Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 23 May 2026
Revised 25 June 2026
Accepted 18 July 2026

Natural Resources for Human Health 2026, Vol. 6 (8s): 570–586

KEYWORDS:
Melanoma detection, dermoscopic images, convolutional neural network, attention mechanism, explainable artificial intelligence, Grad-CAM, skin lesion classification, HAM10000 dataset.

An Explainable Attention-Guided Convolutional Neural Network Framework for Melanoma Detection and Classification

Ashish Jain[1], Rashmi Yadav [2]

[1] Research Scholar, Department of Computer Science and Engineering, SAGE University, Indore, Madhya Pradesh, India
ashishtruba5@gmail.com, ORCID ID: ORCID ID: 0009-0002-6693-1791
[2] Professor and HOI, Department of Computer Science and Engineering, SAGE University, Indore, Madhya Pradesh, India
rashmi.yadav@sageuniversity.in, ORCID ID: 0009-0008-4418-4020,
Scopus ID: 57208815839

ABSTRACT: Melanoma is a more aggressive type of skin cancer and early detection is critical to good clinical outcomes. Interpretation is difficult, however, due to the fact that dermoscopic images of melanomas and benign lesions can present with similar visual features, such as uneven pigmentation, border variation and texture variation. In this study, a CNN with explainable attention guidance framework is designed to detect and classify melanoma skin lesion images from dermoscopic images. The dataset HAM10000 (10,015 images) was split into 1,113 melanoma and 8,902 non-melanoma images in a binary classification task. The proposed framework combines a preprocessing, data augmentation, class-balanced learning, channel–spatial attention-based feature refinement, and explainability via Grad-CAM. Attention mechanism was applied to improve feature learning to focus on lesions and class-balanced learning was applied to prevent the bias of non-melanoma samples. The model's performance on the independent test set resulted in an accuracy of 94.81%, precision of 91.76%, recall/sensitivity of 92.21%, specificity of 95.30%, F1 score of 91.98% and ROC-AUC of 0.972. The results revealed high sensitivity to the detection of malignant lesions, as 154 out of 167 images of melanomas were correctly identified.The confusion matrix indicated a high sensitivity towards melanomas (154/167 images were correctly identified). The proposed framework was evaluated by comparative analysis and it was seen that all the basic CNN, VGG16, ResNet50, DenseNet121, MobileNetV2, and EfficientNetB0 models were outperformed by the proposed model. The ablation study indicated that the contributions of attention-guided learning, data augmentation, and class balancing were each progressively in the direction of performance improvements. The results also revealed that the Grad-CAM visualization primarily highlighted areas of the image that were relevant to the lesions, including the presence of irregular borders, asymmetry, and pigment variations. As such, the proposed framework offered a valid, meaningful and clinically relevant solution for automated melanoma screening.

1. INTRODUCTION

Melanoma is one of the most malignant forms of skin cancer and a public health issue due to the high number of people diagnosed late, when it has a high chance of spreading and a low survival rate. Melanoma makes up a minority of all skin cancers, but it is the cause of a majority of the deaths from skin cancer. Melanoma has continued to be a burden across the globe with variations attributed to genetic, environmental, geographical and lifestyle related risk factors [1]. Early diagnosis can be a key part to the success of treatment: Melanoma diagnosed early is easier to deal with than melanoma in its advanced

metastatic stage. Thus, timely, efficient, and accurate detection of melanomas is of great clinical importance in screening procedures and for clinical decision making, as well as public health surveillance.

Dermoscopy is an important non-invasive imaging modality in the evaluation of pigmented skin lesions. By providing dermatologists with a visual pattern of their lesions such as asymmetry, irregular borders, changes in color, diameter, and structural changes that may not be readily apparent to the naked eye, the treatment can be customized to the individual. The dermoscopic interpretation is, however, strongly related to the clinical experience and might be influenced by the image quality and the similarity of the lesions, inter-observer differences, and subjective visual judgement. Benign nevi and malignant melanoma can have similar appearances in some instances, and in some cases, it may be difficult to make an accurate distinction even by trained clinicians. These diagnostic problems have paved the way for the development of computer-aided diagnostic systems that can help dermatologists with objective, reproducible, image-based melanoma classification.

Conventional image-processing and machine-learning techniques were previously popular in the field of computer-aided melanoma diagnosis. These approaches were mostly based on handcrafted features including colour, shape, texture, border irregularity, lesion symmetry, and geometric feature descriptors. After that, classifiers like support vector machine, k-nearest neighbor, random forest, artificial neural networks and logistic regression were used to classify the lesions. While these methods offered helpful diagnosis information, they were highly sensitive to the quality of the feature extraction, segmentation, illumination variation, hair artifacts, and background noise. Handcrafted features are not always able to capture complex intra-lesion variation and subtle morphological differences between melanomas and benign lesions. Consequently, traditional machine-learning techniques have poor generalization capabilities when trained with large and diverse dermoscopic collections of images obtained under various imaging conditions and clinical settings.

The advent of deep learning, especially convolutional neural networks, has revolutionized the field of medical image analysis by providing the capability to automatically extract features from raw images. CNNs can be trained to learn the hierarchical visual representations, such that the earlier layers are trained to learn low level features like edge, color, textures, and the deeper layers are trained to learn the high-level features associated with the lesions, which are specific to malignancy. This result by Esteva et al. made skin cancer diagnosis with clinical images a promising application of deep learning by showing high performances of deep CNNs equal to dermatologists' performances [2]. Since then, several CNN and transfer learning architectures such as VGG, ResNet, DenseNet, Inception, MobileNet and EfficientNet have been used in the fields of melanoma and multi-class skin lesion classification [5]–[7]. Public datasets, like HAM10000 and challenge datasets from ISIC, have further contributed to the fast rise in research concerning automated melanoma detection by offering a large number of dermoscopic images for training and benchmarking deep-learning models [3] and [4].

Although the existing CNN-based melanoma detection systems have demonstrated great classification accuracy, there are still some crucial drawbacks. First, CNN models are typically utilized as black-box classifiers and a final prediction is delivered without a visual reason. This lack of interpretability also results in a loss of trust with clinical application as dermatologists would like to know whether the model is paying attention to the lesion area or some irrelevant artifacts. Second, there are many melanoma datasets which are class-imbalanced, meaning that there are more benign cases than melanomas cases. This imbalance can lead to the creation of skewed models that have high overall accuracy but low sensitivity to melanoma. A false negative prediction of melanoma in medical diagnosis can be especially problematic as it can cause a delay in treatment. Third, common CNN models can pay attention to non-diagnostic areas, including hair, image boundaries, illumination shadows, colors, calibration marks, or other skin texture. Thus, in addition to the accuracy requirement, for clinical reliability, the model must demonstrate lesion-focused learning, high sensitivity, robust generalization and interpretable decision support.

Attention mechanisms can effectively strengthen the CNN-based melanoma diagnosis system by making the system focus on the parts that are more important for diagnosis, while reducing the importance of the background parts. Channel attention can detect the importance of the feature maps, and spatial attention can detect the location of the most informative lesion regions. In deep learning, attention modules like squeeze-and-excitation blocks and convolutional block attention modules have demonstrated significant promise for enhancing the representation of deep features in an adaptive manner [8, 9]. For the purpose of identifying melanoma, an attention-guided CNN may be utilized to guide the model to focus on the border of the lesion, pigment network, irregular shapes, color variations, etc. discriminative patterns linked to malignancy. This fact is beneficial in the classification of dermoscopic images, where the diagnostic region can be a small part of the image.

