




Leveraging Stacked Ensemble Meta-Learning and Deep Neural Networks for Improved Skin Cancer Diagnosis

Abdulmajeed Alsufyani 

Department of Computer Science, College of Computers and Information Technology, Taif University,
Taif 21944, Saudi Arabia

Corresponding Author Email: a.s.alsufyani@tu.edu.sa

Copyright: ©2025 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license
(<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420426>

ABSTRACT

Received: 13 May 2025
Revised: 10 June 2025
Accepted: 16 June 2025
Available online: 14 August 2025

Keywords:

*deep neural networks, dermoscopic images,
meta-learning, skin cancer diagnosis,
stacked ensemble*

Skin cancer poses a major global health threat, with melanoma being one of the deadliest forms due to its rapid progression and high mortality rate. Timely and precise detection is crucial for enhancing patient survival rates. In this study, we propose a robust stacked ensemble meta-learning model for the classification of various skin cancer types using dermoscopic images. The proposed framework integrates four convolutional neural networks (CNNs) such as custom CNN, InceptionResNetV2, ResNet101V2, and DenseNet201 as base learners, and leverages five meta-learners: Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, and XGBoost for final classification. The system is trained and fine-tuned on a curated subset of the ISIC 2020 dataset, consisting of 2,357 images across nine skin lesion categories. Comprehensive experiments demonstrate the superior performance of the ensemble approach, achieving an accuracy of 98.63%, with both precision and recall reaching 98.64%. The Random Forest-based ensemble emerges as the top-performing configuration. Additionally, the study provides a comparative evaluation with existing methods and highlights the clinical potential of the model in reducing false positives and false negatives. By leveraging transfer learning, data augmentation, and meta-learning strategies, this work contributes a scalable and accurate diagnostic tool for skin cancer detection, especially suitable for deployment in primary care and resource-limited healthcare settings.

1. INTRODUCTION

Skin cancer represents a significant global health issue, affecting millions of individuals each year. According to the World Health Organization (WHO), the incidence of skin cancer continues to rise, particularly in regions with high ultraviolet (UV) radiation exposure [1, 2]. From a clinical perspective, skin cancers are commonly divided into malignant and benign types. Melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC) are the primary malignant skin cancers, varying in their severity and progression. Among these, melanoma is the most life-threatening, as it originates in melanocytes and has a high potential for metastasis if not detected early [3]. Benign lesions, such as seborrheic keratosis and nevi (moles), are non-cancerous but often resemble malignant growth, making accurate diagnosis challenging [4].

The survival rate for skin cancer is highly dependent on early detection. For instance, patients diagnosed at an early stage of melanoma have a five-year survival rate nearing 99%, but this drastically declines in advanced stages [3]. Traditional diagnostic methods include visual inspections by dermatologists and biopsies. While effective, these methods are often subjective, time-consuming, and invasive [5]. Dermatoscopic analysis relying on visual clues like lesion

color, symmetry, and texture can be affected by human error and inconsistency [6]. Moreover, in cases where visual diagnosis is inconclusive, a biopsy is required, which may involve unnecessary surgical procedures, added healthcare costs, and patient discomfort. A major diagnostic challenge is the visual similarity between benign and malignant lesions. Benign conditions such as seborrheic keratosis and nevi can exhibit irregular pigmentation or asymmetry similar to melanoma, leading to frequent misclassification [2]. This increases the risk of unnecessary biopsies (false positives) or, worse, missed cancer diagnoses (false negatives). Additionally, a shortage of expert dermatologists, particularly in rural or under-resourced regions, further impedes timely diagnosis and intervention.

There is a growing need for accurate, accessible, and non-invasive tools that can help diagnose skin cancer early, especially in places where specialist care is limited. In many rural and under-resourced areas, the lack of trained dermatologists often leads to higher rates of missed diagnoses, with some patients not receiving proper care until it's too late. Traditional methods like biopsies or visual checks by experts aren't always practical in these settings—they take time, require expert judgment, and can be costly or uncomfortable for patients. That's where our approach comes in. The stacked ensemble deep learning model we propose is built to reduce

false negatives and improve diagnostic accuracy, even when clinical resources are scarce. By combining the strengths of multiple deep learning models and using advanced meta-classifiers, our system is well-suited for use in mobile apps or cloud platforms—making it easier to deliver reliable diagnosis support to people who need it most.

Given these limitations, AI-assisted diagnostic tools, particularly those utilizing deep learning, have shown great promise as non-invasive, accurate, and scalable solutions for skin cancer screening [7]. In particular, convolutional Neural Networks (CNNs) have demonstrated outstanding performance in the field of medical image analysis, especially in identifying subtle spatial patterns in dermoscopic images without human intervention. CNNs reduce diagnostic subjectivity, improve classification accuracy, and are capable of detecting malignancy-indicative patterns with high precision [3, 5]. To enhance CNN capabilities, numerous architectures have been developed. ResNet introduced residual learning to address the vanishing gradient problem and enabled deeper networks to distinguish between malignant and benign lesions more effectively [3]. EfficientNet and MobileNet optimize for accuracy and computational efficiency, making them suitable for deployment in mobile and low-resource settings [2]. In parallel, Vision Transformers (ViTs) provide a novel approach by analyzing images as sequences of patches, improving long-range pattern recognition critical in skin lesion classification [8].

While individual CNN models have demonstrated strong performance, combining multiple models through ensemble learning, specifically stacked ensemble frameworks, can further improve robustness and predictive accuracy. This approach harnesses the complementary strengths of multiple architectures while reducing the limitations inherent in individual models. Moreover, deep learning systems can be deployed in cloud-based platforms or mobile applications, expanding access to high-quality diagnostic tools in remote or underserved areas [2].

In this work, a stacked ensemble CNN-based approach is proposed for the accurate classification of multiple skin cancer types, incorporating four deep learning models: Custom CNN, InceptionResNetV2, ResNet101V2, and DenseNet201. These models are fine-tuned using transfer learning on a curated ISIC 2020 dataset. For the final classification, we evaluate five meta-classifiers: Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost to combine the prediction outputs of the base CNNs. The dataset used contains 2,357 dermoscopic images from nine classes of both benign and malignant lesions, split into training, validation, and test sets. Our results demonstrate high effectiveness with accuracy, precision, and recall all exceeding 98%, confirming the model's potential for aiding early skin cancer detection. In summary, the key contributions of this paper are:

- We propose a stacked ensemble deep learning framework to classify benign and malignant skin lesions from dermoscopic images with high accuracy.
- We integrate four CNN-based base models and evaluate five different meta-classifiers to determine the optimal ensemble configuration to enhance diagnostic accuracy and reduce false negatives and false positives.
- We utilize transfer learning and fine-tuning techniques to enable learning from a limited dataset while maintaining generalization.

- A comprehensive performance evaluation is conducted using a curated subset of the ISIC 2020 dataset, showing the model's robustness across multiple lesion types.

- We provide a comparative analysis with existing state-of-the-art methods, highlighting the advantages of our approach in terms of diagnostic performance.

This study contributes not only to the development of accurate AI models but also to bridging the gap between artificial intelligence and clinical dermatology. By offering a scalable and interpretable solution for early skin cancer detection, the proposed framework has the potential to improve diagnostic workflows, reduce human error, and enhance patient outcomes, especially in resource-limited healthcare environments [9].

The following sections outline the structure of this study: In Section 2, we present an overview of previous studies focused on deep learning-based approaches for skin cancer detection. Section 3 outlines the design of the proposed system, highlighting the architecture and training procedures of the model. In Section 4, we report the experimental findings, assess model performance, and provide a comprehensive discussion. Lastly, Section 5 provides the conclusion of the study.

2. RELATED STUDIES

In medical image analysis, deep convolutional neural networks have proven highly effective and widely adopted, particularly in classification and object detection tasks [10, 11]. The potential of machine learning and deep learning in early skin cancer detection has been widely studied, with a focus on improving diagnostic outcomes and assisting dermatologists in making informed clinical decisions. Automated image analysis techniques enable these approaches to classify skin lesions as malignant or benign, providing an efficient and scalable alternative to standard diagnostic methods. This section provides an overview of the latest advancements in deep learning, ensemble learning, and meta-learning strategies for skin cancer detection.

Many studies have leveraged pre-trained models like XceptionNet, EfficientNetV2, and ResNet for skin lesion classification. For example, Thapar and Tiwari [12] employed XceptionNet for skin lesion classification, achieving an accuracy of 88.72% and highlighting its effectiveness in medical image analysis. The authors employed multiple advanced deep learning architectures for the diagnosis of skin cancer, specifically leveraging architectures like XceptionNet, EfficientNetV2S, InceptionResNetV2, and EfficientNetV2M. While their work emphasizes the integration of explainable AI (XAI) to clarify model decision-making, it does not specifically address stacked ensemble meta-learning. With XceptionNet reaching an accuracy of 88.72%, the study illustrates how such technologies can elevate diagnostic accuracy, facilitate clinical workflows, and positively impact patient outcomes. Ezeddin et al. [13] presented a method which highlights the optimization of deep ensemble learning for melanoma skin cancer classification, utilizing state-of-the-art convolutional neural networks (CNNs) like ResNet101 and ResNext101. By employing a weighted averaging ensemble approach, the study achieved a classification accuracy of 96.12%, demonstrating significant advancements in detection methods. This integrated architecture enhances diagnostic

precision in medical imaging, indicating its effectiveness for early clinical detection and management of melanoma, with the potential to significantly improve patient care.

A max voting ensemble method for skin cancer classification was proposed by Hossain et al. [14], utilizing the combined strengths of MobileNetV2, VGG16, and ResNet50. With an accuracy of 93.18% and an AUC of 0.9320 on the ISIC 2018 dataset, the ensemble model significantly boosted diagnostic accuracy. This approach leverages multiple models to aid clinicians in making precise and prompt diagnostic decisions, ultimately contributing to improved patient care. It may not always capture the nuances of complex cases where a single model might provide a more accurate prediction than the majority, potentially leading to suboptimal outcomes in certain scenarios.

