Received 27 June 2023, accepted 18 July 2023, date of publication 31 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299850

RESEARCH ARTICLE

TS-CNN: A Three-Tier Self-Interpretable CNN for Multi-Region Medical Image Classification

V. A. ASHWATH[1], O. K. SIKHA[1], AND RAUL BENITEZ[2]
[1]Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India
[2]Department of Automatic Control, Universitat Politécnica de Catalunya (BarcelonaTech), 08034 Barcelona, Spain

Corresponding authors: O. K. Sikha (ok_sikha@cb.amrita.edu) and Raul Benitez (raul.benitez@upc.edu)

The work of Raul Benitez was supported by the Spanish Ministry of Science and Innovation under Grant PID2020-116927RB-C22.

ABSTRACT Medical image classification is critical, where reliability and transparency are crucial for the safe and accurate diagnosis of diseases. Deep Convolutional Neural Networks (DCNNs) are widely used in medical image classification due to their high performance. However, they are often considered black-boxes because they offer little insight into decision-making. Therefore, improving the interpretability of DCNNs is crucial for their adoption in medical diagnoses. This paper proposes a novel three-tier self-interpretable DCNN (TS-CNN) architecture for multi-region medical image classification, which improves classification performance while being inherently interpretable. The proposed TS-CNN architecture is well-suited for medical images with multiple regions, such as images with scattered and randomly shaped lesions. The proposed architecture has three branches: a global branch that learns the relevant patterns from the raw input image; an attention branch that selects the important regions and discards the irrelevant parts for the local branch to learn; and a fusion branch that distills knowledge from both the global and local branches for classification. The proposed architecture is flexible in terms of the backbone CNNs used for classification and post-hoc interpretability methods used for attention capture. We demonstrate the flexibility and generalization of the architecture through a series of experiments involving multiple state-of-the-art CNN architectures such as DenseNet-121, Inception, Xception, and ResNet-50 as the global/local branches, each paired with GradCAM and Saliency maps as attention modules. The proposed architecture outperformed the backbone model in classification tasks on two datasets: a custom-made blob dataset and a publicly available skin lesion PAD-UFES-20 dataset, demonstrating its potential for improving accuracy in medical image classification tasks. The code related to this work can be found at: https://github.com/sikha2552/TS-CNN-A-Three-Tier-Self-Interpretable-CNN-for-Medical-Image-Classification-Empowered-with-Post-hoc.git.

INDEX TERMS Explainable AI, interpretable CNN, medical image classification.

I. INTRODUCTION

Deep neural networks (DNNs), specifically deep convolutional neural networks (DCNNs) [1] are a powerful class of machine learning models that have been widely used in medical image classification and segmentation [2], [3] owing to their exceptional performance. DCNNs are capable of automatically learning complex representations and patterns from input images. This has led to their increasing adoption in the diagnosis of various diseases such as cancer [4], alzheimer's [5], and cardiovascular diseases [6] etc. Despite their high accuracy, one of the most significant challenges in using DCNNs for medical diagnose is their lack of interpretability [7]. They often act as black-boxes, because it is difficult to understand how they attain their predictions. This can make it challenging for medical professionals to trust and rely on the results produced by these models, as they may not be able to understand the reasoning behind the model's output. Interpretability is particularly important in the medical field, where decisions based on machine learning models can have serious consequences [8]. Understanding how a model arrives at its conclusion can help medical

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A. Linte.

professionals identify potential errors, biases, or limitations in the model's design or training [9]. This knowledge can inform the development of more accurate and reliable models and help medical professionals make informed decisions. Efforts are currently underway to address the challenge of interpretability in DCNNs [10], [11]. Researchers have developed several post-hoc techniques to visualize and understand the internal workings of DCNNs [12], [13], such as saliency maps [14], class activation maps [15], layer activations [16], feature maps [16] or attention mechanisms [17].

The saliency map [14] uses the activation values of different layers to identify regions of interest. These regions are identified by computing the gradient of the output with respect to the input image. The magnitude of this gradient indicates the importance of each pixel in the image for the output prediction. Class Activation Maps (CAM) were introduced to identify the localized attention of the DCNNs by creating a spatial map of the regions that contribute to the most to a given prediction [15]. It operates by adding a global average pooling layer on top of the final convolutional layer of the DCNN. The weights learned by this layer were used to compute the class activation map. The Gradient Class Activation Mapping (GradCAM) builds upon the CAM technique using gradient information to compute the class activation map [18]. The gradient computed from the final convolutional block was used to weight the feature maps, producing a visual explanation of the regions that were most relevant to the prediction. Multiple modifications have been made to these methods, including approaches that use multiple layers to compute the saliency map [19] and techniques that use guided backpropagation to improve the interpretability of the saliency map [20]. Some researchers have also explored the use of other visualization techniques, such as occlusion-based methods, which involve masking parts of the input image to observe their effect on the output of the DCNN. Despite the usefulness of post-hoc interpretation methods as detailed above, there have been few efforts to integrate them directly into model architecture design to enhance the interpretability of the model [21].

This paper proposes a novel three-tier, self-interpretable deep neural network architecture for disease classification using post-hoc interpretation models, such as GradCAM or saliency maps, as attention modules to enhance the prediction accuracy of the base model. The proposed classification model consists of three branches: global, attention, and a local. The global branch serves as the initial stage of learning, where it learns and extracts high-level features from the input data. It focuses on capturing overall patterns and characteristics relevant to the classification task. The attention branch acts as an unsupervised segmentation module, leveraging post-hoc attention maps generated from the learned global branch. The attention maps highlight important regions in the input data that contribute to the classification decision. Finally, the local branch utilizes the attention branch to refine the knowledge gained from the global branch. It leverages the highlighted regions identified by the attention branch to obtain more fine-grained information about the disease being classified. The local branch aims to capture intricate details and nuances that may be crucial for accurate disease classification. By embodying the knowledge gained by the global branch and focusing on specific regions of interest, the local branch is equipped to achieve better classification accuracy. Thus, the proposed methodology can help improve the classification performance after the model reaches a training plateau. The proposed architecture not only improves the accuracy of the model, but also provides an opportunity for researchers and practitioners to better understand how the model arrives at its predictions. This level of interpretability can be particularly valuable in fields such as healthcare, where both accuracy and transparency are equally critical. The primary contributions of this paper are as follows:

- A three-tier, self-interpretable CNN architecture for medical image classification (TS-CNN) which can effectively capture attention regions that are disconnected and dispersed throughout the image. By incorporating the global, attention, and local branches, the model can extract relevant features and refine its classification knowledge.
- The incorporation of the attention branch into the TS-CNN architecture: which serves as an unsupervised ROI extraction module, a critical component in enhancing classification accuracy. The attention branch uses the global branch's information to find attention areas or regions of interest. This approach aids in directing the model's attention to essential regions while filtering out irrelevant information.
- Extensive validation experiments were performed using both a synthetic dataset with random blobs and a real-world skin lesion dataset for disease classification. The experimental results show that the proposed model outperforms the state-of-the-art generic CNN classification models, specifically when dealing with diverse image qualities. The model shows remarkable efficacy in handling both low-quality images obtained from sources like smartphones (skin lesion dataset) and high-quality (custom blob dataset) medical images
- Analysis of Noise Impact: The study investigates the effect of noise on the proposed architecture. By subjecting the model to different types and noise levels, the research explores the model's robustness and performance degradation in the presence of noise. This analysis provides insights into the model's limitations and areas for improvement.

II. RELATED WORKS

Several efforts have been made to incorporate attention models and interpretability methods to improve medical image analysis and classification tasks. Guan et al. [22] proposed an attention-driven model for thoracic disease classification using chest X-rays. The proposed method, called attention guided convolution neural network (AG-CNN), uses a three-branch approach that learns from disease-specific regions to avoid noise and improve alignment. It also integrates a global branch to compensate for the lost discriminative cues from the local branch. The global branch produces an attention heatmap, and identifies distinct and informative areas within the image. This localized region was then utilized to train a separate local CNN branch. Finally, to fine-tune the fusion branch, the last pooling layers of both the global and local branches were combined by concatenation. The proposed method is evaluated using the ChestX-ray14 dataset. It achieved a new state-of-the-art performance in thorax disease classification with an average AUC of 0.871 using DenseNet-121 as a backbone. The major drawback of this study is that the proposed architecture can not consider regions of interest that are disconnected and dispersed throughout the image, such as cells in microscopy images. Shen et al. [23] proposed an interpretable classifier for high-resolution breast cancer screening. The model uses a low-capacity network to identify informative regions and a high-capacity network to collect details from these regions. A fusion module was introduced to aggregate global and local information to make a final prediction. The model was trained using image-level labels and pixel-level saliency maps were generated. The model outperforms ResNet-34 and Faster R-CNN in classifying breasts with malignant findings on the NYU Breast Cancer Screening Dataset, and achieves performance on par with state-of-the-art approaches on the CBIS-DDSM dataset. An attention-guided CNN for breast can histopathology image classification was proposed by Yang et al. [24]. The authors proposed a supervised attention mechanism to localize the region of interest. This attention mechanism generates class activation maps that align well with the expectations of expert pathologists. The proposed method has shown promising results on the BACH microscopy test dataset (part A) [25], outperforming the state-of-the-art methods by a significant margin. A novel attention gate mechanism was proposed by Schlempe et al. [26], specifically for medical image analysis. The proposed attention gate (AG) suppresses the irrelevant part of the image under consideration and highlights the region of interest. Liu et al. [27] combined an attention module with a multi-scale latent representation network to identify a specific region of interest, which was then used to construct an accurate attention map. The attention module is used to determine the channel weights. Pacheco et al. stated that the aggregation of clinical data with image data improved the accuracy and interpretability of a DCNN-based skin cancer detection model [28]. The authors claimed that the classification accuracy improved by 7% for most of the state-of-the-art CNNs reported in the literature when combined with the clinical data. Yeh et al. [29] introduced a visual attention learning module to refine feature maps generated by any CNN architecture. The proposed attention module learns a weighting coefficient map to highlight the essential features by suppressing irrelevant pixels. Arshiya et al. [30] proposed adding an attention network as an additional branch to any generic CNN to highlight the region of interest for prediction. They used the learned feature maps from the convolutional blocks to construct an attention map, which was then multiplied by the feature maps for prediction. The authors reported an enhanced accuracy for skin cancer detection using the proposed attention network. In [31], Wang et al. presented a novel ConvNet architecture for glaucoma diagnosis that is clinically interpretable and capable of highlighting the distinct regions recognized by the network. The architecture employs M-LAP, a scheme that aggregates features from multiple scales, to enhance the diagnosis accuracy and generate glaucoma activations that provide a link between global semantic diagnosis and precise location. The method achieved superior performance compared to state-of-the-art approaches, with an AUC of 0.88, and demonstrated effectiveness in optic disc segmentation and local disease focus localization. Xing et al. [32]proposed a two-branch attention guided deformation network (AGDN) to improve the accuracy of WCE image classification. The AGDN utilizes attention maps to identify and amplify the regions of interest, specifically lesions, and also incorporates Third-order Long-range Feature Aggregation (TLFA) modules to capture long-range dependencies and contextual features. To refine the attention maps and promote interaction between the two branches, the authors introduced a novel Deformation-based Attention Consistency (DAC) loss. The global feature embeddings obtained from both branches were merged to predict image labels. With an overall classification accuracy of 91.29% on two publicly available WCE datasets, the proposed AGDN model outperformed the other state-of-the-art methods. This study addresses the challenge of limited CNN capacity in WCE image classification owing to small lesions and background interference.