Another critical aspect of clinical use of deep learning systems for diagnosing melanoma is explainable AI. Explainability techniques, like Grad-CAM, LIME and saliency-based visualization, can be used to show image regions contributing to the final prediction of the model [10, 11]. The CNN model used for medical imaging is particularly convenient for grad-CAM because the class-discriminative heatmaps are produced by the last convolution layers, and the highlighted regions have affected the prediction. Such heatmaps can also be used to indicate that the model focuses on the lesion area and not on other skin or artifacts in the classification of melanoma. This visual evidence is useful to help improve transparency, aid clinical interpretation and to be able to recognize when the model is making decisions based on irrelevant features.

Motivated by these research requirements, the current work suggests a transparent attention-guided CNN structure for melanomas detection and categorization from dermoscopic skin lesion photos. The novelty of the proposed framework is that attention-guided feature refinement, class-balanced learning and explainable visual interpretation are integrated in a single CNN-based diagnostic model. The proposed framework seeks to alleviate the class imbalance induced bias, gain melanoma sensitivity, and offer visual explanations of the model's predictions while emphasizing the accuracy of classification. The attention module is developed to enhance the diagnostic features of the lesions and the class-balanced learning is employed to avoid false-negative melanoma classification. Furthermore, explainability based on Grad-CAM is added to visualize the region that was the basis of the melanoma or benign predictions.

This study has four major contributions. First, an intelligent CNN-based system is designed for automatic melanoma detection and classification for the dermoscopic image. Second, an attention-guided feature refinement mechanism is incorporated for enhancing the lesion-focused representation learning. Third, class-balanced training is implemented to reduce the imbalance of the datasets and enhance melanoma sensitivity. Fourth, AI visualization of the model is used to offer heatmap interpretation of the model decision, enhancing clinical transparency. The proposed framework is validated with conventional diagnostic parameters such as accuracy, precision, recall, specificity, F1-score, ROC-AUC, confusion matrix, and ablation analysis. It also validates the proposed model with a comparison with the conventional CNN model and transfer learning model to showcase the importance of attention, class balancing and explainability in the classification of melanoma.

2. RELATED WORK

The initial research for melanoma diagnosis has been primarily focused on clinical observation, visual assessment based on dermoscopy, and handcrafted image descriptors. The diagnostic results of dermoscopy were dependent on dermatologist experience and image quality, and it had enhanced the visibility of pigment network, border of the lesion, color distribution, streaks, dots and globules, and other morphological structures. Several computers aided diagnostic techniques had extracted handcrafted features that were related to asymmetry, irregular boundary, color difference, texture, and lesion geometry, in order to reduce subjectivity. The features were commonly classified with support vector machine, k nearest neighbor, decision tree, random forest, artificial neural networks and ensemble classifiers. These approaches had been proven to be beneficial in synthetic data sets, but they were significantly impacted by segmentation error, changes in illumination, artifacts due to hair color, background skin color, and the limited expressive power of handcrafted descriptors of complex melanoma patterns [13, 14].

As public dermoscopic dataset increased, deep learning had become a prevalent method to classify skin lesions. Very deep residual networks had been applied to automatic melanoma recognition by Yu et al. [15] and they demonstrated the superiority of deep CNN features over handcrafted descriptors. The potential of deep neural networks in the field of skin cancer classification at a level similar to that of dermatologists demonstrated by Esteva et al. [2] had popularized the use of CNNs in dermatological image analysis. The ISIC challenge framework for melanoma detection had been reported by Codella et al. [4] where CNN-based methods had gained importance for the classification of lesions. Such studies demonstrated that CNNs could automatically extract hierarchical representations from dermoscopic images, such as low-level edge and color gradient, and high-level patterns of malignant lesions.

Transfer learning also became popular because of the fact that medical image data sets were typically smaller than natural image data sets. Pre-trained CNN architectures like VGG, Inception, ResNet, DenseNet, MobileNet and EfficientNet had been adapted for Melanoma and Multi-class classification of skin lesions. Harangi [16] had explored the use of deep CNN ensembles for skin lesion classification and demonstrated that multi-networks prediction stability. Hosny et al. [17] have used transfer learning and data augmentation for skin lesion classification and reported that pre-trained networks were able to decrease the need for very large training sets. Gessert et al. [18] had applied CNN-based techniques for skin cancer diagnosis coupled with

diagnosis guided loss weighting and patch-based attention, and had suggested model optimization and attention to the lesion regions could boost the classification performance. Many transfer learning models were however, still remained to be high computational, and sometimes didn't offer enough clinical interpretation.

Augmentation, ensemble learning and class balancing were used in several studies to enhance the classification of melanoma. Class imbalance had impacted the sensitivity of trained models due to the fact that dermoscopic datasets had typically more benign lesions than melanoma cases. The overall accuracy was important as it could potentially overlook poor melanoma recall. Chaturvedi et al. [19] had achieved skin lesion analysis using a deep learning approach on dermoscopic images and they had highlighted the importance of data augmentation and data preparation. The deep convolutional approach was suggested by Kaur et al. [20] in the context of melanoma classification, and it was reported that better results in classification were obtained by improving the preprocessing and feature learning. However, there had remained a strong concern of a bias for imbalance and a large number of false-negatives in automated diagnosis for melanoma prediction.

Later, attention-based deep learning was introduced to boost the learning of features from lesions. CNN models were capable of focusing more attention on diagnostically important regions and outright suppressing irrelevant background information thanks to attention mechanisms. Channel attention had focused on salient feature maps, while spatial attention had highlighted the salient lesion locations in the image. Hu et al. [8] proposed squeeze-and-excitation mechanism for adaptive channel-wise feature recalibration, and Woo et al. [9] proposed the convolutional block attention module (CBAM) that fused channel-attention and spatial-attention. For skin lesion classification, attention-based CNN models were found to have enhanced the ability of the models to discriminate between visually similar classes by focusing on the irregular borders, pigment networks, color heterogeneity and lesion-specific structures. He et al. [21] proposed deep metric attention learning for dermoscopic skin lesion classification and demonstrated that attention-guided representation learning led to better inter-class separation and intra-class compactness.

Though there were improvements in classification performance, many of the deep learning models were still facing the black-box problem. The limitation was important in medical diagnosis since doctors wanted to have evidence supporting the prediction of a lesion being a melanoma or benign tumor. To enhance transparency in classification of skin lesions, methods of explainable artificial intelligence had been, therefore, proposed. It was known that there are ways to identify image regions that affect deep learning predictions, such as Grad-CAM, LIME, SHAP, and saliency maps. The LIME approach was proposed by Ribeiro et al. [11] for the explanation of the predictions of classifiers, while the Grad-CAM approach was proposed by Selvaraju et al. [10] for the localization of the class discriminative information in CNNs. In the dermoscopic image analysis, Grad-CAM-based heatmaps have been used to prove if the model had marked the lesion areas, or not, as compared to the ruler, borders, illumination artefacts, or the surrounding skin.

The current study brought together explainability and deep learning in the diagnosis of skin cancer in an increasing number of recent studies. In the past, Shah et al. [22] had proposed an explainable AI-based skin cancer detection system based on CNN models and Grad-CAM visualization. Attalah et al. [23] suggested an explainable deep learning-based computer-aided diagnosis model for skin cancer classification and tested the model on benchmark datasets. These studies had suggested that the explainability helped to make the AI assisted diagnosis clinically acceptable. But, most of the explainable models merely added visualization after classification and didn't sufficiently incorporate attention-guided learning, class-balanced training, and explainable interpretation in one consistent diagnostic framework.