Maurya et al. [15] introduced a hybrid model combining topological data analysis (TDA) and deep learning (DL) for basal cell carcinoma (BCC) diagnosis, achieving 97.4% accuracy and an AUC of 0.995. Persistent homology is utilized to capture complex topological patterns from telangiectasia and skin lesions, contributing to improved performance of deep learning models. While it focuses on BCC, the integration of TDA with DL represents a significant technological advancement in skin cancer detection. This study utilizes a dataset of 395 skin lesion images, with a focus on telangiectasia features that are processed through automated segmentation and analyzed using both topological data analysis (TDA) and deep learning techniques. A larger dataset could enhance the robustness and generalizability of the model. A limited dataset may not capture the full variability of BCC presentations in different populations. In their study, Daghrir et al. [16] introduced a hybrid strategy that merges deep learning and classical machine learning techniques to improve the accuracy of skin cancer diagnosis. The approach integrates a CNN for extracting deep features automatically from dermoscopic images. Simultaneously, KNN and SVM classifiers were trained on manually crafted features describing the lesions' texture, shape, and color. To improve classification accuracy, the outputs of these models are aggregated using a majority voting approach. The experiment was carried out on a publicly accessible ISIC dataset, where the authors selected 640 representative images, both benign and malignant, from a total of 23,000 melanoma samples. Experimental results indicate that CNN achieved an accuracy of 85.5%, while SVM and KNN reached 71.8% and 57.3%, respectively. The ensemble model, leveraging majority voting, further improved the accuracy to 88.4%. However, the study highlights key challenges, including a limited amount of labelled training data, which the authors suggest addressing through semi-supervised learning in future work. To enhance melanoma classification, Filali et al. [17] later proposed a method that integrates handcrafted features with deep representations extracted from pre-trained CNNs. The proposed approach leverages feature-level fusion to combine the discriminative power of both feature types, aiming to improve classification accuracy. To ensure robustness, the model was evaluated on datasets of different sizes, namely the PH2 and the ISIC challenge datasets. The fusion approach yielded significant improvements in classification accuracy, particularly on the PH2 dataset. The integration of handcrafted and deep features in this study demonstrates a robust and effective approach for melanoma detection. Amin et al. [18] further enhanced skin cancer detection by integrating deep feature fusion techniques for both localization and

classification. The study employs multiple deep learning models, including AlexNet and VGG-16, to extract important features from segmented skin lesion images. Principal Component Analysis (PCA) is applied to the extracted features to retain the most relevant features, thereby improving classification performance. Deep feature integration, when trained and validated on a curated dataset, proves effective in boosting both the detection accuracy and localization precision of skin cancer.

Devadhas et al. [19] presented a deep learning framework based on stacking ensemble techniques, attaining a 97% accuracy rate in detecting skin cancer. It utilizes a combination of CNN and deep neural networks (DNN) as input for a Long Short-Term Memory (LSTM) meta-classifier. The approach includes preprocessing of skin images, segmentation using Fuzzy-C-Means clustering, and feature extraction via Local Binary Pattern (LBP). The absence of a thorough evaluation limits understanding of the model's performance, which is critical in medical applications where diagnostic errors can have significant consequences. In another study, Chiu et al. [20] developed an AI-driven skin cancer diagnosis model using a two-stage voting ensemble approach, significantly reducing false negatives and improving diagnostic accuracy in both ISIC and CSMUH datasets, with potential to aid resource-limited medical settings. Mary et al. [21] presented an advanced ensemble model combining ResNet, EfficientNet, and MobileNet to enhance skin cancer diagnosis. It addresses limitations of previous models by leveraging the strengths of these CNNs, integrating advanced preprocessing, data augmentation, and transfer learning. With an accuracy of 96.33% on the HAM10000 dataset, the model demonstrates strong potential for precise and early skin cancer detection.

The study by Kaur et al. [22] highlights the implementation of advanced deep learning models tailored for melanoma detection, with particular emphasis on automating the early diagnostic process. The method involves a hybrid pipeline, starting with morphological and context aggregation techniques for preprocessing, and concluding with a segmentation network to extract lesion areas. The classification model achieves a 93.40% accuracy rate using cleaned and segmented images. The BEDLM-CMS model proposed by Shah et al. [23] employs an ensemble of LSTM, BLSTM, and GRU networks to detect mutations responsible for cutaneous melanoma. It addresses challenges like data scarcity and overfitting, achieving an accuracy rate of 97% in independent tests and 94% in tenfold cross-validation. Liu et al. [24] developed a hybrid deep network combining Resnet-50 and CrossViT, which enhances feature representation and diagnostic accuracy through improved classification performance and stable feature fusion. This approach aims to refine early diagnostic accuracy in skin cancer, ultimately contributing to better patient survival rates. Qureshi and Roos [25] addressed the challenge of class imbalance in skin cancer detection by proposing an ensemble CNN model that integrates diverse CNN architectures along with metadata. Using 33,126 dermoscopic images, it combines pre-trained and custom-trained models with an SVM meta-learner, achieving enhanced accuracy. In their work, Rahman et al. [26] utilized anisotropic diffusion filtering to enhance image quality and applied a fast-bounding box (FBB) method for efficient segmentation of skin lesions. It integrates hybrid feature extractor (HOG, LBP, SURF) and VGG19-based CNN for feature fusion. The approach achieves 91.65% sensitivity and 95.70% specificity using ISIC 2017 and the academic

torrents datasets. In a subsequent effort, Hemantkumar et al. [27] focuses on a deep learning feature fusion approach combined with an extreme learning machine (ELM) to classify benign and malignant skin lesions. The study highlights the use of a curated dataset from HAM10000 and Dermis.net, achieving a high accuracy of 98.0% for the validation set.

Additionally, several studies have tackled the challenge of multi-class skin cancer detection using optimized hybrid deep learning models [28-30]. Table 1 provides a summarized comparison of the reviewed literature, highlighting the models used, key findings, and datasets employed in various skin cancer detection studies.

Table 1. Summary of reviewed literature on skin cancer detection methods

| Authors | Models Used | Key Findings | Dataset Used |
|-------------------------|--|---|---|
| Thapar and Tiwari [12] | XceptionNet, EfficientNetV2S, InceptionResNetV2, EfficientNetV2M | XceptionNet achieved 88.72% accuracy, explored XAI for model explainability Weighted averaging ensemble achieved | Curated dataset |
| Ezeddin et al. [13] | ResNet101, ResNext101 (Ensemble) | 96.12% accuracy, improved melanoma classification | Custom dataset |
| Hosaain et al. [14] | MobileNetV2, VGG16, ResNet50 (Max Voting Ensemble) | Achieved 93.18% accuracy, AUC 0.9320, enhanced diagnostic accuracy using ISIC 2018 | ISIC 2018 |
| Maurya et al. [15] | Topological Data Analysis (TDA) + Deep Learning | 97.4% accuracy, AUC 0.995 for basal cell carcinoma (BCC) detection | 395 skin lesion images (Telangiectasia focus) |
| Daghrir et al. [16] | CNN, KNN, SVM (Majority Voting Ensemble) | CNN: 85.5%, SVM: 71.8%, KNN: 57.3%, ensemble improved to 88.4% Feature-level fusion improved | ISIC repository (640 images) |
| Filali et al. [17] | Handcrafted + Pretrained CNN Features | classification, better performance on PH2 dataset | PH2, ISIC Challenge |
| Amin et al. [18] | AlexNet, VGG-16 (Feature Fusion) | PCA-selected feature fusion improved skin cancer detection and localization | Curated dataset |
| Devadhas et al. [19] | Stacking-based CNN + DNN with LSTM meta-classifier | Achieved 97% accuracy but lacked detailed performance evaluation | Custom dataset |
| Chiu et al. [20] | Two-stage Voting Ensemble | Reduced false negatives, improved accuracy in ISIC and CSMUH datasets | ISIC, CSMUH |
| Mary et al. [21] | ResNet, EfficientNet, MobileNet (Ensemble) | 96.33% accuracy on HAM10000, used preprocessing, data augmentation, and transfer learning | HAM10000 |
| Kaur et al. [22] | Deep Neural Networks with Morphological Preprocessing | 93.40% accuracy using segmented images for melanoma detection | Segmented images dataset (private) |
| Shah et al. [23] | LSTM, BLSTM, GRU (BEDLM-CMS) | 97% accuracy in independent tests, 94% in tenfold cross-validation | Custom dataset |
| Liu et al. [24] | ResNet-50, CrossViT (Hybrid Model) | Enhanced feature representation and classification performance for skin cancer | Curated dataset |
| Qureshi and Roos [25] | Ensemble CNN + Metadata with SVM meta-learner | Trained on 33,126 images, improved classification accuracy | 33,126 dermoscopic images |
| Rahman et al. [26] | Hybrid Feature Extractor (HOG, LBP, SURF) + VGG19 | Achieved 91.65% sensitivity, 95.70% specificity on ISIC 2017 & Academic Torrents | ISIC 2017, Academic Torrents |
| Hemantkumar et al. [27] | Deep Learning Feature Fusion + Extreme Learning Machine (ELM) | 98.0% accuracy using HAM10000 and Dermis.net datasets | HAM10000, Dermis.net |

3. MATERIALS AND METHODS

This section details the methodology developed for the detection and classification of skin cancer. Figure 1 outlines the overall structure of the framework, emphasizing the sequential stages of image preprocessing, deep feature extraction, and final classification. The preprocessing stage involves resizing, normalization, and data augmentation of dermoscopic images to ensure consistent input dimensions and improved model generalization. Next, pre-trained CNN models trained on the ImageNet dataset are fine-tuned to adapt to the specific characteristics of skin lesion images, improving feature extraction and classification accuracy. The prediction outputs from the fine-tuned base CNN architectures are aggregated and passed to the meta-learner, which refines the predictions for final classification. As depicted in Figure 2, the proposed stacked ensemble learning framework trains multiple base classifiers and a meta-classifier on the ISIC dataset, which includes images of both benign and malignant

skin lesions. Once training is complete, the ensemble model is employed to classify unseen skin lesion images as either benign or malignant. The following sections provide a detailed explanation of each step in the proposed methodology.

3.1 Skin lesion dataset and pre-processing steps

This study utilizes a subset of the ISIC 2020 challenge dataset [31, 32], which offers a varied collection of dermoscopic images for classifying skin lesions. This dataset includes images from nine distinct categories, covering both benign and malignant skin lesions: actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, pigmented benign keratosis, seborrheic keratosis, squamous cell carcinoma, and vascular lesions. These categories encompass a wide range of skin conditions, ensuring that the model learns to differentiate between malignant and benign cases effectively. Figure 3 presents sample images from the dataset, showcasing various skin lesion categories.

