similarity between the values of three chosen methods, it was decided to evaluate it from two perspectives: how they correlate and how they overlap. To achieve that, methods were evaluated pairwise on a subset of data, calculating Pearson correlation coefficient (PCC) and Intersection over Union (IoU) measure [7].

PCC is defined as

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

Where X and Y are random variables, *cov* is covariance, σ is the standard deviation.

IoU for sets A and B is defined as

$$IoU(A, B) = \frac{A \cap B}{A \cup B}.$$

Only positive values were considered, as it allows values to be calculated in the same way as the set membership function. Values for each method were normalized (scaled to a range from 0 to 1), as each method returns values of a different scale and can be hard to compare directly [10].

After that, the IoU value for each pixel can be calculated as follows:

$$IoU(f, g, x) = \frac{\min\big(f(x), g(x)\big)}{\max\big(f(x), g(x)\big)}.$$

Data description

For this task, two datasets were used, CBIS-DDSM breast cancer dataset [11] was used for training and evaluation of one model, and Imagenet Object Classification Challenge (IOCC) [12].

CBIS-DDSM consists of 10239 scanned film mammography images. It's split into mass and calcification type findings, and for the purposes of this paper, only mass type was considered, for which there is a total of 1696 images. It contains normal, benign, and malignant cases with verified pathology information. For simplicity, the data was used as a binary classification problem with the malignant class staying as is, and normal and benign classes united into "non-malignant".

IOCC consists of 1.4 million images with 1000 object categories, with one or more objects present in the image with corresponding bounding boxes. For the experiment, 500 random images were selected, and for each, the first object was selected as the class to be evaluated on.

Image classification models. In the experiment, for IOCC, a VGG-16 CNN [13] was used. VGG-16 is a CNN which has achieved great results in image classification tasks in research and machine learning competitions. Its parameters can be found in Table 1.

TABLE 1. VGG-16 network details

Input shape	244 x 244 x 3
Total number of convolution layers	13
Total number of fully connected layers	3
Total number of parameters	14,977,857

For CBIS-DDSM, the process of transfer learning from ResNet-18 was used. This model uses residual connections connected in basic blocks, which are then combined into residual blocks [14]. The model was imported with the weights from the Imagenet challenge [15], all fully connected layers were removed and replaced with new ones to be trained for the classification task. Details can be found in Table 2.

TABLE 2. Modified ResNet-18 network details

Input shape	244 x 244 x 3
Total number of residual blocks	8
Total number of fully connected layers	3
Total number of parameters	11,572,546

The model was trained on 1318 images and evaluated on 378 images. It was trained over 12 epochs with Adam [16], with the first 6 epochs having a 0.001 learning rate, and the following 6 having a 0.0001 learning rate.

Figures 1 and 2 show the dynamics of loss and accuracy on both train and validation data for the ResNet-18 model per training epoch. The model has achieved 81.4 % accuracy on the training data and 61.9 % accuracy on the validation data.

