BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 3
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0021

Early Skin Cancer Detection Using the SLICE-3D Dataset: A Transfer Learning Model and DCGAN Approach to Address Data Imbalance

Youssera Z. Mecifi[1], *Mohamed Merzoug*[1], *Abdelhak Etchiali*[1], *Mohamed M'hamedi*[1,2], *Fethallah Hadjila*[1], *Amina Bekkouche*[1]

[1]*LRIT Laboratory, Department of Computer Science, Faculty of Science, University of Abou-Bakr Belkaid, Tlemcen, Algeria*
[2]*Ecole Supérieure en Sciences Appliquées de Tlemcen, BP 165 RP Bel Horizon, Tlemcen, Algeria*
E-mails: youssera zoukha.mecifi@univ-tlemcen.dz (*corresponding author*)
mohammed.merzoug@univ-tlemcen.dz abdelhak.etchiali@univ-tlemcen.dz
mohamed.mhamedi@univ-tlemcen.dz fethallah.hadjila@univ-tlemcen.dz
amina.bekkouche@univ-tlemcen.dz

***Abstract*:** *Early detection of skin cancer is crucial for improving patient outcomes, as the disease progresses rapidly when left untreated. Recent advancements in artificial intelligence have revolutionized the field of early detection, giving clinicians more accurate and efficient diagnostic tools. In this paper, two convolutional neural network-based classifiers using transfer learning are proposed to improve early skin cancer detection. These models were trained and tested on the novel ISIC-2024 dataset. To mitigate the class imbalance in this Dataset, a Generative Adversarial Network* (*DCGAN*) *is adopted to synthesize malignant samples. Additionally, the pre-trained VGG-16 and MobileNetV2 models were fine-tuned to improve feature learning and classification performance. Our MobileNetV2-based model outperformed the VGG16-based model, achieving an accuracy of 96.87%, a precision of 98.97%, and a recall of 94.7%. These results highlight the impact of deep learning in early skin cancer detection, and most importantly, they lead to better patient outcomes.*

***Keywords*:** *Skin cancer, ISIC-2024, DCGAN, VGG16, MobileNetV2.*

1. Introduction

Skin cancer remains a significant public health concern; the number of new skin malignant tumors diagnosed is projected to increase by 5.9% in 2025, and skin cancer fatalities will increase by 1.7% [1]. However, correct and early detection can greatly reduce mortality, as most skin lesions are extremely treatable if identified in their early stage.

Skin cancer is a runaway growth of abnormal cells in the epidermis, the outer layer of the skin. The abnormal growth results from unrepaired DNA damage,

leading to genetic mutations that make the cells grow and divide rapidly and create tumors. Basal cell carcinoma, squamous cell carcinoma, melanoma, and Merkel cell carcinoma are the primary types of skin cancer [2]. This type of cancer looks different depending on the type of skin cancer a person might have, but keeping in mind the ABCDE rule: Asymmetry – irregular shape; Border – jagged or fuzzy edges; Color – mole with more than one color; Diameter – more than a pencil eraser (6 mm); Evolution – increasing, shape, color or size change (this is the most important indicator) [3], any person can check for some concerning lesions just to be safe, and if any lesion looks concerning, it is important to visit a specialized dermatologist, where after a thorough visual examination, he will decide whether a biopsy is necessary or not. In a biopsy, a sample of tissue is removed and sent to a laboratory where a pathologist examines it under a microscope, and then he will be able to conclude: if the skin lesion is cancerous, what type of cancer you have, and will list treatment options for you [4]. Unfortunately, misdiagnosis of skin cancer remains a significant concern in medical practice. Studies have shown that a notable percentage of skin lesions are incorrectly identified, leading to delayed treatment and potentially poorer patient outcomes. For instance, research indicates that approximately 21.9% of cases clinically identified as seborrheic keratosis were later found to be misdiagnosed, with 5.7% of these cases being skin cancers such as basal cell carcinoma, squamous cell carcinoma, or melanoma [5]. Additionally, there have been instances where skin cancer was mistakenly identified as less severe conditions like eczema, underscoring the critical need for accurate diagnosis [6].

On the other hand, AI has introduced some life-changing technologies that made the doctors work easier and patient lives safer, and that is by integrating deep learning models in their diagnosis process, particularly Convolutional Neural Networks (CNNs), which have shown significant success in automating skin lesion classification. However, for AI to be effectively incorporated into clinical workflows, it has to be trained on vast, diverse, and correctly labeled datasets; this helps to guarantee that it is accurate for multiple skin types, imaging conditions, as well as cancer types.

One of the most significant skin disease problems in AI is the lack of skin cancer datasets publicly available for model training. Unlike general medical images, skin lesion datasets are often small in sample size, low in diversity, and poor in annotation quality. Some of the most commonly used datasets include: HAM10000 [7], PH2 [8], PAD-UFES-20 [9].

Compared to existing datasets, our study leverages the novel SLICE-3D dataset from the ISIC-2024 challenge, which uniquely incorporates 3D volumetric imaging, offering enhanced lesion depth and texture analysis. This new dataset was released by the International Skin Imaging Collaboration (ISIC) but remains relatively unvisited in the literature. Other works employing the SLICE-3D dataset have employed different methodologies from our own.

For instance, P i n t e l a s et al. [10] utilized the dataset without any cleaning or data augmentation but maximized model performance via an ensemble of MobileNet specialists; however, their accuracy was not as impressive as ours

because our study focused not only on fine-tuning the model but also enhanced the dataset distribution and quality as well. Similarly, Syed and Albalawi [11] used a CNN and trained it on 10000 3D-TBP images of SLICE-3D with geometric augmentations (translation, rotation, zooming). However, their results were not as favorable as ours.