In summary, several studies have proposed attention-driven models and interpretability methods to improve medical image analysis and classification. Most models use attention mechanisms to highlight important features and regions of interest in medical images, resulting in improved accuracy and interpretability. However, some models have limitations in considering disconnected and dispersed regions of interest. The current study aims to address these limitations and advance the field of medical image analysis and classification by developing novel attention mechanisms using post-hoc interpretability methods. By specifically focusing on disconnected and dispersed regions of interest, the proposed model seeks to capture important details that may be missed by traditional approaches. In addition to addressing the limitations of disconnected and dispersed regions of interest, the current study also focuses on improving the classification of medical images obtained from low-quality sources such as mobile devices. By incorporating advanced attention mechanisms using post-hoc interpretability methods, the model seeks to identify and leverage relevant features and regions of interest. This allows the model to extract meaningful information and make accurate predictions despite the limitations of low-quality images.

III. PROPOSED MODEL

The proposed three-tiered architecture, consisting of a global branch, an attention branch, and a local branch is detailed in this section. The global branch can be any custom or pre-trained CNN, responsible for extracting high-level features from the input data. During the initial training, the global branch was trained using the classification loss for few epochs. The weights of the global module were then frozen to ensure that the knowledge gained during the initial training was retained during the training of the local and fusion modules. To engage with the knowledge gained by the global branch, attention maps are generated by the attention branch, using post-hoc methods such as GradCAM and Saliency maps. The attention branch highlights the region of interest by suppressing irrelevant pixels from the feature map learned by the global branch. The interpretability map was then binarized and applied as a mask on the original input image, with a pre-defined threshold value τ and an ROI window size Ω to retain only the significant pixels and their local neighborhood. The threshold value and ROI window size should be established with respect to the backbone architecture, attention strategy, and the dataset. A threshold value of $\tau = 0.75$ for GradCAM and $\tau = 0.6$ for the Saliency map were used and found to be a good fit for the experiments detailed in this work. An ROI window Ω of size 30×30 pixels was found to be optimal in terms of information capture. A detailed explanation on selecting the optimal values for τ and Ω is detailed in Section.IV-A5. The masked image is then fed into the local module, which uses the same architecture as that of the global model. Unlike cropping the input image, as in [22], using a mask allows for capturing disjoint regions of interest, making it suitable for detecting a broader range of pathologies. The masked input approach removed noise from the image, enabling the local module to focus on the most critical regions of the image. This feature is especially useful in lesion detection applications, where the region of interest is small and sparsely distributed compared with the usual object detection problems.

Finally, the pooling layer outputs from the global and local layers are concatenated and fed into the fusion module, which is a fully connected classification network. This architecture allows the local module to learn from the knowledge extracted by the global module, thereby enhancing classification accuracy. To demonstrate the flexibility of the proposed architecture, it is shown that diverse global and local module architectures can be utilized interchangeably in conjunction with various post-hoc interpretability models in the attention branch. In conclusion, the proposed architecture not only enhances classification accuracy, but also provides interpretability, making it useful in medical image analysis and decision-making, where transparency and accuracy are critical. Figure. 1 shows the architecture of the proposed model and further detailed description is as follows.

The global branch f_g and the local branch f_l are DCNNs that learn to reduce a classification loss $\mathcal{L}$ defined as

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{\mathrm{N_{im}}} \mathcal{L}\left(\hat{\omega}_i, \omega_i\right) \text{ for each } I_i, \tag{1}$$

where I_i is the ith input image, ω_i is the corresponding ground truth classification label, $\hat{\omega}_i$ is the predicted label from the branch, $\mathcal{L}$ is the classification loss function (categorical cross entropy, binary cross-entropy, etc.), N_{im} is the total number of images in the dataset, and W are the learned weights of the CNN model along with their respective classification layers. The input image I is first passed to the global branch feature extractor f_g. pool_g are the deep features extracted by the learned global branch and h_g are the activation output an input image I, defined as

$$\mathrm{pool}_g, h_g = f_g(I). \tag{2}$$

The global branch is trained for k epochs, after that the weights are frozen. For every input image processed by the learned and frozen global branch, the corresponding attention map, denoted as map_I, is generated using the attention module $\mathcal{A}$ as

$$\mathrm{map}_I = \mathcal{A}(h_g), \tag{3}$$

where $\mathcal{A}$ represents the chosen post-hoc interpretability method (GradCAM, Saliency Maps, Occulsion-Sensitivity, etc.). A binarized mask mask_I is then generated by clipping the values in the map according to a fixed threshold τ. For each pixel that is significant, the mask also includes all neighboring pixels within a set window. The attention mask is calculated as

$$\mathrm{mask}_I = \mathcal{R}_\Omega(\mathcal{T}_\tau(\mathrm{map}_I)), \tag{4}$$

where $\mathcal{T}_\tau$ represents a hard thresholding with threshold τ and $\mathcal{R}_\Omega$ applies a max-pooling operation within a local neighbourhood Ω. In the following experiments, we demonstrate that the optimal Ω is a 30×30 neighbourhood. The attention mask is then multiplied pixel-wise with the input image I to get the filter input I_l to the local branch

$$I_l = I \odot \mathrm{mask}_I. \tag{5}$$

The masked input image I_l is then fed to the local branch f_l to obtain the corresponding pooling output pool_l

$$\mathrm{pool}_l = f_l(I_l). \tag{6}$$

Finally, the fusion branch combines the features extracted by the global and local branch pool_g and pool_l and sequentially applies max-pooling, fully connected and activation layers to generate a final prediction $\hat{\omega}_f$ as represented in Figure. 1.

A. TRAINING STRATEGY

A two-stage training strategy was employed to train the proposed three-tier architecture. During the initial stage of training, the global branch was trained on the raw input images to minimize the classification loss. We used binary

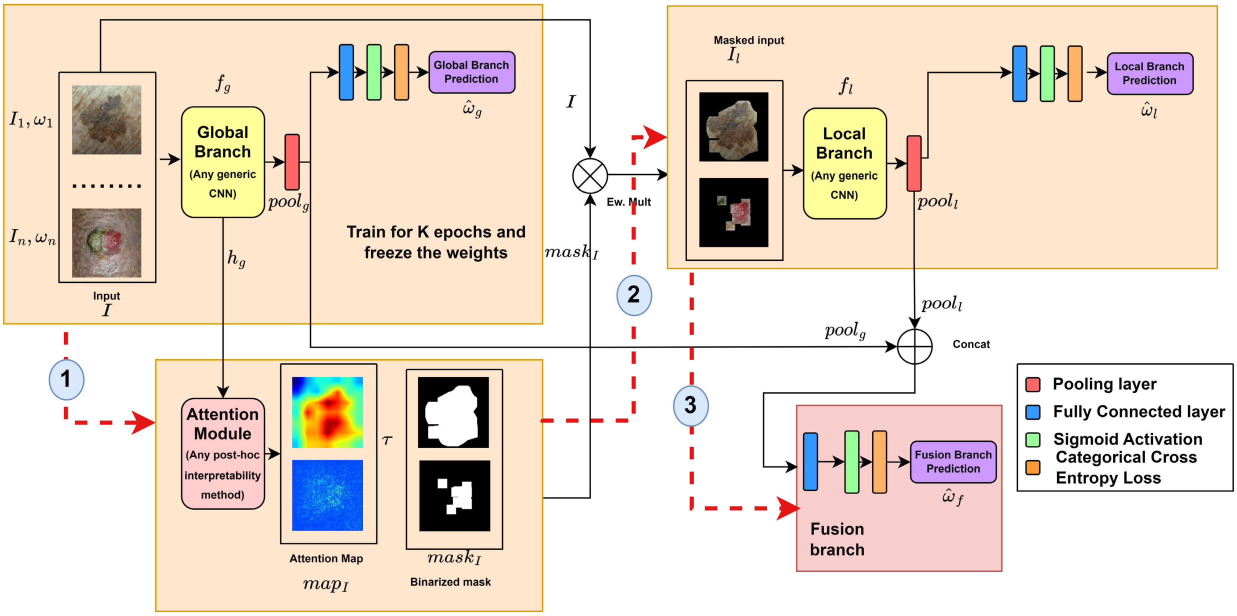


FIGURE 1. Proposed 3-tier Architecture. The global branch is a base CNN model. The attention module provides localization of relevant structures to the local branch using a post-hoc visual explainability method. The fusion of the global and local branches obtains the final prediction. Red dashed lines represent the major process flow in the order 1, 2 and 3.

cross-entropy (for the skin lesion's dataset) and categorical cross-entropy (for the custom blob dataset) to optimize the weight parameters of the global branch. The following subsections detail the training of all the branches in the architecture.

1) GLOBAL BRANCH

The global branch was trained on the dataset in a completely supervised manner by minimizing the classification loss between the final dense layer output with sigmoid activation and the ground truth labels. The final sigmoid normalization of the output vector $p(c|I)$ is given by

$$\widetilde{p}(c|I) = \frac{1}{1 + e^{-p(c|I)}}, \tag{7}$$

where I denote the input image and $\widetilde{p}(c|I)$ represents the probability score of I that belongs to class $c \in \{1, 2, \ldots, C\}$. The global branch was trained by minimizing the categorical cross entropy (CCE) loss defined as

$$\text{CCE} = \sum_{i=1}^{N_{im}} \sum_{j=1}^{C} \omega_{ij} \log(p_{ij}), \tag{8}$$

where ω_{ij} is a one-hot encoded class label corresponding to the ground truth of the ith image with respect to the jth class and p_{ij} represent the posterior probability score of the ith image belonging to the jth class. The global branch is trained for k epochs, after which the validation loss increases, and the weights W_g are frozen. In the present study, we fixed k as 30 epochs for the synthetic blob classification, and for the skin lesion classification, we fixed k as 120, 60, 80 and 60 for different backbones DenseNet-121, InceptionV3, Xception and ResNet50, respectively.

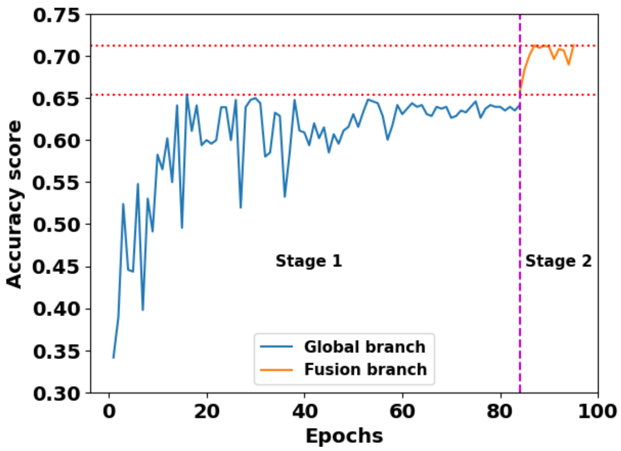


FIGURE 2. Accuracy trend during training of Xception backbone and GradCAM on the skin lesions dataset.

