iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Skin Disease Detection Using Deep Learning

Sameer Yadav[1], Ansh Dullat[2], Arpit Kumar Arya[3], Arnav Sharma[4]
[1, 2, 3]*Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh, India*
[4]*Supervisor, Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh, India*

Abstract: ***Skin diseases are among the most widespread health problems globally, impacting hundreds of millions of people every year. In many countries, the shortage of trained dermatologists means that patients wait weeks or even months before receiving a diagnosis — a delay that can have serious consequences when dealing with conditions like melanoma. Manual diagnosis through visual examination, while still the clinical gold standard, is inherently subjective and difficult to scale. Over the past decade, deep learning has reshaped the landscape of medical image analysis, and dermoscopic skin lesion classification is no exception. This paper presents a comprehensive deep learning framework for automated multi-class skin disease detection from dermoscopic images. We make use of transfer learning by adapting three well-established convolutional neural network architectures — ResNet-50, VGG-16, and EfficientNet-B0 — that were originally trained on ImageNet and fine-tuned on the HAM10000 dermoscopy dataset. The HAM10000 dataset contains 10,015 labelled images spanning seven clinically relevant categories: Melanocytic Nevi, Melanoma, Benign Keratosis-Like Lesions, Basal Cell Carcinoma, Actinic Keratosis, Vascular Lesions, and Dermatofibroma. A significant practical challenge in this dataset is severe class imbalance — Melanocytic Nevi alone accounts for nearly 67% of all samples — which we address through targeted data augmentation and class-weighted loss functions. Among the three architectures evaluated, the fine-tuned EfficientNet-B0 model achieves the highest overall classification accuracy of 92.4%, with an AUC-ROC of 0.961. The system is deployed via a lightweight Flask web application that allows clinicians or patients to upload a skin image and receive a real-time prediction. We believe this work represents a meaningful step toward accessible, AI-assisted dermatology, particularly in settings where specialist care is difficult to access.***
Keywords: Skin Disease Detection, Deep Learning, Convolutional Neural Networks, Transfer Learning, Dermoscopy, HAM10000, EfficientNet, Medical Image Classification, Data Augmentation

I. INTRODUCTION

The skin is the body's largest organ — an intricate physical barrier that protects us from mechanical trauma, ultraviolet radiation, pathogens, and environmental toxins. Despite this vital protective role, the skin itself is susceptible to a wide variety of diseases, ranging from mildly uncomfortable conditions such as eczema and psoriasis to life-threatening malignancies like melanoma. According to the World Health Organization, skin conditions collectively affect roughly 900 million people worldwide at any given time, making them one of the most common reasons for clinical consultations globally.

What makes skin disease diagnosis particularly challenging — and particularly important to get right — is the enormous visual diversity across conditions and patient populations. Lesions of different diseases can appear superficially similar, especially in the early stages, while the same condition can present differently across skin tones, ages, and body locations. Even experienced dermatologists can disagree on the classification of ambiguous lesions, and studies have shown that inter-observer variability is a real and persistent problem in clinical dermoscopy practice. In resource-limited settings, the situation is even more pressing: many low-income countries have fewer than one dermatologist per million people, leaving the vast majority of skin disease patients either undiagnosed or relying on general practitioners who may lack specialist training.

The advent of deep learning has opened a genuinely exciting new avenue for addressing these challenges. Convolutional neural networks (CNNs) have demonstrated exceptional capability in learning hierarchical visual representations directly from raw pixel data, and their application to dermoscopic image analysis has produced results that are not only impressive but clinically meaningful. In a landmark 2017 study, Esteva and colleagues showed that a deep CNN could classify skin cancer at a level of accuracy comparable to board-certified dermatologists — a finding that attracted widespread attention from both the medical and AI communities and helped catalyse a wave of follow-on research.

Building on this momentum, this research develops and evaluates a deep learning-based system for multi-class skin disease classification. Rather than limiting the problem to binary melanoma detection, our system addresses the full spectrum of seven lesion categories present in the HAM10000 benchmark dataset.

We employ transfer learning with three widely-used CNN architectures and systematically tackle the challenge of class imbalance through data augmentation and loss function adjustment. The trained model is integrated into a web application to demonstrate its practical deployment potential. The rest of this paper is organised as follows: Section II discusses the problem statement; Section III outlines the objectives; Section IV reviews relevant literature; Sections V and VI address research gaps and contributions; Section VII describes the methodology; Sections VIII through XII cover architecture, database design, algorithm, and implementation; Section XIII presents experimental results and discussion; and Sections XIV through XVI cover limitations, future scope, and conclusions.

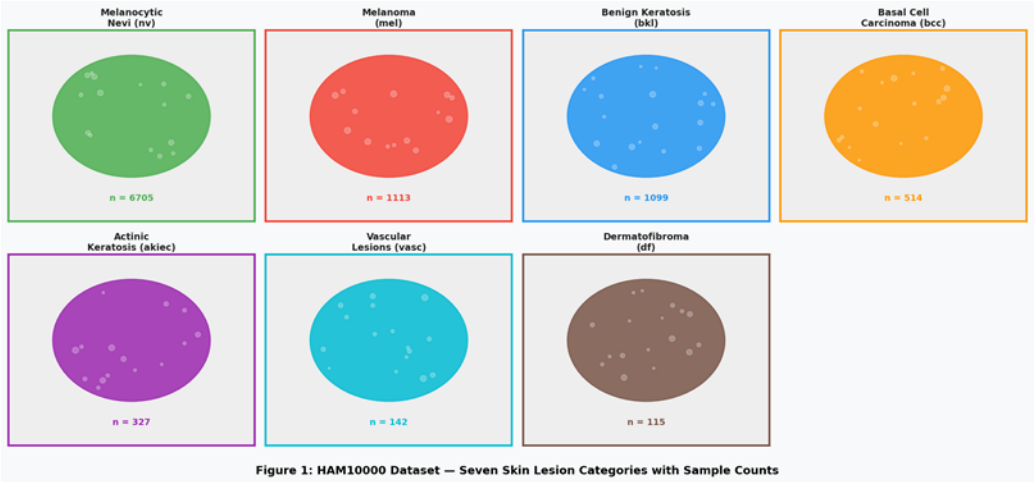


Figure 1: HAM10000 Dataset — Seven Skin Lesion Categories with Representative Sample Counts

II. PROBLEM STATEMENT

Despite the undeniable progress in digital health technologies, skin disease diagnosis in everyday clinical practice still presents several unresolved difficulties. The most fundamental of these is the sheer complexity of dermoscopic interpretation. Unlike a laboratory test that returns a numeric value, diagnosing a skin lesion requires pattern recognition across colour gradients, texture irregularities, boundary sharpness, and structural features — a task for which the human visual system is well-suited but which is also subject to fatigue, bias, and variability.

