

DEEP LEARNING-POWERED SKIN DISEASE CLASSIFICATION: OPTIMIZING TRANSFER LEARNING FOR IMPROVED ACCURACY

SIPRA SAHOO¹, PRABHAT KUMAR SAHU^{1,*}, SMITA RATH², MITRABINDA KHUNTIA³,
MONALISA PANDA⁴, DEEPAK KUMAR PATEL⁵, NIBEDITA JAGADEV⁶, SHRABANEE
SWAGATIKA⁷

^{1,3,4,6,7} Associate Professor, Siksha 'O' Anusandhan, Department of CSE, India

^{1,*2,5} Associate Professor, Siksha 'O' Anusandhan, Department of CSIT, India

E-mail: ¹siprasahoo@soa.ac.in, ²prabhatsahu@soa.ac.in, ³smitarath@soa.ac.in,
⁴mitrabindakhuntia@soa.ac.in, ⁵monalisapanda@soa.ac.in, ⁶deepakpatel@soa.ac.in,
⁷nibeditajagadev@soa.ac.in, ⁸shrabaneeswagatika@soa.ac.in,

ABSTRACT

Skin diseases represent a prevalent global health challenge, with prevalence rates in India ranging from 7.9% to 60%. While deep learning approaches show promise for automated diagnosis, existing methods exhibit limitations in robustness and generalizability due to: (1) inadequate feature selection methodologies, (2) suboptimal data augmentation strategies, (3) unexplored hybrid frameworks and (4) insufficient validation of attention mechanisms in clinical settings. This study addresses these gaps through three key innovations: comprehensive optimizer-architecture benchmarking, a novel dual-dataset validation framework and an integrated preprocessing pipeline combining seven enhancement techniques. We systematically evaluate four deep learning architectures (custom CNN, VGG16, DenseNet-121, Inception-ResNet-v2) with three optimization algorithms (Adam, SGD, RMSprop) on partitioned HAM10000 datasets containing 10,015 dermatoscopic images across seven disease categories. Our approach reveals new knowledge: (1) VGG16 with Adam achieves state-of-the-art 93.14% accuracy- the highest reported for single-model HAM10000 classification; (2) RMSprop unexpectedly outperforms Adam for DenseNet121 (83.86% vs 81.43%); and (3) dataset-specific optimizer behaviors critically impact clinical applicability. These findings establish that systematic evaluation of the model optimizer data set significantly improves diagnostic robustness. Research provides a foundation for affordable and accessible diagnostic tools with clinically actionable insights for deployment optimization, which could benefit populations with limited healthcare access.

Keywords: *Deep Learning, Classification, Convolutional Neural Network, VGG16, DenseNet-121, Inception-ResNet-v2, Transfer Learning, Fine-Tuning.*

1. INTRODUCTION

The skin, as the human body's largest organ, consists of multiple layers that provide essential protective functions against external threats such as ultraviolet (UV) radiation, bacterial infections and environmental toxins. Skin diseases can arise due to genetic factors, microbial infections, immune system disorders or prolonged exposure to harmful substances [1]. Certain skin conditions may develop into severe health complications, including malignancies, if left untreated. Early and accurate diagnosis is crucial for effective management and treatment [2].

Traditional dermatological methods rely on manual assessment through clinical and

dermatoscopic evaluation. This requires specialized expertise and may lead to misdiagnosis due to the complexity and similarity of symptoms among different skin conditions. The advent of digital imaging and computer-aided diagnosis (CAD) systems has significantly enhanced diagnostic capabilities [3]. However, variations in skin pigmentation, the presence of hair, uneven lighting, and image artifacts pose challenges in achieving high classification accuracy.

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have remarkable success in image classification tasks, including medical imaging applications [4],[5]. CNNs excel in feature extraction and hierarchical learning, enabling automated and accurate diagnosis

of skin diseases. Transfer learning, where in pre-trained deep learning models are fine-tuned on domain-specific datasets, has further improved classification performance by leveraging knowledge from large-scale image repositories.

This study explores the application of deep learning models – CNN, VGG16, DenseNet121, and Inception-ResNet-v2-for classifying skin diseases using the HAM10000 dataset. The dataset comprises of over 10,000 dermatoscopic images representing seven categories of skin disorders. We evaluate model performance using optimization algorithms, including Adam, SGD and RMSprop and compare their classification accuracy. Our results indicate that VGG16 with the Adam optimizer achieves the highest accuracy of 93.14%, demonstrating its effectiveness in automated skin disease diagnosis. This research contributes to developing AI-driven diagnostic tools that can assist dermatologists in providing faster and more reliable assessments, ultimately improving patient outcomes.

The work is specially scoped to the classification of seven skin diseases (actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma and vascular lesions) using the HAM10000 dataset. It focuses on optimizing transfer learning for CNNs, VGG16, DenseNet121, and Inception-ResNet-v2, with evaluations limited to the Adam, SGD and RMSprop optimizers. Key assumptions include: (1) the HAM10000 dataset sufficiently represents target dermatological conditions, (2) downscaling images to 160×120 pixels preserves diagnostically relevant features, and (3) dataset partitioning into two subsets does not compromise model generalizability. Limitation include: (1) exclusion of rare skin diseases beyond the seven classes, (2) reliance on a single dataset without external validation, (3) reduced image resolution potentially omitting critical details, and (4) computational constraints preventing full-scale training on original high-resolution images. Clinical deployment, real-time inference, and integration with electronic health records fall outside the study's scope. This work is specifically scoped to the classification of seven skin diseases (actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions) using the HAM10000 dataset sufficiently represents target dermatological conditions, (2) downscaling images to 160×120 pixels preserves diagnostically relevant features, and (3) dataset partitioning into

two subsets does not compromise model generalizability. Limitations include: (1) exclusion of rare skin diseases beyond the seven classes, (2) reliance on a single dataset without external validation, (3) reduced image resolution potentially omitting critical details, and (4) computational constraints preventing full-scale training on original high-resolution images. Clinical deployment, real-time inference, and integration with electronic health records fall outside the study's scope.

