Efficient Combination of Deep Learning Models for Skin Disease Detection

Mohamed Massamba Sene(✉), Ndeye Fatou Ngom, and Michel Seck

LTISI Laboratory, Ecole Polytechnique de Thiès, Thies, Senegal
smmassamba@ept.sn

Abstract. Identifying skin diseases is challenging due to their similar visual appearance, making it difficult to select features. Despite significant progress in the effective identification of skin diseases, the problem remains unsolved. This paper presents a new method for accurate skin disease detection, which combines the Attention U-Net architecture for image segmentation and a customized Convolutional Neural Network (CNN) for image classification. The first step of the proposed approach is training the segmentation model on a dataset of segmented skin disease images. The resulting model is then used to segment images from a skin disease images classification dataset before training the customized CNN using the segmented images to classify skin diseases. With the proposed approach, we were able to achieve an accuracy of 99% as well as high precision, recall and F1 score of 99% on the HAM10000 dataset. Comparative analysis with other similar studies demonstrates its effectiveness in accurately identifying these diseases.

Keywords: Skin diseases identification · Image segmentation · Deep learning · Image classification · Computer vision

1 Introduction

The skin is an essential organ of the human body, protecting us from harmful radiation, injuries and infections caused by UV rays. The prevalence of skin cancer is increasing worldwide and represents a significant health risk. Skin diseases are generally detected by time-consuming procedures such as skin biopsy, clinical screening, dermoscopic analysis and histopathological diagnosis [1]. Melanoma is the most dangerous type of skin cancer and is responsible for the majority of skin cancer deaths. However, there are other diseases that are visually similar to malignant melanoma, such as blue nevus and spitz nevus. Dermoscopic images are essential for identifying skin cancer, but visual inspection can be challenging. The use of computerized automatic medical image segmentation technology will have a significant impact on the problem of diagnostic errors by clinicians as well as providing strong support for automatic classification [1]. In recent years, CNNs have shown promising results when it comes to computer vision tasks like

K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 287–298, 2025.
https://doi.org/10.1007/978-3-031-86493-3_23

pattern localization or features learning. In addition, semantic segmentation is proving to be well suited to biomedical image segmentation, allowing a more flexible and comprehensive approach to capturing the diverse structures within biomedical images.

This study aims to propose a new approach that improves the accuracy of skin disease identification using deep learning based techniques. The proposed workflow starts with segmentation using Attention U-Net to identify the region of interest, followed by the use of a customized CNN to identify skin disease. The key contributions of this paper are:

- A new skin disease detection model that combines the Attention U-Net and a customized Convolutional Neural Network.
- By using the Attention U-Net architecture instead of classical methods such as region growing [15] or other descriptors for segmentation, we enhance the model's ability to capture context, resulting in more accurate segmentation.
- By using a customized Convolutional Neural Network with optimized hyperparameters, we directly extract features and identify the skin disease without manual feature extraction.

With the proposed approach, we were able to achieve an accuracy of 99% on the HAM10000 dataset with high precision, recall and F1 score of 99%. Comparative analysis with other similar studies demonstrates its effectiveness in accurately identifying skin diseases.

The rest of this paper is organized as follows. Section 2 discusses previous work on skin disease classification. Section 3 discusses the Attention U-Net Architecture. In Sect. 4, we describe the design of the proposed approach, as well as the datasets and techniques used for segmentation and classification of the skin disease images. Section 5 examines the process used to build the models and discusses the results of the experiment. Finally we conclude our work in Sect. 6.

2 Related Work

The development of artificial intelligence technology has opened up many possibilities in the field of automated disease detection from image analysis.

Shetty et al. [12] applied machine learning and deep learning techniques using CNNs to classify the skin lesion images on the HAM10000 dataset, achieving an accuracy of 95.1% using a customized CNN. After training different state-of-the-art models (MobileNet, VGG19, ResNet50, InceptionV3) using transfer learning, Valasco et al. [2] found that MobileNet had the highest accuracy (94.1%) while VGG16 had the lowest accuracy (44.1%). In [11], Ali et al. trained the EfficientNets B0-B7 on the HAM10000 dataset by performing transfer learning on pre-trained weights from ImageNet and fine-tuning the Convolutional Neural Networks. Wei et al. [3] proposed an approach based on the combination of DenseNet and ConvNet. Compared to other state-of-the-art models, the proposed model achieved good results due to its accuracy and F1 score of 95.29%

and 89.99% respectively. However, it is computationally expensive, which limited it to performing classification on only three skin diseases.