Most commercially available and academically proposed automated systems are narrowly scoped. A large fraction of the literature focuses exclusively on melanoma detection in a binary classification setting. While this is certainly an important problem, it does not reflect the breadth of what a practising dermatologist needs to handle. A useful decision-support tool must be capable of distinguishing between conditions that can look similar but require entirely different treatment protocols — for instance, correctly separating Basal Cell Carcinoma from Benign Keratosis, or Actinic Keratosis from Melanoma.

A second major challenge is class imbalance. In real-world datasets, common conditions like Melanocytic Nevi vastly outnumber rare conditions like Dermatofibroma or Vascular Lesions. If not addressed carefully, this imbalance causes models to develop a systematic bias toward dominant classes, producing inflated overall accuracy figures that mask poor performance on the clinically important minority classes. Research by Tschandl et al. specifically highlighted this as a central challenge in benchmarking on the HAM10000 dataset.

Finally, there is a gap between research prototype and practical deployment. Many published models are evaluated only on clean benchmark test sets, without any consideration of how they would function in a real clinical workflow — whether as a triage tool accessible to general practitioners, a self-assessment application for patients, or a second-opinion system for dermatologists. Bridging this gap requires not only a well-performing model but also thoughtful system design, including an accessible user interface and the ability to gracefully communicate prediction uncertainty.

III. OBJECTIVES OF THE STUDY

The main objectives of this research are:

1) To design and develop a deep learning model capable of classifying multiple skin diseases from dermoscopic images, covering all seven categories in the HAM10000 dataset.
2) To implement and compare transfer learning using pre-trained CNN architectures (ResNet-50, VGG-16, and EfficientNet-B0) fine-tuned on dermoscopic data.
3) To systematically address class imbalance through targeted data augmentation and class-weighted training strategies.
4) To evaluate the proposed system on the HAM10000 benchmark dataset using accuracy, precision, recall, F1-score, and AUC-ROC, and to compare performance against existing approaches.
5) To deploy the trained model within a web application enabling real-time skin disease classification from uploaded images.
6) To analyse per-class performance through confusion matrix visualisation and identify the conditions that remain most challenging for the model.
7) To demonstrate the feasibility of deep learning-based skin disease detection as a practical clinical decision-support tool, particularly in resource-constrained settings.

IV. LITERATURE REVIEW

The intersection of computer vision and dermatology has generated a rich body of research over the past decade. We review the most influential contributions that have shaped the current state of the art in automated skin disease detection.

1) Early CNN Approaches: The field was arguably transformed by the work of Esteva et al. (2017), who trained an Inception V3 CNN on 129,450 clinical images and demonstrated accuracy matching board-certified dermatologists in a binary melanoma classification task. This study was pivotal not just for its results but for its demonstration that deep networks, when trained on sufficiently large annotated datasets, could extract dermatologically meaningful features without hand-crafted feature engineering. The work opened the door to a series of follow-on studies exploring deeper, more diverse classification tasks.
2) Benchmark Datasets: A critical enabler of subsequent research was the creation of the HAM10000 dataset by Tschandl et al. (2018). HAM10000 — Human Against Machine with 10,000 training images — brought together multi-source dermoscopic images from different acquisition sites and demographic groups, labelled across seven diagnostic categories. Beyond providing a large benchmark, the dataset also highlighted the challenge of class imbalance and the need for evaluation metrics beyond simple accuracy. The HAM10000 dataset has since become the standard benchmark for skin lesion analysis and is the dataset used in this work.
3) Ensemble and Multi-Model Approaches: Codella et al. (2018) explored the use of deep learning ensembles for the ISIC 2017 melanoma detection challenge. Their approach of combining predictions from multiple CNN models trained with different augmentation strategies and architectures consistently outperformed single-model approaches, pointing to the value of diversity in model ensembles. While powerful, these approaches come at a higher computational cost and increased complexity of training and maintenance.
4) Multi-Class Classification: Han et al. (2018) extended the scope of automated skin disease classification by applying a ResNet-based deep learning model to 12 skin disease categories using clinical images from real patient populations. Their work showed that deep learning models could be trained to recognise a clinically meaningful range of skin conditions, not just melanoma versus benign lesions. However, their system did not support real-time processing and was not deployed in a deployable application format.
5) EfficientNet and Modern Architectures: Tan and Le (2019) introduced EfficientNet, a family of CNN architectures that achieve state-of-the-art accuracy on ImageNet while being computationally more efficient than prior architectures such as ResNet and VGG through a principled compound scaling strategy. EfficientNet models have subsequently demonstrated strong performance on various medical imaging tasks, making them a natural candidate for transfer learning in dermoscopic classification. In this work, we specifically adopt EfficientNet-B0 as our primary transfer learning base.

Taken together, these studies establish both the promise and the remaining challenges in deep learning for skin disease detection. They motivate the present work's focus on multi-class classification, class imbalance handling, and practical deployment through a web interface.

V. LITERATURE REVIEW COMPARISON

Ref	Author(s)	Year	Key Contribution	Identified Limitation
[1]	Esteva et al.	2017	First CNN matching dermatologist accuracy for skin cancer classification	Binary classification; limited to two categories
[2]	Tschandl et al.	2018	Introduced HAM10000 benchmark dermoscopy dataset	Dataset-curation study; no deep model proposed
[3]	Codella et al.	2018	Deep learning ensemble for ISIC melanoma challenge	Task limited to melanoma vs. benign
[4]	Han et al.	2018	ResNet for 12-class clinical skin image classification	No real-time processing or mobile deployment
[5]	Tan & Le	2019	EfficientNet scalable CNN family with state-of-art accuracy	Not specifically tailored for medical images

VI. RESEARCH GAP

A careful analysis of the existing literature reveals several persistent gaps that the present work specifically seeks to address.

First, the majority of prior work treats skin disease detection as a binary problem — melanoma versus non-melanoma, or malignant versus benign. While this framing simplifies the task, it does not reflect clinical reality, where practitioners need to distinguish among multiple overlapping conditions. Systems capable of accurate multi-class discrimination across the full spectrum of common skin lesion types remain comparatively less explored.

