Al-Nahrain Journal for Engineering Sciences

Al-Nahrain J. Eng. Sci. 29(1) 141-151, 2026
Original Article

YOLOv11 with spatial attention and preprocessing enhancements for accurate skin cancer classification

Aythem Khairi Kareem[1]●, Ahmed Adil Nafea [2]● and Afrig Aminuddin [3]●

1Department of Heet Education, General Directorate of Education in Anbar, Ministry of Education, Iraq
2Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Anbar, Iraq.
3Faculty of Computer Science, Universitas Amikom Yogyakarta Sleman, Indonesia.
●*These authors contributed equally to this work*

DOI: http://doi.org/10.29194/NJES.29010141

Received: October 31, 2025 **Revised:** Desember29, 2025 **Accepted:** January 18, 2026 **Published:** March 20, 2026

Abstract

This work suggests a Deep Learning (DL) architecture based on You Only Look Once YOLOv11 for Skin Cancer (SC) detection. The similarity between malignant and benign lesions makes visual inspection a failure to distinguish between them. To solve this problem, the proposed approach uses a 3-step pre-processing stage, namely hair removal, color normalization, and Contrast Limited Adaptive Histogram Equalization (CLAHE) contrasts, has been conducted to eliminate artifacts and improve image quality. Balanced data augmentation on the training set of the PROVe-AI dataset. In this process, YOLOv11 with C3k2 module and C2PSA module showed significant results in optimized multi-scale feature collection and spatial interest. The experimental outcome demonstrates that the proposed model has a classification accuracy of 93.09% and led the baseline models, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Artificial Neural Network (ANN). The proposed optimized YOLOv11 architecture allows for skin cancer detection in a computationally efficient framework with promising preliminary results so that the proposed approach can be a beneficial Artificial intelligence (AI) tool for early diagnosis, particularly in a lack of high-tech medical facilities.

Keywords: *Deep Learning, Medical Image Analysis, Melanoma Detection, Skin Cancer Classification, YOLOv11.*

Corresponding author: Provide the corresponding author information and publisher here. E-mail address: ayt19c1004@uoanbar.edu.iq

1. Introduction

The skin, the largest organ of the human body, serves as a central interface between the external environment and the internal systems most critical to human functioning. Focused primarily on its anatomical composition, it consists of three primary layers: the hypodermis as the inner layer responsible for energy storage and insulation; the dermis, which provides structural support alongside housing critical components such as blood vessels and sensory receptors; and the epidermis protected by the first two layers, the outer layer responsible for forming a protective barrier for the parts below. Concerning its physiological purposes, the skin acts as a protective barrier against reducing the risk of harmful microorganisms entering, protecting the body from mechanical trauma, and regulating exposure to extreme heat and cold [1, 2].

Skin lesions are situated anomalies in the skin's structure that are distinct from the adjacent healthy tissue. These alterations can be congenital from birth, arise later in life, or present as the first indication of a medical disease. These skin lesions can be non-cancerous or cancerous. Among the most prevalent kinds is SC, with an estimated 1.2 million cases documented in 2020. There has been a significant increase in the SC rate of occurrence across various countries [3]. The primary reasons for SC include extensive use of tanning and solarium beds, infection, allergies, excessive Ultraviolet (UV) light exposure, older age, and a family history of it. Basal cell carcinoma, actinic keratoses, melanoma, and squamous cell carcinoma are the most widespread conditions of SC. Basal cell carcinoma appears in the inferior region of the epidermis and is seldom life-threatening. Squamous-cell skin carcinoma arises in the outermost epidermis layer and can progress to life-threatening conditions without treatment [4].

Being able to diagnose SC early is still considered a great benefit. Although there have been significant advancements due to clinical screening and medical imaging, the detection of SC faces a major challenge. Due to the fact that SC has similarities to benign conditions in the texture and color of the skin, there is a significant number of cases of misdiagnosis or prolonged detection. In order to address these issues, the traditional methods, such as visual diagnosis, require the attention and skills of a dermatologist. Hence, the potential for subjectivity and the lack of accessibility are concerns in resource-scarce areas. For this reason, there is a great need for

efficient, reliable, and accurate diagnostic tools to discriminate between SC and benign lesions much earlier [5, 6].

Generally, medical image analysis techniques can be categorized into Machine Learning (ML) and DL techniques. Traditional ML often employs manual feature engineering, which is difficult due to the complexity of medical images. In contrast, DL has proven itself in numerous research fields, showing superiority over ML by learning relevant features to classify medical images. For example, in DL, significant progress has been achieved with the help of CNN, RNN, LSTM, and ANN [7, 8]. YOLO is one of the most popular DL-based object detectors. Ongoing developments within the YOLO family provide a high level of accuracy for real-time systems because of their architectural advancements. Specifically, the YOLOv11 framework improved performance when the model architecture was through architectural advancements within the backbone, neck, and detection head. In other words, the YOLOv11 model was developed and integrated with the Region-based Convolutional Neural Network (R-CNN) framework to improve its performance [9, 10].

In order to resolve these issues, a YOLOv11 model was developed specifically for SC classification. By introducing C2PSA and C3k2 blocks in this model, improved feature extraction and classification efficiency are shown. To estimate the performance of the developed model, the test was conducted on 603 high-resolution images in a binary class scenario, and the results show that the proposed architecture achieved the optimization level of existing models.

The main contributions are as follows:

- Balanced Dataset Preparation on training data: The study standardized the PROVe-AI dataset (603 biopsy-confirmed images) by applying symmetric offline augmentation, creating ~5,000 balanced images (2,500 per class) to address class imbalance and enrich variability.
- Core Preprocessing Pipeline: Introduced a three-step preprocessing method (hair removal, color constancy, and CLAHE) to improve lesion visibility, reduce noise, and normalize lighting conditions, ensuring robust model learning.
- Integration of YOLOv11 for Classification: Adapted the YOLOv11 architecture, traditionally used for detection, into a high-speed and accurate skin cancer classification framework, leveraging backbone (C3k2 + C2PSA), neck (up-sampling + concatenation), and head modules.
- Efficient Architectural Enhancements: Employed advanced modules (C3k2 block for computational efficiency, C2PSA for spatial attention) to enhance multi-scale feature representation and improve performance in distinction between benign and malignant lesions.

The rest of this paper is organized as follows: Section 2 explains related work, Section 3 presents the methodology, Section 4 describes experiments and results, discusses findings, and Section 6 concludes with future directions.

2. Related Works

Recent improvements in image processing, DL techniques deliver advantageous effects in early SC detection. Determining the skin lesion area is one of the essential steps in creating a practical decision aid system for specialists. With this objective, multiple techniques have been presented in recent years. Recent and valuable studies conducted to optimize the effectiveness and accuracy of automated SC segmentation were analyzed to demonstrate the contributions of the presented research.

Ibrahim et al. [11] present an ensemble transfer-learning model for multiclass SC classification. Five comparative models were employed as the base of the ensemble model: DenseNet201, ResNeXt101, Xception, Efficient-NetV2B2, and MobileNet. Moreover, the 10-fold cross-validation is used to reduce data bias. A Test Time Augmentation (TTA) and annealing learning rate scheduler were even employed to improve the model performance. Publicly dermatoscopy Human Against Machine 10000 (HAM10000) dataset were utilized to evaluate the presented models. The presented model achieved 94.49% accuracy. These outcomes suggest that this technique can be effective in enhancing the performance of SC classification.