From the literature reviewed, the CNN and transfer learning models had been seen to provide promising results in melanoma classification, but some research gaps were not resolved. First, there were a number of models that focused on accuracy and failed to consider issues of sensitivity, specificity, false negative and clinical reliability. Second, class imbalance is still an issue with melanoma prediction, due to the prevalence of benign images typically included in the public datasets. Third, the conventional CNNs were not always targeting on lesion specific region and learned misleading background artifacts. The fourth was that many studies were missing any visual interpretable evidence, resulting in a decrease of clinical trust. Hence, an explainable attention-guided CNN framework was required that could integrate lessons, class-balanced learning, and Grad-CAM-guided interpretation, for accurate melanoma detection and classification.

3. MATERIALS AND METHODS

This study aimed to be a dermoscopic image-based melanoma detection and classification experiment. A fully labeled dataset was created from the HAM10000 dermoscopic image dataset with 10,015 skin lesion images from seven diagnostic categories. This work is based on the reorganized data set where the original problem is transformed into a binary classification between melanoma and non-melanoma. The melanoma class comprised all images that were annotated as melanoma and the non-melanoma class comprised all images annotated as melanocytic nevus, basal cell carcinoma, benign keratosis-like lesion, dermatofibroma, vascular lesion, and actinic keratoses. This binary organization was clinically adequate as the primary aim of the framework proposed was to differentiate malignant melanoma cases from other types of skin lesions. Table 1 shows the description of the dataset and the distribution of the classes. Figure 1 illustrates some representative dermoscopic images of melanoma and non-melanoma lesions.

Table 1. Prepared dataset description and class distribution

Parameter / Class	**Description / Count**
Dataset used	HAM10000 dermoscopic image dataset
Total number of images	10,015
Image type	Dermoscopic skin lesion images
Original diagnostic classes	7
Proposed classification type	Binary classification
Input image size	224 × 224 × 3
Image format	JPG
Melanoma images	1,113
Non-melanoma images	8,902
Melanoma percentage	11.11%
Non-melanoma percentage	88.89%

The original HAM10000 dataset had seven diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, benign keratosis-like lesion, dermatofibroma, vascular lesion, and actinic keratoses. The prepared binary dataset consisted of images of melanoma as the positive class and all other types of lesions as the non-melanoma class. This conversion enabled the proposed model to focus on clinically important melanoma screening. The binary distribution was also highly imbalanced as melanoma only accounted for 11.11% of the dataset. To avoid the majority class bias in non-melanoma predictions, therefore, learning was class-balanced in the model training.

Model training was preceded by all dermoscopic images preprocessing. The original images had different resolutions, illumination conditions and visual properties and each image was resized to 224 × 224 × 3 pixels. Intensity values of the image are scaled from 0 to 1 to stabilize the training of the CNN. The images were enhanced for contrast and normalized for color to minimize the variation in the illumination between images. Hair artifacts, dark borders and background disturbance were treated as unwanted visual noise as it may mislead the model to non-lesion regions. The proposed preprocessing workflow and preprocessing and augmentation strategy are depicted in Figure 2 and summarized in Table 2, respectively.

Natural Resources for Human Health
eISSN: 2583-1194

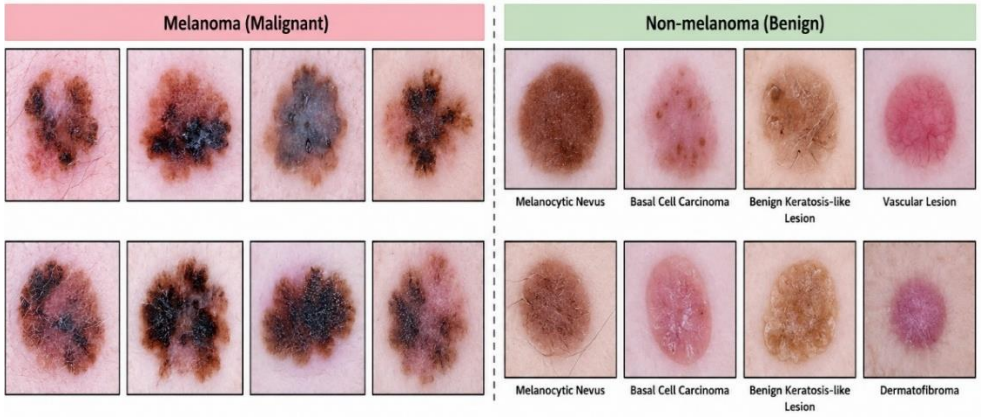

Figure 1. Representative dermoscopic images of melanoma and non-melanoma skin lesions

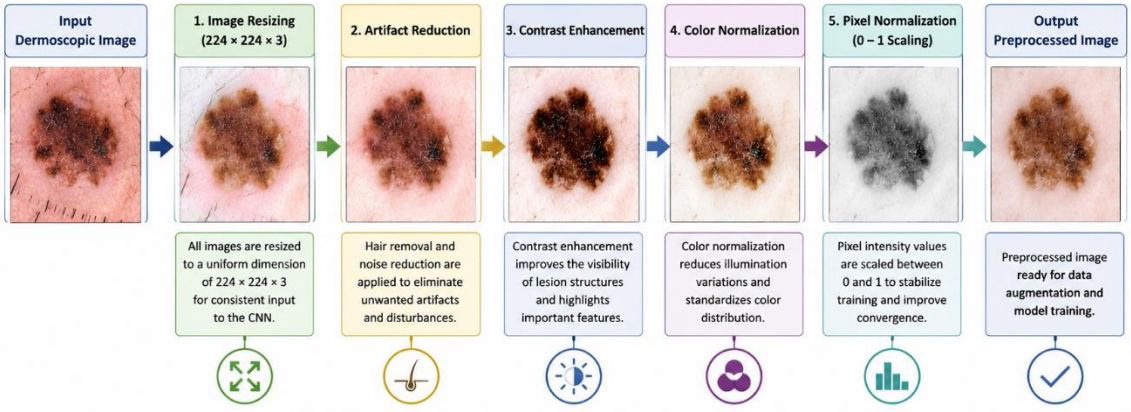

Figure 2. Preprocessing workflow for dermoscopic skin lesion images.

Table 2. Preprocessing and augmentation strategy

Operation	**Method / Parameter**	**Purpose**
Image resizing	224 × 224 × 3	Uniform CNN input size
Pixel normalization	0–1 scaling	Stable model training
Contrast enhancement	Intensity adjustment	Improved lesion visibility
Color normalization	Color correction	Reduced illumination variation
Artifact reduction	Hair/border/noise suppression	Reduced non-lesion interference
Rotation	0°–30°	Improved orientation robustness
Horizontal/vertical flipping	Enabled	Increased image variability
Zooming	0.80–1.20	Improved scale robustness
Width/height shifting	0.1	Improved position robustness
Brightness variation	0.80–1.20	Improved illumination robustness

Natural Resources for Human Health
eISSN: 2583-1194

Data augmentation was used just on the training set to enhance model generalization and avoid overfitting. The augmentation process produced modified versions of the dermoscopic images, but not altered the diagnostic labels of the images. Through the process of rotation, flipping, zoom, shifting and brightness variation, the model learned the features of the lesions in different orientations, scales, positions and illumination. In order to increase the variability of the minority-class, images from the melanoma class were augmented, and images from the non-melanoma class were not augmented as much as those from the melanoma class. Figure 3 shows the data augmentation pipeline that was used in this study.

