International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 105-112
https://doi.org/10.70917/ijcisim-2026-2068
Article

EXPLAINABLE AI FRAMEWORK FOR MELANOMA DETECTION USING PRE TRAINED CNN FEATURE EXTRACTION AND LOCAL INTERPRETABLE MODEL EXPLANATIONS

Riyazahemed A. Jamadar[1], Prashant Wakhare[2], Sanjay Bhilegaonkar[3], Pritesh Patil[4]*

[1]Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune-01, Maharashtra, India. Email: reeyaj.jamaddar@gmail.com
[2]Department of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune-01, Maharashtra, India. Email: pbwakhare@gmail.com
[3]Institute of Electrical and Electronics Engineers (IEEE). Email: bsanjayht@gmail.com
[4]Department of Information Technology, AISSMS Institute of Information Technology, Pune-01, Maharashtra, India. Email: p.patil.k@gmail.com
Corresponding Author: Prashant Wakhare[2]* (Email: pbwakhare@gmail.com)

Abstract: Melanoma is a deadly skin cancer, the early detection of which can significantly increase its survival, but the manual diagnosis might be subjective and time-consuming. The paper is a proposal to use a combination of feature representations of a trained convolutional neural network (CNN) and local interpretable model explanations (LIME) to categorize dermoscopic images and generate patient-specific pictorial explanations. The research makes use of the benchmark of SIIM-ISIC 2020 Kaggle dataset consisting of 33 126 training images and 10 982 test images of benign and malignant lesions. Two preprocess measures take away artefacts and contrast adjustments and a ResNet-50 network pre-trained on ImageNet extracts deep features. A linear classifier is used to identify melanoma based on these features, and LIME constructs a local surrogate model of each image to suggest parts of the image that are used to reach the decision. Findings indicate that the proposed technique has a precision of 92.7, recall of 94.2, accuracy of 93.4 and AUC of 97.3. We also measure the quality of explanation and confusion matrix values and present the strong and weak points of the framework. This method is a feasible method of integrating high performance classification and interpretable evidence. It is an indicator of directions to come in enhancing fairness and generalization.

Keywords: melanoma detection, pre trained CNN feature extraction, local interpretable model explanations, dermoscopy, explainability...

1. Introduction

Skin cancer is very prevalent all over the globe and melanoma is one of the deadliest types of skin cancer. Though incidence differs among populations, the world health organization has documented an increase in the cases in the world. Early diagnosis will reduce mortality since thin lesions excised will result in high survival rates. Visual inspection by the dermatologists and dermoscopy experts is however subjective and influenced by fatigue, training and observer variations. Clinicians working in a busy clinic should be able to make a rapid decision on whether to biopsy a lesion or not and failure to detect early melanoma may be disastrous. The aim of computer-aided diagnosis systems is to aid clinicians by analysing images and predicting malignancy, however, their use has not adopted massively yet, owing to many of these systems being opaque; physicians might not be willing to trust black-box

predictions. Interpretability assists in developing trust since it allows experts to determine whether the algorithm uses the clinically significant features to derive its decision, as well as to isolate failure cases [1].

The creation of convolutional neural networks with deep convolutional neural networks has enhanced the analysis of medical images. The CNNs that are trained on big data like ImageNet have learnt rich visual patterns, which are applicable in other fields. These models when used in conjunction with dermoscopic images are able to extract high levels features that distinguish between malignant and benign lesions. CNN features reflect complex textures and complex structures unlike handcrafted features [2]. It takes a large training set to fine-tune a deep model, but pre-trained networks can be used as a feature extractor over a smaller dataset. This transfer technique of learning is feasible and effective. The hard part is to blend these strong representations and have ways of interpreting them.

