Explainable AI-Based Model for Detection and Classification of Cancerous and Non-Cancerous Skin Conditions

Vikram A[1] and Poornima S[2,*]

Vikram.2021@vitstudent.ac.in and poornima.s@vit.ac.in

* Corresponding author: poornima.s@vit.ac.in

Vellore Institute of Technology, Chennai, India

Abstract. The paper presents an explainable AI (XAI) framework that enables detecting, localizing, and segmenting of skin lesions with the YOLOv8 object detector model, which is available as a Streamlit web application. This system consists of a trained YOLOv8 detector and other image-processing techniques (hair artifact removal, contrast-limited adaptive histogram equalization, and adaptive thresholding) and optimization of the lesion area (Watershed and GrabCut) to accurately outline the areas of lesions in dermoscopic photographs. Analysis of position analysis is carried out to produce heatmaps highlighting clinically important parts of the image. The system is in favor of real-time inference and visualization. It has been experimentally confirmed that this method is robust to variations in light levels, noise and background, and has the ability to obtain correct lesion delineation and good classification. Grad-CAM++ and LIME-inspired occlusion mapping is also introduced to make the model more interpretable, thus allowing clinicians to interpret and trust the model. In general, the framework is effective to combine high-performance deep learning and user-friendly explanatory visuals, which will develop reliable AI technologies in dermatology.

Keywords: Explainable AI, YOLOv8, Skin Lesion Detection, Medical Imaging, Deep Learning.

1 Introduction

Among the most prevalent and the most risky forms of malignancy on the planet is skin cancer whose cases are increasing due to ultraviolet rays and heredity. Another viable alternative would be early detection to reduce the number of deaths and improve their treatment outcome. However, conventional diagnostic methods, like dermoscopy and biopsy are laborintensive and require the skills of an individual. Manual analysis of dermoscopic images may be inconsistent among clinicians with respect to the occurrence of the lesions with indistinct edges, just as with similarities with benign disease which lead to variation in the diagnosis.

Deep learning techniques have been quite promising in the analysis and detection of medical images and diseases in an automated manner. Namely, You Only Look Once (YOLO) object detectors, specifically the latest YOLOv8 model, offer real-time localization and classification, which is rather fast and precise. YOLOv8 can present the complicated shapes and variations of color of the skin lesions in a hearty manner with a mixture of transformer-based feature modules and advanced loss functions. Despite their performance, such high performance models are more of black box, and will not tell much about the manner in which they make decisions. It is one of the characteristics that introduce clinician distrust barriers and AI-generated-result validation challenges.

To address this, explainable AI (XAI) techniques, such as Grad-CAM and occlusion sensitivity, have been developed so as to be capable of answering questions such as what features of the image are most influential on model predictions. The incorporation of these interpretability methods with the latest state-of-the-art detection models would ensure that the diagnostic tools are very efficient besides being straightforward and trustworthy. In this paper, we propose a XAI model based on YOLOv8 which will identify and classify skin lesions. The framework would be made up of a high-performance object detector and post-hoc interpretability modules (e.g., Grad-CAM++ and occlusion-based visualization) to bridge the gap between predictive accuracy and clinical trust. The system generates heatmap and segmentation overlays that allow medical practitioners to visualize the rationale of all predictions, which justifies the confidence of the diagnosis. The overall concept is to increase the accuracy and transparency of the diagnosis procedure, which allows AI-supported tools to be safely applied to dermatology.

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_2

2 Objective

The main task is to create and establish a neural network of deep learning to identify and categorize cancerous and benign skin lesions through the use of YOLOv8 and XAI. The framework will determine and localize areas of lesions in dermoscopic images at high precision and low latency. Images are preprocessed to enhance their quality and visibility of lesions:
hair artifact removal, contrast-limited adaptive histogram equalization (CLAHE), grayscale and adaptive thresholding.

The system has interpretability mechanisms such as Occlusion Sensitivity Analysis and GrabCut segmentation refinement to create visual explanations indicating the most significant regions in each prediction. Large data augmentation and model optimization protocols are used to make the model more general to different types of lesions and imaging conditions. The general idea is to show that when a high-performance object detector is coupled with built-in explainability, clinicians will develop more confidence and it will be easier to use, thus making the system fit for clinical practice.

