Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1214

A Robust Explainable Deep Learning Ensemble for Early Skin Cancer Diagnosis

Hammad Ali[1*], Muhammad Rizwan Rahsid Rana[1], and Abdul Sami[1]

[1]Department of Robotics & Artificial Intelligence, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan.
[*]Corresponding Author: Hamad Ali. Email: hammadaly.6229@gmail.com

Received: November 02, 2025 **Accepted:** January 14, 2026

Abstract: Skin cancer is one of the most common types of malignancies around the world, and the ability to detect skin cancers in an early stage is crucial for improving overall patient outcomes. This study introduces a hybrid deep learning framework that utilizes self-supervised pretraining, multi-architecture ensemble learning, and explainable AI approaches to enable accurate and interpretable skin cancer diagnosis. This framework uses SimCLR-based contrastive learning techniques to generate powerful feature representations from large data sets of unlabeled images of dermatoscopic images before implementing either supervised fine-tuning processes or feature-level fusion processes on three different types of architectures (EfficientNetV2-L, Swin Transformer, and ConvNeXt). In order to classify patients using the features derived from the different architectures, a meta-learning classifying component based on LightGBM is built into the model and provides explainability through the Grad-CAM and SHAP explainable AI methods. The results of the experiments performed with benchmark datasets (ISIC, and HAM10000) demonstrate the proposed method outperformed previously established baseline models by a wide margin, achieving 94.5% accuracy, 92.55% precision, and 93.26% recall, providing evidence of the robustness, high sensitivity, and reliability of the proposed method in the early detection of skin cancer.

Keywords: Skin Cancer Diagnosis; Deep Learning; Self-Supervised Learning; Multi-Architecture Ensemble; Explainable AI

1. Introduction

There are many cases of skin cancer in the world, Non-Melanoma Skin cancer is the most common cancer in the World, estimated at 1,234,533 cases of Non-Melanoma Skin Cancer and 69,416 deaths from the disease [1]. An estimated 330,000 new cases of melanoma were diagnosed in the world in 2022 and related deaths totaled almost 60,000 [2]. Even with considerable improvements in various treatments, early detection of Skin Cancer is still most critical determining factor for how well patients are able to survive and recover from their illness. With this urgent need for Early Detection, we can expect that as the demand for better diagnostic approaches continues to grow, current limitations associated with Traditional Methods will need to be overcome with more advanced methods [4].

Traditional dermatological examination relies heavily on visual inspection and clinical expertise, often supplemented by dermoscopy for enhanced visualization of skin structures [5]. However, this conventional approach faces substantial challenges including inter-observer variability, diagnostic subjectivity, and the limited availability of specialized expertise, particularly in resource-constrained healthcare settings. These diagnostic limitations have created pressing demands for objective, accurate, and accessible tools that can augment clinical decision-making while reducing dependency on subjective interpretation [6]. The evolution toward computational solutions represents a natural progression in addressing these clinical challenges.

Artificial Intelligence (AI) and Deep Learning (DL) technologies have drastically changed how we use medical imaging by providing the ability to improve diagnostic accuracy through automation [7]. By using tendon-like models (Deep Learning algorithms) that are comparable in accuracy to board-certified dermatologists, DL will allow us to better classify skin lesions accurately (benign vs malignant) than humans will most likely produce. Such algorithms use large numbers of imaging (deep imaging) datasets to identify complex patterns that may not immediately be apparent to a human eye; thus, making them an opportunity to improve inaccuracy [8]. The transition of automated methods from traditional Convolutional Neural Networks (CNNs) to more advanced architectures of CNN has shown further opportunity to improve our current capabilities in dermoscopy applications.

Emerging methodologies such as Vision Transformers and Hybrid Models in computer vision are promising methods to further advance and/or automate some of the dermoscopy applications we see today [9]. The advent of novel AI methodologies will continue to create advantages for working with Dermoscopy, yet several limitations remain, such as data scarcity and model interpretability. Most AI-driven systems currently continue to primarily utilize single-model (single architecture) methodologies, which do not encompass the entire feature representation spectrum needed for comprehensive lesion analysis. Because of this limitation, the creation of more sophisticated ensemble methodologies incorporating multiple complementary architectures continues to be needed.

The emergence of Self-Supervised Learning (SSL) has provided at least one potential solution to the issue of being able to build ML systems using SSL, thus, providing a means for learning to generalize features from a dataset where labeled images do not exist. For example, the SimCLR Framework has produced excellent performance levels associated with learning to generalize features from unlabeled images by generating augmented views of images and distinguishing between positive and negative image pairs (comparing). Thus, by allowing ML systems to learn robust feature representations without needing explicit metadata associated with the image (explicit annotations), SSL is a methodology that addresses some of the Data Limitation Challenges of the current Supervised Learning methodologies. By combining SSL techniques of pretraining with Supervised Fine-Tuning, we can now maximize the usefulness of the medical imaging dataset.

Combining multiple complementary architectures through an Ensemble Learning approach has proven to be another key area of advancement in improving the diagnostic accuracy of the upcoming AI-driven diagnostic systems beyond what a single architecture can currently produce. However, the biggest deterrent to using AI within the clinical/healthcare arena is the general lack of Identifiable/model Explainability. Clinicians need an explanation of how a model arrived at its conclusion(s) so that trust can be established in these new automated systems. To resolve this issue, the incorporation of Explainable AI techniques takes this need into consideration. For example, Grad-CAM, Shapley Value Explanation, and Attention Mechanisms allow clinicians to visually understand how AI makes decisions to help them establish trust in AI systems within the clinical environment.

The proposed research will investigate these multi-dimensional limitations within an integrated framework that effectively synergistically integrates Self-Supervised Pretraining, multi-resolution Ensemble Architecture strategies, and Advanced Explainable AI (XAI) Techniques in order to develop a clinically viable AI diagnostic system. This Hybrid Model will provide the opportunity to Learn Robust Feature Representations from Unlabeled Dermoscopy Data using Contrastive Learning techniques (SimCLR) and learn to utilize Supervised Fine-Tuning with Dynamic Hyperparameter Optimization techniques (Optuna). These Hybrid Models will be utilized via a combination of Incremental/Feature Level Fusion, and Meta Learning approaches, along with Regressor-based Ensembles (Example: LightGBM), to arrive at a Final Classification.

The major contributions of this research include:

- Integration of self-supervised pretraining using SimCLR with supervised fine-tuning to effectively leverage both labeled and unlabeled dermoscopic images, addressing the critical data scarcity problem in medical imaging.
- Development of a novel multi-architecture ensemble approach that combines EfficientNetV2-L, Swin Transformer, and ConvNeXt architectures through advanced feature-level fusion and meta-learning using LightGBM for enhanced diagnostic accuracy.
- Implementation of comprehensive explainable AI framework integrating Grad-CAM, SHAP, and attention visualization to provide clinical interpretability essential for trust-building and adoption in dermatological practice.

