Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

WaveInvNet: A Wavelet-Separated Dynamic Involution Network for Multimodal Skin Lesion Classification

Mohamed Amine Ibrahimi

mohamedamine.ibrahimi@usmba.ac.ma

Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University

Abdelghafour Abbad

Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University

Rachid Ben Abbou

Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University

Khalid Abbad

Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University

Research Article

Keywords: Skin lesion classification, Deep learning, Involution, Stationary Haar Wavelet Transform, Multimodal learning

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10678636/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

WaveInvNet: A Wavelet-Separated Dynamic Involution Network for Multimodal Skin Lesion Classification

Mohamed Amine Ibrahimi[1*], Abdelghafour Abbad[1†], Rachid Ben Abbou[1†], Khalid Abbad[1†]

[1*]Laboratory of Intelligent Systems and Applications, Department of Computer Science, Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University, Fez, 30000, Morocco.

*Corresponding author(s). E-mail(s): mohamedamine.ibrahimi@usmba.ac.ma;
Contributing authors: Abdelghafour.abbad@usmba.ac.ma; rachid.benabbou@usmba.ac.ma; khalid.abbad@usmba.ac.ma;
†These authors contributed equally to this work.

Abstract

As the global burden of skin cancer continues to rise, early and accurate diagnosis has become increasingly critical for timely intervention and reducing mortality. Recent advances in artificial intelligence, particularly deep learning, have shown considerable potential for automating dermoscopic image analysis and supporting accurate skin lesion diagnosis. However, conventional convolutional neural networks rely on spatially shared kernels that process all image regions uniformly, limiting their ability to adapt to heterogeneous lesion structures and capture subtle diagnostic features. In this study, we propose WaveInvNet, a lightweight multimodal architecture that integrates dermoscopic images with patient metadata through a novel Wavelet-Separated Dynamic Involution (WSDI) block. Using a stationary Haar wavelet transform, the WSDI block separates structural information from fine-grained details and employs these complementary components to generate spatially adaptive involution kernels. The blocks are organized within a four-stage hierarchical encoder to progressively extract frequency-aware visual representations. These features are subsequently concatenated with encoded patient metadata before final prediction. Experimental evaluation on the HAM10000 dataset showed that WaveInvNet outperformed the compared deep learning architectures, achieving an accuracy of 93.63%, an AUC of 98.46%,

and an F1-score of 94.81%, while requiring a relatively small number of trainable parameters.

Keywords: Skin lesion classification, Deep learning, Involution, Stationary Haar Wavelet Transform, Multimodal learning

1 Introduction

Skin cancer is one of the most commonly diagnosed malignancies worldwide and includes a broad range of diseases with different clinical behaviors and prognoses. In 2022, approximately 1.5 million new cases were reported globally, including around 330,000 cases of melanoma, which was associated with nearly 60,000 deaths [1]. Because melanoma can spread to distant organs, early detection is essential, as survival is considerably higher when the disease is identified at a localized stage [2]. In practice, however, dermoscopic assessment still depends heavily on clinician expertise and may vary between observers, leading to inconsistent interpretations. Manual review of dermoscopic images can also be time-consuming and adds to the clinical workload [3]. These challenges have encouraged the development of automated skin lesion classification systems that can analyze dermoscopic images and support clinical decision-making [4].

Artificial intelligence has significantly advanced computer-assisted diagnosis systems. Among the available approaches, deep learning methods, particularly convolutional neural networks (CNNs), have achieved remarkable efficiency [5]. Rather than relying on manually designed features, CNNs learn hierarchical and discriminative representations directly from images, allowing them to distinguish among different skin lesion types with high accuracy [6].

Despite their strong performance, conventional CNNs still face challenges in representing the wide range of visual information found in dermoscopic images. Such data contain global lesion characteristics, such as overall shape and color distribution, alongside subtle diagnostic patterns, including irregular borders, streaks, dots, and fine textures. Conventional feature extraction processes these complementary components jointly without explicitly separating structural information from fine-grained details. Successive downsampling may further weaken diagnostically relevant information and hinder the preservation of subtle lesion patterns [7]. Moreover, standard convolution applies spatially shared kernels across the image, limiting its ability to adapt to the heterogeneous characteristics of lesion centers, boundaries, internal structures, and surrounding skin [8]. Skin lesion datasets also commonly exhibit pronounced class imbalance, causing models to favor majority classes and reducing their ability to recognize underrepresented lesion types [9].

Recent studies have explored multiresolution methods to better capture the different types of information present in an image. One widely used approach is the discrete wavelet transform (DWT), which separates an image into low- and high-frequency components. Low-frequency components retain the overall structure of the image, while high-frequency components highlight finer patterns and local variations

[10]. Integrating this decomposition into deep learning architectures allows each frequency component to be processed separately and can help preserve diagnostically important information as the feature maps are progressively downsampled [11].

Involution has also been explored to enable spatially adaptive feature aggregation. Instead of applying the same learned kernels across all spatial locations, it generates input-dependent kernels for each location while sharing them across channel groups [12]. This enables feature aggregation to adapt to local image content, allowing different lesion regions to be processed according to their distinct characteristics.

Motivated by these considerations, we propose WaveInvNet, a multimodal architecture that combines wavelet decomposition with dynamic involution for multiclass skin lesion classification. Within the visual branch, the proposed Wavelet-Separated Dynamic Involution (WSDI) blocks decompose feature maps into low-frequency structural information and high-frequency detail components, which guide the generation of location-specific kernels for adaptive feature aggregation. The resulting visual features are then fused with patient metadata, including age, sex, and anatomical location, to incorporate complementary non-visual information before classification. Furthermore, targeted data augmentation is applied to underrepresented classes to increase sample diversity, mitigate bias toward majority classes, and improve model generalization.

The main contributions of this study are as follows:

- A Wavelet-Separated Dynamic Involution block is introduced, in which wavelet-derived structural and detail components directly guide the generation of location-specific involution kernels for adaptive feature aggregation. To the best of our knowledge, this is the first application of this mechanism to skin lesion classification.
- A lightweight four-stage hierarchical encoder, termed WaveInvNet, is developed by organizing WSDI blocks across multiple spatial resolutions to progressively learn frequency-aware and spatially adaptive visual representations.
- A comprehensive evaluation is conducted against several established convolution-based architectures under a unified experimental protocol, comparing classification performance, computational efficiency, and model complexity.
- The proposed model achieves competitive performance on the HAM10000 dataset, with an average accuracy of 93.63% and an AUC of 98.46%.

The remainder of this paper is structured as follows. Section 2 reviews relevant studies in the field. Section 3 presents the proposed WaveInvNet architecture. Section 4 outlines the experimental setup and presents and discusses the results. Finally, Section 5 summarizes the principal findings and concludes the paper.

2 Related work

2.1 Convolutional neural network-based models

CNNs continue to serve as the foundation of many skin lesion classification systems, while recent advances focus on redesigning conventional backbone architectures and integrating complementary learning mechanisms to enhance feature representation [13].