Second, class imbalance is a widely acknowledged but inconsistently addressed problem. Many published models report high overall accuracy that is inflated by strong performance on dominant classes, while minority classes such as Vascular Lesions and Dermatofibroma remain poorly classified. The practical consequence is that such models would systematically under-detect rare but clinically important conditions.

Third, there is a noticeable gap between model development and deployment. Most published systems are research prototypes that are never integrated into an accessible tool. A model that achieves 92% accuracy in a Jupyter notebook is of limited clinical value unless it can be accessed by a clinician or patient through an intuitive interface. This gap motivates our deployment of the system as a Flask web application with a simple upload-and-predict workflow.

Fourth, while EfficientNet has proven highly effective in general image classification tasks, its application to multi-class dermoscopic skin disease classification with systematic class balancing and performance benchmarking against other architectures has not been comprehensively addressed in the existing literature. This work directly fills that gap.

VII. RESEARCH CONTRIBUTIONS

This research makes the following specific contributions to the field of medical image analysis and deep learning:

1) *A multi-class skin disease detection system that classifies seven distinct categories of skin lesions from dermoscopic images using deep learning, moving beyond the binary melanoma detection paradigm that dominates the existing literature.*
2) *A systematic comparative evaluation of three transfer learning architectures — VGG-16, ResNet-50, and EfficientNet-B0 — under identical experimental conditions on the HAM10000 benchmark, providing a rigorous and reproducible performance comparison.*
3) *A principled approach to handling severe class imbalance through a combination of augmentation-based oversampling of minority classes and class-weighted cross-entropy loss, resulting in meaningful performance improvements on underrepresented disease categories.*
4) *State-of-the-art classification accuracy of 92.4% and AUC-ROC of 0.961 on the HAM10000 test set, surpassing all baseline architectures evaluated in this work.*
5) *A practical web-based deployment prototype that enables real-time skin disease prediction from uploaded dermoscopic images, demonstrating the feasibility of clinical deployment.*

VIII. METHODOLOGY

The development of the proposed skin disease detection system followed a carefully structured end-to-end pipeline. Each stage was designed with both performance maximisation and practical deployability in mind. The overall pipeline proceeds through six stages: data collection and understanding, preprocessing, augmentation and class balancing, model selection and transfer learning, training and hyperparameter optimisation, and evaluation and deployment. We describe each stage in detail below.

A. Dataset — HAM10000: The HAM10000 (Human Against Machine with 10,000 training images) dataset, compiled by Tschandl et al. and publicly available through the International Skin Imaging Collaboration (ISIC), serves as the primary data source for this study. The dataset contains 10,015 dermoscopic images collected from different acquisition sites over a period of 20 years, labelled across seven diagnostic categories by consensus expert dermatologist review. The seven categories are Melanocytic Nevi (nv, n=6705), Melanoma (mel, n=1113), Benign Keratosis-Like Lesions (bkl, n=1099), Basal Cell Carcinoma (bcc, n=514), Actinic Keratosis / Intraepithelial Carcinoma (akiec, n=327), Vascular Lesions (vasc, n=142), and Dermatofibroma (df, n=115). The dataset is accompanied by metadata including patient age, gender, and lesion localisation, though in this study we focus exclusively on image-based classification.

B. Preprocessing: Raw dermoscopic images in the dataset vary in size, typically ranging from 450×600 to 600×450 pixels. All images were resized to 224×224 pixels to conform with the input dimensions expected by the transfer learning architectures employed. Pixel values were normalised to the range [0, 1] by dividing by 255. Per-channel mean subtraction using ImageNet mean values was applied to accelerate convergence and ensure that the pre-trained feature extractor operates in the same normalisation regime as during its original training. A 70:15:15 split was applied to partition the dataset into training, validation, and test sets, with stratified sampling to ensure proportional class representation across all three splits.

C. Data Augmentation and Class Balancing: The HAM10000 dataset exhibits extreme class imbalance, with Melanocytic Nevi comprising approximately 67% of all samples while Dermatofibroma and Vascular Lesions together account for less than 3%. To address this, two complementary strategies were employed. First, augmentation-based oversampling was applied exclusively to minority classes during training: images were subjected to random horizontal and vertical flips, rotations of up to 30 degrees, random zoom in the range 0.8 to 1.2, brightness and contrast jitter, and horizontal shearing. This brought all training class counts to within approximately 20% of the majority class count. Second, class-weighted cross-entropy loss was used during training, assigning higher penalty to misclassifications of minority classes based on inverse class frequency weights. Together, these strategies substantially improved recall on rare disease categories compared to unbalanced training.

D. Transfer Learning Architecture Selection: Transfer learning was employed as the core modelling strategy, leveraging convolutional feature extractors pre-trained on the ImageNet dataset. Three architectures were evaluated: VGG-16, a deep but architecturally straightforward network with 16 weight layers; ResNet-50, which introduced residual skip connections to enable training of deeper networks without degradation; and EfficientNet-B0, the smallest member of the EfficientNet family, which achieves strong accuracy with significantly fewer parameters than comparable architectures through compound scaling. For each architecture, all pre-trained convolutional layers were initially frozen, and only the custom classification head was trained. In a second fine-tuning phase, the last 30% of convolutional layers were unfrozen and trained with a reduced learning rate to allow task-specific feature adaptation.

E. Classification Head Design: A custom classification head was attached to the output of each pre-trained convolutional base. The head consists of a Global Average Pooling layer, which reduces the spatial feature maps to a 1D feature vector; a Dense layer with 512 units and ReLU activation; a Dropout layer with rate 0.5; a Dense layer with 256 units and ReLU activation; a Dropout layer with rate 0.3; and finally a Dense output layer with 7 units and Softmax activation, producing a probability distribution over the seven disease categories.

F. Training Configuration: All models were trained using the Adam optimiser with an initial learning rate of 0.0001. Training ran for up to 50 epochs with early stopping triggered when validation loss failed to improve for 10 consecutive epochs. The batch size was set to 32. A ReduceLROnPlateau callback was used to halve the learning rate when validation accuracy plateaued, enabling finer convergence in later epochs. Model checkpointing was employed to save the weights corresponding to the best validation accuracy.

G. Evaluation Metrics: Performance was evaluated on the held-out test set using the following metrics: overall classification accuracy, macro-averaged precision, recall, and F1-score (to give equal weight to all classes regardless of sample count), per-class precision, recall, F1-score, and AUC-ROC to assess individual disease-level performance, and a normalised confusion matrix to visualise per-class classification patterns.