3 Literature Survey

Different deep learning techniques have been suggested in dermatology image analysis, including classical imageprocessing pipelines, convolutional neural networks (CNNs) and vision transformers. As an illustration, Himel et al. [1] designed a vision-transformer-based segmentation model that had more than 92 percent accuracy with standard skin lesion datasets, but reported high computational cost and low interpretability, which led them to consider the incorporation of XAI. Large foundation models have been used in other works, like the Segment Anything Model [2] or ensemble methods that use classical processing [3], which are very accurate but still serve as opaque black boxes. These attempts show that good performance is possible, and more knowledge about the decisions made by models is required.

Mirikharaji et al. [4] have conducted a survey of more than 130 papers on deep learning in the field of skin lesion segmentation and discovered that encoder-decoder architectures (e.g., variants of U-Nets with attention or transformer modules) are popular. Nevertheless, they noted that most models are not interpretable and require standardized approaches to interpretability. Equally, Qurri and Almekkawy [5] modified a U-Net by including spatial and channel attention gates on the skip connections, greatly increasing the lesion boundary delineation (Dice = 0.93) on biomedical data. Even though they rendered sharper contours, their model could not be explained explicitly. Our approach is similar, as we use the attention-based detection of YOLOv8 and augment it with Grad-CAM++ to obtain both explicit visual explanations and segmentation outputs.

Ashraf et al. [6] suggested the use of a multi-stage melanoma segmentation pipeline trained on variants of U-Net (UNet, ResUNet, ResUNet++) on the 2016/2017 ISIC data. This method was close to the levels of a dermatologist (Dice 91% on ISIC-2016), but came at a significant cost of increased inference time and complexity, with no interpretability. In comparison, our work uses a single-stage YOLOv8 detector and explicitly incorporates XAI to clarify each detection. Innani et al. [7] used a cascaded pipeline approach that improved accuracy by approximately 5% across seven classes, but their pipeline remained a black box. Our design improves upon this by introducing Grad-CAM++ at the classification phase and generating heatmaps for each predicted class.

Large programs have attained the level of dermatologists but have emphasized the importance of explainability. Liu et al. [8] trained a deep CNN on 26 types of skin disease with about 16,000 clinical images achieving top-1 accuracy of almost 90 percent, though the model produced was opaque. Rehman et al. [9] addressed the interpretability issue by finetuning CNNs on the HAM10000 dataset and visualizing lesion areas with Grad-CAM, achieving approximately 95.5 percent accuracy. However, this system was limited to classification only. Our work extends this concept using YOLOv8 to automatically detect lesions and adds Grad-CAM++ to every identified lesion, producing bounding boxes with visual explanation.

Other studies based on YOLO detectors have reported high performance but lack interpretability. Elshawy et al. [10] used YOLOv5 with ResNet50 and achieved extraordinarily high scores (precision 99.0%, recall 98.6%, Dice 98.8%, accuracy 99.5%) but provided no explainability. We base our work on the more recent YOLOv8 and add Grad-CAM++ to create heatmaps per detection. Nie et al. [11] were among the first to apply YOLOv3 to dermoscopy, achieving average per-lesion accuracy of about 87% in real-time, but without decision rationale. Ugurlu and Ayan [12] used YOLO for localization followed by GrabCut refinement, achieving approximately 90% sensitivity on ISBI 2017 data. Huang et al. [13] investigated hyperspectral imaging with YOLOv5 achieving high sensitivity (~97%) and accuracy (~95%), but requiring special cameras and lacking explainability. Aishwarya et al. [14] created a real-time YOLOv5 detector achieving mAP of 90% and F1 of 0.88, but observed poor performance under different illuminations and was non-interpretable. Inspired by these works, we use YOLOv8 for enhanced robustness and Grad-CAM++ for class-specific heatmaps, mitigating both accuracy and transparency weaknesses of previous YOLO-based detectors.