Another study by Teymouri et al. [12] where the authors uses several models, such as the Data-efficient Image Transformer 3 and Light Gradient Boosting Machine to improve performance, in addition to an ensemble approach. Using a thorough preprocessing approach, they addressed class imbalance and yielded promising results. In our work, different methodological approaches were adopted, which further refined the performance on the same dataset. To summarize, our key contributions are as follows:

- We proposed 2 enhanced architectures based on pretrained models, which are VGG-16 and MobileNetV2, and they achieved state-of-the-art results.
- We generated realistic, diverse, and clear synthetic images of malignant lesions using a DCGAN.
- We handled this novel dataset imbalance through advanced data augmentation and thorough data preprocessing.

We will be presenting a review of related works in the rest of this article in Section 2. After that, in Section 3, we will present our methodology by first describing the SLICE 3D dataset and then its preprocessing journey. Furthermore, we will address data augmentation, classification models, training, and model evaluation. Finally, in Section 4, we will provide the empirical findings of the proposed model, its evaluation, and a comparative analysis before concluding this article in Section 6.

2. Related works

This section discusses recent relevant studies on skin cancer detection using deep neural networks, with only three specifically applying the same dataset, ISIC-2024. For instance, Aljohani and Turki [13] investigated the performance of 8 deep learning models, such as VGG16, ResNet50, and GoogleNet, for melanoma classification on the ISIC 2019 dataset in their work; The results indicated that GoogleNet had the highest test accuracy of 76.08% and outperformed dermatologists in sensitivity, with results of 84.5% compared to 73.3%. Djaroudib et al. [14] analyzed the impact of image quality rather than quantity in melanoma classification using transfer learning with VGG-16 on the HAM10000 dataset. The outcomes revealed that a model learned from fewer images of high quality yielded higher accuracy (94%) compared to large datasets. Sikandar et al. [15] presented SCDet, a 32-layer CNN for efficient skin lesion detection with reduced computational cost compared to pre-trained models like ResNet50; trained on DermIS data, SCDet achieved 99.6% accuracy, and external validation on HAM10000 reached 85% accuracy. The proposed model is a reasonable compromise between effectiveness and efficiency, showing the promise of simplified CNN architectures in medical image analysis. Yashwant, Ingle and

Shaikh [16] compared the classification of skin cancer using a custom CNN versus VGG16 with transfer learning on the HAM10000 dataset; the results indicated that VGG16, augmented by the inclusion of more dense layers and dropout regularization, performed better than the CNN model with 89% accuracy and 89% weighted F1-score. The use of pre-trained architectures to improve classification over custom CNNs is emphasized in this study. In a related work, K et al. [17] proposed a deep learning approach to skin lesion classification using a sequential CNN architecture trained on a large Kaggle dataset. The model worked nicely, achieving 95% accuracy and an F1-score of 96.3%, making it efficient in distinguishing between seven types of skin cancer. Their research indicated the reliability of CNNs in computer-aided cancer detection in medical images. In a parallel study, M'hamedi et al. [18] improved melanoma classification using data augmentation to address class imbalance in the SIIM-ISIC 2020 dataset. With fine-tuning VGG-19 and MobileNetV2, their approach achieved a 95.16% accuracy, where MobileNetV2 performed best; the results show the impact of transfer learning and augmentation in enhancing diagnostic accuracy. Gouda et al. [19] enhanced the diagnosis of skin cancer by coupling ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) for preprocessing with the assessment of different deep learning models using the ISIC-2018 dataset. The best performer was InceptionV3 at 85.8% accuracy, demonstrating the utility of super-resolution techniques to improve classification accuracy. Their contribution shows the importance of data preprocessing in deep learning diagnosis. Mehr and Ameri [20] merged lesion images and patient metadata (gender, age, anatomical location) using Inception-ResNet-v2 to improve classification performance on ISIC2019, PAD-UFES-20, and Fitzpatrick17k datasets. The model achieved 94.5% accuracy for benign versus malignant lesion discrimination, demonstrating the influence of patient data on diagnostic performance. Alrabai, Echtioui and Kallel [21] compared the performance of InceptionV3 and Xception for the detection of skin cancer through transfer learning using a Kaggle dataset. InceptionV3 performed better than Xception, with 89.1% accuracy and better classification performance in comparison with state-of-the-art approaches. Islam and Panta [22] employed 5 pre-trained transfer learning models for skin cancer binary classification on the Kaggle ISIC dataset (3297 images). The best accuracy (93.5%) with an F1-score of 0.86 was achieved by ResNet-50, demonstrating the benefit of fine-tuning activation functions and layers. Pope et al. [23] investigated tone bias in skin cancer diagnosis on an imbalanced (3623 images) and a balanced dataset (~500 images). Their system consistently favored lighter skin tones, with selection rates of 27.5% vs 15.9% (imbalanced) and 50.0% vs 34.2% (balanced), confirming significant bias below the 0.80 fairness threshold. Imran, Alghamdi and Alkatheiri [24] improved skin cancer classification with EfficientNetB0 and Ant Colony Optimization (ACO) as feature selection. Their CB-SVM model reached more than 98% accuracy on a self-created ISIC dataset. Naeem et al. [25] introduced SNC_Net, a fusion of handcrafted and deep learning-based features for classification. It outperformed baseline models in terms of accuracy, precision, and F1-score with 97.81%, 98.31%, and 98.10% on ISIC 2019, respectively. Sedigh,

Sadeghian and Masouleh [26] employed a CNN to detect cancer from the skin and used a GAN to generate synthetic images to augment the dataset. Their small ISIC dataset of 97 images was improved from 53% to 71% accuracy after synthesizing additional samples. Furthermore, Teodoro et al. [27] also presented EfficientAttentionNet for skin cancer classification, integrating a GAN to balance data and a U-Net to be employed for RoI-based attention. It was trained on ISIC 2020, HAM10000, BCN20000, and MKS datasets with an accuracy of 97.9%, a recall of 99.5%, and a precision of 94.5%.