This trend is reflected in several recent studies. H. Zuo et al. [14] proposed ZACP, a modified EfficientNet architecture in which the conventional MBConv blocks are replaced with CPMB units. These units combine squeeze-and-excitation weighting with a zero-attention mechanism to refine channel responses during feature extraction. Sharma et al. [15] introduced a hybrid skin lesion classification framework that combines CNN-based feature extraction with Transformer layers to capture local visual patterns and long-range contextual dependencies, while incorporating clinical metadata to enhance classification performance. X. Li et al. [16] presented SCTFD, a framework that incorporates MARD-Net for lesion classification. MARD-Net combines residual convolutional blocks with spatial–channel attention and a Transformer encoder to capture local discriminative patterns and long-range dependencies, respectively. Similarly, Hussain et al. [17] developed a hybrid CNN–Transformer architecture that employs ResNet50V2 for local feature extraction and DeiT for global representation learning. The extracted features are integrated through an attention-based fusion module. In another study, P. C. Córcoles et al. [3] introduced MEL-IA, a multimodal architecture that uses EfficientNet-B4 to extract visual features and a dense branch to encode clinical metadata. The visual and metadata features are concatenated before multiclass classification.

2.2 Wavelet-based models

Wavelet-based models have gained increasing attention for incorporating frequency information into skin lesion analysis. Since wavelet transforms use localized basis functions, they preserve the spatial location of frequency components, unlike global spectral methods such as the Fourier transform [18]. Previous studies have introduced wavelet operations at different stages of deep learning pipelines, including image preprocessing, intermediate feature extraction, late-stage feature decomposition, and lesion segmentation.

As a related frequency-domain approach, O. Attallah [19] introduced SCaLiNG, which processes the original image alongside four directionally Gabor-filtered representations using parallel ResNet-18, MobileNet, and ShuffleNet backbones, followed by feature fusion, mRMR-based selection, and SVM classification. Moving toward direct wavelet integration, A. Aboulmira et al. [20] developed Wave-EfficientNet, which performs multilevel wavelet decomposition and integrates weighted subband features into selected stages of EfficientNet-B0 through element-wise addition. Its final configuration employs a two-level decomposition to combine spatial representations with structural and textural information. In contrast, R. Mousa et al. [21] introduced wavelet processing at a later stage by applying DWT to the final feature maps extracted by MobileNet, DenseNet-121, Inception, or Xception. The resulting LL, LH, HL, and HH subbands are concatenated and refined using self-attention before classification, while swarm-based optimization algorithms are employed to tune the model configuration. Wavelet processing has also been incorporated into lesion segmentation. X. Zheng et al. [22] combined LW-XNet with ISA-DenseNet, using an LWT channel-concatenation block to generate eight Legendre multiwavelet coefficient maps within the segmentation encoder before classifying the segmented lesion with a densely connected network incorporating improved soft attention. More recently, S. Cherukuri and S. D. V. Prasad

[23] applied wavelet-based denoising before lesion segmentation using Tversky-CA-Net. Following segmentation and data augmentation, wavelet-based texture features were extracted and classified using an optimized ensemble of SpinalNet, a quantum dilated CNN, and a deep Kronecker network.

2.3 Involution-based models

Involution has been investigated in medical image analysis because it enables feature aggregation to adapt to spatially heterogeneous patterns while maintaining channel-shared computation [24]. Several studies have incorporated this operator into diagnostic architectures for a variety of medical conditions; however, its application to skin lesion classification remains comparatively limited.

H. M. A. Mohammed et al. [25] developed ISAFusionNet, a multimodal architecture comprising parallel clinical and dermoscopic image branches built from involutional identity-residual and convolutional-residual blocks. Features from corresponding stages are progressively concatenated and refined through a four-stage soft-attention fusion pathway. The fused visual representation is combined with encoded patient metadata to jointly predict the lesion diagnosis and seven dermoscopic criteria. Beyond skin lesion analysis, O. M. Alshehri et al. [26] introduced HICNN for the classification of leukemia stages and leukocyte subtypes. Following an initial convolutional layer, the network employs three successive hybrid blocks, each consisting of parallel convolution and involution branches for feature extraction. Their outputs are fused by element-wise addition, followed by adaptive average pooling and final classification. In eye-tracking analysis, M. F. Islam et al. [27] developed a lightweight network in which three involution layers capture location-specific patterns, followed by three convolutional blocks incorporating max pooling. The final architecture contains 356,624 trainable parameters. For histopathological cancer classification, S. M. N. Nobel et al. [28] introduced CancerNet, which integrates convolution for local feature extraction, involution for spatially adaptive feature refinement, and Transformer-based self-attention for modeling long-range dependencies. The extracted features are aggregated through global average pooling before classification, while LIME and Grad-CAM variants are used to interpret the model's predictions. Similarly, M. Cihan and M. Ceylan [29] developed HybridCISN for neonatal disease detection using hyperspectral images. The architecture combines 2D and 3D convolutions with involution to extract complementary spatial, spectral, and location-specific features. Image features are fused with haemoglobin and bilirubin measurements for binary and multiclass classification.

3 Method

The proposed WaveInvNet is a multimodal architecture designed for multiclass skin-lesion classification by combining frequency-aware visual features with patient metadata. The model takes two inputs: a dermoscopic image and the corresponding metadata, including age, sex, and anatomical location. Missing age values are replaced with the median age, while sex and anatomical location are converted into numerical labels. These three variables are then standardized to zero mean and unit variance.

On the image side, a patch-based stem first projects the input into an initial feature map. The resulting representation is then passed through four hierarchical stages of Wavelet-Separated Dynamic Involution blocks, with spatial downsampling applied between successive stages.

Within each WSDI block, the input features are decomposed using a stationary Haar wavelet transform into one low-frequency structural component and three high-frequency detail components. These frequency-separated representations are then used to generate spatially adaptive involution kernels, allowing the network to respond to lesion structure, texture, and local variations at different resolution levels. After the final stage, global average pooling converts the visual feature map into a compact representation, which is concatenated with the encoded metadata. The resulting fused vector is finally passed through a dense layer with softmax activation to predict one of the seven skin-lesion classes.

By integrating frequency-separated visual processing with spatially adaptive aggregation and patient metadata, WaveInvNet combines complementary image-derived and clinical information within a unified classification framework. Figure 1 illustrates the proposed model. The following subsections describe the grouped involution operator and the proposed WSDI block.

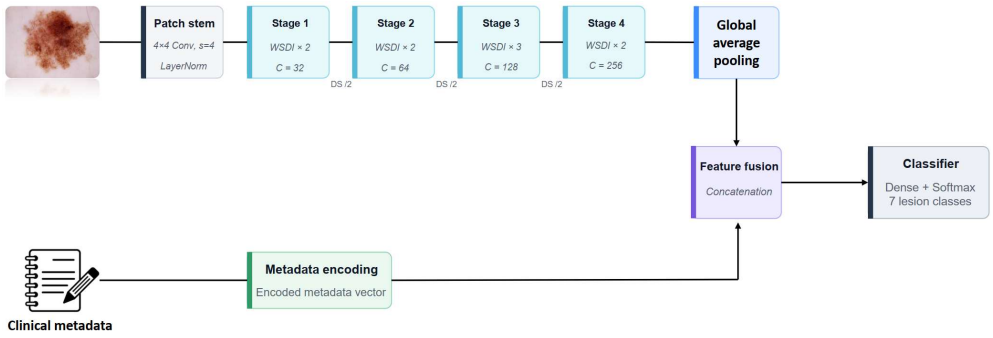


Fig. 1 Overall architecture of the proposed WaveInvNet for multimodal seven-class skin-lesion classification. WSDI denotes Wavelet-Separated Dynamic Involution block, and C denotes the number of channels.