Pandey et al. [12] present an effective SC detection method that utilizes image processing strategies, non-local means denoising, and a CNN substantiated by sparse dictionary learning. The most significant merit of Non-Local Mean (NLM) noise elimination was realized with the sparse dictionary learning in CNN-based techniques to image categories. The two integrated strategies successfully handle feature extraction, noise resistance, and classification performance, thus remarkably robust incomplete SC image investigation tasks. The CNN-based model is efficient in the classification of the SC image because it was trained using a CNN method to analyze the SC image and achieved an adequate level of improved classification of cancer SC classes. The International Skin Imaging Collaboration 2019 (ISIC-2019) dataset accuracy that was approved and accepted is: 81.23% and 85.61% for HAM-10000 dataset.

Diptho et al. [13] presented multiple DL techniques, including VGG16, ResNet50, YOLOv8, and YOLOv11. A dataset of 900 cutaneous images, meticulously selected for nine SC disease classes, was employed for the analysis. Generalizability and robustness were ensured by utilizing data hyperparameters and augmentation optimization. YOLOv11 achieved an optimal accuracy of 80.72%. The performance of YOLOv11 represents an encouraging step in the advancement of highly performing SC prediction models.

Huang et al. [14] presented an effective Spectrum-Aided Vision Enhancer (SAVE) that changes images into generated narrowband imaging models in an unmarried phase. The presented approach employed five YOLO versions, including YOLOv5,8, 9, 10, and 11 models, which were evaluated over three SC classes using SAVE and White Light Imaging (WLI) diagnostic tools. The models were trained individually on WLI and SAVE images. YOLOv10 achieved the highest prevalence effectiveness, specifically under the SAVE diagnostic tools. Integrating the SAVE diagnostic tools with YOLOv8 significantly improves the performance of SC detection corresponding to traditional WLI.

This work Erbay et al. [15] concerns a DL-based model for SC forecasting that employs the HAM10000 dataset. Different types of advanced CNN models, such as Xception, Inception-ResNetV2, and DenseNet201, were included in the investigation in terms of a fully fine-tuned and frozen configuration. The latter was also compared with the performance of Vision Transformer (ViT) framework to

determine the level of its potential influence on SC research field. Ensemble learning approaches, including soft voting, weighted soft voting, and hard voting, were applied to enhance classification performance. The results show that fully finetuned configurations are significantly better than frozen ones, and the highest accuracy of 0.88% was reached by Inception-ResNetV2. With regard to the most outstanding comprehensive performance, the weighted soft voting ensemble demonstrated the best potential of 0.89% accuracy. Therefore, one can possibly state that ensemble models and transfer learning-based methods are beneficial in the classification of SC.

A lightweight hybrid strategy is proposed by Ozdemir et al. [1] with the multicast of ConvNeXtV2 blocks and focal self-attention to tackle the model complexity and unbalancing data. Additionally, the model utilizes ConvNeXtV2 in the first two steps to enhance local feature extraction, while in the next few steps of focal self-attention, it concentrates on the significant regions, which may diagnose for better accuracy. The model results on the ISIC 2019 dataset that comprises eight SC types and imbalances in the class. Although there exists a vast gap with the comparison model, our approach would perform a core vital performance over all classes with 93.60% accuracy. This novel technique will act as an effective benchmark for SC detection; it produces flexible, scalable, and accurate prediction that is expanded for diagnostic assistance to improve the clinical outcome.

This paper Shaik et al. [16] proposed a hybrid model via integration of advanced CNNs with BiLSTM improved by temporal attention, channel, and spatial mechanisms to enhance the SC classification. The presented model is developed to determine among different SC classes with high accuracy. The presented model demonstrated optimal performance, with an accuracy of 92.73%. The primary objective is to empower healthcare professionals by delivering a reliable diagnostic model that improves performance and substantiates anticipatory management systems.

While this paper Nanda et al. [17] proposed an exhaustive strategy based on DL models for SC diagnosis. The presented strategy utilized the HAM10000 dataset, which comprises over 10,015 dermatoscopic images, and functions as a criterion for assessing SC prediction models. The presented strategy employed three DL models, namely, ResNet, GoogleNet, and CNN. Performance will be estimated using standard metrics, including recall, precision, accuracy, and F1-score. The obtained outcomes prove that the CNN model has a high accuracy of 73%.

Asriani et al. [18] present a technology-based approach via classifying SC utilizing a CNN technique with a ResNet50 framework. The HAM10000 dataset was utilized for model training. Different testing strategies were performed to select the effective parameter integration. The optimal outcomes were reached with an accuracy of 83.84%.

In this work, Mondal et al. [19] present a different hybrid CNN model for SC detection, harnessing the complementary advantages of different models. Investigation contains varieties such as VGG with ResNet, MobileNet with VGG, ResNet with MobileNet, etc. Further, an effective hybrid model is presented, consisting of GoogleNet with MobileNet as the feature extraction backbone. Extracted features are then provided to a Feed-Forward Neural Network (FFNN) for classification. The significance of each hybrid model was evaluated based on the specificity, sensitivity, and accuracy via exhaustive testing of an SC dataset. The final presented model, GoogleNet with MobileNet, reaches an accuracy of 87.9%, indicating its significance in the accurate classification of SC.

Ilyas et al. [20] present a robust DL-based SC classification approach employing the ResNet Framework. The analysis utilizes the HAM10000 dataset, a comprehensive exhibition of cutaneous imaging depicting different SC types. The method presented here comprises several steps, including dataset preprocessing, extracted features using pre-trained DL models, and classification using ResNet. Data augmentation techniques are employed to enhance model performance and mitigate class imbalances. Feature extraction leverages ResNet's hierarchical feature presentation, which extracts complicated patterns in dermatoscopic images to reinforce accurate classification. The approach is fine-tuned and optimized with the optimal hyperparameters to reach the highest classification accuracy with reduced computational overhead. Experimental findings demonstrate that the presented approach has a superior classification accuracy of 92.5%.

Indeed, the existing studies in SC classification have yielded promising outcomes, thanks to ensemble learning, hybrid CNN models, and earlier YOLO versions. However, they have several critical limitations, including dataset imbalance, single public data sets such as HAM1000, inability to incorporate noise and illumination variations, and constrained generalization to diverse clinical scenarios. Coupled with complex architectures or overemphasis on transfer learning, none of these models explicitly address the artifacts, such as hair, lighting imbalance, or the smallness of lesions, presented by dermoscopic image data from the real world. The developed framework thus bridges these gaps and limitations, offering enhanced accuracy, with balance and an augmented dataset produced from PROVe-AI images that have been diagnosed by biopsy to enable actual robust training despite the dataset's initial imbalance. I vendored a core preprocessing pipeline, including hair removal, color constancy and CLAHE for enhanced diagnostic feature visibility and reduced misclassification due to artifacts. At the architectural level, we repurposed the YOLOv11 detection model to specialize for the SC classification pipeline, equipped with C3k2 blocks for computational efficiency, and C2PSA for spatial attention. These innovations not only overcome prior work limitations but yield substantial performance enhancements, with 93.09% accuracy, indicating a faster, more accurate, and clinically adaptive SC detection solution.

3. Research Methodology

Figure 1 shows the General SC classification method with the YOLOv11 approach. Note: The image pipeline generally includes: configuration of the dataset; division into training, validation, and test; augmentation of the training set; fine-tuning of YOLOv11 and the four DL models with suboptimal hyperparameters. The model is tested with performance metrics like the confusion matrix, f1-score, precision, accuracy, and recall.