The most pertinent comparison should be made with works that use the same novel dataset, even though some of the previously mentioned studies have reported slightly better results than our models. Of these, Teymouri et al. [12] proposed a sophisticated two-step method for classifying skin lesions through the use of the SLICE-3D dataset. To improve overall diagnostic performance, the authors use an ensemble framework that combines the Light Gradient Boosting Machine (LGBM) and the Data-efficient Image Transformer 3 (DeIT3). Also, they used stratified sampling, the Synthetic Minority Over-sampling Technique (SMOTE), and data augmentation techniques like random Gaussian blur, rotation, and flipping as part of a multi-phase preprocessing strategy to address data imbalance. Better model training is made possible by this all-encompassing approach, which contributes to the creation of a balanced dataset. Outperforming individual models, the combined model identified malignant lesions with an accuracy of 89% and a recall of 90%.

In a separate work using the same dataset, a new method of detecting skin cancer was created by Syed and Albalawi [11], combining sophisticated artificial intelligence algorithms with 3D Total Body Photography (3D-TBP). Careful preprocessing methods like zooming, normalization, translation, and rotation were used to improve the dataset's diversity and representativeness in order to improve model performance. Focusing on feature extraction via transfer learning and rigorous augmentation techniques, the researchers created a specialized Convolutional Neural Network (CNN) architecture that is optimized for analyzing single-lesion crops from 3D-TBP images. The study’s findings showed that the model was effective in accurately differentiating between benign and malignant lesions, with a true positive rate of over 80% and a partial Area Under the ROC Curve (pAUC) of 85%.

Finally, in the work of Pintelas et al. [10], the authors present a brand-new ensemble model designed to minimize computational demands and maximize performance. In order to generate a diverse population of MobileNets, the model employs a two-step Expand-and-Squeeze mechanism. These are subsequently trimmed according to performance metrics. An important benchmark was the ISIC 2024 dataset, there are roughly 400,000 benign and almost 400 malignant cases. With an accuracy of 89%, an AUC of 90.5%, and a Geometric Mean (GM) of 0.809, the MobileNet-HeX demonstrated noteworthy performance. These results show that it can effectively handle the difficulties posed by unbalanced data.

3. Methodology

This study demonstrates our modified skin cancer classifier based on the VGG16 model and MobileNetV2, and showcases the effectiveness of a DCGAN framework in generating synthetic images resembling the original malignant samples from the ISIC 2024 dataset.

3.1. Dataset

For our research, the novel SLICE-3D (Skin Lesion Image Crops Extracted from 3D Total Body Photography) dataset [28] was used, published in the summer of 2024 for the ISIC-2024 challenge hosted on Kaggle. This rich dataset has 401,059 standardized, de-identified, and diagnostically labelled skin images, with 393 malignant images and 400,552 benign images. Each image was accompanied by its metadata, that were exploited for data preparation. Fig. 1 shows some interesting representations of the metadata offered by this dataset.

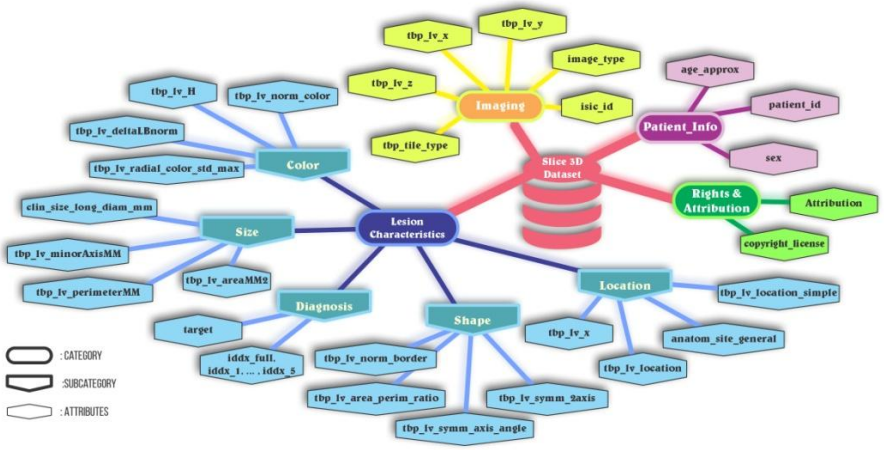


Fig. 1. The SLICE-3D dataset attribute classification

These 3D images were captured using VECTRA WB360, a 3D Total Body Photography imaging system designed to document skin pathologies. It captures the entire exposed body in a single shot using 92 fixed cameras equipped with xenon flashes for polarized and non-polarized lighting. Moreover, the DermaGraphix software was employed to enable clinicians to tag and track lesions within a secure database and to link them to pathology reports, as well as enabling clinicians to manually tag the lesions. On the other hand, the Lesion Visualizer (LV) research tool in DermaGraphix was introduced to automatically detect lesions in 3D TBP images using AI, estimating their size, shape, color, nevus confidence, and asymmetry. And finally, all Lesions in 3D TBP images were automatically detected and cropped using the ISIC2024 Tile Export Tool, with visual confirmation of a primary lesion in each image. So, in conclusion, lesions are identified either through manual tagging (by clinicians for attribution) or automated detection by Lesion Visualizer.