Malignant skin cancers include melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC). Melanoma arises from melanocytes and is characterized by its aggressive nature and strong tendency to metastasize if not detected early. BCC is the most frequently diagnosed skin cancer, known for its slow progression and low likelihood of spreading beyond the epidermis. SCC is more aggressive than BCC and can metastasize if left untreated. On the other hand, benign skin lesions include nevus, pigmented benign keratosis (PBK), seborrheic keratosis (SK), dermatofibroma, actinic keratosis (AK), and vascular lesions. A nevus (mole) is a common benign pigmented lesion. PBK and SK are non-cancerous growths that can resemble melanoma but do not pose a serious threat. Dermatofibromas are firm, fibrous, harmless skin nodules. Vascular lesions encompass benign growths like hemangiomas and angiomas, which result from blood vessel proliferation. Actinic Keratosis (AK) is a precancerous lesion caused by prolonged sun exposure, with

the potential to progress into SCC if untreated. While benign lesions are generally harmless, detecting malignant skin cancers at an early stage plays a vital role in enabling prompt treatment and improving patient survival rates.

This dataset includes 2,357 dermoscopic images of various dimensions, covering a range of benign and malignant skin lesion types. It was divided into three subsets to facilitate effective training and evaluation of the proposed model. Specifically, 1,792 images (approximately 76% of the dataset) were allocated for training, 447 images (around 19%) for validation, and 106 images (about 5%) for testing. Model parameters were learned from the training set, while the validation set supported hyperparameter tuning and helped mitigate overfitting. An independent test set was used at the final stage to objectively assess the model’s performance on unseen samples, ensuring generalization capability. Several preprocessing steps were applied to improve the quality and consistency of the input images.

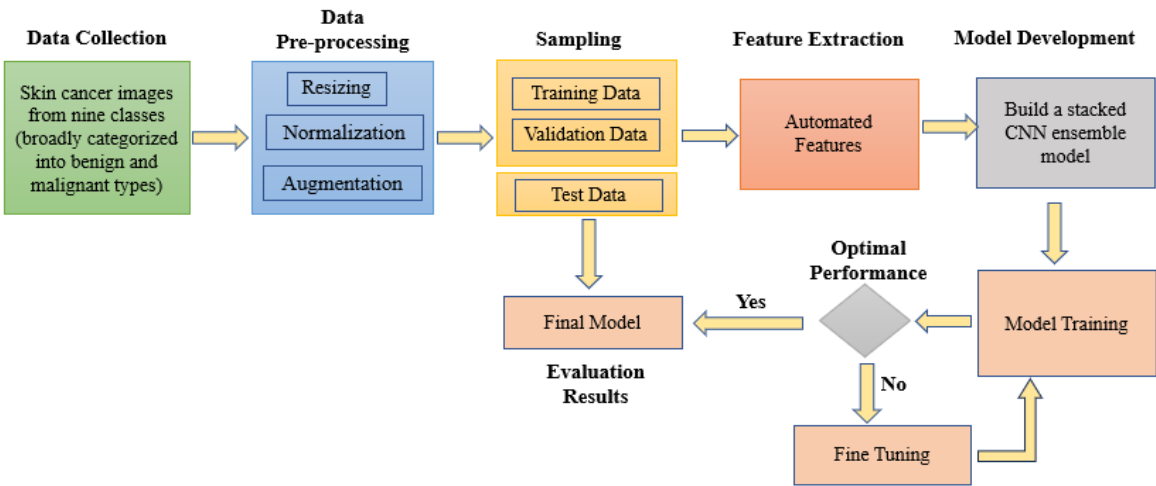


Figure 1. Proposed system design for automatic classification of skin lesions

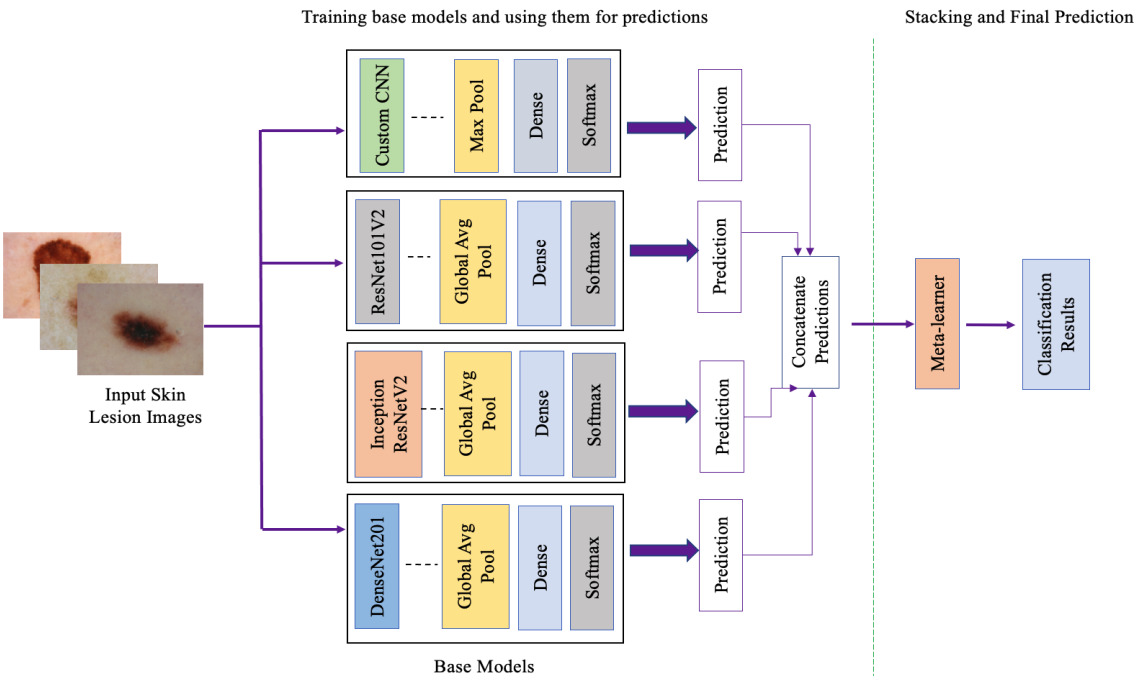


Figure 2. Schematic representation of the deep learning ensemble framework, featuring stacked CNN models and a meta-classifier for improved skin lesion classification

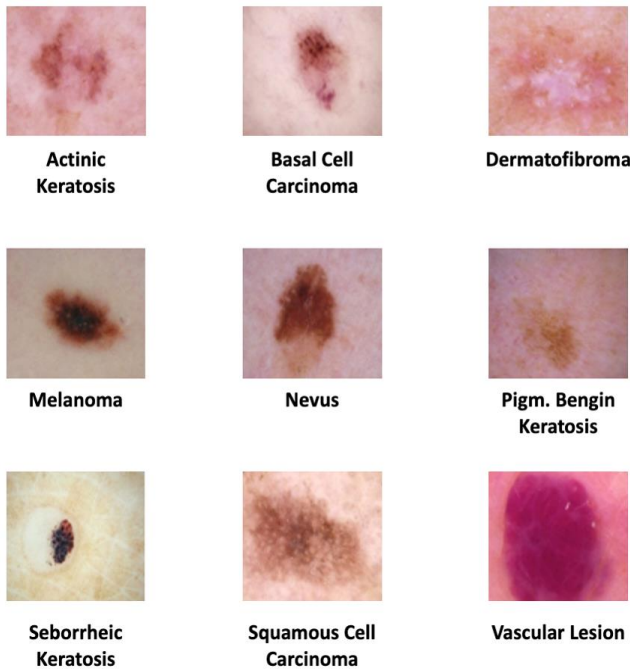


Figure 3. Sample images from the dataset showing benign and malignant skin lesions

Since dermoscopic images come in varying resolutions, all images were resized to a fixed dimension (180 x 180) to standardize input size and optimize CNN performance. To prevent biases during training, the dataset was randomly shuffled before splitting, ensuring a balanced distribution of different lesion classes across training, validation, and testing sets. Additionally, pixel values were normalized to the [0,1] range to ensure consistent feature scaling, thereby stabilizing training and improving convergence speed.

Due to the relatively small size of the dataset, data augmentation was used to expand the training set, thereby promoting better generalization of the model. The augmentation strategies included rotation, flipping, translation, and zooming to introduce variations in image orientation, position, and scale. Rotation ensured the model became invariant to different lesion orientations, while flipping (both horizontal and vertical) introduced diversity in lesion placement. Translation slightly shifted images in different directions, making the model robust to variations in lesion positioning. Zooming simulated different levels of magnification, helping the model handle real-world variations in dermoscopic images. An outline of the augmentation methods used, along with their parameter settings, is provided in Table 2. These preprocessing steps collectively enhanced the model’s ability to generalize effectively, reducing overfitting and improving classification performance across different types of skin lesions.

Table 2. Image augmentation settings

| Method | Amount Value |
|--------------------|-----------------|
| width translation | 0.1 |
| height translation | 0.1 |
| rotation range | 36 |
| shuffle | true |
| zoom range | 0.2 |
| vertical flip | true |
| horizontal flip | true |

3.2 Architecture of the stacked ensemble meta-learner model

As depicted in Figure 2 the proposed ensemble framework utilizes several deep learning models in parallel to improve the accuracy and robustness of automatic skin cancer classification. Four fine-tuned CNN sub-models are first employed to generate prediction probabilities from input skin lesion images: a custom CNN, InceptionResNetV2 [33], ResNet101V2 [34], and DenseNet201 [35], which are stacked together at level-0. These base models extract diverse feature representations, capturing both low-level and high-level patterns crucial for distinguishing different skin lesion types. The prediction outputs from these networks are then aggregated and passed to five different meta-classifiers at level-1, including Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost. The meta-classifiers refine the decision-making process, and for the final prediction, the classifier with the highest performance is chosen to distinguish between different skin cancer classes. A detailed breakdown of each module in the proposed system is provided in the following subsections.

3.2.1 Base (Level 0) CNN architectures and fine-tuning

In the proposed stacking ensemble system, we have utilized one custom CNN model along with three pre-trained CNN architectures—InceptionResNetV2, ResNet101V2, and DenseNet201—which have been fine-tuned to generate level-0 prediction probabilities from input skin lesion images. The custom CNN model, as shown in Figure 4, begins with an input preprocessing layer, where images are normalized using a rescaling layer that scales pixel values to the [0,1] range, ensuring better convergence during training [36, 37]. The network employs a series of convolutional blocks, each consisting of a 2D convolutional layer (Conv2D), a max-pooling layer (MaxPool2D), and, in the later stages, a dropout layer.