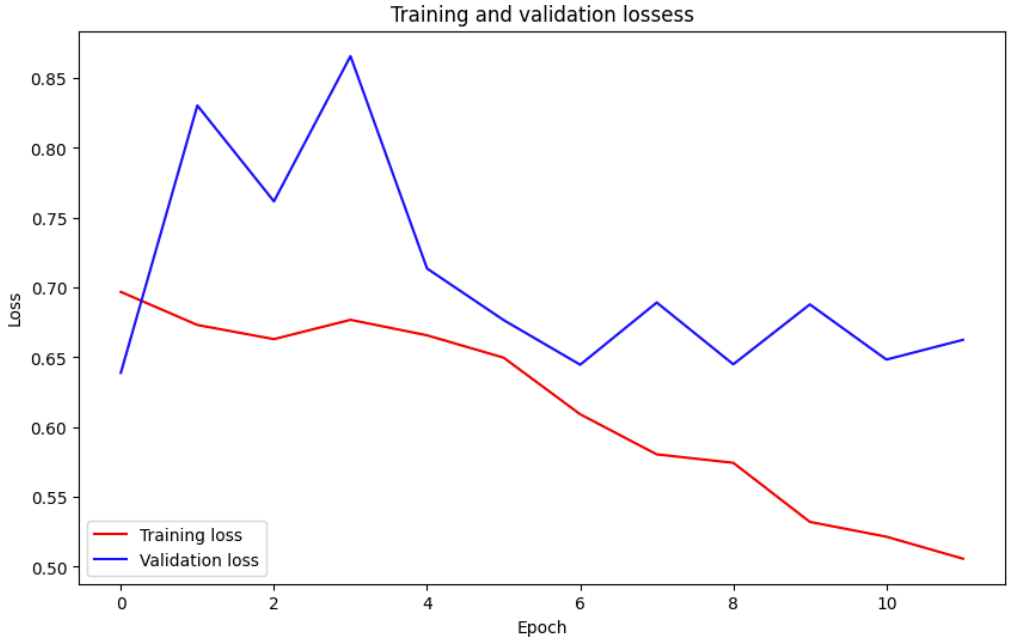

FIG. 1. Loss of the model per training epoch. Red line corresponds to training data, blue to validation data

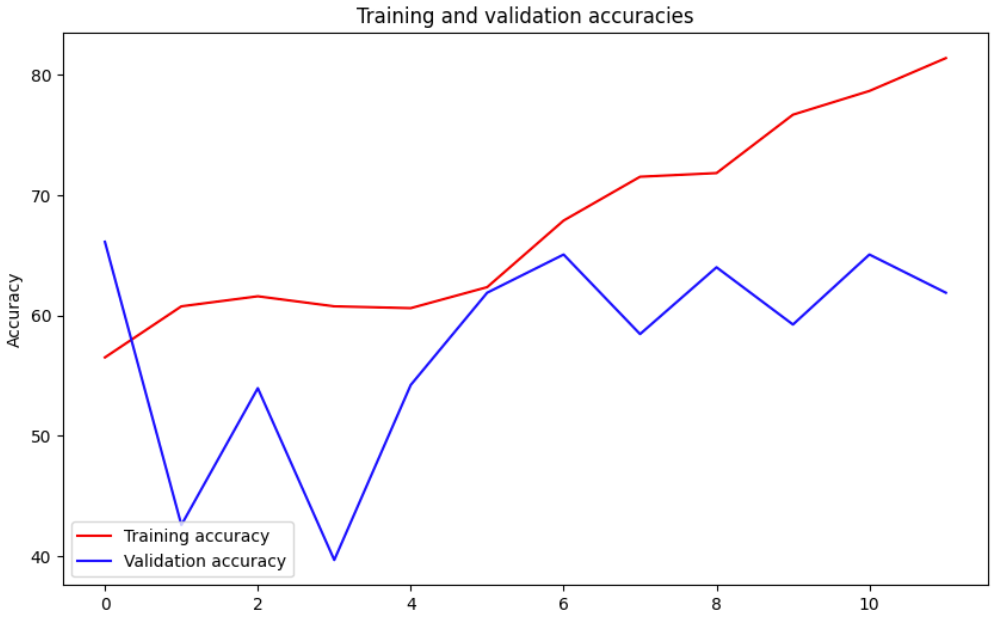

FIG. 2. Accuracy of the model per training epoch. Red line corresponds to training data, blue to validation data

Evaluation of similarity. The models were used to evaluate previously described PC and IoU measures. For SHAP, it was selected to use 100 evaluations as a limit for the number of subset evaluations. For LIME, the number of perturbed images to train linear classifiers described in the algorithm was set to 100, and for the Grad-CAM last convolutional layer was selected to be reviewed and scaled to 224 x 224 with bilinear interpolation. VGG-16 model was evaluated on 500 sampled images, while ResNet-18 model was evaluated on 378 validation images. Results for them are shown in tables 3 and 4 respectively.

TABLE 3. Correlation coefficients and IoU metrics for LIME, SHAP, and Grad-CAM for the VGG-16 model for the IMAGENET dataset

Pearson statistic	LIME	SHAP	Grad-CAM
LIME	1	0.236707	0.386474
SHAP	0.236707	1	0.328904
Grad-CAM	0.386474	0.328904	1

IoU	LIME	SHAP	Grad-CAM
LIME	1	0.299562	0.357367
SHAP	0.299562	1	0.430439
Grad-CAM	0.357367	0.430439	1

TABLE 4. Correlation coefficients and IoU metrics for LIME, SHAP, and Grad-CAM for the Resnet-18 model on the Breast Cancer dataset

Pearson statistic	LIME	SHAP	Grad-CAM
LIME	1	0.078041	0.129545
SHAP	0.078041	1	0.125460
Grad-CAM	0.129545	0.125460	1

IoU	LIME	SHAP	Grad-CAM
LIME	1	0.088992	0.184091
SHAP	0.088992	1	0.125865
Grad-CAM	0.184091	0.125865	1

To get an intuition behind the numbers, fig. 3 visualizes samples and relative explanations given by each method.