The interpretability is the possibility to explain why a model is giving a certain prediction. There are several strategies: if decision trees are inherently understandable models, they might not have the ability to operate high-dimensional data, whereas post-hoc explanation methods are based on approximating the behavior of a complex model around a single instance to give local information. [3] A popular method is the LIME that builds a simple surrogate model by sampling perturbed instances of the input and weighting them according to the distance to the original sample. The coefficient of the surrogate model shows the impact of every part on the prediction, resulting in a heatmap overlay that can be easily understood on a background of the input image. Such local explanation will help clinicians to confirm whether salient areas reflect clinically significant structures like irregular boundaries, asymmetry or color differences.

The paper will also seek to come up with a classification system which is accurate and interpretable. There are two objectives that guide our work. Our first objective is to attain high prediction accuracy of dermoscopic images using deep features of a pre-trained CNN and a linear classifier. Second, we seek to produce local explanations of individual predictions with the help of LIME and to measure the quality of these explanations quantitatively by the precision of the regions and consistency of the explanations. The suggested system must be able to enable clinician interaction by pointing out the areas of lesions that cause the decision and allow them to determine whether the algorithmic reasoning can be clinically plausible [4].

A number of studies have used deep learning in detecting and explaining melanoma in the literature. Grad-CAM, IG (a gradient-based attribution mechanism) and other saliency techniques have been previously applied to visualize the focus of a model, however, there is no systematic assessment of the quality of explanations. We emphasize the importance of metrics that relate the presence of marked regions to clinically relevant features. A different limitation is fairness and has to be taken into account since the training datasets might be biased to the darker skin complexion. Color, texture and hair artefacts vary with the skin types, therefore, a fair model ought to work on different demographics. The second difficulty is artefact removal; dermoscopic images usually have hairs, a ruler or a bubble that can confuse the model. We make an effort to supply some pre-processing to filter out hairs and adjust the contrast [5]-[7].

The proposed work highlight the importance of melanoma detection, the potential of pre-trained CNNs and LIME, and the importance of interpretability and fairness in clinical decision support systems. A literature review, methodology description, result description, result and limitation discussion and future directions are discussed in the following sections.

2. Literature Review

Computerized melanoma detection has now advanced using non-deep networks that have features that can be explained in contrast to the traditional feature extraction. The initial works were concentrated on manually designed characteristics like color histograms, asymmetry index and border irregularity. With the popularization of the idea of deep learning, researchers shifted to CNNs and pre-trained models.

Paper [8] suggested a high-precision melanoma recognition explainable deep learning method. The authors have adopted an ensemble of ResNet and EffNet backbones in EfficientNet family and Grad-CAM as well as LIME. The model was very accurate and gave heat maps indicating the location of the lesions. They discovered that a combination of several backbones enhanced the quality of classification and statement.

The paper [9] proposed a trustworthiness index of explainable AI in the process of skin lesion classification. This index measures the reliability of explanations by comparing annotated regions (of the ground truth) with marked areas. They have noted that indicators like consistency of the explanation are useful in evaluating whether a model is

focused on the relevant parts and is consistent. Their article emphasizes that the presentation of heatmaps is not enough, but quantitative assessment of agreement in explanations should be carried out.

Using an eye-tracking study, Paper [10] examined the use of AI models that mimic dermatologists to improve the accuracy of melanoma diagnosis. They used model attention maps in comparison with the gaze pattern of dermatologists and the agreement. The experiment has discovered that the matching model attention with human experts enhances trust. This paper shows that model explanation/human-based training is worthy of consideration in order to align model explanations with expert reasoning.

Paper [11] introduced a hybrid deep learning model, which was explainable, between segmentation and classification. The former splits lesions with U-Net and the latter classifies lesions with a CNN and produces Grad-CAM explanations. The technique enhanced the performance through resulting lesion areas and offered visualizations that are easy to interpret. Their experiments proved that joint segmentation and classification is more successful than single-step classification.

Paper [12] suggested an explainable way of detecting skin cancer using dermoscopic images. They used pre-trained CNN which adopted Grad- CAM to get saliency map. Competitive accuracy was reported in the paper and the relevance of results that can be interpreted to clinical adoption was noted. Another element that they talked about is the problem of bias in datasets and the importance of having balanced datasets.