3.2. Data preprocessing

So, given the great imbalance between the images of the two classes, a profound dataset cleaning was needed. Basic triage and selection methods allowed us to reduce the benign images from 400,666 to 4084 images, and preserve all 393 malignant images. Fig. 2 shows the details of the data cleaning process. Before anything, the dataset was divided into two classes, using the binary target attribute (0: benign, 1: malignant), from here on, only the benign images were taken into consideration for treatment, the malignant cases stayed untouched since we already had a very small number of those images, so, we started by removing rows with null values of some chosen columns, that were considered important for analysis and model performance (age_approx, sex, lesion_id, anatom_site_general, tbp_tile_type), After that, columns content with more than 90% missing values were removed, and then duplicate images were dropped, keeping only the first image for each patient and body part. And for the last important step, only one type of lighting modality of the 3D TBP source image (tbp-tile-type) was considered, as it is the primary technical factor affecting variations in hue, tint, tone, and shade among the otherwise standardized images within this dataset. So overall, this data cleaning process helped us reduce the number of benign cases from +400,000 to 4084 images (when choosing 'tbp_tile_type' = '3D: XP).

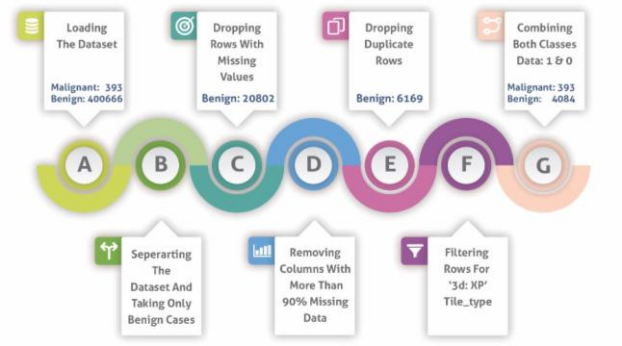


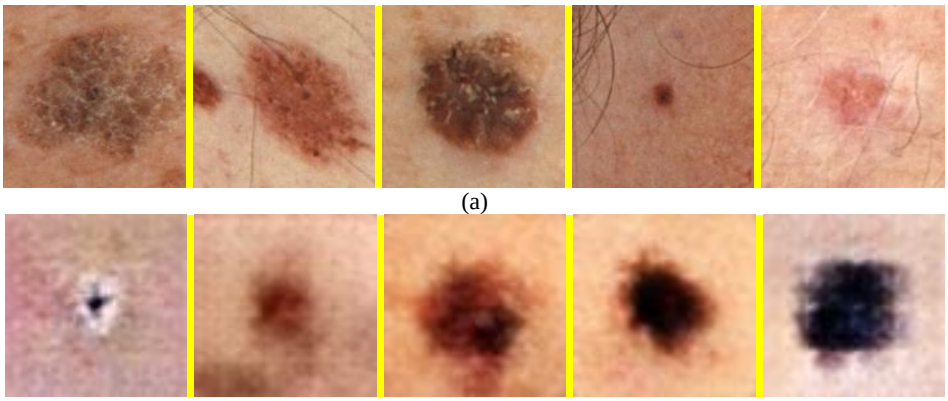
Fig. 2. Data preprocessing for the SLICE-3D dataset

3.3. Data augmentation using DCGAN

After reducing the large number of benign images as much as possible, the remaining data imbalance was addressed through data augmentation using DCGANs. In the 2016 paper, R a d f o r d, M e t z and C h i n t a l a [28] define Deep Convolutional Generative Adversarial Networks as a class of Convolutional Neural Networks (CNNs) with specific architectural constraints, demonstrating their effectiveness in unsupervised learning; they show That DCGANs can capture a hierarchical structure of representations, ranging from object components to entire scenes, within both the generator and discriminator networks. The DCGAN applied consists of two competing neural networks: first, we have the generator, which rescales a 100-dimensional latent vector into a high-resolution RGB image. This happens through a series of convolutional layers with batch norm and ReLU activation, while the final layer employs a Tanh activation to normalize pixel

values. Conversely, there is the discriminator, which verifies whether an image is real or fake, using LeakyReLU activation and progressively down-sampling by subsequent convolutional layers until producing a binary classification output via a sigmoid function.

Moreover, to encourage diversity and realism, we applied data augmentation techniques such as random horizontal and vertical flipping of the 393 malignant images during training to enhance the robustness of the discriminator, we didn't opt for the zoom technique as it isn't favoured in medical imaging because it has the potential to modify the scale and body shape of such key features and generate deceptive training images and clinically unreliable models, so we made sure to exclude it to guarantee a more accurate synthetic images. Both networks were trained using an Adam learning rate of 0.0001 and betas of (0.5, 0.999). DCGAN Training was then carried out for 350 epochs. Loss functions of both networks were monitored to attain convergence, and synthetic images were saved at intervals. Once trained, the DCGAN model was applied to the minority class and generated a total of 3691 malignant images to balance our dataset. Below in Fig. 3 are the lesion images used for our model training, where the good quality of the images generated, compared to the original ones, can be observed.



(a)

(b)

Fig. 3. Dataset skin lesions sample: original lesion images (a); generated lesion images (b)

To better assess the quality of the generated images, 2 metrics were calculated to confirm the integrity of the generated medical images; first, we calculated the Structural Similarity Index Metric (SSIM), which is a widely recognized metric for assessing image quality by evaluating the similarity between two images. It was proposed originally by W a n g et al. [29]; this metric is very much in line with how the Human Visual System (HVS) perceives image quality. Unlike other methods that are based on the calculation of error values, SSIM deals with image distortion by observing it from three perspectives: correlation loss, changes in luminance, and contrast alterations. The results of comparing original malignant images vs the same amount of the synthetic images were equal to 0.6, which is a fair result considering the limited number of original training malignant images.

