algorithms

MDPI

Article

Integrating Explanations into CNNs by Adopting Spiking Attention Block for Skin Cancer Detection

Inzamam Mashood Nasir, Sara Tehsin, Robertas Damaševičius * and Rytis Maskeliūnas

Centre of Real Time Computer Systems, Kaunas University of Technology, 51368 Kaunas, Lithuania
* Correspondence: robertas.damasevicius@ktu.lt

Abstract: Lately, there has been a substantial rise in the number of identified individuals with skin cancer, making it the most widespread form of cancer worldwide. Until now, several machine learning methods that utilize skin scans have been directly employed for skin cancer classification, showing encouraging outcomes in terms of enhancing diagnostic precision. In this paper, multimodal Explainable Artificial Intelligence (XAI) is presented that offers explanations that (1) address a gap regarding interpretation by identifying specific dermoscopic features, thereby enabling (2) dermatologists to comprehend them during melanoma diagnosis and allowing for an (3) evaluation of the interaction between clinicians and XAI. The specific goal of this article is to create an XAI system that closely aligns with the perspective of dermatologists when it comes to diagnosing melanoma. By building upon previous research on explainability in dermatology, this work introduces a novel soft attention mechanism, called Convolutional Spiking Attention Module (CSAM), to deep neural architectures, which focuses on enhancing critical elements and reducing noise-inducing features. Two instances of the proposed CSAM were placed inside the proposed Spiking Attention Block (SAB). The InceptionResNetV2, DenseNet201, and Xception architectures with and without the proposed SAB mechanism were compared for skin lesion classification. Pretrained networks with SAB outperform state-of-the-art methods on the HAM10000 dataset. The proposed method used the ISIC-2019 dataset for the crossdataset validation process. The proposed model provides attention regarding cancer pixels without using an external explainer, which proves the importance and significance of the SAB module.

Keywords: skin cancer recognition; spatial attention; channel attention; convolutional spiking attention; AI explainability; explainable artificial intelligence

check for updates

Citation: Nasir, I.M.; Tehsin, S.; Damaševičius, R.; Maskeliūnas, R. Integrating Explanations into CNNs by Adopting Spiking Attention Block for Skin Cancer Detection. *Algorithms* **2024**, *17*, 557. https://doi.org/10.3390/a17120557

Academic Editors: Mario Rosario Guarracino, Laura Antonelli and Pietro Hiram Guzzi

Received: 18 June 2024
Revised: 22 November 2024
Accepted: 2 December 2024
Published: 5 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Skin cancer is a prevalent kind of cancer worldwide, with melanoma being responsible for the highest yearly mortality rate. Melanoma is an extremely dangerous kind of skin cancer that exhibits rapid growth and can metastasize to other organs. The World Health Organization (WHO) has forecasted that there will be around 7990 fatalities caused by melanoma in 2025, along with an estimated 97,610 new cases in the United States [1]. To prevent negative outcomes from occurring, it is necessary to promptly identify and treat the situation.

Artificial intelligence (AI)-driven decision support systems have recently experienced a notable increase in interest across many domains owing to their remarkable capabilities. However, the use of AI in sensitive domains, especially those that impact human decision-making, such as healthcare, has raised ethical concerns about the lack of transparency in AI-driven decisions [2]. In skin image classification, the decision-making process of deep learning models lacks interpretability, thus making the interaction between the AI system and practitioner difficult. Therefore, it is imperative to expand current classification architectures with explainability features to enable more profound relations and offer supplementary diagnostic tools [3]. In conventional Computer-Aided Diagnostic (CAD)

systems for skin cancer prediction, the training data are fed into a learning model, typically consisting of a convolutional neural network (CNN) model with multiple layers. This model predicts if the input image is malignant or non-cancerous [4–6]. A limitation of these CAD systems is their black box nature, which means that the end-user is unaware of the inner workings of the trained model. The end-user receives the outcome with satisfactory accuracy but lacks knowledge of the underlying mechanisms of the CAD system, which determines why the image is classified as malignant rather than non-cancerous. What factors contribute to the accuracy of the trained model in predicting the class of an image? What prevents it from making incorrect predictions? These questions have significantly diminished the end-user's faith in the CAD system. Unlike conventional CAD systems, XAI-based CAD solutions for skin cancer prediction address these inquiries. XAI-based CAD systems not only yield outcomes but also provide justifications for their decision-making process. Figure 1 illustrates a juxtaposition between a conventional CAD system and the XAI-based CAD system for the purpose of predicting skin cancer.

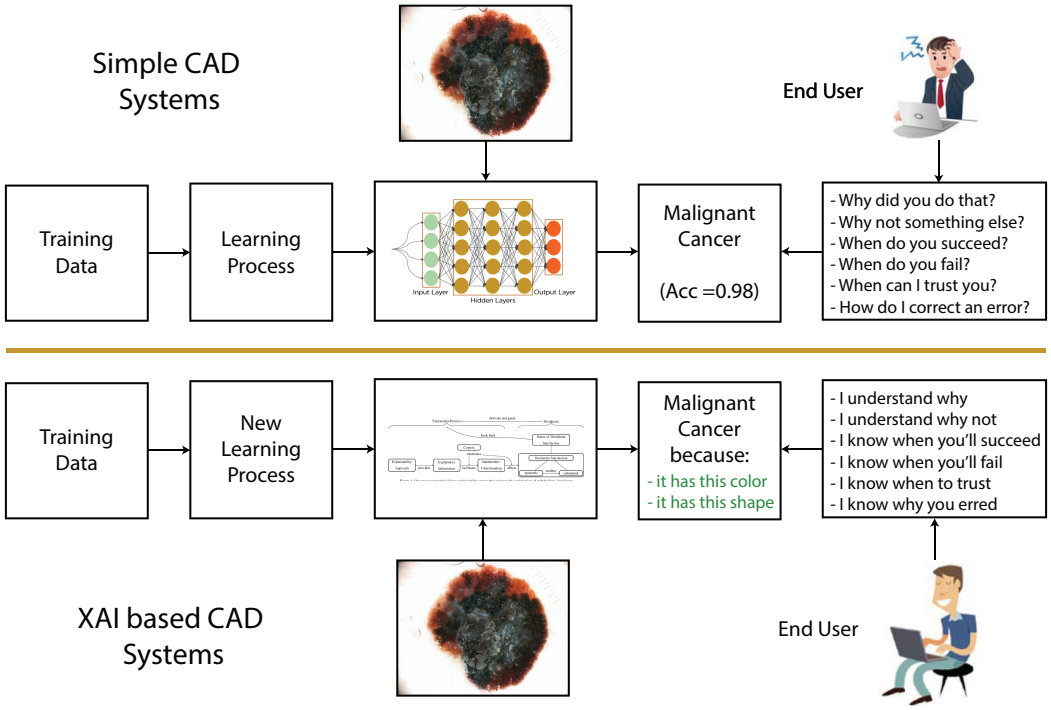

Figure 1. Comparison of traditional CAD systems and XAI-based CAD systems.

A range of XAI methods have been proposed to rectify the deficiencies of CNN models. These methods aim to enhance the transparency of reasoning of AI systems. There are two main branches of XAI techniques: first, post-hoc algorithms, which are specifically developed to provide retrospective explanations for decisions made by a given CNN, such as Grad-CAM [7], Layer-wise Relevance Propagation (LRP) [8], and other similar methods [9]. These techniques are generally used for groups of deep neural networks that have similar components, such as convolutional layers, and are, therefore, mainly independent of the specific model. Depending on the context, post-hoc XAI can be utilized with varying degrees of model parameter access. Some methods are as simple as needing to know the input and output of the model. Inherently interpretable algorithms make up the second group of XAI approaches [10]. These algorithms are specifically designed to be easily comprehensible, which allows them to address issues about the accuracy of explanations that arise in post-hoc XAI [11]. However, there are additional constraints on both the training process and the model construction due to the interconnected nature of these XAI methods. Aside from the integration of explanatory algorithms into a CNN, XAI approaches differ in providing explanations either at a global level, applicable to the entire

dataset, or at a class level, applicable to specific classes. Alternatively, they may provide explanations at a local level, meaning for each individual image.

Moreover, XAI methods vary in their target audience for explanation. Each group has distinct requirements for machine explanations and varying levels of comprehension in the relevant field. To investigate the reliability of automated predictions individually, clinicians require local and highly user-centric explanations from a diagnosis help system. On the other hand, most deep learning software libraries, including post-hoc XAI techniques, do not require expert-annotated data [12], leading to their widespread adoption. Nevertheless, these solutions are frequently dismissed as inadequate for addressing the issue of openness in the medical field since they provide a potential risk of generating confirmation bias during the interpretation of the explanations [13].