2) LOCAL AND FUSION BRANCHES

The local branch is fed with I_l, which is the masked image produced by the global and attention branches for the input image I and a set threshold τ. The local branch is trained for fewer epochs than the global branch as it refines the knowledge gained by the global branch. The results from the local and global branches were then fed into the fusion module for the final classification. Algorithm 1 summarizes the proposed two-stage training strategy. Figure.2 shows the training curve for global (stage 1) and fusion branches (stage 2) of Xception on the skin lesions dataset. The figure indicates that the global branch plateaued after 70 epochs, as evidenced by the minimal variation in the accuracy curve. On the other hand, the local branch (stage 2) demonstrates a noteworthy improvement in accuracy of, approximately 5%, over the global branch.

Algorithm 1 Two-Stage Training Algorithm

```
Input: $I$ (input image), Ground truth label vector $\omega$, Threshold $\tau$
Output: Probability score $\widetilde{p}_f(c|[I, I_l])$
Initialization: Initialize the weights of the global and local branches.

Step 1: Train global branch
    Learn $W_g$ using Image $I$
    Compute $p_g(c|I)$, and optimize using Categorical Cross-Entropy (CCE) for k epochs

Step 2: Obtain attention map
    Compute the attention map $map_I$ using the convolutional outputs of the global branch.
    Compute the mask $mask_I$ and the masked image $I_l$ using threshold $\tau$.

Step 3: Train local branch
    Learn $W_l$ with $I_l$
    Compute $p_l(c|I_l)$ and optimize using CCE with the frozen $W_g$.

Step 4: Model fusion:
    Concatenate $pool_g$ and $pool_l$,
    Learn $W_f$ (Weights of fusion network)
    Compute $\widetilde{p}_f(c|[I, I_l])$, optimize by CCE.

Output the Probability score
```

IV. EXPERIMENTS, RESULTS AND ANALYSIS

This section describes an evaluation of the proposed TS-CNN architecture. This paper presents two sets of experiments.

1) Experiment 1 (Synthetic blobs): The first set uses a custom shallow CNN as the backbone for both the global and local branches and employed GradCAM and saliency maps as the attention modules.
2) Experiment 2 (Skin lesions' database): The second experiment adopted transfer learning, leveraging cutting-edge convolutional neural networks (CNNs) as the foundation for both global and local branches. It also integrates the GradCAM and saliency maps as attention modules. This approach is employed in response to the scarcity of data in many medical image classification tasks, which often require transfer learning. These models typically achieve near-optimal performance, and the experiment aims to improve this ceiling.

TABLE 1. CNN models used as the global and the local branches for the experiments (1 & 2).

CNN	Trainable parameters	Training Modality	Hyperparameters
Custom CNN	172,707	End-End Training	Learning Rate: 0.001 Batch Size: 4
DenseNet-121	8.1M	Transfer Learning ImageNet weights	Learning Rate: 0.0001 Batch Size: 4
InceptionV3	23.9M	Transfer Learning ImageNet weights	Learning Rate: 0.0001 Batch Size: 4
Xception	22.9M	Transfer Learning ImageNet weights	Learning Rate: 0.0001 Batch Size: 4
ResNet-50	25.6M	Transfer Learning ImageNet weights	Learning Rate: 0.0001 Batch Size: 4

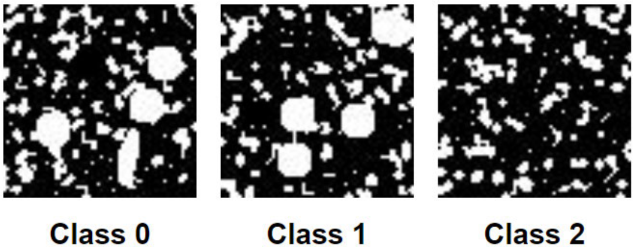


FIGURE 3. Sample images from the custom blob dataset.

Table.1 tabulates the details of different CNN architectures considered in the experiments detailed above. All the experiments described in this paper were conducted using a machine with the following configuration: 16 GB RAM, an RTX GeForce 3060 Laptop GPU with 6 GB VRAM, and a Ryzen 7 CPU. This hardware setup provided the necessary computational resources to effectively train and evaluate the proposed model.

A. SYNTHETIC BLOBS

This experiment used a synthetic blob dataset of three classes to evaluate the proposed TS-CNN model. The dataset was designed such that each class was represented equally in the dataset. Furthermore, the complexity of the patterns in the synthetic dataset was carefully controlled to ensure that they were neither too simple nor too complex. This is important because if the images are too simple, the model may overfit and perform poorly on real-world images. However, if the patterns are too complicated, the model may struggle to learn the relevant features and perform poorly on synthetic and real-world images.

1) DATASET DESCRIPTION

The blob dataset was constructed as a representative of a wide variety of medical and biological images with scattered ROIs. The custom blob dataset consists of 6000, 64×64 images with random blob-like structures of three different types. Images of class 2 consist of small blobs that cover 20% of the image, with an average area fraction of approximately 0.36% of the image area. Class 1 images are similar to Class 2, but it adds four large circular blobs with a radius of 5 pixels to the image. Class 0 differs from Class 1 in that one of the large circular

TABLE 2. Train Test split of the synthetic blob dataset.

Class	Training set	Testing set	Total
0	400	100	500
1	400	100	500
2	400	100	500
$\sum$	**1200**	**300**	**1500**

blobs is replaced by a large elliptical blob that is randomly oriented with a major axis length of 13 pixels and a minor axis length of 3 pixels. The dataset was balanced between the three classes to ensure that the evaluation was fair and unbiased. Figure. 3 shows sample images from the three classes (class 0,1,2 respectively) and Table. 2 shows the data distribution across all three classes.

2) GLOBAL AND LOCAL CNN ARCHITECTURE

A custom, 8-layered shallow CNN was used to classify the custom blob dataset by employing alternating convolutional and max-pooling layers. This model was selected because of the simplicity of the dataset in terms of classification complexity. The aim of using a simple model is to facilitate the interpretation of predictions and reduce the computational complexity. Employing a straightforward model allows for an easier analysis of the attention mechanism and its impact on the classification output, leading to a better understanding of the behavior of the model and the factors influencing its predictions. The simplified model also facilitates efficient training and testing, leading to a more effective evaluation of its performance.

3) ATTENTION MODULE

The attention branch of the proposed model was designed to be adaptable to various post-hoc interpretability methods. To demonstrate this adaptability, two widely used interpretability techniques were utilized in the experiment: GradCAM [18] and saliency map [14]. GradCAM, or Gradient Class Activation Mapping, employs gradients to determine the region of interest in the model for making classification decisions. This technique provides accurate but coarse localization of the model's attention and closely resembles human attention in various tasks. On the other hand, the saliency map is a basic method for capturing the attention of the model by providing pixel-wise importance of image regions in classification. This method uses feature maps to generate attention maps, making them more computationally efficient.

4) TRAINING, RESULTS AND ANALYSIS

The custom shallow CNN, as detailed in Section IV-A2, was trained using the custom blob dataset outlined in Section IV-A1, employing an end-to-end training strategy. The custom CNN that forms the global branch was pre-trained for 60 epochs with an Adam optimizer to minimize a categorical cross-entropy loss function. The learning rate was initially set to 0.001 and decreased by a factor of 0.2 upon reaching a plateau. The weights and biases obtained from the global branch were used to initialize the local model, which helped expedite the training process of the entire architecture. Subsequently, the complete architecture was trained for 30 epochs using an Adam optimizer with a learning rate of 0.001. Table. 3, tabulates the results obtained from the proposed TS-CNN architecture with the custom shallow CNN as the backbone for global and local branches and GradCAM as the attention module. The table shows that the local module slightly improves the classification accuracy with respect to the global reference model (97% to 97.33%). Indeed, the local model is trained on the relevant regions, as shown in Figure. 4. The fusion model further improves accuracy (97% to 98.33%) because it considers the patterns learned by both the local and the global modules to make predictions. The same trend can be observed in all the other evaluation metrics considered for the study (Precision: an increment of 1.15% between global and fusion, Recall: an increment of 1.33%).

Figure. 4 visually shows the region of interest considered by the global and local branches to make predictions. The

TABLE 3. Results from the classification of blob dataset with custom CNN and GradCAM.

custom CNN + GradCAM	Accuracy	Precision	Recall	F1	AUC
Global	97	97.25	97	96.99	1
Local	97.33	97.44	97.33	97.31	98.72
Fusion	98.33	98.4	98.33	98.33	99.66

TABLE 4. Results from the classification of blob dataset with custom CNN and saliency map.

Custom CNN + Saliency	Accuracy	Precision	Recall	F1	AUC
Global	97	97.25	97	96.99	1
Local	86.33	86.59	86.33	86.23	96.88
Fusion	99	99.1	99	98.99	99.99

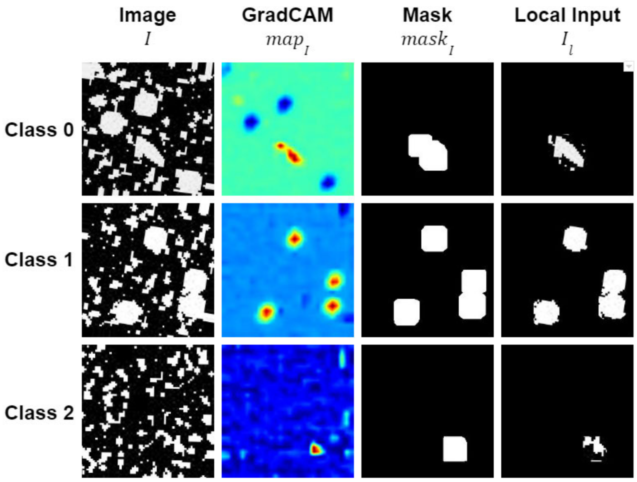


FIGURE 4. Outputs with custom CNN as backbone and GradCAM as the attention module.

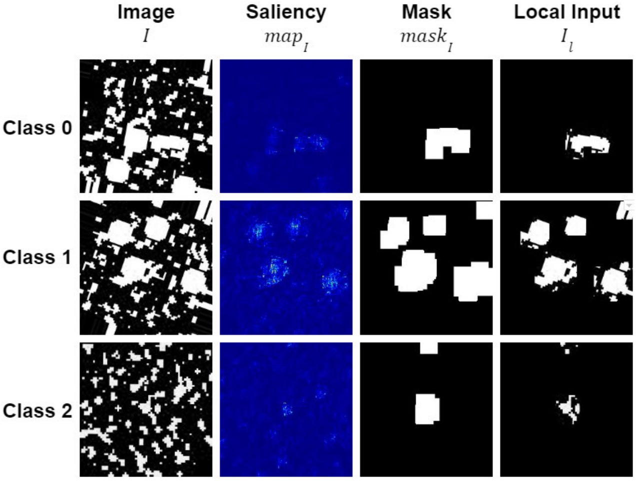


FIGURE 5. Outputs with customCNN backbone and saliency map.