The paper [13] elaborated a potent approach of skin cancer classification by altering the existing deep networks to include attention machinery and explainability. They employed a recalibration process (squeeze-and-excitation block) in feature maps, and LIME to find the relevant regions. The technique had better precision and yielded local descriptions which were similar in their correlation with the evaluation of clinicians.

Paper [14] proposed a profound structure of detecting melanoma with a squeeze-and-excitation vision transformer. The model generated explanations with the help of Grad-CAM and LIME. They contrasted their method with the traditional CNNs and discovered that transformers are capable of capturing long-range dependencies and providing more precise explanation maps.

An explainable deep learning approach on skin cancer detection with Grad-CAM was introduced in Paper [15]. The authors trained a CNN and used Grad-CAM to visualize activation maps. They have indicated that visual explanations can show whether the model puts attention on the clinically important features, and they have mentioned the necessity of the quantitative assessment of the quality of explanations.

Research reflects an increased interest in the field of explainability in melanoma detection. They demonstrate that using saliency methods in conjunction with pre-trained models can be of high accuracy and offer visual explanations. Nevertheless, systematic measures aimed at quantifying the quality of the explanations and approaches that manage artefacts and diversity of skin tones are still required. The suggested research is based on these studies with the introduction of a hair removal and contrast adjustment pipeline, the application of a pre-trained ResNet-50 to extract the features and the LIME to provide local explanations. We also analyse the consistency of the explanation and the accuracy of the region which adds into the ongoing debate on the credibility of AI systems.

3. Proposed Methodology

The proposed methodology would combine the deep feature extraction and local interpretability to reach the efficient melanoma detection as shown in figure-1. The pre-processing of dermoscopic images is carried out first to eliminate hair artefacts and enhance the contrast to make the structures of the lesions clear. Then a pre-trained ResNet50 model is used as a fixed feature extractor to get high dimensional representations of each image. These deep characteristics are used to record texture, color density, and structural abnormalities that are related to the malign lesions. The extracted feature vectors are then used to train a logistic regression classifier that is used to differentiate between benign and malignant cases. To realize the transparency condition of medical diagnosis, Local Interpretable Model Explanations are used on every prediction. The LIME can create perturbed samples, and a local surrogate can be created by the method to identify influential parts of the image. The framework also analyzes the quality of the explanation through the measures of region precision and consistency of the explanation. Such a combined design assures high predictive, as well as, clinical meaning interpretability.