The remainder of this paper is organized as follows: The "Literature Review" discusses the literature-based methods of diagnosing skin cancer that have been developed using various techniques. The use of materials and techniques used for the development of our model is presented in the "Proposed Methodology" where we describe the processes used in the development of our model along with the results obtained from our model discussed in "Results". Finally, the "Conclusion" section presents the conclusions drawn from the results of this study as well as directions for future work.

2. Literature Review

In the past decade, dermatologists have seen a dramatic increase in how artificial intelligence can help identify and differentiate skin cancer. The state of the art in automated detection and classification of skin cancer via the use of machine learning techniques based on deep learning models has achieved unprecedented success. This article is a thorough review of the literature and will explore how computational methods have evolved over time in the diagnosis of skin cancer, with a focus on self-supervised learning techniques, ensemble-based systems, transfer learning applications, and explainable AI techniques that form the basis of current research in this area.

One of the most exciting developments in self-supervised learning techniques is that they represent a major advancement in the field of medical image analysis. For many years, dermatologists and other medical practitioners have been challenged by the lack of sufficient labeled data to create robust diagnostic systems [11]. The application of self-supervised pre-training followed by a supervised fine-tuning approach has been shown to have tremendous success in the area of natural image classification and has also begun to be applied in the area of medical images [12]. Recent systematic reviews of the literature have detailed the many ways self-supervised learning could contribute to the development of robust medical imaging models by allowing for the analysis of vast amounts of medical data without the use of labeled data.

The SimCLR framework has generated interest among medical imaging practitioners, based on its success in developing generalized feature representations that can be used across a wide variety of medical imaging tasks [13]. Azizi et al. demonstrated the superiority of using a self-supervised pre-training approach on medical images over the use of ImageNet as a pre-training source for the training of medical image classification models, particularly in the field of dermatology [14]. They demonstrated that the domain of study was more restricted with the use of domain-specific self-supervised pre-training, resulting in improved performance on downstream classification tasks because the domain of study was closer to the source.

Over the past several years, many deep learning architectures have been implemented and have advanced the development of skin cancer detection systems, each possessing distinct network architectures that have advantages in the processing of dermoscopic images. Specifically, convolutional neural networks have historically ruled the field of medical image analysis. Many of the traditional network architectures, including VGGNet, GoogleNet, and ResNet, are employed for the classification of skin lesions [15]. These architectures have established benchmarks for performance and laid the groundwork for the development of more recent and sophisticated approaches to the classification of skin images. In particular, the introduction of the EfficientNet architecture has provided a considerable advancement in the development of a network that is capable of generating a balanced trade-off between the accuracy of skin cancer diagnosis and the time and computational resources necessary for training the model [16].

Although many advances have been made in AI applications for skin cancer diagnosis, some of the limitations and fault lines present in current literature regarding the potential applicability of AI to skin cancer diagnosis are still unresolved. For example, the vast majority of existing systems for skin cancer diagnosis rely on single architecture models and/or simple ensemble techniques to evaluate skin lesions, resulting in a missed opportunity to capitalize on the complementary characteristics that multiple different architectural paradigms possess. Additionally, there has been limited research conducted with respect to the application of self-supervised pre-training techniques in dermatological applications. As a result, there remains an abundance of unlabeled skin imaging data that could be used to improve feature learning through the use of self-supervised learning. Furthermore, the lack of comprehensive XAI (explainable artificial intelligence) frameworks that are capable of integrating the many techniques related to XAI severely limits the clinical adoption of AI for skin cancer diagnosis. Moreover, the use of only single technique-based XAI frameworks limits the depth and breadth of the interpretability of the model. Finally, the integration of DPHO (dynamic hyperparameter optimization) with multi-architecture ensemble learning has attracted relatively little research attention in published literature and represents a significant research gap.

3. Materials and Methods

This research proposes an integrated system of Hybrid Deep Transfer (HDT) learning for improved Skin Cancer Diagnostic (SCD) capabilities utilizing Self-Supervised Pretraining, Ensemble Learning and Explainable Artificial Intelligence (EAI). The methodology includes pre-training using the SimCLR algorithm to extract image representations from extensive collections of unlabeled dermoscopy images. The HDT Model then employs end-to-end training, whereby a model will use the features extracted from EfficientNetV2-L, Swin Transformer and ConvNeXt networks, combine these into a combined representation and predict the skin pathologies via an ensemble classification model specifically designed for skin pathology detection, LightGBM. The integration of three networks allows for increased variability in image representations, reduces bias occurring from a single model and enhances the overall accuracy of diagnosing pathology. Finally, this framework applies explainable AI (EAI) techniques, utilizing techniques such as Grad-CAM and SHAP to generate interpretable predictions, facilitate decisions based on clinical reliability and ensure the model can provide transparency in the automated detection of skin cancer.