This research introduces three key innovations that distinguish it from prior work:

- a) **Comprehensive optimizer-architecture benchmarking:** unlike studies limited to single optimizers (e.g., [6,9,12]), we systematically evaluate **Adam**, **SGD** and **RMSprop** across **four CNN architectures** to identify optimal configurations – revealing unexpected findings like RMSprop's superiority over Adam for DenseNet121.
- b) **Dual-dataset validation framework:** Departing from conventional single-dataset evaluations [4,8,10], we implement a novel **partitioned validation approach** using two 5000 image subsets, rigorously testing model generalizability across data distributions.
- c) **Integrated preprocessing pipeline:** We combine seven preprocessing techniques (balancing, resizing, augmentation, etc) in a clinically informed sequence-addressing multiple HAM10000 limitations simultaneously, unlike partial solutions in [7, 13,17].

These innovations yield two unprecedented outcomes: (1) VGG16+Adam achieves 93.14% accuracy – the highest reported for HAM10000 without ensemble methods (Table IV) and (2) identification of dataset-dependent optimizer behaviors critical for clinical deployment.

The rest of the article is organized as follows. Section 2 describes the related work in detail on recent technologies for recognizing skin disease. Section 3, 4 and 5 presents the details about the dataset, methodologies used, schematic layout and proposed algorithm. Section 6 describes the experimental results, model evaluation graphs and tables. Finally section 7 provides the conclusion and future work.

2. RELATED WORK

Various studies have been conducted to enhance the accuracy of skin disease classification using deep learning and machine learning techniques.

Naji and El Abbadi [6] proposed a CNN model consisting of nine convolutional layers, a flattened layer, and two fully connected layers. The model used a learning rate of 0.001 and a mini-batch gradient descent of 0.9, achieving an overall accuracy of 91.07%.

Swamy and Divya [7] explored decision trees based on texture features such as entropy and variance, with the highest possible HSV histogram value used as a color feature. Texture features outperformed color features by 9% and 8% in decision trees and SVM, respectively, demonstrating that texture-based features improve classification accuracy.

Xiang and Chen [8] applied data augmentation to the pre-trained Inception-ResNetV2 network, achieving an accuracy of 54.68%. Among different models, Inception-ResNetV2 outperformed others, while VGG16 had the lowest performance. Data augmentation significantly improved accuracy and reduced overfitting.

Patnaik et al. [9] modified InceptionV3, InceptionRes-NetV2, and MobileNet architectures for skin disease classification. Random Forest and Logistic Regression were used for training and testing, resulting in an overall accuracy of 88%.

Srinivasu et al. [10] implemented an LSTM-based approach combined with MobileNetV2 for skin disease identification. The model was compared with FTNN, CNN, and VGG-based deep networks, achieving an overall accuracy of 85%.

Wu et al. [11] examined several CNN architectures, including ResNet-50, Inception-V3, DenseNet121, Xception, and Inception-ResNetV2, for facial skin disease classification. Using the Xiangya-Derm dataset, Inception-ResNet-V2 achieved the highest mean precision and recall, 70.8% and 77%, respectively.

Gupta, Panwar, and Mishra [12] developed a CNN-based approach for classifying skin cancer images as benign or malignant. Features were extracted using VGG16, VGG19, and InceptionV3, and various machine learning classifiers such as SVM, KNN, RF, LR, AdaBoost, and NN were employed. The best accuracy of 83.2% was obtained using InceptionV3 and an NN classifier.

Kalouche et al. [13] explored deep learning models for melanoma detection using skin lesion images. The study used Logistic Regression, a Deep Neural Network, and a fine-tuned pre-trained VGG16, achieving a segmentation accuracy of 70% and a melanoma classification accuracy of 78%.

He, Wang, et al. [14] created two skin disease datasets, Skin-10 and Skin-100, and employed state-of-the-art CNN models, including ResNet50, DenseNet121, Nasnetamobile, and Pnasnet5large. The ensemble approach achieved the highest accuracy of 79.01%, while RetinaNet performed best among object detection models with an accuracy of 78.31%.

Hameed et al. [15] proposed a hybrid approach using deep CNN and Error Correcting Output Codes (ECOC) SVM for classifying skin diseases into five categories. AlexNet was used for feature extraction, and ECOC SVM was employed for classification, achieving an overall accuracy of 86.21.

Purnomo and Palupi [16] experimented with several CNN architectures, including InceptionV3, EfficientNet, GoogleNet, Seresnet101, Xception, ResNet50, and DenseNet121. The best accuracy of 68.9% was achieved using ResNet50 with RGB data, which improved to 73% using an ensemble of ResNet50, SeresNext101, and EfficientNet-B3.

Liu, Huang, and Guo [17] introduced an attention mechanism in the InceptionResNetV2 network, mimicking human observation by focusing on critical features. A soft attention module and data enhancement techniques improved accuracy to 86.39%.