Transfer learning and Generative Adversarial Networks (GANs) have been employed to recognize and classify skin cancer [14,15]. Although there has been notable advancement, the classification of skin cancer remains a challenging endeavor. This can be attributed to the scarcity of annotated data and the limited diversity between different classes. Varieties of skin lesion size, color, shape, and contrast, along with additional artifacts, such as hair and veins, further complicate the process.

This article investigates the impact of incorporating a novel soft attention mechanism, called Convolutional Spiking Attention Module (CSAM), into a CNN. A single CSAM improves the performance of pretrained CNN models, but when the proposed CSAM is integrated twice, called Spiking Attention Block (SAB), the performance of pretrained models is boosted further. Deep learning architectures classify images by acquiring knowledge of the prominent characteristics and relationships. SAB enhances performance by directing its attention exclusively towards pertinent regions of the input; it also enhances the transparency of the image classification process for medical staff and effectively identifies and maps the specific sections of the input that the network relies on to classify the image. As a result, this mechanism increases the level of trust in the classification model.

The remainder of the article is structured such that Section 2 provides the literature review, Section 3 provides an in-depth review of the proposed approach, Section 4 provides an overview of the classification and comprehensibility of the proposed method, and Section 5 concludes the article by discussing potential future directions.

2. Literature Review

CNN architectures have demonstrated significant capabilities in analyzing spatio-temporal patterns, including in the domain of skin disease recognition [16,17], thanks to the accumulation of information over time through the collective computations of each convolutional neuron. In [18], the authors employed two multimodal datasets consisting of photos illustrating skin problems and metadata pertaining to clinical trials. The model they implemented consisted of a decoder and two encoders, which were influenced by the transformer architecture. Multiple Vision Transformers (ViTs) were compared to several commonly used CNN designs. Among them, NesT achieved a better accuracy rate of 75% compared to DenseNet121 and ResNet101, which reached 66% accuracy. NesT served as the network's support structure during the analysis of the private dataset. In this situation, the SLE and MA block methods were found to be more appropriate for the metadata than the one hot encoder technique. When the suggested model was compared to the alternative model that accepts input photos, the accuracy increased from 0.75 to 0.816. It is crucial to acknowledge that image processing and segmentation have a significant function in the medical field. The melanoma parameters can be utilized for image segmentation and feature extraction phases to detect melanoma, a kind of skin cancer. The image is classified as either cancer-free or as a melanoma cancer lesion based on the texture, size, and form obtained from feature parameters.

The authors of [19] proposed that computer vision may help diagnose images in healthcare. Computerized diagnostic methods are needed to diagnose melanoma early and solve the problem. Segmenting skin lesions is the initial step. Performing feature and

pattern analysis is crucial to diagnosing the afflicted area. Combining CNN structures with multi-scale data has led to many skin lesion segmentation methods. These systems either require extra labeling or use cumbersome, ineffective components. Researchers combined ViT and ConvNet to create a hybrid segmentation method. This method pinpointed infections and skin lesions with great precision.

To classify skin abnormalities, numerous researchers opt to carry out comprehensive studies on the readily accessible skin cancer dataset known as Human Against Machine (HAM10000) [20]. In [21], a basic transformer framework was used to classify skin cancer using this dataset and a clinical dataset. The suggested model outperformed state-of-the-art techniques on the HAM10000 dataset, achieving 94.3% accuracy. The contrastive learning-based multi-scaled network outperformed CNN models, such as MobileNetV2, ResNet50, and InceptionV2, with 94.1% accuracy.

In [22], the authors created a two-tier approach to handle skin disease classification issues in separate studies. Their MVT model was thoroughly tested on the HAM10000 dataset to classify SC images. The training set had 98% accuracy, and the validation set had 90% accuracy without data preprocessing. The authors trained the same dataset with the GrabCut-stacked CNN (GC-SCNN) model with a fuzzy foundation. SVM and fuzzy GC-SCNN achieved 99.75% classification accuracy in this case.

The authors of [23] asserted that pathologists interpret skin biopsy specimens by using visual judgments, which can frequently be prone to inaccuracies. A Scale-Aware Transformer Network (ScAtNet) was used to categorize melanocytic skin lesions in digital Whole-Slide Images (WSIs) to resolve this issue. To accomplish this, CNN was employed to independently create mappings for each individual data scale in a patch-wise manner. Subsequently, the model utilized a transformer to acquire inter-scale and inner patch representation from the combined multi-scale patch embeddings. The network was evaluated against patch-based classification, ChikonMIL, weighted feature aggregation, a streaming CNN, and MS-DA-MIL. The test set attained an accuracy of 64% across various input scales.

A large, documented dataset, as described in [24], is crucial for deep learning to work in medical data analysis. However, generating extensive interpretations of this magnitude is an ambitious task, particularly for histopathology images that possess unique descriptions. Consequently, the researchers utilized the most extensive publicly available datasets of histo-pathological imaging to create a transformer-based unsupervised feature extractor. The system's CNN-multi-scale Swin Transformer hybrid architecture was its key component.

Numerous scholars have been engaged in developing CAD techniques for classifying skin cancer. Prior to the emergence of deep learning, machine learning (ML) methods were predominantly utilized in CAD systems. ML-based techniques have limitations in identifying skin disorders because of the challenges in constructing features and the restrictions of manually created features. Deep learning algorithms excel in their ability to autonomously acquire semantic features from extensive datasets, resulting in superior accuracy and effectiveness.

3. Proposed Methodology

Since most CNN frameworks fail to account for how various inter-dependencies affect spatial dimensions, the problem is solved by the proposed CSAM model. This block incorporates spatial attention and channel attention components to selectively capture and emphasize meaningful inter-dependencies from spatial channels. Figure 2 presents an overview of the proposed CSAM model; Figure 3 illustrates the internal architectures of spatial attention and channel attention, and Figure 4 outlines the architecture of the proposed SAB, which consists of two CSAM models. The total number of CSAMs embedded in the proposed SAB is selected after extensive experiments.

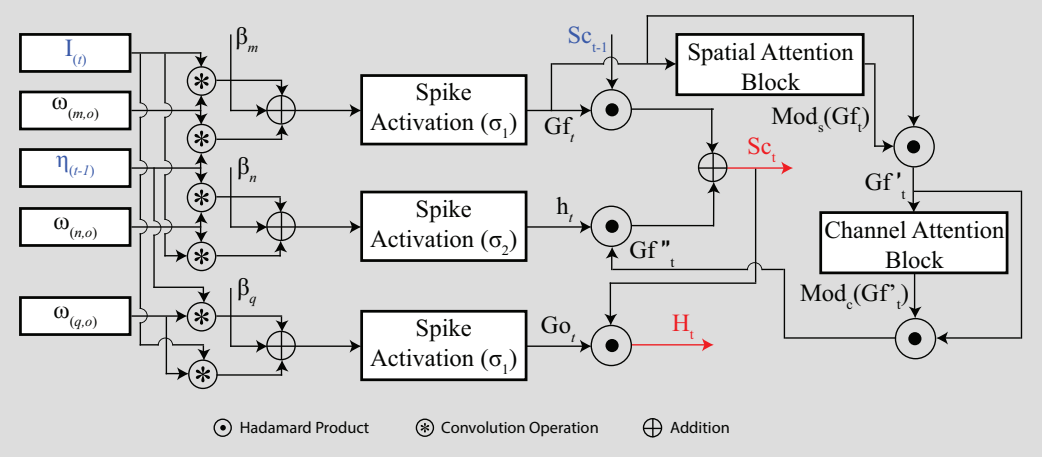

Figure 2. Overall architecture of the proposed CSAM model.