Conclusion and discussion

In this study, we evaluated the similarity between image explanations generated by SHAP, LIME, and Grad-CAM using two different models trained for specific image classification tasks. The evaluation was performed on two datasets, with one fine tuned and one pre-trained model. The datasets were the CBIS-DDSM breast cancer dataset with fine tuned ResNet-18 model, and the Imagenet Object Classification Challenge (IOCC) with a VGG-16 pre-trained model. Our analysis revealed that while all of the methods aim to approximate feature importance, their outputs significantly differ, which makes it difficult to define the true reasoning of the model. Quantitative similarity metrics confirmed that these methods were most often independent, with less than half overlap on average. To add to that, metrics were also significantly different depending on the dataset or the model. The definition of what should be the ground truth or has the best practical use for these methods is complicated, as research contains both numerous variations of fidelity metrics and significantly varies in human-based evaluation perspectives. Future work can include evaluation of the impact of method parameters on the overlap, further investigation on the impact of the dataset and the selected model on the similarity, or quantitative comparison of the models with human-based metrics, such as comparing saliency maps with segmentation masks.

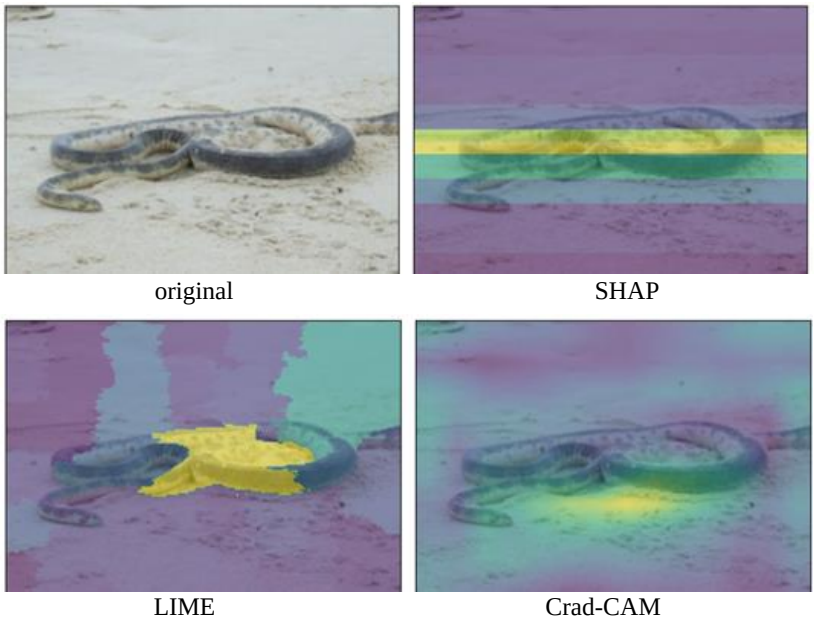

FIG. 3. Examples of visualizations produced by different algorithms. Picture in the upper left corner is the image used for the demonstration; bottom left is the original picture overlayed with heatmap produced by LIME; upper right is the image with SHAP heatmap; and bottom-left is image with Grad-CAM heatmap. Heatmaps show high importance with lighter colors, so yellow is the region with the highest importance, green colors indicate medium importance and dark red and purple – low importance

Author contributions. Vladyslav Yavtukhovskyi – collecting and visualizing the data, selecting neural network architectures, fine-tuning a model for classification, developing software for the experiment, calculating experimental metrics; Violeta Tretynyk – conceptual structure, data analysis, writing introduction and conclusion.

Data availability. Data used for the experiment is available with the following links:
https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset CBIS-DDSM;
https://www.kaggle.com/competitions/imagenet-object-localization-challenge CBIS-DDSM: Breast Cancer Image Dataset.