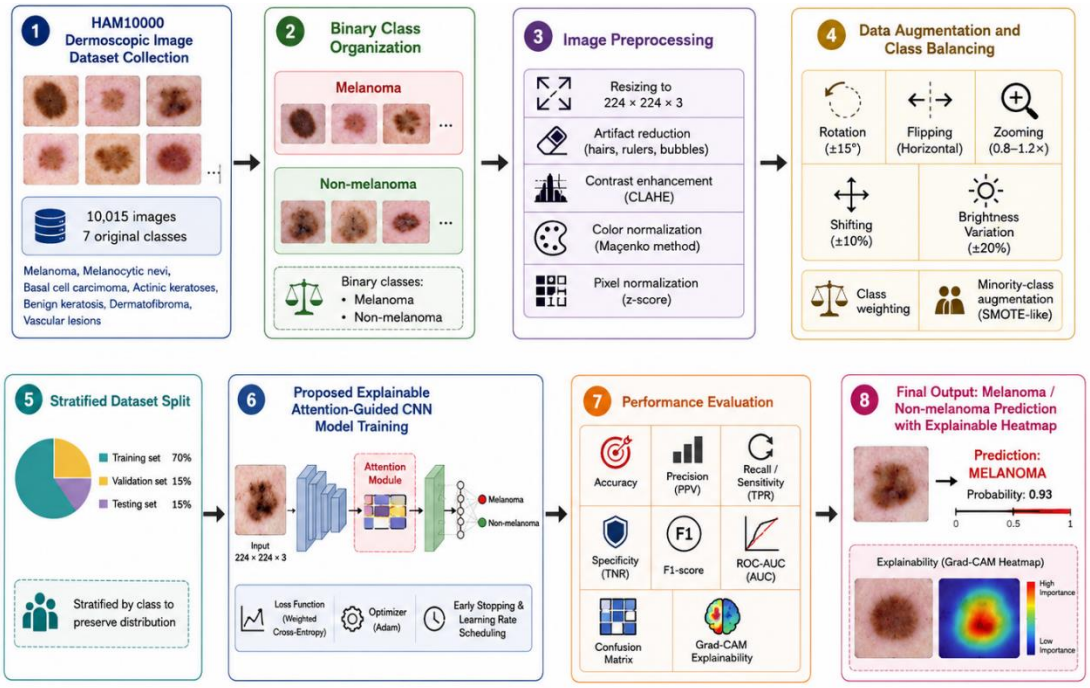

Figure 3. Data augmentation pipeline used for melanoma and non-melanoma images

The prepared dataset was split into a training, validation and testing set in a stratified manner. The classes of melanoma and non-melanoma were evenly represented across the subsets due to stratification. For model-learning, the training set was employed; hyperparameter tuning and monitoring overfitting were carried out using the validation set; and the testing set was employed for final independent performance evaluation. The splitting ratio was chosen to be 70:15:15 as indicated in Table 3.

Table 3. Training, validation, and testing split of the prepared dataset

Dataset subset	Melanoma images	Non-melanoma images	Total images	Purpose
Training set	779	6,231	7,010	Model training
Validation set	167	1,335	1,502	Hyperparameter tuning
Testing set	167	1,336	1,503	Final model evaluation
Total	1,113	8,902	10,015	-

A class-weighted training strategy was considered to minimize the impact of class imbalance. Melanoma images were given a higher training weight due to being the minority class. This enhanced the CNN model to pay more attention to the melanoma samples when calculating loss. In the training phase, images from the melanoma class were also augmented with minority class augmentation to increase the number of images in the class. It was a significant strategy because it was important to look at melanoma sensitivity, rather than accuracy alone. Moreover, focal loss is introduced as an optional loss function to enhance learning from challenging and wrongly classified lesion sets.

The entire experimental process involved dataset preparation, conversion of binary labels, dataset preprocessing, data augmentation, dataset balancing, model training and evaluation. The first step was to reclassify all the HAM10000 images into melanoma and non-melanoma. Second, all images were resized and normalized. Third, the dataset was split into training, validation, and testing subsets with stratified splitting. Additionally, the model was made more robust by adding augmentation and class weighted learning methods (fourth). Lastly, the generated images were fed into the proposed explainable attention guided CNN model for training.

The model was developed in a language of deep learning called Python and TensorFlow/Keras. The Adam optimizer was chosen due to its stable and adaptive learning when optimizing a CNN. The model was trained for a maximum of 100 epochs, using a batch size of 32. Validation loss was monitored and training was halted by the early stopping technique when it failed to improve continuously. For standard binary classification, binary cross-entropy was used, and focal loss was investigated in the class balanced setting. To prevent overfitting and to make the convergence more stable, dropout and batch normalization were applied. The experimental setup and training hyperparameters are given in Table 4.

Table 4. Experimental configuration and training hyperparameters

Parameter	**Value**
Programming language	Python
Deep learning framework	TensorFlow/Keras
Input image size	224 × 224 × 3
Classification type	Binary classification
Optimizer	Adam
Learning rate	0.0001
Batch size	32
Maximum epochs	100
Loss function	Binary cross-entropy / focal loss
Activation function	ReLU and sigmoid
Regularization	Dropout and batch normalization
Class balancing	Class weighting and minority-class augmentation
Validation monitoring	Early stopping
Evaluation mode	Independent test set

The data set and the experimental design gave a good framework to build the proposed explainable attention-guided CNN. Stratified splitting, pre-processing, minority-class augmentation and class-weighted learning were applied to increase the accuracy of the melanoma prediction, and to decrease the bias towards the majority class (non-melanoma). The proposed CNN architecture, attention mechanism and explainability module are detailed in the following section.

4. PROPOSED EXPLAINABLE ATTENTION-GUIDED CNN FRAMEWORK

The proposed framework aims to detect and classify melanoma from dermoscopic images of skin lesions using an attention-guided convolutional neural network and providing an explanation in the form of a visual interpretation. Three key challenges with traditional CNN-based melanoma diagnosis were overcome in the design of this framework: inadequate lesion-centric feature learning, class imbalance bias, and lack of deep learning decision interpretability. The entire structure was divided into five main phases: image input, feature extraction using convolutional layers, attention mechanism for feature refinement,

melanoma classification, and explainable visualization. The overall architecture of the proposed explainable attention-guided CNN is presented in Figure 4.