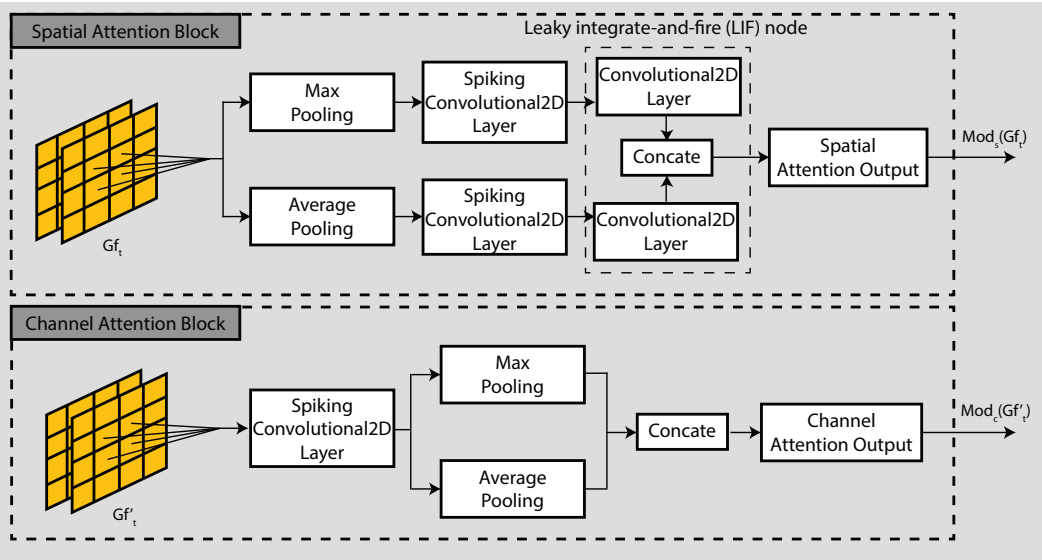

Figure 3. Internal architecture of Spatial Attention Block and Channel Attention Block.

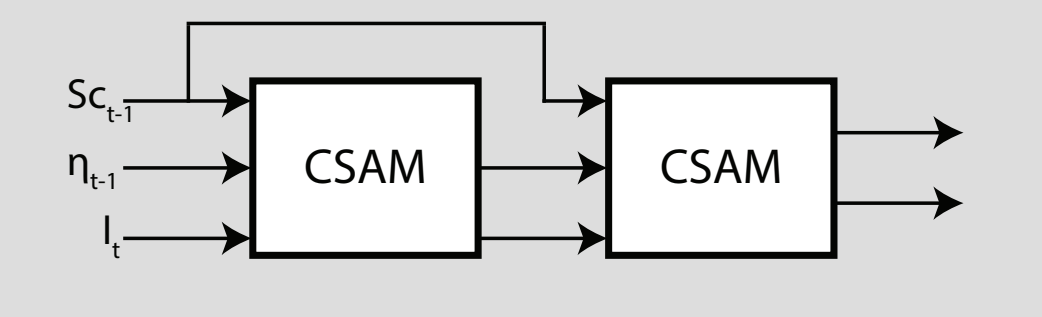

Figure 4. Architecture of the proposed SAB.

The addition of a spiking convolutional layer [25] to CSAM is motivated by its capacity to extract spatial features, enabling precise processing of the input data. Spiking convolution is analogous to CNN convolution and addresses the limitations of the normal convolution layer in handling three-dimensional data. As spiking convolution uses event-based computation that processes only significant changes, this approach naturally incorporates temporal dependencies and efficiently handles sparse data, making spiking convolutions suitable for tasks that require capturing both spatial and temporal patterns. The fundamental principle of spiking convolution, unlike traditional convolution, centers around the gate structure and cell state, c_t, which effectively transmits pertinent informa-

tion throughout the entire sequence-processing procedure. The gate structure acquires the ability to selectively retain or discard information during the training process. For the three-dimensional tensor, the last two dimensions stand for the physical dimensions of weight and height, and the first dimension represents depth, such as the number of channels, and these constitute the input to the spiking convolution. Data are processed iteratively at each step, T, where the convolution operation is substituted as part of the connection procedure to refine features progressively within the network. This convolutional procedure leverages the current input and spatial dependencies from surrounding regions in a structure that facilitates transitions within feature maps, enhancing contextual representation. Although it loses some of the detail in high-resolution images with tiny pixels, striking convolution is better at dealing with dynamic spatial patterns, such as cancerous patches with unique variations in texture, color, and shape within a single image. These spatial distributions represent the irregular patterns or gradients that spiking convolution can more effectively capture due to its sensitivity to spatial dynamics within the image. To be more precise, when provided with a collection of spiking neurons, $I_1, I_2, \ldots, I_t$, the gates and states can be described as follows:

$$Gf_t = \sigma_1((\omega_{(m,o)} * \eta_{(t-1)}) + (\omega_{(m,o)} * I_t) + \beta_m) \tag{1}$$

$$h_t = \sigma_2((\omega_{(n,o)} * \eta_{(t-1)}) + (\omega_{(n,o)} * I_t) + \beta_n) \tag{2}$$

$$Go_t = \sigma_1((\omega_{(q,o)} * \eta_{(t-1)}) + (\omega_{(q,o)} * I_t) + \beta_q) \tag{3}$$

$$Sc_t = (Gf_t \odot Sc_{(t-1)}) + (\mathrm{Gf}_t'' \odot h_t) \tag{4}$$

$$H_t = Go_t \odot Sc_t \tag{5}$$

Here, the symbol t represents the current state, m is the subscript for the forget gate, o is the subscript for the spiking neuron, n is the subscript for the auxiliary input gate, and q is the subscript for the output gate. The symbol $\odot$ denotes the Hadamard product, and $*$ symbolizes the convolution operation; the spike activations σ_1 and σ_2 transform a neuron's activation value into a spike when it exceeds the threshold value, respectively. Additionally, ω and β represent the corresponding weights and biases. For this work, the value of ω is randomly initiated at the start, and β for all inputs is fixed at 1.05 after extensive experiments. For the network, $\eta_{(t-1)}$ represents the preceding hidden state. The forget gate, represented as Gf_t, determines which information should be disregarded. The input gate, after applying spatial attention, Gi_t, is responsible for regulating the flow of information into the cell and includes an additional auxiliary input layer denoted as h_t, which is affected by an additional activation peak σ_2. A spiking convolution's output is defined by its cell state, h_t, and output gate, Go_t.

The use of spiking convolution is made possible by improving the forget gate through the attention mechanism. The proposed method, known as the spiking convolution with the CSAM component, aims to dynamically incorporate previously extracted spatial features to refine the representation. When using spiking convolution, the forget gate is designed by merging the current pulse input with the hidden state of the previous red state, followed by segmentation and convolution. The proposed method derives the one-dimensional channel attention map and the two-dimensional space attention map from the CSAM component's forget gate. Because of this, the forget gate's channel properties and spatial aspects can be extracted. Additionally, spatial adjustments are applied to refine channel features based on their contextual relevance. The comprehensive process of attention can be succinctly summarized as

$$\mathrm{Gf}_t' = \mathrm{Mod}_s(\mathrm{Gf}_t) \odot \mathrm{Gf}_t \tag{6}$$

$$\mathrm{Gf}_t'' = \mathrm{Mod}_c(\mathrm{Gf}_t') \odot \mathrm{Gf}_t' \tag{7}$$

The spatial attention module is referred to as Mod_s, whereas the channel attention module is referred to as Mod_c. Channel features and spatial features are extracted sequen-

tially from the forget gate, resulting in the generation of $Mod_c(\mathrm{Gf}_t)$ and $Mod_s(\mathrm{Gf}'_t)$. Gf''_t represents the output of the spiking CSAM. Figure 2 illustrates the computation method of each attention diagram. The CSAM channel attention component uses max pooling and average pooling to generate a one-dimensional channel attention map of the forget gate. The calculation for the channel attention is as follows:

$$\mathrm{Mod}_c(\mathrm{Gf}'_t) = \mathrm{AvgPooling}(\mathrm{SConv}^{(3\times3)}(\mathrm{Gf}'_t)) + \mathrm{MaxPooling}(\mathrm{SConv}^{(3\times3)}\mathrm{Gf}'_t)) \quad (8)$$

Here, $SConv^{(3\times3)}$ denotes the spiking convolutional operation. The weights of the models in the average and max pooling feature maps are the same. After passing it through the convolution process, the spatial attention part of CSAM pools the results and does average and max pooling operations along the channel axis. The forget gate's two-dimensional spatial attention feature map is created by passing the result through the spike activation function. The calculation of spatial attention is as follows:

$$MP = \mathrm{MaxPooling}(\mathrm{SConv}^{(3\times3)}(\mathrm{Conv}^{(7\times7)}(\mathrm{Gf}_t))) \quad (9)$$

$$AP = \mathrm{AvgPooling}(\mathrm{SConv}^{(3\times3)}(\mathrm{Conv}^{(7\times7)}(\mathrm{Gf}_t))) \quad (10)$$

$$\mathrm{Mod}_s(\mathrm{Gf}'_t) = AP + MP \quad (11)$$

Here, $Conv^{(7\times7)}$ denotes a convolution operation using a filter of size 7×7. As a result of these attention operations, the forget gate's spatial attention map and one-dimensional channel attention map are generated. These maps extract the spatial aspects of the forget gate, as well as its channel features.