References

1. Rangel G., Cuevas-Tello J.C., Nunez-Varela J., Puente C., Silva-Trujillo A.G. A survey on convolutional neural networks and their performance limitations in image recognition tasks. *Journal of sensors*. 2024. 1. 2797320. https://doi.org/10.1155/2024/2797320
2. Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13*. P. 1135–1144. https://doi.org/10.1145/2939672.293977
3. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017. 30. https://doi.org/10.48550/arXiv.1705.07874
4. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision 2017*. P. 618–626. https://doi.org/10.1109/ICCV.2017.74
5. Hasan M.M. Understanding model predictions: a comparative analysis of SHAP and LIME on various ML algorithms. *Journal of Scientific and Technological Research*. 2023. **5** (1). P. 17–26. https://doi.org/10.59738/jstr.v5i1.23(17-26).eaqr5800

6. Paccotacya-Yanque R.Y., Bissoto A., Avila S. Are Explanations Helpful? A Comparative Analysis of Explainability Methods in Skin Lesion Classifiers. *In 2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM) 2024 Nov 13*. P. 1–5. https://doi.org/10.1109/SIPAIM62974.2024.10783606
7. Rezatofighi H., Tsoi N., Gwak J., Sadeghian A., Reid I., Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019*. P. 658–666. https://doi.org/10.1109/CVPR.2019.00075
8. Dzindolet M.T., Peterson S.A., Pomranky R.A., Pierce L.G., Beck H.P. The role of trust in automation reliance. *International journal of human-computer studies*. 2003. **58** (6). P. 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7
9. Shapley L.S. A value for n-person games. https://doi.org/10.1515/9781400881970-018 (accessed: 10.05.2025)
10. Adebayo J., Gilmer J., Muelly M., Goodfellow I., Hardt M., Kim B. Sanity checks for saliency maps. *Advances in neural information processing systems*. 2018. 31. https://doi.org/10.48550/arXiv.1810.03292
11. Lee R.S., Gimenez F., Hoogi A., Miyake K.K., Gorovoy M., Rubin D.L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*. 2017. **4** (1). P. 1–9. https://doi.org/10.1038/sdata.2017.177
12. Howard A., Park E., Kan W. Imagenet object localization challenge. Kaggle. 2018. https://www.kaggle.com/competitions/imagenet-object-localization-challenge (accessed: 10.05.2025)
13. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. https://doi.org/10.48550/arXiv.1409.1556 (accessed: 10.05.2025)
14. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *InProceedings of the IEEE conference on computer vision and pattern recognition*. 2016. P. 770–778. https://doi.org/10.1109/CVPR.2016.90
15. Deng J., Dong W., Socher R., Li L.J., Li K., Fei-Fei L. ImageNet: A large-scale hierarchical image database. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, Jun 20. P. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
16. Kingma D.P. Adam: A method for stochastic optimization. 2014. https://doi.org/10.48550/arXiv.1412.6980 (accessed: 10.05.2025)

Received 15.05.2025

Yavtukhovskyi Vladyslav,
Postgraduate Student at the Department of Applied Mathematics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
https://orcid.org/0009-0001-7668-386X
yavtukhovskyi.vladyslav@lll.kpi.ua

Tretynyk Violeta,
Candidate of Physical and Mathematical Sciences (Ph. D.), Docent at the Department of Applied Mathematics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.
https://orcid.org/0000-0002-3538-8207
viola.tret@gmail.com

UDC 519.67

Vladyslav Yavtukhovskyi, Violeta Tretynyk *

Evaluation of Similarity of Image Explanations Produced by SHAP, LIME and Grad-CAM

The National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
* *Correspondence: viola.tret@gmail.com*

Introduction. Convolutional neural networks (CNNs) are a subtype of neural networks developed specifically to work with images [1]. They have achieved great success both in research and in practical applications in recent years, however, one of the major pain points when adopting them is the lack of ability to interpret what is the reasoning behind their conclusion. Because of this, various explainable artificial intelligence (XAI) methods have been developed; however, it is unclear if they show reasoning or the same aspects of reasoning of CNNs. In recent years some of the most popular methods, LIME[2], SHAP[3], and Grad-CAM [4], were evaluated using tabular data and it was showed how significantly different results are [5] or some were evaluated on a matter of trustworthiness with human evaluation on medical images [6], there is still a lack of measure of how different these methods are on image classification models. This study uses correlation and a popular segmentation measure, Intersection over Union (IoU) [7], to evaluate their differences.

The purpose of the article. The aim of this work is to evaluate the level of differences between SHAP, LIME, and Grad-CAM on an image classification task.

Results. In this study, we evaluated the similarity between image explanations generated by SHAP, LIME, and Grad-CAM using two different models trained for specific image classification tasks. The evaluation was performed on two datasets, with one fine tuned and one pre-trained model. The datasets were the CBIS-DDSM breast cancer dataset with fine tuned ResNet-18 model, and the Imagenet Object Classification Challenge (IOCC) with a VGG-16 pre-trained model. Our analysis revealed that while all of the methods aim to approximate feature importance, their outputs significantly differ, which makes it difficult to define the true reasoning of the model. Quantitative similarity metrics confirmed that these methods were most often independent, with less than half overlap on average. To add to that, metrics were also significantly different depending on the dataset or the model. The definition of what should be the ground truth or has the best practical use for these methods is complicated, as research contains both numerous variations of fidelity metrics and significantly varies in human-based evaluation perspectives. Future work can include evaluation of the impact of method parameters on the overlap, further investigation on the impact of the dataset and the selected model on the similarity, or quantitative comparison of the models with human-based metrics, such as comparing saliency maps with segmentation masks.

Keywords: computer vision, convolutional neural network, Grad-CAM, LIME, SHAP, saliency maps, explainable AI, XAI.

УДК 519.67

В.С. Явтуховський, В.В. Третиник *

Оцінка схожості методів SHAP, LIME і Grad-CAM для пояснення зображень при класифікації

НТУУ «КПІ ім. Ігоря Сікорського»
** Листування: viola.tret@gmail.com*

Вступ. Згорткові нейронні мережі (CNN) – це підтип нейронних мереж, розроблений спеціально для роботи з зображеннями. Вони досягли значного успіху як у наукових дослідженнях, так і в практичних застосуваннях протягом останніх років. Проте одним із ключових викликів при їх впровадженні залишається відсутність можливості інтерпретувати логіку прийняття рішень моделлю. Через це були розроблені різні методи пояснюваного штучного інтелекту (XAI), однак досі залишається неясним, чи справді ці методи відображають логіку міркувань моделей CNN або принаймні ті самі її аспекти. У нещодавніх дослідженнях деякі з найпопулярніших методів – LIME, SHAP та Grad-CAM – були оцінені на табличних даних, де було показано суттєві відмінності між результатами, або ж аналізувалися з точки зору довіри користувача на основі медичних зображень. Водночас, все ще відсутня оцінка того, наскільки різняться ці методи при роботі із задачами класифікації зображень. У цій роботі ми застосовуємо кореляційний аналіз і популярну сегментаційну метрику – перетин над об'єднанням (Intersection over Union, IoU) – для оцінки їхніх відмінностей.

Мета роботи. Метою цього дослідження є оцінка рівня відмінностей між методами SHAP, LIME та Grad-CAM у задачі класифікації зображень.

Результати. У цьому дослідженні ми оцінили схожість візуальних пояснень, згенерованих методами SHAP, LIME та Grad-CAM, використовуючи дві різні моделі, натреновані на конкретні задачі класифікації зображень. Оцінювання проводилося на двох датасетах: CBIS-DDSM (датасет зображень раку молочної залози) із донавченим ResNet-18, та ImageNet Object Classification Challenge (IOCC) із попередньо натренованою VGG-16. Наш аналіз показав, що, попри спільну мету – апроксимацію важливості ознак, результати цих методів суттєво відрізняються, що ускладнює визначення справжньої логіки рішень моделі. Кількісні метрики схожості підтвердили, що ці методи найчастіше працюють незалежно, із середнім перекриттям менш ніж 50 %. Крім того, отримані значення метрик суттєво змінювалися залежно від використаного датасету або моделі. Визначення того, який із методів слід вважати «еталоном» або найбільш практичним, є складним завданням, адже наукові роботи пропонують численні варіації метрик відповідності, а також демонструють істотну варіативність у підходах до оцінювання з боку людини. Перспективними напрямами майбутніх досліджень можуть стати: вивчення впливу параметрів методів на ступінь перекриття, глибший аналіз впливу вибору датасету та архітектури моделі на схожість результатів, або ж кількісне порівняння із використанням метрик, заснованих на людській оцінці, наприклад, зіставлення карт важливості із сегментаційними масками.

Ключові слова: комп'ютерний зір, згорткові нейронні мережі, Grad-CAM, LIME, SHAP, мапи важливості, пояснювальний штучний інтелект, XAI.