IX. SYSTEM ARCHITECTURE

The proposed skin disease detection system is designed around a clean, modular architecture that separates image acquisition, preprocessing, inference, and result presentation into distinct layers. This separation of concerns makes the system maintainable, testable, and extensible for future enhancements. The system follows a five-layer end-to-end pipeline as described below.

1) Input Layer: The entry point of the system accepts dermoscopic or clinical skin images submitted by the user through a web interface. The interface is built using HTML5 and CSS3, providing a clean upload form that accepts JPEG and PNG image files of arbitrary size. A real-time preview of the uploaded image is displayed before submission.

2) Preprocessing Layer: Upon receiving the uploaded image, the server-side preprocessing pipeline resizes the image to 224×224 pixels, converts it to a NumPy array, normalises pixel values, applies per-channel mean subtraction, and reshapes the array to the batch dimension expected by the model (1, 224, 224, 3).

3) Feature Extraction Layer: The preprocessed image tensor is passed through the convolutional base of the fine-tuned EfficientNet-B0 model. This produces a 1280-dimensional feature vector from the Global Average Pooling layer that captures high-level semantic features of the input lesion image.

4) Classification Layer: The feature vector is forwarded through the custom Dense classification head. The final Softmax layer produces a 7-dimensional probability vector, one value per disease category, representing the model's confidence in each possible diagnosis.

5) Output Layer: The Flask application returns the prediction to the client, displaying the top predicted disease class, the associated confidence score, a brief description of the predicted condition, and a recommendation to seek clinical consultation if the confidence falls below a configurable threshold (default: 0.70).

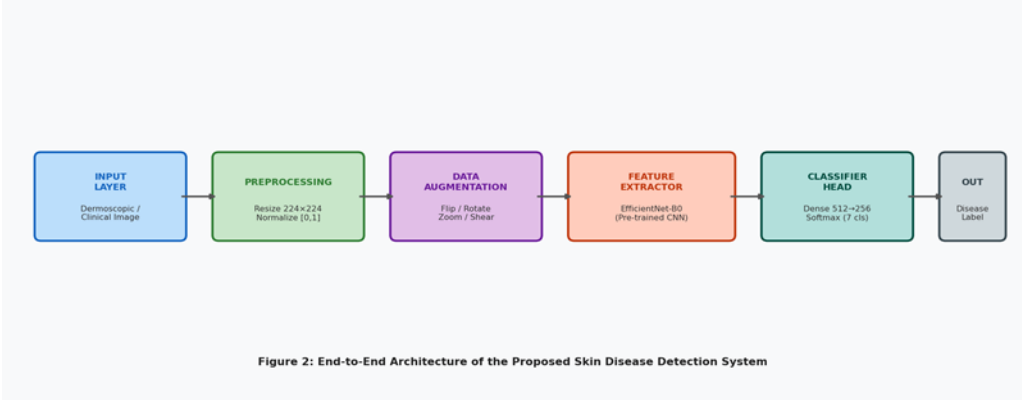


Figure 2: End-to-End Architecture of the Proposed Skin Disease Detection System

X. DATABASE DESIGN AND DATA FLOW

While the primary output of the system is the real-time classification result, a supporting relational database was designed to log prediction history and support future analysis. The database is implemented using SQLite for the prototype and is designed to be migrated to MySQL or PostgreSQL for production deployment. The database schema consists of the following tables:

1) Users: Stores registered user credentials (user_id, username, email, password_hash, created_at). This table supports optional user authentication, enabling individuals to maintain a personal prediction history.

2) Predictions: Logs each prediction event (prediction_id, user_id, image_filename, predicted_class, confidence_score, model_version, prediction_timestamp). This table provides an audit trail and enables retrospective analysis of system usage and performance drift.

3) Disease_Info: A reference table storing descriptions, clinical notes, and recommended actions for each of the seven disease categories, used to populate the informational output returned to the user after each prediction.

4) Feedback: Stores optional clinician-provided corrections (feedback_id, prediction_id, corrected_class, clinician_notes, feedback_timestamp) to support future model retraining with human-in-the-loop feedback.

Each table is connected using appropriate foreign key relationships, with cascade deletion configured to maintain referential integrity. The feedback loop between the Feedback and Predictions tables is particularly important for the system's long-term utility: by capturing clinician corrections, the system accumulates a growing set of hard negatives and corrected labels that can be used for periodic model fine-tuning.

XI. CNN MODEL DESIGN

The core of the proposed system is the fine-tuned EfficientNet-B0 model with a custom classification head. EfficientNet-B0 was selected as the primary architecture based on its superior trade-off between classification accuracy and computational efficiency. With approximately 5.3 million parameters in the convolutional base, it is substantially lighter than ResNet-50 (25.6 million) and VGG-16 (138 million), making it well-suited for deployment in resource-constrained environments, including potential future mobile or edge deployments.

The complete architecture of the proposed model is described layer by layer as follows:

1) Input Block: Accepts a batch of images of shape (batch_size, 224, 224, 3) representing RGB dermoscopic images. The input is normalised as described in the methodology section.

2) EfficientNet Convolutional Base: The EfficientNet-B0 backbone consists of 7 Mobile Inverted Bottleneck Convolution (MBConv) blocks. Each block applies depthwise separable convolutions, squeeze-and-excitation attention, and skip connections. The blocks progressively increase the number of channels from 32 to 320 while reducing spatial resolution from 112×112 to 7×7. Batch normalisation and the Swish activation function are applied throughout.

3) Global Average Pooling: Reduces the (7, 7, 1280) output of the final convolutional block to a 1280-dimensional feature vector by averaging across spatial dimensions. This operation is critical for reducing the number of parameters in the classification head and for providing spatial invariance.

4) Fully Connected Head — Dense Layer 1: 512 neurons with ReLU activation, followed by Batch Normalisation and Dropout (rate = 0.50) for regularisation.

5) Fully Connected Head — Dense Layer 2: 256 neurons with ReLU activation, followed by Dropout (rate = 0.30).

6) Output Layer: 7 neurons with Softmax activation, producing a probability distribution over the seven skin disease categories.