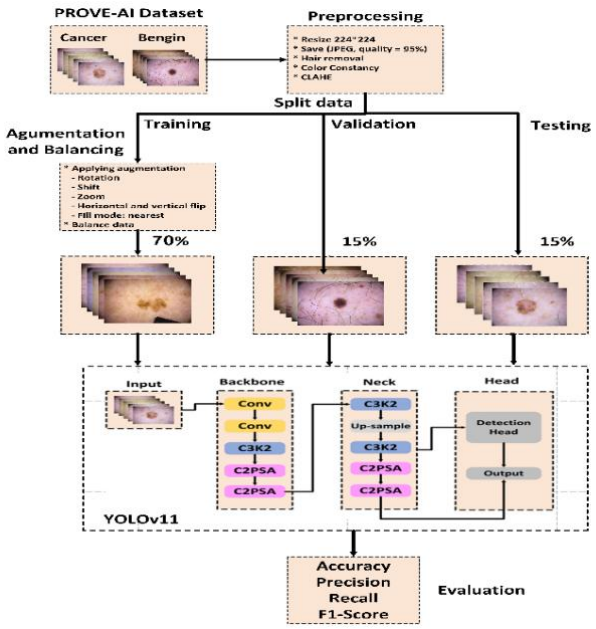

Figure 1. Block diagram of the presented model

3.1 Dataset Description

The proposed method uses "PROVe-AI," a non-commercial dataset. The 603 images, all of which were thermoscopic photographs of skin lesions histopathologically verified by skin biopsy, were obtained from patients in which the biopsy was performed to exclude the possibility of melanoma [21]. Thus, this dataset is specifically designed to be a reference standard against a novel open-source AI algorithm for clinical, real-world diagnostic test efforts to detect melanoma. Of them, 486 cases were benign cases, and 117 were skin malignancies. This data set is already integrated into the latest model reviews and validated as a universal reference data set in papers on automatic melanoma morbidity detection.

3.2 Preprocessing

The preprocessing step is vital to ensure the comparability among datasets. First, all images are resized to 224 × 224 pixels in order to have a uniform input size. This resolution was chosen because it is the default input size for the YOLOv11 model, balancing computational efficiency and diagnostic feature preservation. In order to keep image quality, they were saved using Joint Photographic Experts Group (JPEG) 95% quality compression to prevent loss of clinically useful patterns. Then, three preprocessing operations are applied to enhance images, including hair removal, color constancy, and CLAHE. These preprocessing steps guaranteed equal input sizes, smaller disk space requirements, and suitable conditions for subsequent augmentation, splitting, and classification operations.

Hair Removal

Hair-removal preprocessing focuses on thin, dark strands that surround lesion borders and achieve a texture that can mislead segmentation and classification. Early pipelines identify hairs by morphological filtering and inpainting over these hairs to recover the pixels beneath; more recent methods, such as SharpRazor, generalize this approach to work well across varying hair thicknesses [22]. Since most of the images used in the presented approach contain hair, which can lead to a malfunction in the model's performance, it is vital to address this issue. The presented approach uses the traditional pipeline to perform the hair removal operation. Morphological black-hat filtering is first performed on the grayscale dermoscopy image with a 17×17 structuring element to enhance fine dark structures like hairs and ruler regions. Subsequently, an Otsu thresholding is applied to create a binary hair mask (computed using OpenCV implementation cv2. Threshold (blackhat, 0, 255, cv2. THRESH_BINARY + cv2. THRESH_OTSU)). Masked pixels are filled in, finally with Telea fast-marching inpainting (cv2. INPAINT_TELEA) with radius = 1 to recover the underlying skin texture and suppress hair/ruler artifacts. An example of the hair-removal operation is shown in Fig. 2.

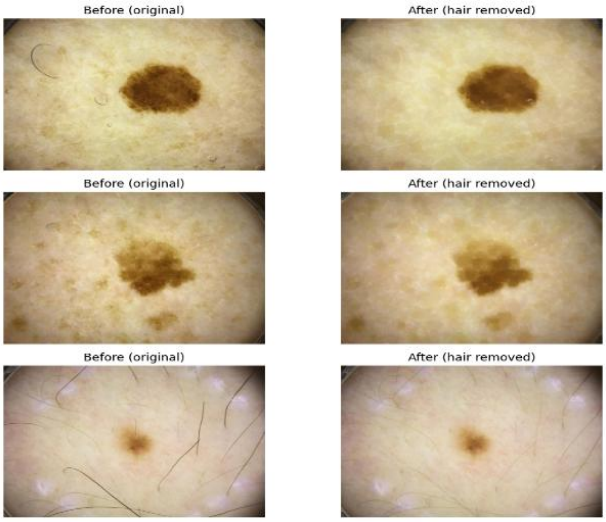

Figure 2. A sample hair removal operation.

Color constancy

Data is typically collected at different times, using different imaging devices and under different lighting conditions. Therefore, the images will vary in color and lighting. Therefore, color constancy normalizes the color of lesions with respect to one another by correcting for illumination, facilitating the model to emphasize the true chromatic properties of tissue and not be affected by lighting changes. Gray-World: normalize Red-Green- Blue (RGB) channels so the average color in an image is achromatic; Shades-of-Gray: generalization of gray-world using a Minkowski-norm estimate of illuminant normalize; both are simple, practical preprocessing steps [23, 24]. In this study, Shades-of-gray color constancy estimates the illuminant via a Minkowski-p norm using p=6 per RGB channel and rescales them toward an achromatic gray to reduce lighting bias. Fig. 3 displays a sample of a hair removal operation.

Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE optimizes contrast with an adjustable clip limit by dividing images into tiles, clipping histograms to a given clip limit, redistributing excess over all gray value histogram bins, and efficiently interpolating bilinearly between tiles to make the image more detailed and without undesired noise increase [25]. In this work, the CLAHE is applied on the luminance component after color constancy, with clipLimit 2.0 and tileGridSize 8×8. CLAHE balances local contrast by preserving image content, first by clipping overly high histogram bins in tiles, redistributing the clipped mass, and then stitching tiles without creating seams, to simultaneously

enhance delicate lesion structures (borders, streaks, pigment networks) while suppressing noise and halos. Having applied to luminance, it preserves chromaticity set by color constancy; but in practice, we tune clipLimit (≈1.5–3.0) for the best trade-off and keep identical settings across train/val/test to prevent distribution shift.