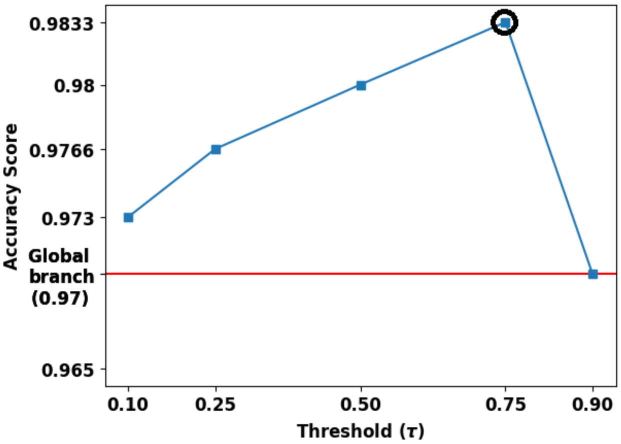


FIGURE 6. Accuracy vs Threshold plot for the Blob dataset classification using custom CNN. The red line shows the effect of thresholding on the global branch, and blue line represents that of the local branch.

first column represents the image input into the global branch, and the GradCAM heat map generated from the final convolutional block of the global branch is shown in the second column. The heat map illustrates that the global branch focuses on elliptical structures to determine class 0 (refer to the first row of the second column), larger blobs to detect class 1 (refer to the second row of the second column), and relatively larger blobs (less in number compared to class 1) to determine the class 2. As in the third column, a 30 × 30 neighborhood was generated from the pixels of the obtained ROI (regions with high intensity in the heatmaps) as the attention masks. The neighborhood was fixed at 30 × 30 experimentally. The attention masks are applied to the original image to remove noisy pixels that do not contribute to disease identification and to highlight the region of interest, as in column.4. The noise-free enhanced images were then fed into the local branch for further classification. The local branch only considers the relevant image regions for prediction, making it more accurate and less complex. The same trend was observed visually and quantitatively for the model with custom CNN as the backbone and saliency map in the attention module, as shown in Figure. 5 and Table. 4, respectively. Remarkably, although saliency maps provide poorer localization of the relevant structures, the proposed architecture still provides an improved classification accuracy, thus indicating that the approach is robust under different attention strategies.

5) HYPERPARAMETER TUNING

The selection of hyperparameters plays a crucial role in the performance of the proposed model. In this study, two key hyperparameters were investigated:

- **Threshold value** (τ): This hyperparameter is used to threshold the heatmap generated by the attention module. It determines the number of pixels retained to construct the mask ($mask_I$) from the attention map (map_I).
- **Neighbourhood size** (Ω)**:**This hyperparameter determines the size of the relevant neighbourhood selected from the mask ($mask_I$) generated.

A series of experiments were conducted to determine the optimal value for Threshold (τ) and the neighbourhood Ω. **Selection of Threshold value** (τ)**:** The threshold value determines the number of pixels retained to construct the mask ($mask_I$) from the generated attention map (map_I). The proposed model was trained for various threshold values in intervals of 0.25 in the range [0,1]. To avoid passing the exact input image (threshold = 0) or obtaining a completely blank image (threshold = 1), the endpoints of the range were replaced with 0.1 and 0.9, respectively. The results, as shown in Figure.6, demonstrated the model's performance across different threshold values. Notably, a threshold value of 0.75 emerged as the best-performing hyperparameter setting.

Selection of Neighbourhood size (Ω)**:** The optimal neighbourhood size is determined by experimenting with window sizes from (10 × 10) up until (50 × 50). The aim was to determine the optimal size for capturing relevant information from the original image. As depicted in Figure.7, the experimental results showcased the model's performance under different neighbourhood sizes. Among the tested sizes, a neighbourhood size of (30 × 30)was observed to be the most suitable fit, as it yielded the best results in terms of classification accuracy.

B. SKIN LESIONS' DATASET

The primary objective of this experiment was to demonstrate the generalizability and adaptability of the proposed TS-CNN architecture for real-world image classification tasks. A real-world skin lesion dataset and state-of-the-art deep CNN architectures were used for evaluation. The results show that the proposed architecture is a suitable solution for real-world problems and can be used on top of any generic classification frameworks.

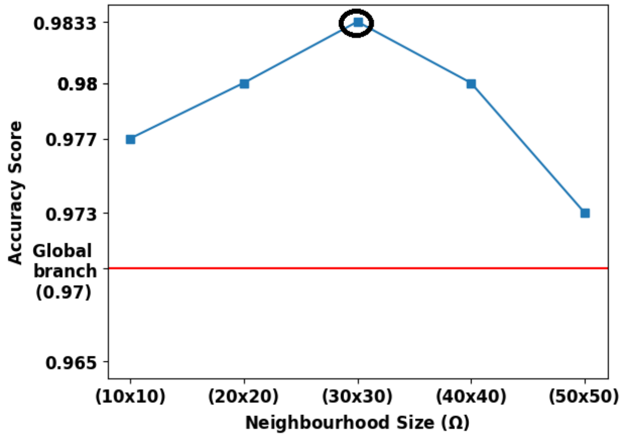


FIGURE 7. Accuracy vs Neighbourhood size plot for the Blob dataset classification using custom CNN. The red line shows the effect of neighbourhood on the global branch, and the blue line represents that of the local branch.

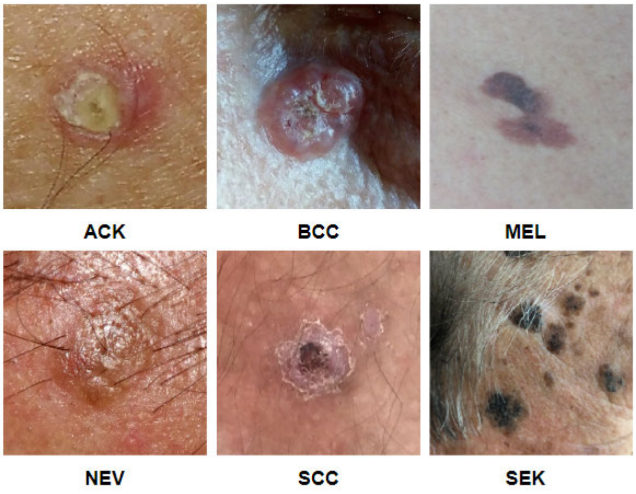


FIGURE 8. Sample Images from the PAD-UFES-20 dataset corresponding of six different types of lesion.

1) DATASET DESCRIPTION

PAD-UFES-20 [28], [33] is a collection of patient data and clinical images of skin lesions collected using smartphones. It was developed by the Universidade Federal do Espirito Santo (UFES) in Brazil specifically to develop and evaluate algorithms for automated classification of skin lesions. The dataset includes images of six types of skin lesions, namely Basal Cell Carcinoma (BCC), Actinic Keratosis (ACK), Nevus (NEV), Seborrheic Keratosis (SEK), Squamous Cell Carcinoma (SCC), Melanoma (MEL) and among which 3 are cancerous (BCC, MEL, and SCC). The lesions are from over 120 anatomical regions of the body and thus provide a rather representative collection. The images were collected from 160 patients; each patient contributed between 1 and 11 images. Figure. 8 represents sample images from the dataset and Table. 5 represents the data distribution. Table. 5 tabulates the train/test split up for each class. It is worth noting that the base reference paper [28] utilized a subset of the PAD-UFES-20 dataset, which comprised 1612 samples. In contrast, this study employs the complete dataset of 2298 skin lesion samples. The selection of this dataset was motivated by its inclusion of low-quality images, enabling an investigation into the effectiveness of the proposed model in handling such challenging image conditions.

TABLE 5. Train Test split of PAD-UFES-20 dataset.

Clinical Diagnosis	Training set	Testing set	Total
ACK	584	146	730
BCC	676	169	845
MEL	42	10	52
NEV	195	49	244
SCC	153	39	192
SEK	188	47	235
$\sum$	**1838**	**460**	**2298**

2) GLOBAL AND LOCAL CNN ARCHITECTURE

To demonstrate the generalizability of the proposed 3-tier architecture, various state-of-the-art DCNNs were tested for global and local branches, including DenseNet-121 [34], InceptionV3 [35], Xception [36], and ResNet-50 [37]. DenseNet-121 is a dense convolutional network that connects each layer to every other layer in a feed-forward manner. It is known to be one of the most accurate models while being computationally light. InceptionV3 is a faster and less computationally expensive model than its ancestors. It has a deeper network than its predecessors without compromising on speed. Xception network is inspired by the Inception network architecture, but it replaces the Inception modules with depthwise separable convolutions, which help outperform it in classification tasks while having the same number of parameters. ResNet-50 uses residual connections to train much deeper neural networks effectively. The residual connections allowed the model to perform better by protecting it from the vanishing gradient problem common in DCNNs. The CNNs mentioned above utilize transfer learning using ImageNet weights. This involves the addition of a global max pooling layer, followed by a dense layer, batch normalization, leaky ReLU activation, and finally, a dense layer with softmax activation. This approach facilitates a comprehensive evaluation of the performance of different CNN architectures on the same dataset, providing insights into the most effective models for the task at hand. By employing multiple models, we can assess the robustness and transferability of the proposed architecture, thus expanding its potential applications to a broader range of tasks. GradCAM and saliency maps were employed as attention modules, as described in the previous experiment.

3) TRAINING, RESULTS AND ANALYSIS

For each of the CNN backbones, namely DenseNet-121, InceptionV3, Xception, and ResNet-50, both GradCAM

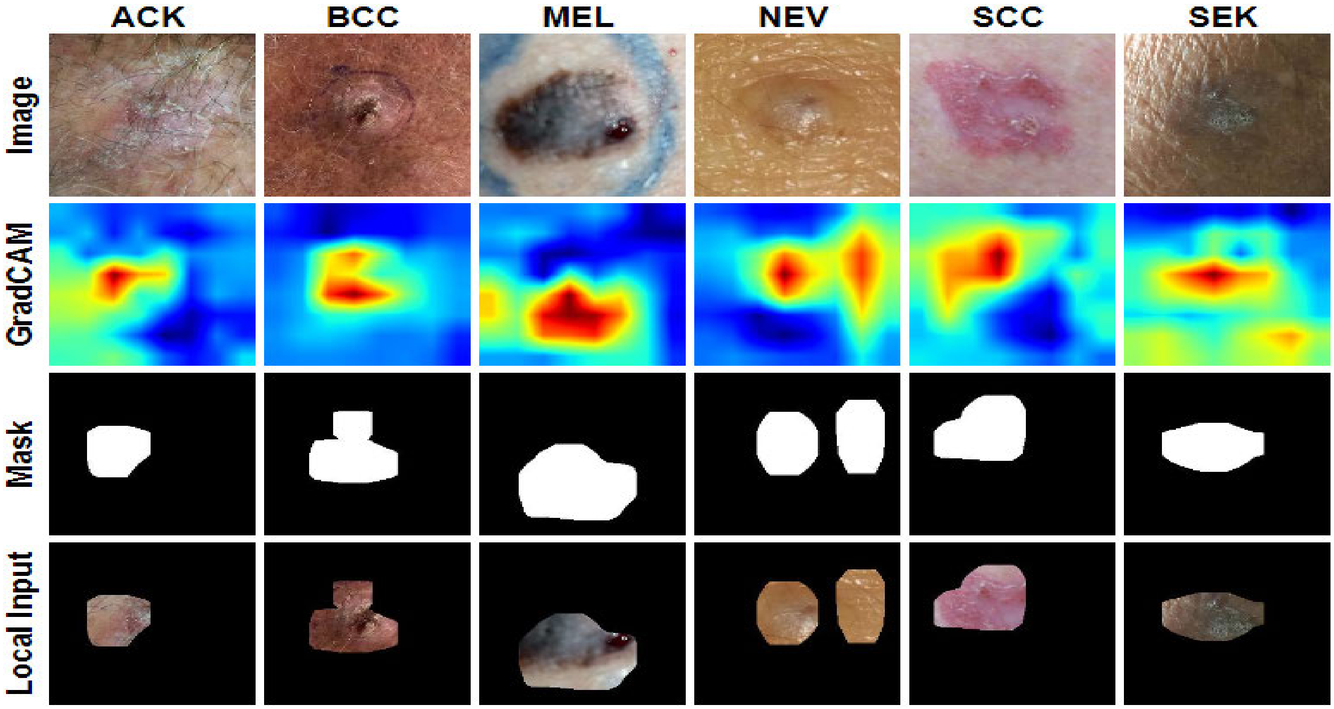


FIGURE 9. Outputs with DenseNet backbone and GradCAM.