In [13], Oktay et al. proposed a skin disease detection approach using pre-trained Convolutional Neural Network features before multi-class SVM classification. Reddy et al. [1] proposed an approach that utilizes optimized region-growing-based segmentation and an autoencoder-based classification to identify skin diseases. The accuracy of the proposed system was 94.2%. However it was limited to detecting the presence or absence of skin disease and not identification on a wide range of diseases. Soh et al. [9] proposed a hybrid U-Net transformer to improve the efficiency of segmentation of single-modality lesion and multi-modality brain tumour images. Recently, the connection between the U-Net architecture and the attention mechanism has been the focus of many authors. The Attention U-Net architecture is based on the U-Net architecture which boasts an expansive symmetric path that effectively captures spatial context, enabling accurate edge detection and detailed segmentation [9]. By introducing the attention mechanism, it allows the network to focus on relevant image features and suppress irrelevant noise, resulting in cleaner and more accurate segmentation [13]. Hasrh et al. [18] use Attention U-Net for dental segmentation to further increase sensitivity and prediction accuracy with minimal computational overhead.

Based on these previous works, we propose a novel approach aimed at improving the accuracy of detection.

3 Attention U-Net Architecture

In this work, unlike in [1], we don't use optimized region growing segmentation. Instead, we use Attention U-Net segmentation. U-Net is a Fully Convolutional Neural Network designed for accurate and efficient biomedical image segmentation. Its name comes from its U-shaped architecture, which consists of two main pathways: a Contracting path (Encoder/Left side) and an Expanding path (Decoder/ Right side).

Contracting path: the upper part of the "U" is used to extract higher level features from the image through a gradual down-sampling of the image [6]. To achieve this, similar to a classical CNN architecture, we repeated the block of two 3×3 convolutions with step 1 and ReLu activation, separated by a Droupout layer with rate 0.1 and followed by a max pooling layer. We start with 32 filters for the first block, and in each step we double the number of filters we use. For the last block, which acts as a bottleneck, we use 5 encoder blocks with pooling removed.

Expanding path: the lower path of the "U" is used to restore the spatial resolution of the image while performing semantic segmentation. It uses transposed convolutions to increase the resolution of the image and up-sample the feature maps learned in the contracted path [6].

A key feature of the U-Net architecture is the use of skip connections. This bridges the gap between high-level and low-level features, improving segmentation accuracy with the advantage of preserving a large amount of spatial information. However, it also results in poor feature representation from the initial layers. In the Attention U-Net, soft attention is implemented at the skip connections with attention gates to actively suppress activation at irrelevant regions, as shown in Fig. 1. Two inputs are taken, a gating signal from the lower layer with better feature representation and the skip connection with better spatial information. The two inputs are passed through convolution layers, with the skip connection being down-sampled to allow it to be added. Aligned weights become larger while unaligned weights become smaller. The result is then passed to a ReLu activation before convolution with a filter that allows the weights to be retrieved. The resulting array is then up-sampled to the skip connection size before element by element multiplication [13].

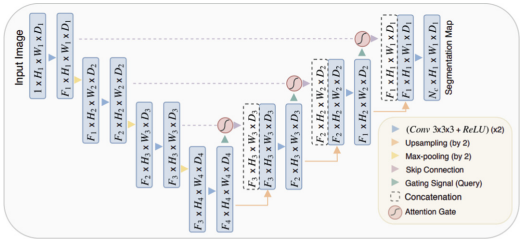

Fig. 1. Attention U-Net architecture as seen in [13]

4 Methodology

4.1 System Design

The proposed workflow consists of a combination of segmentation using the Attention U-Net and classification using a CNN, as shown in Fig. 2. Unlike in [1], we don't use optimized region growing segmentation. Instead, we use Attention U-Net segmentation. This is a crucial shift from traditional region growing segmentation, which is sensitive to noise and relies heavily on setting thresholds and seed points [15], which can be tedious. By replacing region growing with Attention U-Net, the proposed workflow enjoys significant advantages in terms of accuracy, robustness and automation [9].

We also remove manual feature extraction and rely on feature extraction provided by the CNN, as it is able to learn features from raw data in a hierarchical manner. In fact, manual feature definition requires domain expertise and meticulous effort, often becoming a bottleneck in the development process and struggling to adapt to new data or variations in existing data, hindering the generalizability of the model [14]. By doing so, we are able to reduce development time and effort and minimize the risk of human bias.