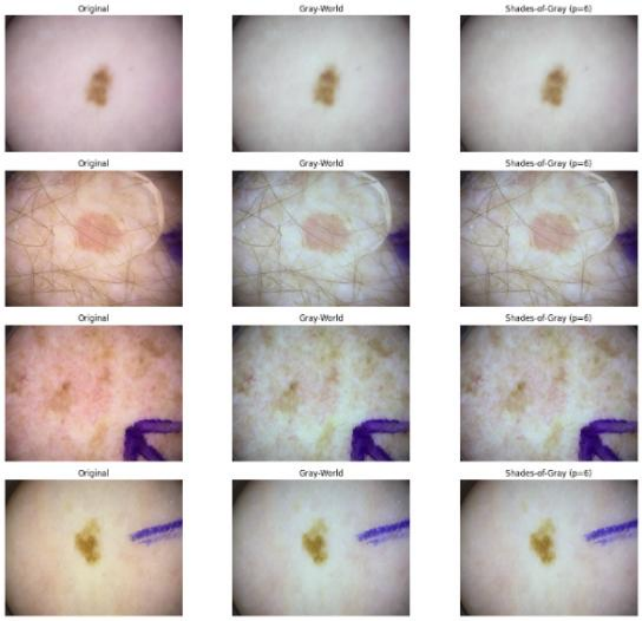

Figure 3. A sample of color constancy

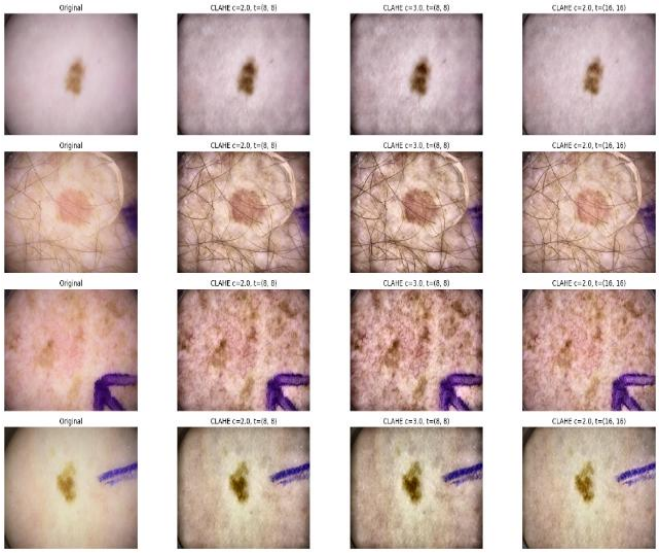

Figure 4. A sample of CLAHE

3.3 Splitting data

Data partitioning is an important step before entering it into the YOLOv11 and DL models. The data was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The training subset is used to fit the model, the validation subset is used to tune hyperparameters and avoid overfitting, and the testing subset is employed to assess the achieved performance. Each subset includes two subfolders, "benign" and "SC", confirming class balance is maintained across all splits.

3.4 Augmentation

Offline Augmentation (OFA) has demonstrated a vital component in addressing class imbalance in SC datasets. It synthetically enlarges minority classes by enriching variability. Geometric and color-based augmentations include cropping, scaling, rotations, flips, and contrast adjustments, which are commonly applied offline. The augmentation of the spectral domain spatially and contextually pools enable balancing the class distributions before training to minimize model bias towards majority classes [26]. This OFA helps create robust and generalizable models. It enhances detection performance on underrepresented lesion categories [27]. In this study, symmetric OFA was performed on the training set to balance the data between the two classes: benign and SC. Both classes had varying numbers of images produced. For the process to commence, the number of oral samples was duplicated for each class (thrashing). But, for classes that operated less, more images were synthetically produced by the augmentation process. The kinds of augmentations used include flipping, rotation, shifting, zooming, and shearing. For the latter, the number of images produced was beyond 5,000. By performing augmentation offline, the dataset was standardized, varied, and balance-free, allowing the DL and YOLOv11 models to learn more invariant and generalizable features from SC classification.

3.5 Deep Learning Models

Baseline DL frameworks serve as the basis for assessing the performance of developed classification architectures of SC. Among these, the CNN serves as a core benchmark due to its practical spatial feature extraction abilities, allowing it to learn discriminatory patterns from dermoscopic images. The RNN and its variation, the LSTM, are also employed as baselines to capture sequential contextual and dependency associations within image data when represented as pixel sequences or feature vectors. Furthermore, while ANN provides a simple feed-forward structure to represent intricate non-linear relationships, they are not spatially informed, unlike CNNs. Even though these baseline models are quite powerful for the selected applications, they tend to underperform due to a lack of feature generalization, presence of class imbalance, and noise robustness, resulting in poor diagnostic accuracy. These simple models, therefore, are essential for direct comparison with the progress given by the improved models, such as YOLOv11, which exploit spatial attention and multi-scale feature fusion for more accurate and robust SC classification.

Artificial Neural Network (ANN)

An ANN is a statistical and nonlinear forecast model. It is based on the design of the human brain's biological structure. ANN comprises three layers: the Input Layer (IL), the Hidden Layer (HL), and the Output Layer (OL). The function of IL is to transfer data to the HL of neurons. The primary architecture of an ANN comprises several HLs. Calculations are learned at each layer operating backpropagation, which is utilized to comprehend the complicated association between IL and OL. The primary function of ANNs in SC prediction is to categorize the extracted characteristics. The total number of HLs present in ANN relies on the number of input images [28]. The architecture of the presented ANN includes a Flatten layer, followed by three Fully Connected Layers (FCLs) with 1024,512 and 256 neurons, Rectified Linear Unit (ReLU) activated (with Batch Normalization (BN) and Dropout after each layer). The network is concluded with a final single-neuron OL with sigmoid activation for binary classification.

Convolutional Neural Network (CNN)

The CNN is a developed Version of the Multilayer Perceptron (MLP), particularly created for conducting two-dimensional data.

The CNN is a subset of DL, which is commonly used in image data implementations. The CNN architecture comprises a Convolution Layer (CL), a Pooling Layer (PL), and a Fully Connected Layer (FCL). The CL handles the weights that PLs reduce to create output from and decrease the input size ratio. Then the PL is employed and provided to the FCL. A crucial part of CNN is the CL, which includes a range of weights tailored for different implementations, such as multiple 2D matrices and image segmentation [29]. The architecture of the presented CNN is built with three CLs of 32, 64, and 128 filters (each block contains two Conv2D layers; BN + ReLU are then repeated and followed by MaxPooling; the latter is then followed by Dropout of probability 0.2/0.3/0.4). Then Dense (128) ReLU, and 0.4 Dropout at the end, SoftMax out for two classes after feature extraction.

Recurrent Neural Network (RNN)

RNNs are an ANN in which links between nodes in each layer of a controlled graph are formed with the following variables. The architecture of an RNN comprises an IL, one or more HLs, and an OL. RNN has an architecture equivalent to a chain of replicating modules. The module is utilized as a memory to keep critical data from the previous phase of the process. Additionally, RNNs provide performance feedback, allowing the neural network to handle sequential data. Therefore, the outcome of the previous phase will be returned to the network, affecting the outcomes of the current and subsequent phases. An input, an output, and hidden units that replicate [30]. The architecture of the presented RNN is the first layer being Time Distributed Dense (128) with ReLU activation, which projects each row of an image before two-stacked Bidirectional RNN layers of 128-unit returned sequences and 64 units, respectively. A sigmoid 1-neuron output for binary classification terminates it.

Long Short-Term Memory (LSTM)

The LSTM model is employed for its potential to promote learning in the SC classification. LSTM was developed for its effectiveness in avoiding potential problems, including the diminishing gradient that occurs during training. Further, LSTM can learn order dependency in sequence forecast tasks. LSTMs improve RNNs by handling the exploding and vanishing gradient issues. LSTMs preferentially discard or retain data operating four strategies: the Forget Gate (FG), the Input Gate (IG), the Tanh function, and the output gate (OG) [31]. The architecture of the presented LSTM consists of a Time Distributed Dense projection layer (256 units) with ReLU activation, and a stack of two layers of LSTM with 128 and 64 units, respectively. It is followed by a single-neuron OL with sigmoid activation for binary classification.