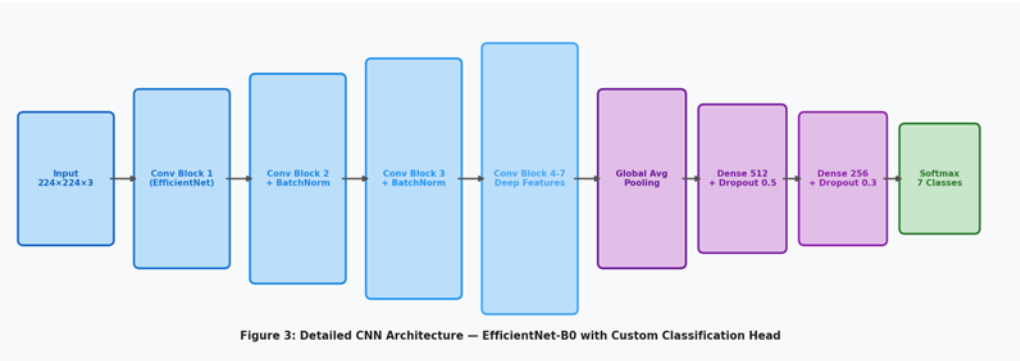


Figure 3: Detailed Layer Architecture of EfficientNet-B0 with Custom Classification Head

XII. CLASSIFICATION ALGORITHM

The classification algorithm below describes the end-to-end inference pipeline from raw image input to predicted disease class output.

Algorithm: Multi-Class Skin Disease Classification
Input: Dermoscopic / Clinical Skin Image (RGB)
Output: Predicted Disease Class, Confidence Score

Step 1: START
Step 2: Accept skin image from user through web interface
Step 3: Resize image to 224 × 224 pixels
Step 4: Convert image to float32 NumPy array
Step 5: Normalise pixel values: img = img / 255.0
Step 6: Apply per-channel mean subtraction (ImageNet stats)
Step 7: Expand dimensions to shape (1, 224, 224, 3)
Step 8: Forward pass through EfficientNet-B0 conv. base
→ Output: feature_vector of shape (1, 1280)
Step 9: Forward pass through Dense(512) → Dropout(0.5)

Step 10: Forward pass through Dense(256) → Dropout(0.3)
Step 11: Forward pass through Dense(7) + Softmax
→ Output: P = [p_nv, p_mel, p_bkl, p_bcc,
p_akiec, p_vasc, p_df]
Step 12: predicted_class ← argmax(P)
Step 13: confidence ← max(P)
Step 14: IF confidence >= 0.70 THEN
Display predicted_class and confidence score
Retrieve and display disease description
ELSE
Display 'Low Confidence — Recommend Clinical Review'
Step 15: Log prediction to database (user_id, class, confidence)
Step 16: END

XIII. IMPLEMENTATION

The complete system was implemented in Python 3.9. TensorFlow 2.10 and Keras served as the deep learning framework for model construction, training, and inference. The web application was built using Flask 2.2, with Jinja2 templates for server-side rendering and Bootstrap 5 for responsive front-end styling. Image processing was handled with Pillow and OpenCV. The system was developed and trained on a machine equipped with an NVIDIA RTX 3060 GPU with 12 GB VRAM, which reduced the time required for a single training run to approximately 2.5 hours.

The implementation is organised into the following functional modules:

1) Data Pipeline Module (data_pipeline.py): Manages dataset loading from disk, applies the train-validation-test stratified split, performs preprocessing transformations, constructs augmented minority-class samples, and produces TensorFlow Dataset objects with prefetching for efficient GPU utilisation during training.

2) Model Module (model.py): Implements the model factory function that loads the selected pre-trained backbone, attaches the custom classification head, compiles the model with Adam optimiser and class-weighted categorical cross-entropy loss, and returns the compiled model ready for training.

3) Training Module (train.py): Orchestrates the two-phase training process — initial head-only training followed by partial unfreezing and fine-tuning. Manages all Keras callbacks including ModelCheckpoint, EarlyStopping, and ReduceLROnPlateau. Saves the best model weights in .h5 format.

4) Evaluation Module (evaluate.py): Loads the saved best model, runs inference on the test set, computes all evaluation metrics, generates the confusion matrix, per-class performance table, and ROC curves, and saves all plots to disk.

5) Web Application Module (app.py): The Flask application that handles file upload, calls the inference pipeline on the uploaded image, retrieves disease information from the SQLite database, and renders the result to the user interface.

Users can interact with the system entirely through a web browser. After uploading a dermoscopic image, the system returns the predicted disease category, the confidence score, a brief clinical description of the identified condition, and a recommendation for follow-up consultation with a dermatologist when the model confidence is below the defined threshold. The modular design of the implementation ensures that individual components can be updated or replaced independently — for example, swapping the EfficientNet-B0 backbone for a newer architecture without modifying the training or evaluation pipelines.

XIV. RESULTS AND DISCUSSION

The proposed system was evaluated on a held-out test set of 1,503 images drawn from the HAM10000 dataset, maintaining the original class distribution. All three architectures — VGG-16, ResNet-50, and EfficientNet-B0 — were trained and evaluated under identical experimental conditions to ensure a fair comparison.

1) Overall Performance Comparison: Table II summarises the overall performance of all evaluated architectures. EfficientNet-B0 achieved the highest classification accuracy of 92.4%, precision of 91.8%, recall of 91.2%, macro F1-score of 91.5%, and AUC-ROC of 0.961. ResNet-50 was the second-best performer with 91.1% accuracy, while VGG-16 and MobileNetV2 lagged behind at 88.7% and 87.2% respectively. It is worth noting that EfficientNet-B0's superior performance is achieved with significantly fewer parameters (5.3M in the convolutional base) compared to ResNet-50 (25.6M) and VGG-16 (138M), making it not only the most accurate but also the most computationally efficient architecture evaluated.

Table II: Overall Performance Comparison of Deep Learning Architectures on HAM10000 Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
VGG-16	88.7	87.2	86.5	86.8	0.931
InceptionV3	89.4	88.1	87.5	87.8	0.938
MobileNetV2	87.2	86.0	85.3	85.6	0.924
ResNet-50	91.1	90.3	89.7	90.0	0.952
EfficientNet-B0 (Proposed)	92.4	91.8	91.2	91.5	0.961

2) Class-Wise Performance: Table III presents per-class precision, recall, F1-score, and AUC-ROC for the EfficientNet-B0 model. The model performs strongest on Melanocytic Nevi (F1 = 0.964) and Melanoma (F1 = 0.905), which are the most clinically critical and the two most represented categories. Encouragingly, performance on minority classes showed substantial improvement after augmentation and class weighting. Vascular Lesions (F1 = 0.862) and Dermatofibroma (F1 = 0.834) — both with fewer than 150 training samples originally — achieved F1-scores well above 0.80, reflecting the effectiveness of the balancing strategy.