The gradient substitution approach employs the Heaviside step function in the forward-propagation process, represented as $j = \theta(i)$. During the process of backpropagation, the substitution function utilized is $\frac{dj}{di} = \sigma(i)'$. Here, $\sigma(i)$ is a function that has a similar form to $\theta(i)$ but is characterized by being smooth and continuous. The gradient of the Gaussian error function (gef) spike function is the main function used in backpropagation. The function that corresponds to this is

$$h(i) = 0.5(1 - \mathrm{gef}(-\tau i)) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\tau i} e^{-t^2}\, dt \quad (12)$$

The parameter τ determines the level of smoothness in the gradient throughout the process of backpropagation. Backpropagation utilizes the gradient descent algorithm based on the derivatives:

$$h'(i) = \frac{\tau}{\sqrt{\pi}} \times e^{-\tau^2 \times i^2} \quad (13)$$

The above formula can be used to approximate the derivatives of the variables' gradients, denoted by τ. In the end, for the seven classes of skin cancer, the neural network is optimized via Categorical Cross-Entropy Loss (CCEL), given as

$$CCEL = -\sum_{x=0}^{l} gt_x \times \log(A(S_x)) \quad (14)$$

$$A(S_x) = \frac{e^{s_x}}{\sum_{y=1}^{l} e^{s_j}} \quad (15)$$

There are seven classes, denoted as $l \in [0, 6]$. Each class x in l has a ground truth gt_x and a CNN score S_x. The softmax activation function $A(S_x)$ is also applied to the scores.

CNN Model Setup

This section provides a comprehensive analysis of the complex structure of all the models. The activation, dropout, flatten, and dense layers are added following the embed-

ding of the SAB with pretrained models to incorporate regularization. The experiments are conducted using two approaches: (a) training and testing on the HAM10000 dataset and (b) training on the HAM10000 dataset and testing on the ISIC2019 dataset. The experiments were performed using the Keras framework.

Three pretrained CNN models, i.e., InceptionResnetV2 [26], DenseNet201 [27], and Xception [28], are used as a backbone model for this work. The architectures of these three selected pretrained CNN models are presented in Figures 5–13, where, for each model, several building blocks are defined; then, a minimized actual version of a pretrained model is presented, and lastly, the proposed architecture of the pretrained model after embedding the SAB module is presented. In Figures 7, 10, and 13, some differences are observed. These changes are intentional and reflect architectural modifications made to improve performance and adapt the network structure to better suit the requirements of the proposed model. As shown in Figure 6, in the original architecture of InceptionResnetV2, IRB8 has 9 IRB_6 blocks and then 1 IRB_1 block, which feeds the output to the GAP block. In the proposed InceptionResnetV2 with SAB, the architecture is modified such that the IRB_8 block has 4 IRB_6 blocks, and then the IRB_7 block is integrated, the output of which for the 3D tensor is forwarded to the SAB block. The three outputs of SAB are then concatenated to be fed to a max pooling block, as shown in Figure 7. Similarly, in Figure 9, the original DensenNet201 architecture has the DNB_4 block before 32 blocks of DNB_3 and then a batch normalization layer. When compared to this, in Figure 10, the proposed DenseNet201 with SAB has 29 DNB_3 blocks after DNB_4 and then a single DNB_2 block is integrated. After that, batch normalization and ReLU operations are performed before embedding the proposed SAB block. In Figure 12, it can be seen that the proposed Xception model has 10 XEB_4 blocks and a single XEB_5, with XEB_6 blocks following them; however, in the proposed Xception with SAB architecture, as shown in Figure 13, there are 6 XEB_4 blocks, followed by a single XEB_5 block, an activation function, and then the SAB block is integrated.

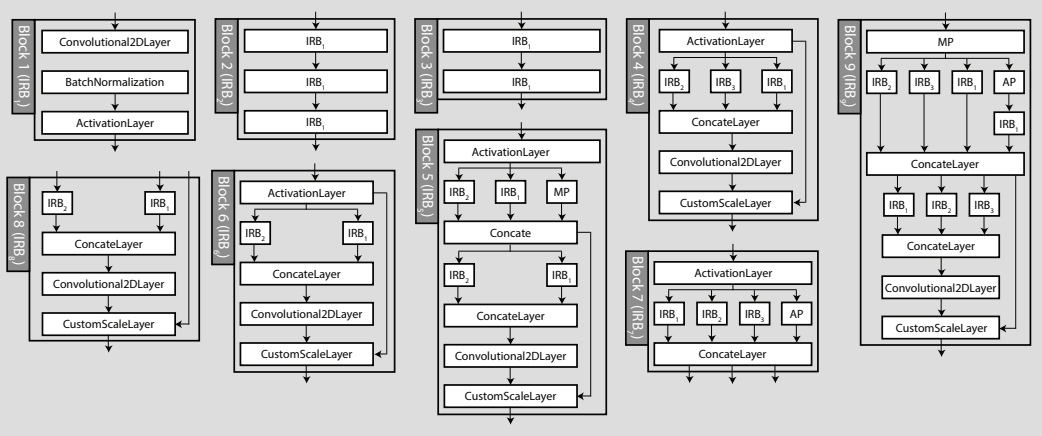

Figure 5. Building blocks of the InceptionResnetV2 model.

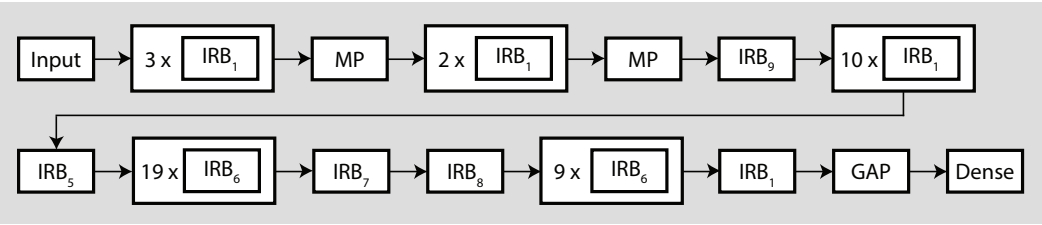

Figure 6. Minimized architecture of the InceptionResnetV2 model.

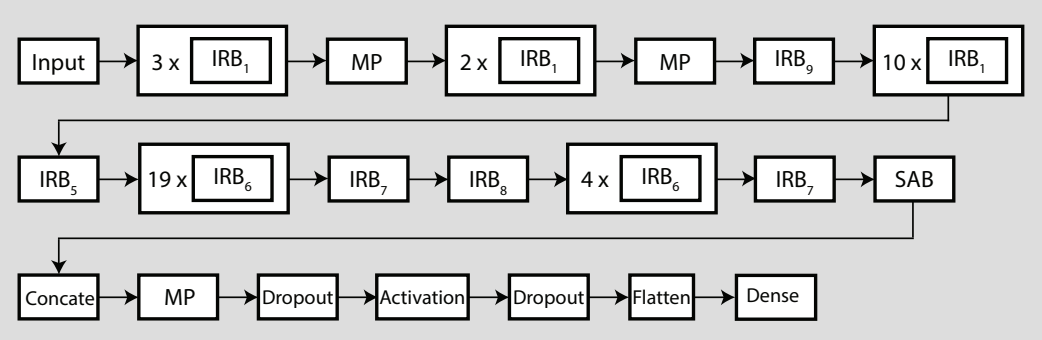

Figure 7. The proposed architecture of the InceptionResnetV2 model after embedding SAB.