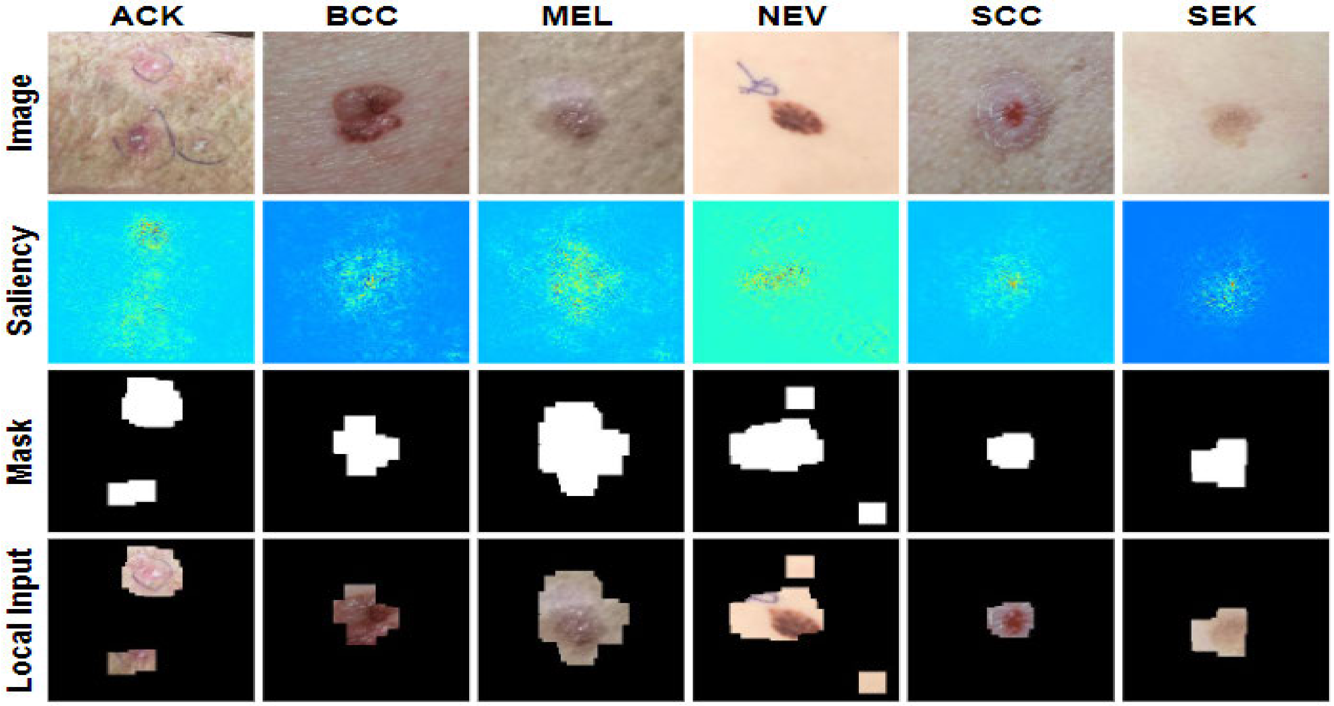


FIGURE 10. Outputs with DenseNet backbone and saliency map.

and saliency maps were used, and their respective outputs are shown in Figure.9, through Figure.14 respectively. The figures show various images, each displaying a different type of lesion. This intentional selection of various lesion types serves two purposes. Firstly, it allows for a comprehensive evaluation of the proposed models' performance across different categories, assessing their ability to classify diverse skin lesions accurately. Secondly,

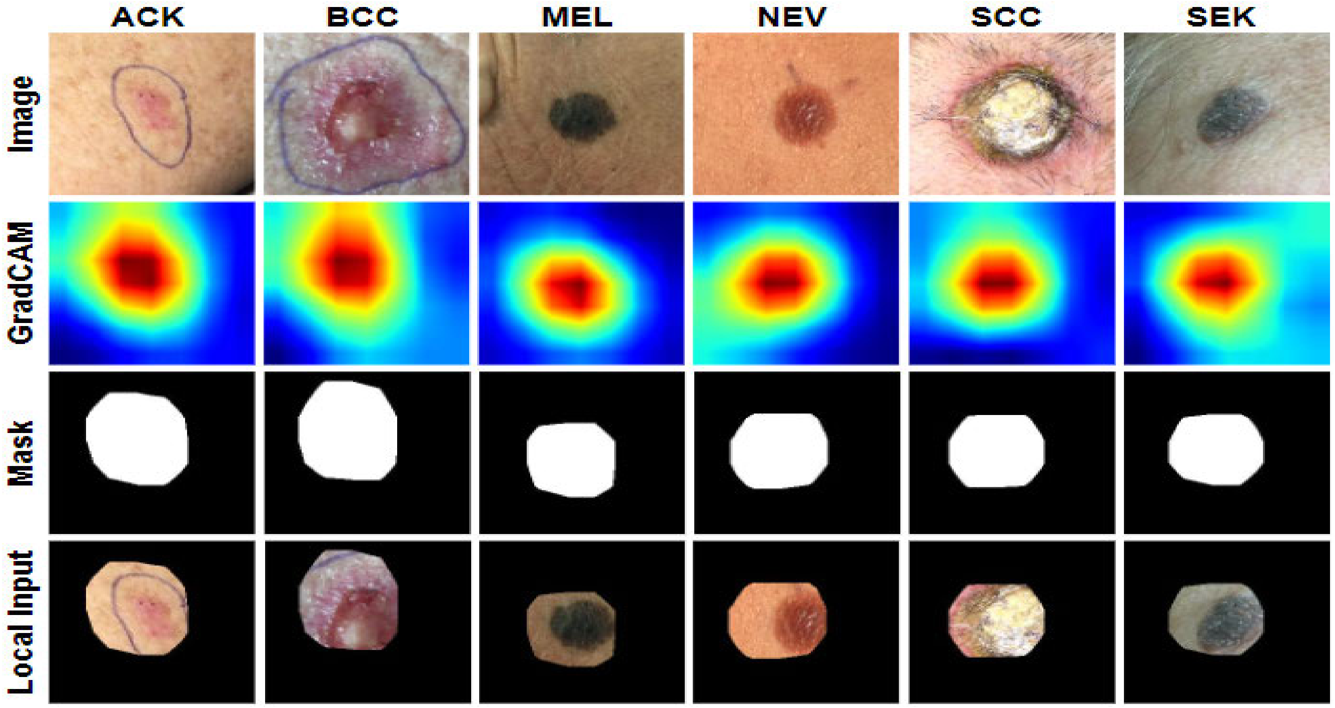


FIGURE 11. Outputs with InceptionV3 backbone and GradCAM.

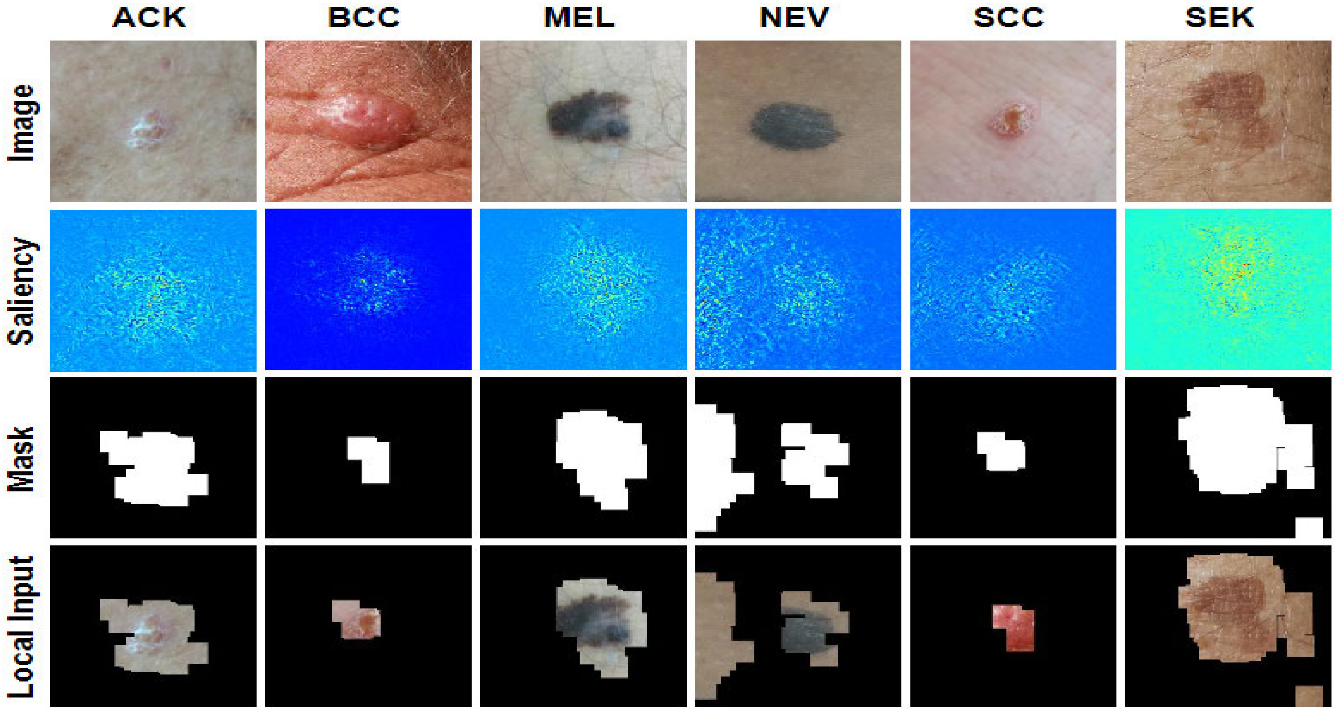


FIGURE 12. Outputs with InceptionV3 backbone and saliency map.

it demonstrates the generalizability of the models, as they are expected to perform well on the specific lesion types present in the training data and unseen lesion categories also.

Similar to custom CNNs, the outputs of the three different models exhibited comparable outcomes. Additionally, the local model integrates attention maps to eliminate image noise, enabling a more precise focus on the relevant regions.