3.1 Grouped involution operator

Involution was introduced by Li et al. [24] as an alternative to conventional convolution. It employs location-dependent kernels shared across channels within each group, allowing aggregation weights to adapt to local feature content while reducing channel-wise kernel redundancy.

In our implementation of the involution operator, let $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$ denote the input feature map, where B, H, W, and C represent the batch size, spatial dimensions, and number of channels, respectively. The channels are divided into $G = C/C_g$ groups, with C_g denoting the number of channels per group. In WaveInvNet, $C_g = 16$. At

each spatial location, a $K \times K$ neighborhood is extracted using unit stride and `SAME` padding. The resulting patches are arranged as $\mathbf{P} \in \mathbb{R}^{B\times H\times W\times G\times K^2\times C_g}$, while the corresponding dynamic kernels are represented by $\mathbf{K} \in \mathbb{R}^{B\times H\times W\times G\times K^2}$. The grouped involution operation is defined as :

$$\mathbf{Y}_{b,i,j,g,q} = \sum_{p=1}^{K^2} \mathbf{K}_{b,i,j,g,p} \mathbf{P}_{b,i,j,g,p,q} \tag{1}$$

where p indexes the K^2 positions within the local neighborhood and q identifies a channel within group g. Thus, all channels belonging to the same group share a position-specific kernel, whereas different spatial locations and channel groups can perform distinct local transformations.

The resulting grouped response $\mathbf{Y} \in \mathbb{R}^{B\times H\times W\times G\times C_g}$ is reshaped into an output feature map of shape $\mathbb{R}^{B\times H\times W\times C}$. Because unit stride and `SAME` padding are used, the spatial resolution is preserved. The proposed implementation uses $K = 7$, corresponding to a 7×7 aggregation neighborhood.

This operator performs only the grouped neighborhood aggregation. The position-dependent kernels are generated separately within the proposed WSDI block from wavelet-separated structural and detail representations, as described in the following subsection.

3.2 Wavelet-Separated Dynamic Involution Block

The proposed WSDI block generates position-dependent involution kernels from frequency-separated feature representations through stationary Haar decomposition, frequency-specific kernel generation, adaptive kernel integration, grouped involution, and residual channel mixing.

Stationary Haar decomposition

Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times H\times W\times C}$, layer normalization is applied across the channel dimension:

$$\mathbf{X}' = \mathrm{LayerNorm}(\mathbf{X})$$

The normalized feature map is then decomposed using a single-level stationary Haar transform:

$$(\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}) = \mathcal{H}_{\mathrm{stat}}(\mathbf{X}') \tag{2}$$

where $\mathcal{H}_{\mathrm{stat}}$ is implemented using fixed 2×2 depthwise Haar filters with unit stride and `SAME` padding. The LL subband preserves low-frequency structural information, whereas LH, HL, and HH capture directional and diagonal variations. Since the transform does not perform downsampling, all four subbands retain the original spatial resolution.

Frequency-specific kernel generation

The block employs separate generators for the structure, detail, and gating branches. Each generator consists of a 1×1 convolution that reduces the channel dimension to

$C_r = \max(1, \lfloor C/r \rfloor)$, followed by GELU activation and a second 1×1 convolution. The structure and detail generators each produce GK^2 coefficients, whereas the gating generator produces G coefficients at every spatial position.

The structure branch generates the raw kernel logits $\mathbf{Z}_{\text{str}} \in \mathbb{R}^{B\times H\times W\times G\times K^2}$ from the low-frequency subband $\mathbf{F}_{LL}$. These logits are normalized over the kernel-position dimension:

$$\mathbf{K}_{\text{str}} = \text{Softmax}_p\left(\mathbf{Z}_{\text{str}}\right) \tag{3}$$

where p indexes the K^2 positions of the local kernel. Consequently, the coefficients of $\mathbf{K}_{\text{str}}$ are non-negative and sum to one for each sample, spatial location, and channel group, providing normalized aggregation of structural information.

The detail branch receives the channel-wise concatenation of the high-frequency subbands:

$$\mathbf{F}_{\text{det}} = \text{Concat}_c\left(\mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\right)$$

Let $\mathbf{Z}_{\text{det}}$ denote the raw coefficients generated from $\mathbf{F}_{\text{det}}$. The logits are bounded using the hyperbolic tangent function, $\widetilde{\mathbf{K}}_{\text{det}} = \tanh(\mathbf{Z}_{\text{det}})$, and subsequently mean-centered and L_1-normalized:

$$\mathbf{K}_{\text{det}} = \frac{\widetilde{\mathbf{K}}_{\text{det}} - \mu_p\left(\widetilde{\mathbf{K}}_{\text{det}}\right)}{\left\|\widetilde{\mathbf{K}}_{\text{det}} - \mu_p\left(\widetilde{\mathbf{K}}_{\text{det}}\right)\right\|_{1,p} + \varepsilon} \tag{4}$$

where $\mu_p(\cdot)$ and $\|\cdot\|_{1,p}$ denote the mean and L_1-norm over the kernel-position dimension, respectively. Mean centering produces a signed, zero-sum kernel that emphasizes local variations, while L_1-normalization controls its magnitude.

Adaptive kernel integration

A spatially varying, group-specific gate controls the contribution of the detail kernel. The gating branch receives

$$\mathbf{F}_{\text{gate}} = \text{Concat}_c\left(\mathbf{F}_{LL}, |\mathbf{F}_{LH}|, |\mathbf{F}_{HL}|, |\mathbf{F}_{HH}|\right)$$

The absolute high-frequency coefficients allow the gate to respond to the strength of local detail independently of its orientation polarity. If $\mathbf{Z}_{\text{gate}}$ denotes the G logits generated at each spatial location, the gate is obtained as $\boldsymbol{\beta} = \sigma(\mathbf{Z}_{\text{gate}})$. Each value $\beta_{b,i,j,g} \in (0, 1)$ controls the detail contribution for one sample, spatial location, and channel group.

The final WSDI kernel is defined as

$$\mathbf{K}_{\text{wsdi}} = \mathbf{K}_{\text{str}} + \boldsymbol{\beta} \odot \mathbf{K}_{\text{det}} \tag{5}$$

where $\odot$ denotes element-wise multiplication. The gate is broadcast over the corresponding K^2 kernel positions. The bias of the final gating convolution is initialized to -2, biasing the block toward a limited high-frequency contribution during the initial training stages. Because $\mathbf{K}_{\text{str}}$ has unit sum and $\mathbf{K}_{\text{det}}$ has zero sum, $\mathbf{K}_{\text{wsdi}}$ also retains a unit sum.