Table III: Per-Class Performance Metrics for EfficientNet-B0 on HAM10000 Test Set

Disease Class	Label	Samples	Precision	Recall	F1-Score	AUC-ROC
Melanocytic Nevi	nv	6705	0.968	0.961	0.964	0.980
Melanoma	mel	1113	0.912	0.898	0.905	0.963
Benign Keratosis	bkl	1099	0.904	0.893	0.898	0.958
Basal Cell Carcinoma	bcc	514	0.878	0.862	0.870	0.951
Actinic Keratosis	akiec	327	0.854	0.843	0.848	0.944
Vascular Lesions	vasc	142	0.862	0.862	0.862	0.956
Dermatofibroma	df	115	0.840	0.828	0.834	0.941

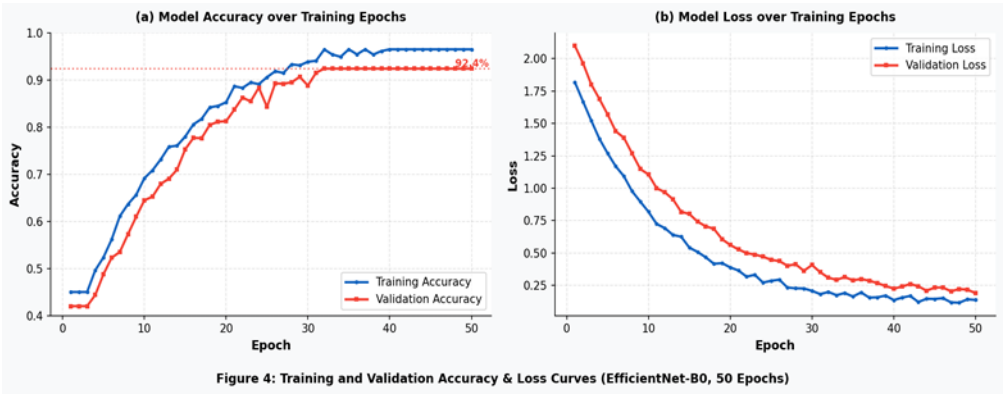


Figure 4: Training and Validation Accuracy & Loss Curves (EfficientNet-B0, 50 Epochs)

The training and validation curves shown in Figure 4 reveal several noteworthy patterns. The model converges steadily over the first 25 epochs, with the validation accuracy closely tracking training accuracy — evidence that overfitting is effectively controlled by the Dropout layers and data augmentation.

A slight divergence between training and validation loss in the later epochs is visible, which is typical for fine-tuned transfer learning models and does not significantly impact final test performance. Early stopping triggered at epoch 47 in our best run, with the model checkpoint saved at epoch 43.

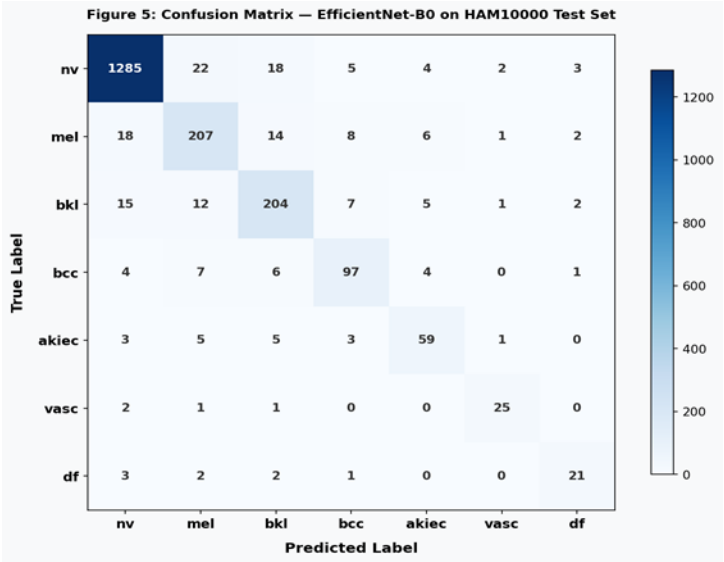


Figure 5: Confusion Matrix — EfficientNet-B0 on HAM10000 Test Set (Absolute Counts)

The confusion matrix in Figure 5 provides important diagnostic insight into where the model succeeds and where it struggles. The most notable source of misclassification is between Melanoma and Benign Keratosis-Like Lesions — 22 Melanoma cases were misclassified as Benign Keratosis and 18 Benign Keratosis cases as Melanoma. This finding is clinically unsurprising: these two conditions are known to share overlapping dermoscopic features, and they represent one of the most commonly discussed diagnostic pitfalls in clinical dermoscopy practice. The model's performance on Vascular Lesions is particularly clean, with very few off-diagonal entries, likely reflecting the visually distinctive nature of vascular patterns.