We also used another metric, which is the Learned Perceptual Image Patch Similarity metric (LPIPS metric) [30], which evaluates the resemblance between

two images using feature representations obtained from deep neural networks. As opposed to traditional measures such as SSIM, which perform direct pixel-wise comparisons, LPIPS relies on intermediate layer activations of neural networks to measure visual similarity in a way that is more like human judgment. Given two images A and B, the measure computes their deep feature embeddings and approximates perceptual closeness. The results given were equal to 0.4, which is also considered a good result.

3.4. Data splitting

Once the dataset was balanced (4084 malignant + 4084 benign images), we primarily conducted an even data splitting to ensure a fair model evaluation. We took 80% of the total dataset for training and 20% for testing. Then, from the training set, we dedicated 85% for pure training and 15% for validation. After balancing the dataset, we can now see an even number of images for each subset. Table 1 demonstrates the final exact number of images used for each phase.

Table 1. Balanced dataset splitting summary

Dataset	Benign	Malignant
Training (60%)	2450	2450
Validation (20%)	817	817
Testing (20%)	817	817

This augmentation and splitting approach ensures that the model learns effectively from both real and synthetic malignant images, ultimately improving classification performance.

3.5. Modal training

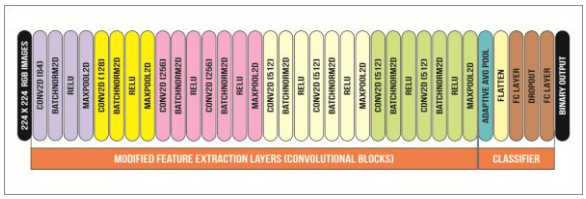
Now that our dataset is ready, we proceed with the final step, which is training. For this important task, transfer learning was applied, as it's proven to be very effective in similar tasks; additionally, the VGG16 [31] and MobileNetV2 [32] models were selected for our transfer learning.

The experiments were conducted on the Kaggle platform using a GPU-accelerated environment. The model was developed with Python using the PyTorch framework, leveraging the Adam optimizer with a learning rate of 0.0001 and the Binary Cross-Entropy with Logits loss function.

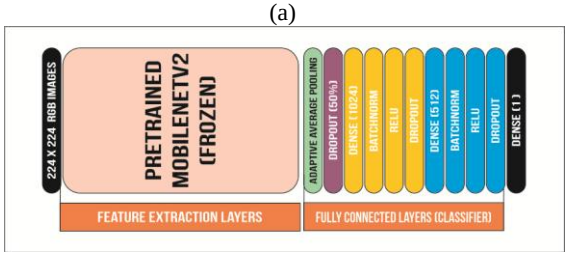
For the first experiment, the VGG16 model was trained, and it was adapted to classify skin cancer images with greater stability and performance. The original VGG16 architecture that was initially created to perform large-scale image classification was particularly adapted to our needs, with batch normalization and dropout. Primarily, the modifications relied on the feature extraction layers. Instead of using the pre-trained convolutional layers directly, it was redefined with batch normalization after each convolutional operation. This modification works to stabilize training by reducing internal covariate shifts and improving generalization. The convolutional layers themselves are identical to the conventional VGG16 architecture, but with their hierarchical feature extraction capability, along with the benefit of normalized activations. Furthermore, dropout layers were added, which serve in reducing the overfitting by randomly deactivating neurons during training,

so the model is not relying too much on specific features. The classifier section was also redesigned to better suit our binary classification task. Instead of the original fully connected layers pre-tuned to multi-class ImageNet classification, we implemented a streamlined approach using an adaptive average pooling layer, followed by a reduced fully connected network. The classifier consists of a global average pooling layer that condenses spatial features, a fully connected layer with 128 neurons, and a ReLU activation function. To enhance regularization and prevent overfitting, a dropout rate of 0.6 was used before the final classification layer, which consists of a single neuron for binary prediction; also, the Binary Cross-Entropy with Logits loss function was used, ensuring compatibility with our classification objective. The Adam optimizer, known for its adaptive learning rate, was employed to accelerate convergence and mitigate the risk of vanishing gradients. Model performance was evaluated for 100 training epochs.

After evaluating the performance of the first model, the second experiment was conducted using MobileNetV2, a lightweight and efficient architecture optimized for mobile and embedded devices. This approach aimed to compare the effectiveness of a more computationally efficient model while maintaining competitive classification performance. Unlike the first approach, this time the feature extraction layers were frozen, and there were some modifications to the classifier layer. Similarly, the classifier starts with a fully connected layer of 1024 neurons, and batch normalization is applied followed by ReLU activation, to help with generalization, dropout (0.5) was also integrated at various locations; after the first layer, another dense layer was added with 512 neurons, again followed by batch normalization, ReLU, and another dropout layer before the final output neuron that does binary classification. We used BCEWithLogitsLoss in this approach also to train and the Adam optimizer with a learning rate of 0.0001, keeping the training consistent from our previous approach. We kept track of accuracy, precision, recall, and AUC at each epoch for both experiments, giving an overall evaluation of the model's performance. The two network architectures are summarized in Fig. 4.