Kernel application and residual refinement

The integrated kernel is applied to local neighborhoods of $\mathbf{X}'$ using the grouped involution operation defined in Eq. (1):

$$\mathbf{Y} = \mathrm{Inv}\left(\mathbf{X}', \mathbf{K}_{\mathrm{wsdi}}\right)$$

The resulting features are expanded to αC channels, activated using GELU, and projected back to C channels:

$$\mathbf{Z} = \mathrm{Conv}_{1\times1}^{C}\left(\mathrm{GELU}\left(\mathrm{Conv}_{1\times1}^{\alpha C}\left(\mathbf{Y}\right)\right)\right)$$

Finally, LayerScale, stochastic depth, and residual addition are applied:

$$\mathbf{X}_{\mathrm{out}} = \mathbf{X} + \mathrm{DropPath}\left(\boldsymbol{\gamma} \odot \mathbf{Z}\right) \tag{6}$$

where α is the channel-expansion ratio and $\boldsymbol{\gamma} \in \mathbb{R}^{C}$ is a learnable LayerScale vector initialized to 10^{-6}. The reduction and channel-expansion ratios are set to $r = 4$ and $\alpha = 4$, respectively.

At the network level, WaveInvNet comprises four stages with depths $(2, 2, 3, 2)$ and channel dimensions $(32, 64, 128, 256)$. Spatial downsampling is performed between consecutive stages using stride-2 layers.

4 Experimental results and discussion

4.1 Dataset

The proposed model was evaluated on the HAM10000 (Human Against Machine with 10,000 training images) dataset, a publicly available multi-source collection containing 10,015 dermoscopic images of pigmented skin lesions [30]. It includes seven diagnostic classes: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). More than half of the lesions were confirmed histopathologically, while the remaining cases were diagnosed through follow-up examinations, expert consensus, or in vivo confocal microscopy. The available metadata include patient age, sex, and lesion location. HAM10000 is also highly imbalanced, with melanocytic nevi representing the majority of the samples. Representative dermoscopic images from each diagnostic class are shown in Fig. 2.

Table 1 summarizes the class distribution before and after data augmentation, along with the numbers of training, validation, and test samples.

4.2 Implementation details

The dataset was partitioned into training, validation, and test sets using a 70%/10%/20% split. Model training was performed on a system equipped with an Intel Core i9-14900HX processor and 32 GB of RAM. The model weights corresponding to the highest validation accuracy were restored for final evaluation on the test set. The training configuration and hyperparameter settings are summarized in Table 2.

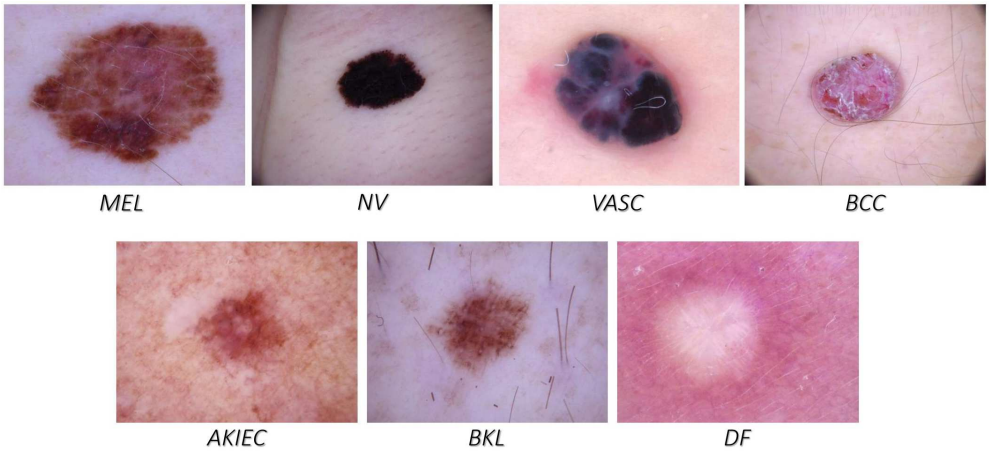


Fig. 2 Representative dermoscopic images of the seven HAM10000 classes.

Table 1 Class distribution of the HAM10000 dataset before and after data augmentation, with a 70/10/20 training/test/validation split.

Class	Original samples	Augmented samples	Training set	Testing set	Validation set
NV	6,705	6,705	4,693	1,341	671
MEL	1,113	3,000	2,100	600	300
BKL	1,099	2,800	1,960	560	280
BCC	514	2,500	1,750	500	250
AKIEC	327	1,800	1,260	360	180
VASC	142	1,500	1,050	300	150
DF	115	1,400	980	280	140

4.3 Evaluation metrics

Performance was evaluated using accuracy (ACC), precision (PRE), recall (REC), F1-score (F1), and the area under the receiver operating characteristic curve (AUC). Accuracy was computed over all test samples, whereas PRE, REC, F1, and AUC were calculated for each class using a one-vs-rest formulation.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$PRE = \frac{TP}{TP + FP} \tag{8}$$

Table 2 Training configuration and hyperparameter settings.

Parameter	Setting
Input image size	$299 \times 299 \times 3$
Dataset shuffling	Deterministic
Implementation framework	TensorFlow/Keras
Number of epochs	75
Batch size	16
Optimizer	Adam
Initial learning rate	1×10^{-4}
Loss function	Categorical cross-entropy
Output activation	Softmax
Checkpoint criterion	Highest validation accuracy

$$REC = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 \times \frac{PRE \times REC}{PRE + REC} \tag{10}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Given the class imbalance in HAM10000, PRE, REC, F1, and AUC were macro-averaged as the unweighted mean of the class-specific scores so that each class contributed equally to the evaluation [31]. To account for variability introduced by stochastic training, each experiment was repeated five times with different random seeds, and the results are reported as mean ± standard deviation [32].

4.4 Performance evaluation of the proposed method

This subsection evaluates the classification performance of WaveInvNet. Quantitative results are summarized in Table 3, with the corresponding confusion matrix and ROC curves shown in Figs. 3 and 4, respectively.

Table 3 Classification performance of the proposed model on the HAM10000 test set. Performance metrics are reported as the mean ± standard deviation over five independent runs with different random seeds.

Method	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-score (%)
WaveInvNet	93.63 ± 0.16	98.46 ± 0.18	94.45 ± 0.15	95.20 ± 0.16	94.81 ± 0.20

The proposed model achieved a macro-averaged AUC of 98.46%, while the remaining metrics ranged from 93.63% to 95.20%. The low standard deviations indicate consistent performance across the five independent runs.