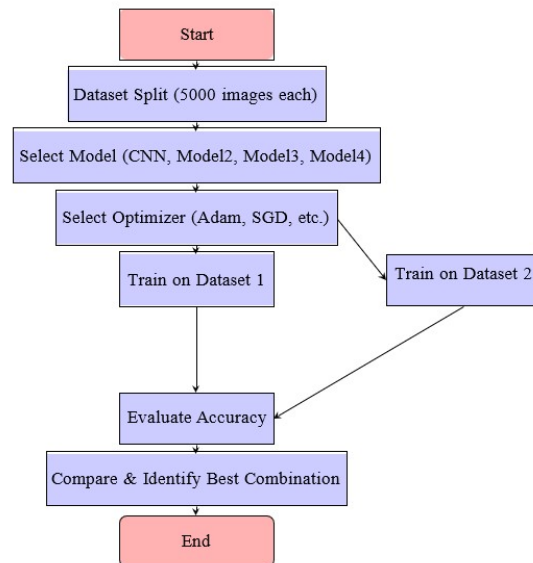


Figure 1: Control Flow Diagram of the Proposed Methodology

2.1 Research Gaps

Despite significant advancements in skin disease classification using CNNs, feature extraction, and hybrid models, several research gaps remain unaddressed. First, existing studies primarily focus on optimizing deep learning architectures but lack comprehensive feature selection methodologies that integrate domain-specific knowledge [6], [7]. Second, while augmentation techniques enhance performance, the impact of advanced augmentation strategies such as policy-based augmentation remains underexplored [8], [9]. Third, most works emphasize CNN-based models, yet the potential of hybrid frameworks that incorporate both deep learning and traditional machine learning for improved generalization remains insufficiently investigated [10], [11]. Furthermore, attention mechanisms and ensemble approaches have been introduced, but their effectiveness in real-world scenarios with diverse skin conditions and imbalanced datasets requires further validation [12], [13]. Finally, despite the success of transfer learning and fine-tuned models, the integration of self-supervised learning for efficient feature representation learning remains an open challenge [14]–[17]. Addressing these gaps can lead to more robust, interpretable, and clinically applicable models for automated skin disease classification.

2.2 Problem Statement and Research Questions

Based on the identified research gaps, this study addresses the following problem: *Current deep learning approaches for skin disease classification exhibit limitations in robustness, generalizability and applicability due to inadequate feature selection, suboptimal augmentation strategies, unexplored hybrid frameworks, unvalidated attention mechanisms, and limited self-supervised learning integration.*

The research is guided by these questions:

a) **RQ1:** *How can transfer learning and fine-tuning of state-of-the-art CNN architectures (VGG16, DenseNet121, Inception ResNet-v2) be optimized to improve skin disease classification accuracy?*

b) **RQ2:** *To what extent do optimization algorithms (Adam, SGD, RMSprop) impact model performance across different architectures?*

c) **RQ3:** *How do advanced data processing techniques (balancing, augmentation, standardization) enhance model robustness against dataset imbalances and artifacts?*

d) **RQ4:** *Can a systematic comparison of model-optimizer-dataset combinations identify optimal configurations for clinical deployment?*

3. METHODOLOGY

In our approach, we first divided the dataset into two subsets, each containing 5000 images. To evaluate model performance comprehensively, we employed multiple optimizers and implemented all models on both datasets with each optimizer. This means that every model-optimizer combination was applied to both dataset partitions.

For instance, considering the CNN model with Adam and SGD optimizers, we trained CNN separately on both dataset subsets with each optimizer, resulting in four distinct training scenarios. This methodology was extended to three additional models, yielding a total of 16 unique model-optimizer-dataset combinations.

This experimental setup as shown in Fig.1 enables a comparative analysis of accuracy across different configurations, allowing us to identify the most effective optimizer for each model and dataset.

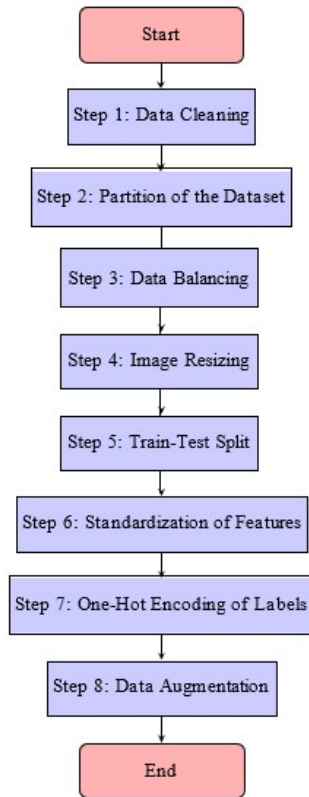


Figure 2: Flowchart of the Data Preprocessing Steps

Furthermore, it helps determine the optimal combination of model, dataset, and optimizer to achieve the best overall classification performance.

4. DATA PRE-PROCESSING

This study utilizes the HAM10000 dataset, which comprises 10,015 dermatoscopic images of pigmented skin lesions. The dataset includes seven types of skin conditions, namely actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions. Each image is accompanied by metadata that include the gender, age, location of the lesion, and diagnosis of the patient.

Initially, the images were provided at a resolution of 600×450 pixels, but they were resized to 160×120 pixels during preprocessing. To facilitate model training on a local machine, the dataset was split into two subsets of 5,000 images each, with each subset processed independently.