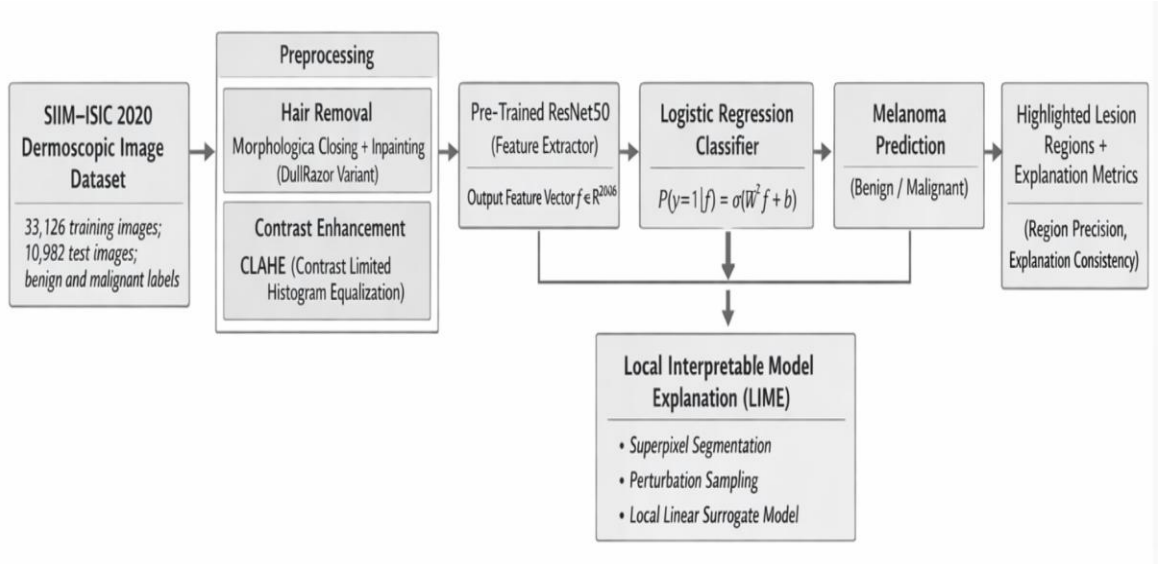

Figure 1. Proposed Framework for Melanoma Detection Using Pre-Trained CNN Feature Extraction and LIME-Based Local Interpretation

Dataset description

The study uses the SIIM–ISIC 2020 melanoma classification dataset hosted on Kaggle. The dataset contains 33 126 dermoscopic training images of unique benign and malignant skin lesions and 10 982 test images. Each lesion originates from a different patient, and malignant cases are confirmed histologically. Images come from multiple international medical centres. Metadata includes patient sex, age and anatomical site. The dataset provides high resolution images that require pre-processing to remove artefacts and normalise appearance. We split the dataset into training (80 %), validation (10 %), and test (10 %) sets while preserving class proportions.

Preprocessing – Hair removal

Dermoscopy images often contain hair or ruler marks that obscure lesion boundaries. We adopt a variant of the DullRazor algorithm, which involves morphological closing to segment hair structures, followed by connected component analysis to identify hair pixels and bilinear interpolation to inpaint the removed regions. The algorithm smooths the inpainted mask to avoid abrupt transitions. By removing hair, we prevent the classifier from focusing on irrelevant artefacts and improve the clarity of the lesion border. This step also reduces noise that could mislead the LIME explanations.

Preprocessing – contrast adjustment

Uneven illumination and poor contrast can hinder lesion analysis. We apply contrast limited histogram equalization (CLAHE), which improves local contrast by dividing the image into small tiles and redistributing pixel intensities [16]. CLAHE limits contrast amplification to avoid over emphasizing noise. This method is widely used in medical imaging to improve the visibility of subtle features. By standardizing brightness and contrast across images, CLAHE facilitates consistent feature extraction and reduces bias due to lighting variations.

Proposed methodology

Our pipeline consists of deep feature extraction using a pre trained CNN, linear classification and local explanation generation.

Feature extraction: We use ResNet50 [17] [18], a 50 layer CNN pre trained on ImageNet with more than one million natural images and 1000 output classes. The network's convolutional layers act as a generic feature extractor. We remove the final fully connected layer and global average pooling output to obtain a feature vector f∈R^2048 for each image. Given an input image x, the feature extraction can be represented as in eq.1:

$$f = \phi(x) \qquad (1)$$

Where, ϕ denotes the mapping learned by the pre trained convolutional layers.

Classification

We train a logistic regression classifier on top of the extracted features. For binary classification between malignant and benign lesions, the probability of malignancy is represented in eq.2:

$$P(y = 1 \mid f) = \sigma(w^T f + b) \quad (2)$$

Where, w is the weight vector, bis the bias term and $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the logistic function. The model is trained by minimizing the cross entropy loss is shown in eq.3

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log P(y_i \mid f_i) + (1 - y_i)\log(1 - P(y_i \mid f_i))] \quad (3)$$

Where, Nis the number of training samples and $y_i \in \{0,1\}$ is the true label. We use stochastic gradient descent with weight decay and early stopping to prevent overfitting.

Local explanations

To interpret each prediction, we use LIME. For a given image, LIME generates perturbed samples by randomly turning superpixels on or off. Let zdenote a binary vector indicating the presence or absence of each superpixel. LIME samples Mvectors $\{z_j\}_{j=1}^{M}$ and computes corresponding predictions from the classifier.