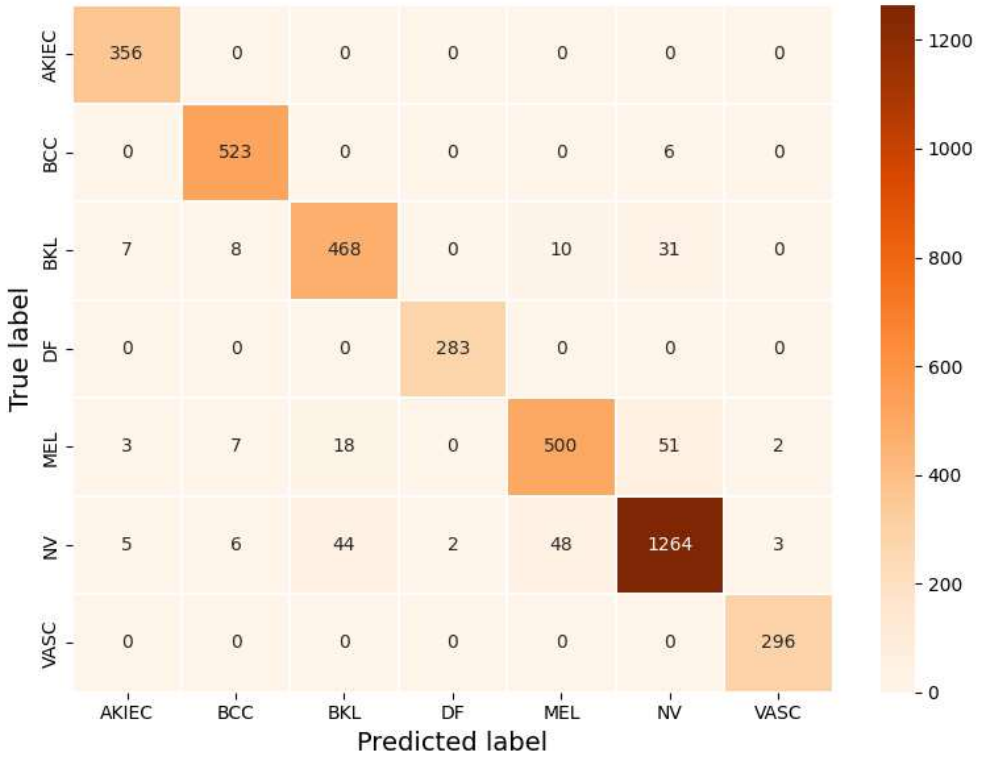


Fig. 3 Confusion matrix illustrating the class-wise prediction performance of the proposed model on the HAM10000 test set.

Figure 3 shows that 3,690 of the 3,941 test samples were classified correctly. AKIEC, DF, and VASC were identified without error, while BCC was also recognized with very few mistakes. Most of the remaining errors involved MEL and NV, with 51 MEL samples classified as NV and 48 NV samples classified as MEL. Some confusion was also observed between BKL and NV, where 31 BKL samples were predicted as NV and 44 NV samples were predicted as BKL.

Figure 4 shows class-wise AUC values between 95.7% and 100%. BCC, DF, and VASC each reached 100%, followed by AKIEC at 99.6%. NV and MEL achieved 97.2% and 96.8%, respectively, while BKL had the lowest AUC at 95.7%. Overall, all classes remained well above the random-guess reference line.

4.5 Comparative analysis with existing approaches

WaveInvNet was compared with five established deep learning architectures under identical preprocessing, data augmentation, and concatenation-based fusion settings. Their performance is reported in Table 4 and illustrated in Fig. 5. Table 5 compares the architectures in terms of trainable parameters, weight storage, and optimizer-state memory.

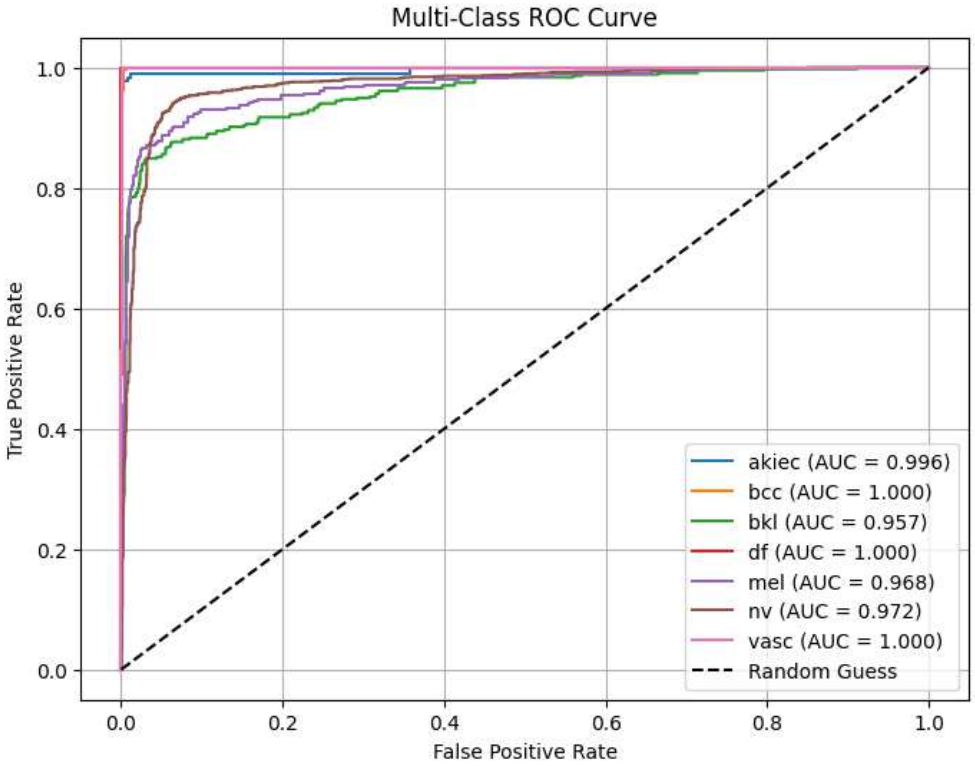


Fig. 4 Per-class ROC curves of the proposed model on the HAM10000 test set. The AUC for each class is reported.

WaveInvNet achieved the best overall performance among the evaluated baselines across all five metrics. Relative to the strongest baseline for each metric, the gains were 1.76 percentage points in accuracy, 0.21 in AUC, 1.89 in precision, 2.08 in recall, and 1.55 in F1-score. The small standard deviations across runs also show that these results were consistent.

As reported in Table 5, WaveInvNet was the most parameter-efficient architecture, requiring only 2.54 million trainable parameters, 10.16 MB for FP32 weights, and 20.32 MB for Adam's two moment states. Compared with EfficientNetB0, the smallest baseline, WaveInvNet reduced the number of trainable parameters and the associated memory requirements by approximately 42.6%. In contrast, ConvNeXt-Base had the highest resource requirements, with 87.92 million trainable parameters. The differences in trainable parameter counts are visualized in Fig. 6.

Table 4 Comparison of WaveInvNet and the baseline architectures on the HAM10000 test set. Results are reported as the mean ± standard deviation over five independent runs.