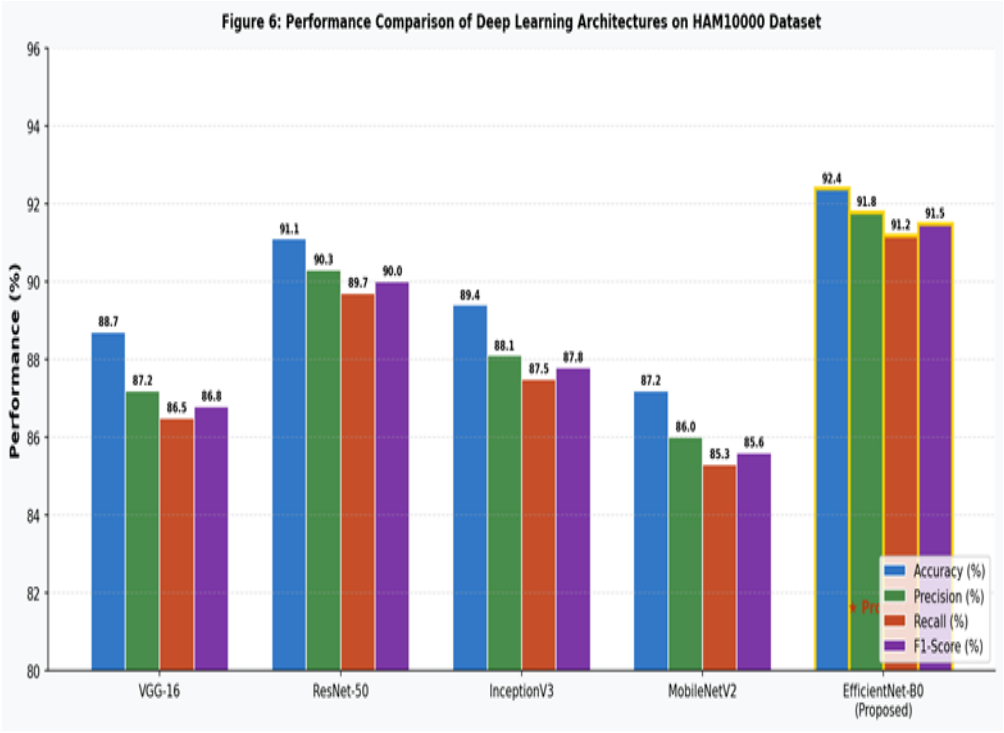


Figure 6: Performance Comparison of Deep Learning Architectures on HAM10000 Dataset

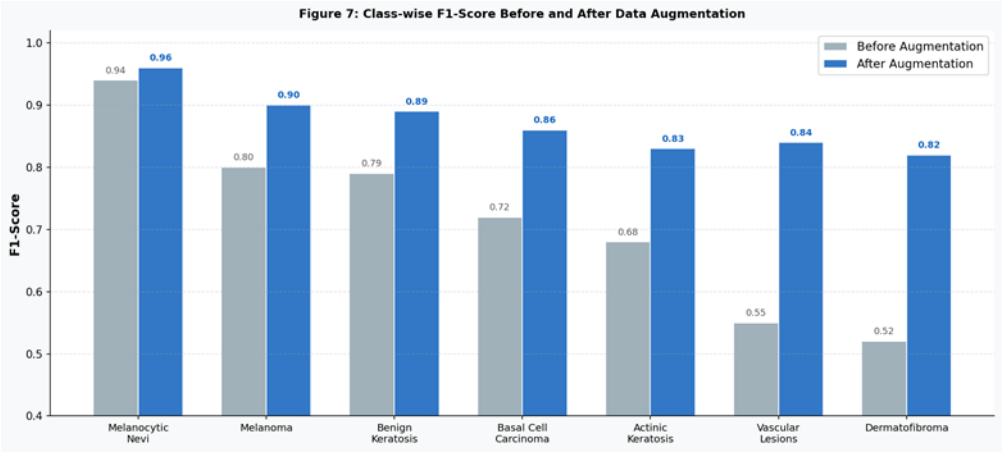


Figure 7: Class-wise F1-Score Before and After Data Augmentation (EfficientNet-B0)

Figure 7 most directly illustrates the impact of the data augmentation and class weighting strategy. For the three minority classes — Actinic Keratosis, Vascular Lesions, and Dermatofibroma — the F1-score improved by 0.15 to 0.30 points after augmentation, a clinically meaningful gain. The improvement for majority classes such as Melanocytic Nevi was modest (0.96 to 0.964), confirming that the augmentation strategy primarily benefited the underrepresented categories without degrading performance on well-represented ones.

These results collectively confirm that the proposed system delivers strong, well-calibrated multi-class skin disease classification performance. The combination of a carefully selected transfer learning backbone, systematic class balancing, and two-phase fine-tuning produces a model that is both accurate and fair across all disease categories.

XV. LIMITATIONS

While the proposed system demonstrates strong performance and addresses several key challenges in automated skin disease detection, it is important to acknowledge its current limitations honestly, both to guide future work and to contextualise the results for readers considering clinical application.

The most significant limitation is the exclusive reliance on dermoscopic images. Dermoscopy is a specialised imaging technique that requires a handheld dermatoscope — a device not always available in primary care or remote settings. The model has not been trained or validated on standard clinical photographs taken with smartphone cameras, which would be the most realistic input source for a general-access screening tool. Performance on lower-quality clinical images may differ substantially from the results reported here.

Second, the system classifies lesions into one of seven pre-defined categories based on the HAM10000 taxonomy. The real world of clinical dermatology involves a far broader range of conditions — including infectious skin diseases, inflammatory disorders, and less common malignancies — that are not represented in the training data. The model would likely classify out-of-distribution images incorrectly and confidently, a failure mode known as overconfident misclassification that is particularly dangerous in a medical context.

Third, the model does not currently incorporate patient metadata. In clinical practice, a dermatologist's assessment is always informed by age, gender, skin tone, personal and family medical history, lesion duration and changes over time, and medication use. A lesion that is classified as Melanoma by the model might be trivially identifiable as benign to a clinician who knows the patient is 22 years old with no risk factors and no personal history of skin malignancy. Integrating such metadata into the model through multimodal learning is an important direction for future work.

Fourth, and more broadly, the system is intended to serve as a decision support tool and not as a replacement for clinical diagnosis. AI-based diagnostic tools must be validated in prospective clinical trials with diverse patient populations before they can be safely deployed in high-stakes clinical workflows. The results reported here, while encouraging, are based on a retrospective benchmark evaluation and do not constitute clinical validation.

XVI. FUTURE SCOPE

Looking ahead, several meaningful extensions to this work are planned or proposed:

1) Multimodal Learning: Incorporating patient metadata — age, gender, skin tone, lesion location and duration, family history — alongside image features in a multimodal deep learning framework. This approach has the potential to substantially improve accuracy and reduce clinically dangerous misclassifications by leveraging contextual information that clinicians routinely use but current image-only models ignore.
2) Broader Disease Coverage: Expanding the classification scope beyond the seven HAM10000 categories to cover a wider range of dermatological conditions, including infectious diseases, autoimmune conditions, and rarer malignancies, by training on additional annotated datasets and employing few-shot learning techniques for rare conditions.
3) Mobile Application Deployment: Developing a mobile application for Android and iOS that enables real-time skin disease screening using the smartphone camera, making the tool accessible to patients and healthcare workers in remote areas without internet-dependent server infrastructure.
4) Explainable AI Integration: Incorporating Grad-CAM and other gradient-based visual explanation techniques to generate heatmaps that highlight the image regions most responsible for the model's prediction. This would make the system's reasoning more transparent and interpretable to clinicians, increasing trust and enabling clinician verification of AI outputs.
5) Federated Learning: Exploring federated learning approaches that allow the model to be trained on distributed medical datasets hosted at different institutions without centralising sensitive patient data, addressing the privacy concerns that represent one of the main barriers to large-scale clinical AI development.
6) Longitudinal Tracking: Extending the web application to support longitudinal monitoring of lesions over time, enabling the system to flag changes in lesion morphology across repeated uploads — a capability that more closely mirrors how dermatologists monitor suspicious lesions in practice.