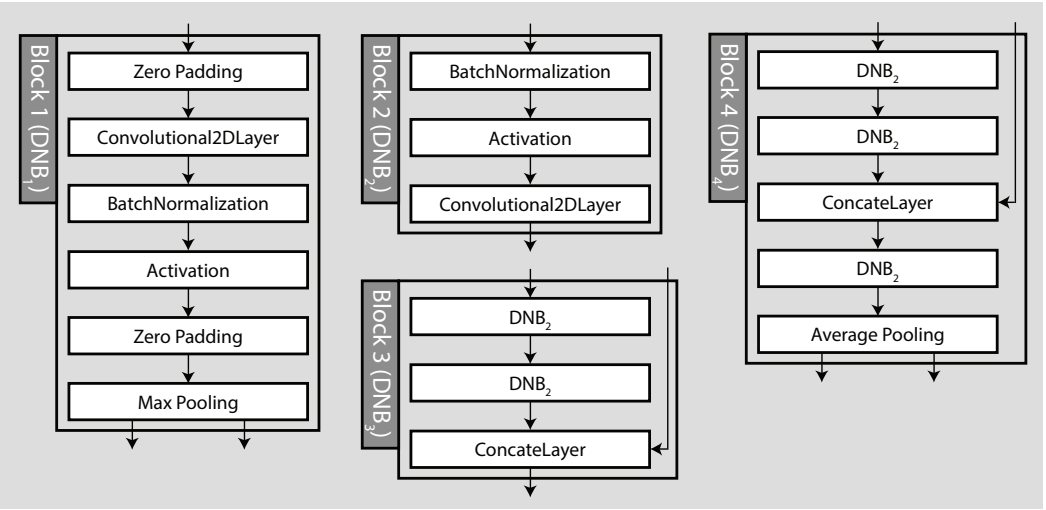

Figure 8. Building blocks of the DenseNet201 model.

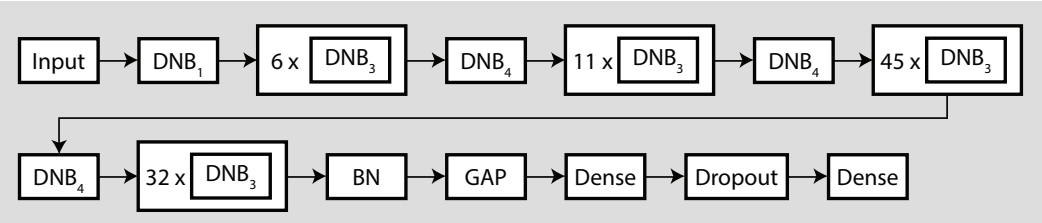

Figure 9. Minimized architecture of the DenseNet201 model.

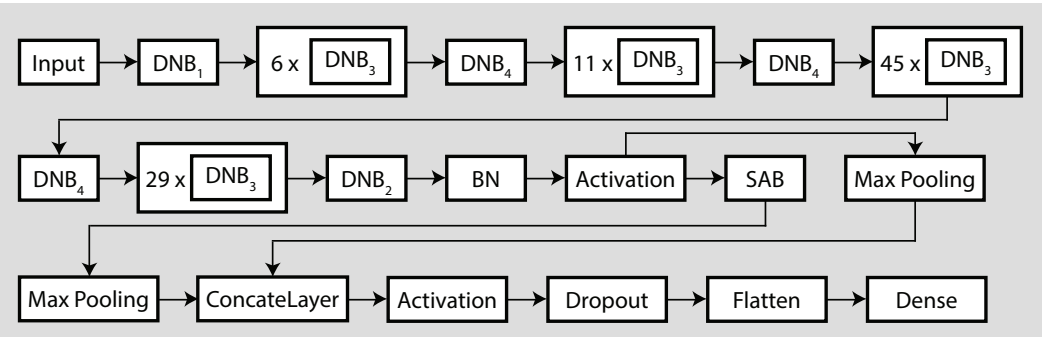

Figure 10. The proposed architecture of the DenseNet201 model after embedding SAB.

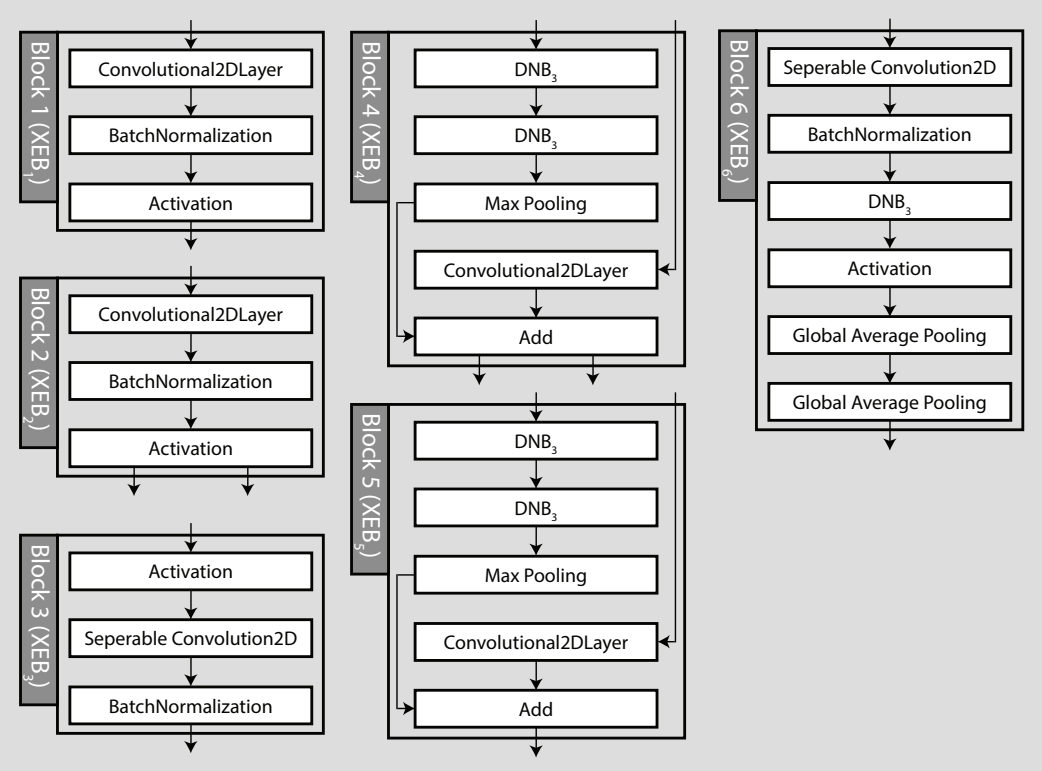

Figure 11. Building blocks of the Xception model.

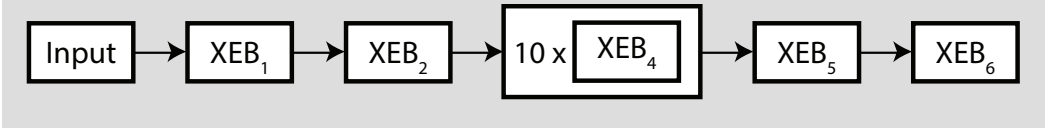

Figure 12. Minimized architecture of the Xception model.

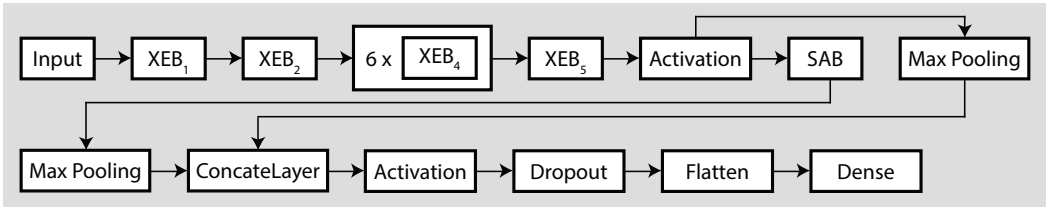

Figure 13. The proposed architecture of the Xception model after embedding SAB.

Algorithm 1 summarizes the method described in the provided text, using a structured and step-by-step approach that can be readily understood and implemented. It encapsulates the CSAM block, the processing steps within the spiking convolution layer, the application of the attention mechanisms, and the final classification step using Categorical Cross-Entropy Loss.