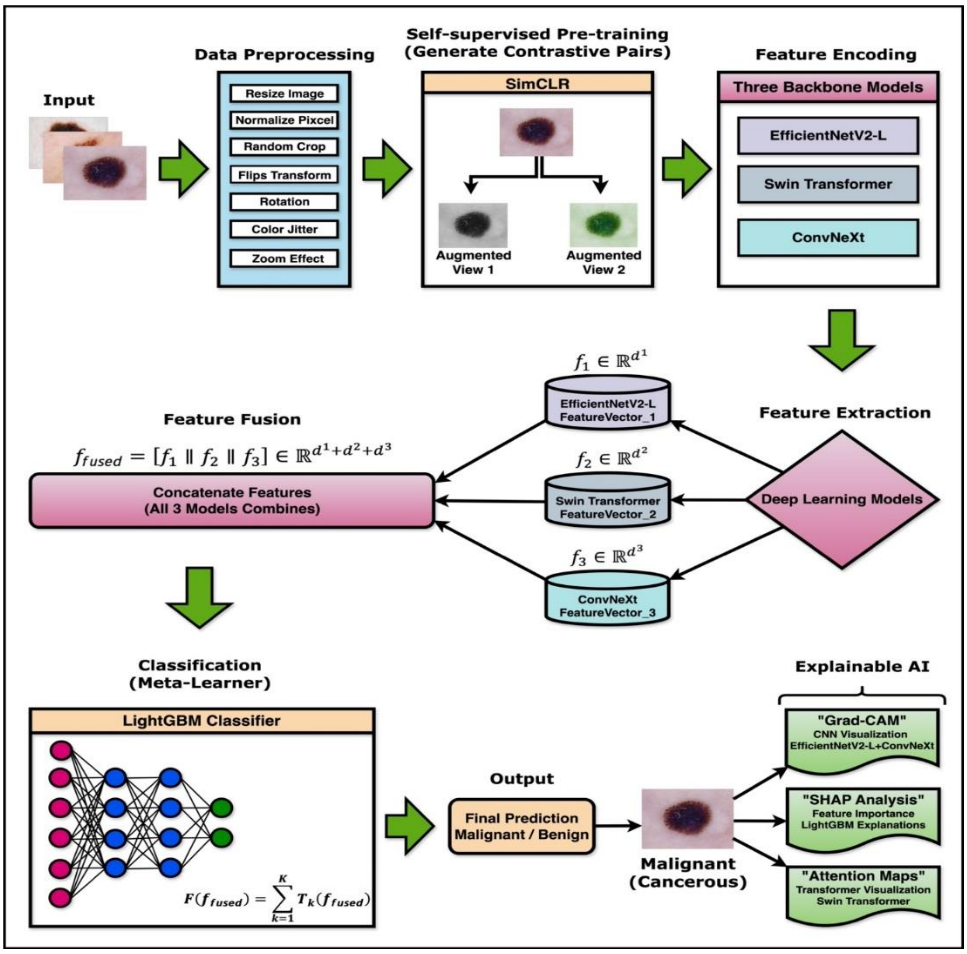


Figure 1. Proposed Model

3.1. Dataset Description

Two benchmark datasets were selected to comprehensively evaluate the methodology: the International Skin Imaging Collaboration (ISIC) 2019 database and the HAM10000 database. The ISIC 2019 database consists of 25,331 dermatoscopic photographs of skin abnormalities. Each photograph is labelled by an expert to assist in the determination of whether the skin abnormality is malignant or benign. There are eight diagnostic categories represented within the ISIC 2019 database: melanoma, melanocytic nevi, basal cell carcinoma, actinic keratoses, benign keratoses, dermatofibromas, vascular lesions, and squamous cell carcinomas [17]. The HAM10000 database adds another 10,015 images of pigmented skin lesions that also came from dermatoscopic imaging. The HAM10000 database contains images across seven skin lesion categories [18]. Additionally, the HAM10000 database contains several complementary features: diverse imaging conditions, patient demographics, and varying lesion appearances that enhance the overall robustness of the training process. The two datasets combined yield a total of 35,346 images that may be used to train and evaluate models on a diverse set of features, resulting in robust performance.

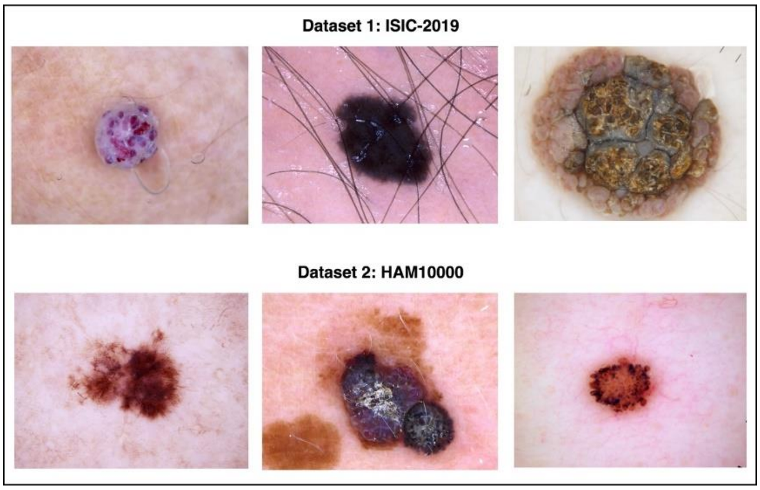


Figure 2. Dataset Image Samples.

3.2. Data Preprocessing

The use of preprocessing techniques is very beneficial for the development of strong computer vision algorithms and models. Preprocessing techniques allow for a more comprehensive method to prepare the images for further use by standardizing and augmenting them throughout training under varying image acquisition conditions. The preprocessing of images occurs in stages to ensure that the requirements of each type of modelling architecture are met while preserving the critical diagnostic information found in the image. Initially, the images will be subject to standardization procedures, which will include a resizing process to produce a standardized target spatial resolution size of 384/384 pixels, which is the resolution that provides the optimal balance between operational efficiency, computational performance and image detail. The resizing of images will be completed using the bicubic interpolation method to provide a more gradual and uniform transition from the original image size to the target resolution. Following the resizing, the pixel values of each image will be standardized by applying a normalization step based on removing the influence of colour format and/or colour space used to initially capture the images, as identified via colour space reference statistics provided by ImageNet (μ = [0.485, 0.456, 0.406]) and (σ = [0.229, 0.224, 0.225]) respectively for the red, green and blue channels of each image.

The normalization transformation is mathematically expressed as:

$$x_{norm} = \frac{x-\mu}{\sigma} \tag{1}$$

Where x represents the original pixel values and x_{norm} represents the normalized values.

Comprehensive data augmentation strategies are implemented to enhance model robustness and generalization capabilities. The augmentation pipeline includes geometric transformations such as random cropping with scale factorss $\in [0.8, 1.0]$, horizontal and vertical flipping with probability$p_{flip} = 0.5$, and rotation angles$\theta \in [-20^\circ, 20^\circ]$. Color space augmentations include brightness adjustment$\beta \in [0.8, 1.2]$, contrast modification. $\gamma \in [0.8, 1.2]$, saturation variation$\delta \in [0.8, 1.2]$, and hue shifts $\epsilon \in [-0.1, 0.1]$ [66].

The augmentation probability for each transformation is controlled by the parameter $p_{aug} = 0.7$, ensuring that augmentations are applied stochastically during training. The mathematical formulation for geometric transformations can be expressed as:

$$T_{geo}(x) = R_\theta \circ S_s \circ F_{p_{flip}}(x) \tag{2}$$

Where R_θ represents rotation, S_s denotes scaling, and $F_{p_{flip}}$ indicates flipping operations.