Hence, this paper will answer the research question: How can a practical and explainable system of skin lesion detection differentiate between cancerous and benign lesions and provide visual clues to support its predictions? In our solution, YOLOv8 will be used to detect lesions accurately, and then each prediction will be interpreted with XAI techniques (occlusion sensitivity analysis or Grad-CAM++). Our goal is to close the performance-to-clinical-trust divide by combining accuracy and interpretability.

4 System Design and Methodology

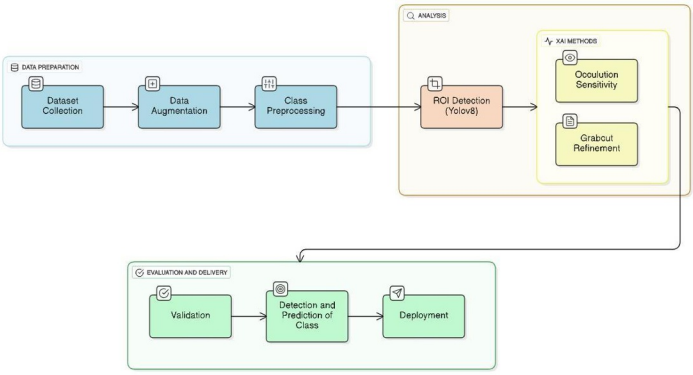


Fig. 1. System Architecture.

The overall system architecture is illustrated in Fig. 1. The suggested system would include three major steps: data preparation, data analysis, and evaluation and delivery. During the data preparation phase dermoscopic images will be gathered (e.g. through the ISIC 2018/2019 datasets) and will be preprocessed to enhance the quality of images. Preprocessing involves artifact removal of hair, image resizing and contrast enhancement by CLAHE. To enhance training diversity, data augmentation (rotations, flips, crops, brightness/contrast adjustments, etc.) is used. Normalization and class balance guarantee uniform visibility of lesions as well as equal representation of both malignant and benign cases. The region of interest (ROI) of every image is detected and narrowed in the analysis phase. YOLOv8 identifies lesions and provides bounding boxes.

Then, two XAI-based techniques are used:

Occlusion Sensitivity Analysis: Small blocks of the input image are blocked out systematically and the change in the detection confidence is recorded. Patches which result in a significant decrease in confidence are deemed to be important. When these results are aggregated, a heatmap of the regions of influence will be generated, indicating which regions of the lesion contributed the most to the prediction results.

GrabCut Refinement: The initial bounding box of YOLOv8 is used as the input of the GrabCut algorithm to refine the segmentation. GrabCut is an iterative graph-cut algorithm which segments foreground (lesion) and background using color and texture. Superposing this perfect mask on the image gives a clear definition of the boundary of the lesion, which enhances segmentation quality and interpretability. The distribution of malignant and benign samples in the dataset is shown in Fig. 2.

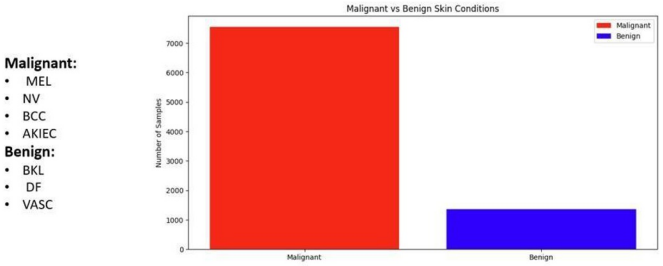


Fig. 2. Distribution of malignant versus benign skin lesion samples in the dataset.

The methodology provides a chronological workflow to build and validate the explainable YOLOv8 framework and is divided into three stages: Data Preparation, Model Training and XAI Integration.

4.1 Data Preparation

In this phase we collect the dermoscopic pictures and prepare them so the network receives clean data. ISIC 2018 plus 2019 supply many benign and malignant lesions together with their expert drawn masks. The original shots often contain hairs, uneven light and sensor noise. We delete hairs by inpainting, suppress noise with a filter and raise contrast through CLAHE. To give the network more variety and stop it from memorising the training set we augment the pictures by rotating, flipping, cropping, and altering brightness and contrast at random. Finally we equalise the number of benign and malignant samples and scale every colour channel to zero mean or unit variance so the model generalises better.