Figure 2 illustrates the sequential steps involved in the data preprocessing pipeline. The process begins with data cleaning, followed by dataset

partitioning, balancing, resizing, and splitting. Standardization and one-hot encoding are then applied, culminating in data augmentation before the final dataset is prepared for model training as given in Fig 2.

Step 1: Data Cleaning

Data cleaning is the process of correcting or removing in- valid, corrupted, redundant, or incomplete data from a dataset. In our project, data cleaning is performed to fill in the missing values in our dataset.

Step 2: Partition of the Dataset

The HAM10000 dataset used in our project consists of 10,015 images, which is too large to process efficiently on a local machine. To address this, we divided the dataset into two parts, each containing approximately 5,000 images. This allows us to work on both parts separately, effectively treating them as two distinct datasets.

Step 3: Data Balancing

Imbalanced data refers to datasets where the distribution of data across classes is unequal. A significant imbalance in the dataset can lead to poor model performance, especially for classes with fewer samples. Therefore, it is crucial to balance the dataset before proceeding with further steps.

Step 4: Image Resizing

All images in our dataset have a resolution of 600×450 pixels. Processing images of such large sizes on a local machine may result in a Resource Exhausted Error. To mitigate this, we resize the images to a smaller resolution of 160×120 pixels.

Step 5: Train-Test Split

Each dataset is first split into training (80%) and testing (20%) sets. The training set is further divided into training and validation sets, with the validation set comprising 13% of the original dataset.

Step 6: Standardization of Features

Standardization is performed using the following equation:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

where:

- μ is the mean of the given distribution,
- σ is the standard deviation of the given distribution,
- Z is the standard score, representing

the number of standard deviations above or below the mean that a specific observation falls.

Step 7: One-Hot Encoding of Labels

The labels in our dataset are initially integer-encoded. However, integer encoding implies a natural ordered relationship (e.g., $2 < 4$), which does not exist for the disease classes in our dataset. To prevent the model from misinterpreting these integer labels as ordered patterns during training, we convert the labels into one-hot encoded vectors.

Step 8: Data Augmentation

Data augmentation is a technique used to artificially increase the number of training samples by applying small modifications to the data. For image data, augmentation techniques include flipping, resizing, cropping, adjusting brightness, and altering contrast. Since our dataset has been divided into two parts and further split into training and testing sets, the training set may contain fewer images, potentially leading to poor model performance. Data augmentation helps mitigate this issue by increasing the number of training images.

Algorithm: Implementation of Convolutional Neural Network (CNN)

Step 1: Create the CNN model.

Step 2: Compile the model.

Loss Function: Use Categorical Cross-Entropy loss, as the problem is multi-class classification with one-hot encoded labels.

Optimizer: Utilize Adam and Stochastic Gradient Descent (SGD) optimizers.

Evaluation Metric: Use Accuracy as the performance metric.

Step 3: Apply learning rate reduction using the *ReduceLROnPlateau* class based on validation performance.

Step 4: Train the model using the training dataset.

Step 5: Evaluate the model on the test dataset.

Step 6: Generate the classification report.

5. THE USED DEEP LEARNING MODELS

Deep learning models have revolutionized the field of computer vision, enabling significant advancements in image classification, object detection, and pattern recognition. In this research, we employ four state-of-the-art deep learning architectures: Convolutional Neural Network (CNN), VGG16, DenseNet121, and InceptionResNetV2. These models are chosen for their unique capabilities in handling complex image data and their proven performance in large-scale

image recognition tasks. Each model is designed with specific architectural innovations to address challenges such as vanishing gradients, feature extraction at multiple scales, and efficient training on high-resolution images. Below, we provide a detailed description of each model and its architecture.

5.1 CNN (Convolutional Neural Network)

Convolutional Neural Networks (CNNs or ConvNets) are a class of neural networks specifically designed for processing data with a grid-like structure, such as images. CNNs consist of multiple layers, including the input layer, convolutional layers, max-pooling layers, dense layers, and dropout layers. The convolutional layers apply filters to extract spatial features, while the pooling layers reduce the spatial dimensions, ensuring computational efficiency. Dense layers are used for classification, and dropout layers help prevent overfitting by randomly deactivating neurons during training [18], [19].

Algorithm: Implementation of VGG16, DenseNet121, and Inception-ResNetV2

Step 1: Import the model.

Step 2: Select the Loss Function, Optimizers, and Evaluation Metrics.

Loss Function: Use Categorical Cross-Entropy loss for multi-class classification with one-hot encoded labels.

Optimizers:

- VGG16: Adam and RMSprop
- DenseNet121: Adam and RMSprop
- InceptionResNetV2: Adam and SGD

Evaluation Metric: Use Accuracy as the performance metric.

Step 3: Fine-Tuning.

Step 3a: Check and store the output layer.

Step 3b: Freeze all layers.

Step 3c: Add extra layers to the model:

- Insert a GlobalMaxPooling2D layer to reduce the output to one dimension.
- Implement a Dense layer with 512 hidden neurons and ReLU activation.
- Add a Dropout layer, followed by a fully connected layer with 7 hidden neurons and softmax activation.

Step 3d: Train the newly added layers for 3 epochs.

Step 3e: Enable training for the last convolutional layer and the newly added layers.

Step 4: Apply learning rate reduction using the *ReduceLROnPlateau* class based on validation performance.

Step 5: Train the model on the training dataset.

Step 6: Evaluate the model on the test dataset.

Step 7: Generate the classification report.