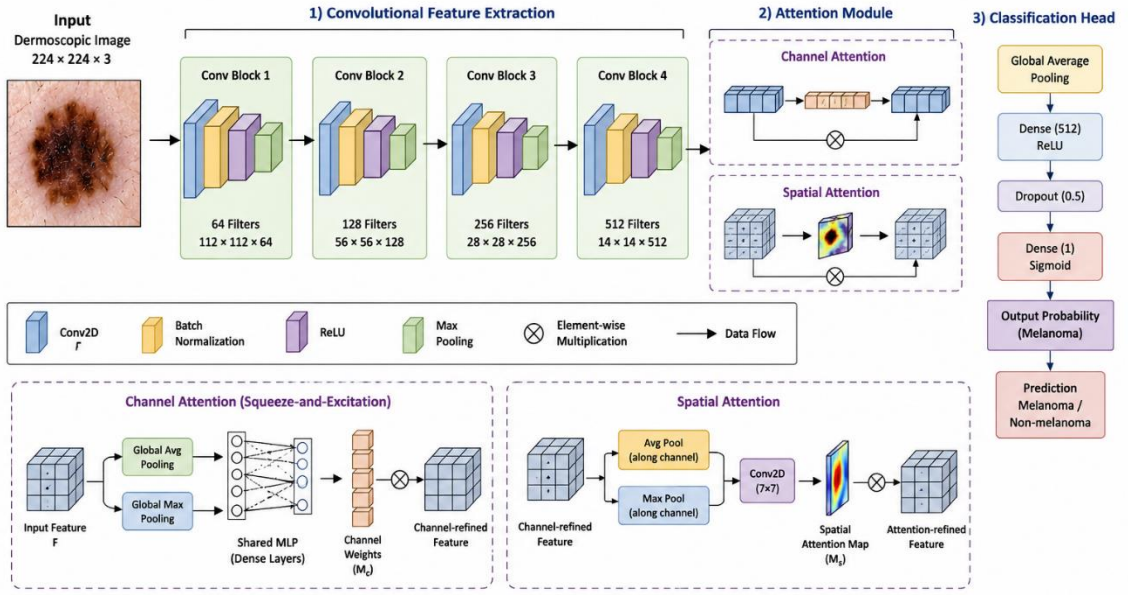

Figure 4. Architecture of the proposed explainable attention-guided CNN framework

The input to the proposed model was a preprocessed dermoscopic image of size 224×224×3. A series of convolutional blocks were used for hierarchical feature extraction in each image. Low-level visual features like edges, color gradients, lesion boundaries, and texture variation were obtained from the first convolutional layers. High-level lesion-specific patterns, such as irregular border structure, pigment distribution, asymmetry, color heterogeneity and malignant texture characteristics, were learned in the deeper layers of the convolutional network. The convolutional blocks contained a Convolutional layer, Batch Normalization, ReLU activation and Max-Pooling operation. The learning process was stabilized using batch normalization and dominant feature responses were maintained by max-pooling the spatial dimensions.

The layer-wise architecture of the proposed CNN is given in Table 5. The architecture was designed to be efficient with minimal depth but still allow for discrimination features to be learned from lesions. After dense layers dropout was added to decrease overfitting particularly because the melanoma class was smaller than the non-melanoma class.

Table 5. Layer-wise configuration of the proposed attention-guided CNN model

Block / Layer	Operation	Output role
Input layer	224 × 224 × 3 dermoscopic image	Standardized image input
Conv Block 1	Conv2D + BN + ReLU + MaxPool	Low-level edge and color features
Conv Block 2	Conv2D + BN + ReLU + MaxPool	Texture and boundary features
Conv Block 3	Conv2D + BN + ReLU + MaxPool	Intermediate lesion patterns
Conv Block 4	Conv2D + BN + ReLU + MaxPool	High-level melanoma-specific features
Attention module	Channel attention + spatial attention	Lesion-focused feature refinement
Global average pooling	Feature vector generation	Dimensionality reduction

Dense layer	Fully connected learning	Class-discriminative representation
Dropout layer	Regularization	Overfitting reduction
Output layer	Sigmoid activation	Melanoma / non-melanoma prediction

The proposed framework's main feature refinement component was the attention module. This was placed at the end of the convolutional layers with the goal of enhancing features of the lesions which were relevant to diagnosis and reducing background responses that were not relevant. A channel and spatial attention strategy was used. Channel attention was used to distinguish between which feature maps were useful for melanoma classification and spatial attention to determine the location of important lesion regions in the image. This design was applicable to dermoscopic images, since the information related to melanoma was not evenly spread throughout the image, but was primarily grouped around the area of the lesion, the borders of the lesion, variations in colour, and any irregular internal structures.

The input feature map was processed by the global average pooling and global max pooling layers in the channel attention stage to extract complementary information from various channels. The channel attention weights were obtained by passing these descriptors through common dense layers and then fusing them together. Generated weights were multiplied with feature maps to highlight the important diagnostic channels. This process assisted in giving higher weights to maps of features that had lesion-specific color and texture and structural data.

A spatial descriptor was generated by using average pooling or max pooling in the spatial attention stage along the channel dimension. These descriptors were then combined together and subjected to a convolutional process and then sigmoid activation to produce a spatial attention map. The channel-refined feature map was then multiplied to the spatial attention map. This enabled the network to concentrate on the key locations of the lesions, and minimize the contribution of healthy skin, artifacts, borders, and non-diagnostic areas. The final feature map with cleaned up attention was then fed into the classification layers.

Global average pooling was used to transform the last feature maps into a compact feature vector after feature refinement using an attention mechanism. This decreased the number of trainable parameters and helped to avoid overfitting. The pooled feature vector was fed into a dense layer with a ReLU activation function and dropout regularization. Finally, a binary classification was carried out using a sigmoid output neuron. Output probability was defined as the probability of having melanoma. If the probability value was over the threshold selected, the image was classified as melanoma and if the value was less, it was classified as non-melanoma.

A summary of the classification workflow of the proposed framework is presented in Table 6. First of all, the model took a dermoscopic image as its input, extracted the high-level features of the image through the hierarchical CNN, refined the high-level features through the attention mechanism, calculated the classification probability, and finally generated an explainable heatmap to explain visually.

Table 6. Functional workflow of the proposed framework

Stage	**Process**	**Purpose**
Stage 1	Input of preprocessed dermoscopic image	Standardized image feeding
Stage 2	Convolutional feature extraction	Learning lesion features
Stage 3	Channel attention	Selecting important feature maps
Stage 4	Spatial attention	Localizing important lesion regions
Stage 5	Global average pooling	Compact feature representation
Stage 6	Dense classification layer	Melanoma/non-melanoma decision
Stage 7	Grad-CAM visualization	Explainable interpretation

In order to enhance the clinical transparency, the proposed CNN framework was integrated with an explainable AI module. Visual heatmaps were created using the last convolutional layer in the network, with the help of Grad-CAM. These heat maps showed where the image was most useful in making the final prediction of melanoma or non-melanoma. The goal of Grad-CAM wasn't to actually boost classification accuracy, but to offer visual proof of what the model is thinking. This was significant since a clinically useful melanoma detection model should not only be a prediction, it must also reflect whether the decision was made based on the regions of the lesion.

The Grad-CAM process used the gradient of the predicted class score with respect to the final convolutional feature maps. Global average pooling of gradient was used to calculate the importance weight of each feature map. These weights were then added to the feature maps to produce class-discriminative localization map. The Grad-CAM heatmap was finally superimposed on the original dermoscopic image to visually highlight the important regions. Explainability process is presented in Figure 5.