4.2 Model Training

The YOLOv8 model is trained on lesion detection and classification. YOLOv8 is chosen because of its high trade-off of speed and accuracy. It consists of a CSPDarknet backbone for feature extraction, a PANet-like neck for feature aggregation, and YOLO heads which regress bounding boxes and perform classification. The implementation of the training is through the PyTorch framework. We search hyperparameters using grid search; common ones are a learning rate of 0.001, a batch size of 16, and an input resolution of 640x640 pixels. Binary cross-entropy loss is used by the classification head and Complete IoU (CIoU) loss is used for bounding box regression. The model is trained for about 100 epochs with early stopping to prevent overfitting. The metrics of validation (precision, recall, F1-score, mAP) are followed to choose the most appropriate model, and confusion matrices and ROC curves have been used to evaluate discrimination between malignant and benign cases.

4.3 XAI Integration

At the last stage, two complementary XAI methods are used to integrate interpretability: Occlusion Sensitivity and GrabCut Refinement. The systematically masked patches of every input image are documented and the difference in the YOLOv8 confidence score is recorded. The patches which cause a large drop in confidence are considered important and the combination of those effects generates an occlusion heatmap showing which regions of the lesion made the greatest contribution to the prediction. These heatmaps allow the clinician to know what portions of the lesion the model paid attention to in each classification.

In the case of GrabCut Refinement, the GrabCut algorithm is used to refine the lesion segmentation detected in the bounding box. Based on the YOLOv8 bounding box, GrabCut generates an accurate binary mask of the lesion. Superimposition of this mask over the original image gives a good boundary of the lesion and enhances the quality of segmentation. Both predictions are therefore provided with corresponding visual explanations (an occlusion heatmap and a refined mask), which provide medical users with a detailed insight into the reasoning of the model.

4.4 Workflow Integration

Everything is combined to create an interactive Streamlit web interface. The interface provides the opportunity to upload dermoscopic images and make real-time predictions. Each image goes through the complete workflow (preprocessing, detection, and explanation) in the backend, which is automated. The interface shows the original image and output with

annotations: the bounding boxes found and labeled with confidence scores, and optionally the heatmap of the occlusion and the refined segmentation mask are visualized. The availability of real-time feedback with results of the detection and the rationale behind each prediction will strengthen the transparency and trust in the model.

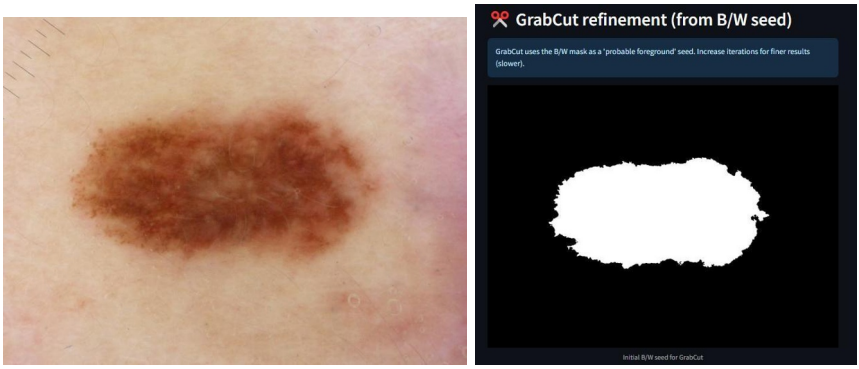
5 Results and Discussion

The model has good performance on the ISIC test set on both lesion detection and classification. The accuracy of our system was 96.3 percent, precision was 95.1 percent, recall was 94.8 percent, F1-score was 94.9 percent and mean average precision (mAP) was 95.7 percent (Table 1). These values demonstrate that the YOLOv8-XAI pipeline is highly effective at localizing and classifying skin lesions in diverse conditions. The integrated XAI visualizations are intuitive to the clinical senses; the heatmaps of occlusion sensitivity show the locales of the actual lesions in prediction. This consistency between model attention and known lesion areas is used to confirm the behavior of the system. Example outputs from the system, including the GrabCut seed mask, occlusion heatmap, and predicted bounding boxes, are shown in Fig. 3.