The architecture follows a progressive increase in the number of filters, starting with 32 filters in the first convolutional layer and doubling with each subsequent block (64, 128, 256, 512). With increasing depth, the network is able to learn more abstract and complex feature representations. Each Conv2D layer utilizes a 3×3 kernel with 'same' padding to maintain spatial dimensions and ReLU activation for non-linearity. MaxPool2D layers are employed to reduce spatial dimensions and introduce translation invariance. Dropout layers with rates of 0.15, 0.20, and 0.25 are introduced after the third, fourth, and fifth convolutional blocks to prevent overfitting. This progressive increase in the dropout rate helps to regularize the deeper layers of the network. Following the convolutional blocks, a Flatten layer transforms the feature maps into a one-dimensional vector. Following feature extraction, the vector is forwarded to a Dense layer with 1024 neurons, activated using the ReLU function. To complete the classification process, a Dense layer with nine neurons and softmax activation is employed to output class probabilities for the nine skin lesion categories.

InceptionResNetV2 serves as the second base model, combining Inception modules with residual connections to improve the depth and efficiency of feature extraction. The model is fine-tuned for skin lesion classification by removing the original top layers and adding task-specific layers. To improve generalization, the model integrates a preprocessing pipeline with random flips and rotations, followed by rescaling, global average pooling, dropout, and Dense layers.

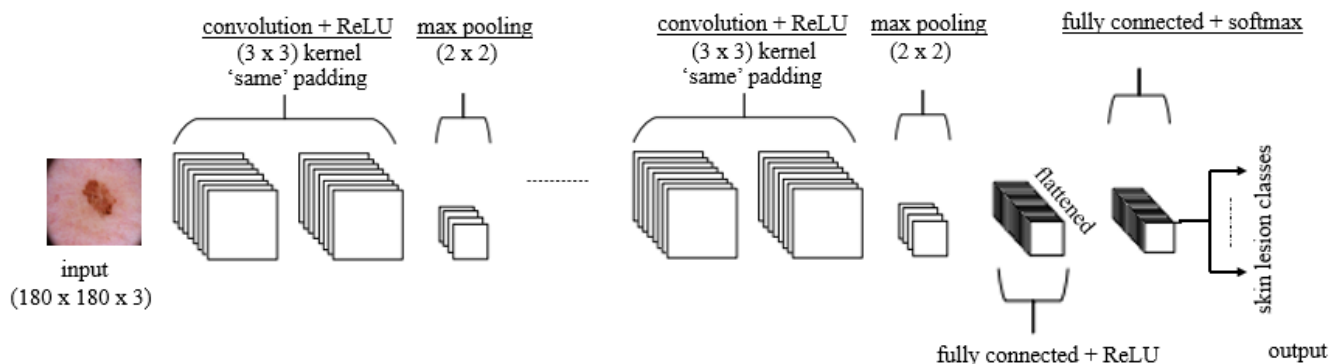


Figure 4. Tailored convolutional neural network (CNN) architecture commonly used for image classification tasks
(The figure is conceptually inspired by standard CNN design [36, 37])

The final softmax output layer assigns probability scores across the nine skin lesion categories. Training is performed in two stages: first, freezing the pre-trained layers for 15 epochs to train the new layers, and second, unfreezing selected deep layers to fine-tune high-level features. Specifically, the first 100 layers remain frozen, while the rest are trainable, ensuring optimal feature adaptation. The *AdamW* optimization algorithm is used during training, paired with the *SparseCategoricalCrossentropy* loss and accuracy as the evaluation criterion to monitor model performance. A reduced learning rate is applied during fine-tuning to prevent disruptive weight updates, effectively leveraging transfer learning to enhance classification performance on the skin cancer dataset.

The third model in the ensemble system is ResNet101V2, a deep residual network known for its efficient feature extraction through skip connections. The base model remains frozen during the initial training phase, followed by global average pooling, dropout layers (0.3), and dense layers (1024 and 512 neurons with ReLU activation). A softmax output layer classifies images into nine skin lesion categories. Again, the model is trained using the *AdamW* optimizer with *SparseCategoricalCrossentropy* loss for 15 epochs while keeping the pre-trained layers frozen.

In the fine-tuning phase, the top 50 layers of ResNet101V2 are unfrozen, allowing deeper feature adaptation to the skin lesion dataset. The model is recompiled with a lower learning rate ($1e-5$) to avoid disrupting previously learned features. Fine-tuning continues for 35 additional epochs, gradually improving classification performance. This two-stage training strategy effectively leverages transfer learning to refine deep feature extraction while preventing overfitting, ensuring a robust and adaptive model for skin cancer detection.

Finally, we use DenseNet201 that enhances feature propagation through dense connections, allowing efficient gradient flow and improved feature reuse. The pre-trained DenseNet201 model is loaded without its top layers and is followed by flattening and dropout layers (0.5) to reduce overfitting. Fully connected dense layers with 1024 and 512 neurons (ReLU activation) further refine feature representation, and a softmax output layer classifies images into nine skin lesion categories. To ensure stable convergence during training, the model leverages the SGD optimizer configured with a 0.001 learning rate and 0.9 momentum.

A learning rate scheduler (*ReduceLROnPlateau*) is implemented to optimize training by reducing the learning rate when validation accuracy shows no improvement over successive epochs. Using a batch size of 32, the model is trained for 50 epochs, incorporating transfer learning and fine-

tuning to effectively repurpose DenseNet201 for classifying skin lesions. This structured training approach improves classification accuracy by refining high-level representations while preventing overfitting.

3.2.2 Level-1 meta classifiers

The Level-1 meta-classifier in the stacking ensemble learns to effectively aggregate predictions from the Level-0 models to enhance classification accuracy. The meta-classifier is trained on the outputs of the base models, leveraging their diverse feature representations and decision patterns to improve final classification accuracy. To build the Level-1 model effectively, we employ five different machine learning classifiers: Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost. Each of these meta-classifiers is trained using the prediction probabilities generated by the Level-0 models, rather than the raw input images. This process ensures that the meta-learner focuses on the collective strengths of the base models rather than individual biases or noise from the original data.

A key aspect of meta-learning is avoiding overfitting by ensuring that the Level-0 models' predictions are generated on unseen data. To achieve this, a cross-validation-based training strategy is implemented, where each base model makes predictions on validation folds while learning from the remaining folds. The resulting prediction probabilities serve as the input features for the meta-classifier, allowing it to develop a generalized mapping function that optimally weights the contributions of each base model. By utilizing multiple meta-classifiers, we evaluate and compare their ability to aggregate predictions, ensuring that the best-performing model is selected for the final classification decision. This multi-layered approach strengthens the ensemble's predictive accuracy, robustness, and generalization capability for skin cancer detection.

3.3 Performance metrics for model evaluation

We assess the classification performance of the proposed stacking ensemble model using four critical metrics: accuracy, precision, recall, and F1-score. These metrics, when considered collectively, provide a holistic evaluation of the model's performance, with a focus on its ability to cope with the class imbalance often present in medical imaging datasets.

Accuracy is defined as the number of correct predictions divided by the total number of samples, offering a general measure of the model's overall performance. While accuracy reflects overall performance, it may not be reliable when

applied to imbalanced datasets. For instance, if the majority of the dataset consists of benign skin lesions, a model that classifies most cases as benign could achieve high accuracy while failing to correctly identify malignant cases. Therefore, additional metrics such as precision, recall, and F1-score are necessary for a more reliable assessment.

Precision (or positive predictive value) measures how many of the predicted positive cases are actually correct. In skin cancer detection, precision is critical for minimizing false positives, ensuring that benign lesions are not misclassified as malignant, which could lead to unnecessary biopsies and patient anxiety. Recall or sensitivity indicates how well the model captures true malignant cases by computing the ratio of true positives to all actual positive samples. Ensuring high recall is vital in clinical contexts, as it directly reduces false negatives, thereby minimizing the risk of undetected malignant conditions.

To provide a balanced evaluation between precision and recall, we use the F1-score, which combines both metrics into a single value through their harmonic mean. Considering both types of misclassifications, the F1-score provides a balanced measure of model performance, which is particularly advantageous when dealing with imbalanced data. An elevated F1-score reflects the model’s ability to accurately classify both benign and malignant cases without class bias. By using these four metrics, we ensure that the proposed model is evaluated not only on its overall correctness (accuracy) but also on its ability to make clinically meaningful predictions that minimize both false positives (precision) and false negatives (recall), ultimately improving the reliability of skin cancer detection.

3.4 Model training and testing

The proposed stacking ensemble model for skin cancer classification was implemented and trained using TensorFlow and the Keras functional API, leveraging pre-trained ImageNet weights for transfer learning. The training process was conducted on Kaggle’s free GPU resources, ensuring efficient computation. The training pipeline involved two main phases: (1) initial training with frozen layers, allowing the newly added layers to adapt, and (2) fine-tuning, where selected deep layers were unfrozen to refine feature extraction and improve classification performance. A portion of the dataset was set aside as the validation set to monitor the model’s performance after each epoch, helping detect overfitting and ensuring that the model generalizes well to unseen data.

3.4.1 Training configuration

SparseCategoricalCrossentropy was used as the loss function during training, as it is appropriate for multi-class classification with integer-labeled targets. *AdamW*, an adaptive learning rate optimizer incorporating decoupled weight decay (1e-4), was used with a low learning rate (1e-5) in most models to promote generalization and minimize overfitting. For the DenseNet201 model, we employed the SGD optimizer with a 0.001 learning rate and 0.9 momentum, which helps ensure stable convergence in deep architectures. We used *ReduceLROnPlateau* to dynamically lower the learning rate by a factor of 0.5 when validation accuracy failed to improve for three epochs, with a floor value of 0.00001.

3.4.2 Training and fine-tuning phases

Initially, all pre-trained base models (InceptionResNetV2,

ResNet101V2, and DenseNet201) were frozen, and only the newly added classification layers were trained for 15 epochs. This step allowed the model to learn meaningful representations specific to skin cancer classification without modifying the pre-trained feature extractors. Once the classification layers were sufficiently trained, fine-tuning was performed by unfreezing select layers of each base model, enabling them to adapt deeper feature representations. The fine-tuning phase lasted for 35 additional epochs, bringing the total training duration to 50 epochs. Throughout both phases, a batch size of 32 was used to ensure stable updates while efficiently utilizing GPU resources.