(a)



(b)

Fig. 4. Our network architectures for: VGG-16-based model (a); MobileNetV2-based model (b)

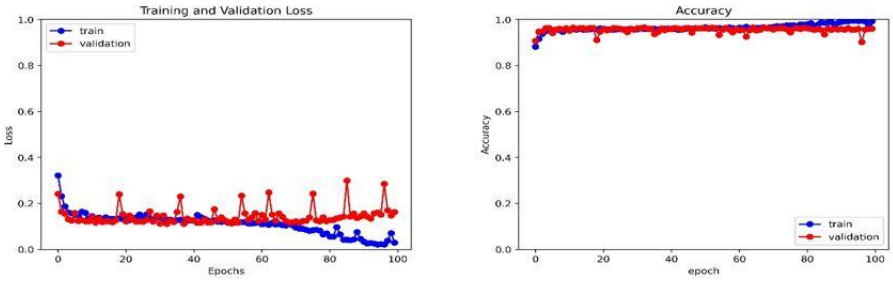
Despite these differences, both models have the same input shape of 224×224×3, both models follow the same general training approach, and use the same cleaned-augmented dataset in order to allow fair comparison between performance and efficiency.

4. Results and evaluation

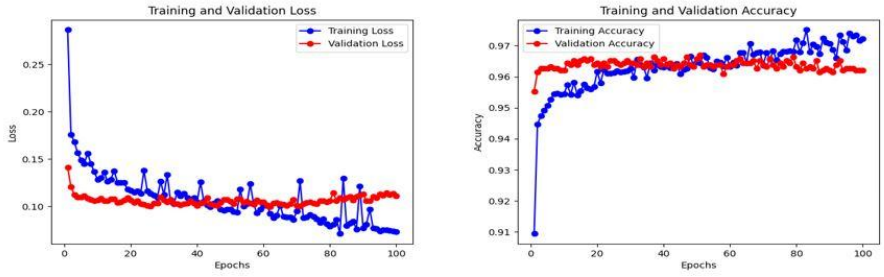
For the final results, we can say that our meticulously crafted models and our work results concluded that the MobileNetV2-based model performed slightly better than the VGG16-based model. The effectiveness of the enhanced model's training/validation is presented below in Fig. 5 regarding accuracy and loss. Furthermore, we had compared the effectiveness of our VGG16/MobileNetV2-based skin cancer classifier models using the confusion matrix illustrated in Fig. 6 on our test set, as well as the most critical metrics of performance were calculated during this testing (Loss, Accuracy, Precision, Recall, AUC) and plotted in the histogram illustrated in Fig. 7.

Our first VGG16-based model had achieved a test accuracy of 96.20%, reflecting exemplary overall classification power, with a test loss of 0.113, indicating relatively low error in predictions. The precision (97.4%) suggests that the model is predicting very few false positives, and the recall (94.8%) suggests its ability to detect most of the positive cases. In addition, the AUC score of 99% confirms the model's strong discriminatory power between classes.

The performance of the second model on the test set is also assessed on a number of measures; the model attained a test loss of 0.0926, revealing a negligible error in its prediction.



(a)

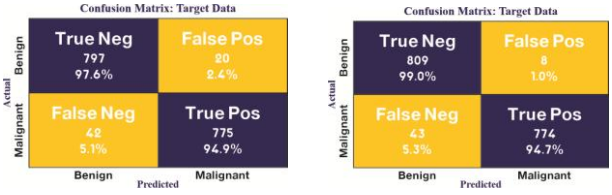


(b)

Fig. 5. The accuracy & loss functions vs epochs of: VGG-16 (a); MobileNetV2 based models (b)

Additionally, the test accuracy was recorded at 96.88%, further backing its high ability in accurately classifying instances. Besides the accuracy of 98.98% and recall of 94.74%, it reveals that the model has an impressive balance in predicting positive cases correctly and in avoiding false negatives and false positives. Moreover, the AUC of 99.13% demonstrates the model's excellence in separating the classes. The confusion matrix plotted in Fig. 6 explains the performance of the model by describing its classification decisions.

Concisely, the findings indicate both models performed outstandingly; nonetheless, the MobileNetV2-based model topped the VGG16-based model regarding loss, precision, and accuracy. In addition, the high AUC values also confirm the accurate classification performance of both models. The finding underscores the potential of the models to identify fine differences between benign and malignant lesions, which speaks to their advanced pattern recognition capacities.



(a) (b)

Fig. 6. Confusion matrix of: VGG-16 (a); MobileNetV2-based models (b)

Finally, to properly evaluate our model, we had previously reported cutting-edge results in the context of our chosen ISIC-2024 dataset [10-12], and we classified their test result in Table 2.

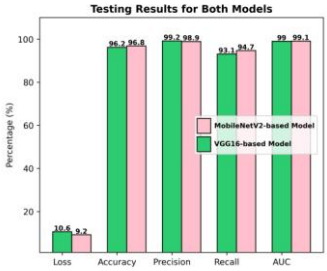


Fig. 7. Testing results of our proposed models: VGG16-based model vs MobileNetV2-based model

Table 2. State-of-the-art models and their test results vs our two models

Model	Accuracy (%)	Precision (%)	True Positives (%)	AUC (%)	Recall (%)
Pintelas et al. (Pintelas et al. 2024)	89	-	-	90	-
Syed & Albawi (Syed and Albalawi 2024)	87.5	-	80	85	-
Teymouri et al. (Teymouri et al. 2024)	89	-	-	-	90
Based on VGG-16	96.20	99.2	93.1	99	93.1
Based on MOBILENETV2	96.87	98.9	94.7	99.1	94.7

In addition, we validated the best performing model with exclusively real (not synthetic) skin cancer images, unlike previous experiments with synthetic malignant samples to augment the dataset. For testing, we removed 79 benign and

79 malignant images from the original, un-augmented dataset, for a total of 158 true images.