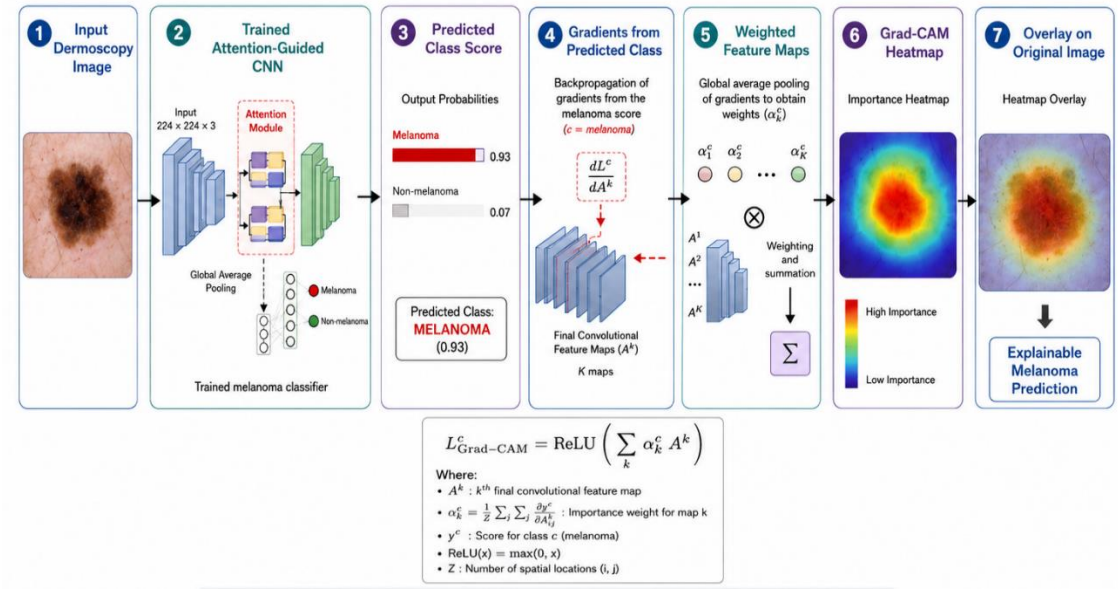

Figure 5. Grad-CAM-based explainability process for melanoma prediction

This was the reason for the proposed explainable attention-guided CNN framework that incorporated prediction performance and interpretability. The attention mechanism enhanced and focused the learning of features that are relevant to the lesion.The attention mechanism helped to focus and emphasize the learning of features that are relevant to the lesion, while suppressing other background information. Class balanced training was found to minimize the impact of majority class dominance and enhance the melanoma sensitivity. Grad-CAM visualization was used to generate a visual representation of the class predicted and as an aid to determining if the model was concentrating on clinically relevant lesion regions. This combination proved to be more suitable for computer-aided melanoma screening than a regular black-box CNN classifier.

5. RESULTS AND DISCUSSION

The proposed explainable attention-guided CNN architecture was tested by using the independent test set on the prepared HAM10000 binary dataset classification. The model performance was evaluated by accuracy, precision, recall, specificity, F1-score, ROC-AUC, confusion matrix, training-validation curve analysis, comparative analysis, ablation study and explainability analysis with Grad-CAM. Due to the clinical sensitivity of melanoma detection, the sensitivity (true positive rate) was also taken into consideration and was thought to be of special importance as false-negative melanoma prediction could lead to delayed diagnosis and treatment.

The accuracy curve of the training and validation set are plotted in Figure 6(a), and the loss curve of the training and validation set are plotted in Figure 6(b). The accuracy in the training phase was improved gradually after each epoch of training, and gradually became stabilized after the convergence. The validation accuracy was similarly low, suggesting that the model was actually learning useful features specific to the lesions and not just learning to recognize the training images. The loss curve also

monotonically declined and there was also no significant gap between training and validation loss. This suggested that the usage of preprocessing, augmentation, dropout, batch normalization, and early stopping contributed to minimize the overfitting problem.

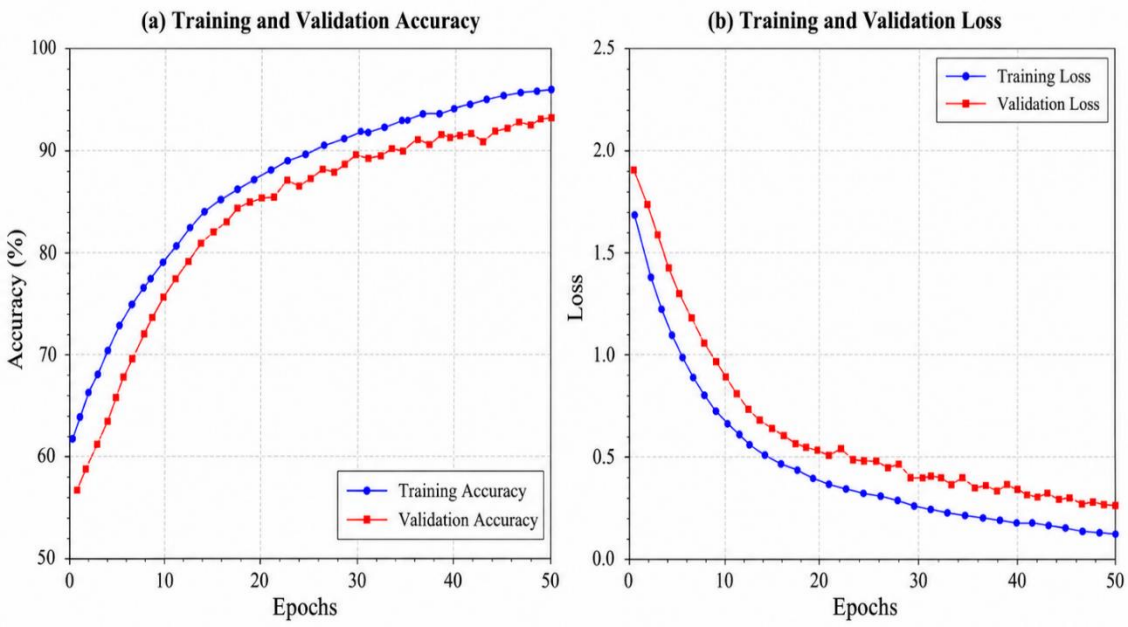

Figure 6. Training and validation performance of the proposed model: (a) accuracy curve and (b) loss curve

Table 7 shows the performance of the proposed model in terms of classification. The model achieved an accuracy of 94.81%, precision of 91.76%, recall/sensitivity of 92.21%, specificity of 95.30%, F1-score of 91.98%, and ROC-AUC of 0.972. The outcome of these results showed that the proposed model was good at distinguishing between the melanoma and non-melanoma lesion. The high recall value indicated that a high proportion of melanoma cases were correctly recalled, which is crucial for clinical screening. The high specificity also suggested the model correctly predicted the majority of the non-melanoma cases, thus minimizing unnecessary false alarms.

Table 7. Classification performance of the proposed model

Metric	Value
Accuracy	94.81%
Precision	91.76%
Recall/Sensitivity	92.21%
Specificity	95.30%
F1-score	91.98%
ROC-AUC	0.972

The proposed framework's confusion matrixes are presented in Figure 7(a). Of the 167 melanoma images in the test set, 154 were correctly identified as melanoma and 13 were wrongly identified as non-melanoma. The 1273 out of 1336 non-melanoma images were correctly classified as non-melanoma, with 63 being misclassified as melanoma. The relatively low rate of false-negative melanomas indicated that the class-balanced learning strategy was more effective at increasing the sensitivity of the model. But there were also 13 cases that were negative that did not classify automatically well, indicating that these lesions look alike and were still a difficult task for automated classification.