3.6 Proposed YOLOv11 Model

The YOLOv11 model constitutes a substantial development in the YOLO series, specifically for classification assignments, based on the improvements of prior versions such as YOLOv10, YOLOv9, and YOLOv8. While YOLOv11 is commonly correlated with object detection assignments, it also achieves very well in classification by making adjustments to the architecture. In a classification structure, YOLOv11 employs its backbone network, which is accountable for feature extraction tasks, comparable to DL like ANN, CNN, RNN or LSTM techniques. This backbone operates on the input image, extracting hierarchical characteristics and developing a feature map. Instead of predicting bounding boxes, the classification network uses a global average pooling layer to convert a high spatial resolution feature map into a single tight feature presentation. This feature vector is then processed through several FCLs, which output class probabilities and box predictions, respectively, and thus classification is done accurately. Hence, YOLOv11 can be considered to accurately and speedily classify which can be ideal for the medical image classification task [32]. The overall YOLOv11 architecture primarily included three key parts which are a backbone, neck, and head, and each part is defined as [33, 34].

3.6.1 Backbone:

The backbone is a vital part of the YOLOv11 architecture, accountable for extracting characteristics from the input image at numerous scales. This method employs stacking CLs and technical blocks to create feature maps at miscellaneous resolutions. Backbone includes several components, which can be described as follows:

Convolutional Layers (CLs): YOLOv11 possesses a configuration equivalent to its predecessors, using initial CLs to down-sample the image. CLs create the basis of the feature extraction method, gradually decreasing spatial dimensions while expanding the channel numbers.

C3K2 block: A substantial advancement in YOLO11 is the utilization of the C3k2 block, which substitutes the C2f block used in earlier versions. The C3k2 block is a computationally enhanced performance of the Cross Stage Partial (CSP) Bottleneck. It uses two smaller convolutions as a substitute for one large convolution. The "k2" in C3k2 denotes a smaller kernel size, which facilitates rapid processing while preserving performance.

Cross Stage Partial with Spatial Attention (C2PSA): YOLO11 introduces a new C2PSA block, which is considered a substantial expansion that optimizes spatial attention in the feature maps. This spatial attention instrument permits the YOLOv11 model to concentrate more efficiently on vital regions within the input image. The C2PSA block allows YOLO11 to focus on specific interest areas, potentially enhancing performance across varying positions and sizes.

3.6.2 Neck

The neck integrates features at various scales and transfers them to the head for forecasting. This mechanism generally applies up-sampling and concatenation of feature maps from multiple levels, allowing the YOLOv11 model to capture multi-scale information efficiently. However, the neck includes several components, which can be described as follows:

C3K2 blocks: YOLO11 presents an effective C3K2 block in the neck, developed to be faster and more effective, thereby improving the prevalent performance of the feature aggregation operation.

Up-sampling layers: Up-sample in YOLOv11 expands the spatial resolution of feature maps, allowing deeper semantic features to align with shallower, high-resolution features. This process facilitates multi-scale fusion, optimizes the model's capability to detect small objects.

Concatenation operations (Concat): Concatenation in YOLOv11 merges feature maps from different scales or paths, combining fine-grained spatial details with deeper semantic information. This operation enriches feature representation and strengthens multi-scale object detection.

3.6.3 Head:

The head of YOLOv11 is accountable for developing the ultimate predictions in terms of classification. It operates the feature maps fed from the neck, finally outputting class labels within the image.

The designed YOLOv11 architecture, represented in Fig. 5, consists of three main stages: Backbone, Neck, and Head. The current architecture starts with the backbone performing two CLs, then applying C3k2 and C2PSA modules that gradually extract hierarchical features, integrating cross-stage connections and parallel spatial attention to achieve more detailed salient information representation. Then comes the neck stage where more C3k2 blocks and Up-sample and Concatenation operations are utilized for multi-scale feature fusion, enhancing spatial context through the combination of shallow and deep feature levels. Finally, the Head stage performs classification tasks – with the refined features translated into task-dependent outputs like class probabilities, bounding box predictions, or category names. Such modular design of YOLOv11 enables the central unit of the designed approach to balance between high accuracy and computational speed, making it suitable for skin cancer classification applications.

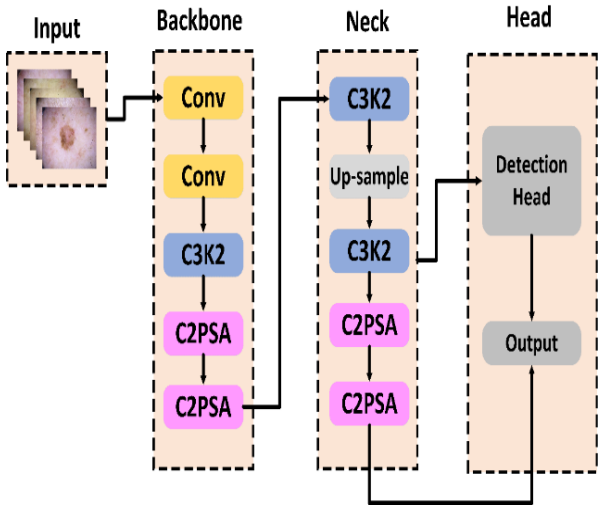

Figure 5. Proposed YOLOv11 architecture.

3.7 Training Setup and Implementation Details

All experiments were conducted under a unified protocol to ensure a fair comparison across models. Specifically, all methods used an input resolution of 224×224, a fixed random seed (SEED=42), epochs= 100 with batch size= 16, and early stopping at 7. The optimization function is "Adam" with a base learning rate was set to 1×10^{-3} across all implementations. All experiments were performed on a personal computer with Windows 11 Pro (64-bit) using an Intel Core i3-3217U CPU @1.8 GHz and 8 GB of RAM. Baselines (ANN/CNN/RNN/LSTM) were developed in TensorFlow/Keras, whereas the YOLOv11 classifier was trained using the Ultralytics (PyTorch-based) framework with corresponding package versions presented in the revised manuscript.

3.8 Evaluation Metrics

The performance of the presented YOLOv11 model is evaluated by using the primary criteria, including precision, F1-score, recall, and accuracy. These possess a more comprehensive understanding of how the model effectively distinguishes between SC. Table 1 shows these metrics with a mathematical presentation [35, 36].

Table 1. Evaluation metrics

Metric	Mathematical equation
Accuracy	$= \frac{TP + TN}{TN + FN + TP + FP}$
Precision	$= \frac{TP}{FP + TP}$
Recall	$= \frac{TP}{FN + TP}$
F1-score	$= \frac{2 \times (Recall * Precision)}{Recall + Precision}$

Where TP is True Positive, FN is False Negative, TN is True Negative, and FN is False Negative.

4. Results and Discussion

The experimental evaluation indicates the comparative performance of the proposed YOLOv11 classification approach with some baseline architectures, specifically (RNN, LSTM, ANN, and CNN). Tables (2) present the results of five models on the PROVe-AI dataset before preprocessing and OFA. The results before preprocessing and OFA in Table 2 are highly diverse: RNN performs with weak accuracy=25.27%, then LSTM reaches moderate performance improvement, accuracy=48.35%. ANN achieves fairly good precision 80.05% but low accuracy =36.26, suggesting the conservative prediction and missing positives in the ANN technique. CNN has a more competitive baseline accuracy=59.34. The optimal baseline model is YOLOv11, which achieves accuracy=82.41. After preprocessing and augmentation shown in Table 3, the RNN baseline shows the weakest performance with an accuracy of 54.71% and a relatively low F1-score of 59.3%, demonstrating its limited ability to capture discriminative features in skin lesion images. LSTM and ANN provided sensible improvements reaching accuracies of 64.41% and 65.74%, respectively. CNN, as expected, performed better due to its ability to extract spatial patterns, achieving 69.26% accuracy.