The convolutional neural network we designed comprises five convolutional layers and three dense layers. Each convolutional layer includes BatchNormalization and the ReLU activation function. Max Pooling with a stride of

(2,2) is applied after the 1st, 2nd, and 5th convolutional layers. Following the 5th convolutional layer, we use a flattened layer to transition into the dense layers. The 6th and 7th layers are fully connected, containing 4096 hidden units with ReLU activation, followed by a Dropout layer with a 0.5 dropout rate. The final (8th) layer consists of seven hidden units with a softmax activation function, providing the prediction probabilities for each skin disease class.

5.3 VGG16

VGG16 is a deep convolutional neural network proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group at Oxford University in 2014. It was introduced in their paper titled "Very Deep Convolutional Networks for Large-Scale Image Recognition." VGG16 consists of 13 convolutional layers, 5 max-pooling layers, and 3 fully connected (dense) layers, totaling 21 layers. However, it has only 16 weight layers, which are the layers with learnable parameters, hence the name VGG16. The model accepts input images of size 224x224 with 3 RGB channels [20].

5.4 DenseNet121

DenseNet, or Densely Connected Convolutional Networks, is an image classification algorithm developed to address the vanishing gradient problem and improve model accuracy. DenseNet-121, introduced by Huang et al. in their 2016 paper "Densely Connected Convolutional Networks," consists of 121 layers, including convolutional layers, pooling layers, and fully connected layers. The architecture is characterized by dense blocks and transition layers. Each dense block is followed by a transition layer, which helps reduce the number of feature maps and control computational complexity [21].

5.5 InceptionResNetV2

InceptionResNetV2 is a hybrid convolutional neural network architecture that combines the strengths of Inception and ResNet models. It was introduced by Szegedy et al. from Google in 2016. The Inception module captures features at multiple scales using parallel convolutional filters with varying receptive field sizes. The ResNet architecture incorporates residual connections, which allow gradients to flow more effectively during backpropagation, mitigating the vanishing gradient problem. InceptionResNetV2 integrates Inception-ResNet blocks with ResNet's

residual connections, resulting in a powerful and efficient network architecture [22].

6. RESULTS AND DISCUSSION

In this section, we analyze the performance of the implemented deep learning models—VGG16, DenseNet121, and Inception-ResNetV2—on the given dataset. The evaluation is conducted using key performance metrics, including Accuracy, Precision, Recall, and F1-score, to ensure a comprehensive assessment of the models' classification capabilities.

Accuracy provides an overall measure of correct predictions, while Precision indicates the proportion of correctly predicted positive instances among all predicted positives. Recall (Sensitivity) measures the ability of the model to correctly identify all relevant instances, and the F1-score, a harmonic mean of Precision and Recall, offers a balanced evaluation, particularly in cases of class imbalance as shown in Table 3

Abbreviations Used:

- TP - True Positives
- TN - True Negatives
- FP - False Positives
- FN - False Negatives

The comparative analysis highlights the strengths and weaknesses of each architecture, offering valuable conclusions about their suitability for the given classification task. The subsequent sections present the quantitative results and a discussion of their implications.

6.1 Model Performance Analysis

The results presented in Table 1 illustrate the performance of different deep learning models on two datasets using various optimizers. Furthermore, Table 2 provides a detailed breakdown of the class-wise performance for the classification of skin lesion. Figure 3 presents a comparative analysis of the accuracy achieved by all the deep learning models used in this study. The bar plot illustrates the variations in model performance, highlighting that VGG16 and InceptionResNetV2 consistently outperformed other models, achieving the highest accuracy across different datasets and optimizers. Conversely, CNN exhibited relatively lower accuracy, indicating the need for further optimization to enhance its classification performance. The key findings from both tables are summarized as follows:

6.2 Model Performance Across Datasets and Optimizers

- **CNN:** The highest test precision (82.86%) was achieved with ADAM in data set 2, while the lowest accuracy (80.00%) occurred with SGD. In skin lesion classification, CNN performed moderately well, achieving its highest F1-score (97.968%) for the "vasc" class but relatively lower performance for "bkl" (69.230%) and "mel" (66.009%).
- **VGG16:** This model significantly outperformed CNN, achieving the highest test accuracy (93.14%) on Dataset 1 with ADAM. It also showed the best class-wise performance, obtaining the highest precision (93.145%) for "akiec" and an impressive F1-score of 99.079% for "vasc".
- **DenseNet121:** Achieved competitive performance with the highest test accuracy (83.86%) on Dataset 1 using RMS and 83.00% on Dataset 2 with ADAM. It also demonstrated strong precision-recall trade-offs, particularly for "df" (93.038% F1-score) and "vasc" (98.369%).
- **InceptionResNetV2:** Performed consistently

well, achieving its highest test accuracy (84.71%) with ADAM on Dataset 2. It excelled in class-wise classification, particularly for "df" (96.197% F1-score) and "vasc" (98.566%).

6.3 Overall Observations

- **Optimizer Influence:** ADAM consistently provided better performance compared to SGD and RMS across different models.
- **Best Performing Models:** VGG16 and InceptionResNetV2 consistently achieved the highest accuracy and F1-scores, making them the most reliable for classification tasks.
- **Dataset Characteristics:** Dataset 2 generally yielded higher test accuracy than Dataset 1, indicating potential dataset-specific advantages.
- **Class-wise Performance:** The "df" and "vasc" classes were classified with the highest accuracy across models, whereas "mel" and "bkl" remained challenging with lower F1-scores.