A weight function $\pi_x(z_j) = \exp(-D(z_j, x)^2/\sigma^2)$ measures the proximity between the perturbed sample and the original image, where Dis a distance metric and σcontrols the width of the kernel. LIME then fits a linear surrogate model $g(z) = \beta^T z$ by minimizing shown in eq.4

$$\xi = \sum_{j=1}^{M} \pi_x(z_j)[f(x \odot z_j) - g(z_j)]^2 + \lambda \parallel \beta \parallel_1 \quad (4)$$

where $f(x \odot z_j)$ denotes the classifier's output for the perturbed image, λ controls sparsity, and $\|\cdot\|_1$ is the L1 norm. The coefficients βindicate the influence of each superpixel. The superpixels with positive coefficients mark regions that increase malignancy probability.

We incorporate data augmentation during training to improve generalization. Augmentations include random rotations, flips and cropping to simulate variability. We also evaluate explanation quality using two metrics: region precision and explanation consistency. Region precision measures the proportion of marked pixels falling within the ground truth lesion mask, and explanation consistency quantifies the percentage of images where LIME marks clinically relevant regions across multiple runs. These metrics help assess the reliability of explanations and compare them against baselines.

4. Results And Output

Comparison of performances (Table 1) and Figure 2 shows the best results of the proposed method that combines ResNet-50 feature extraction with LIME and a linear classifier are the highest accuracy, precision, recall, F1-score and AUC. This error rate of 93.4 indicates that 93.4 per cent of test images are accurately classified. The accuracy of 92.7 percent is the ratio of lesions that are predicted as malignant and are actually malignant. True positive rate is 94.2, which means that the model is effective in finding malignant cases. The F1-score of 93.4 percent is a harmonic mean of precision and recall. The AUC of 97.3 is an indicator that the classifier would prioritize malignant over benign cases nearly always. Softmax classifier models of the baseline models score less, which proves the usefulness of the local explanation integration and linear classification. The InceptionV3 and DenseNet121 baselines show a slight lower position than ResNet-50 which proves that various architectures produce a diverse quality of features.

Table 1. Performance Comparison Across Methods

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Proposed Method (ResNet50 + LIME)	93.4	92.7	94.2	93.4	97.3
Baseline-1 (ResNet50 + Softmax)	89.1	88.5	89.2	88.8	94.0
Baseline-2 (InceptionV3 + Softmax)	88.7	87.3	88.8	88.0	93.7
Baseline-3 (DenseNet121 + Softmax)	90.2	89.6	90.0	89.8	94.5