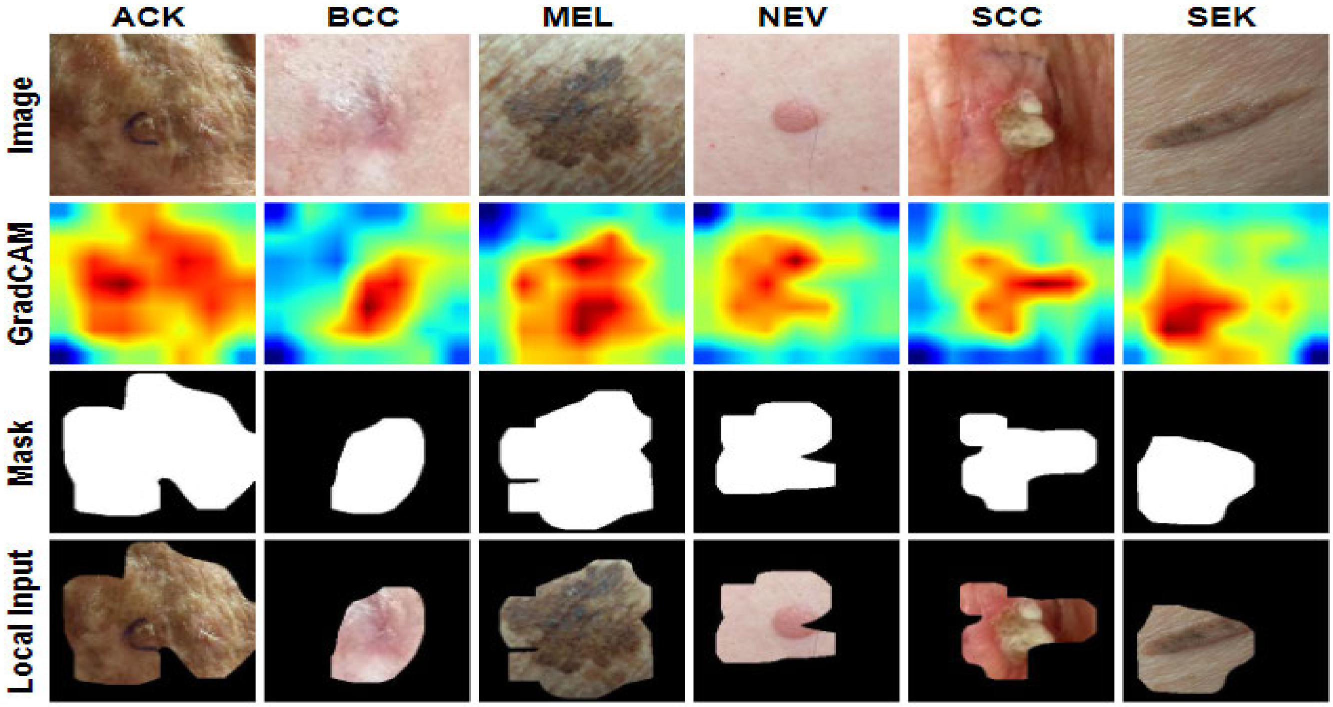


FIGURE 13. Outputs with Xception backbone and GradCAM.

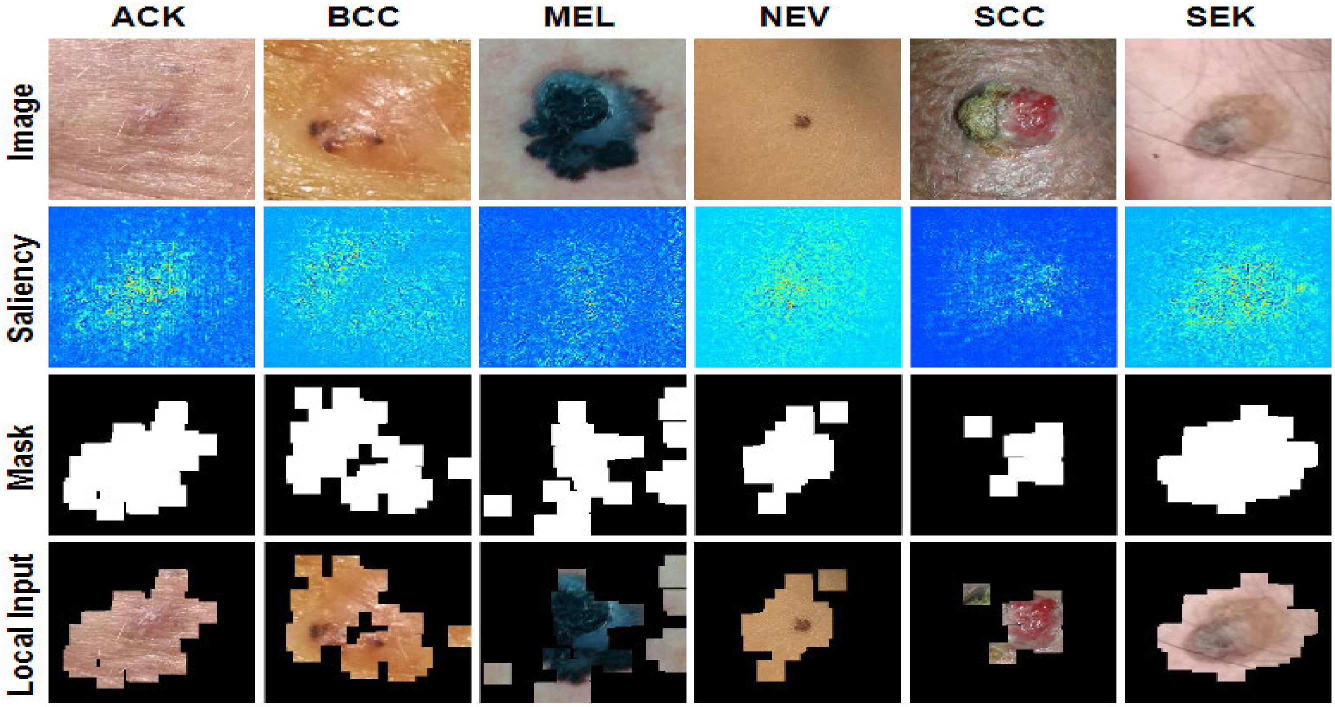


FIGURE 14. Outputs with Xception backbone and saliency map.

As a result, the classification performance of the fusion model was enhanced, as shown in Table. 7 and Table. 8. Most of the images have a single centralized region of interest. This is especially seen in Figure. 9 through Figure. 16. Both GradCAM and the saliency maps captured the region of interest. However, the advantage of using pixel-wise attention maps such as GradCAM and saliency maps becomes more relevant when the images present disjoint regions

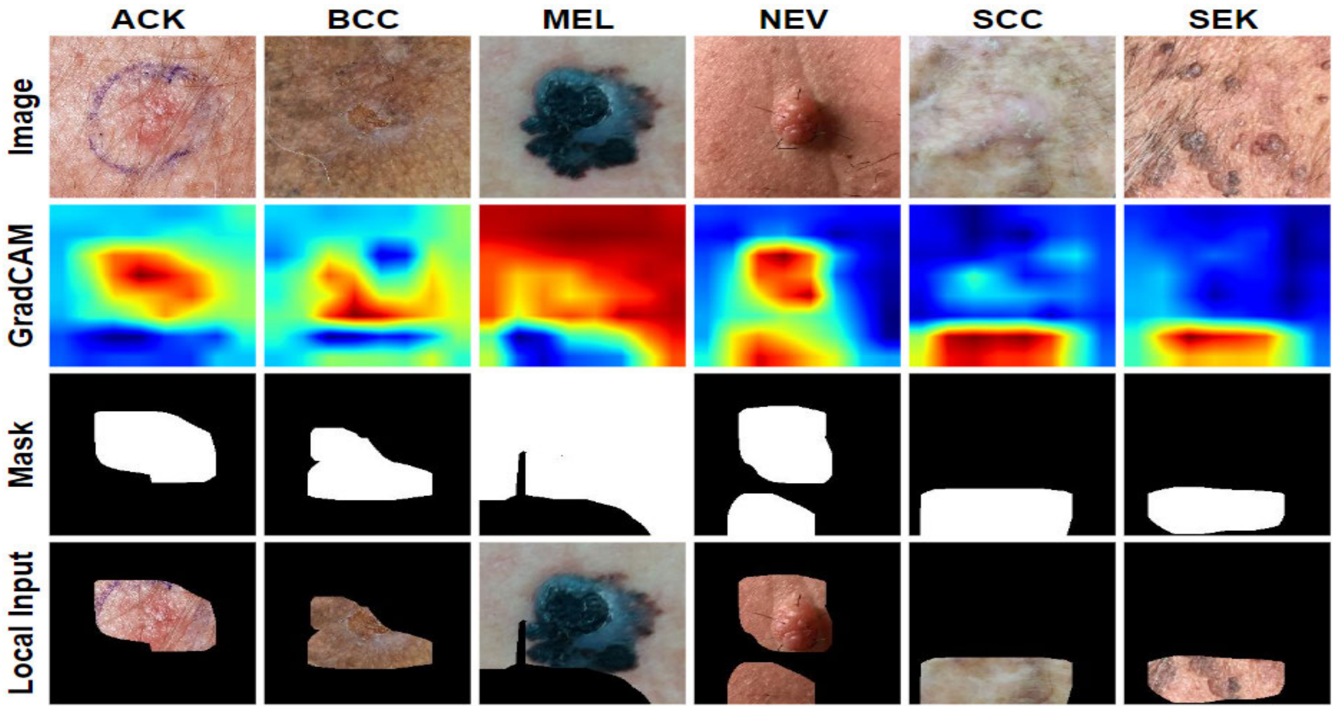


FIGURE 15. Outputs with ResNet-50 backbone and GradCAM.

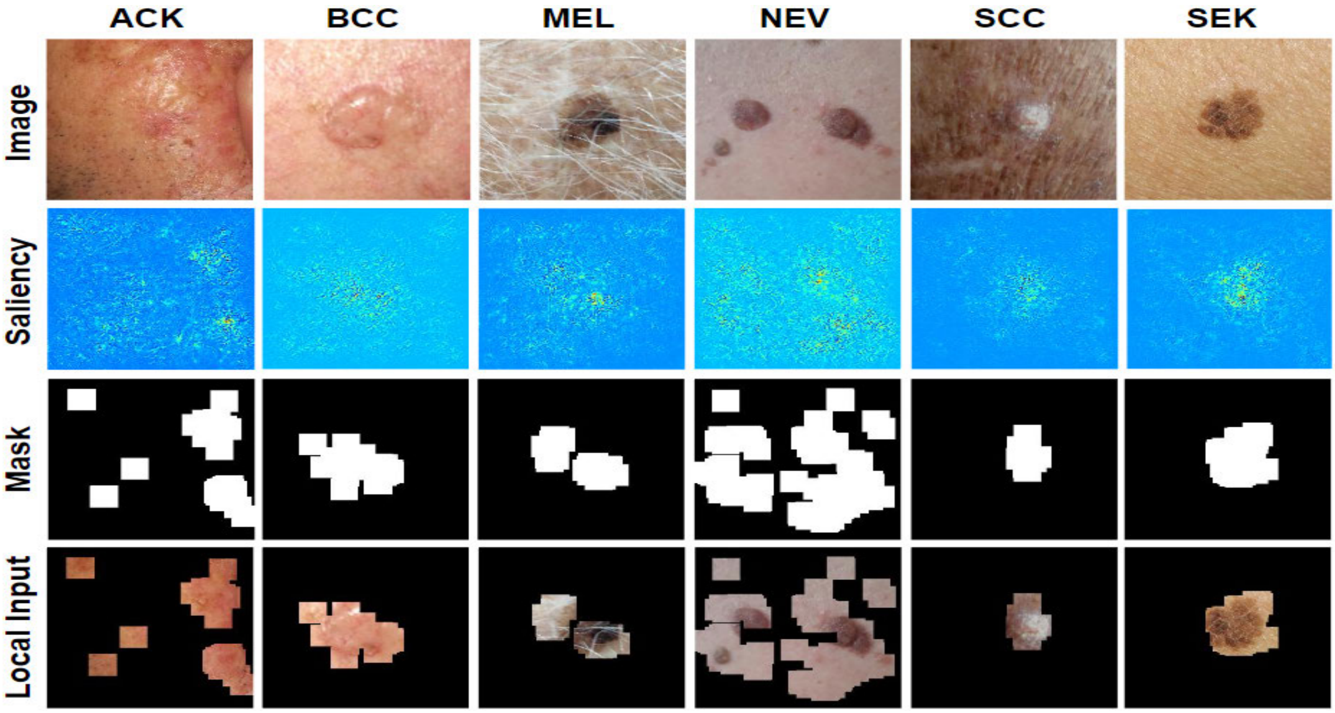


FIGURE 16. Outputs with ResNet-50 backbone and saliency map.