Algorithm 1: Data Preprocessing for Skin Cancer Diagnosis

```
Require: Raw images $x \in R^{H\times W\times C}$
Ensure: Preprocessed and augmented images $\tilde{x} \in R^{384\times 384\times 3}$
1: Input: Images from ISIC-2019 and HAM10000 datasets
2: Output: Preprocessed images $\tilde{x}$
3: for each image $x$ in datasets do
4: Resize $x$ to 384 × 384 using bicubic interpolation:
        $x \leftarrow \text{Resize}(x, 384, 384)$
5: Normalize pixel values using ImageNet statistics:
        $x_{\text{norm}} = \frac{x-\mu}{\sigma}$, $\mu = [0.485,0.456,0.406]$, $\sigma = [0.229,0.224,0.225]$
6: Apply stochastic augmentation with probability $p_{\text{aug}} = 0.7$:
  • Geometric transformations: $T_{\text{geo}}(x) = R_\theta \circ S_s \circ F_{p_{\text{flip}}}(x)$
    Where;
    – $R_\theta: \theta \in [-20°, 20°]$
    – $S_s: s \in [0.8,1.0]$
    – $F_{p_{\text{flip}}}$: horizontal/vertical flip with probability $p_{\text{flip}} = 0.5$
  • Color augmentations:
    – Brightness $\beta \in [0.8,1.2]$
    – Contrast $\gamma \in [0.8,1.2]$
    – Saturation $\delta \in [0.8,1.2]$
        Hue $\epsilon \in [-0.1,0.1]$
7: Store the preprocessed and augmented image $\tilde{x}$
8: end for
9: Return: Preprocessed and augmented images $\tilde{x}$
```

In addition to the above preprocessing steps, a critical consideration involves the handling of the multi-class nature of the original datasets. Both ISIC 2019 and HAM10000 datasets contain multiple diagnostic categories (eight for ISIC 2019 and seven for HAM10000). For the purposes of the current study, these categories were mapped to binary labels ("benign" vs. "malignant") to streamline the classification task. However, it is important to note that this mapping is performed explicitly, and each original category is carefully assigned to one of the binary classes based on clinical guidelines and prior literature. The mapping ensures consistency across datasets and facilitates direct comparison during model training and evaluation. Following the binarization, the resulting class distributions were examined and recorded to ensure transparency and to highlight potential imbalances. For instance, some categories initially representing rare conditions could result in an under-representation of the corresponding binary class, potentially affecting model performance. Therefore, class distribution statistics were computed post-mapping to guide subsequent training strategies, such as applying class-weighted loss functions or targeted data augmentation to mitigate imbalance effects. This step ensures that all images entering the model are standardized and comparable, while also providing a clear, reproducible framework for evaluation and reporting of performance metrics on both the original datasets and the derived binary classification task.

3.3. Model Architecture

The multi-tiered framework of the architecture utilizes self-supervised and supervised pre-training techniques, extraction of features via the use of multiple backbone networks, ensembling based on the extracted features, and creating classifications with an array of features to provide an optimal solution for extracting and identifying high-quality feature representations of a dataset.

3.3.1. Self-Supervised Pretraining using SimCLR

The self-supervised pretraining phase employs the SimCLR (Simple Framework for Contrastive Learning of Visual Representations) framework to learn robust feature representations from unlabeled dermoscopic images [19]. SimCLR learns representations by maximizing agreement between differently augmented views of the same data example through a contrastive loss function in the latent space.

For each input image x_i, the framework generates two augmented views $\left(x_i^{(1)}, x_i^{(2)}\right)$ using the stochastic augmentation pipeline described in Section 3.2. These augmented pairs form positive examples, while all other images in the batch serve as negative examples. The contrastive learning objective encourages the model to produce similar representations for positive pairs while maximizing the distance between negative pairs.

Each backbone architecture is modified by removing the classification head and adding a projection head $\mathrm{g}(\cdot)$ consisting of a two-layer MLP with ReLU activation and batch normalization. The projection head maps the backbone feature representations h_i to a normalized 128-dimensional contrastive learning space z_i:

$$z_i = g(h_i) = W_2 \cdot \mathrm{ReLU}\big(\mathrm{BN}(W_1 \cdot h_i + b_1)\big) + b_2 \tag{3}$$

where W_1, W_2 denotes the weight and b_1, b_2 are bias vectors.

The contrastive loss function employed is theNT-Xent loss, which encourages positive pairs to have similar representations while pushing negative pairs apart. For a batch of N examples generating 2N augmented views, the loss for a positive pair $\left(z_i, z_j\right)$ is computed as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j) / \tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k) / \tau)} \tag{4}$$

where $\mathrm{sim}\left(z_i, z_j\right) = \frac{z_i^T z_j}{||z_i||\,||z_j||}$ represents cosine similarity, $\tau = 0.07$ is the temperature parameter, and $\mathbf{1}_{[\boldsymbol{k} \neq \boldsymbol{i}]}$ is an indicator function excluding the anchor sample.

The total contrastive loss over the entire batch is:

$$\mathcal{L}\ \text{contrastive} = \frac{1}{2N} \sum_{i=1}^{2N} \mathcal{L}_{i,j(i)} \tag{5}$$

where j(i) indicates the positive pair index for sample i.

Algorithm 2: SimCLR Self-Supervised Pretraining for Skin Cancer Diagnosis

```
**Require:**
• Unlabeled dataset $D = \{x_i\}_{i=1}^{N}$
• Data augmentations $\mathcal{T}$ (e.g., crop, flip, jitter)
• Encoder network $f(\cdot)$ and projection head $g(\cdot)$
• Temperature scaling factor $\tau$
**Ensure:** Pretrained encoder $f(\cdot)$ with generalized representations
1: Create two transformed versions for each $x_i$:
       $\widetilde{x_i^{(1)}} = t_1(x_i), \quad \widetilde{x_i^{(2)}} = t_2(x_i), \quad \text{where } t_1, t_2 \sim \mathcal{T}$
2: Pass both views through the encoder:
       $h_i^{(1)} = f\left(\widetilde{x_i^{(1)}}\right), \quad h_i^{(2)} = f\left(\widetilde{x_i^{(2)}}\right)$
3: Project features to contrastive latent space:
       $z_i^{(1)} = g\left(h_i^{(1)}\right), \quad z_i^{(2)} = g\left(h_i^{(2)}\right)$
4: Compute cosine similarity between projections:
       $\mathrm{sim}\left(z_i, z_j\right) = \frac{z_i^{\top} z_j}{||z_i|| \cdot ||z_j||}$
5: Contrastive loss for a positive pair $(\boldsymbol{i}, \boldsymbol{j})$:
       $\boldsymbol{\mathcal{L}_{i,j}} = -\mathbf{log} \sum \mathbf{exp}\left(\frac{\mathrm{sim}\left(\boldsymbol{z_i}, \boldsymbol{z_j}\right)}{\boldsymbol{\tau}}\right)$
6: Average the loss over all positive pairs:
       $\boldsymbol{\mathcal{L}} = \frac{\mathbf{1}}{\mathbf{2N}} \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{N}} \left[\boldsymbol{\mathcal{L}_{2k-1,2k}} + \boldsymbol{\mathcal{L}_{2k,2k-1}}\right]$
7: Update parameters of $\boldsymbol{f}(\cdot)$ and $\boldsymbol{g}(\cdot)$ using backpropagation (e.g., Adam)
```

3.3.2. Supervised Fine-tuning with Dynamic Hyperparameter Optimization

Following self-supervised pretraining, the learned feature representations are adapted for the specific skin cancer classification task through supervised fine-tuning. The pretrained backbone networks are loaded with their SimCLR weights, and the projection heads are replaced with task-specific classification heads designed for binary classification.