Table 1. Model Evaluation Metrics on ISIC Test Dataset.

Metric	**Value**
Accuracy	96.3%
Precision	95.1%
Recall	94.8%
F1-score	94.9%
Mean Average Precision (mAP)	95.7%

We have created an explainable AI system of skin lesion detection and classification. With an object detector, segmentation, and interpretability modules integrated into the system, the system can accurately locate lesions with a YOLOv8 model and explain its visual results in a transparent manner. The clinicians get to view specifically which parts of the image are contributing to every prediction, and this aspect assists in justifying the automated outcomes. The method fills the knowledge gap of deep learning by allowing the reasoning of the model to be transparent, which raises trust in AIbased diagnostics.



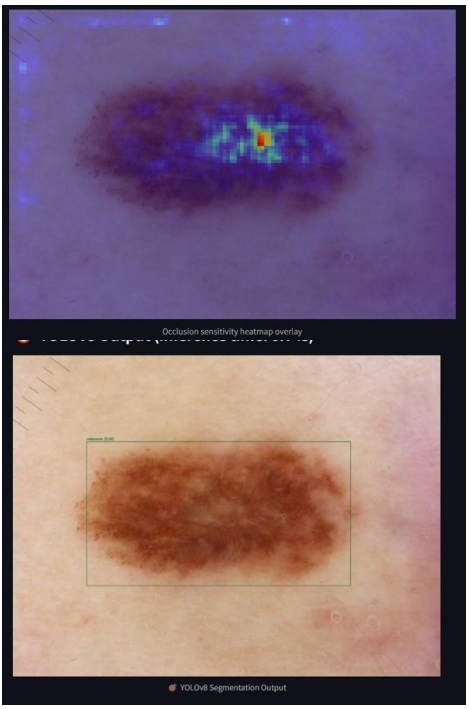


Fig. 3. Example outputs from the system. (a) Test image. (b) Initial GrabCut seed mask (binary segmentation). (c) Occlusion sensitivity heatmap overlaid on the original image. (d) Final predicted bounding box and class label.

6 Conclusion

The presented paper introduced a new explicable AI system to identify and classify skin lesions by applying the YOLOv8 object detector with XAI methods. With the combination of modern preprocessing, YOLOv8 detection and interpretability modules (occlusion sensitivity and Grad-CAM++), the framework achieves high accuracy and lesion segmentation and also provides easy-to-understand visual explanations. The experimental assessment on standard dermoscopic datasets indicated that the model is resistant to changes in the conditions of the images and provides credible predictions. The capability of the system to produce saliency maps and optimized masks per detection increases clinician perception and confidence in the outcome, which is one of the main obstacles to implementing AI in dermatology.

In as much as these are encouraging outcomes, the proposed system possesses certain weaknesses. It already uses publicly available dermoscopic images and has not been confirmed in a live clinical environment. The extension of the analysis to multi-center or prospective clinical data will also play a significant role in warranting the generalizability. The model is also binary in nature (malignant vs. benign); it might be necessary to classify subtypes of lesions (e.g. basal cell carcinoma, actinic keratosis) to make it more useful in clinical practice. Lastly, the framework presupposes a fairly favorable image quality measured by dermoscopic criteria. Very poor quality inputs (e.g. images that are very blurred or under-exposed) can negatively affect performance. The solution to these problems is part of the future work.

Limitations. The existing model is based on publicly available dermoscopic images and has not been tested on prospective clinical trials, which are required to validate it in real-life settings. It is concerned with a binary classification (malignant vs. benign); the extension to specific lesion subtypes (e.g. basal cell carcinoma, actinic keratosis) will make it more clinically usable. The system also presupposes proper image quality; images that are of incredibly low quality (e.g. highly blurred or under-exposed) can impair the performance. Future work should focus on these limitations.