While the YOLOv11 approach significantly outperformed all baselines, achieving 93.09% accuracy, along with balanced precision, F1-score, and recall values for each. This performance reflects both robust sensitivity to true positives (recall) and high specificity in avoiding false positives (precision), confirming the efficiency of the proposed pipeline that integrated preprocessing, augmentation, and a balanced training strategy.

Table 2. A Comparative performance before preprocessing and OFA of different models on the PROVe-AI dataset

Models	Accuracy	Precision	Recall	F1-score
RNN	25.27	66.25	25.27	20.03
LSTM	48.35	66.3	48.35	53.56
ANN	36.26	80.05	36.26	35.83
CNN	59.34	71.66	59.34	63.32
YOLOv11	82.41	80.2	82.42	80.36

Table 3. A Comparative performance after preprocessing and OFA of different models on the PROVe-AI dataset

Models	Accuracy	Precision	Recall	F1-score
RNN	80.22	64.35	80.22	71.42
LSTM	80.22	64.35	80.22	71.42
ANN	78.02	64	78.02	70.32
CNN	81.32	78.15	81.32	77.86

YOLO v11	89.01	88.64	89.01	88.76

Figs. 6 to 10 show ANN, CNN, LSTM, RNN, and YOLOv11 before preprocessing and OFA. While Figs. 11 to 15 show ANN, CNN, LSTM, and RNN confusion matrices after preprocessing and OFA with frequent misclassifications. In difference, Fig. 10 highlights YOLOv11's robustness, achieving strong diagonal dominance, minimal false predictions, and balanced recognition of benign and malignant lesions, confirming superior generalization.

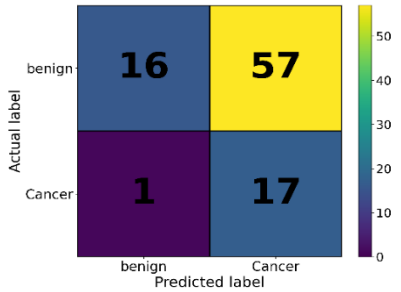

Figure 6. Confusion matrix of ANN before preprocessing and OFA

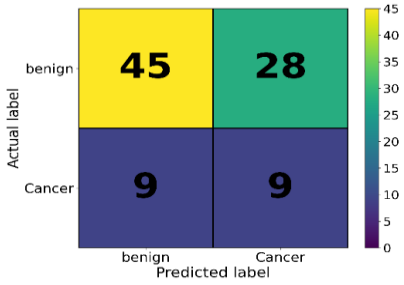

Figure 7. Confusion matrix of CNN before preprocessing and OFA

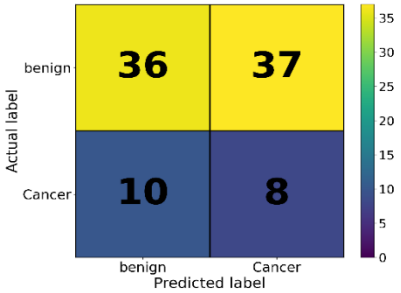

Figure 8. Confusion matrix of LSTM before preprocessing and OFA

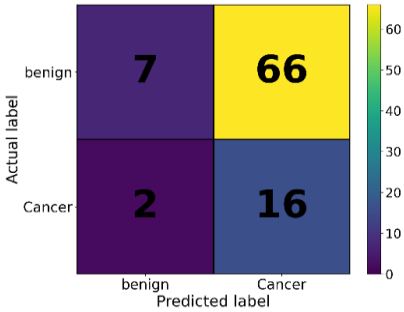

Figure 9. Confusion matrix of RNN before preprocessing and OFA

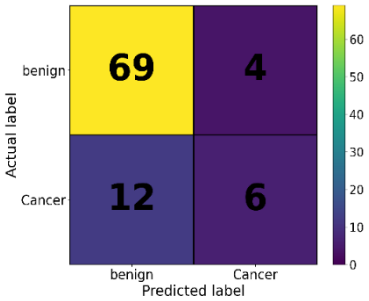

Figure 10. Confusion matrix of YOLOv11 before preprocessing and OFA

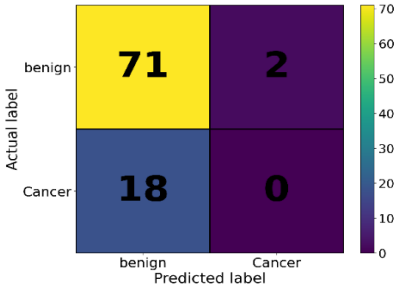

Figure 11. Confusion matrix of ANN after preprocessing and OFA

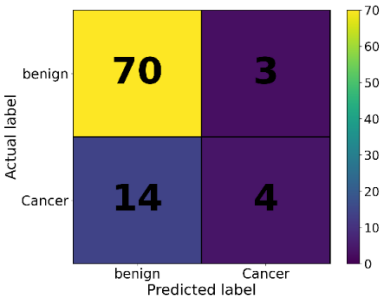

Figure 12. Confusion matrix of CNN after preprocessing and OFA

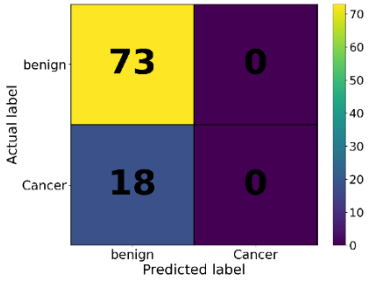

Figure 13. Confusion matrix of LSTM after preprocessing and OFA

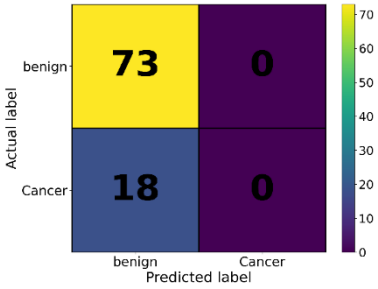

Figure 14. Confusion matrix of RNN after preprocessing and OFA

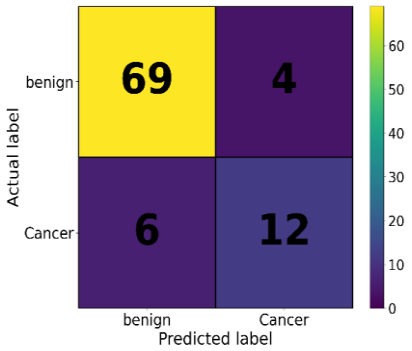

Figure 15. Confusion matrix of YOLOv11 after preprocessing and OFA

Table 4 shows compared to modern works the proposed YOLOv11 framework achieved 89.01% accuracy, outperforming several state-of-the-art approaches. For example, Pandey et al. [12] sparse dictionary learning with CNN (85.61%), Diptho et al. [13] YOLOv11 (80.72%), Huang et al. [14] YOLOv8 (72.6%), and Asriani et al. [18] ResNet-50 (83.84%). It also surpassed Mondal et al. [19] hybrid CNN–GoogleNet–MobileNet–FFNN (87.9%) and Erbay et al. [15] weighted soft voting ensemble (89%). These comparisons emphasize that the optimized YOLOv11 not only outperforms standard CNN- and ensemble-based methods but also achieves accuracy on par with the best hybrid attention-driven architectures, establishing its effectiveness for reliable clinical skin cancer classification.