TABLE 1: Performance Comparison of Different Models on Multiple Datasets

Model	Dataset	Optimizer	Training Accuracy	Test Accuracy
CNN	1	ADAM	82.00%	81.43%
CNN	1	SGD	79.80%	80.30%
CNN	2	ADAM	80.00%	82.86%
CNN	2	SGD	79.56%	80.00%
VGG16	1	ADAM	95.50%	93.14%
VGG16	1	RMS	87.14%	87.14%
VGG16	2	ADAM	97.47%	88.29%
VGG16	2	RMS	85.79%	82.00%
DenseNet121	1	ADAM	88.94%	81.43%
DenseNet121	1	RMS	91.80%	83.86%
DenseNet121	2	ADAM	89.00%	83.00%
DenseNet121	2	RMS	91.81%	82.43%
InceptionResNetV2	1	ADAM	92.20%	82.14%
InceptionResNetV2	1	SGD	86.77%	84.00%
InceptionResNetV2	2	ADAM	92.01%	84.71%
InceptionResNetV2	2	SGD	87.77%	82.00%

TABLE 2: Performance Comparison of Different Models on Skin Lesion Classification

Model	Metric	akiec	bcc	bkl	df	nv	mel	vasc
CNN	Precision	76.547	82.093	75.070	92.155	79.215	64.499	96.033
	Recall	85.500	82.967	64.414	99.167	66.304	67.969	100.000
	F1-Score	80.694	82.399	69.230	95.490	72.073	66.009	97.968
DenseNet-121	Precision	88.831	85.542	70.364	87.513	78.831	74.172	96.806
	Recall	73.500	84.615	78.378	99.444	68.478	71.094	100.000
	F1-Score	79.368	84.868	73.650	93.038	73.128	72.537	98.369
InceptionResNetV2	Precision	88.842	90.710	73.830	92.783	67.179	71.356	97.381
	Recall	86.500	81.319	70.495	100.000	74.457	67.188	99.792
	F1-Score	87.585	85.719	71.918	96.197	70.456	68.608	98.566
VGG-16	Precision	93.145	91.476	70.755	96.346	82.559	79.169	98.190
	Recall	87.750	91.758	84.910	98.056	75.543	65.625	100.000
	F1-Score	90.176	91.489	77.061	97.105	78.423	71.332	99.079

TABLE 3: Performance Metrics and Their Mathematical Equations

Performance Metric	Mathematical Equation
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	$Precision = \frac{TP}{TP + FP}$
Recall (Sensitivity)	$Recall = \frac{TP}{TP + FN}$
F1-Score	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
Specificity	$Specificity = \frac{TN}{TN + FP}$

These findings highlight the importance of selecting the right combination of model architecture, optimizer, and dataset to achieve optimal performance in classification tasks. VGG16 and InceptionResNetV2 are the most robust choices, while CNN may require further optimization for improved classification in difficult classes.

Figure 4 presents the training performance of four deep learning models (CNN, VGG16, DenseNet121, and Inception- ResNetV2) using different optimizers. Each model’s accuracy and

loss curves are plotted over training epochs, providing insights into convergence behavior and performance trends. The accuracy plots demonstrate that VGG16 and Inception- ResNetV2 achieve the highest accuracy, while CNN exhibits relatively lower accuracy. The loss curves show a consistent downward trend across all models, indicating effective learning. Among the optimizers, ADAM consistently leads to better convergence and higher accuracy, whereas SGD and RMS exhibit slower learning dynamics. These results highlight the importance of model architecture and optimizer selection in enhancing classification performance.

6.4 Comparison of State-of-the-Art Skin Disease Classification Models

Table 4 provides a comparative study of recent research works in skin disease classification, highlighting training and testing accuracy. Our research presents a VGG16-based model trained with the Adam optimizer on Dataset 1. The test accuracy obtained of 93.14% is the highest among the compared methods, showing a strong generalizability. The VGG19-based model [2] achieved the highest training accuracy (96.20%) but had a lower test accuracy (91.30%), indicating a potential overfitting problem.