The classification head architecture consists of global average pooling followed by dropout regularization and a linear classifier. The mathematical formulation for the classification head is:

$$\hat{y} = \text{softmax}\big(W_c \cdot \text{Dropout}\big(\text{GAP}(h)\big) + b_c\big) \quad (6)$$

where $\text{GAP}(\cdot)$ represents global average pooling, W_c and b_c are the classifier weights and bias, and the dropout probability p_{drop} is optimized during hyperparameter tuning. The fine-tuning process employs weighted cross-entropy loss to address class imbalance:

$$L_{CE} = -\sum_{i=1}^{N} w_{y_i} \log\big(\hat{y}_{i,y_i}\big) \quad (7)$$

where w_0 and w_1 are class weights computed as $w_j = \frac{N}{2 \cdot N_j}$ with N_j being the number of samples in class j. Dynamic hyperparameter optimization is implemented using the Optuna framework with Tree-structured Parzen Estimator (TPE) algorithms. The optimization space includes learning rates $\eta \in [10^{-5}, 10^{-2}]$ with logarithmic scaling, batch sizes $B \in \{16, 32, 64, 128\}$, weight decay values $\lambda \in [10^{-6}, 10^{-2}]$, and dropout rates $p_{\text{drop}} \in [0.1,0.5]$.

The optimization objective function is defined as:

$$\theta^* = \arg\min_{\theta} E\,[L_{\text{val}}(\theta)] \quad (8)$$

where θ represents the hyperparameter vector and $L_{\text{val}}(\theta)$ is the validation loss.

Dynamic optimizer selection evaluates four different optimizers: AdamW with decoupled weight decay, RAdam with rectified adaptive learning rates [20], Ranger combining RAdam with Lookahead optimization, and SGD with momentum and Nesterov acceleration. The selection criterion is based on validation performance after initial training epochs.

3.3.3. *Feature Extraction from Fine-tuned Models*

After supervised fine-tuning, discriminative features are extracted from each of the three backbone architectures for subsequent ensemble learning. The feature extraction process removes the final classification layers while preserving the learned feature representations from the penultimate layers.

For EfficientNetV2-L, features $f_E \in R^{1280}$ are extracted from the global average pooling layer. The Swin Transformer produces features $f_S \in R^{1536}$ from the final normalization layer before classification [21]. ConvNeXt generates features $f_C \in R^{2048}$ from the global average pooling layer preceding the classifier. The extracted features undergo standardization using z-score normalization to ensure consistent scales across different architectures:

$$f_{\text{norm}} = \frac{f - \mu_f}{\sigma_f} \quad (9)$$

where μ_f and σ_f are the mean and standard deviation computed from the training set features.

3.3.4. *Ensemble Learning through Feature Fusion and Meta-learning*

The ensemble learning component uses a two-stage process to be a sophisticated approach that combines feature-level fusion and meta-learning for the final classification using LightGBM as the classifier [22]. The early fusion approach is paired with a stacked ensemble methodology that takes advantage of the complementary strengths of the three backbone architectures in order to maximize performance.

Feature-level fusion is implemented through concatenation of the standardized feature vectors:

$$f_{\text{fused}} = \big[f_{E,\text{norm}} \oplus f_{S,\text{norm}} \oplus f_{C,\text{norm}}\big] \in R^{4864} \quad (10)$$

where $\oplus$ denotes concatenation operation and the resulting fused feature vector has dimensionality $d_{\text{fused}} = 1280 + 1536 + 2048 = 4864$.

A meta-learner utilizing LightGBM, which implements gradient boosting using advanced optimization methods, processes these combined features. By utilizing a combination of leaf-flourishing tree growth and histogram-based algorithms, LightGBM is able to provide highly efficient training options for models formed from data that has been pre-combined into features. The mathematical formulation for the gradient boosting process is:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (11)$$

where $F_m(x)$ is the ensemble prediction after mm m iterations, $h_m(x)$ is the m-th weak learner, and γ_m is the step size determined through line search.

The objective function for LightGBM optimization includes both loss and regularization terms:

$$L_{LightGBM} = \sum_{i=1}^{N} l\big(y_i, F_m(x_i)\big) + \sum_{j=1}^{m} \Omega\big(h_j\big) \quad (13)$$

where $l(\cdot)$ is the loss function and $\Omega(\cdot)$ represents regularization terms controlling model complexity.

Hyperparameter optimization for LightGBM explores parameters including number of leaves $L \in [31, 511]$, learning rate $\eta_{gbm} \in [0.01, 0.3]$, maximum depth $d_{max} \in [3, 15]$, feature fraction $f_{\text{frac}} \in [0.6, 1.0]$, and regularization parameters $\lambda_{L1}, \lambda_{L2} \in [0, 10]$.

Algorithm 3: LightGBM Classification for Skin Cancer Diagnosis

```
Require:
• Labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$
• Number of boosting rounds $T$
• Learning rate $\eta$
• Maximum tree depth $d$
• Loss function $L(y, \hat{y})$
Ensure: Trained LightGBM classifier $F_T(x)$
1: Initialize the model:
       $F_0(x) = \arg\min_{c} \sum_{i=1}^{N} L(y_i, c)$
2: for $t = 1$ to $T$ do
3: Compute gradients and hessians for each $x_i$:
       $g_i = \frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)}, \quad h_i = \frac{\partial^2 L(y_i, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)^2}$
4: Train regression tree $h_t(x)$ using $(g_i, h_i)$:
• Histogram-based split finding
• Leaf-wise tree growth
• Gradient-based One-Side Sampling (GOSS)
• Depth limited to $d$
5: Update the ensemble:
       $F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x)$
6: end for
7: Output the final model:
       $\hat{y} = F_T(x)$
```

3.3.5. *Classification Output*

The final classification output combines the probabilistic predictions from the LightGBM meta-learner with confidence estimation mechanisms. The model produces binary classification probabilities $P(y = 1 \mid x)$ and $P(y = 0 \mid x)$ for malignant and benign classes respectively.

The final prediction is determined by:

$$\hat{y} = \arg\max_{c \in \{0,1\}} P(y = c \mid f_{\text{fused}}) \tag{14}$$

Confidence scores are computed using prediction entropy:

$$\text{Confidence} = 1 - H(P) = 1 + \sum_{c \in \{0,1\}} P(y = c \mid x) \log P(y = c \mid x) \tag{15}$$

where $H(P)$ represents the entropy of the prediction distribution.