Table 4. A comparative performance of the proposed YOLOv11 with state-of-the-art studies

Ref/ year	Dataset	DL techniques	Accuracy
[12] 2024	HAM-10000 and ISIC-2019	Sparse Dictionary Learning + CNN with NLM Denoising	85.61
[13] 2025	Kaggle – Skin Disease Classification Image Dataset (900 images, 9 categories)	YOLOv11	80.72
		YOLOv8	79.51
		ResNet50	79.07
		VGG16	72.09
[14] 2025	ISIC – 2203 images	YOLOv11	57.9
		YOLOv10	70.06
		YOLOv9	65.9
		YOLOv8	72.6
		YOLOv5	67.4
[15] 2025	HAM10000	DenseNet201	85
		Inception-ResNetV2	88
		ViT	79
		Ensemble – Soft Voting	88
		Ensemble – Weighted Soft Voting	89
[17] 2025	HAM10000	CNN	73
		GoogleNet	71
		ResNet	47
[18] 2025	HAM10000	ResNet-50	83.84
[19] 2025	Kaggle Skin Cancer Dataset (13,879 images; benign & malignant, dermatoscopic and camera images)	Hybrid CNN using GoogleNet + MobileNet for feature extraction + Feedforward Neural Network (FFNN)	87.9
Proposed 2025	PROVe-AI	RNN	80.22
		LSTM	80.22
		ANN	78.02
		CNN	81.32
		YOLOv11	89.01

From the findings described above, the proposed YOLOv11 framework emerges as a robust and viable automated SC classification solution. The integration of OFA reasonably solved the class-imbalance problem, meaning that the model did not learn to overlook malignant lesions by being taught on the majority class. The three-stage preprocessing pipeline made lesions more visible and helped lower other disruptive artifacts, allowing the model to focus on more clinically relevant morphologies. And finally, the architectural strengths of YOLOv11's C3k2 and particularly C2PSA modules enabled effective multifocus pattern-based feature extraction and spatial attention, leading to much higher accuracy, precision, and recall compared to baseline DL approaches. These features make the optimized YOLOv11 model a promising decision support tool that one could rely upon for any early melanoma diagnosis for SC disease. While the new model thus demonstrated promise, it has some limitations that warrant consideration. First, owing to the relatively limited size of the datasets and their specific origins in the clinical source they came from, which limits their generalization to the broadest possible populations or device and clinical testing conditions. Second, the OFA without synthetic data, although balancing, does not fully seem to capture the heterogeneous extent of real-world SC lesions. Finally, Yeah, the computational efficiency, though already high for YOLOv11 on a GPU, should also have involved a clinical center with multiple dermoscopy devices for fully comprehensive validation. These considerations and the positive results clearly show that should be among the next steps for such frameworks.

5. Conclusion

This study introduced an DL framework for SC classification using enhanced YOLOv11 architecture, OFA, and a flexible preprocessing pipeline. The suggested strategy obtained an impressive accuracy of 93.09%, which outperforms regular CNN, RNN, and several state-of-the-art methods. The combination of hair removal, color constancy, and CLAHE dramatically improved lesion visibility, and the architectural updates of YOLOv11, C3k2, and C2PSA modules ensured modifiers-free multi-scale feature presentation and spatial consideration, respectively. Combined, these advancements enabled unprecedented generalization between benign and malignant lesions, and confirmed the model's capacity to function as a support for future clinical decision support after further prospective validation and deployment-oriented testing model for early melanoma recognition. Despite the excellent results obtained from YOLOv11's proposed framework, future work will extend the benchmarking to ResNet-50, DenseNet-121, EfficientNet, and transformer-based models (ViT/Swin) under the same PROVe-AI split and evaluation protocol. It will include a standard YOLOv11 baseline without preprocessing to quantify the incremental benefit of the proposed pipeline. Also, we focus on dataset dimensions and diversity, multi-class classification, and generative augmentation for more natural variability. Adding explainable AI design and optimizing deployment at a larger scale that includes various imaging devices and settings to ensure scalability, real-time operation, and integration into routine dermatology practice.

6. References

[1] B. Ozdemir and I. Pacal, "An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms," Results Eng., vol. 25, p. 103692, 2025.
https://doi.org/10.1016/j.rineng.2024.103692

[2] H. J. Mohammed, A. A. Nafea, H. K. Almulla, S. A. S. Aliesawi, and M. M. Al-Ani, "An effective hybrid model for skin cancer detection using transfer learning," in Proc. 2023 16th Int. Conf. Developments eSystems Eng. (DeSE), IEEE, 2023, pp. 840-845.
https://doi.org/10.1109/DeSE60595.2023.10468994

[3] R. Karthik, R. Menaka, S. Atre, J. Cho, and S. V. Easwaramoorthy, "A hybrid deep learning approach for skin cancer classification using Swin Transformer and dense group shuffle non-local attention network," IEEE Access, 2024.
https://doi.org/10.1109/ACCESS.2024.3485507

[4] A. Kumar, A. Vishwakarma, V. Bajaj, and S. Mishra, "Novel mixed domain hand-crafted features for skin disease recognition using multiheaded CNN," IEEE Trans. Instrum. Meas., vol. 73, pp. 1-13, 2024.
https://doi.org/10.1109/TIM.2024.3370772

[5] H. L. Gururaj, N. Manju, A. Nagarjun, V. N. M. Aradhya, and F. Flammini, "DeepSkin: a deep learning approach for skin cancer classification," IEEE Access, vol. 11, pp. 50205-50214, 2023.
https://doi.org/10.1109/ACCESS.2023.3274848

[6] V. S. S. B. T. Sathvika et al., "Pipelined structure in the classification of skin lesions based on AlexNet CNN and SVM model with bi-sectional texture features," IEEE Access, vol. 12, pp. 57366-57380, 2024.
https://doi.org/10.1109/ACCESS.2024.3387533

[7] I. F. Jassam, A. A. Mukhlif, A. A. Nafea, M. A. Tharthar, and A. I. Khudhair, "A review of breast cancer histological image classification: challenges and limitations," Iraqi J. Comput. Sci. Math., vol. 6, no. 1, p. 1, 2025.
https://doi.org/10.52866/2788-7421.1232

[8] A. A. Nafea, A.-M. Manar, K. M. A. Alheeti, M. S. I. Alsumaidaie, and M. M. Al-Ani, "A hybrid method of 1D-CNN and machine learning algorithms for breast cancer detection," Baghdad Sci. J., vol. 21, no. 10, pp. 3333-3343, 2024.
https://doi.org/10.21123/bsj.2024.9443

[9] Q. Yao, D. Zhuang, Y. Feng, Y. Wang, and J. Liu, "Accurate detection of brain tumor lesions from medical images based on improved YOLOv8 algorithm," IEEE Access, 2024.
https://doi.org/10.1109/ACCESS.2024.3472039

[10] Y. Lin, C. Dong, H. Long, J. Liang, M. Liu, and Y. Liu, "Algorithm research for melanoma detection based on improved YOLOv11," in Proc. 2025 10th Int. Conf. Intell. Comput. Signal Process. (ICSP), IEEE, 2025, pp. 557-560.
https://doi.org/10.1109/ICSP65755.2025.11086968

