©2025 Published in 11th International Symposium on Intelligent Technologies in Engineering and Science 16-18 May 2025 (ISITES2025 Diyarbakır - Turkey)

Classification of Dermatological Lesions Utilizing ProtoPNet and an Analysis Through Explainable Artificial Intelligence (XAI)

[1,2,*]Büruce ÖZTÜRK and [1,3]Mustafa Zahid YILDIZ
[1]Biomedical Technologies Application and Research Center (Biyotam), Sakarya University of Applied Sciences, Turkey
[2*]Graduate Education Institute, Department of Biomedical Engineering, Sakarya University of Applied Sciences, Turkey
[3]Faculty of Technology, Department of Electrical-Electronics Engineering, Sakarya University of Applied Sciences, Turkey

Abstract

In this study, an explainable deep learning model combining EfficientNetB0 and a Prototype Layer (ProtoPNet) was developed for the classification of skin lesions. The model assigned two prototypes per class to learn specific class features and employed both Grad-CAM and a novel Most Similar Patch method to explain decision mechanisms following prediction. During Most Similar Patch analysis, LIME contour maps were generated for patches with high edge density, whereas color histograms were used for regions characterized predominantly by color and texture. The model's ability to achieve both accurate classification and causal reasoning was evaluated, showing promising results in learning intra-class representations and providing meaningful explanations. Future work envisions extending this approach to more complex datasets such as histopathological images, enabling reliable interpretation of critical structures like tumor microenvironments. This hybrid framework is expected to contribute to the development of clinically trustworthy, explainable AI systems applicable to medical image analysis.

Keywords: Skin Lesion, XAI, Classification, Hybrid Models, Grad-CAM

1. Introduction

In recent years, deep learning (DL)-based artificial intelligence (AI) systems have achieved remarkable advancements in the field of medical image analysis, significantly improving both diagnostic accuracy and automation. Convolutional neural networks (CNNs) and transformer-based architectures are now widely used in clinical tasks such as dermatological lesion classification and histopathological tumor segmentation, often achieving accuracy exceeding 90% [1,2]. However, despite these impressive performances, such systems frequently operate as "black boxes," offering little to no insight into the reasoning behind their decisions. In healthcare, where decisions must be not only accurate but also ethically and clinically defensible, explainability has become an essential requirement. Explainable Artificial Intelligence (XAI) addresses this issue by making the decision-making process of AI models transparent and interpretable [3]. Techniques such as Grad-CAM++, SHAP, and LIME have been used to provide visual or numerical justifications for model predictions. For instance, in skin cancer classification, Grad-CAM++ can

*Büruce Öztürk: *Sakarya University of Applied Sciences, Graduate Education Institute, Biomedical Engineering Department, 54050,* Sakarya TURKEY. E-mail address: *brc.oztrk98@gmail.com*

highlight pixel regions most influential in diagnosing melanoma, while SHAP can quantify the contribution of color and texture patterns to a specific output [4,5]. Similarly, in histopathology, post-segmentation saliency maps and Grad-CAM++ are employed to visualize which morphological regions in a tissue slide influenced the classification.

Nonetheless, the majority of existing XAI studies are limited to surface-level visual explanations, falling short of providing causal, conceptual, or counterfactual reasoning. While methods like Grad-CAM++ indicate areas of attention, they do not determine whether these regions are causally responsible for the prediction. SHAP and LIME generate feature attribution scores, yet these often lack clinical interpretability or stability across repeated runs [6]. Furthermore, clinical studies have shown that such explanation maps are not always perceived as diagnostically relevant by physicians, particularly in low-grade tumor cases [7,8].

Notable studies further reveal these limitations. For instance, SmartSkin-XAI (2025) achieved high classification accuracy but provided only spatial attention without addressing why a particular region led to the decision [1]. Similarly, GRAPHITE (2025) introduced a graph-based attention model for breast cancer histology but lacked counterfactual analysis [2]. In NDG-CAM (2022), nucleus-level segmentation was effective, yet explanation maps focused solely on central regions, ignoring the diagnostic value of peripheral structures [4]. In hist2RNA (2023), a Transform-based model predicted gene expression from pathological images but failed to clarify the morphological basis of those predictions [5]. Even more concerning, explanation methods have been shown to produce inconsistent results across identical inputs, undermining trust in clinical settings [6]. These findings indicate that visual heatmaps or numerical feature scores are insufficient for clinical integration. XAI must evolve from isolated, post-hoc modules into systems embedded within the decision process, capable of causal, conceptual, and counterfactual explanations [9-11].