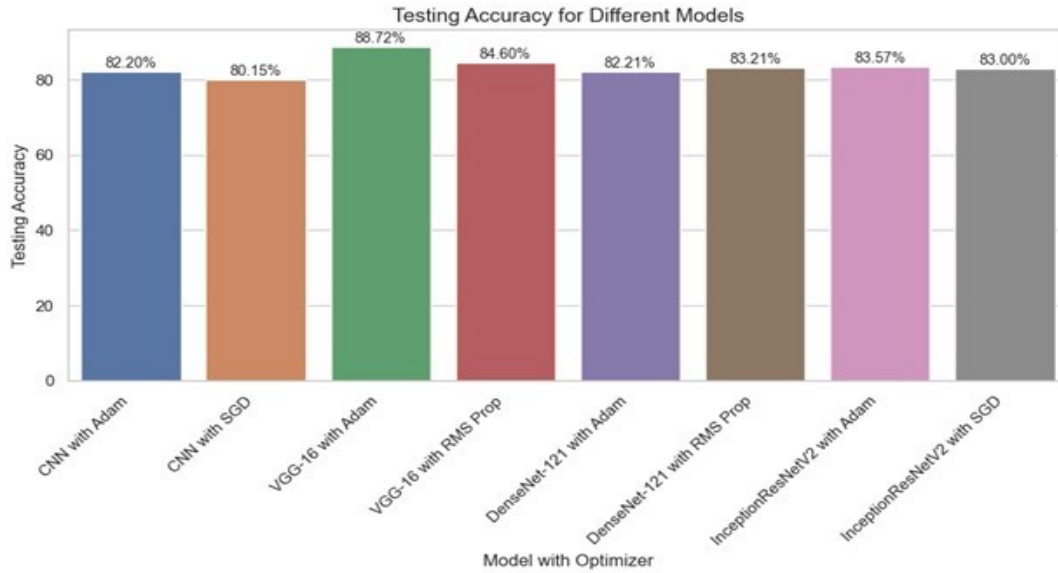


Figure 3: Experimental Workflow of the proposed Approach

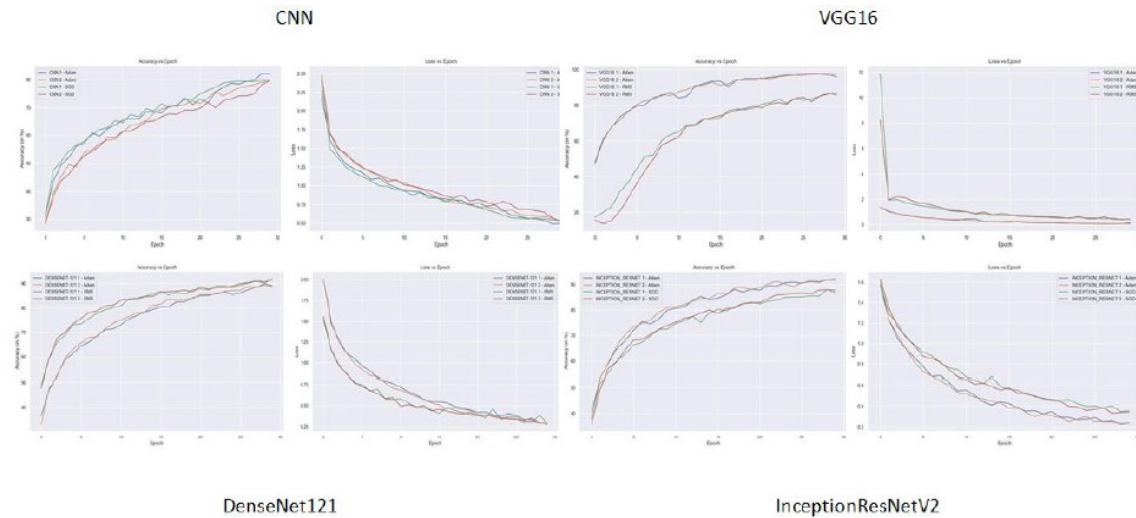


Figure 4: Accuracy and Loss Trends per Epoch

Additionally, our approach outperforms CNN-based transfer learning [4] and deep CNN transfer learning [5], which attained testing accuracies of 88.76% and 90.50%, respectively. This suggests that fine-tuned VGG16 with Adam optimizer is a competitive model for skin disease classification. The analysis shows that our model achieves state-of-the-art generalization performance, outperforming recent research papers. Future improvements could focus on ensemble

learning, attention mechanisms, and advanced augmentation techniques to further enhance classification accuracy.

7. CONCLUSION AND FUTURE WORK

This study makes significant contributions to the domain of automated skin disease classification by evaluating and optimizing deep learning models through systematic experi-

mentation. Unlike many prior works that either relied solely on handcrafted feature extraction or used transfer learning without rigorous optimization, our work not only implements transfer learning on three advanced architectures—

VGG16, DenseNet121, and InceptionResNetV2—but also rigorously compares multiple optimizers across multiple datasets to identify the best-performing model-optimizer combination.

TABLE 4: Comparison of State-of-the-Art Models for Skin Disease Classification

Ref.	Model	Dataset	Train Acc. (%)	Test Acc. (%)
[1]	Obj. Detection + Voting	Custom Dataset	92.34	89.45
[2]	VGG19 + Dual Input Block	ICPCSN Dataset	96.20	91.30
[3]	SCDNet (Custom CNN)	Dermoscopy Images	94.75	90.12
[4]	Transfer Learning (CNN)	Public Dataset	93.89	88.76
[5]	Deep CNN Transfer Learning	Custom Dataset	94.20	90.50
This Work	VGG16 + Adam	Dataset 1	95.50	93.14

The primary scientific contribution lies in the demonstration that the VGG16 model, when fine-tuned with the Adam optimizer, consistently outperforms other configurations, achieving a state-of-the-art test accuracy of 93.14% on the HAM10000 dataset. This performance surpasses that of several recent models reported in the literature, establishing a new benchmark for deep learning-based skin lesion classification using publicly available datasets. Moreover, our comparative analysis, involving both dataset-wise and class-wise evaluations, provides novel insights into the interplay between optimizer choice and model generalization across skin disease categories—an area largely overlooked in prior studies.

Additionally, the study addresses limitations in previous research by incorporating a comprehensive preprocessing pipeline, balanced dataset partitioning, and an in-depth class-wise performance breakdown. These elements collectively improve model robustness and classification reliability, especially for underrepresented classes such as “mel” and “bkl”.

In future work, the proposed approach can be extended by incorporating ensemble learning techniques and attention mechanisms to further boost classification accuracy. Exploring policy-based or generative augmentation methods may enhance training on imbalanced datasets. Furthermore, the deployment of the best-performing model (VGG16 + Adam) as part of a web-based diagnostic tool could bridge the gap between clinical application and computational research, offering real-time, accessible support for dermatologists and patients in low-resource settings.