3.4. Explainable AI (XAI) Implementation

Explainable AI has integrated several techniques to enable comprehensive insight into how the model is making its decisions. In order to fulfil the requirement for interpretability in clinical applications, a multi-modal approach to explainability combines both visual explanations and an overall quantitative measure of the importance of each feature within the model. Through this combination of both visual and quantitative metrics, clinical users can improve their clinical understanding of and build their trust in the model output. Gradient-weighted Class Activation Mapping (Grad-CAM) is implemented for CNN-based architectures (EfficientNetV2-L and ConvNeXt) to generate visual explanations. Grad-CAM computes the gradient of the target class score with respect to feature maps in the final convolutional layer:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \quad (16)$$

where α_k^c denotes the importance weight for feature map k and class c, $A_{i,j}^k$ is the activation at spatial location (i,j) in feature map k, and Z is the factor of normalization.

$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (17)$$

For the Swin Transformer architecture, attention visualization leverages the multi-head self-attention mechanism. The attention weights from the final transformer block are aggregated across heads and spatial dimensions:

$$A_{\text{avg}} = \frac{1}{H} \sum_{h=1}^{H} A^{(h)} \quad (18)$$

where H is the number of attention heads and $A^{(h)}$ represents the attention matrix for head h.

SHAP (SHapley Additive exPlanations) analysis is applied to the LightGBM meta-learner to quantify feature contributions. SHAP values satisfy the efficiency property:

$$\sum_{i=1}^{d_{\text{fused}}} \phi_i = f(x) - E[f(X)] \quad (19)$$

where ϕ_i represents the SHAP value for feature i, and $f(x)$ is the model prediction.

4. Experiments and Results

This section presents a detailed analysis of the experimental findings derived from an extensive series of evaluations conducted to assess the performance and clinical utility of the proposed diagnostic framework. The system's effectiveness was rigorously tested across three benchmark dermoscopic datasets obtained from previously published studies, enabling a comprehensive comparison of its diagnostic accuracy, robustness, and generalization capability.All experiments were conducted in a computationally intensive environment using NVIDIA RTX 4090 Graphics Processing Units (GPUs) to maximize training and evaluation efficiency. The framework was implemented in PyTorch 2.0 with CUDA 11.8 support, enabling full GPU acceleration. Mixed precision training was employed to reduce memory overhead and accelerate training while maintaining numerical stability.

Hyperparameter optimization was performed using Optuna. After 100 trials with median pruning, the optimal values identified for the final model were: learning rate = 1e-4, batch size = 32, weight decay = 1e-5, and dropout = 0.3. Cross-entropy loss on the validation set was used as the objective metric, and validation splits guided the hyperparameter selection. The data preprocessing and augmentation pipeline was executed with the following optimal settings: images were resized to 384×384 pixels using bicubic interpolation and normalized with ImageNet statistics (μ = [0.485, 0.456, and 0.406], σ = [0.229, 0.224, 0.225]). Geometric augmentations included random cropping with scale s = 0.9, horizontal and vertical flipping with probability p_flip = 0.5, and rotation θ = ±15°. Color augmentations included brightness β = 1.1, contrast γ = 1.0, saturation δ = 1.05, and hue shift ε = 0.05. All augmentations were applied stochastically with probability p_aug = 0.7.These optimal settings were used consistently across all datasets to ensure reproducibility and maximize model performance, and all reported metrics reflect the results obtained with this configuration.

The proposed skin cancer diagnostic model was evaluated on both the ISIC 2019 and HAM10000 datasets using multiple performance metrics, including Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC, to provide a comprehensive assessment of its diagnostic capability. On the ISIC dataset, the model achieved an accuracy of 93.66% (95% CI: 92.10–95.12%), precision of 91.12%, recall of 91.97%, F1-score of 91.54%, specificity of 94.21%, and ROC-AUC of 0.967. Similarly, on the HAM10000 dataset, the model obtained an accuracy of 95.34% (95% CI: 94.10–96.58%), precision of 93.98%, recall of 94.56%, F1-score of 94.27%, specificity of 96.12%, and an ROC-AUC of 0.981.Bootstrap resampling with 1,000 iterations was used to compute the confidence intervals, ensuring that the reported metrics reflect their statistical reliability. Threshold selection for binary classification was determined using the Youden's J statistic to balance sensitivity and specificity. The results indicate that the inclusion of self-supervised pretraining, multi-architecture ensemble learning, and explainable AI techniques not only improves overall accuracy but also maintains high sensitivity and specificity, highlighting the robustness and generalizability of the framework across diverse dermoscopic datasets.

In another experiment, confusion matrices were utilized to assess the effectiveness of the proposed framework in accurately classifying benign and malignant skin lesions, as shown in Figure 5. The model demonstrated a notable average accuracy of 94.50% across both ISIC and HAM10000 datasets, highlighting its strong capability to distinguish between different types of skin lesions with high reliability.

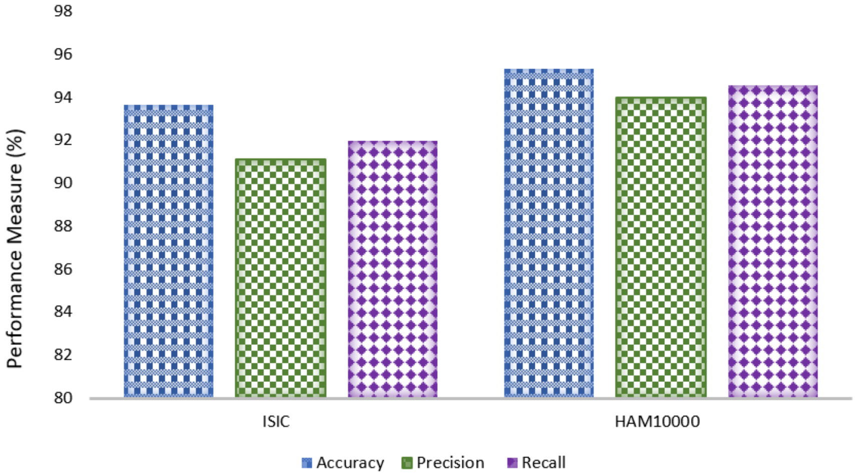
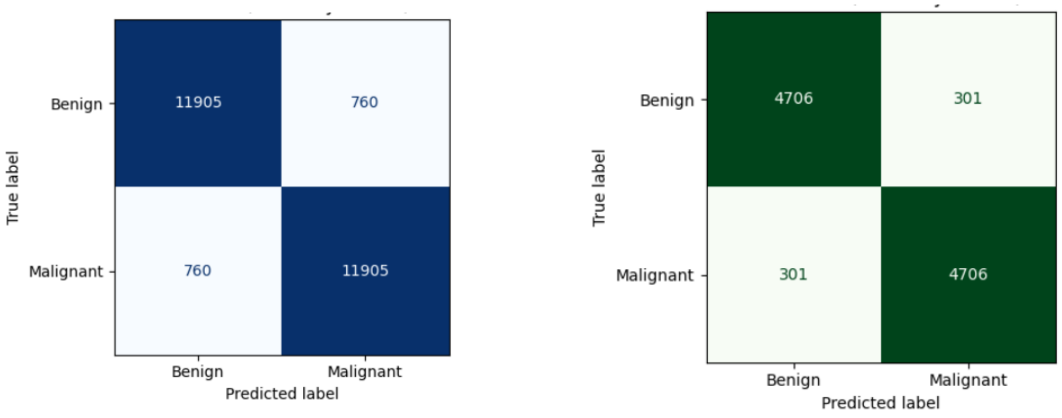


Figure 3. Proposed Model results in terms of Accuray, Precision and Recall.