Motivated by these gaps, the present study proposes an explainable classification framework built upon ProtoPNet, a model designed to provide both visual and concept-based justifications [12,13]. ProtoPNet classifies inputs by comparing them to class-specific learned prototypes, enabling users to ask not only “What was predicted?” but also “Why this prediction?” Through this architecture, the study aims to evaluate the interpretability, clinical relevance, and conceptual alignment of ProtoPNet explanations in digital pathology, with a particular focus on skin and breast cancer diagnosis tasks [12].

2. Materials and Method

In this study, the publicly available ISIC 2018 dataset was utilized to perform both multi-class skin lesion classification and semantic segmentation [14]. The dataset includes a total of 9 diagnostic categories, each representing a distinct type of skin lesion with clinical relevance: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis / Intraepithelial Carcinoma (AKIEC), Pigmented Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), Squamous Cell Carcinoma (SCC), Seborrheic Keratosis (SK). As the class distribution

within the dataset was significantly imbalanced, a comprehensive data augmentation strategy was implemented to ensure that each class consisted of approximately 4500 images. Each original image was artificially augmented multiple times for underrepresented classes to achieve the target volume. For instance, seborrheic keratosis (SK) originally had 77 images, each of which was augmented 59 times; squamous cell carcinoma (SCC) images were augmented 25 times, basal cell carcinoma (BCC) 12 times, nevus (NV) 13 times, pigmented benign keratosis (BKL) 10 times, melanoma (MEL) 11 times, dermatofibroma (DF) 48 times, vascular lesion (VASC) 33 times, and actinic keratosis (AKIEC) 40 times. Augmentation techniques included horizontal and vertical flipping, random rotation, cropping, zoom, color jitter, and elastic deformation. These transformations ensured structural consistency while improving the model's generalizability, especially for rare lesion types. A prototype-based explainability mechanism was developed on top of an EfficientNetB0 backbone to enhance both the classification accuracy and interpretability of skin lesion diagnoses. Normalized 224×224 images were fed into the EfficientNetB0 model, where the first 150 layers were frozen to prevent overfitting during low-level feature extraction. A 1×1 Conv2D layer with 128 filters was added to the output, creating compact feature maps that were directly connected to a Prototype Layer, thus preserving spatial information for prototype matching. The Prototype Layer computed Euclidean distances between input feature maps and learnable prototype vectors, outputting minimum distance activations. A subsequent Dense layer with 64 units was introduced to strengthen prototype-to-class associations, followed by a softmax output layer predicting 9 classes. Prototype vectors were initialized using two representative samples per class, selected from the test set, thereby enhancing interpretability. The model's performance was evaluated on the training and validation sets using accuracy and categorical cross-entropy loss metrics. During training, early stopping was applied to monitor improvements in validation accuracy. For XAI analysis, Grad-CAM was used to visualize the most influential regions for each test sample, prototype activation maps were extracted, and the "Most Similar Patch" was identified on each image [15]. Whether these regions were edge- or color-based was analyzed using LIME explanations and color histograms. The quality of explanations was quantitatively assessed using normalized heatmap intensity and prototype similarity scores. This integrated approach achieved both high classification accuracy and interpretable decision-making processes.

2.1. Calculation

In skin lesions, the balance of classes before XAI analysis is important; almost every class should give an equal number of samples, thus preventing class imbalance. Here, there was a serious imbalance in the data at the beginning within 9 classes, while the number of samples of some classes was 1800, some were as few as 300. For this reason, a threshold value (4500) was determined while augmenting the classes, so that almost the same number of samples were prepared for XAI for each class. Fig. 1.A shows the numbers of classes after balancing. After the training process, the accuracy and loss values of the model were used to measure reliability based on the literature. The model was created to stop automatically at the highest accuracy value, based on early stopping at 50 epochs. This process is visualized in Figure 1.B.