Algorithm 1 Convolutional Spiking Attention Module (CSAM)

Input: Collection of spiking inputs $I_1, I_2, \ldots, I_t$
Output: Processed features with attention mechanisms
Initialize spiking convolution layer with parameters ω, β
for each timestep T **do**
    $Gf_t = \sigma_1(\omega_{(m,o)} * \eta_{(t-1)} + \omega_{(m,o)} * I_t + \beta_m)$
    $h_t = \sigma_2(\omega_{(n,o)} * \eta_{(t-1)} + \omega_{(n,o)} * I_t + \beta_n)$

Algorithm 1 *Cont.*

$Go_t = \sigma_1(\omega_{(q,o)} * \eta_{(t-1)} + \omega_{(q,o)} * I_t + \beta_q)$
$Sc_t = (Gf_t \odot Sc_{(t-1)}) + (Gf_t'' \odot h_t)$
$H_t = Go_t \odot h_t$
end for
▷ Integrate CSAM for attention modulation:
$Gf_t' = \text{Mod}_s(Gf_t) \otimes Gf_t$
$Gf_t'' = \text{Mod}_c(Gf_t') \otimes Gf_t'$
▷ Channel attention (CSAM component):
$\text{Mod}_c(\text{Gf}_t') = \text{AvgPooling}(\text{SConv}^{(3\times3)}(\text{Gf}_t')) + \text{MaxPooling}(\text{SConv}^{(3\times3)}\text{Gf}_t'))$
▷ Spatial attention (CSAM component):
$MP = \text{MaxPooling}(\text{SConv}^{(3\times3)}(\text{Conv}^{(7\times7)}(\text{Gf}_t)))$
$AP = \text{AvgPooling}(\text{SConv}^{(3\times3)}(\text{Conv}^{(7\times7)}(\text{Gf}_t)))$
$\text{Mod}_s(\text{Gf}_t') = AP + MP$
▷ Employ gradient substitution:
$j = \theta(i)$ ▷ Forward-propagation
$\frac{dj}{di} = \sigma'(i)$ ▷ Backpropagation
▷ Calculate Categorical Cross-Entropy Loss (CCEL) for classification:
$CCEL = -\sum_{x=0}^{l} gt_x \times \log(A(S_x))$
$A(S_x) = \frac{e^{S_x}}{\sum_{y=1}^{l} e^{S_j}}$

4. Experimental Results

During this research, two publicly available datasets, HAM10000 and ISIC2019, were used. The proposed model was solely trained on the HAM10000 dataset, and the ISIC2019 dataset was only used to cross-validate the proposed model.

Since the HAM10000 dataset is the benchmark for skin lesion categorization, it was utilized in this study [13]. There is a total of 10,015 RGB thermoscopic images in the dataset, with dimensions of $450 \times 600 \times 3$. These images are divided into seven classes: melanoma (mel), melanocytic nevi (nv), basal cell carcinoma (bcc), benign keratosis (bkl), dermatofibroma (df), vascular lesions (vasc), and actinic keratosis and intraepithelial carcinoma (akiec). The number of images for each class is as follows: 327, 541, 1099, 155, 6705, 1113, and 142. Class imbalance and artifacts (skin hair, dark areas, etc.) are the primary issues with this dataset.

A total of 11 different classes of skin lesions—akiec (867), bcc (3323), bkl (2624), df (239), mel (4522), nv (12,875), SCC (628), and vasc (253)—are represented in the 25,331 dermoscopic images that make up the ISIC2019 dataset.

The proposed model was trained on NVIDIA GeForce RTX 2060. The model was trained using 150 epochs, with a batch size of 16, epsilon = 0.1, and an initial learning rate of 0.01. The Adam optimizer was used, with an initial weight of 1.0 for each class. During the training, training was stopped if, for the 30 epochs, the validation loss did not change from 0.001. During the training process, the weights of the selected pretrained CNN models were learned from scratch.

4.1. Classification Results

Table 1 presents the precision of all the models on the HAM10000 dataset. The abbreviation "+SAB" in this table represents models that include SAB. InRe represents InceptionResNetV2, Xe is short for Xception, and De represents DenseNet201. By examining the table, it is evident that when InRe is combined with SAB, there are notable enhancements in the results. The precision of this combination is 94%, which is also the highest score compared to all other models. In addition, it is evident that SAB enhances the precision of the InRe model by 5% compared to the baseline InRe model. This trend is also applicable to the Xe and De models as well. For example, SAB enhances the accuracy of DenseNet201, ResNet34, ResNet50, and VGG16 by 0.5%, 0.8%, 1.2%, and 2%, respectively. When the

SA block is incorporated into the networks, we observe a comparable pattern in the AUC scores. Specifically, the performance of Xe increased by 5%, and the performance of De increased by 9%.

Table 1. Comparison of selected pretrained CNN models with and without SAB on HAM10000.

Class	Precision					
	InRe	**InRe + SAB**	**Xe**	**Xe + SAB**	**De**	**De + SAB**
akiec	0.94	1.00	0.62	0.75	0.63	0.78
bcc	0.87	0.95	0.59	0.70	0.69	0.81
bkl	0.86	0.92	0.53	0.65	0.58	0.70
df	0.93	1.00	0.72	0.89	0.73	0.86
nv	0.78	0.89	0.51	0.64	0.49	0.60
mel	0.90	0.97	0.70	0.74	0.80	0.91
vasc	0.96	1.00	0.88	0.93	0.81	0.90
Macro average	0.81	0.89	0.65	0.75	0.67	0.79
Weighted average	0.89	0.96	0.71	0.76	0.73	0.82

While InRe + SAB achieves the highest weighted average performance, a detailed analysis reveals that SAB significantly improves the efficiency of the original InRe algorithm for classifying the akiec, bcc, bkl, df, nv and mel classes. Specifically, SAB increases precision by 6% for akiec, 8% for bcc, 6% for bkl, 7% for df, 11% for nv, 7% for mel, and 4% for vasc.

From Figure 14, it is evident that the inclusion of SAB generally enhances the precision of the models across all classes, with several intriguing observations. Notably, the Xception (Xe) model shows substantial improvements when augmented with SAB, particularly for the dermatofibroma (df) and actinic keratosis and intraepithelial carcinoma (akiec) classes, suggesting that this model benefits significantly from attention mechanisms in handling the complex textural and morphological features of these skin conditions. InceptionResNetV2 (InRe) and DenseNet201 (De) also display notable precision gains with SAB, especially for melanoma (mel) and vascular lesions (vasc), indicating their effectiveness in leveraging temporal-spatial information for these highly varied and nuanced classifications. The uniformly positive increments across models and classes, especially in challenging classes such as melanoma and basal cell carcinoma, underscore the robustness and utility of adding SAB to these architectures, potentially due to better feature isolation and prioritization during model training.

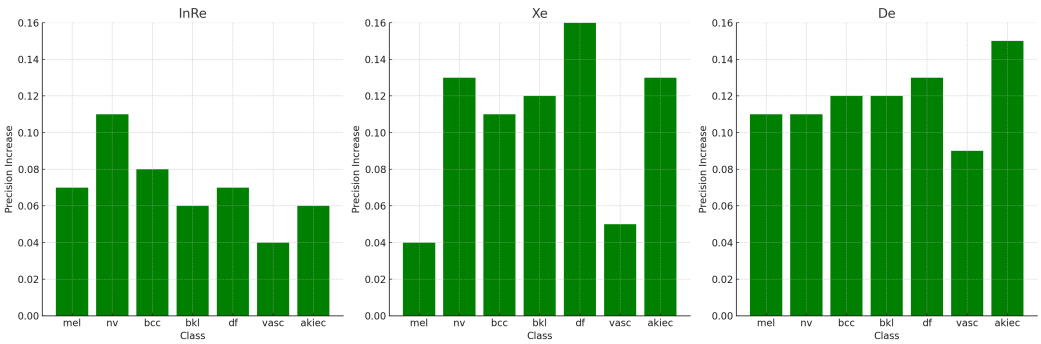

Figure 14. Performance increase using models that include the soft attention block.

The performance of InRe + SAB as an explainer is also illustrated in Figure 15, where the proposed SAB has clearly enabled the CNN model to provide accurate attentions for its decisions.