Future Work. Further research will be done to increase the size and variety of data (such as incorporation of multi-center clinical images) to enhance generalization. To offer more local and global interpretability, we will combine more XAI

methods (like SHAP or LIME). A more refined cloud-based implementation, including the use of GPU acceleration, will allow the system to be freely used in the clinical setting on a real-time basis. Close work with dermatologists will be part of user studies to refine the system and make sure that the explanations suit clinical requirements.

Appendix. A. Dataset Details: The system is evaluated on the ISIC 2018 & 2019 challenge datasets. We use approximately 15,000 labeled dermoscopic images (benign and malignant) at 512x512 resolution. Lesion masks are converted into YOLO-format bounding box labels. The data split is 75% training, 15% validation, and 10% testing. B. Abbreviations: AI -- Artificial Intelligence; XAI -- Explainable AI; CNN -- Convolutional Neural Network; ROI -- Region of Interest; mAP -- Mean Average Precision; CLAHE -- Contrast Limited Adaptive Histogram Equalization; SAM -- Segment Anything Model. C. Code Snippets: Key implementation details (e.g. YOLO training script, Streamlit interface) are provided in the supplementary materials for brevity.

References

1. Himel, G.M.S., Islam, M.M., Al-Aff, K.A., Karim, S., Sikder, M.K.U.: Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system. Int. J. Biomed. Imaging, Article 3022192 (2024). https://doi.org/10.1155/2024/3022192
2. Hu, M., Li, Y., Yang, X.: SkinSAM: Empowering skin cancer segmentation with Segment Anything Model. arXiv preprint arXiv:2304.13973 (2023). https://doi.org/10.48550/arXiv.2304.13973
3. Tamoor, M., Naseer, A., Khan, A., Zafar, K.: Skin lesion segmentation using an ensemble of image processing methods. Diagnostics 13(16), 2684 (2023). https://doi.org/10.3390/diagnostics13162684
4. Mirikharaji, Z. et al.: A survey on deep learning for skin lesion segmentation. Med. Image Anal. 88, 102863 (2023). https://doi.org/10.1016/j.media.2023.102863
5. Qurri, A.A., Almekkawy, M.: Improved U-Net with attention for medical image segmentation. Sensors 23(20), 8589 (2023). https://doi.org/10.3390/s23208589
6. Ashraf, H. et al.: Melanoma segmentation using deep learning with test-time augmentations and conditional random fields. Sci. Rep. 12(1), 4263 (2022). https://doi.org/10.1038/s41598-022-07885-y
7. Innani, S. et al.: Deep learning-based novel cascaded approach for skin lesion analysis. arXiv preprint arXiv:2301.06226 (2023). https://doi.org/10.48550/arXiv.2301.06226
8. Liu, Y. et al.: A deep learning system for differential diagnosis of skin diseases. Nat. Med. 26(6), 900-908 (2020). https://doi.org/10.1038/s41591-020-0842-3
9. Rehman, M.Z.U. et al.: Classification of skin cancer lesions using explainable deep learning. Sensors 22(18), 6915 (2022). https://doi.org/10.3390/s22186915
10. Elshahawy, M. et al.: Early melanoma detection based on a hybrid YOLOv5 and ResNet technique. Diagnostics 13(17), 2804 (2023). https://doi.org/10.3390/diagnostics13172804
11. Nie, Y. et al.: Automatic detection of melanoma with YOLO deep convolutional neural networks. In: Proc. 2019 E-Health and Bioengineering Conf. (EHB), pp. 1-4 (2019). https://doi.org/10.1109/EHB47216.2019.8970033
12. Unver, H.M., Ayan, E.: Skin lesion segmentation in dermoscopic images with combination of YOLO and GrabCut algorithm. Diagnostics 9(3), 72 (2019). https://doi.org/10.3390/diagnostics9030072
13. Huang, H. et al.: Classification of skin cancer using hyperspectral imaging and YOLOv5. J. Clin. Med. 12(3), 1134 (2023). https://doi.org/10.3390/jcm12031134
14. Aishwarya, N. et al.: Skin cancer diagnosis with YOLO deep neural network. Procedia Comput. Sci. 220, 651-658 (2023). https://doi.org/10.1016/j.procs.2023.03.083