Finally, as seen in [3], we exploit the power of fusion models by combining the Attention U-Net and the CNN model. This improves the performance of the proposed model by exploiting the strengths of both architectures.

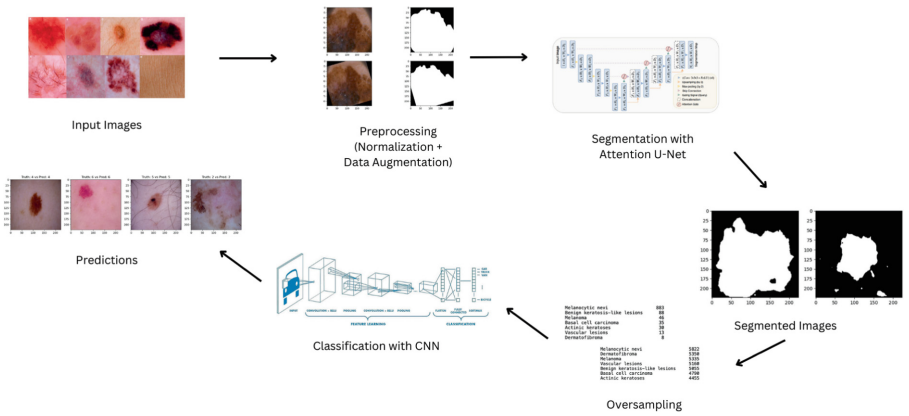

Fig. 2. Proposed skin disease classification model

4.2 Datasets

Two datasets are used in our study: the PH^2 dataset [4] for segmentation and the HAM10000 (Human-Against-Machine with 10000 training images) dataset for classification [5].

The PH^2 dataset was created for research and comparative analysis purposes and consists of a collection of 200 images acquired at the Dermatology Service of the Pedro Hispano Hospital (Matosinhos, Portugal). They are 8-bit RGB color images together with their segmentation masks at 768×560 pixels resolution [4].

The HAM10000 dataset is a collection of 10015 dermoscopic images from different populations, acquired and stored using different modalities. The cases include a representative collection of all major diagnostic categories of pigmented lesions: Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevi (NV) and Vascular Lesions (VASC) [5].

4.3 Data Preprocessing

Image normalization was applied to both datasets and the images were resized to a resolution of 224×224 pixels using OpenCV's resize function with the INTER_AREA strategy. This strategy prioritizes maintaining smooth transitions and minimizing aliasing effects, making it a preferred choice for downsampling images in applications where visual quality is important [7].

To increase the size of the PH^2 dataset to 600 images, we applied data augmentation by flipping and rotating the 200 images. Data augmentation helps to learn more robust features and reduces the risk of over-fitting by exposing our model to a wider range of data variations. This leads to better performance on unseen data, resulting in more accurate and generalizable models.

Class imbalance was present in the HAM10000 dataset, with the Melanocytic Nevi class represented in more than 70% of our images. We apply over-sampling random over-sampler, while adjusting the factor based on the number of images for each class. This simply duplicates the minority class by the specified factor in the training dataset to increase its representation, artificially balances the class distribution and gives the model more opportunities to learn from examples of the minority class.

4.4 Adapted Attention U-Net Architecture

We implement the encoder blocks of the attention U-Net architecture as seen in [13] with four repeated combinations of attention gates followed by decoder blocks. The attention gates consist of three convolution layers followed by an up-sampling layer and batch normalisation. For the first two layers we use a 3×3 convolution with ReLu activation. The first layer uses stride 1 to keep the dimension of the gating signal and the second uses stride 2 to reduce the dimension of the skip connection. The last layer uses 1×1 convolution with sigmoid activation. The decoder blocks use an up-sampling layer followed by two 3×3 convolutions with stride 1 and ReLu activation. We start with 256 filters and at each step we divide the number of filters by 2.

4.5 Convolutional Neural Network

Convolutional Neural Networks (CNNs), also known as ConvNets, are a class of a class of deep learning models specifically designed to solve computer vision problems such as image classification, object detection, and image segmentation. We propose a CNN model that is built from scratch and modified by fine-tuning the parameters to find the best classifier for skin disease classification. The architecture of the proposed CNN can be seen in Fig. 3. The network takes as input images of size 224×224 pixels and provides as output a probability distribution for the seven output classes. The preprocessing block implements the transformations described in Subsect. 4.3, such as normalization, data augmentation, and class balancing.