As indicated in Fig. 8, the resulting confusion matrix is observed to have the model accurately detecting benign instances (True Negatives = 69, 87.6%), though it fared worse with malignant instances, registering more False Negatives (23.4%). This may be because of the absence of synthetic data that had earlier balanced out more difficult-to-detect malignant patterns. This leads us to observe a decline in evaluation metrics under these constraints.

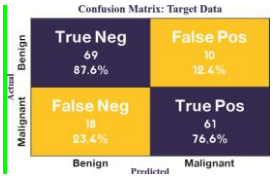


Fig. 8. Confusion matrix results after evaluation of only real images on the MobileNetV2 Model

In terms of quantitative evaluation, the confusion matrix (Fig. 7) reveals that the model achieved an accuracy of 82.3%, reflecting its overall ability to distinguish between benign and malignant lesions in the test set. The precision reached 85.9%, indicating that the majority of lesions predicted as malignant were indeed true positives, while the recall was 77.2%, suggesting that the model successfully identified over three-quarters of the actual malignant cases. This balance between precision and recall highlights a robust classification behavior, although a slight bias towards missing malignant cases (i.e., false negatives) is still present. These results underline the importance of data diversity and quantity – especially for minority classes like malignant lesions – in achieving optimal sensitivity without compromising specificity.

5. Conclusion

Although deep learning techniques have demonstrated remarkable success across diverse domains [33, 34], this paper specifically investigates their transformative impact on skin cancer classification through leveraging the power of transfer learning and data augmentation. Founded on the SLICE-3D dataset offered in the ISIC-2024 challenge, we have addressed the extreme class imbalance problem by utilizing a GAN framework in synthesizing malignant images and thereby creating a more balanced and representative dataset. Our solution has explored two popular pre-trained models: VGG16 and MobileNetV2, each of which has been altered with architectural modifications to enhance performance and generalization.

By thorough experimentation, we have determined that both models have achieved high classification accuracy, with the MobileNetV2-based model having shown slight superiority to VGG16 in fundamental evaluation metrics, including AUC, recall, and precision. Importantly, the MobileNetV2 model has shown improved efficiency while maintaining good predictive performance, thus making it a viable candidate for deployment in real-world clinical settings. Overall, our results have validated the potential of deep learning models for dermatological diagnostics,

particularly when we enhanced and balanced the dataset. By leveraging data augmentation and architectural optimization, we have demonstrated an effective strategy for improving skin cancer classification, achieving state-of-the-art performance compared to related studies on the ISIC-2024 dataset. Future work can explore other potential enhancements, such as ensemble learning or multimodal data fusion, to enhance the diagnostic performance of the model and its applicability across diverse clinical environments.

References

1. Skin Cancer Facts & Statistics, 2025 (Accessed 16.02.2025). **https://www.skincancer.org/skin-cancer-information/skin-cancer-facts**
2. Skin Cancer 101, 2025 (Accessed 16.02.2025). **https://www.skincancer.org/skin-cancer-information**
3. What Tests Will Be Done to Diagnose Skin Cancer? 2025 (Accessed 16.02.2025). **https://my.clevelandclinic.org/health/diseases/15818-skin-cancer**
4. 1 in 5 People Get Skin Cancer, Diseases & Conditions, 2023 (Accessed 15.03.2025). **https://my.clevelandclinic.org/health/diseases/15818-skin-cancer**
5. Skin Cancer Misdiagnosis, Paul and Perkins (Accessed 15.03.2025). **https://paulandperkins.com/skin-cancer/**
6. Zhang, J., Y. Wang, W. Zhang, L. Cai, J. Feng, Y. Zhu et al. Clinical Misdiagnosis of Cutaneous Malignant Tumors as Melanocytic Nevi or Seborrheic Keratosis. – Clin Cosmet Investig Dermatol, Vol. **17**, 2024, No 2, pp. 465-476.
7. Tschandl, P., C. Rosendahl, H. Kittler. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. – Sci. Data, Vol. **5**, 2018, No 1, 180161.
8. Mendonca, T., P. M. Ferreira, J. S. Marques, A. R. S. Marcal, J. Rozeira. PH2 – A Dermoscopic Image Database for Research and Benchmarking. – In: Proc. of 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13), IEEE, Osaka, 2013, pp. 5437-5440.
9. Pacheco, A. G. C., G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo et al. PAD-UFES-20: A Skin Lesion Dataset Composed of Patient Data and Clinical Images Collected from Smartphones. – Data Brief, Vol. **32**, 2020, No 10, 106221.
10. Pintelas, E., I. E. Livieris, V. Tampakas, P. Pintelas. MobileNet-HeX: Heterogeneous Ensemble of MobileNet eXperts for Efficient and Scalable Vision Model Optimization. – Big Data Cogn. Comput., Vol. **9**, 2024, No 1, 2.
11. Syed, S., E. M. Albalawi. Improved Skin Cancer Detection with 3D Total Body Photography: Integrating AI Algorithms for Precise Diagnosis. – Research Square [Preprint], 2024.
12. Teymouri, S., A. A. Larki, M. Mohammadi, S. H. Klidbary. Hybrid Vision Transformer-Boosting Model for Skin Lesion Classification on SLICE-3D. – In: Proc. of 12th RSI International Conference on Robotics and Mechatronics (ICRoM'24), IEEE, Tehran, 2024, pp. 586-592.
13. Aljohani, K., T. Turki. Automatic Classification of Melanoma Skin Cancer with Deep Convolutional Neural Networks. – AI, Vol. **3**, 2022, No 2, pp. 512-525.
14. Djaroudib, K., P. Lorenz, R. Belkacem Bouzida, H. Merzougui. Skin Cancer Diagnosis Using VGG16 and Transfer Learning: Analyzing the Effects of Data Quality over Quantity on Model Efficiency. – Appl. Sci., Vol. **14**, 2024, No 17, 7447.
15. Sikandar, S., R. Mahum, A. E. Ragab, S. Y. Yayilgan, S. Shaikh. SCDet: A Robust Approach for the Detection of Skin Lesions. – Diagnostics, Vol. **13**, 2023, No 11, 1824.
16. Ingle, Y, S., N. F. Shaikh. Deep Learning for Skin Cancer Classification: A Comparative Study of CNN and VGG16 on HAM10000 Dataset. – Commun. Appl. Nonlinear Anal., Vol. **31**, 2024, No 4s, pp. 490-499.