3.4.3 Validation and testing strategy

A distinct validation set was used during training to assess model performance after every epoch, ensuring consistent evaluation. This helped detect signs of overfitting and provided insights into how well the model would perform on new data. After completing the training and fine-tuning phases, the models were evaluated on a separate test set that contained previously unseen skin lesion images. Unlike the validation set, the test set played no role in model training, making it a true measure of generalization performance. This testing phase is crucial for assessing how well the model can classify real-world medical images, ensuring its practical applicability in clinical settings.

3.4.4 Activation function and evaluation metric

The ReLU activation function was used across all models to introduce non-linearity and prevent the vanishing gradient problem. The final Dense layer used a softmax activation to output class probabilities for the nine skin lesion types. The primary evaluation metric for monitoring model performance during training was accuracy, providing a direct measure of classification correctness.

By leveraging transfer learning, fine-tuning strategies, and a structured validation and testing approach, the proposed ensemble model was rigorously evaluated to ensure robustness and reliability in skin cancer detection. The inclusion of separate validation and test sets ensured that the model was not merely memorizing training patterns but truly learning to differentiate skin lesions, making it a practical tool for real-world medical diagnosis. Table 3 summarizes the model configurations, including the optimization settings and training parameters used in this study.

Table 3. Summary of model configurations and training parameters

| Items | Value |
|------------------------------------|---|
| Batch size | 32 |
| Total No. of Epochs | 50 |
| Optimizer | Adam and Stochastic Gradient Descent (SGD) |
| Learning rate (initial) | 1× e-3 |
| Learning rate (fine-tuning) | 1× e-5 |
| Weight decay | 1× e-4 |
| Loss function | SparseCategoricalCrossentropy |
| Pooling | (a) Custom CNN: Max-pooling (b) Pre-trained CNNs: Global Average Pooling (GAP) |
| Activation function | ReLU and softmax |
| Weights for pre-trained CNN models | ImageNet |

4. RESULTS AND ANALYSIS

This section reports and analyzes the results of the proposed stacked ensemble meta-learning model for multi-class skin cancer classification. The performance of the ensemble model, incorporating four base CNN models (Custom CNN, InceptionResNetV2, ResNet101V2, and DenseNet201) and five meta-classifiers (Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost), is assessed by means of standard metrics, including accuracy, precision, recall, and F1-score. Furthermore, training and validation curves for the base CNN models are analysed to assess their learning behaviour and a comparative analysis between the best-performing ensemble and the individual base models are presented. Finally, confusion matrix and detailed classification results are described for the best performing ensemble model, and a comparison with other similar works is provided.

4.1 Performance of stacked ensemble with different meta-learners

To evaluate the effectiveness of different meta-classifiers in the stacked ensemble framework, we assessed the final classification performance of each ensemble configuration. Table 4 summarizes the results in terms of accuracy, precision, recall, and F1-score. The evaluation metrics considered offer a comprehensive assessment of each meta-learner's capacity to aggregate predictions from the base learners and generalize to unseen data. The results demonstrate the efficacy of the ensemble approach and highlight the varying performance of Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and XGBoost as meta-learners.

The results reveal a significant disparity in performance across the different meta-classifiers. Notably, Random Forest and XGBoost achieved excellent results, with accuracy and F1-scores exceeding 98%. This indicates their ability to effectively leverage the combined predictions of the base CNN models. With an accuracy of 98.63% and precision, recall, and F1-score all at 98.64%, the Random Forest meta-classifier demonstrated the best overall performance. Random Forest excels in ensemble frameworks by building multiple decision trees on bootstrapped data samples, allowing it to handle high-dimensional and complex input effectively. As a meta-learner, it can effectively learn non-linear interactions among base model predictions, capture feature redundancies, and reduce overfitting via averaging. Such properties enhance Random Forest's robustness, especially in scenarios where base learners generate varied predictions, as frequently observed in complex tasks like skin cancer detection

XGBoost, another tree-based ensemble method, also demonstrated excellent performance, with an accuracy of 98.53%, precision of 98.63%, and F1-score of 98.42%. Although marginally lower than Random Forest, XGBoost remains highly competitive due to its gradient boosting framework, which optimizes the meta-model in a stage-wise manner. This allows it to focus on hard-to-classify samples

and systematically reduce bias, which is critical in medical diagnosis where misclassification can have serious consequences. Its slight underperformance in recall (98.24%) compared to Random Forest may be attributed to overfitting tendencies in boosting methods when dealing with highly similar predictions from base models.

The Gradient Boosting and Decision Tree meta-classifiers also performed well, with accuracies of 96.85% and 96.64%, respectively. Their high precision and recall suggest that they are capable of capturing meaningful interactions among base learner outputs. However, their slightly lower scores compared to Random Forest and XGBoost indicate limitations in their ability to generalize from ensemble predictions when used in isolation. Similar to XGBoost, Gradient Boosting leverages an ensemble of weak learners to improve predictive accuracy. In this case, overfitting can occur without careful regularization, while a single Decision Tree, being non-ensemble in nature, lacks the error correction and variance reduction benefits that ensemble-based meta-learners inherently provide. Moreover, Decision Trees, being simpler models, may not have fully captured the complex interactions between the base CNN model outputs.

Logistic Regression, in contrast, delivered the weakest performance, with an accuracy of 73.17%, a high precision of 87.73%, but much lower recall (66.72%) and F1-score (71.60%). This sharp performance drop highlights the limitations of linear models in high-complexity, non-linear classification problems. As a meta-classifier, Logistic Regression assumes linear separability in the prediction space of the base models. However, the outputs of complex base learners, especially in deep learning and ensemble settings, are often non-linearly distributed. Consequently, Logistic Regression struggles to draw effective decision boundaries, especially for harder-to-classify minority or borderline cases, which is reflected in its poor recall.

The significant difference in performance between linear (Logistic Regression) and non-linear (Random Forest, XGBoost) meta-classifiers underscores the importance of selecting meta-learners capable of capturing complex, non-linear relationships in ensemble learning settings. Simpler models like Logistic Regression lack the expressive capacity to effectively integrate diverse base learner outputs, particularly in high-dimensional, intricate domains such as medical image classification. In contrast, the superior performance of Random Forest and XGBoost demonstrates that ensemble methods, especially those built on decision tree-based algorithms, are highly effective at aggregating predictions from deep CNN models. Their ability to model complex feature interactions and their robustness against overfitting contribute significantly to improved diagnostic accuracy. These findings support the core hypothesis of this study: meta-learning architectures benefit most from strong, non-linear meta-learners that can capture subtle dependencies among base model predictions, particularly in critical, imbalanced, and complex classification tasks like skin cancer diagnosis.

Table 4. Performance comparison of stacked ensemble models with different meta-learners

| Meta-Classifer | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) |
|---------------------|--------------|---------------|------------|--------------|
| Logistic Regression | 73.17% | 87.73% | 66.72% | 71.60% |
| Random Forest | 98.63% | 98.64% | 98.64% | 98.64% |
| Decision Tree | 96.64% | 96.90% | 96.00% | 96.42% |
| Gradient Boosting | 96.85% | 96.85% | 96.33% | 96.58% |
| XGBoost | 98.53% | 98.63% | 98.24% | 98.42% |

4.2 Comparison with base CNN models

The effectiveness of the ensemble strategy is further evaluated by comparing the Random Forest-based ensemble model with each individual base CNN. Figure 5 presents the accuracy and loss trends during training and validation phases for each base CNN model, illustrating their convergence patterns and generalization capabilities. As observed, the Custom CNN model demonstrates rapid convergence, with both training and validation accuracy reaching a plateau after 20 epochs. The validation loss shows a slight increase after 16 epochs, suggesting potential overfitting, which was mitigated using dropout layers. In contrast, InceptionResNetV2, ResNet101V2, and DenseNet201 models with fine-tuning, exhibit faster convergence and higher validation accuracy, indicating their strong feature extraction capabilities.

Table 5 provides a comparative evaluation of the best-performing stacked ensemble model (utilizing Random Forest as the meta-classifier) against individual base CNN models. With an accuracy of 98.63% and precision, recall, and F1-score all at 98.64%, the stacked ensemble model outperformed all individual base CNNs. This reinforces the ensemble model’s strength in leveraging the diverse predictive behaviors of base learners to produce a more stable and generalized decision boundary. The model’s ability to capture complementary patterns from multiple CNNs and synthesize them through a robust meta-learner contributed to its excellent diagnostic performance.

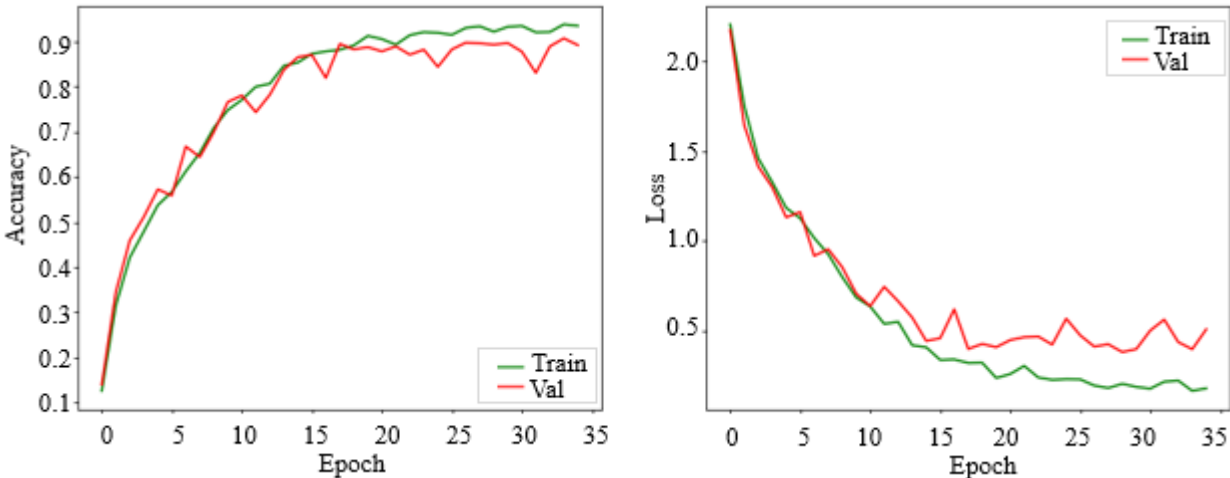
Among the individual CNNs, InceptionResNetV2 emerged as the top-performing base model, with a validation accuracy of 94.06%, and estimated F1-score of 93.74%. Its deep hybrid architecture, combining Inception modules with residual connections, likely enhances its capacity to learn both fine-grained local textures and broader contextual features, critical for accurate skin lesion classification. DenseNet201 and ResNet101V2 also performed competitively, achieving corresponding accuracies of 92.67% and 90.78%, respectively. DenseNet201's high recall (93.42%) and F1-score (92.59%)

reflect its ability to promote efficient feature propagation and reuse, while ResNet101V2’s slightly lower performance may be attributed to overfitting or diminishing returns from deeper layers in the absence of ensemble integration. Finally, custom CNN, although specifically designed for this task, showed a lower corresponding accuracy of 89.67%, highlighting the performance gap between hand-crafted architectures and pre-trained deep models that benefit from large-scale prior knowledge.