[11] A. T. Ibrahim, M. Abdullahi, A. F. D. Kana, M. T. Mohammed, and I. H. Hassan, "Categorical classification of skin cancer using a weighted ensemble of transfer learning with test time augmentation," Data Sci. Manag., vol. 8, no. 2, pp. 174-184, 2025.
https://doi.org/10.1016/j.dsm.2024.10.002

[12] A. Pandey, M. S. Teja, P. Sahare, V. Kamble, M. Parate, and M. F. Hashmi, "Skin cancer classification using non-local means denoising and sparse dictionary learning based CNN," J. Electr. Syst. Inf. Technol., vol. 11, no. 1, p. 36, 2024.
https://doi.org/10.1186/s43067-024-00162-0

[13] R. A. Diptho and S. Basak, "Enhancing dermatological diagnosis through medical image analysis: how effective is YOLO11 compared to leading CNN models?," NDT, vol. 3, no. 2, p. 11, 2025.
https://doi.org/10.3390/ndt3020011

[14] N.-C. Huang, A. Mukundan, R. Karmakar, S. Syna, W.-Y. Chang, and H.-C. Wang, "Novel snapshot-based hyperspectral conversion for dermatological lesion detection via YOLO object detection models," Bioengineering, vol. 12, no. 7, p. 714, 2025.
https://doi.org/10.3390/bioengineering12070714

[15] H. Erbay, Y. M. Abulgasim, D. E. Özer, and F. Ertürk, "Enhancing multi-class skin lesion diagnosis through ensemble learning of CNN and transformer architectures," Eng. Sci. Technol. Int. J., vol. 70, p. 102145, 2025.
https://doi.org/10.1016/j.jestch.2025.102145

[16] A. Shaik, S. S. Dutta, I. M. Sawant, S. Kumar, A. Balasundaram, and K. De, "An attention-based hybrid approach using CNN and BiLSTM for improved skin lesion classification," Sci. Rep., vol. 15, no. 1, p. 15680, 2025.
https://doi.org/10.1038/s41598-025-00025-2

[17] P. Nanda, D. Rout, and S. Kumari, "Multi-class skin cancer detection using CNN-architecture based deep learning models," Procedia Comput. Sci., vol. 260, pp. 226-235, 2025.
https://doi.org/10.1016/j.procs.2025.03.197

[18] A. Asriani, N. T. Lapatta, D. W. Nugraha, A. Amriana, and W. Wirdayanti, "Implementation of ResNet-50-based convolutional neural network for mobile skin cancer classification," J. Appl. Informatics Comput., vol. 9, no. 4, pp. 1577-1969, 2025.
https://doi.org/10.30871/jaic.v9i4.9696

[19] S. Mondal, R. Bonthagorla, H. Kotapati, H. Tummala, and D. Khatua, "Deep learning for skin cancer: hybrid feature extraction with GoogleNet and MobileNet," in Proc. 2025 Int. Conf. Adv. Modern Age Technol. Health Eng. Sci. (AMATHE), IEEE, 2025, pp. 1-6.
https://doi.org/10.1109/AMATHE65477.2025.11080891

[20] U. Ilyas, M. Fuzail, M. K. Abid, T. F. Khan, A. Naeem, and N. Aslam, "Multiclass skin cancer classification using ResNet and dermoscopic imaging," Spectr. Eng. Sci., vol. 3, no. 6, pp. 671-685, 2025.

[21] M. A. Marchetti et al., "Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study)," NPJ Digit. Med., vol. 6, no. 1, p. 127, 2023.
https://doi.org/10.1038/s41746-023-00872-1

[22] R. Kasmi et al., "SharpRazor: automatic removal of hair and ruler marks from dermoscopy images," Skin Res. Technol., vol. 29, no. 4, p. e13203, 2023.
https://doi.org/10.1111/srt.13203

[23] X. Shen, L. Wei, and S. Tang, "Dermoscopic image classification method using an ensemble of fine-tuned

convolutional neural networks," Sensors, vol. 22, no. 11, p. 4147, 2022.
https://doi.org/10.3390/s22114147
[24] S. Barın and G. E. Güraksın, "An automatic skin lesion segmentation system with hybrid FCN-ResAlexNet," Eng. Sci. Technol. Int. J., vol. 34, p. 101174, 2022.
https://doi.org/10.1016/j.jestch.2022.101174
[25] S. Joseph and O. O. Olugbara, "Preprocessing effects on performance of skin lesion saliency segmentation," Diagnostics, vol. 12, no. 2, p. 344, 2022.
https://doi.org/10.3390/diagnostics12020344
[26] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in Proc. Int. Workshop Comput.-Assisted Robot. Endoscopy, Springer, 2018, pp. 303-311. https://doi.org/10.1007/978-3-030-01201-4_33
[27] M. Alsaidi, M. T. Jan, A. Altaher, H. Zhuang, and X. Zhu, "Tackling the class imbalanced dermoscopic image classification using data augmentation and GAN," Multimed. Tools Appl., vol. 83, no. 16, pp. 49121-49147, 2024.
https://doi.org/10.1007/s11042-023-17067-1
[28] A. Shah et al., "A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN)," Clin. eHealth, vol. 6, pp. 76-84, 2023.
https://doi.org/10.1016/j.ceh.2023.08.002
[29] A. Ashtikar and S. Shastri, "A CNN model for skin cancer detection and classification by using image processing techniques," J. Sci. Res. Technol., pp. 251-263, 2025.
[30] S. S. Zareen, G. Sun, M. Kundi, S. F. Qadri, and S. Qadri, "Enhancing skin cancer diagnosis with deep learning: a hybrid CNN-RNN approach," Comput. Mater. Contin., vol. 79, no. 1, 2024.
https://doi.org/10.32604/cmc.2024.047418
[31] I. R. Putri, A. Amalia, and T. F. Abidin, "Hybrid algorithm using CNN and LSTM to improve performance in skin lesion image classification," in Proc. 2024 Ninth Int. Conf. Informatics Comput. (ICIC), IEEE, 2024, pp. 1-6.
https://doi.org/10.1109/ICIC64337.2024.10956620
[32] H. K. Verma, N. D. Londhe, A. Kumar, and R. S. Sonawane, "YOLOv11 for psoriasis classification: enhancing feature extraction for accurate severity differentiation," in Proc. 2025 Int. Conf. Electron., AI Comput. (EAIC), IEEE, 2025, pp. 1-6.
https://doi.org/10.1109/EAIC66483.2025.11101292
[33] R. Khanam and M. Hussain, "YOLOv11: an overview of the key architectural enhancements," arXiv Prepr. arXiv2410.17725, 2024.
[34] Q. Zhao and J. Zhu, "An improved YOLOv11 architecture with multi-scale attention and spatial fusion for fine-grained residual detection," Results Eng., p. 107061, 2025.
https://doi.org/10.1016/j.rineng.2025.107061
[35] A. K. Kareem and K. M. A. Alheeti, "Hybrid approach for fall detection based on machine learning," in Proc. Int. Conf. New Trends Inf. Commun. Technol. Appl., Springer, 2021, pp. 111-130.
https://doi.org/10.1007/978-3-030-93417-0_8
[36] K. M. A. Alheeti, A. Alzahrani, M. Alamri, A. K. Kareem, and D. Al_Dosary, "A comparative study for SDN security based on machine learning," Int. J. Interact. Mob. Technol., vol. 17, no. 11, 2023.
https://doi.org/10.3991/ijim.v17i11.39065