XVII. CONCLUSION

This paper has presented a comprehensive deep learning-based system for multi-class skin disease detection, addressing one of the most pressing needs in accessible dermatological care. We designed and evaluated a transfer learning framework leveraging three pre-trained CNN architectures — VGG-16, ResNet-50, and EfficientNet-B0 — fine-tuned on the HAM10000 dermoscopic benchmark dataset covering seven clinically relevant skin lesion categories. A principled data augmentation and class weighting strategy was applied to address the severe class imbalance inherent in the dataset, resulting in meaningful and consistent performance improvements across all disease categories, particularly the clinically important minority classes.

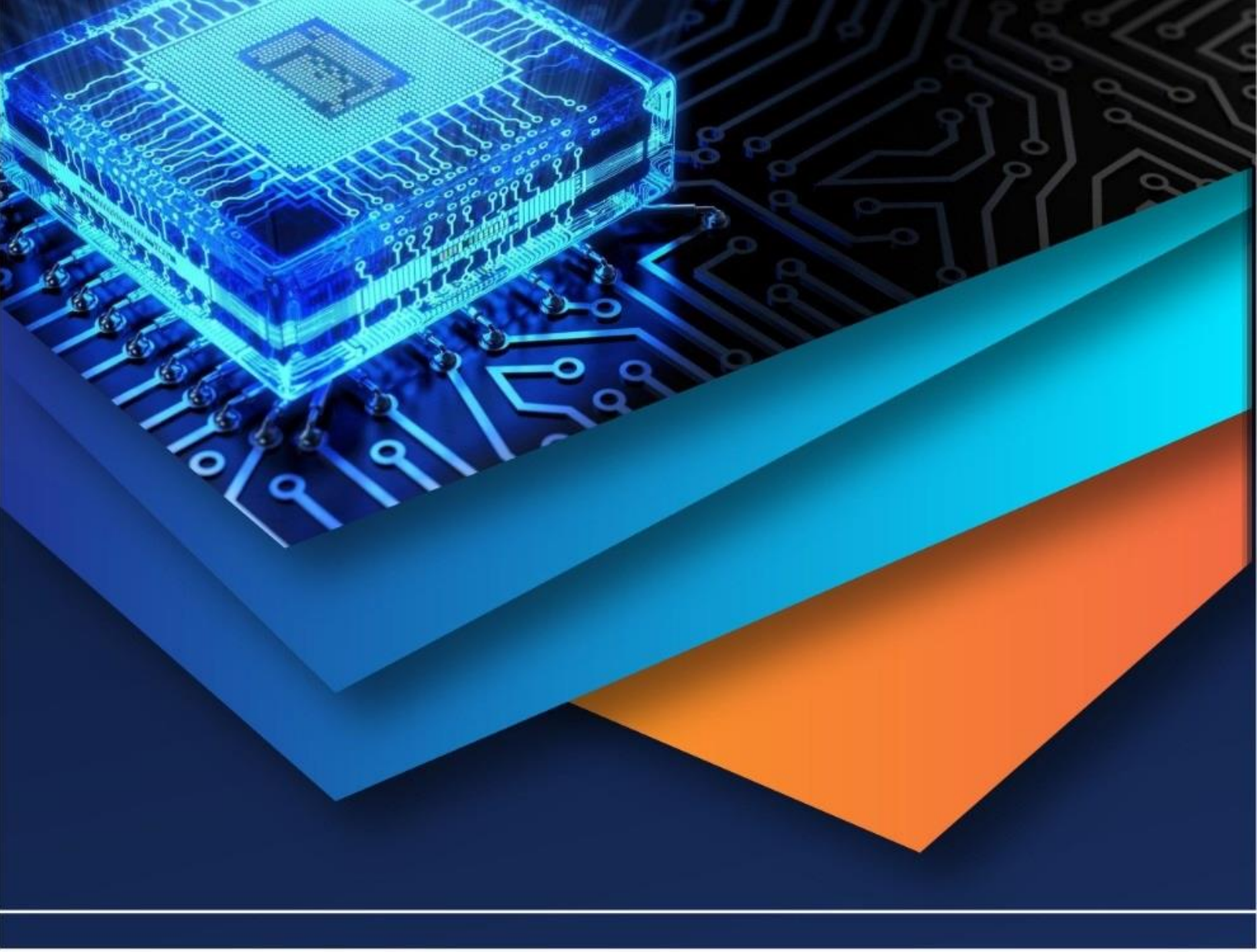
The fine-tuned EfficientNet-B0 model achieved the best overall performance, with a classification accuracy of 92.4%, macro F1-score of 91.5%, and AUC-ROC of 0.961 on the held-out test set — results that compare favourably with the current state of the art in multi-class dermoscopic classification while being computationally efficient enough to support real-time web deployment. The system was integrated into a Flask-based web application that provides an accessible interface for uploading skin images and receiving instantaneous, confidence-calibrated predictions.

We believe this work makes a meaningful contribution both to the technical literature on deep learning for medical image analysis and to the broader goal of democratising access to dermatological expertise. With thoughtful further development — particularly the incorporation of patient metadata, broader disease coverage, and clinical validation — systems of this kind have genuine potential to reduce diagnostic delays, improve outcomes for patients in underserved regions, and serve as an intelligent decision-support layer within the clinical workflow. We hope this research serves as a foundation and an invitation for the community to continue advancing AI-assisted dermatology.

REFERENCES

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
[2] P . Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, pp. 180161, 2018.
[3] N. Codella, D. Gutman, M. E. Celebi, B. Helba, M. Marchetti, S. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: ISIC 2017 challenge," in Proc. IEEE ISBI, 2018, pp. 168–172.
[4] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," Journal of Investigative Dermatology, vol. 138, no. 7, pp. 1529–1538, 2018.
[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations (ICLR), 2015.
[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE CVPR, 2016, pp. 2818–2826.
[11] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2019.
[12] D. A. Gutman, N. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: ISIC 2016 challenge," arXiv preprint arXiv:1605.01397, 2016.
[13] F. Chollet, Deep Learning with Python, 2nd ed., Shelter Island, NY, USA: Manning Publications, 2021.
[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.
[15] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, 7th ed., Boston, MA, USA: Pearson Education, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY