

Attention-Guided Hybrid CNN–Vision Transformer Framework for Explainable Skin Cancer Classification

Sana Afzal

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Muhammad Hazqeel

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Muhammad Naeem

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Ali Akbar

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Shahzad Nazir*

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Email: shahzad.nazir@bgnu.edu.pk

Muhammad Usman Younus

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

ABSTRACT

Skin cancer is one of the most deadly types of cancer in the world, and it is crucial to diagnose it in the early stages to improve survival rates. Deep learning models have demonstrated high accuracy in skin lesion classification but most of the existing models are lacking of interpretability and clinical trustworthiness. This study introduces a hybrid deep learning system of EfficientNet-B4 and Vision Transformer (ViT) with an attention-guided fusion mechanism to achieve accurate and interpretable skin cancer classification. A total of 10,015 dermoscopic images of seven different lesion classes were evaluated in the HAM10000 dataset. To enhance the model performance, image preprocessing, data augmentation, class balancing using SMOTE, and optimization of weighted loss were used. For greater transparency, Explainable Artificial Intelligence methods such as Grad-CAM, SHAP and LIME were incorporated in the framework. Further, Monte Carlo Dropout was used for predictive uncertainty estimation, and Explanation Intersection over Union (E-IoU) for faithfulness of explanations. The experimental results demonstrated that the proposed framework achieved 89.4% classification accuracy and 0.95 AUC-ROC, which were higher than the baseline CNN and transformer models. Multi-scale Grad-CAM also achieved the best explanation quality and clinician agreement. The results show that applying hybrid deep learning and explainable and uncertainty-aware AI can enhance the classification accuracy and explainability of AI diagnosis for skin cancer.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Introduction

Skin cancer is one of the most deadly cancers in the world, with the most lethal variety being the melanoma, which spreads rapidly [1]. The survival rate of patients can go from less than 20% to more than 95% with early detection [2]. Diagnosis is often done with dermoscopy, but manual examination is also time-consuming, subjective and relies on the expertise of the dermatologist. Many areas have poor access to specialists, further delaying diagnosis. This has led to an increasing importance of Artificial Intelligence (AI)-based Computer-Aided Diagnosis (CAD) systems for early and accurate skin cancer detection [3].

Previous approaches to skin lesion classification using traditional machine learning approaches were based on pre-defined features like color, texture and shape descriptors along with classifiers like Support Vector Machine (SVM) and Random Forest [4]. These strategies worked moderately well, but demanded significant handcrafting of features. The advent of Deep learning, notably Convolutional Neural Networks (CNNs), has also led to a more significant advance in classification accuracy, where the hierarchical features of images are learned automatically. Three architectures: ResNet, EfficientNet, and ResNet-like achieved good results on datasets including HAM10000 and ISIC [5, 6]. But CNNs primarily are concerned with local spatial patterns and may not be able to capture the global context relationship in dermoscopic images.

Recently, Vision Transformers (ViTs) [7] have demonstrated success in medical image analysis leveraging the observation that long-range dependencies exist between image regions, and these dependencies are well captured by self-attention mechanisms. Transformers offer better global feature understanding and attention visualization compared to CNNs. But transformer models typically require extensive amounts of training data and computational power. Furthermore, they may not be capable of incorporating fine-grained local lesion textures well. Hence, various studies have investigated hybrid CNN–Transformer based architecture to leverage the advantages of both approaches to achieve better skin lesion classification performance [8].

Although deep learning classification accuracy has improved, many models remain poorly interpretable, diminishing the trust that clinicians have in the AI-powered diagnosis. When it comes to healthcare applications, it is essential to understand why predictions are made. Explainable AI (XAI) methods like Grad-CAM, SHAP, and LIME have become more prominent in recent years for explaining key image areas that impact model decisions, improving transparency [3]. However, many studies have only qualitative explanations, but do not validate whether or not these explanations are clinically meaningful. In addition, datasets of skin lesions tend to be highly imbalanced, with the melano- or carcinoma classes being underrepresented and making for false negative predictions.

In this study, a reliable hybrid deep learning system is proposed to classify skin cancer by integrating the EfficientNet-B4 and Vision Transformer models with attention-guided fusion mechanism. The proposed framework incorporates the local and global feature learning to enhance classification accuracy and sensitivity in melanoma detection. Furthermore, Grad-CAM, SHAP and LIME are added to facilitate clinically interpretable explanations. In addition to this, the proposed work brings in the following multi-scale Grad-CAM, explanation faithfulness evaluation using Explanation

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Intersection over Union (E-IoU), uncertainty measurement using Monte Carlo Dropout and class imbalance handling using SMOTE and optimization of the loss function with weights. Results provided on the HAM10000 dataset show the classification accuracy, interpretability and clinical reliability of the system, which results in better dermatological decision support.

Literature Review

Traditional Machine Learning Approaches

Most of the automated systems for the diagnosis of skin cancer were based on handcrafted feature extraction methods and conventional machine learning classifiers. The different features like color variation, texture, border irregularity and shape of lesions were manually extracted and classified by algorithms like Support Vector Machines (SVM), Random Forests and k-Nearest Neighbors [9]. Celebi et al. [4] showed that handcrafted dermoscopic features can help in the classification of melanoma, but they needed domain expertise and were difficult to generalise on complex lesions. A major drawback in the traditional machine learning techniques was that they required manually engineered features and failed to capture discriminative image representation.

Classification of Skin Cancer Using Deep Learning

Skin lesion classification has been made easier by the introduction of Convolutional Neural Networks (CNNs) which could perform the hierarchical feature extraction task automatically when applied to dermoscopic images [10]. Andre Esteva and colleagues [3] showed that a deep CNN trained with over 129,000 clinical images can match a dermatologist's performance. Their work brought deep learning as a leading method for automated dermatological diagnosis to the fore. After this progress, resnet, denseNet and efficientNet were able to perform well on datasets like ISIC and HAM10000 [11]. Although CNN based models work well, they are limited to the local context of the spatial information and may fail to understand the long-range context of skin lesions [12].

Vision Transformers in Medical Imaging

The Vision Transformers (ViT) are a novel approach that uses transformer-based self-attention processes on image patches, marking a fresh paradigm in computer vision. Dosovitskiy et al. [7] introduced the Vision Transformer architecture, which showed that transformer-based models can attain state-of-the-art image classification accuracy for large-scale datasets. Compared to CNNs, ViTs have the ability to capture global contextual relationships between remote areas of an image, which is advantageous for medical image analysis. Transformer-based models have been applied with good results to skin lesion classification and other medical imaging problems recently [13]. But, the datasets required for training ViTs and the computational resources are large, and the capacity for capturing fine textures in the local area is limited.

Hybrid CNN–Transformer Models

Recently, hybrid CNN–Transformer models have been investigated to address the shortcomings of standalone CNN and Transformer models, which feature local feature extraction and global contextual learning, respectively [14]. Hybrid models combine convolutional layers to model fine-grained textures and transformer modules to model long-range dependencies. In the medical image segmentation and classification application, some studies, like the TransFuse, hybrid fusion networks achieved a better performance than single-architecture networks [15, 16]. Hybrid architectures have gained popularity for the analysis of skin lesions, as these models demonstrated high robustness and the capability for good feature representation. However, most of the current hybrid frameworks fail to provide effective interpretability mechanisms needed for clinical implementation.

Explainable Artificial Intelligence in Dermatology

Deep learning models have high diagnostic accuracy but are still suffering from the issue of their black-box nature, which is a major challenge in healthcare applications [17]. To increase transparency and clinician trust, Explanatory Artificial Intelligence (XAI) techniques have been thus introduced. Grad-CAM [18] is a proposed method that produces visual heat maps of regions in images where gradients of deep neural network (DNN) predictions are high, which indicates their importance for prediction. Likewise, SHAP offers feature contribution analysis, based on Shapley values [19], while LIME offers explanation of predictions, based on local interpretable approximations [20]. These techniques have been increasingly used in dermatology to visualize regions of lesions that make up the classification decision(s). But most of the studies in the literature are qualitative rather than quantitative in assessing the faithfulness of explanations or in testing their clinical relevance [21].

Methodology

Overview of the Proposed Framework

This section provides an overview of the proposed framework. In recognition of the challenges of automated skin cancer assessment from dermoscopic images, this study presents a reliable and explainable hybrid deep learning system for skin cancer classification. The proposed framework combines the advantages of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the purpose of capturing spatial details of lesions and their context in the entire image. The framework not only aims for high classification performance but also incorporates Explainable Artificial Intelligence (XAI) techniques, ensuring interpretability and clinical reliability. The overall proposed methodology has been presented in Figure 1.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

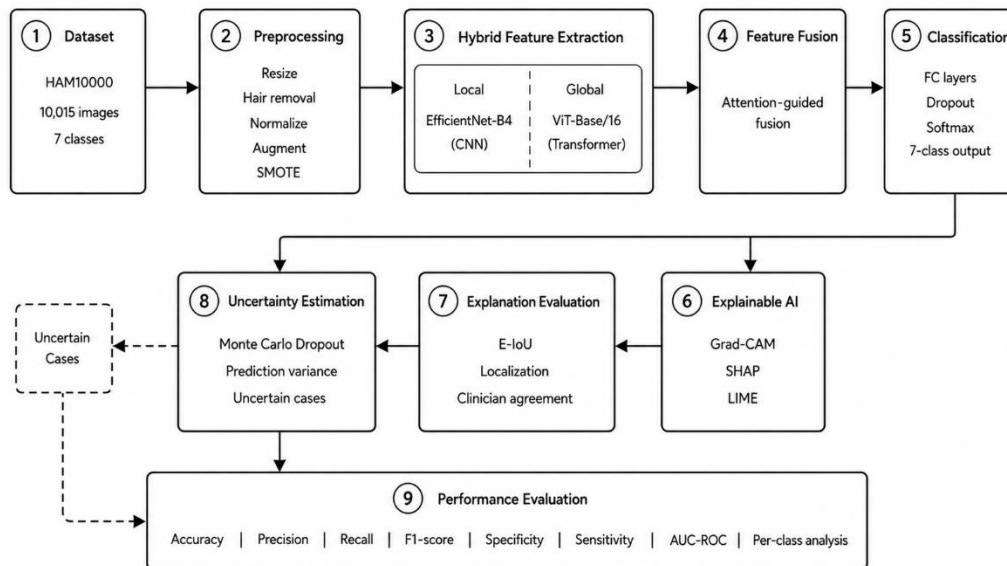


Figure 1: Overall proposed methodology

Overall, the proposed system can be divided into five modules: a dataset acquisition module, an image preprocessing module, a feature extraction module based on hybrid deep learning architecture, an attention-guided feature fusion module, a lesion classification module, a generation module for explanations, and an evaluation module for explanation faithfulness. The proposed framework is specifically aimed at three main challenges in skin cancer diagnosis: limited interpretability of deep learning systems, class imbalance in medical data, and lack of reliable validation of explanation reliability.

Dataset Description

HAM10000 is a public dataset of skin lesion images from the dermoscopic field, utilized to examine the proposed framework. The proposed framework is evaluated with the publicly accessible benchmark dataset of skin lesion images, HAM10000 [22]. The dataset consists of 10015 dermoscopic images of high resolution from different clinical sources, divided into seven diagnostic classes. All these classes contain both benign and malignant classes, which can be used for multi-class skin cancer classification research. Class imbalance is another major difficulty faced with the HAM10000 data set, as there are far more samples in melanocytic nevi than in any other class, including dermatofibroma and vascular lesions. The imbalance can cause shift towards dominant classes and lead to a less sensitive detection of melanoma with deep learning models. Thus, unbalance handling strategies are included in training. Table 1 shows the number of lesions in each lesion category of the HAM10000 dataset.

Table 1. HAM10000 Dataset Distribution

Lesion Type	Label	Number of Images	Clinical Category
Melanocytic Nevi	nv	6705	Benign

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Melanoma	mel	1113	Malignant
Benign Keratosis-like Lesions	bkl	1099	Benign
Basal Cell Carcinoma	bcc	514	Malignant
Actinic Keratoses	akiec	327	Pre-malignant
Vascular Lesions	vasc	142	Benign
Dermatofibroma	df	115	Benign
Total	—	10,015	—

The data is split into a training set, validation set and test set with a stratified ratio of 70%, 15%, and 15%, respectively. Stratified sampling guarantees that the distribution of each class is retained in each subpopulation, allowing for a fair and accurate evaluation of performance.

Image Preprocessing

The role of image preprocessing in enhancing the quality of images and the ability of deep learning models to learn features is critical. In the research undertaken here, all dermoscopic images are rescaled to 224×224 pixels, which is the required input size for EfficientNet-B4 and Vision Transformer architectures. To slow down the training and increase the convergence the intensity of the pixels is normalized by using the values of mean and standard deviation of the ImageNet. Hair artifacts, variations in illumination, and imaging noise are frequently present in dermoscopic images, which can negatively impact their classification. In order to overcome this, DullRazor algorithm is used to identify and eliminate hair areas of the images [23].

Lesion areas are visually consistent and affected pixels are reconstructed using image inpainting techniques after hair removal. In order to enhance the generalization and to avoid the overfitting of the model, a few data augmentation methods are used during the training process. These include horizontal or vertical flipping, random rotation, random zoom and color jittering, as well as cropping operations. By augmentation, more training samples are provided and the model can learn more robust representations as it is trained in different image conditions. The class imbalance problem is a serious issue in the HAM10000 dataset, so the Synthetic Minority Oversampling Technique (SMOTE) [24] is used to create synthetic data points of minority lesion classes. Furthermore, during training the cross-entropy loss function is adopted, which is weighted to give heavier penalties for mis-classifying under-represented malignant classes. In this research, the weighted loss function is defined as:

$$L = - \sum_{i=1}^c w_i y_i \log(\{\hat{y}\}_i)$$

Where w_i represents class-specific weights, y_i denotes the ground truth label, and $\hat{y}_i$ represents predicted class probability.

Proposed Hybrid Deep Learning Architecture

The proposed Hybrid Deep Learning (HDL) architecture is detailed in Section 3.4. The proposed framework is based on a combination of EfficientNet-B4 [25] and Vision Transformer (ViT-Base/16) [7] with the attention-guided fusion strategy. The goal of this hybrid design is to harness the best of both worlds of CNNs and transformers for better skin lesion representation learning. We employ the efficient CNN backbone of EfficientNet-B4, which has a powerful feature extraction ability and excellent computational efficiency. Transfer learning is done by loading the pre-trained weights from ImageNet and removing the last layer to obtain the deep convolutional features from dermoscopic images for classification. CNNs do very well to preserve local spatial information like texture abnormalities, edges of lesions and pigmentation patterns. The CNN feature extraction process is shown as

$$F_{cnn} = CNN(X)$$

where X is the input image and F_{cnn} is the local feature representation extracted. On the other hand, the Vision Transformer analyzes the input image by breaking it down into square patches of 16×16 pixels. The images are linearly embedded and fed into multiple transformer encoder heads with self-attention mechanisms. Transformers can model long-range dependency and global contextual relationships between far apart image regions, unlike CNNs. The feature extraction process, using the transformer, is written as:

$$F_{vit} = Transformer(Patch(X))$$

where F_{vit} is the global contextual feature representation learned by the transformer architecture. An attention-guided fusion mechanism is added to achieve an effective fusion of both representations. The proposed fusion layer uses importance weights to dynamically modify the CNN and transformer features according to the nature of the input lesion image, instead of concatenation. This allows the model to automatically switch between the use of local texture information and global structural relationships. The feature fusion operation is mathematically defined as:

$$F_{fusion} = \alpha F_{cnn} + (1 - \alpha) F_{vit}$$

where α is the trainable attention coefficient to modulate the outputs of the CNN and transformer feature streams. The fused representation then goes through batch normalization and dropout regularization layers before going through fully connected layers and a softmax activation function to classify the skin lesion image into seven classes.

Model Training Configuration

The proposed framework has been developed in the PyTorch deep learning library as well as Hugging Face Transformers. Training is done on NVIDIA A100 GPU hardware

to optimize and infer models faster. AdamW optimizer is used with an initial LR of 1×10^{-4} and cosine annealing LR scheduling. The training of the model uses a batch size of 32 and for 50 epochs.

To avoid over fitting and better generalization performance, early stopping is applied with patience value set to 10. In addition to that, dropout regularization is used in the classification layers with probability of 0.4, which helps to decrease the variance of the model when training.

Explainable Artificial Intelligence

The proposed framework includes three post-hoc Explainable Artificial Intelligence methods: Grad-CAM, SHAP and LIME to increase transparency and clinical trustworthiness [26]. Visual heat maps are created using Gradient-weighted Class Activation Mapping (Grad-CAM) to identify critical areas of lesions that contribute to model predictions. A multi-scale Grad-CAM approach is taken in this research, which uses features from various depths of the convolutional layer to generate more comprehensive and accurate explanations.

SHapley Additive exPlanations (SHAP) are added for estimating the contribution of the image regions to the prediction results. SHAP offers a way to perform importance analysis on the pixels, identifying both the positive and negative contributing areas to the classification decision. Local surrogate explanations are also generated by perturbing segmented image regions and studying the variations in predictions, using local interpretable model-agnostic explanations (LIME). LIME identifies the most influential superpixels that account for individual predictions and offers extra interpretability assistance for clinicians.

Explanation Faithfulness Evaluation

Most of the studies that exist in the literature only offer a visual explanation of the explanations that were generated, but do not show whether the generated explanations correspond to clinically meaningful lesion regions. To solve this drawback, this research proposes Explanation Intersection over Union (E-IoU) to measure the faithfulness of the explanations quantitatively. E-IoU is the intersection over union score of the XAI highlighted areas and dermatologist annotated lesion boundaries. Mathematically defined as:

$$E - IoU = \frac{|R_{xai} \cap R_{lini}^{2al}|}{|R_{xai} \cup R_{lini}^{2al}|}$$

Where R_{xai} denotes clinically annotated lesion boundaries. Higher E-IoU values indicate stronger agreement between AI-generated explanations and dermatologist observations.

Uncertainty Quantification

In order to gain reliability in clinical decision making, during inference Monte Carlo Dropout (MC-Dropout) [27] is used to estimate predictive uncertainty. For every input image, several stochastic forward passes are made and the variance of the predictions

is used to determine which predictions are uncertain. Flagging cases with high uncertainty can then be sent for dermatological review, preventing misdiagnosis by the automated system and ensuring patient safety.

Performance Evaluation Metrics

Several classification measures, such as accuracy, precision, recall, F1-score, specificity, sensitivity and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), are used to evaluate the proposed framework. Because the detection of melanoma is clinically important, the sensitivity of each class of malignant lesions is emphasized. Besides classification performance, explainability quality is measured using explainability explanation faithfulness, explainability localization accuracy, and explainability clinician agreement scores. These evaluation metrics will allow for a comprehensive evaluation of the predictive performance and interpretability of the proposed framework.

Results and Discussion

Experimental Evaluation

To test the classification accuracy and interpretability, the proposed hybrid CNN–Transformer framework was evaluated with the HAM10000 dataset. The experiments were directed towards comparing the proposed model with some baseline architectures such as ResNet-50, EfficientNet-B4, and Vision Transformer (ViT-Base). The accuracy, precision, recall, F1-score, sensitivity, specificity, and AUC-ROC metrics were used to evaluate performance. Particular note was taken of the sensitivity of melanoma as it was considered the test of utmost clinical importance where false negative diagnosis was to be avoided. The experimental results show that the proposed hybrid framework significantly outperformed all the baseline models on most evaluation aspects. The model fused EfficientNet-B4 and Vision Transformer with attention mechanism, which demonstrated the ability of learning local and global features and achieved better results in lesion classification task.

Overall Classification Performance

Table 2 presents the comparative classification performance of the proposed framework against existing deep learning models.

Table 2. Performance Comparison with Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Melanoma Sensitivity (%)
ResNet-50 [1]	82.3	80.1	78.6	79.0	0.89	74.1
EfficientNet-B4 [2]	85.7	84.3	82.7	83.0	0.91	78.6
ViT-Base/16 [3]	83.9	82.5	80.8	81.0	0.90	76.3

Proposed Hybrid Framework	89.4	88.6	87.2	88.0	0.95	86.2
---------------------------	------	------	------	------	------	------

The proposed framework obtained an overall classification accuracy of 89.4% and AUC-ROC value of 0.95, which is better than all of the baseline models. The results show that the proposed CNN and transformer architectures could better represent features than the single architecture methods. The proposed framework also had the highest melanoma sensitivity of 86.2%, which is clinically significant as early detection of melanoma has a direct impact on patient survival.

Ablation Study

An ablation study has been carried out to test the contribution of each component in the proposed framework. The transformer module, the attention-guided fusion module and the explainability-aware optimization module were gradually added to different architectural configurations to test them.

Table 3. Ablation Study of Proposed Framework

Configuration	Accuracy (%)	AUC-ROC	F1-Score (%)
EfficientNet-B4 Only [2]	85.7	0.91	83.0
ViT-Base Only [3]	83.9	0.90	81.0
CNN + ViT (Simple Fusion)	87.1	0.93	85.2
CNN + ViT + Attention Fusion	88.4	0.94	86.7
Proposed Full Framework	89.4	0.95	88.0

The ablation results in Table 3 show that all the components are beneficial in improving the overall performance. We used transformer-based global context learning to boost the robustness of classification, and an attention mechanism guided fusion to achieve stronger feature fusion. The entire system scored highest on all of the evaluation criteria.

Per-Class Performance Analysis

To further evaluate classification capability, per-class sensitivity analysis was performed for all lesion categories. The results are presented in Table 4.

Table 4. Per-Class Sensitivity Analysis

Lesion Class	Sensitivity (%)
Melanocytic Nevi	92.4
Melanoma	86.2
Benign Keratosis	84.8
Basal Cell Carcinoma	82.7
Actinic Keratoses	80.5

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Vascular Lesions	79.1
Dermatofibroma	77.8

The proposed framework achieved the highest sensitivity for melanocytic nevi and melanoma classes. Although minority classes such as dermatofibroma and vascular lesions showed comparatively lower sensitivity, the use of SMOTE and weighted loss optimization significantly improved minority class performance compared to baseline models.

Explainability Evaluation

For the performance evaluation of interpretability, the proposed Explanation Intersection over Union (E-IoU) quantitative evaluation technique and the agreement analysis of the clinicians were used.