Method	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNetB0	88.45 ± 0.36	96.72 ± 0.61	89.92 ± 0.31	89.94 ± 0.35	89.88 ± 0.32
DenseNet121	89.94 ± 0.53	98.07 ± 0.28	90.14 ± 0.54	91.43 ± 0.49	90.66 ± 0.57
ResNet50	90.87 ± 0.19	98.25 ± 0.30	91.91 ± 0.21	92.80 ± 0.26	92.33 ± 0.18
ConvNeXt-Base	91.52 ± 0.71	98.10 ± 0.58	91.99 ± 0.70	93.07 ± 0.63	92.89 ± 0.69
InceptionV3	91.87 ± 0.32	98.23 ± 0.24	92.56 ± 0.30	93.12 ± 0.35	93.26 ± 0.31
WaveInvNet	**93.63 ± 0.16**	**98.46 ± 0.18**	**94.45 ± 0.15**	**95.20 ± 0.16**	**94.81 ± 0.20**

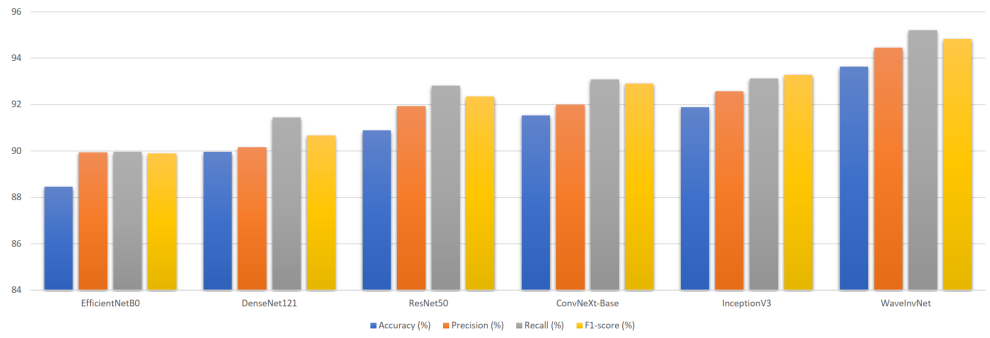


Fig. 5 Performance comparison of WaveInvNet and the baseline architectures in terms of mean accuracy, precision, recall, and F1-score under identical experimental settings.

4.6 Overview of tate-of-the-art approaches

To position the proposed model within the existing literature, this section reviews recent classification approaches evaluated on the HAM10000 dataset. Table 6 summarizes their reported performance across commonly used evaluation metrics. Since the studies differ in metric reporting, dataset partitioning, preprocessing, input resolution, class-balancing strategies, and metadata integration, the comparison is intended to provide context rather than a strict performance ranking.

MAKNet100–WN2 [37] achieved 90.50% accuracy, with precision, recall, and F1-score values of 91.58%, 90.55%, and 91.06%, respectively. GC-ViT-E [33] reached 92.81% accuracy and a 90.29% F1-score, while the XGBoost Stacking Ensemble [36] obtained 86.38% accuracy and 74.85% F1-score. The Hybrid CNN–Transformer [15] reported 92.80% accuracy, 91.80% precision, 93.33% recall, and 92.56% F1-score. CNN–BiLSTM with Attention [38] showed similarly balanced performance, with 92.73% accuracy and a 92.70% F1-score.

Table 5 Comparison of trainable parameter counts and FP32 parameter-related memory requirements. The Adam optimizer state comprises the first- and second-moment tensors.

Model	Trainable parameters	Weights (MB)	Adam states (MB)
ConvNeXt-Base	87,915,143	351.66	703.32
ResNet50	24,147,463	96.59	193.18
DenseNet121	7,302,535	29.21	58.42
EfficientNetB0	4,422,275	17.69	35.38
InceptionV3	22,381,223	89.52	179.05
WaveInvNet	2,540,131	10.16	20.32

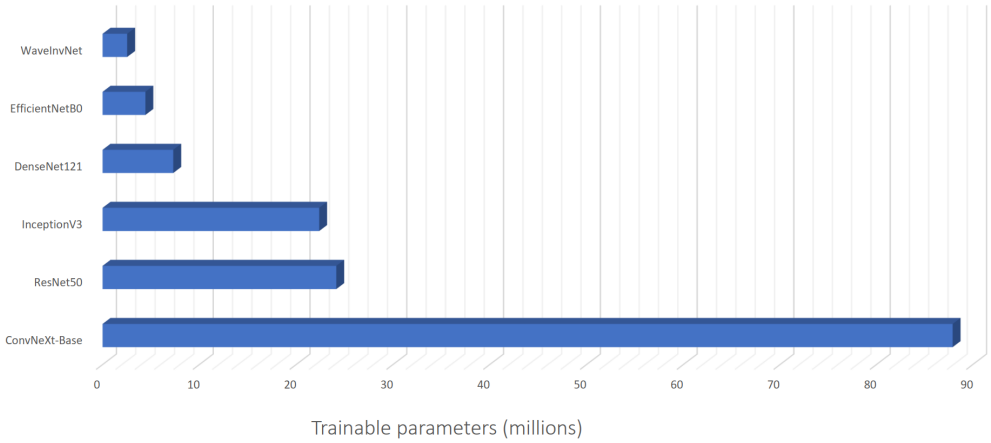


Fig. 6 Comparison of trainable parameter counts across WaveInvNet and the baseline models.

Explainable Deep Inherent Learning [39] reported the highest accuracy among these studies at 92.89%, although its precision, recall, and F1-score were 76.85%, 58.57%, and 63.14%, respectively. StAlK [34] achieved 90.46% accuracy and an F1-score of 86.50%, while precision and recall were not reported. MiSC [35] reached 90.00% accuracy, 89.80% precision, and 90.50% recall, with no F1-score provided.

4.7 Discussion

The similar values of macro precision, recall, and F1-score suggest balanced performance across the seven classes (Table 3). Macro recall was slightly higher than macro precision, indicating somewhat stronger sensitivity while precision remained comparably high. Since these metrics give equal weight to each class, the results reflect

Table 6 Reported classification performance of recent methods evaluated on the HAM10000 dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Hybrid CNN–Transformer [15]	92.80	91.80	93.33	92.56
GC-ViT-E [33]	92.81	90.61	90.03	90.29
StAlK [34]	90.46	–	–	86.50
MiSC [35]	90.00	89.80	90.50	–
XGBoost Stacking Ensemble [36]	86.38	83.90	69.96	74.85
MAKNet100–WN2 [37]	90.50	91.58	90.55	91.06
CNN–BiLSTM with Attention [38]	92.73	92.84	92.73	92.70
Explainable Deep Inherent Learning [39]	92.89	76.85	58.57	63.14

performance beyond the dominant NV category. The small variation across independent runs also shows that the model produced consistent results under different random initializations.

The class-wise results help clarify where the remaining errors occurred. As shown in Fig. 3, most misclassifications involved BKL, MEL, and NV. MEL was most often predicted as NV, while NV was mainly confused with MEL and BKL, indicating that these classes were more difficult to separate. In the representative confusion matrix, AKIEC, DF, and VASC were classified without error, and only six BCC samples were misclassified. The ROC curves show a similar pattern (Fig. 4), with BKL obtaining the lowest AUC (0.957), followed by MEL (0.968) and NV (0.972), while BCC, DF, and VASC each reached 1.000. Despite these class-specific errors, the high AUC values indicate strong one-versus-rest discrimination across all seven categories.