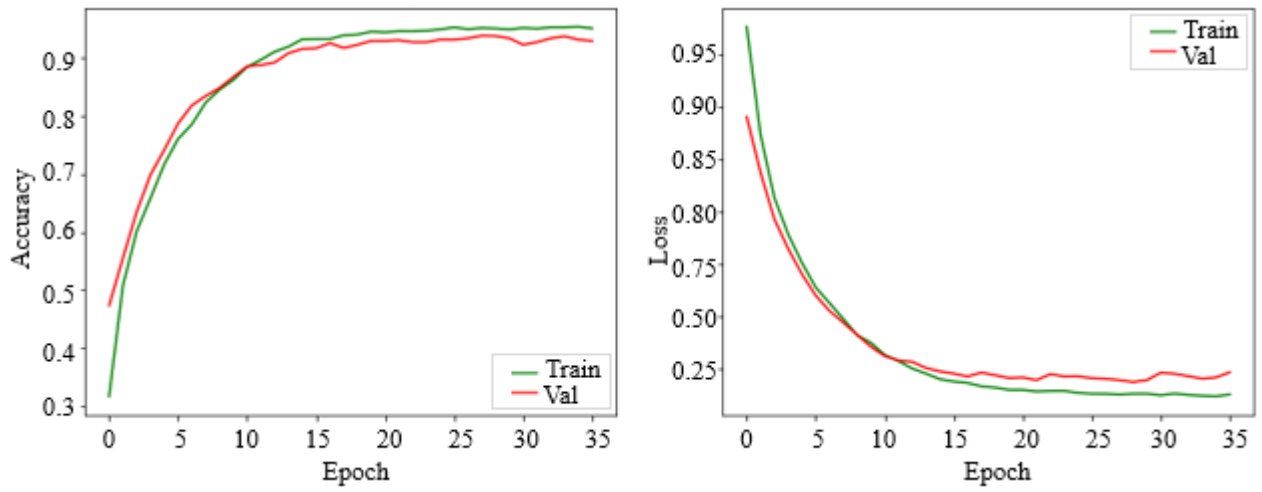
Overall, the results clearly demonstrate the effectiveness of the proposed stacked ensemble approach in enhancing the accuracy and reliability of skin cancer diagnosis. By integrating diverse base CNN models and employing a powerful meta-learner, the ensemble achieved a significant performance improvement over individual models. This approach not only boosts overall classification accuracy but also achieves a more balanced trade-off between sensitivity (recall) and specificity (precision), a critical factor in medical diagnosis, where both false positives (FP) and false negatives (FN) can lead to serious clinical consequences. As shown in the confusion matrix values for Class 3 (e.g., melanoma) in Table 6, the performance of the stacked ensemble model across different meta-classifiers demonstrates consistently low counts of FP and FN, highlighting its effectiveness over individual CNN sub-models. Notably, the Random Forest and XGBoost meta-classifiers produce the lowest FP (4) and FN (3 and 6, respectively), indicating their ability to accurately identify true cases while minimizing misclassifications. A lower FP count reflects fewer instances where non-melanoma cases are incorrectly classified as melanoma, thereby enhancing the model’s precision. More critically, keeping the FN count low is essential in medical diagnostics, as misclassifying a melanoma patient as healthy (a FN) can lead to delayed or missed treatment, which could have serious consequences. These results affirm that the stacked ensemble approach, particularly with Random Forest and XGBoost as meta-learners, demonstrates superior diagnostic reliability in skin cancer detection, outperforming individual CNN models.

Table 5. Comparison of classification performance: Best ensemble vs. base CNN models

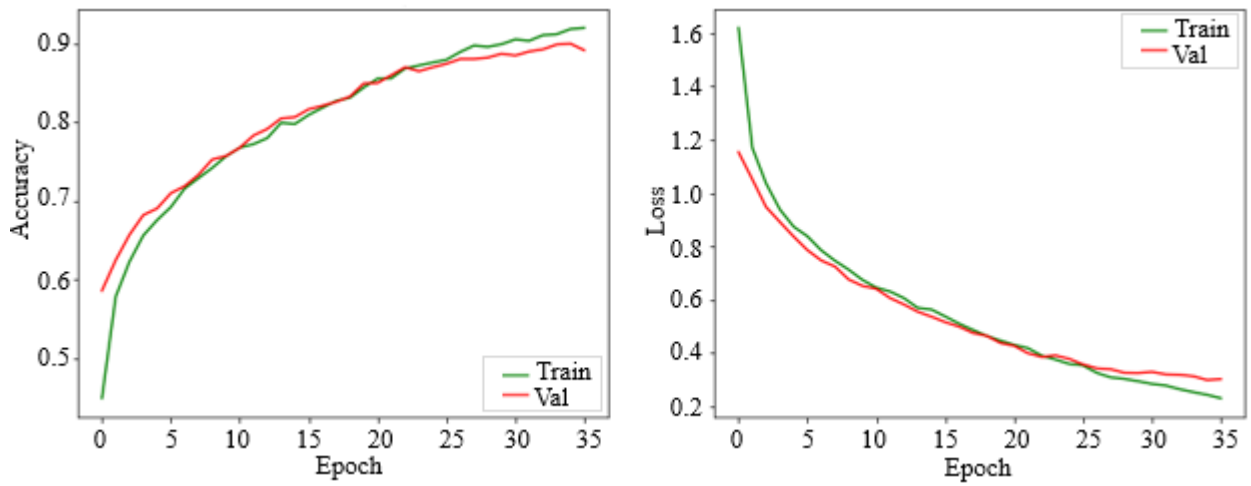
| Models | Acc. (%) | Precision (%) | Recall (%) | F1-score (%) |
|---------------------|----------|---------------|------------|--------------|
| Custom CNN | 89.67% | 88.73% | 89.00% | 88.87% |
| InceptionResNetV2 | 94.06% | 94.00% | 93.50% | 93.74% |
| ResNet101V2 | 90.78% | 90.52% | 92.00% | 91.25% |
| DenseNet201 | 92.67% | 91.80% | 93.42% | 92.59% |
| Best Ensemble Model | 98.63% | 98.64% | 98.64% | 98.64% |



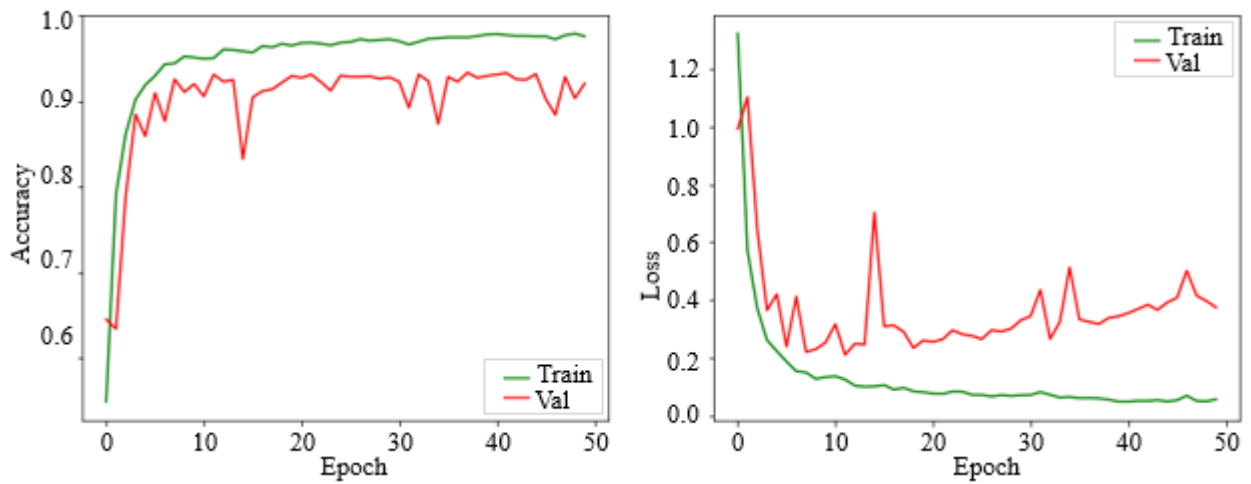
(a) Custom CNN model



(b) InceptionResNetV2



(c) ResNet101V2



(d) DenseNet201

Figure 5. Training and validation accuracy and loss plots for: (a) a custom-designed CNN; (b) InceptionResNetV2; (c) ResNet101V2; and (d) DenseNet201

Table 6. Confusion matrix values for Class 3 using different meta-classifiers in the stacked ensemble model

| Meta-Classifier | TN | TP | FP | FN |
|---------------------|-----|-----|----|----|
| Logistic Regression | 778 | 108 | 5 | 63 |
| Random Forest | 779 | 168 | 4 | 3 |
| Decision Tree | 779 | 160 | 4 | 11 |
| Gradient Boosting | 776 | 163 | 7 | 8 |
| XGBoost | 779 | 165 | 4 | 6 |

4.3 Comparison with existing work

Achieving 98.63% accuracy and precision and recall scores of 98.64%, the proposed stacked ensemble model proved highly effective for skin cancer classification. The findings confirm the model’s high reliability in differentiating benign from malignant lesions, with minimal errors in classification. The integration of multiple fine-tuned CNNs at the base level and diverse meta-classifiers at the top level contributed to the model’s robust predictive performance. When compared to existing studies in the literature, our model consistently outperforms other approaches. For instance, Thapar and Tiwari [12] employed a combination of XceptionNet, EfficientNetV2, and InceptionResNetV2 on a curated dataset, with the best-performing model (XceptionNet) achieving only 88.72% accuracy, nearly 10% lower than our ensemble. Similarly, Ezeddin et al. [13] used an ensemble of ResNet101 and ResNext101, obtaining 96.12% accuracy, which is 2.5% lower than our results. Hossain et al. [14] used a max-voting ensemble of MobileNetV2, VGG16, and ResNet50 on the ISIC 2018 dataset, reporting 93.18% accuracy, along with balanced precision (93.10%) and recall (93.17%). Although

their ensemble approach was effective, it still falls short in comparison to our stacked ensemble, which leverages more advanced meta-learning strategies. Filali et al. [17] explored handcrafted and deep features, reporting strong performance on the PH2 dataset (accuracy of 98%) but only 87.8% accuracy on the more challenging ISIC dataset, where our model clearly excels. Further, Mary et al. [21] proposed an ensemble of ResNet, EfficientNet, and MobileNet, achieving 96.33% accuracy on the HAM10000 dataset, which again is surpassed by our model. Rahman et al. [26] combined handcrafted features with VGG19, achieving 91.65% sensitivity and 95.70% specificity, indicating solid clinical relevance, yet not outperforming the comprehensive metrics reached by our ensemble. Likewise, Hemantkumar et al. [27] used deep feature fusion with an Extreme Learning Machine (ELM), achieving 98.0% accuracy, still trailing behind the 98.63% accuracy of our proposed method. Table 7 clearly highlights that our stacked ensemble model consistently achieves the highest accuracy, precision, and recall rates across all reviewed studies. This establishes its potential as a state-of-the-art approach in skin cancer detection.