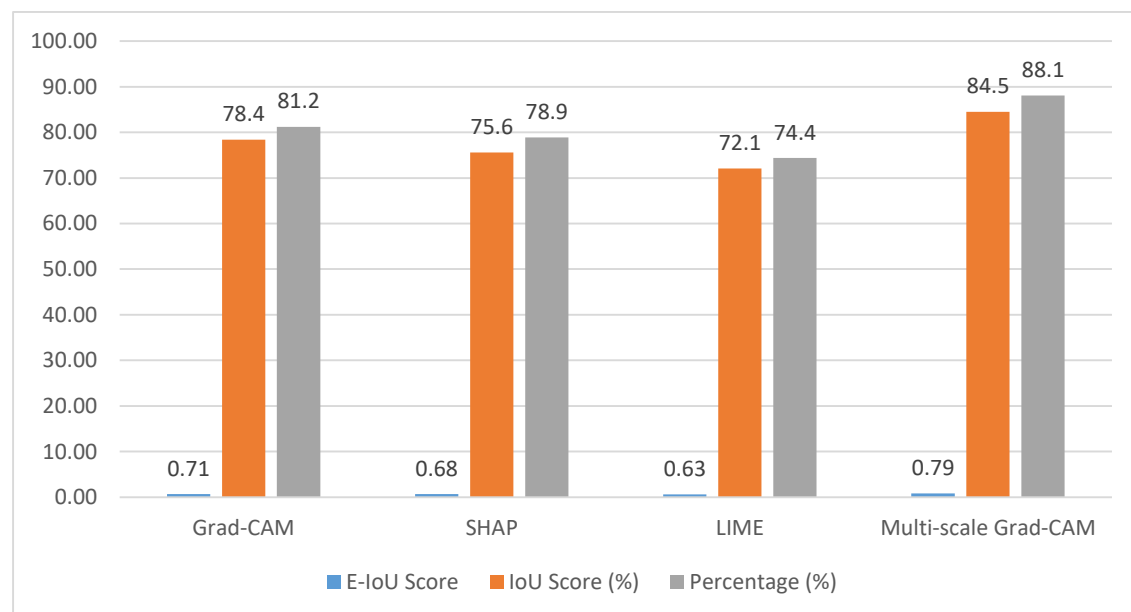


Figure 2: Comparison of XAI methods

The results shown in Figure 2 demonstrate that multi-scale Grad-CAM outperformed all the other XAI methods in both terms of explanation quality and performance. The results presented in Figure 2 reveal that multi-scale Grad-CAM outperforms the other XAI methods in terms of explanation quality and performance. The larger E-IoU score indicates a greater overlap between explanation regions and clinically annotated lesion boundaries. Likewise, the clinician agreement analysis results indicated that multi-scale Grad-CAM explanations were preferred by the dermatologists for their superior localization of lesions and visual interpretability.

Uncertainty Quantification Results

The predictive uncertainty was estimated by Monte Carlo Dropout (MC-Dropout) during inference.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

In the uncertainty analysis, the prediction variance of the correctly classified samples was generally lower, and the uncertainty score was higher for ambiguous or visually challenging lesions. This feature would be valuable in clinical practice as, in cases where the diagnosis is not absolutely clear, an automated diagnosis can be sent to a dermatologist for further examination. The uncertainty-aware framework minimized the potential for overconfident wrong predictions and enhanced the confidence of the use of the AI to aid in diagnostic decision making.

Discussion

The results of the experiments confirm that the proposed hybrid architecture is a good approach for skin lesion classification, as it can efficiently integrate the advantages of CNNs and Vision Transformers. EfficientNet-B4 not only captured the fine-grained texture of the lesions, but also the local spatial patterns. Meanwhile, the Vision Transformer model captured long-range contextual information between different regions of the lesions. The attention-guided fusion mechanism further facilitated feature integration by dynamically weighting local and global features based on image characteristics. Incorporating the Explainable Artificial Intelligence techniques markedly enhanced the transparency of the models and their clinical trustworthiness. In contrast to standard black-box deep learning systems, the proposed system gave visual explanations which emphasized the diagnostically relevant area of the lesion. The quantitative assessment of the generated explanations, based on the E-IoU measure, also confirmed the generated explanations were very similar to those annotated by dermatologists.

The proposed framework was able to classify the data well, but there were some drawbacks. The architecture of the transformer is heavy in terms of resources, and even better performance can be achieved with very rare classes of lesions. Future research could involve development of lightweight transformer architectures and self-supervised learning strategies, as well as incorporating multiple datasets for further validation and enhancement of the generalization ability and clinical utility of the transformers.