The rest of our CNN consists of two parts: the convolution module, which transforms the input images into feature vectors, and the classification module, which determines which class the input image belongs to. It is composed of

– **Convolution Layers**: serve as the backbone of a CNN. Filters also called kernels are used to perform operations on the input images. Each filter goes through the image and computes the weighted sum of the pixel values. The results are then stored in feature maps. Filters are then moved along the

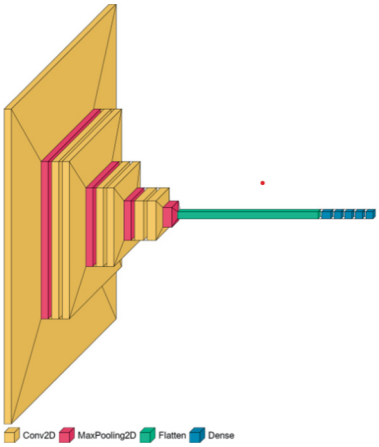

Fig. 3. CNN model architecture

images storing important features like edges, textures, shapes. The size of the Kernel used here is 3×3 with He-normal as kernel initializer.

- **Activation Layers**: after each convolution, an activation function is used on the resulting feature map. This allows the introduction of non-linearity allowing the capture of more complex relations between features [8]. The ReLu activation function was used for our Convolution layers.
- **Pooling Layers**: are used to reduce the dimensions of the feature maps while preserving the important information. Pooling reduces the model complexity and makes features insusceptible to small changes in position on the image [8]. The MaxPooling2D layer was used in our implementation. We use three blocks of double convolution followed by pooling.
- **Fully Connected Layers**: after repeated convolution and pooling, they are used to perform classification. These layers are similar to those of traditional Neural Networks and are used to combine the extracted feature maps to perform classification [8].

The features extracted from the convolutional module are passed on to the classification module, which consists of 5 Fully Connected Layers using Dense Layers. The number of neurons in the first and second layers is 64. The third and fourth layers have 16 neurons. The output layer consists of 7 neurons, corresponding to the number of output classes in the network. The ReLu activation is used for each of the Dense Layers, except for the last layer where a softmax activation function is used to provide a score for the 7 classes.

4.6 Evaluation Metrics

The metrics used in this study to evaluate the performance of our model are accuracy, precision, recall and F1 score.

Accuracy is simply the ratio of the number of correctly predicted observations to the total number of observations. It is calculated using

$$\text{Accuracy} = \frac{TP}{TP + FP + TN + FP}$$

where

- TP (True Positives) are correctly predicted positives, the actual class values are yes and prediction class values are also yes.
- TN (True Negatives) represent the correctly predicted negative values. This means that the value of the actual class is no and the value of the predicted class is also no.
- FP (False Positives) represent when the actual class is no and the predicted class is yes.
- FN (False Negatives) denotes when the actual class is yes but the predicted class is no [3].

Recall is the proportion of true positives that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision is the ratio of the correctly predicted positive observations to the total number of positive observations predicted.

$$\text{Precision} = \frac{TP}{TP + FP}$$

The F1 score takes into account both precision and recall.

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We also use IoU (Intersect over Union) to evaluate our segmentation model, which allows us to compare our segmented masks with the ground truth masks. It calculates the overlap between the predicted and ground truth regions by calculating the ratio of the intersection area to the union area [19].

5 Experimental Results

The experiments were performed on the two datasets described in 4.2. The metrics described in the Subsect. 4.6 were used for performance evaluation. The effectiveness of the proposed methodology is compared with previous studies based on deep learning for skin disease classification.

5.1 Experimentation

The experiments were carried out on a 14" Macbook Pro M2 with an Apple M2 Pro chip with a combined 10-core CPU and 16-core GPU. The code was written using Tensorflow, OpenCV and NumPy libraries with Python 3.8 as the programming language.

We first load and preprocess the images from the PH^2 dataset using the steps described in Subsect. 4.3 such as normalization, data augmentation and class-balancing. We then build and train the segmentation model using the Attention U-Net architecture. Training lasted for 45 epochs, and the results shown in Fig. 4, Fig. 5 and Fig. 6 were obtained after training and evaluation.