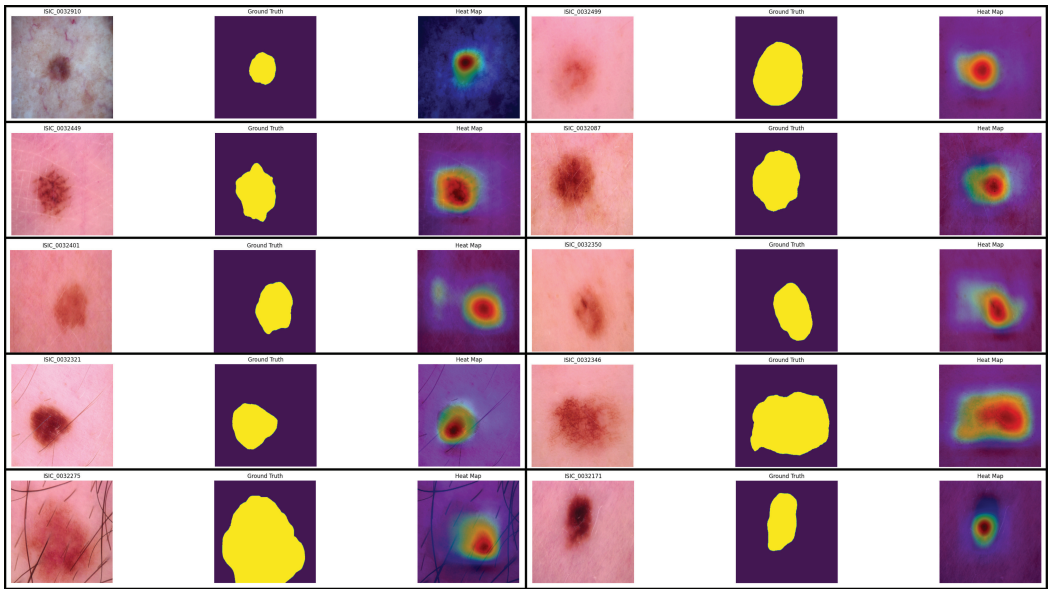

Figure 15. Attention maps generated using InRe+SAB on the HAM10000 dataset.

4.2. Ablation Analysis

When the model was evaluated using various train–test splits on the HAM10000 dataset using the InRe + SAB model, the model trained using 85% of the data outperformed the models developed using 80% and 70% of the data, as indicated in Table 2. Therefore, we picked the $85 - 15$ dataset for the purpose of the training–validation split to conduct the experiments.

Table 2. Comparison of different split ratios using the InRe + SAB model on the HAM10000 dataset.

Class	Split = 15		Split = 20		Split = 30	
	AUC	**Precision**	**AUC**	**Precision**	**AUC**	**Precision**
akiec	1.0	1.00	0.97	0.95	0.98	0.89
bcc	0.99	0.95	0.92	0.83	0.99	0.78
bkl	0.94	0.92	0.96	0.68	0.96	0.62
df	0.96	1.00	0.98	0.94	0.93	0.89
nv	0.92	0.89	0.91	0.59	0.94	0.53
mel	1.0	0.97	0.99	0.88	0.97	0.81
vasc	1.0	1.00	1.0	0.93	0.98	0.87
Macro Average	0.97	0.89	0.96	0.82	0.96	0.77
Weighted Average	0.98	0.96	0.96	0.92	0.93	0.91

The scatter plots in Figure 16 reveal a general trend of higher precision associated with higher AUC values, indicating robust model performance in certain scenarios. For instance, the 'akiec' class consistently shows high precision and AUC, particularly in the Split = 15 scenario, reflecting the model's strong capability to accurately classify actinic keratosis and intra-epithelial carcinoma with minimal false positives. However, there are classes and splits where the precision significantly drops while maintaining a relatively high AUC, such as the 'nv' (melanocytic nevi) in Split = 30, suggesting a possible overfitting or imbalance in the model's ability to generalize this class under varied data conditions. The variability in performance across different splits and classes underlines the importance of tailored model training and validation strategies, especially for medical diagnostic applications where precision is critical. The visual markers enhance interpretability by

distinctively showcasing the performance metrics of each class, enabling rapid assessment and comparison across different experimental conditions.

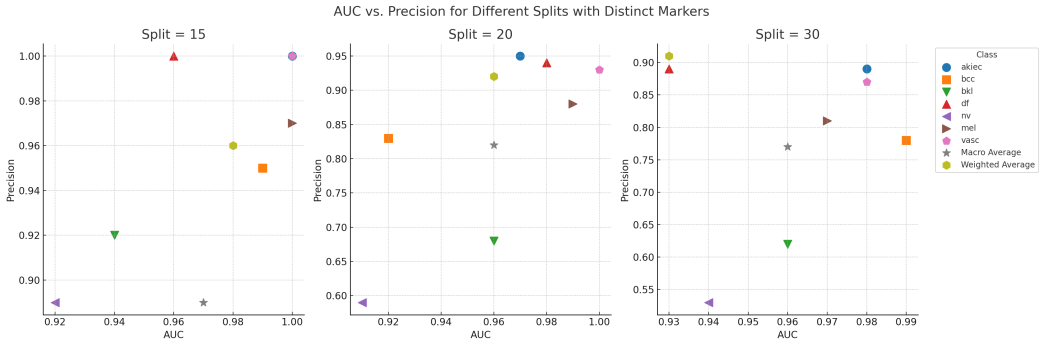

Figure 16. Comparison of class performance using different split ratios.

In addition, the proposed approach is compared to the state-of-the-art models for skin cancer classification on the HAM10000 dataset, as shown in Table 3. In [29], the authors introduced an innovative Multi-weighted New Loss technique to address class imbalance and enhance the accuracy of detecting critical classes, such as melanoma. They suggested an innovative end-to-end cumulative learning technique that more successfully balances representation learning and classifier learning without incurring extra processing costs. The suggested technique achieved 92% accuracy and 94% AUC on the HAM10000 dataset. In [30], the authors introduced a high-performance data augmentation technique for a search space, which can be integrated with any model via a plug-and-play approach to identify the optimal augmentation method for a medical database with minimal resource expenditure. The integration with EfficientNets resulted in an accuracy of 86% and an AUC of 97% on the HAM10000 dataset.

Table 3. Comparison of the proposed model to state-of-the-art techniques.

Model	Accuracy	Precision	AUC
Ensemble-based loss balancing [29]	0.92	-	0.94
Deep learning model [30]	0.86	-	0.97
Augmentation based classification [31]	0.85	-	0.97
Two-way CNN model [32]	0.88	-	-
Pretrained CNN models [33]	-	0.89	0.97
Proposed Model	**0.97**	**0.96**	**0.98**

The authors of [31] introduced an automated model for skin lesion categorization with dermoscopic pictures, employing convolutional neural networks (CNNs) and including augmentation methods such as cut-out, random shear, random scaling, random rotation, random flipping, contrast adjustments, and random brightness modifications. The suggested model attained 85% accuracy and 97% AUC. In [32], the efficacy and proficiency of convolutional neural networks have been examined in the categorization of eight dermatological conditions. Various pretrained state-of-the-art architectures (DenseNet 201, ResNet 152, Inception v3, and InceptionResNet v2) were used to construct a bidirectional CNN model, attaining 88% accuracy on the HAM10000 dataset. The authors of [33] tackled skin lesion classification with an ensemble of deep learning models, including EfficientNets, SENet, and ResNeXt WSL, achieving an accuracy of 89% and an AUC of 97% on the HAM10000 dataset.

The proposed CSAM approach has superior performance compared to the standard method, with a precision increase of 4.7%. The soft attention-based technique demonstrates superior performance compared to all other approaches, with AUC scores surpassing them by a margin ranging from 0.5% to 4.3%. As shown in Figure 17, the performance of the proposed model is evaluated by utilizing the crossdataset validation technique. This

technique involves training the model on the HAM10000 dataset but then providing it with inputs from the ISIC-2019 dataset to determine how well it performs. Based on the data presented in Figure 17, it is evident that the model demonstrates consistent performance across both datasets and offers effective explanations.

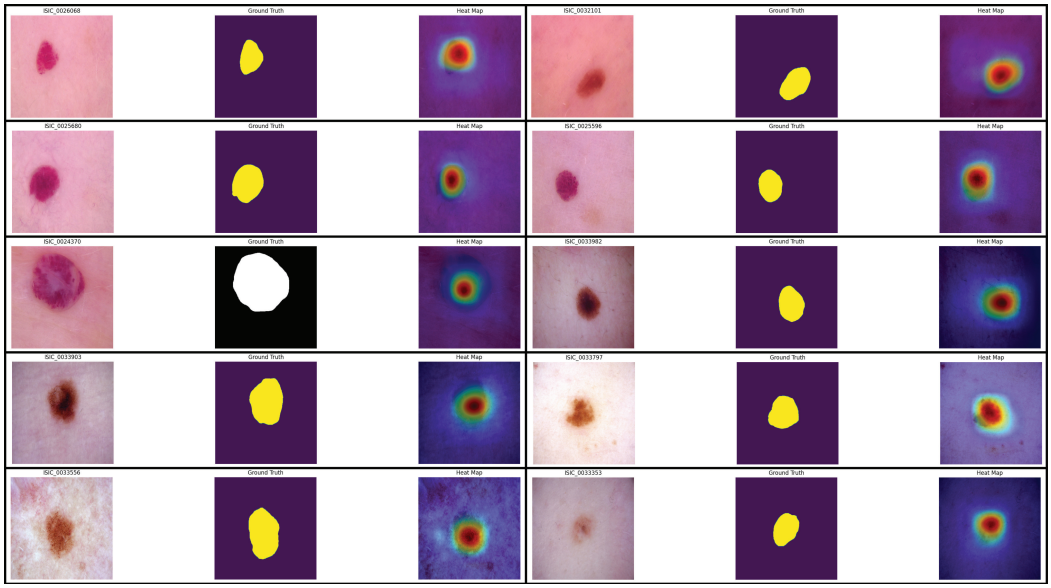

Figure 17. Attention maps generated using InRe + SAB when trained on the HAM10000 dataset and validated on the ISIC-2019 dataset.