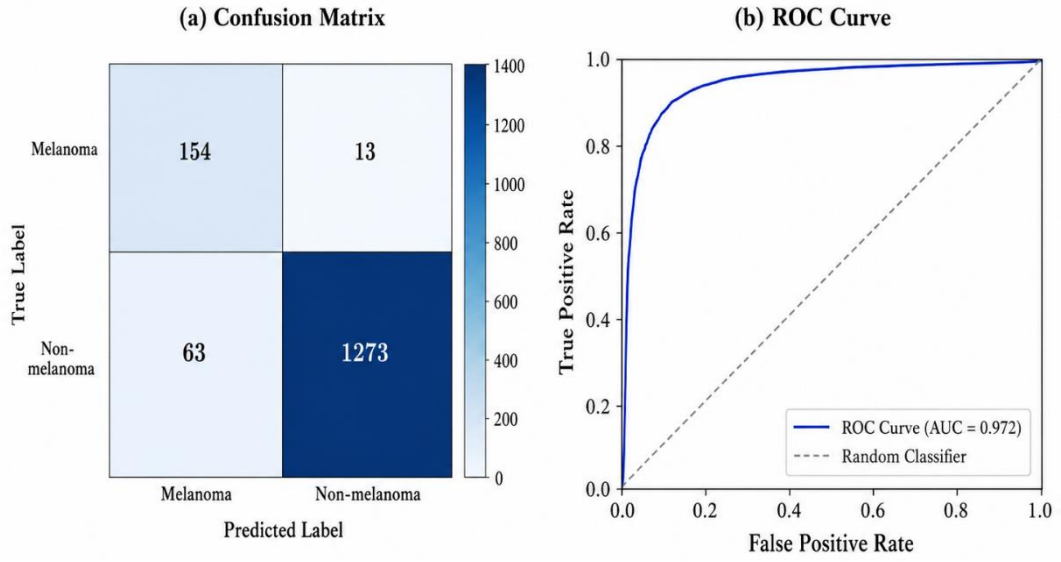

Figure 7. Classification performance of the proposed model: (a) confusion matrix and (b) ROC curve

Table 8 shows the overall values in the confusion matrix. The false negative cases would be more significant than the false positive cases, as melanoma that is diagnosed as non-melanoma could lead to delayed treatment. On the other hand, false-positive cases can result in further clinical screening but are not as severe as the cases of melanoma that are not being identified. Hence, the proposed model was optimized to achieve high sensitivity towards melanoma alongside with good specificity.

Table 8. Confusion matrix values of the proposed model

Actual / Predicted	Melanoma	Non-melanoma
Melanoma	154	13
Non-melanoma	63	1273

The proposed model ROC curve is presented in Figure 7(b). A ROC-AUC value of 0.972 showed the high discriminative power between the melanoma and non-melanoma class. The model showed good classification accuracy at various threshold values, as a ROC-AUC was obtained that was close to 1. This was important as the classification threshold may be changed when needed during clinical screening – either to increase sensitivity or increase specificity.

The proposed framework was evaluated with the standard CNN and transfer learning-based architectures as is seen in Table 9. The comparison consisted of Basic CNN, VGG16, ResNet50, DenseNet121, MobileNetV2, EfficientNetB0, and the algorithm proposed in the paper known as attention-guided CNN. The proposed model had better recall, F1 score and ROC AUC than compared models. This improvement was primarily attributed to the attention-guided feature refinement, class-weighted learning and minority-class augmentation. The proposed model was also well balanced between the sensitivity of melanoma and overall classification reliability, with EfficientNetB0 and DenseNet121 also performing well.

Table 9. Comparative performance analysis with existing CNN models

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC
Basic CNN	89.42%	82.35%	80.84%	90.50%	81.59%	0.914
VGG16	91.28%	85.71%	85.03%	92.06%	85.37%	0.936
ResNet50	92.15%	87.42%	87.43%	92.74%	87.42%	0.948
DenseNet121	93.21%	89.18%	89.82%	93.64%	89.50%	0.958

MobileNetV2	92.67%	88.62%	88.02%	93.25%	88.32%	0.951
EfficientNetB0	93.74%	90.48%	90.42%	94.16%	90.45%	0.963
Proposed model	94.81%	91.76%	92.21%	95.30%	91.98%	0.972

An ablation study was carried out to assess the importance of each part of the proposed mechanism. Table 10 shows the results. The baseline CNN without augmentation performance was worse in terms of recall and F1-score, showing poor generalization. The model performance was better when augmentation was applied, due to the more diversified training data. The inclusion of the attention module also enhanced the performance of recall and ROC-AUC, suggesting that the attention mechanism was useful for giving the model an insight into the diagnostically relevant regions. Class-balanced learning achieved an additional increase in the sensitivity of melanoma detection by minimizing the majority-class bias. Attention, class balancing and Grad-CAM explainability in the complete proposed model yielded the overall best performance.

Table 10. Ablation study of the proposed framework

Model variant	**Accuracy**	**Recall**	**Specificity**	**F1-score**	**ROC-AUC**
CNN without augmentation	88.96%	78.44%	90.27%	79.88%	0.902
CNN + augmentation	91.02%	84.43%	91.84%	84.97%	0.931
CNN + augmentation + attention	93.16%	89.22%	93.65%	89.56%	0.956
CNN + attention + class balancing	94.08%	91.02%	94.46%	90.96%	0.965
Proposed model	94.81%	92.21%	95.30%	91.98%	0.972

The ablation results showed that each component had a significant contribution to the overall performance. Augmentation enhanced robustness against rotation, scale, illumination and position variations. The attention mechanism enabled the model to concentrate on the lesion area instead of the surrounding background skin. Class-balanced learning helped give melanoma samples more weight during training, and also helped to lower the number of false-negative predictions. The explainability module gave visualization of the prediction which increased transparency of the prediction process.

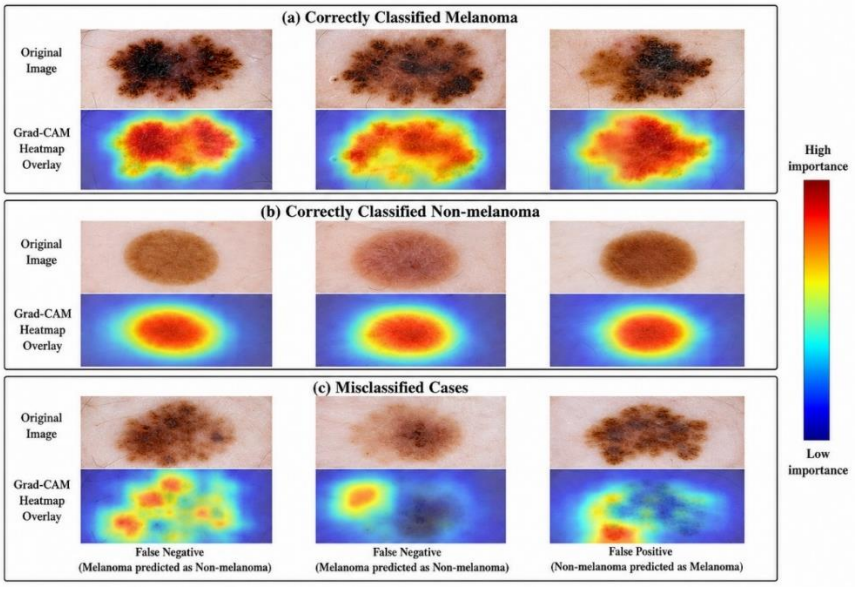

Figure 8. Grad-CAM visualization analysis: (a) correctly classified melanoma, (b) correctly classified non-melanoma, and (c) misclassified cases

The prediction behavior of the proposed model was explained using the Grad-CAM visualization approach. The correctly classified melanoma cases heatmaps are displayed in Figure 8(a). The highlighted areas were primarily located near the edges of the lesions, around areas of irregular pigmentation, around the edges of asymmetric formations and in darker areas within the lesions. This suggested that the model concentrated on clinically relevant image areas and did not include irrelevant background areas. This visualization is helpful as it gives visual evidence that the CNN prediction was made based on lesion-relevant features.