Under identical training and evaluation conditions, WaveInvNet achieved the highest mean score for all reported metrics (Table 4). InceptionV3 was the strongest overall baseline, while ResNet50 obtained the highest baseline AUC. The largest improvement was seen in macro recall, where WaveInvNet exceeded InceptionV3 by 2.08 percentage points. Since macro averaging gives equal weight to each class, this gain reflects improved sensitivity across the class set rather than performance dominated by NV. Precision and F1-score also increased, showing that the improvement in recall was accompanied by similarly strong overall classification performance. The AUC gain over ResNet50 was smaller at 0.21 percentage points, which is consistent with the already high discrimination achieved by the baseline models. Overall, the main advantage of WaveInvNet was more accurate final classification while preserving strong one-versus-rest discrimination.

The model complexity results put these performance gains into perspective (Table 5). WaveInvNet used 2.54 million trainable parameters, which is 42.6% fewer than EfficientNetB0, the smallest baseline, and 88.7% fewer than InceptionV3, the strongest overall baseline. The corresponding FP32 weight storage and Adam-state

memory were 10.16 MB and 20.32 MB, respectively. This shows that the improved performance was achieved with a relatively small parameter footprint.

While the results are promising, several directions remain for future evaluation. WaveInvNet was assessed on HAM10000 under a controlled experimental protocol, and testing on additional datasets would provide broader evidence of generalization across different clinical and imaging conditions. Stratified k-fold cross-validation could also be used to examine variability across alternative data splits. In addition, the metadata branch currently uses age, sex, and anatomical location, but could be extended with other clinical variables when these are consistently available. Future studies may also explore explainability, uncertainty estimation, and probability calibration to improve the transparency and clinical interpretability of WaveInvNet.

5 Conclusion

This study introduced WaveInvNet, a lightweight multimodal architecture for multiclass skin-lesion classification that integrates dermoscopic images with patient metadata. The proposed Wavelet-Separated Dynamic Involution block combines stationary Haar wavelet decomposition with spatially adaptive involution, while concatenation-based fusion incorporates age, sex, and anatomical location into the learned visual representation. Across five independent runs on HAM10000, WaveInvNet achieved a mean accuracy of 93.63%, a macro-averaged AUC of 98.46%, and a macro F1-score of 94.81%. Under identical experimental conditions, it outperformed all evaluated baseline architectures while requiring only 2.54 million trainable parameters. Overall, these findings demonstrate that combining frequency-aware visual processing with patient metadata enables accurate and parameter-efficient multiclass skin-lesion classification.

Acknowledgements. The authors gratefully acknowledge the support and contribution of the Department of Information Technology, Faculty of Sciences and Technology of Fez, to this research.

Author contributions. Mohamed Amine Ibrahimi: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Validation, Visualization, Writing – Original Draft. Abdelghafour Abbad and Rachid Ben Abbou: Methodology, Validation, Supervision, Writing – Review & Editing. Khalid Abbad: Conceptualization, Validation, Supervision, Writing – Review & Editing.

Data availability. The dataset used in this study is available at: https://api.isic-archive.com/collections/212/.

Statements and declarations

Funding. The authors acknowledge the support of the Moroccan National Centre for Scientific and Technical Research (CNRST) through the Doctoral Monitor Scholarship Program "PhD-Associate Scholarship – PASS" (Digital domain), Horizon 2027, National Strategy Morocco 2030.

Consent for publication. Not applicable.

Competing interests. The authors declare that they have no competing interests.

Ethics approval and consent to participate. This study used the publicly available and de-identified HAM10000 dataset and did not involve the recruitment of human participants or the collection of new patient data.

References

[1] Wang, M., Gao, X., Zhang, L.: Recent global patterns in skin cancer incidence, mortality, and prevalence. Chinese Medical Journal **138**(2), 185–192 (2025) https://doi.org/10.1097/CM9.0000000000003416

[2] Okobi, O.E., Abreo, E., Sams, N.P., Chukwuebuni, O.H., Amoako, L.A.T., Wiredu, B., Uboh, E.E., Ekechi, V.C., Okafor, A.A.: Trends in melanoma incidence, prevalence, stage at diagnosis, and survival: An analysis of the united states cancer statistics (USCS) database. Cureus **16**(10), 70697 (2024) https://doi.org/10.7759/cureus.70697

[3] Córcoles, P.C., De Ramón Fernández, A., Saval Calvo, M., Salinas Serrano, J.M., Guijarro Peral, D., Ruiz Fernández, D.: MEL-IA: An interoperable AI system for multimodal skin lesion classification in hospital settings. Journal of Medical Systems **50**, 95 (2026) https://doi.org/10.1007/s10916-026-02424-y

[4] Strzelecki, M., Kociołek, M., Strkakowska, M., Kozłowski, M., Grzybowski, A., Szczypiński, P.M.: Artificial intelligence in the detection of skin cancer: State of the art. Clinics in Dermatology **42**(3), 280–295 (2024) https://doi.org/10.1016/j.clindermatol.2023.12.022

[5] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., Zhao, S.: Skin cancer classification with deep learning: A systematic review. Frontiers in Oncology **12**, 893972 (2022) https://doi.org/10.3389/fonc.2022.893972

[6] Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Transactions on Medical Imaging **36**(4), 994–1004 (2017) https://doi.org/10.1109/TMI.2016.2642839

[7] Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I.M., Werner, R., Schlaefer, A.: Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Transactions on Biomedical Engineering **67**(2), 495–503 (2020) https://doi.org/10.1109/TBME.2019.2915839

[8] Hou, J., Guo, Z., Feng, Y., Diao, W.: SPANet: Spatial adaptive convolution based content-aware network for aerial image semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **16**, 2192–2204 (2023) https://doi.org/10.1109/JSTARS.2023.3244207

[9] Noaman, A., Ahmad, R., Khan, M.F., Mohammed, A.S., Farooq, M., Adnan, K.M.: Beyond binary: Multi-class skin lesion classification with alexnet transfer learning—towards enhanced dermatological diagnosis. Discover Applied Sciences **7**, 35 (2025) https://doi.org/10.1007/s42452-024-06448-2

[10] Serte, S., Demirel, H.: Wavelet-based deep learning for skin lesion classification. IET Image Processing **14**(4), 720–726 (2020) https://doi.org/10.1049/iet-ipr.2019.0553

[11] Wu, Q.-e., Yu, Y., Zhang, X.: A skin cancer classification method based on discrete wavelet down-sampling feature reconstruction. Electronics **12**(9), 2103 (2023) https://doi.org/10.3390/electronics12092103

[12] Meng, Z., Zhao, F., Liang, M., Xie, W.: Deep residual involution network for hyperspectral image classification. Remote Sensing **13**(16), 3055 (2021) https://doi.org/10.3390/rs13163055

[13] Hasan, M.K., Ahamad, M.A., Yap, C.H., Yang, G.: A survey, review, and future trends of skin lesion segmentation and classification. Computers in Biology and Medicine **155**, 106624 (2023) https://doi.org/10.1016/j.compbiomed.2023.106624

[14] Zuo, H., Weng, Z., Fu, Z., Ding, F., Chai, J., Chen, L., Gong, C., Zheng, Z.: ZACP: Enhancing skin lesion classification ability using zero attention and complete perception. Visual Computing for Industry, Biomedicine, and Art **9**, 12 (2026) https://doi.org/10.1186/s42492-026-00223-3