of interest, as in the examples shown in Figure. 10 (first column, class ACK), Figure.14 (columns 2 and 5, classes BCC and SCC) and Figure.16 (columns 1 and 4, classes ACK and NEV). In these examples, the attention maps captured disjoint regions of interest, making the approach versatile and applicable to various datasets. Comparing the results in Table 6 and Table. 7, it is evident that the proposed model using DenseNet-121 as its backbone is able to outperform

TABLE 6. Results from the classification of PAD-UFES 20 dataset with the previous state-of-the-art model [28].

Model	Accuracy	Precision	Recall	F1	AUC
VGGNet-13	70.70	73.40	70.80	71.00	93.20

the existing state-of-the-art classification performance by a margin of 2.13% in terms of accuracy. Although the fusion branch significantly improves classification accuracy, we observe that the local branch experiences a decline compared to the global branch, as shown in Table. 7 and Table. 8. We interpret this observation by noting that the local model contains less information than the global branch, since the attention module removes some regions that could aid in accurate classification. However, the purpose of the local model in this architecture is not to outperform the global model, but to complement the global model to achieve better classification results in the final fusion branch. This phenomenon is not observed in the blob dataset reported in the previous section because the regions of interest are too simple, and almost everything removed by the attention module is just noise. It is also notable from Table. 8 that the drop in performance of the local model is larger when using saliency maps to generate the mask as compared to the GradCAM version, which is consistent with the fact that a poorer localization of the structures provides less meaningful information to the local branch. It is known that GradCAM captures the model's attention better than the saliency map [38]. However, it is worth noting that the saliency map version performs better than the GradCAM version when using the InceptionV3 backbone (69.56% with GradCAM vs 70.65% with Saliency), demonstrating the adaptability of the architecture to different backbone CNN models and attention capture methods. Additionally, this suggests that the architecture can improve on the global branch even with less effective interpretability techniques. Table. 9 summarizes the performance gain of the proposed TS-CNN model over different backbones (DenseNet-121, InceptionV3, Xception, and ResNet-50) and post-hoc methods (GradCAM and Saliency) on the PAD-UFES-2 skin lesion dataset. The results show that it consistently improves the performance of all the underlying classification frameworks by a high margin across all evaluation metrics (accuracy, precision, recall, F1 score, and AUC). The InceptionV3 backbone with the saliency map version achieved the highest overall performance gain with the highest increase in accuracy, precision, recall, F1 score, and AUC compared with other backbone CNNs and attention capture methods. These results demonstrated the adaptability and effectiveness of the proposed TS-CNN architecture with different backbone CNNs and attention mechanisms.

V. ROBUSTNESS TO NOISE

The ablation studies conducted in this research aim to assess the model's robustness in the presence of noise in the input images. Two types of noise were chosen for testing purposes:

- **Salt and Pepper noise:** A fixed amount of random pixels are set to white or black. The experiment utilized three noise levels, namely 1%, 5%, and 10%.
- **Gaussian noise:** Gaussian noise introduces random variations to each pixel by sampling from a normal distribution. The distribution's mean(μ) is set to 0, and

TABLE 7. Results from the classification of PAD-UFES 20 dataset with Transfer-learned SOTA CNNs and GradCAM.

Backbone CNN	Branch	Accuracy	Precision	Recall	F1	AUC
DenseNet-121	**Global**	69.13	72.54	69.13	69.72	90.23
	Local	60.21	61.59	60.21	60.58	82.64
	Fusion	**72.83**	**72.61**	**72.83**	**72.34**	89.68
InceptionV3	**Global**	63.48	67.84	63.48	64.13	86.35
	Local	58.26	58.35	58.26	57.07	80.89
	Fusion	**69.56**	**69.09**	**69.56**	**68.32**	**87.07**
Xception	**Global**	65.43	71.12	65.43	66.73	87.59
	Local	60.65	64.79	60.65	61.92	83.33
	Fusion	**71.3**	70.45	**71.3**	**70.13**	**89.42**
ResNet-50	**Global**	68.04	68.43	68.04	67.95	87.51
	Local	59.13	63.24	59.13	59.83	78.23
	Fusion	**69.13**	**68.92**	**69.13**	**68.56**	85.44

TABLE 8. Results from the classification of PAD-UFES 20 dataset with Transfer-learned SOTA CNNs and Saliency.

Backbone CNN	Branch	Accuracy	Precision	Recall	F1	AUC
DenseNet-121	**Global**	69.13	72.54	69.13	69.72	90.23
	Local	49.78	53.31	49.78	50.21	76.98
	Fusion	**72.6**	71.7	**72.6**	**71.5**	90.12
InceptionV3	**Global**	63.48	67.84	63.48	64.13	86.35
	Local	53.69	53.89	53.69	52.31	77.89
	Fusion	**70.65**	**69.96**	**70.65**	**69.22**	**86.92**
Xception	**Global**	65.43	71.12	65.43	66.73	87.59
	Local	57.39	59.82	57.39	58.3	80.86
	Fusion	**71.52**	70.45	**71.52**	**70.11**	**89.7**
ResNet-50	**Global**	68.04	68.43	68.04	67.95	87.51
	Local	45.43	55.21	45.43	47.19	75.09
	Fusion	**68.69**	**68.51**	**68.70**	67.15	86.37

TABLE 9. Improvement of performance in Fusion branch over Global branch for each backbone and posthoc method combination.

Backbone CNN	Accuracy	Precision	Recall	F1	AUC
DenseNet-121 + GradCAM	+3.7	+0.07	+3.7	+2.62	-0.55
DenseNet-121 + Saliency	+3.47	-0.84	+3.47	+1.78	-0.11
InceptionV3 + GradCAM	+6.08	+1.25	+6.08	+4.19	+0.72
InceptionV3 + Saliency	+7.17	+2.12	+7.17	+5.09	+0.57
Xception + GradCAM	+5.87	-0.67	+5.87	+3.4	+1.38
Xception + Saliency	+6.09	-0.67	+6.09	+3.38	+2.11
ResNet-50 + GradCAM	+1.09	+0.49	+1.09	+0.61	-2.07
ResNet-50 + Saliency	+0.65	+0.08	+0.66	-0.80	-1.14

TABLE 10. Results on the robustness of the proposed TSCNN architecture with DenseNet-121 as the back bone on PAD-UFES-20 datase.

Attention Maps	Salt and Pepper Noise			Gaussian Noise (μ = 0)		
	1%	5%	10%	σ = 0.01	σ = 0.05	σ = 0.10
GradCam	49.13	10.65	10.65	20.22	11.09	10.65
	46.52	26.96	14.35	30.87	20.22	23.69
	54.35	30.43	10.65	38.69	7.61	3.91
Saliency	49.35	10.65	10.65	20.22	11.09	10.65
	38.04	20.43	11.74	24.78	16.09	20.22
	56.52	20.65	10.65	44.35	21.74	22.61

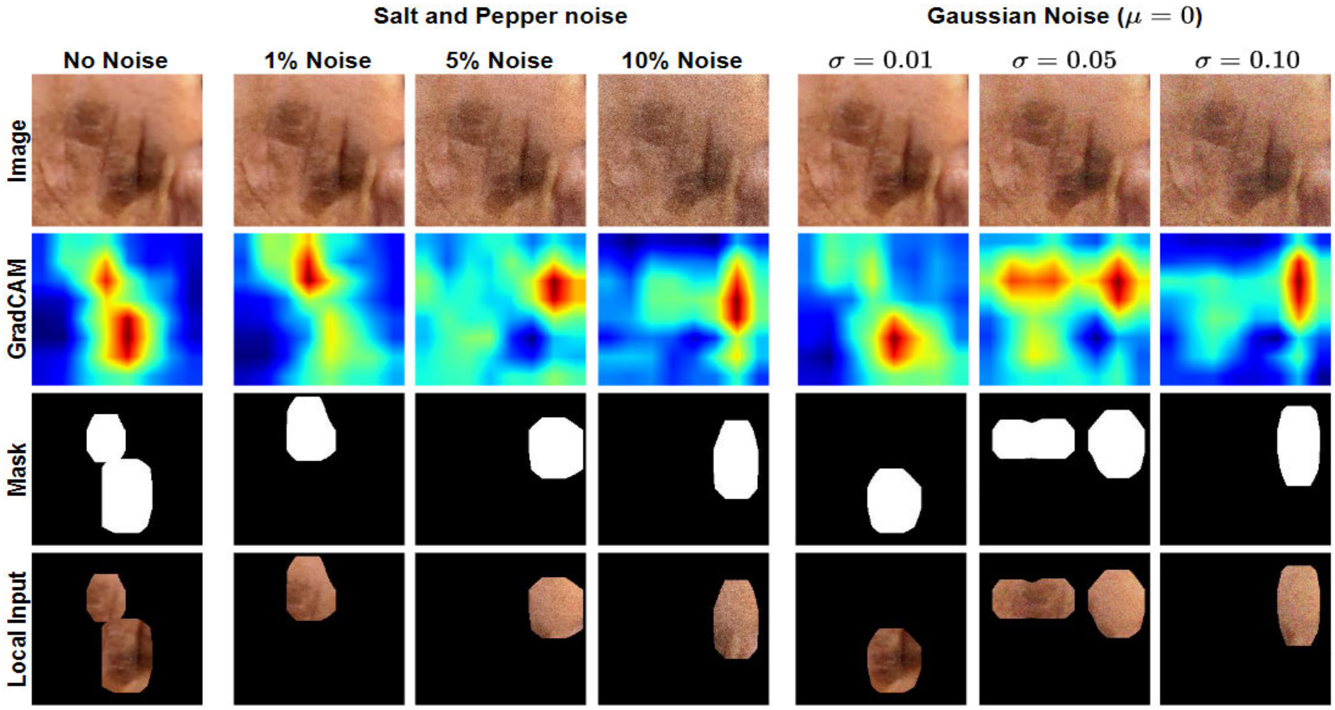


FIGURE 17. Outputs of ablation study with Denset-121 and GradCAM.