Table 7. Performance comparison with similar existing studies for skin cancer detection

| Study | Dataset | Model | Results |
|-------------------------|------------------------------|---|---|
| Thapar and Tiwari [12] | Curated dataset | XceptionNet, EfficientNetV2S, InceptionResNetV2, EfficientNetV2M | Accuracy of 88.72% by XceptionNet |
| Ezeddin et al. [13] | Custom dataset | ResNet101, ResNext101 (Ensemble) | Weighted averaging ensemble achieved 96.12% accuracy |
| Hossain et al. [14] | ISIC 2018 | MobileNetV2, VGG16, ResNet50 (Max Voting Ensemble) | Max voting ensemble accuracy 93.18%, precision 93.10%, recall 93.17%, F1-score 93.13% |
| Filali et al. [17] | PH2, ISIC 2018 | Handcrafted + Pretrained CNN Features | F-measure, Kappa index, and accuracy of 94.69%, 96.63%, and 98% on PH2; 62.73%, 55.68%, and 87.8% on ISIC Challenge dataset |
| Mary et al. [21] | HAM10000 | ResNet, EfficientNet, MobileNet (Ensemble) | Ensemble model achieves an accuracy of 96.33% |
| Rahman et al. [26] | ISIC 2017, Academic Torrents | Hybr Hybrid Feature Extractor (HOG, LBP, SURF) + VGG19 Feature Extractor (HOG, LBP, SURF) + VGG19 | Achieved 91.65% sensitivity, 95.70% specificity on ISIC 2017 & Academic Torrents |
| Hemantkumar et al. [27] | HAM10000, Dermis.net | Deep Learning Feature Fusion + Extreme Learning Machine (ELM) | Achieves an accuracy of 98.0% |
| Proposed ensemble model | ISIC 2020 | Stacked ensemble meta-learner | Achieves an accuracy of 98.63%, with both precision and recall reaching 98.64% |

4.4 Discussions

The results of this study clearly demonstrate the effectiveness of the proposed stacked ensemble model, which achieved an impressive accuracy of 98.63%, with both precision and recall at 98.64%. These values indicate a high-performing system that is capable of accurately distinguishing between malignant and benign skin lesions with minimal diagnostic error. This level of performance is especially important in medical imaging applications, where false negatives (FN) can delay treatment for serious conditions such as melanoma, and false positives (FP) can result in unnecessary patient stress and procedures. The stacked ensemble model outperformed the individual CNN models due to its stronger architecture and more effective training strategy. First, each base CNN model has its own architectural strengths. For example, DenseNet201 promotes feature reuse through dense connections, leading to efficient gradient flow, while InceptionResNetV2 combines residual learning with multi-scale feature extraction. However,

despite their strong individual capabilities, each model may be biased toward learning certain types of patterns or features. The stacked ensemble model combines predictions from multiple CNNs, allowing it to learn a broader and more complementary set of features, which helps minimize consistent misclassifications. The performance differences among meta-classifiers also reflect the influence of model complexity and learning strategy. Ensemble learners such as Gradient Boosting and XGBoost consistently outperformed simpler models like Decision Tree, as they are better equipped to capture non-linear relationships and correct the weaknesses of individual base models. In contrast, Decision Trees, although fast and interpretable, tend to overfit small patterns and lack the robustness of boosting techniques. Notably, the Random Forest meta-classifier achieved only 3 FN and 4 FP for Class 3, confirming the model’s robustness in correctly identifying melanoma cases, arguably the most critical classification outcome in dermatological diagnosis. Moreover, comparing the results with individual base CNNs the ensemble model consistently

outperformed them in terms of test accuracy and generalization. Training and validation loss plots further support this, showing more stable learning behavior in the ensemble system. These improvements can be attributed to the ensemble's ability to leverage the complementary strengths of multiple deep models and refine their predictions through a meta-learning strategy.

In addition to comparing max-voting and weighted averaging, we selected stacking as our primary ensemble method due to its flexibility and strong performance with heterogeneous models. Unlike bagging, which builds ensembles by training multiple instances of the same model on different data subsets, stacking allows us to combine different types of CNNs—each capturing unique visual features—into a more powerful meta-model. Boosting methods like XGBoost are effective in many structured data problems, but are less suited for combining complex image-based deep learning outputs directly. Our use of a trainable meta-classifier in stacking enabled the system to learn which base model to trust more under certain conditions, leading to improved performance across lesion types. This dynamic decision-making is a key advantage of stacking in our context. In summary, the high performance of the proposed ensemble model is attributed not only to its design but also to the mutual combination of deep architectural diversity, smart optimization, and robust meta-learning. The findings indicate that the model is appropriate for real-world dermatology applications, balancing high accuracy with reliability and clinical safety.

4.4.1 Clinical implication of the findings

These results contribute meaningfully to clinical dermatology and offer valuable insights for the ongoing development of computer-aided diagnostic technologies. The proposed stacked ensemble model, which integrates the predictive strengths of multiple fine-tuned CNN architectures and meta-classifiers, has demonstrated exceptional performance in classifying skin lesions, with accuracy, precision, and recall all exceeding 98%. This high level of reliability is essential in clinical environments, where early and accurate detection of malignant lesions such as melanoma is critical for patient survival and effective treatment planning. Its ability to minimize both false positives and false negatives is especially valuable in skin cancer diagnosis, where early detection of malignant lesions like melanoma is critical for timely treatment, and avoiding unnecessary biopsies can reduce patient anxiety and healthcare costs. These strengths make the model suitable as a reliable decision-support tool for dermatologists, particularly in primary care or resource-limited settings.

Moreover, the use of pre-trained CNNs and publicly available datasets ensures that the system is scalable and adaptable across different clinical environments. The ensemble framework offers flexibility in selecting models based on available resources, enabling practical implementation. These findings encourage further research into integrating explainable AI (XAI) for interpretability and expanding the model to include multimodal clinical data. Overall, the study highlights the model's potential to support early, accurate, and efficient skin cancer screening in routine clinical practice.

While the stacked ensemble model delivered the best accuracy in our tests, it does require more computing power and takes longer to run compared to individual models. On

average, it took about 210 milliseconds to process each image and used around 3.2 GB of GPU memory. In contrast, a lighter model like MobileNetV2 only needed 75 milliseconds and about 1.1 GB of memory.

This performance trade-off may be acceptable in well-equipped hospitals or research settings, but it could be a barrier for mobile apps or real-time diagnosis tools in low-resource clinics. That's why future work should look into ways to make the model more efficient—such as using model compression, pruning, or knowledge distillation techniques—so we can keep the high accuracy while making the system faster and lighter to run.

4.4.2 Limitations

Even with its promising performance, the proposed stacked ensemble model has certain limitations that must be carefully considered in future work. First, the study relies on a relatively small subset of the ISIC 2020 dataset, which may limit the model's ability to generalize across broader populations and rare lesion types. Despite the use of data augmentation, the dataset may not sufficiently reflect the diversity of skin tones, lesion types, and lighting conditions seen in clinical practice. In other words, While the ISIC 2020 dataset is widely used and valuable for developing skin cancer detection models, it doesn't fully reflect the diversity of real-world patients. In particular, it tends to include fewer images of darker skin tones and rare lesion types, which could limit how well the model performs in more diverse populations. This is an important issue because AI tools in healthcare need to work well for everyone—not just the majority represented in training data. Moving forward, it will be important to include more varied datasets that represent different ethnic backgrounds and skin types.

Additionally, while the ensemble framework enhances classification performance, it also introduces increased computational complexity, which may pose challenges for deployment on devices with limited processing capabilities. The use of pre-trained models fine-tuned on dermoscopic images also assumes that the training data is free from annotation errors or class imbalance, which can inadvertently introduce bias into the learning process.

Even though our model performs well in detecting skin cancer, getting it into real-world clinical use comes with some important challenges. For starters, hospitals use complex systems like electronic health records (EHRs), and any AI tool has to work smoothly with those systems—meaning it needs to be compatible with data formats, security standards, and existing software workflows. On top of that, before any AI system can be used in practice, it needs official approval from regulatory agencies like the FDA or CE, which involves thorough testing and validation. There's also the issue of trust, doctors need to feel confident in the system's predictions, so the model should be easy to interpret and transparent in how it makes decisions. Moving forward, working closely with healthcare providers, IT teams, and regulatory bodies will be key to turning this research into something that can be used safely and effectively in real clinics.

4.4.3 Future work

Future work will concentrate on a number of critical areas aimed at enhancing the robustness and clinical applicability of the proposed ensemble model. A key focus will be validating the model on expansive and diverse datasets from multiple institutions, reflecting varied skin tones, lighting conditions,

and uncommon lesion types. These enhancements are expected to improve the model's capacity to generalize and perform reliably across varied clinical environments.