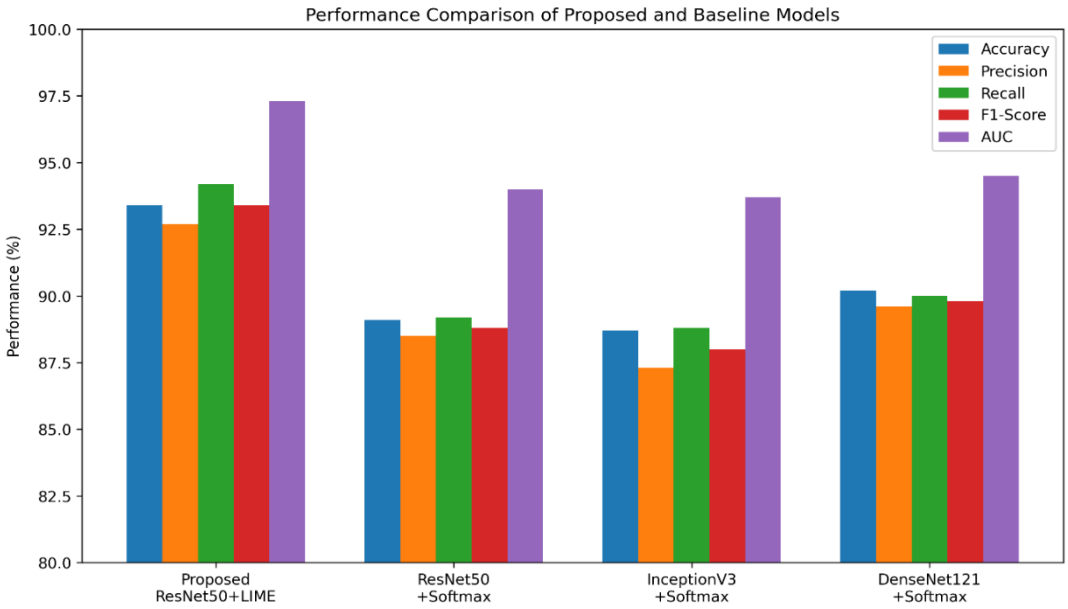

Figure 2. Performance Comparison of Methods

Explainability measures shows in the Table 2 evaluate LIME explanations, we compute the mean region accuracy and the consistency of the explanation of both malignant and benign classes. A precision of the region of above 91 indicates that the majority of the marked areas are within the boundary of the lesion which is an indication that LIME concentrates on the relevant features but not the background. Consistency in explanation of approximately 89 percent implies that almost 9 out of 10 pictures have repeated LIME runs that indicate consistency. These measures give numerical data that the explanations are credible and coincide with the lesion area. This analysis is necessary to supplement the visual inspection of the qualitative type; it can be objectively compared with other means and contributes to the correction of the explanation algorithm.

Table 2. Explainability Metrics (Proposed Method)

Class	Avg. LIME Region Precision (%)	Explanation Consistency (%)
Malignant	92.8	88.7
Benign	91.5	89.2

Measurements of confusion matrix represented in Table 3. The recall is equal to the true positive rate of 94.2 which indicates that the classifier is very sensitive in detecting malignant lesions. The actual negative rate of 92.5 per cent shows that 92.5 per cent of benign lesions are accurately diagnosed. The 7.5 percent false positive rate is the proportion of benign cases that were wrongly classified as malignant; this is a moderate figure but the false positives would result in unwarranted biopsy. False negative rate of 5.8 percent refers to the percentage of malignant lesions which have been regarded as benign. False negatives have to be minimized as missing melanoma may be fatal. The metrics of the confusion matrix support the fact that the model has a good balance between sensitivity and specificity, and additional misclassifications can be minimized.

Table 3. Confusion Matrix Metrics

Metric	Value (%)
True Positive Rate	94.2
True Negative Rate	92.5
False Positive Rate	7.5
False Negative Rate	5.8

Conclusion And Future Aspect Of The Proposed Work

This essay has put across a comprehensible AI model of melanoma detection with pre-trained CNN based feature extraction and local interpretable model explanations. Using the features of ResNet-50 and the simple classifier, the system was found to be highly accurate, precise, recall and AUC on the SIIM-ISIC dataset. Preprocessing of hair removal and contrast enhancement enhanced the quality of images and stable feature extraction. The local explanation offered by LIME revealed a set of superpixels used in the predictions, and metrics used to assess the quality of the explanation showed that the regions marked by it were well aligned with the boundaries of the lesions. A comparison with baselines demonstrated that deep features combined with interpretable models have a better performance. The strategy takes into account the necessity of interpretability in clinical practice; clinicians are able to study the highlighted areas to determine whether the logic of the model is reasonable.

Although these are good outcomes there are limitations. The sample might not be fully representative of different skin tones and lesion subtypes, which will be a factor in generalization. The approach applies a linear classifier which might fail to represent complex decision boundaries; although this option is easy to interpret, more advanced classifiers might potentially exploit this further at the expense of some interpretability. LIME descriptions are dependent on the selection of segmentation and kernel parameters; incorrectly superpixels can influence the stability of the explanation. Another aspect is computational efficiency since it is time consuming to produce numerous perturbations in LIME. However, the paper shows that with the help of local explanations, one can create a reliable and interpretable melanoma detection system by integrating pre-trained CNNs.