ISIC 2019 Dataset HAM10000 Dataset

Figure 4. Confusion Metrics of ISIC 2019 and HAM10000 Datasets

Grad-CAM and attention visualizations were generated for sample dermoscopic images to assess the clinical relevance of the explainability modules. SHAP values from the meta-learner were analyzed to interpret the contribution of individual features in ensemble predictions. The highlighted regions were compared with lesion segmentation masks using the Dice coefficient, achieving an average overlap of 0.82 ± 0.05, indicating that the model focuses on clinically relevant areas. Example heatmaps and attention overlays are shown in Figure X. Failure cases were also examined, highlighting instances where the model attended to non-lesion regions, providing insight into limitations and areas for future improvement. These results demonstrate that the proposed framework is both accurate and interpretable in a clinically meaningful manner.

For comparative analysis, three baseline approaches were selected from published studies to provide a fair reference: Baseline 1 [23] proposed a Computer-Assisted Diagnosis (CAD) framework using a lightweight CNN architecture for early detection of skin diseases; Baseline 2 [24] evaluated skin lesion classification using four CNN-based architectures—DenseNet, MobileNetV2, Xception, and InceptionResNetV2—assessing their effectiveness in a comprehensive framework; Baseline 3 [25] employed a combination of CNNs, Residual Networks, and Xception models to detect skin diseases early, emphasizing both accuracy and robustness. All baseline models were re-implemented under identical experimental conditions, including the same dataset splits, preprocessing pipeline, and augmentation strategies used for the proposed framework. Each model was trained and evaluated over five independent runs to account for stochastic variability, and the mean ± standard deviation of the performance metrics was recorded.

The proposed framework achieved an accuracy of 94.5% ± 0.62%, outperforming Baseline 1 (88.45% ± 0.75%), Baseline 2 (90.15% ± 0.68%), and Baseline 3 (92.98% ± 0.55%). The precision of the proposed model was 92.55% ± 0.71%, compared with 86.45% ± 0.80%, 88.62% ± 0.72%, and 91.61% ± 0.60% for the respective baselines, while the recall reached 93.26% ± 0.65%, exceeding Baseline 1 (87.45% ± 0.78%), Baseline 2 (89.22% ± 0.70%), and Baseline 3 (91.45% ± 0.59%).These improvements are attributable to the inclusion of self-supervised pretraining (SimCLR) for robust feature representation, multi-architecture ensemble learning (EfficientNetV2-L, Swin Transformer, ConvNeXt) to reduce model bias, and explainable AI methods (Grad-CAM, SHAP) to focus on clinically relevant regions for accurate lesion localization. By controlling for experimental conditions and reporting variance across runs, the observed performance gains reflect methodological superiority rather than differences in setup.

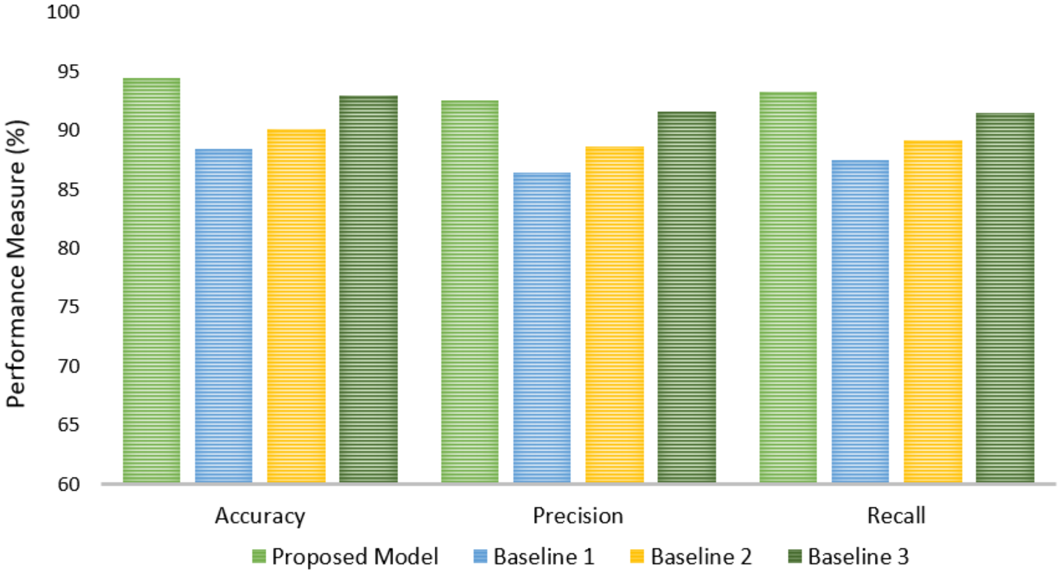


Figure 5. Comparative Analysis of Proposed Model with Baselines.

To evaluate the contribution of each component in the proposed framework, an ablation study was conducted on both ISIC 2019 and HAM10000 datasets. The study systematically removed or modified key components, including self-supervised pretraining (SimCLR), the multi-architecture ensemble, and explainable AI modules (Grad-CAM and SHAP), and measured the resulting impact on classification performance. This approach allows a quantitative assessment of how each module contributes to the overall effectiveness of the framework.

Table 1. Model evaluation and results

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Full Framework	94.92 ± 0.60	93.27 ± 0.68	93.91 ± 0.63	93.59 ± 0.65
Without SimCLR	91.85 ± 0.65	90.05 ± 0.72	91.25 ± 0.68	90.65 ± 0.70
Single Backbone (No Ensemble)	92.55 ± 0.62	90.85 ± 0.68	91.60 ± 0.64	91.22 ± 0.66
Without Explainable AI	93.80 ± 0.61	92.30 ± 0.67	92.75 ± 0.64	92.52 ± 0.65

The results indicate that the full proposed framework consistently outperforms all ablated variants, with the highest average accuracy (94.92%) and F1-score (93.59%). Removing self-supervised pretraining (SimCLR) causes the largest drop in performance, highlighting its critical role in learning robust feature representations from unlabeled dermoscopic images. The multi-architecture ensemble contributes significantly by capturing complementary features from different backbones, while the explainable AI modules further enhance precision and recall by focusing the model on clinically relevant regions. These findings confirm that the superior performance of the proposed framework is the result of the combined contribution of all its components.