Correctly classified non-melanoma cases are presented using Grad-CAM visualization in Figure 8(b). In such instances, the model emphasized benign patterns of growth and rather regular pigmentation. The heat maps revealed that the choice was primarily based on the lesion area, not on artifacts in the hair, corners of the images, or the surrounding skin. It validated the feature learning guided by attention, which suppressed the non-diagnostic background information.

The misclassified cases were also examined with Grad-CAM as illustrated in Figure 8(c). Low contrast, small lesion area, irregular illumination, and visual similarity to benign nevus were some of the false-negative melanoma cases. In these situations, either a weak distribution of model attention or partial attention toward less discriminative lesion regions occurred. Most false-positive cases were due to benign lesions with melanoma-like characteristics (irregular border, dark pigmentation, and color heterogeneity). This analysis also revealed that the visually ambiguous lesions were still difficult to handle even with the attention-guided CNN model.

Overall results demonstrated that the proposed explainable attention-guided CNN framework achieved better performance in the melanoma classification task than the standard CNN and the transfer learning models. The improvement was primarily attributed to a combination of attention guided feature refinement, class balanced learning and a structured pre-processing. The model had high sensitivity, which is significant to minimize missed cases of melanoma. The ability to discriminate between melanoma and non-melanoma lesions was also good and was confirmed by the high ROC-AUC.

The proposed framework could be a supporting computer-aided diagnostic tool for melanoma screening from a clinical viewpoint. It's not a substitute for dermatologists, but can help them by giving them an extra image-based prediction and interpretable heatmap. The attention-guided design allowed the model to concentrate on the diagnostically relevant regions of the lesion and the Grad-CAM visualization techniques helped make it easier to understand. The combination is crucial, as clinical application of a diagnosis based on AI requires not only accuracy but also reliability, sensitivity and interpretability.

Although these are positive developments, there were some limitations to the study. The model was created on a public dermoscopic database, and should be validated with multi-center clinical databases. Other factors that can affect the performance are variations in the image acquisition conditions, lesion types, skin tone diversity, and imaging artifacts. Furthermore, grad-CAM gave a visual explanation but was not a full clinical explanation. Further research is needed to conduct external validation, dermatologist in the loop assessments, incorporate patient data to improve the system, and release a lightweight version for real time melanoma screening.

6. CONCLUSION

In the present study, an explainable attention-guided convolutional neural network (CNN) framework was developed for the detection and classification of melanoma from dermoscopic skin lesion images. The dataset, HAM10000, was split into melanoma and non-melanoma classes and the proposed framework used a combination of preprocessing, augmentation, class-balanced learning, attention-based feature refinement and Grad-CAM visualization. The attention module directed the network to regions of the lesion that were important for diagnosing the type, and the class-weighted learning decreased the learning bias towards the more dominant class of non-melanoma skin cancer. This design overcame significant drawbacks of traditional CNN-based melanoma classification, in particular, low interpretability, imbalanced classes, and poor representation of lesions.

In quantitative terms, the model proposed achieved an accuracy of 94.81%, a precision of 91.76%, a recall/sensitivity of 92.21%, a specificity of 95.30%, a F1-score of 91.98% and an ROC-AUC of 0.972. In the independent test set, the model achieved an accuracy of 91.6% in classification with 154 melanomas correctly identified, demonstrating its ability to detect malignant lesions successfully. The comparative results showed that the proposed framework outperforms the standard CNN and the transfer learning model, and the ablation study showed that the augmentation step, the attention learning step, and the class balancing step provided a step-by-step enhancement in the performance of the model. The qualitatively analysed heatmaps of Grad-CAM showed that the proposed model primarily targeted regions of the image that were clinically relevant, such as borders of the

lesions, irregular pigmentation, asymmetry, variations in the internal texture of the images. This made the prediction process more transparent and alleviated the limitation of CNN-based diagnosis as a black-box.

The results indicated that the model proposed in the paper could serve as a helpful CA-Diagnostic system for the detection of Melanoma and be considered as an explanation-based image interpretation approach. It can help dermatologists with the classification output as well as the visualization of the predicted decision.It can support dermatologists in both classification output and in the visualization of the predicted decision. The study only included a public dermoscopic dataset, however, and needed to be expanded to multi-center clinical datasets to be validated. Future work should comprise external clinical validation, dermatologist-in-the-loop assessment, incorporation of patient data, real-time deployment, and extending to multi-class skin lesion classification.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, Art. no. 180161, 2018.

[4] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration," arXiv preprint arXiv:1902.03368, 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. International Conference on Machine Learning, 2019, pp. 6105–6114.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[9] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. European Conference on Computer Vision, 2018, pp. 3–19.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[13] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," IEEE Transactions on Medical Imaging, vol. 36, no. 4, pp. 994–1004, 2017.

[14] N. C. F. Codella et al., "Deep learning ensembles for melanoma recognition in dermoscopy images," IBM Journal of Research and Development, vol. 61, no. 4/5, pp. 5:1–5:15, 2017.

[15] A. A. Adegun and S. Viriri, "Deep learning-based system for automatic melanoma detection," IEEE Access, vol. 8, pp. 7160–7172, 2020.

[16] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," Journal of Biomedical Informatics, vol. 86, pp. 25–32, 2018.

[17] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with AlexNet," PLOS ONE, vol. 14, no. 5, Art. no. e0217293, 2019.

[18] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion classification using convolutional neural networks with patch-based attention and diagnosis-guided loss weighting," IEEE Transactions on Biomedical Engineering, vol. 67, no. 2, pp. 495–503, 2020.

[19] S. S. Chaturvedi, K. Gupta, and P. S. Prasad, "Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using MobileNet," Advanced Engineering Informatics, vol. 43, Art. no. 101037, 2020.

[20] R. Kaur, H. GholamHosseini, and R. Sinha, "Melanoma classification using a novel deep convolutional neural network with dermoscopic images," Sensors, vol. 22, no. 3, Art. no. 1134, 2022.

[21] Nannuri S., Gantla H. R., Ananya D., Vennela A., Bhuvana B., Neha G. (2026). Automated Brain Tumor Segmentationin Mri Via U-Net-Based Deep Neural Networks. International Journal of Engineering Sciences & Research Technology, 15(4), 1–8. https://doi.org/10.64149/j.ijesrt.15.4.1-8

[22] X. He, Y. Wang, S. Zhao, and J. Zhang, "Deep metric attention learning for skin lesion classification in dermoscopy images," Complex & Intelligent Systems, vol. 8, pp. 1487–1504, 2022.

[23] A. Chaturvedi, A. Tiwari, and V. Janghel, "A brief review on skin cancer diagnostic techniques using artificial intelligence," Artificial Intelligence Review, vol. 53, pp. 4441–4466, 2020.

[24] Alkhatib A. J., (2026). Integrated Computational Retinal Vascular Morphometry Using Fractal, Skeleton-Based, and Phi-Related Features for Diabetic Retinopathy and Glaucoma. International Journal of Artificial Intelligence, Machine Learning and Data Analytics, 1(1), 43-50. https://www.ijaimda.org/index.php/ijaimda/article/view/3

[25] O. Attallah, "Skin-CAD: Explainable deep learning classification of skin cancer from dermoscopic images," Computers in Biology and Medicine, vol. 179, Art. no. 108822, 2024.