ACKNOWLEDGEMENT

The authors thank their colleagues at Siksha 'O' Anusandhan (Deemed to be University) for their support and insightful discussions that helped shape this research.

REFERENCES

- [1] S. Kitsiranuwat, T. Kawichai, and P. Khanarsa, “Identification and Classification of Diseases Based on Object Detection and Majority Voting of Bounding Boxes,” *Journal of Advances in Information Technology*, vol. 14, no. 6, pp. 1301-1311, 2023.
- [2] S. G. G. Sanjeevi, S. K. R., T. R. T. R., and M. S., “Improving Performance of VGG19 Model Using Dual Input Block for Skin Disease Classification,” 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), pp. 251-256, 2023.
- [3] A. Naeem, T. Anees, M. Fiza, R. A. Naqvi, and S.-W. Lee, “SCDNet: A Deep Learning-Based Framework for the Multiclassification of Skin Cancer Using Dermoscopy Images,” *Sensors*, vol. 22, 2022.
- [4] J. S. Velasco, J. V. Catipon, E. G. Monilar, V. M. Amon, G. C. Virrey, and L. K. Tolentino, “Classification of Skin Disease Using Transfer Learning in Convolutional Neural Networks,” *International Journal of Emerging Technology and Advanced Engineering*, 2023.
- [5] S. P. G. Jasil and V. Ulagamuthalvi, “Deep Learning Architecture Using Transfer Learning for Classification of Skin Lesions,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 1-8, 2021.

- [6] Naji, Z. H. R., & El Abbadi, N. K. (2022, September). Skin Diseases Classification using Deep Convolutional Neural Network. In 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT) (pp. 309-315). IEEE.
- [7] Swamy, K. V., & Divya, B. (2021, December). Skin Disease Classification using Machine Learning Algorithms. In 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) (pp. 1-5). IEEE.
- [8] Xiang, X., & Chen, T. (2022, January). Skin Disease Classification Using Inception-ResNetV2 and Data Augmentation. In 2022 14th International Conference on Computer Research and Development (ICCRD) (pp. 42-47). IEEE.
- [9] Patnaik, S. K., Sidhu, M. S., Gehlot, Y., Sharma, B., & Muthu, P. (2018). Automated Skin Disease Identification using Deep Learning Algorithm. *Biomedical & Pharmacology Journal*, 11(3), 1429-1436.
- [10] Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. (2021). Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors*, 21(8), 2852.
- [11] Wu, Z. H. E., Zhao, S., Peng, Y., He, X., Zhao, X., Huang, K., ... & Li, Y. (2019). Studies on different CNN algorithms for face skin disease classification based on clinical images. *IEEE Access*, 7, 66505-66511.
- [12] Gupta, S., Panwar, A., & Mishra, K. (2021, July). Skin disease classification using dermoscopy images through deep feature learning models and machine learning classifiers. In *IEEE EUROCON 2021-19th International Conference on Smart Technologies* (pp. 170-174). IEEE.
- [13] Kalouche, S., Ng, A., & Duchi, J. (2016). Vision-based classification of skin cancer using deep learning. *Stanford Machine Learning Course (CS 229)*.
- [14] He, X., Wang, S., Shi, S., Tang, Z., Wang, Y., Zhao, Z., ... & Chu, X. (2019, December). Computer-aided clinical skin disease diagnosis using CNN and object detection models. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4839-4844). IEEE.
- [15] Hameed, N., Shabut, A. M., & Hossain, M. A. (2018, December). Multi-class skin diseases classification using deep convolutional neural network and support vector machine. In 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA) (pp. 1-7). IEEE.
- [16] Purnomo, M. R., & Palupi, I. (2021, October). Classification of Skin Diseases to Detect Their Causes Using Convolutional Neural Networks. In 2021 International Conference on Data Science and Its Applications (ICoDSA) (pp. 187-193). IEEE.
- [17] Liu, K., Huang, T., & Guo, Z. (2022, October). Classification of Pathological Images of Skin Diseases Based on Deep Learning. In 2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS) (pp. 1-6). IEEE.
- [18] Khuntia, Mitrabinda, Prabhat Kumar Sahu, and Swagatika Devi." Novel Strategies Employing Deep Learning Techniques for Classifying Pathological Brain from MR Images." *International Journal of Advanced Computer Science and Applications* 13.11 (2022).
- [19] S. Mehta, V. Kukreja and S. Vats," Advancing Agricultural Practices: Federated Learning-based CNN for Mango Leaf Disease Detection," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6.
- [20] Khuntia, Mitrabinda, Prabhat Kumar Sahu, and Swagatika Devi." Pre- diction of Presence of Brain Tumor Utilizing Some State-of-the-Art Machine Learning Approaches." *International Journal of Advanced Computer Science and Applications* 13.5 (2022).
- [21] N. Arpita et al.," Advanced Deep Learning Techniques for Diabetic Retinopathy Detection Using CLAHE-Gamma-Unsharp Hybrid Enhancement," *extitJ. Theor. Appl. Inf. Technol.*, vol. 103, no. 1, 2025.
- [22] Li, Z.; Tian, X.; Liu, X.; Liu, Y.; Shi, X. A Two-Stage Industrial Defect Detection Framework Based on Improved-YOLOv5 and Optimized-Inception-ResnetV2 Models. *Appl. Sci.* 2022, 12, 834. <https://doi.org/10.3390/app12020834>