17. P r e m n a t h, K., P. S u n d a r a v a d i v e l, R. A u g u s t i a n I s a a c, K. S r i m a t h i, M. V a i s h n a v i, N. S a t h y a. Deep Learning-Based Skin Lesion Classification for Cancer Detection. – Research Square [Preprint], 2024.
18. M'h a m e d i, M., M. M e r z o u g, M. H a d j i l a, A. B e k k o u c h e. Enhancing Melanoma Skin Cancer Classification through Data Augmentation. – TELKOMNIKA Telecommun. Comput. Electron Control, Vol. **22**, 2024, No 5, 1209.
19. G o u d a, W., N. U. S a m a, G. A l-W a a k i d, M. H u m a y u n, N. Z. J h a n j h i. Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. – Healthcare, Vol. **10**, 2022, No 7, 1183.
20. M e h r, R. A., A. A m e r i. Skin Cancer Detection Based on Deep Learning. – J. Biomed. Phys. Eng., Vol. **12**, 2022, No 6.
21. A l r a b a i, A., A. E c h t i o u i, F. K a l l e l. Skin Cancer Detection Based on Transfer Learning Techniques. – In: Proc. of 7th IEEE International Conference on Advanced Technologies, Signal and Image Processing (ATSIP'24). Sousse, Tunisia: IEEE; 2024, pp. 8-13.
22. I s l a m, M. S., S. P a n t a. Skin Cancer Images Classification Using Transfer Learning Techniques. arXiv; 2024.
23. P o p e, J., M. H a s s a n u z z a m a n, M. S h e r p a, O. E m a r a, A. J o s h i, N. A d h i k a r i. Skin Cancer Machine Learning Model Tone Bias. – arXiv; 2024.
24. I m r a n, T., A. S. A l g h a m d i, M. S. A l k a t h e i r i. Enhanced Skin Cancer Classification Using Deep Learning and Nature-Based Feature Optimization. – Eng. Technol. Appl. Sci. Res., 8 February Vol. **14**, 2024, No 1, pp. 12702-12710.
25. N a e e m, A., T. A n e e s, M. K h a l i l, K. Z a h r a, R. A. N a q v i, S. W. L e e. SNC_Net: Skin Cancer Detection by Integrating Handcrafted and Deep Learning-Based Features Using Dermoscopy Images. – Mathematics, Vol. **12**, 29 March 2024, No 7, 1030.
26. S e d i g h, P., R. S a d e g h i a n, M. T. M a s o u l e h. Generating Synthetic Medical Images by Using GAN to Improve CNN Performance in Skin Cancer Classification. – In: In: Proc. of 7th International Conference on Robotics and Mechatronics (ICRoM'19), Tehran, Iran: IEEE; 2019, pp. 497-502.
27. T e o d o r o, A. A. M., D. H. S i l v a, R. L. R o s a, M. S a a d i, L. W u t t i s i t t i k u l k i j, R. A. M u m t a z et al. A Skin Cancer Classification Approach Using GAN and RoI-Based Attention Mechanism. – J. Signal Process Syst., Vol. **95**, March 2023, No 2-3, pp. 211-224.
28. R a d f o r d, A., L. M e t z, S. C h i n t a l a. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. – arXiv, 2016.
29. W a n g, Z., A. C. B o v i k, H. R. S h e i k h, E. P. S i m o n c e l l i. Image Quality Assessment: from Error Visibility to Structural Similarity. – In: IEEE Transactions on Image Processing, Vol. **13**, April 2004, No 4, pp. 600-612.
30. Z h a n g, R., P. I s o l a, A. A. E f r o s, E. S h e c h t m a n, O. W a n g. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. – In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586-595.
31. S i m o n y a n, K., A. Z i s s e r m a n. Very Deep Convolutional Networks for Large-Scale Image Recognition. – In: Proc. of 3rd International Conference on Learning Representations (ICLR'15), 2015, pp. 1-14.
32. S a n d l e r, M., A. H o w a r d, M. Z h u, A. Z h m o g i n o v, L.-C. C h e n. MobileNetV2: Inverted Residuals and Linear Bottlenecks. – In: Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510-4520.
33. G y i r e s-T ó t h, B. P., M. O s v á t h, D. P a p p, G. S z ű c s. Deep Learning for Plant Classification and Content-Based Image Retrieval. – Cybernetics and Information Technologies, Vol. **19**, 2019, No 1, pp. 88-100.
34. V a n d h a n a, S., J. A n u r a d h a. Spatial and Temporal Variations on Air Quality Prediction Using Deep Learning Techniques. – Cybernetics and Information Technologies, Vol. **23**, 2023, No 4, pp. 213-232.

Received: 27.03.2025. First Revision: 02.06.2025. Second Revision: 24.06.2025. Third Revision: 30.06.2025. Accepted: 04.07.2025