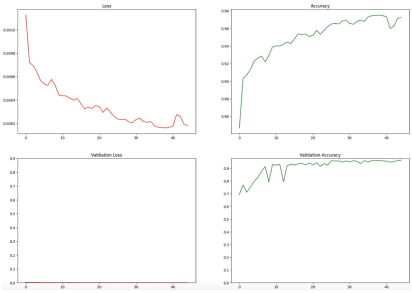

Fig. 4. Loss and Accuracy on the train set with the Attention U-Net model

IOU:	82.61	
Precision:	95.79	
Recall:	90.19	
Accuracy:	94.95	
Loss:	3.64	

Fig. 5. Metrics on the test set with the Attention U-Net model

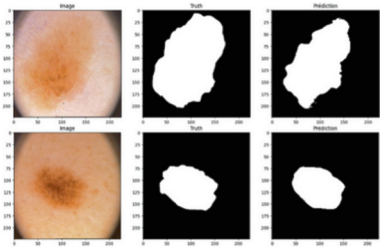

Fig. 6. Intermediate results obtained with the Attention U-Net model

The images from the HAM10000 dataset were then segmented using the trained model and used as input to train our CNN model for classification. To optimize the hyperparameters, we configure a search space with a define-by-run syntax and then use RandomSearch to find the best values for our models. After training and evaluation, the results obtained are shown in the Fig. 7 and Fig. 8 which provides a better understanding of the metrics of our model.

The proposed model has a higher accuracy compared to the model [3] which uses a transfer learning based approach. Experimental results also show that the proposed model achieves better performance than previous CNN based model [11], GAN based model [10] and models based on deep transfer learning with sparrow search algorithm [17]. It also shows similar results to the model based on deep ensemble learning [16]. The models were all tested on the HAM10000 dataset and the results can be seen in Table 1.

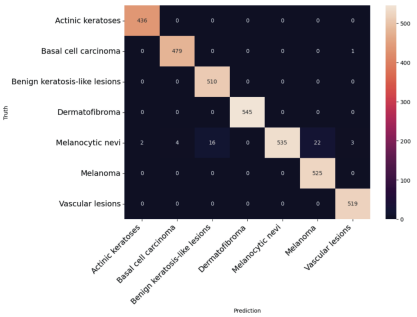

Fig. 7. Confusion matrix of the proposed model

	precision	recall	f1-score	support
0	1.00	1.00	1.00	436
1	0.99	1.00	0.99	480
2	0.97	1.00	0.98	510
3	1.00	1.00	1.00	545
4	1.00	0.92	0.96	582
5	0.96	1.00	0.98	525
6	0.99	1.00	1.00	519
accuracy			0.99	3597
macro avg	0.99	0.99	0.99	3597
weighted avg	0.99	0.99	0.99	3597

Fig. 8. Classification report of the proposed model

Table 1. Comparative performance analysis of the proposed model to other state-of-the-art models

Model	Accuracy	Precision	Recall	F1-score
Our model	0.99	0.987	0.988	0.987
model 1 [3]	0.9529	0.8835	0.9258	0.8999
model 2 [10]	0.9193	0.5050	0.9104	0.6497
model 3 [11]	0.879	0.88	0.88	0.87
model 4 [12]	0.9518	0.88	0.85	0.86
model 5 [16]	0.910	0.9938	0.9927	0.9932
model 6 [17]	0.9883	0.9883	0.9883	0.9883

5.2 Discussion

The proposed approach combines the strengths of Attention U-Net segmentation and CNNs to improve the accuracy of skin disease classification. U-Net architectures excel in semantic segmentation, making them well suited to medical image analysis. By adding the attention mechanism to the architecture, it

leads to more accurate localization by focusing on the relevant regions. It can effectively distinguish between different classes of tissues or lesions, providing detailed segmentation maps. The end-to-end learning capability of CNNs allows the model to understand complex relationships within the data, enabling better representation of relevant features for classification. The combination of these two powerful architectures enhances the model's ability to accurately locate and classify lesions, ultimately contributing to more effective and reliable diagnostic support. Although similar architecture were previously explored in [1,3], our work highlights the importance of model choice when combining different deep learning models.

Although our proposed model approach achieved good classification performance on the HAM10000 dataset, which presents an extreme imbalance, it was not perfect and still had limitations. Therefore, for future work, we recommend applying a lightweight transformation to the proposed model to achieve better performance. We would also recommend trying out other segmentation models such as U-Net3+, V-Net, E-Net to verify their impact on the performance of the model.