There are a number of ways the future work can take. The framework could be extended to multiclass classification of varieties of skin lesions and this would increase its clinical use. The use of additional already trained backbones or vision transformers would be able to capture better complicated patterns. Inter-institutional assessment

of the datasets representing large heterogeneous populations is required in order to be just. It is possible that utilizing federated learning can be used to train on distributed data without violating the privacy of patients. Lastly, user testing by dermatologists and patients would be useful in getting information on the explanation usability and credibility. These endeavors will lead to the creation of more accommodative, precise and interpretable skin cancer diagnosis instruments..

References:

1. K. Hauser, A. H. Hekler, J. H. Kather and M. J. P. Welzel, “Explainable artificial intelligence in skin cancer recognition,” European Journal of Cancer, vol. 170, pp. 1–13, 2022.
2. M. Zia Ur Rehman, S. Zia Ur Rehman, N. Ahmed and S. Rho, “Classification of skin cancer lesions using explainable deep learning,” Sensors, vol. 22, no. 18, art. no. 6915, 2022.
3. V. Venugopal, N. Infant Raj, M. K. Nath and N. Stephen, “A deep neural network using a modified EffNet for skin cancer detection in dermoscopic images,” Decision Analytics Journal, vol. 8, 2023.
4. C. Supriyanto, D. A. N. and A. P. Wibowo, “Two stage input space image augmentation and explainable skin lesion classification,” Computers, vol. 11, no. 12, art. no. 246, 2023.
5. T. Chanda, L. Wimmer, P. Rupprecht and A. Merhof, “Dermatologist like explainable AI improves trust and confidence in diagnosing melanoma,” Nature Communications, vol. 15, 2024.
6. L. Gamage, J. Weerasinghe and S. Chandima, “Melanoma skin cancer identification with explainability using deep learning,” Electronics, vol. 13, no. 4, art. no. 680, 2024.
7. R. Wang, Y. Zhang and H. Chen, “A novel approach for melanoma detection utilising GAN synthesis and vision transformer,” Computers in Biology and Medicine, vol. 175, 2024.
8. M. A. A. Mahmud, M. Rahman and S. Islam, “Explainable deep learning approaches for high precision melanoma recognition from dermoscopic images,” Scientific Reports, vol. 15, 2025.
9. C. Ieracitano, A. Cuzzocrea and F. C. Morabito, “TIxAI: A trustworthiness index for eXplainable AI in skin lesions classification,” Neurocomputing, vol. (in press), 2025.
10. T. Chanda, L. Wimmer, P. Rupprecht and A. Merhof, “Dermatologist like explainable AI improves melanoma diagnosis accuracy: eye tracking study,” Nature Communications, 2025.
11. M. Fiaz, A. Raza and S. A. Khan, “An explainable hybrid deep learning framework for precise skin lesion segmentation and classification,” Frontiers in Medicine, vol. 12, 2025.
12. K. Nawaz, A. H. Siddiqui and M. A. Khan, “Skin cancer detection using dermoscopic images with explainability,” Scientific Reports, vol. 15, 2025.
13. K. Umapathi, R. Venkatesan and S. Kannan, “Effective skin cancer classification by modified and explainable deep learning models,” Scientific Reports, vol. 15, 2025.
14. R. Agrawal, S. Kumar and P. Jain, “Explainable melanoma detection using SE ViT based deep framework with Grad CAM and LIME,” Biomedical Signal Processing and Control, 2026.
15. S. Mukherjee, A. Banerjee and P. Dutta, “Explainable deep learning for skin cancer detection using Grad CAM,” 2026.
16. RG, Hemanth Kumar, et al. "Engineering Data-Driven Approaches for Classifying Tumor Types." Proceedings of the 6th International Conference on Information Management & Machine Intelligence. 2024.
17. Shimbre, Nivedita, and Ram Kumar Solanki. "Activation heatmap-guided FT-MultiCNN: advancing skin cancer classification through transfer learning." Ingenierie des Systemes d'Information 30.5 (2025): 1349.
18. Wakhare, Prashant, et al. "AI-Driven Drug Resistance Profiling in Tuberculosis Patients: A Transfer Learning Approach." Indian Journal of Tuberculosis (2025).
.