5. Conclusions

To classify high-resolution skin cancer images, this article proposes the use of the attention module. On the HAM10000 dataset, the model achieved better results than the state-of-the-art methods now. Another technique used to assess the efficacy of the proposed model is crossdataset validation, specifically with the ISIC-2019 dataset. These findings demonstrate the promise and utility of the deep learning architecture, which utilizes soft attention, in the field of image analysis. In addition, the soft attention method reduces the need to use external explainers such as LIME, SHAP, and GradCAM. Furthermore, it provides the precise place where the model concentrates its attention when classifying a disease, hence improving the overall performance of the primary model. Soft attention has the extra benefit of effectively managing visual noise internally. This model has the potential to be included in future dermoscopy devices, providing valuable support to dermatologists implementing this methodology to classify data from alternative medical approaches.

Author Contributions: Conceptualization, I.M.N., S.T., R.D. and R.M.; Methodology, I.M.N.; Software, I.M.N. and S.T.; Validation, I.M.N., S.T., R.D. and R.M.; Formal Analysis, R.D. and R.M.; Investigation, I.M.N. and S.T.; Resources, I.M.N., S.T., R.D. and R.M.; Data Curation, I.M.N., S.T., R.D. and R.M.; Writing—Original Draft Preparation, I.M.N. and S.T.; Writing—Review and Editing, R.D. and R.M.; Visualization, I.M.N. and S.T.; Supervision, R.D. and R.M.; Project Administration, I.M.N. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Two publicly available datasets were used in this research: HAM10000 and ISIC2019 [13].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghebrial, M.; Wang, Q.; Zhang, R.; Robson, P.J.; Shack, L.; Kopciuk, K.A. Predictors of primary cutaneous melanoma stage at diagnosis: Observations from Alberta's Tomorrow Project. *Ann. Cancer Epidemiol.* **2024**, *8*, 1. [CrossRef]
2. Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F. Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
3. Markus, A.F., Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [CrossRef] [PubMed]
4. Maqsood, S.; Damaševičius, R.; Shahid, S.; Forkert, N. D. MOX-NET: Multi-stage deep hybrid feature fusion and selection framework for monkeypox classification. *Expert Syst. Appl.* **2024**, *255*, 124584. [CrossRef]
5. Bibi, S.; Khan, M.A.; Shah, J.H.; Damaševičius, R.; Alasiry, A.; Marzougui, M.; Alhaisoni, M.; Masood, A. MSRNet: Multiclass Skin Lesion Recognition Using Additional Residual Block Based Fine-Tuned Deep Models Information Fusion and Best Feature Selection. *Diagnostics* **2023**, *13*, 3063. [CrossRef] [PubMed]
6. Hussain, M.; Khan, M.A.; Damaševičius, R.; Alasiry, A.; Marzougui, M.; Alhaisoni, M.; Masood, A. SkinNet-INIO: Multiclass Skin Lesion Localization and Classification Using Fusion-Assisted Deep Neural Networks and Improved Nature-Inspired Optimization Algorithm. *Diagnostics* **2023**, *13*, 2869. [CrossRef]
7. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
8. Hägele, M., Seegerer, P.; Lapuschkin, S.; Bockmayr, M.; Samek, W.; Klauschen, F.; Müller, K.R.; Binde, A. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci. Rep.* **2020**, *10*, 6423. [CrossRef]
9. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA, 27–30 June 2016.
10. Barata, C.; Celebi, M.E.; Marques, J.S. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognit.* **2021**, *110*, 107413. [CrossRef]
11. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
12. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S., Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for pytorch. *arXiv* **2020**, arXiv:2009.07896.
13. Ghassemi, M., L. Oakden-Rayner, and A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef]
14. La Salvia, M.; Torti, E.; Leon, R.; Fabelo, H.; Ortega, S.; Martinez-Vega, B.; Callico, G.M.; Leporati, F. Deep Convolutional Generative Adversarial Networks to Enhance Artificial Intelligence in Healthcare: A Skin Cancer Application. *Sensors* **2022**, *22*, 6145. [CrossRef] [PubMed]
15. Al-Rasheed, A.; Ksibi, A.; Ayadi, M.; Alzahrani, A.I.A.; Zakariah, M.; Ali Hakami, N. An Ensemble of Transfer Learning Models for the Prediction of Skin Cancers with Conditional Generative Adversarial Networks. *Diagnostics* **2022**, *12*, 3145. [CrossRef] [PubMed]
16. Nivedha, S.; Shankar, S., Melanoma Diagnosis Using Enhanced Faster Region Convolutional Neural Networks Optimized by Artificial Gorilla Troops Algorithm. *Inf. Technol. Control.* **2023**, *52*, 819–832. [CrossRef]
17. Ren, G. Monkeypox Disease Detection with Pretrained Deep Learning Models. *Inf. Technol. Control.* **2023**, *52*, 288–296. [CrossRef]
18. Xin, C.; Liu, Z.; Zhao, K.; Miao, L.; Ma, Y.; Zhu, X.; Zhou, Q.; Wang, S.; Li, L.; Yang, F.; Xu, S. An improved transformer network for skin cancer classification. *Comput. Biol. Med.* **2022**, *149*, 105939. [CrossRef]
19. Gulzar; Y.; Khan, S.A. Skin lesion segmentation based on vision transformers and convolutional neural networks—A comparative study. *Appl. Sci.* **2022**, *12*, 5990. [CrossRef]
20. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 1–9. [CrossRef]
21. Naeem, M.A.; Yang, S.; Sharif, A.; Saleem, M.A.; Sharif, M.I. Vision Transformer for Skin Cancer Identification Based on Contrastive Learning and Adaptive-Scale Fragmentation. *Preprint* 2024, *in press*. [CrossRef]
22. Behara, K., E. Bhero, and J.T. Agee, An Improved Skin Lesion Classification Using a Hybrid Approach with Active Contour Snake Model and Lightweight Attention-Guided Capsule Networks. *Diagnostics* **2024**, *14*, 636. [CrossRef]
23. Dong; H; Farid, F.A. Deep learning based patient care application for skin cancer detection. *Int. Res. J. Adv. Eng. Hub* **2022**. [CrossRef]
24. Thomsen, K.; Christensen, A.L.; Iversen, L.; Lomholt, H.B.; Winther, O. Deep learning for diagnostic binary classification of multiple-lesion skin diseases. *Front. Med.* **2020**, *7*, 574329. [CrossRef] [PubMed]
25. Cao, Y.; Chen, Y.; Khosla, D. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comput. Vis.* **2015**, *113*, 54–66. [CrossRef]
26. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

27. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
28. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
29. Gessert, N.; Nielsen, M.; Shaikh, M.; Werner, R.; Schlaefer, A. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX* **2020**. *7*, 100864. [CrossRef]
30. Yao, P.; Shen, S.; Xu, M.; Liu, P.; Zhang, F.; Xing, J.; Shao, P.; Kaffenberger, B.; Xu, R.X. Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE Trans. Med. Imaging* **2021**, *41*, 1242–1254. [CrossRef] [PubMed]
31. Shen, S., Xu, M.; Zhang, F.; Shao, P.; Liu, H.; Xu, L.; Zhang, C.; Liu, P., Yao, P.; Xu, R.X. A low-cost high-performance data augmentation for deep learning-based skin lesion classification. *BME Front.* **2022**, *2022*, 9765307. [CrossRef]
32. Charan, D.S.; Nadipineni, H.; Sahayam, S.; Jayaraman, U. Method to classify skin lesions using dermoscopic images. *arXiv* **2020**, arXiv:2008.09418.
33. Rezvantalab, A.; Safigholi, H.; Karimijeshni, S. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *arXiv* **2018**, arXiv:1810.10348.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.