5. Conclusions

This study proposed a hybrid deep learning framework for skin cancer diagnosis that integrates self-supervised pretraining, multi-architecture ensemble learning, and explainable artificial intelligence (XAI) techniques. The framework leverages SimCLR-based contrastive learning to generate rich and discriminative feature representations from large unlabeled dermoscopic datasets, which are subsequently fine-tuned through supervised learning across three advanced architectures—EfficientNetV2-L, Swin Transformer, and ConvNeXt. A meta-learning classifier based on LightGBM fuses the features from these architectures, achieving enhanced diagnostic performance. Experimental evaluations conducted on benchmark datasets (ISIC and HAM10000) demonstrated that the proposed framework significantly outperformed existing baseline models. These results confirm the model's robustness, sensitivity, and reliability in distinguishing between benign and malignant skin lesions. Moreover, the integration of explainable AI methods such as Grad-CAM and SHAP provided meaningful visual and feature-level explanations, enhancing model interpretability and clinical trust—an essential component for real-world medical applications. Future research will focus on integrating multimodal data (e.g., patient metadata and histopathology images), developing lightweight real-time versions for resource-limited settings, and enhancing model explainability and uncertainty estimation to improve clinical reliability and trust.

References

1. Vidya, M.; Karki, M.V. Skin cancer detection using machine learning techniques. IEEE Int. Conf. Electron. Comput. Commun. Technol. (CONECCT) 2020, 1–5.
2. Murugan, A.; Nair, S.A.H.; Preethi, A.A.P.; Kumar, K.S. Diagnosis of skin cancer using machine learning techniques. Microprocess. Microsyst. 2021, 81, 103727.
3. Naqvi, M.; Gilani, S.Q.; Syed, T.; Marques, O.; Kim, H.C. Skin cancer detection using deep learning—A review. Diagnostics 2023, 13(11), 1911.
4. Bistroń, M.; Piotrowski, Z. Comparison of machine learning algorithms used for skin cancer diagnosis. Appl. Sci. 2022, 12(19), 9960.
5. Bhatt, H.; Shah, V.; Shah, K.; Shah, R.; Shah, M. State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: A comprehensive review. Intell. Med. 2023, 3(3), 180–190.
6. Abdulazeez, A. A review on utilizing machine learning classification algorithms for skin cancer. J. Appl. Sci. Technol. Trends 2024, 5(2), 60–71.
7. Pacal, I.; Ozdemir, B.; Zeynalov, J.; Gasimov, H.; Pacal, N. A novel CNN-ViT-based deep learning model for early skin cancer diagnosis. Biomed. Signal Process. Control 2025, 104, 107627.
8. Farea, E.; Saleh, R.A.; AbuAlkebash, H.; Farea, A.A.; Al-antari, M.A. A hybrid deep learning skin cancer prediction framework. Eng. Sci. Technol. Int. J. 2024, 57, 101818.
9. Naeem, A.; Anees, T.; Khalil, M.; Zahra, K.; Naqvi, R.A.; Lee, S.W. SNC_Net: Skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images. Mathematics 2024, 12(7), 1030.
10. Patil, P.R. Deep learning revolution in skin cancer diagnosis with hybrid Transformer–CNN architectures. Vidhyayana Int. Multidiscip. Peer-Rev. E-J. 2025, 10(SI4).
11. Huang, S.C.; Pareek, A.; Jensen, M.; Lungren, M.P.; Yeung, S.; Chaudhari, A.S. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. NPJ Digit. Med. 2023, 6(1), 74.
12. Shurrab, S.; Duwairi, R. Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Comput. Sci. 2022, 8, e1045.
13. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. Proc. Int. Conf. Mach. Learn. 2020, 1597–1607.
14. Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Norouzi, M. Big self-supervised models advance medical image classification. Proc. IEEE/CVF Int. Conf. Comput. Vis. 2021, 3478–3488.
15. Shahin, A.H.; Kamal, A.; Elattar, M.A. Deep ensemble learning for skin lesion classification from dermoscopic images. Proc. 9th Cairo Int. Biomed. Eng. Conf. (CIBEC) 2018, 150–153.
16. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proc. Int. Conf. Mach. Learn. 2019, 6105–6114.
17. Combalia, M.; Codella, N.; Rotemberg, V.; Carrera, C.; Dusza, S.; Gutman, D.; Malvehy, J. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: The 2019 International Skin Imaging Collaboration Grand Challenge. Lancet Digit. Health 2022, 4(5), e330–e339.
18. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 2018, 5(1), 1–9.
19. Kaczmarek, E.; Szeto, J.; Nichyporuk, B.; Arbel, T. Building a general SimCLR self-supervised foundation model across neurological diseases to advance 3D brain MRI diagnoses. Proc. IEEE/CVF Int. Conf. Comput. Vis. 2025, 1310–1319.
20. Sajid, M., Khan, A. H., Malik, K. R., Khan, J. A., & Alwadain, A. (2025). A new approach of anomaly detection in shopping center surveillance videos for theft prevention based on RLCNN model. PeerJ Computer Science, 11, e2944.
21. Liu, Y.; He, S.; Liu, Z.; Guo, J.; Peng, Z.; Wu, L.; Wu, Y. RAdam–backpropagation-based model for predicting propped fracture conductivity. SPE J. 2025, 30(3), 1024–1036.
22. Yao, D., & Shao, Y. (2024). A data efficient transformer based on Swin Transformer. The Visual Computer, 40(4), 2589-2598.
23. Sajid, M., Malik, K. R., Khan, A. H., Fuzail, M., & Li, J. (2025). TVAE-3D: Efficient multi-view 3D shape reconstruction with diffusion models and transformer based VAE. Cluster Computing, 28(13), 854.
24. Lokker, C.; Abdelkader, W.; Bagheri, E.; Parrish, R.; Cotoi, C.; Navarro, T.; Iorio, A. Boosting efficiency in a clinical literature surveillance system with LightGBM. PLOS Digit. Health 2024, 3(9), e0000299.

25. Malik, S.G.; Jamil, S.S.; Aziz, A.; Ullah, S.; Ullah, I.; Abohashrh, M. High-precision skin disease diagnosis through deep learning on dermoscopic images. Bioengineering 2024, 11(9), 867.
26. Saha, D.K. An extensive investigation of convolutional neural network designs for the diagnosis of lumpy skin disease in dairy cows. Heliyon 2024, 10(14).
27. Khan, A. H., Li, J., Asghar, M. N., & Iqbal, S. (2025). LGD_Net: Capsule network with extreme learning machine for classification of lung diseases using CT scans. Plos one, 20(8), e0327419.
28. Sudar, K.M.; Nagaraj, P.; Muneeswaran, V.; Panda, B.; Bhoi, A.K. DermoClassify: A dermatologist skin disease detection and classification using DCNN. Res. Biomed. Eng. 2025, 41(1), 13.