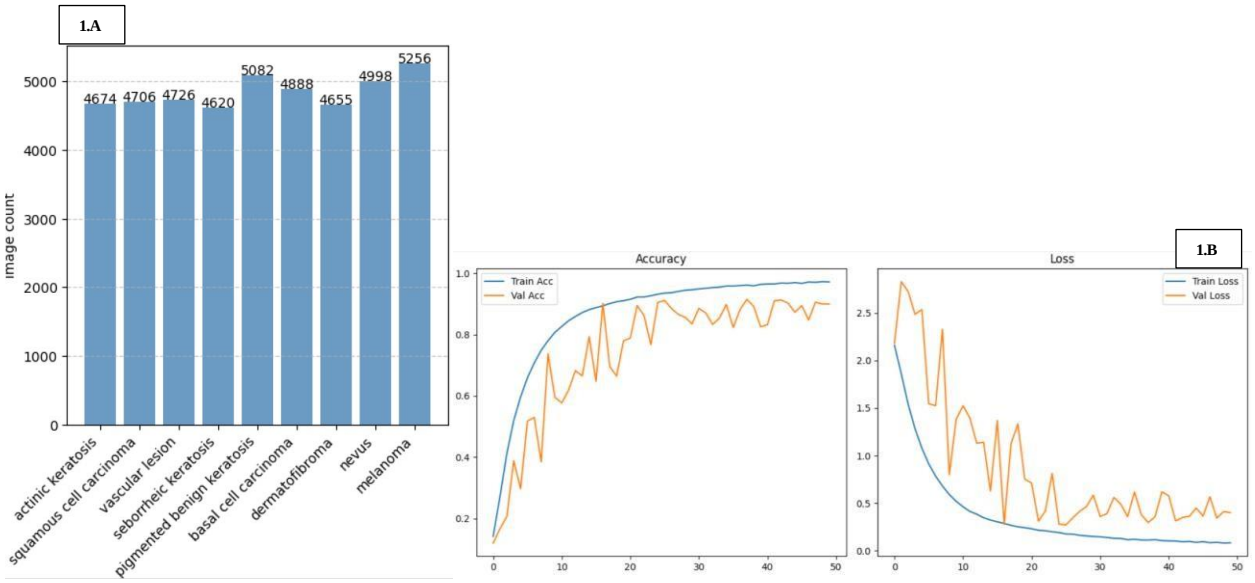

Figure 1.A Class balance after augmented data, and **1.B** Accuracy and loss function of the training process.

The model utilized feature maps extracted from the "block6a_expand_activation" layer, which were subsequently projected into a 128-dimensional feature space through a 1x1 Conv2D layer. To process these features, a custom-designed Prototype Layer containing two prototypes per class was introduced. The Prototype Layer calculated the Euclidean distances between the input features and the learned prototypes, routing the classification through the prototypes with minimum distances. Through this structure, the model was enabled to learn class-specific representative features. During training, extensive data augmentation techniques were applied, including random rotation, shifting, zooming, brightness adjustment, and horizontal/vertical flipping, to enhance the model's generalization capability. The Adam optimization algorithm was employed, categorical cross-entropy was selected as the loss function, and accuracy was used as the evaluation metric. Early Stopping was applied to prevent overfitting, and Model Checkpoint was used to retain the best performing model. To enhance explainability, Grad-CAM, the most widely used method in literature, was integrated into the model. Grad-CAM was utilized to visualize the critical regions that contributed most to the model's predictions through heatmaps. However, relying solely on Grad-CAM was deemed insufficient. Therefore, an additional explanation strategy, specific to the model's architecture, was introduced: Prototype Activation Maps. By extracting prototype-specific activation heatmaps, it became possible to directly visualize which regions activated particular prototypes most strongly, providing a more granular understanding of the model's behavior.