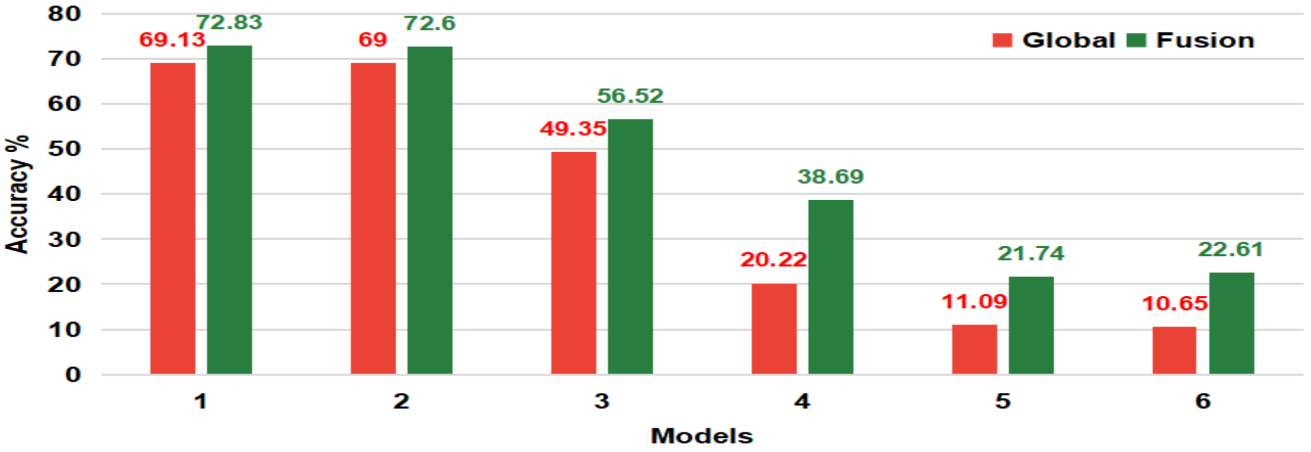


FIGURE 18. Effect of Global Model Accuracy on Fusion Model Performance: Analysis of accuracy achieved by the fusion model in relation to varying levels of accuracy of the global model on the skin lesion dataset.

the variance(σ) is varied between the values of 0.01, 0.05, and 0.10.

To evaluate the impact of noise on the TSCNN architecture, the DenseNet-121 backbone was selected for experimentation on the PAD-UFES-20 dataset. The choice of using DenseNet as the backbone was based on its superior performance in terms of accuracy when integrated into the TSCNN model for skin lesion classification. By examining the behavior of TSCNN with the DenseNet-121 backbone under noisy conditions, the study aims to understand how the architecture's robustness is affected by the presence of noise in the input images. The classification accuracy of the DenseNet-121 model on the noisy PAD-UFES-20 dataset is shown in Figure.10. It is evident that even with small amounts of noise, the model fails to classify the skin lesions properly. Moreover, the model's performance deteriorates proportionally as the noise level increases. The decline in the performance of the global model has a cascading effect on the fusion branch, since both the local and fusion models heavily rely on the activations and attention of the global model to make decisions. This dependency is expected, as the loss of proper localization in the global model's attention can be observed in Figure.17, where an increase in noise results in a loss of focus on the correct regions of interest. This vulnerability in the architecture represents a potential weak point. The performance of the fusion model is highly contingent upon the accurate localization of the global model's attention. Therefore, as noise levels increase, leading to compromised attention localization, the model's ability to make accurate classifications is significantly impacted. In order to gain a deeper understanding of the impact of the global model on the overall TSCNN architecture, an experiment was conducted where the global model was trained to achieve varying levels of accuracy on the skin lesion dataset. The resulting fusion model's accuracy was then analyzed and plotted in Figure.18. The findings demonstrate that the fusion model consistently achieves high accuracy levels compared to the global model. However, it is also observed that when the accuracy of the global model is lower, it has a detrimental effect on the performance of the final fusion model. This emphasizes the crucial role of the global model in influencing the overall accuracy of the TSCNN architecture.

VI. CONCLUSION

In this paper, we introduce a three-tier self-interpretable deep CNN architecture that is highly flexible in terms of its backbone models and post-hoc interpretability techniques. Through several experiments utilizing multiple backbone models and two interpretability techniques (GradCAM and saliency maps), we demonstrate the architecture's adaptability and potential for improving classification metrics while enhancing interpretability. Our findings show that the proposed architecture outperforms the backbone model in terms of classification metrics and offers a significant improvement over traditional black-box models. Moreover, it provides interpretable visualizations that shed light on the model's decision-making process, enhancing our understanding of its reasoning. Furthermore, the flexibility of the proposed architecture is demonstrated through its generalization across two different datasets, indicating its potential for wide applicability and versatility in various domains. Overall, the proposed architecture offers a promising solution to the ongoing challenge of balancing accuracy and interpretability in deep learning models. Its flexibility in both backbone models and interpretability techniques makes it an attractive option for researchers and practitioners seeking to optimize performance and transparency in their deep learning models. The major drawbacks of the proposed model include its vulnerability to noise and the heavy reliance on the effectiveness of the global branch for accurate classification. The experiments conducted with salt and pepper noise and Gaussian noise demonstrated a decline in performance as the noise levels increased, indicating a limitation in handling noisy input images. One potential future direction is to investigate techniques that enhance the model's robustness to noise, such as data augmentation methods or noise reduction algorithms. Additionally, exploring alternative architectures or modifications to the existing model could be valuable.

REFERENCES

[1] O. K. Sikha and B. Bharath, "VGG16-random Fourier hybrid model for masked face recognition," *Soft Comput.*, vol. 26, no. 22, pp. 12795–12810, Nov. 2022.

[2] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Inf. Fusion*, vol. 66, pp. 111–137, Feb. 2021, doi: 10.1016/j.inffus.2020.09.006.

[3] J. Chaki and M. Woźniak, "A deep learning based four-fold approach to classify brain MRI: BTSCNet," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104902.

[4] S. Shrinithi and J. Aravinth, "Detection of melanoma skin cancer using dermoscopic skin lesion images," in *Proc. Int. Conf. Recent Trends Electron., Inf., Commun. Technol. (RTEICT)*, Aug. 2021, pp. 240–245.

[5] D. Sudharsan, S. I. Indhu, K. S. Kumar, L. Karthikeyan, L. Srividhya, V. Sowmya, E. Gopalakrishnan, and K. Soman, "Analysis of machine learning and deep learning algorithms for detection of brain disorders using MRI data," in *Artificial Intelligence on Medical Data*. Cham, Switzerland: Springer, 2022, pp. 39–46.

[6] N. I. Hasan and A. Bhattacharjee, "Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition," *Biomed. Signal Process. Control*, vol. 52, pp. 128–140, Jul. 2019, doi: 10.1016/j.bspc.2019.04.005.

[7] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18069–18083, Dec. 2020, doi: 10.1007/s00521-019-04051-w.

[8] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105111.

[9] P. Costa, A. Galdran, A. Smailagic, and A. Campilho, "A weakly-supervised framework for interpretable diabetic retinopathy detection on retinal images," *IEEE Access*, vol. 6, pp. 18747–18758, 2018.

[10] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Med. Phys.*, vol. 47, no. 5, pp. e218–e227, 2020.

[11] K. Mridha, M. M. Uddin, J. Shin, S. Khadka, and M. F. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, vol. 11, pp. 41003–41018, 2023.

[12] A. Guna, R. Benitez, and S. Ok, "Interpreting CNN predictions using conditional generative adversarial networks," 2023, *arXiv:2301.08067*.

[13] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6254–6263.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[16] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018, doi: 10.1631/FITEE.1700808.

[17] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, doi: 10.1016/j.neucom.2021.03.091.

[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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[19] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided CNN for image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4162–4173, Nov. 2021, doi: 10.1109/TCSVT.2020.3046625.

[20] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[21] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[22] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*.

[23] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, and K. J. Geras, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, Art. no. 101908, Feb. 2021, doi: 10.1016/j.media.2020.101908.

[24] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1306–1315, May 2020, doi: 10.1109/TMI.2019.2948026.

[25] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, and M. Donovan, "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.

[26] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019, doi: 10.1016/j.media.2019.01.012.

[27] X. Liu, L. Zhang, T. Li, D. Wang, and Z. Wang, "Dual attention guided multi-scale CNN for fine-grained image classification," *Inf. Sci.*, vol. 573, pp. 37–45, Sep. 2021, doi: 10.1016/j.ins.2021.05.040.

[28] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Comput. Biol. Med.*, vol. 116, Jan. 2020, Art. no. 103545, doi: 10.1016/j.compbiomed.2019.103545.

[29] C.-H. Yeh, M.-H. Lin, P.-C. Chang, and L.-W. Kang, "Enhanced visual attention-guided deep neural networks for image classification," *IEEE Access*, vol. 8, pp. 163447–163457, 2020, doi: 10.1109/ACCESS.2020.3021729.

[30] A. Aggarwal, N. Das, and I. Sreedevi, "Attention-guided deep convolutional neural networks for skin cancer classification," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, 2019, pp. 1–6, doi: 10.1109/IPTA.2019.8936100.

[31] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 5, pp. 1405–1412, May 2020, doi: 10.1109/JBHI.2019.2949075.

[32] X. Xing, Y. Yuan, and M. Q.-H. Meng, "Zoom in lesions for better diagnosis: Attention guided deformation network for WCE image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4047–4059, Dec. 2020, doi: 10.1109/TMI.2020.3010102.

[33] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr., J. G. Esgario, A. C. Simora, P. B. Castro, F. B. Rodrigues, P. H. Frasson, R. A. Krohling, H. Knidel, M. C. Santos, R. B. do Espírito Santo, T. L. Macedo, T. R. Canuto, and L. F. de Barros, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, Oct. 2020, Art. no. 106221, doi: 10.1016/j.dib.2020.106221.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2017, pp. 1–9.

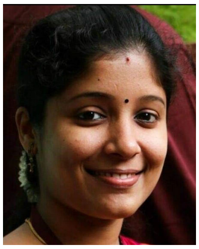
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2017, pp. 1–17.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[38] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

V. A. ASHWATH is currently pursuing the bachelor's degree in computer science and engineering with Amrita Vishwa Vidyapeetham, Coimbatore. His research interests include deep learning in computer vision and deep learning interpretability.



O. K. SIKHA received the B.Tech. degree in information technology from the Calicut University Institute of Engineering and Technology, Calicut, and the M.Tech. degree in computational engineering and networking from the Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore. She is currently an Assistant Professor with the Department of Computer Science, Amrita Vishwa Vidyapeetham. Her current research interests include computer vision for agriculture and healthcare, deep learning interpretability models, and pattern recognition.

RAUL BENITEZ received the B.S. and M.S. degrees in physics from the University of Barcelona and the Ph.D. degree in nonlinear and complex physics from Universitat Politècnica de Catalunya (UPC). He is currently an Associate Professor with the Department of Automatic Control, UPC, leading a research laboratory on biomedical image analysis. His primary research interests include pattern recognition in fluorescence microscopy, biomarker extraction, image segmentation, and deep learning interpretability models.