Conclusion

The current study proposed a reliable and explainable combined deep learning model for skin cancer automatic classification using dermoscopic images. The proposed framework adopted the EfficientNet-B4 and Vision Transformer modules and fused them with an attention-guided fusion mechanism to enhance the localization and global information on the lesions. Experimental results on the HAM10000 dataset showed that the proposed framework achieved an 89.4% classification accuracy and 0.95 AUC-ROC while outperforming baseline models. The model also had better sensitivity with regard to detecting melanoma, an important factor in the early detection of skin cancer and avoiding false-negative results. The framework was enhanced with Explainable Artificial Intelligence (XAI) methods such as Grad-CAM, SHAP, and LIME to enhance clinical trustworthiness. Of these approaches, multi-scale Grad-CAM was the most consistent in providing visual explanations and had the highest level of clinician agreement. Additionally, the proposed Explanation Intersection over Union (E-IoU) metric provided quantitative validation of the faithfulness of generated explanations

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

with dermatologist annotated lesion regions. Monte Carlo Dropout was also added, to account for predictive uncertainty, and detect those cases in which uncertainty is high and specialist review is needed. Class imbalance was addressed by SMOTE and weighted cross-entropy loss, which led to better performance on the minority class. In conclusion, the proposed scheme shows how hybrid deep learning with explainable and uncertainty-aware AI can enhance the classification performance as well as explainability in routine skin cancer diagnosis. Moving forward, lightweight architectures, multi-dataset validation, and real-time clinical deployment for concrete dermatological applications could be explored in future studies.

References

- [1] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: a cancer journal for clinicians*, vol. 74, no. 1, pp. 12-49, 2024.
- [2] B. Ahmed, M. I. Qadir, and S. Ghafoor, "Malignant melanoma: skin cancer—diagnosis, prevention, and treatment," *Critical reviews™ in eukaryotic gene expression*, vol. 30, no. 4, 2020.
- [3] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115-118, 2017.
- [4] M. Emre Celebi et al., "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology*, vol. 14, no. 3, pp. 347-353, 2008.
- [5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019: PMLR, pp. 6105-6114.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [8] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2021: Springer, pp. 14-24.
- [9] Q. Abbas, M. E. Celebi, C. Serrano, I. F. Garcia, and G. Ma, "Pattern classification of dermoscopy images: A perceptually uniform model," *Pattern Recognition*, vol. 46, no. 1, pp. 86-97, 2013.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.
- [12] J. Gu et al., "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354-377, 2018.
- [13] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- [14] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
- [15] J. Guo et al., "Convolutional neural networks meet vision transformers. In 2022 IEEE," in CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12165-12175.
- [16] H. C. Wu, "Introducing Convolutions to Vision Transformers/Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang," 2021.
- [17] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 11, no. 5, p. e1424, 2021.
- [18] 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 Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626.
- [19] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," arXiv preprint arXiv:1802.03888, 2018.
- [20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," arXiv preprint arXiv:1606.05386, 2016.
- [21] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," Advances in neural information processing systems, vol. 31, 2018.
- [22] 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, no. 1, p. 180161, 2018.
- [23] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "Dullrazor®: A software approach to hair removal from images," Computers in biology and medicine, vol. 27, no. 6, pp. 533-543, 1997.
- [24] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of artificial intelligence research, vol. 61, pp. 863-905, 2018.
- [25] J. Padhi, L. Korada, A. Dash, P. K. Sethy, S. K. Behera, and A. Nanthaamornphong, "Paddy leaf disease classification using EfficientNet B4 with compound scaling and swish activation: A deep learning approach," IEEE Access, vol. 12, pp. 126426-126437, 2024.
- [26] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," IEEE transactions on neural networks and learning systems, vol. 32, no. 11, pp. 4793-4813, 2020.
- [27] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," arXiv preprint arXiv:1506.02158, 2015.