Beyond these, an original method not commonly found in the literature, termed "Most Similar Patch," was incorporated into the analysis. In this approach, for each test image, the model identified the patch that bore the highest similarity to one of its assigned prototypes. The model thus explained the rationale behind its prediction by extracting features from the prototypes assigned within each class. If the selected patch exhibited prominent edge structures, such as contours or depth variations, LIME was applied to generate edge contour explanations over the

patch. Conversely, if the patch lacked significant edges but displayed characteristic color or texture distributions, a color histogram was generated to represent the explanation.

This strategy allowed the model to autonomously adapt its explanatory approach based on the structural properties of each selected patch, generating class-specific, individualized explanations for each test instance. Consequently, a comprehensive explainability framework was established, integrating standard methods such as Grad-CAM alongside the novel Most Similar Patch mechanism, thus providing a deeper and more interpretable insight into the model's internal decision-making processes.

3. Results

In order to test the operability and reliability of the model and to measure its decision-making accuracy, it was asked it to assign one random example of a random class from the entire test file and to perform the analysis of this example. The first sample called from the test file came from the vascular lesion class. The model correctly predicted this class and tried to explain the predicted result by creating a heat map with Grad-CAM and Prototype Net. It argued that the reason for this explanation and prediction was the color scale of the lesions belonging to this class. The model saw that the lesions from its prototypes in this class were pink-light red and showed this color tone in a histogram graph as a result explanation. This situation is visualized in Fig.2.

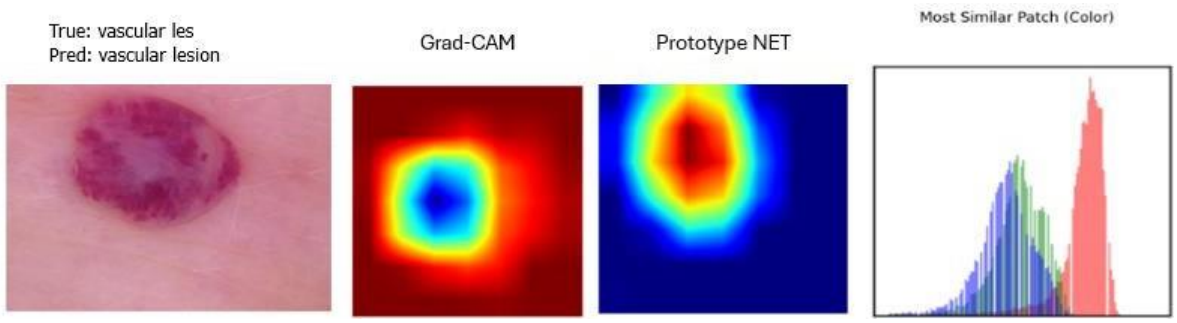

Figure 2. First Prediction of Trained Model

Upon examination of Fig. 2, it is evident that for the class of vascular lesions, the Grad-CAM technique generates a comprehensive density framework during the significant pixel analysis. In contrast, the prototype method yields a heat density map that highlights the most critical regions of the lesion. For Grad-cam, even with high accuracy, there seems to be an overflow into healthy tissue, and there are points where the important pixel does not correspond to the original image. Here, an explanation of the essential pixels that remain almost within the lesion borders for the prototype is presented.

From this point on, the codes were changed so that the model would randomly call a much more aggressive class, and a prediction was requested between the "nevus" and "melanoma" classes, whose color scales were almost similar, and an attempt was made to test whether the result was still based on the random color scale or based on the edge depth or size of the lesion. Figure 3 visualizes the stability of the model between aggressive classes.

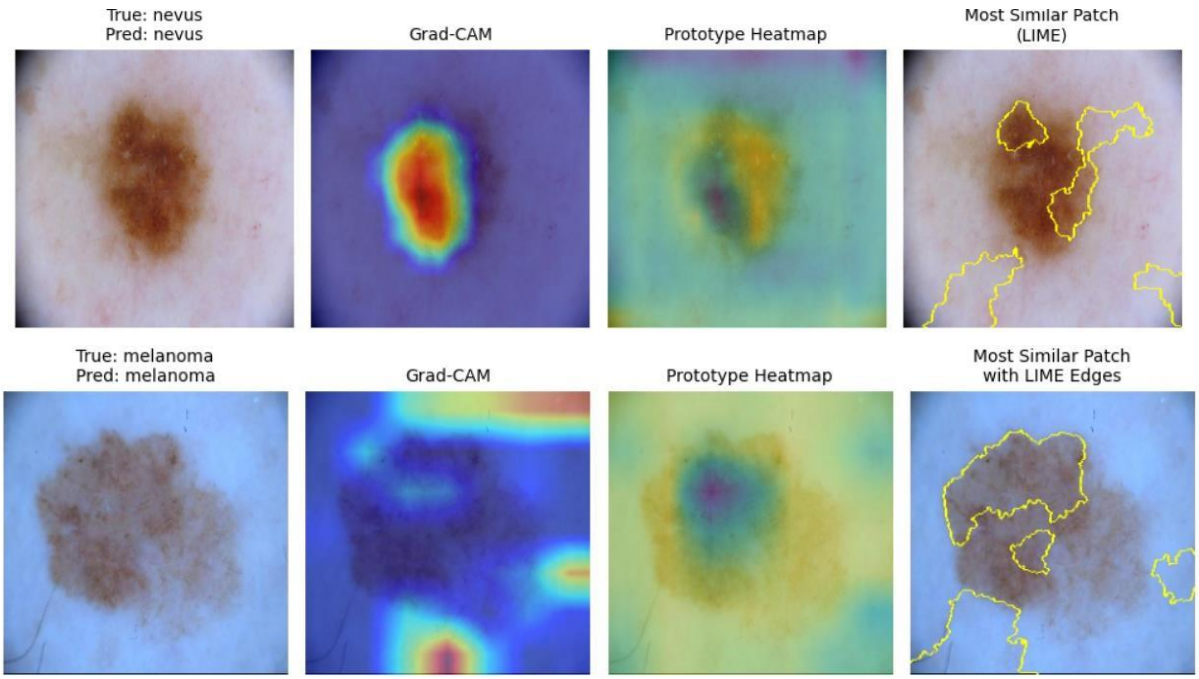

Figure 3. Testing the model for edge analysis between densely colored classes

Upon examination of Fig. 3, it is evident that one sample categorized within the nevus class has been accurately classified as nevus, while one sample within the melanoma class has been appropriately identified as melanoma. When the model is tested again with both Grad-cam and prototype to explain the classification result, it is seen that Grad-cam remains within the lesion borders for the nevus sample and is a bit weak in representing the entire lesion size. However, the prototype shows the whole lesion in yellow and shows that it can be important in general, especially since edge depth and size are distinctive within the class. As expected from the model, the edge depth and parts of other tissue surfaces in the class are indicated with the help of LIME as the reason for the prediction of the prototype heat map. In fact, it is argued that the tissue surfaces outside the lesion are also drawn, and the same situation is encountered in other samples belonging to the nevus class. It is seen that the Grad-cam heat map, which is traditionally used in the melanoma class, behaves quite inconsistently. These inconsistencies are expected for classes where the color tone and lesion sizes are not very distinctive. However, since the prototype model completed the process by specifically seeing the color tone and edge analysis of each class sample, even during the training process, it was able to mark the lesion area and the lesion edge area as yellow, that is, as an important pixel. It did not make the 'prediction by color' error among the aggressive classes, showing that the reason for this prediction result was again the edge features, as expected.

4. Discussion

This study proposes a novel hybrid Explainable AI (XAI) model designed to overcome the limitations inherent in traditional convolutional neural network (CNN)-based classification methods prevalent in current literature. This model is founded on the EfficientNetB0 architecture, seamlessly integrating prototype-based explanations into the learning process. The innovation of this approach resides in its dual capability: the model not only engages in classification tasks but also facilitates automatic, dynamic reasoning of its predictions at the point of decision-making, thereby eliminating the need for post-hoc explanations. In addition to employing the commonly utilized Grad-CAM style heatmaps, our model incorporates prototype representations for each class, enhancing the decision-making process's interpretability. This dual explanatory framework elucidates both the spatial (i.e., "where") and the contextual (i.e., "why") aspects of the model's predictions, grounded in texture and color similarity. Doing so advances the discourse on explainability in AI, providing a clearer understanding of the rationale behind model decisions.

Table 1. Comparison of this study with some current studies

Current Studies	Explanation Method	Type of Explanation	Timing of Explanation	Class-Specific Details	Contribution of This Study
SAB (2024) [16]	Attention Maps Grad-CAM	Global focus	Post-hoc	No	No explicit explanation per class
SHAP and LRP Comparison (2024) [8]	SHAP, LRP	Global contribution scores	Post-hoc	No	No patch-level, cause-specific visualization
Ladybug & LOCapsNet-CNN (2024) [17]	Channel-wise optimization Grad-CAM	Global explanation & Area Based	Post-hoc	Partial	Focused mainly on general accuracy optimization
This Study (2024)	**Prototype + Grad-CAM + LIME/Histogram**	**Cause-specific (color/edge explanation)**	**Integrated with prediction**	**Yes**	**Provides both class-level and cause-specific real-time explanations**

A thorough analysis of the comparative overview presented in Table 1 unequivocally demonstrates that prior methodologies leveraging Explainable Artificial Intelligence (XAI) have made substantial strides in the field. These approaches, reminiscent of robust frameworks established in

literature-based studies, have made a compelling impact by delivering insightful post-hoc explanations of model decisions and significantly enhancing the interpretability of deep learning systems. These methods have established a strong foundation for making AI decisions more transparent, especially by highlighting important regions and feature contributions. Building on this solid foundation, the model proposed in this study aims to expand interpretability by integrating real-time, cause-specific explanations directly into the prediction process. In addition to localizing decision regions, it provides a deeper understanding by distinguishing whether decisions are based on color textures or structural edges. By placing prototype assignments in the training phase, this model offers a complementary perspective that increases the transparency and clinical reliability of AI-driven medical diagnoses.

Conclusions

In recent years, there has been a remarkable increase in the number of studies based on deep learning (DL) and machine learning (ML) techniques, particularly in the domain of healthcare data analysis. However, in critical application areas such as healthcare, merely making predictions has proven insufficient; there has been a growing demand for systems that can also explain the underlying rationale behind their decisions. Consequently, explainable artificial intelligence (XAI) has emerged as a significant research area, yet most studies in the field have remained in early developmental stages. One of the main limitations has been the scarcity of systems capable of providing complete, causally grounded explanations for classified or segmented regions. To address this gap, the present study proposes a hybrid model designed for a 9-class skin lesion dataset. The model was constructed by combining a powerful deep learning backbone, EfficientNetB0, with a specially designed Prototype Layer, assigning two prototypes to each class to facilitate the learning of class-specific distinctive features. Following prediction, a two-stage explainability system was integrated: firstly, Grad-CAM, the most widely used method in the literature, was employed to generate global heatmaps; secondly, a novel Most Similar Patch method, uniquely developed for this study, was applied to analyze the local regions most influenced by the selected prototype. During the Most Similar Patch analysis, if the selected patch exhibited high edge density, LIME was used to generate contour-based explanations, whereas if the patch was characterized predominantly by color or texture distribution, a color histogram was generated to support the explanation. The results demonstrated that the model was able to accurately predict the correct class among multiple categories and effectively learn the specific internal features of each class. Moreover, the model not only provided accurate predictions but also offered causal explanations for its decisions, representing a significant advancement beyond the early-stage XAI efforts commonly found in the literature. Looking forward, this approach is envisioned to extend beyond surface datasets such as skin lesions, toward application in complex histopathological datasets involving cellular structures, enabling models to infer and explain tumor, tumor microenvironment, and lesion boundaries through deeper causal relationships. Additionally, it is anticipated that by allowing internal architectural interventions within models, highly reliable clinical explanations and interpretations could be achieved. Although only accuracy and loss were utilized as evaluation metrics in this study for comparative purposes, future work will incorporate more complex evaluation metrics such as Dice coefficient, micro/macro-ROC AUC. Moreover, advanced optimization techniques, including cosine annealing, learning rate warmup, and

adversarial training, are planned to be employed to enhance both classification and explainability performance further. In this direction, the methods developed are expected to lay a strong foundation for the integration of explainable AI systems into clinical research and real-world deployment scenarios.

References

[1] Hamim SA, Tamim MUI, Mridha MF, Safran M, Che D. SmartSkin-XAI: An interpretable deep learning approach for enhanced skin cancer diagnosis in smart healthcare. *Diagnostics* 2025;15:64. https://doi.org/10.3390/diagnostics15010064.

[2] Mondol RK, Millar EKA, Graham PH, Browne L, Sowmya A, Meijering E. GRAPHite: Graph-based interpretable tissue examination for enhanced explainability in breast cancer histopathology. *arXiv preprint* 2025;arXiv:2501.04206v1. https://doi.org/10.48550/arXiv.2501.04206.

[3] Donmez TB, Kutlu M, Mansour M, Yildiz MZ. Explainable AI in action: a comparative analysis of hypertension risk factors using SHAP and LIME. *Neural Comput Appl* 2025;37:4053–4074. https://doi.org/10.1007/s00521-024-10724-y.

[4] Altini N, Brunetti A, Puro E, Taccogna MG, Saponaro C, Zito FA, De Summa S, Bevilacqua V. NDG-CAM: Nuclei detection in histopathology images with semantic segmentation networks and Grad-CAM. *Bioengineering* 2022;9:475. https://doi.org/10.3390/bioengineering9090475.

[5] Mondol RK, Millar EKA, Graham PH, Browne L, Sowmya A, Meijering E. hist2RNA: An efficient deep learning model to predict gene expression from breast cancer histopathology images. *Cancers* 2023;15:2569. https://doi.org/10.3390/cancers15092569.

[6] Chiaburu T, Bießmann F, Haußer F. Uncertainty propagation in XAI: A comparison of analytical and empirical estimators. *arXiv preprint* 2025;arXiv:2504.03736v1. https://arxiv.org/abs/2504.03736.

[7] Manz R, Bäcker J, Cramer S, Meyer P, Müller D, Muzalyova A, et al. Do explainable AI (XAI) methods improve the acceptance of AI in clinical practice? An evaluation of XAI methods on Gleason grading. *J Pathol Clin Res* 2025;11:e70023. https://doi.org/10.1002/2056-4538.70023.

[8] Özkurt C. Advancing skin cancer diagnosis through the comparison of SHAP and Layer-wise Relevance Propagation (LRP). *Research Square* 2024. https://doi.org/10.21203/rs.3.rs-3920847/v1.

[9] Slack D, Hilgard S, Singh S, Lakkaraju H. Reliable post hoc explanations: Modeling uncertainty in explainability. *arXiv preprint* 2021;arXiv:2008.05030. http://arxiv.org/abs/2008.05030.

[10] Evans T, Retzlaff CO, Geißler C, et al. The explainability paradox: challenges for XAI in

digital pathology. *Future Gener Comput Syst* 2022;133:281–296.

[11] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Trans Neural Netw Learn Syst* 2020;32(11):4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314.

[12] Ribeiro MT, 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. p. 1135–1144.https://doi.org/10.1145/2939672.2939778.

[13] Ribera M, Lapedriza A. Can we do better explanations? A proposal of User-Centered Explainable AI. In: *Joint Proceedings of the ACM IUI 2019 Workshops*; 2019; Los Angeles, USA. ACM: New York, NY. http://hdl.handle.net/10609/99643.

[14] Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2018 ISIC dermoscopic image analysis workshop. *arXiv preprint* 2018;arXiv:1902.03368. https://doi.org/10.48550/arXiv.1902.03368 (accessed Dec 26, 2024).

[15] Selvaraju RR, 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 (ICCV)*; 2017. p. 618–626. https://doi.org/10.1109/ICCV.2017.74.

[16] Nasir IM, Tehsin S, Damaševičius R, Maskeliūnas R. Integrating explanations into CNNs by adopting spiking attention block for skin cancer detection. *Algorithms* 2024;17:557. https://doi.org/10.3390/a17120557.

[17] Pramila RP, Subhashini R. Multi-center validation of ladybug beetle optimized convolutional capsule neural networks with explainable AI for skin cancer classification using dermography images. *Afr J Biomed Res* 2024;27:972–989. https://doi.org/10.53555/AJBR.v27i3.3246.