[15] Sharma, A.K., Bhanodia, P.K., Bhushan, M.: A hybrid deep learning approach for skin lesion classification using convolutional neural networks and transformers. Network Modeling Analysis in Health Informatics and Bioinformatics **15**, 119 (2026) https://doi.org/10.1007/s13721-026-00794-4

[16] Li, X., Ouyang, J., Chen, Y., Diao, Z., Li, K., He, D., Liang, W., Poniszewska-Marańda, A.: Accurate skin lesion classification on imbalanced dermoscopic images with high variance via the SCTFD framework. Scientific Reports **16**, 12302 (2026) https://doi.org/10.1038/s41598-026-37846-8

[17] Hussain, S.R., Saritha, S., Chevuri, A., Karuna, Y.: Enhanced skin cancer classification for minority classes using conditional GAN pipeline and CNN–ViT ensemble. Scientific Reports **16**, 13114 (2026) https://doi.org/10.1038/s41598-026-43339-5

[18] Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory **36**(5), 961–1005 (1990) https://doi.org/10.1109/18.57199

[19] Attallah, O.: Skin cancer classification leveraging multi-directional compact convolutional neural network ensembles and gabor wavelets. Scientific Reports **14**, 20637 (2024) https://doi.org/10.1038/s41598-024-69954-8

[20] Aboulmira, A., Hrimech, H., Lachgar, M., Hanine, M., Garcia, C.O., Mezquita, G.M., Ashraf, I.: Hybrid model with wavelet decomposition and EfficientNet for accurate skin cancer classification. Journal of Cancer **16**(2), 506–520 (2025) https://doi.org/10.7150/jca.101574

[21] Mousa, R., Chamani, S., Morsali, M., Kazzazi, M., Hatami, P., Sarabi, S.: Enhancing skin cancer diagnosis using late discrete wavelet transform and new swarm-based optimizers. Machine Learning with Applications **23**, 100811 (2026) https://doi.org/10.1016/j.mlwa.2025.100811

[22] Zheng, X., Huang, Y., Liu, W., Cai, C.: LW-XNet for segmentation and classification of skin lesions from dermoscopy images. Expert Systems with Applications **255**, 124826 (2024) https://doi.org/10.1016/j.eswa.2024.124826

[23] Cherukuri, S., Prasad, S.D.V.: A clinical decision support system for skin cancer classification using fractional gooseneck barnacle-enabled ensemble classifier. Computational Biology and Chemistry **120**, 108738 (2026) https://doi.org/10.1016/j.compbiolchem.2025.108738

[24] Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., Chen, Q.: Involution: Inverting the inherence of convolution for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12321–12330 (2021). https://doi.org/10.1109/CVPR46437.2021.01214

[25] Mohammed, H.M.A., Omeroglu, A.N., Oral, E.A., Ozbek, I.Y.: ISAFusionNet: Involution and soft attention based deep multi-modal fusion network for multi-label skin lesion classification. Computers & Electrical Engineering **122**, 109966 (2025) https://doi.org/10.1016/j.compeleceng.2024.109966

[26] Alshehri, O.M., Shaf, A., Shakeel, U., Ali, T., Irfan, M., Jalal, M.M., Altayar, M.A., Abu-Alghayth, M.H., Al Shmrany, H.: Dynamic kernel generation through hybrid involution and convolution neural networks for leukemia and white blood cell classification. Scientific Reports **15**, 43844 (2025) https://doi.org/10.1038/s41598-025-28040-3

[27] Islam, M.F., Manab, M.A., Mondal, J.J., Zabeen, S., Rahman, F.B., Hasan, M.Z., Sadeque, F., Noor, J.: Involution fused convolution for classifying eye-tracking patterns of children with autism spectrum disorder. Engineering Applications of Artificial Intelligence **139**, 109475 (2025) https://doi.org/10.1016/j.engappai.2024.109475

[28] Nobel, S.M.N., Sultana, S., Tasir, M.A.M., Mridha, M.F., Aung, Z.: CancerNet: A comprehensive deep learning framework for precise and intelligible cancer

identification. Computers in Biology and Medicine **193**, 110339 (2025) https://doi.org/10.1016/j.compbiomed.2025.110339

[29] Cihan, M., Ceylan, M.: HybridCISN: Integrating 2D/3D convolutions and involutions with hyperspectral imaging and blood biomarkers for neonatal disease detection. Computers & Electrical Engineering **123**, 110193 (2025) https://doi.org/10.1016/j.compeleceng.2025.110193

[30] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**, 180161 (2018) https://doi.org/10.1038/sdata.2018.161

[31] Schaudt, D., Schwerin, R., Hafner, A., Riedel, P., Reichert, M., Schwerin, M., Beer, M., Kloth, C.: Augmentation strategies for an imbalanced learning problem on a novel COVID-19 severity dataset. Scientific Reports **13**(1), 18299 (2023) https://doi.org/10.1038/s41598-023-45532-2

[32] Coakley, K.L., Gundersen, O.E.: Measuring deep learning performance: An empirical study of performance distributions across architectures and tasks. Scientific Reports **16**, 20474 (2026) https://doi.org/10.1038/s41598-026-49656-z

[33] Alhaj, Z., Ozturk, M.: GC-ViT-E: Enhanced global context vision transformers for robust skin cancer classification under severe class imbalance. Journal of King Saud University Computer and Information Sciences **38**, 281 (2026) https://doi.org/10.1007/s44443-026-00680-z

[34] Sharma, S., Kumar, A., Monpara, J., Chandra, J.: Stalk: Structural alignment based self knowledge distillation for medical image classification. Knowledge-Based Systems **304**, 112503 (2024) https://doi.org/10.1016/j.knosys.2024.112503

[35] Cheng, H., Lian, J., Jiao, W.: Enhanced mobilenet for skin cancer image classification with fused spatial channel attention mechanism. Scientific Reports **14**, 28850 (2024) https://doi.org/10.1038/s41598-024-80087-w

[36] Chand, H.V., Sabharwal, S., Jiang, W., Islam, S.M.N., Cengiz, K.: A stacking-based deep learning ensemble for multi-class skin lesion classification. Discover Computing **29**, 417 (2026) https://doi.org/10.1007/s10791-026-10288-6

[37] Arshad, M., Khan, M.A., Górriz, J.M., Baili, J., AlHammadi, D.A., Kim, C., Nam, Y.: Skin cancer segmentation and recognition from dermoscopy images: A novel framework based on improved DeepLabV3+ and network-level fused deep architectures. Journal of Advanced Research **83**, 575–600 (2026) https://doi.org/10.1016/j.jare.2025.08.039 . Available online in 2025

[38] Shaik, A., Dutta, S.S., Sawant, I.M., Kumar, S., Balasundaram, A., De, K.: An attention based hybrid approach using CNN and BiLSTM for improved skin lesion classification. Scientific Reports **15**, 15680 (2025) https://doi.org/10.1038/

s41598-025-00025-2

[39] Hosny, K.M., Said, W., Elmezain, M., Kassem, M.A.: Explainable deep inherent learning for multi-classes skin lesion classification. Applied Soft Computing **159**, 111624 (2024) https://doi.org/10.1016/j.asoc.2024.111624