6 Conclusion

In this paper, we propose an approach for skin disease classification based on a combination of segmentation using Attention U-Net and classification using a customized CNN. The classification performance of the proposed model was further improved by a series of preprocessing transformations such as data augmentation, oversampling and hyperparameters optimization. On the public dataset HAM10000, the accuracy and F1 scores of the proposed model were 99% and 98.7%, respectively. These results were also good compared to the other state-of-the-art models.

References

1. Reddy, D.A., Roy, S., Kumar, S., Tripathi, R.: A scheme for effective skin disease detection using optimized region growing segmentation and autoencoder based classification. Procedia Comput. Sci. **218**, 274–282 (2023). https://doi.org/10.1016/j.procs.2023.01.009
2. Velasco, J.S., Catipon, J.V., Monilar, E.G., Amon, V.M., Virrey, G.C., Tolentino, L.K.S.: Classification of skin disease using transfer learning in convolutional neural networks. arXiv (2023). https://doi.org/10.48550/ARXIV.2304.02852
3. Wei, M., et al.: A skin disease classification model based on DenseNet and ConvNeXt fusion. Electronics **12**(2), 438 (2023). https://doi.org/10.3390/electronics12020438
4. Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J.: PH² - a dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE (2013). https://doi.org/10.1109/embc.2013.6610779

5. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. arXiv (2018). https://doi.org/10.48550/ARXIV.1803.10417
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation (Version 1). arXiv (2015). https://doi.org/10.48550/ARXIV.1505.04597
7. Team, L.: Image Resizing with OpenCV | LearnOpenCV. LearnOpenCV - Learn OpenCV, PyTorch, Keras, Tensorflow With Examples and Tutorials (2023). https://learnopencv.com/image-resizing-with-opencv/
8. Vancappel, K.: Deep Learning: le Réseau neuronal convolutif (CNN). Business & Decision (2023). https://fr.blog.businessdecision.com/tutoriel-deep-learning-le-reseau-neuronal-convolutif-cnn/
9. Soh, W.K., Yuen, H.Y., Rajapakse, J.C.: HUT: hybrid UNet transformer for brain lesion and tumour segmentation. Heliyon **9**(12), e22412 (2023). https://doi.org/10.1016/j.heliyon.2023.e22412
10. Gu, Y., Ge, Z., Bonnington, C.P., Zhou, J.: Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. IEEE J. Biomed. Health Inform. **24**(5), 1379–1393. https://doi.org/10.1109/jbhi.2019.2942429
11. Ali, K., Shaikh, Z.A., Khan, A.A., Laghari, A.A.: Multiclass skin cancer classification using EfficientNets - a first step towards preventing skin cancer. Neurosci. Inform. **2**(4), 100034 (2022). https://doi.org/10.1016/j.neuri.2021.100034
12. Shetty, B., Fernandes, R., Rodrigues, A.P., Chengoden, R., Bhattacharya, S., Lakshmanna, K.: Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci. Rep. **12**(1) (2022). https://doi.org/10.1038/s41598-022-22644-9
13. Oktay, O., et al.: Attention U-net: learning where to look for the pancreas (Version 3). arXiv (2018). https://doi.org/10.48550/ARXIV.1804.03999
14. Workgroup, T.G.M., et al.: SpecBit, DecayBit and PrecisionBit: GAMBIT modules for computing mass spectra, particle decay rates and precision observables. arXiv (2017). https://doi.org/10.48550/ARXIV.1705.07936
15. Iznita Izhar, L., Petrou, M.: Thermal imaging in medicine. Adv. Imaging Electron Phys. 41–114 (2012). https://doi.org/10.1016/b978-0-12-394297-5.00002-7
16. Shehzad, K., et al.: A deep-ensemble-learning-based approach for skin cancer diagnosis. Electronics **12**(6), 1342 (2023). https://doi.org/10.3390/electronics12061342
17. Balaha, H.M., Hassan, A.E.-S.: Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm. Neural Comput. Appl. **35**(1), 815–853 (2022). https://doi.org/10.1007/s00521-022-07762-9
18. Lin, Z., Tsui, P.-H., Zeng, Y., Bin, G., Wu, S., Zhou, Z.: CLA-U-Net: convolutional Long-short-term-memory attention-gated U-Net for automatic segmentation of the left ventricle in 2-D echocardiograms. In: 2022 IEEE International Ultrasonics Symposium (IUS). IEEE (2022). https://doi.org/10.1109/ius54386.2022.9958784
19. Shah, D.: Intersection over Union (IoU): Definition, Calculation, Code. V7 (2024). https://www.v7labs.com/blog/intersection-over-union-guide