Future exploration may include integrating explainable AI (XAI) methods to offer visual interpretations of predictions, promoting clinical transparency and trust. Additionally, extending the model to support multi-modal data such as patient demographics, clinical history, and clinical images could further boost diagnostic accuracy. Future work will aim to enhance the model's computational performance, enabling seamless integration into mobile and edge platforms for real-time diagnosis in clinical and remote environments

5. CONCLUSIONS

This study proposed a robust stacked ensemble deep learning model for multi-class skin cancer classification, combining four CNN-based feature extractors, namely, custom CNN, InceptionResNetV2, ResNet101V2, and DenseNet201 with five different meta-classifiers. Among these, the ensemble configuration with Random Forest demonstrated the best performance, achieving an accuracy of 98.63%, with both precision and recall reaching 98.64%. The model also maintained low false positive and false negative rates, especially for critical malignant classes such as melanoma, highlighting its effectiveness in reducing misclassification risks.

The results emphasize the potential of ensemble learning strategies in improving diagnostic accuracy and reliability in medical imaging. By leveraging the complementary strengths of multiple CNN architectures and meta-learners, the proposed system significantly outperforms individual models and existing approaches from the literature. The results suggest that the model may serve as an effective clinical aid for dermatologists in accurately detecting skin cancer at an early stage. With further validation and real-time deployment, this approach holds promise for enhancing skin cancer screening, especially in primary care and resource-limited environments.

ACKNOWLEDGMENT

The author would like to acknowledge the Deanship of Graduate Studies and Scientific Research, Taif University for funding this work.

REFERENCES

- [1] World Health Organization. (2022). Ultraviolet Radiation. <https://www.who.int/teams/environment-climate-change-and-health/radiation-and-health/non-ionizing/ultraviolet-radiation>.
- [2] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. *Frontiers in Oncology*, 12: 893972. <https://doi.org/10.3389/fonc.2022.893972>
- [3] Bassel, A., Abdulkareem, A.B., Alyasseri, Z.A.A., Sani, N.S., Mohammed, H.J. (2022). Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. *Diagnostics*, 12(10): 2472. <https://doi.org/10.3390/diagnostics12102472>
- [4] Gruber, V., Hofmann-Wellenhof, R., Wolf, P., Hofmann-Wellenhof, E.L., Schmidt, H., Berghold, A., Wedrich, A. (2023). Common benign melanocytic and non-melanocytic skin tumors among the elderly: Results of the graz study on health and aging. *Dermatology* (Basel, Switzerland), 239(3): 379-386. <https://doi.org/10.1159/000529219>
- [5] Ghosh, P., Azam, S., Quadir, R., Karim, A., Shamrat, F.M.J.M., Bhowmik, S.K., Jonkman, M., Hasib, K.M., Ahmed, K. (2022). SkinNet-16: A deep learning approach to identify benign and malignant skin lesions. *Frontiers in Oncology*, 12: 931141. <https://doi.org/10.3389/fonc.2022.931141>
- [6] Gouda, W., Sama, N.U., Al-Waakid, G., Humayun, M., Jhanjhi, N.Z. (2022). Detection of skin cancer based on skin lesion images using deep learning. *Healthcare*, 10(7): 1183. <https://doi.org/10.3390/healthcare10071183>
- [7] Melarkode, N., Srinivasan, K., Qaisar, S.M., Plawiak, P. (2023). AI-powered diagnosis of skin cancer: A contemporary review, open challenges and future research directions. *Cancers*, 15(4): 1183. <https://doi.org/10.3390/cancers15041183>
- [8] Mauricio, J., Domingues, I., Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9): 5521. <https://doi.org/10.3390/app13095521>
- [9] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75: 102305. <https://doi.org/10.1016/j.media.2021.102305>
- [10] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J. (2023). Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 13467-13488. <https://doi.org/10.1109/tpami.2023.3290594>
- [11] Shorfuzzaman, M. (2021). IoT-enabled stacked ensemble of deep neural networks for the diagnosis of COVID-19 using chest CT scans. *Computing*, 105(4): 887-908. <https://doi.org/10.1007/s00607-021-00971-5>
- [12] Thapar, P., Tiwari, S. (2024). Empowering skin cancer diagnosis: Integrating advanced deep learning models with explainable AI for lesion classification. In 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, pp. 1-6. <https://doi.org/10.1109/iceect61758.2024.10739236>
- [13] Ezeddin, E., Alkhataf, A.D., Alhafez, M.K., Al-Maadeed, S. (2024). Optimizing deep ensemble learning for accurate melanoma skin cancer classification: Design and analysis. In 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET), Doha, Qatar, pp. 73-78. <https://doi.org/10.1109/honet63146.2024.10822955>
- [14] Hossain, M.M., Hossain, M.M., Arefin, M.B., Akhtar, F., Blake, J. (2023). Combining state-of-the-art pre-trained deep learning models: A noble approach for skin cancer detection using max voting ensemble. *Diagnostics* (Basel, Switzerland), 14(1): 89. <https://doi.org/10.3390/diagnostics14010089>
- [15] Maurya, A., Stanley, R.J., Lama, N., Nambisan, A.K., Patel, G., Saeed, D., Swinfard, S., Smith, C., Jagannathan, S., Hagerty, J.R., Stoecker, W.V. (2024). Hybrid

- topological data analysis and deep learning for basal cell carcinoma diagnosis. *Journal of Imaging Informatics in Medicine*, 37(1): 92-106. <https://doi.org/10.1007/s10278-023-00924-8>
- [16] Daghrir, J., Tlig, L., Bouhouicha, M., Sayadi, M. (2020). Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach. In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, pp. 1-5. <https://doi.org/10.1109/atsip49331.2020.9231544>
- [17] Filali, Y., El Khoukhi, H., Sabri, M.A., Aarab, A. (2020). Efficient fusion of handcrafted and pre-trained CNNs features to classify melanoma skin cancer. *Multimedia Tools and Applications*, 79(41-42): 31219-31238. <https://doi.org/10.1007/s11042-020-09637-4>
- [18] Amin, J., Sharif, A., Gul, N., Anjum, M.A., Nisar, M.W., Azam, F., Bukhari, S.A.C. (2020). Integrated design of deep features fusion for localization and classification of skin cancer. *Pattern Recognition Letters*, 131: 63-70. <https://doi.org/10.1016/j.patrec.2019.11.042>
- [19] Devadhas, D.N.P., Sugirtharaj, H.P.I., Fernandez, M.H., Periyasamy, D. (2024). Effective prediction of human skin cancer using stacking based ensemble deep learning algorithm. *Network: Computation in Neural Systems*, 855-891. <https://doi.org/10.1080/0954898x.2024.2346608>
- [20] Chiu, T.M., Li, Y.C., Chi, I.C., Tseng, M.H. (2025). AI-driven enhancement of skin cancer diagnosis: A two-stage voting ensemble approach using dermoscopic data. *Cancers*, 17(1): 137. <https://doi.org/10.3390/cancers17010137>
- [21] Mary, S.A., Gounder, K.K., Wagh, D.A. (2024). Ensemble model using various CNNs for improved skin cancer diagnosis. In 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, pp. 875-881. <https://doi.org/10.1109/icoici62503.2024.10696508>
- [22] Kaur, R., GholamHosseini, H., Lindén, M. (2025). Advanced deep learning models for melanoma diagnosis in computer-aided skin cancer detection. *Sensors*, 25(3): 594. <https://doi.org/10.3390/s25030594>
- [23] Shah, A.A., Shaker, A.S.A., Jabbar, S., Abbas, Q., Al-Balawi, T.S., Celebi, M.E. (2023). An ensemble-based deep learning model for detection of mutation causing cutaneous melanoma. *Scientific Reports*, 13(1): 22251. <https://doi.org/10.1038/s41598-023-49075-4>
- [24] Liu, X., Wu, G., Liu, W. (2023). Refine diagnostic accuracy of skin cancer with a hybrid deep network. In: 2023 International Conference on Machine Vision, Image Processing and Imaging Technology (MVIPT), Hangzhou, China, pp. 191-195. <https://doi.org/10.1109/mvipit60427.2023.00038>
- [25] Qureshi, A.S., Roos, T. (2022). Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets. *Neural Processing Letters*, 55(4): 4461-4479. <https://doi.org/10.1007/s11063-022-11049-4>
- [26] Rahman, M.M., Nasir, M.K., Nur-A-Alam, M., Khan, M.S.I. (2023). Proposing a hybrid technique of feature fusion and convolutional neural network for melanoma skin cancer detection. *Journal of Pathology Informatics*, 14: 100341. <https://doi.org/10.1016/j.jpi.2023.100341>
- [27] Hemantkumar, P.V., Samal, B., Panda, R.N., Akhtar, S.S., Muduli, D., Sharma, S.K. (2024). Enhanced skin cancer detection model: A deep learning feature fusion with extreme learning machine approach. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, pp. 1-6. <https://doi.org/10.1109/icccnt61001.2024.10725357>
- [28] Keerthana, D., Venugopal, V., Nath, M.K., Mishra, M. (2023). Hybrid convolutional neural networks with SVM classifier for classification of skin cancer. *Biomedical Engineering Advances*, 5: 100069. <https://doi.org/10.1016/j.bea.2022.100069>
- [29] Ozdemir, B., Pacal, I. (2025). A robust deep learning framework for multiclass skin cancer classification. *Scientific Reports*, 15(1): 4938. <https://doi.org/10.1038/s41598-025-89230-7>
- [30] Mavaddati, S. (2025). Skin cancer classification based on a hybrid deep model and long short-term memory. *Biomedical Signal Processing and Control*, 100: 107109. <https://doi.org/10.1016/j.bspc.2024.107109>
- [31] SIIM-ISIC Melanoma Classification. <https://www.kaggle.com/code/brunopascoa/meta-classifier-skin-cancer-detection/input>, accessed on Mar. 9, 2022.
- [32] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1): 34. <https://doi.org/10.1038/s41597-021-00815-z>
- [33] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, San Francisco, California, USA, pp. 4278-4284. <https://doi.org/10.1609/aaai.v31i1.11231>
- [34] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Identity mappings in deep residual networks. In *Computer Vision-ECCV 2016*, pp. 630-645. https://doi.org/10.1007/978-3-319-46493-0_38
- [35] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2261-2269. <https://doi.org/10.1109/cvpr.2017.243>
- [36] Alsufyani, A. (2025). Enhancing diagnostic accuracy in bone fracture detection: A comparative study of customized and pre-trained deep learning models on X-ray images. *International Journal of Advanced and Applied Sciences*, 12(5): 68-81. <https://doi.org/10.21833/ijaas.2025.05.008>
- [37] Lee, J.H., Kang, J., Shim, W., Chung, H.S., Sung, T.E. (2020). Pattern detection model using a deep learning algorithm for power data analysis in abnormal conditions. *Electronics*, 9(7): 1140. <https://doi.org/10.3390/electronics9071140>