International Journal of Artificial Intelligence and Digital Innovation
https://journals.hu.edu.so/index.php/ijaidi

ResNet50-Based Multi-Class Skin Lesion Classification with Image Suitability Screening in a Web-Based Decision-Support System

Abdirahim Omar Hassan[1], Nuur Abdulahi Gacal[1], Asli Ali Osman[1], Sumaya Bashir Osman[1], Alhaji-Mohamud Abdullahi Sheikh[2]*

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Hormuud University, Mogadishu, Somalia
[2]Department of Information Technology, Taaj, Mogadishu, Somalia

*Corresponding author's email: Mohamud.abdullahi@paysii.com

Abstract
Limited access to specialist dermatology and the variability of image-based assessment motivate the use of artificial intelligence as a preliminary decision-support tool. This study evaluated five convolutional neural network architectures—MobileNetV2, EfficientNetB0, ResNet50, DenseNet121 and InceptionV3—for seven-class skin-lesion classification using the HAM10000 dataset, and integrated the selected classifier into a web-based prototype with image-suitability screening. The evaluation used test accuracy, precision, recall, F1-score and training time. ResNet50 produced the highest initial test accuracy (72.06%) and F1-score (74.00%), although MobileNetV2 and EfficientNetB0 reported slightly higher precision. ResNet50 was therefore selected and retrained for a further 30 epochs. Its test accuracy increased to 80.24% and its F1-score to 80.85%, corresponding to gains of 8.18 and 6.85 percentage points, respectively. The implemented inference workflow converts uploaded images to RGB, resizes them to 224 × 224 pixels, normalises them, screens their suitability through a Gatekeeper stage, and then applies the ResNet50 classifier when an image is accepted. The web application further supports secure login, email one-time-password verification, role-based access control, patient records, diagnostic history and report generation. The study demonstrates comparative model selection and end-to-end system integration, but it does not establish clinical effectiveness. No external dataset, locally collected Somali clinical images, dermatologist comparison or prospective patient evaluation was used. The system should therefore be interpreted as a research-stage decision-support prototype requiring external and clinical validation before use in patient care.

Keywords: skin lesion classification, ResNet50, deep learning, transfer learning, HAM10000, clinical decision support

1. Introduction

Skin disease contributes substantially to global morbidity, while access to specialist assessment remains uneven. A recent Global Burden of Disease analysis confirms the continuing scale of skin and subcutaneous disease worldwide and highlights the importance of improving access to care in settings with limited resources [1]. The workforce gap is particularly visible in dermatology: sub-Saharan Africa has been reported to have approximately 0–3 dermatologists per million population, compared with about 34 per million in the United States [2]. The present project was motivated by this access problem in Mogadishu, Somalia, using DR AK Derma

Clinic as a case context in which routine assessment depends primarily on visual inspection and specialist resources are limited.

Deep learning offers a practical route to image-based decision support because convolutional neural networks (CNNs) can learn visual representations directly from labelled images. In dermatology, CNNs have been used for lesion classification, triage and related image-analysis tasks, and systematic reviews continue to report strong experimental performance across a wide range of architectures [3]–[5]. Transfer learning is especially relevant because medical image datasets are much smaller than large natural-image collections; pretrained feature extractors can be adapted to lesion classification rather than trained entirely from scratch [5], [6].

Public datasets made comparative work possible. HAM10000 contains 10,015 dermatoscopic images across seven commonly used diagnostic categories and was created to support machine-learning research on pigmented skin lesions [7]. The ISIC challenges subsequently provided shared evaluation settings and demonstrated an important methodological point: algorithms that perform similarly on one benchmark can differ in their ability to generalise to data from another source [8]. Reported accuracy therefore describes performance under a particular dataset, split, preprocessing pipeline and evaluation procedure rather than a universal diagnostic capability.

This distinction is important because the dermatology AI literature still has significant evidence gaps. A systematic review of deep-learning image analyses found that external validation was uncommon and that reporting of participant ethnicity or Fitzpatrick skin type was limited [3]. Reviews of public skin-image datasets likewise show substantial under-representation of darker skin types and incomplete demographic metadata [9], [10]. Direct evaluation on the Diverse Dermatology Images dataset has shown that state-of-the-art dermatology models can lose substantial performance on diverse skin tones and uncommon diseases [11]. These findings are directly relevant to a system intended for Somalia: performance measured on HAM10000 cannot be assumed to transfer to the target population without external and local validation.

A second deployment problem is the uploaded image itself. A fixed seven-class classifier will still return one of its known labels when given an unrelated, blurred, dark, poorly framed or otherwise unsuitable photograph. Teledermatology studies report that patient-submitted images frequently have quality or usefulness problems [12], [13], and classifier performance can degrade when images come from uncontrolled patient-driven sources rather than protocol-based acquisition [14]. Automated dermatological image-quality assessment has therefore emerged as a separate research problem [15], and clinical work has shown that real-time image-quality feedback can reduce poor-quality submissions [16].

A third gap concerns the workflow around a prediction. Browser and mobile systems have already demonstrated that skin-image classifiers can be delivered through accessible interfaces [17], [18]. However, a clinically interpretable workflow requires more than a single predicted label: authentication, patient registration, diagnostic history, access control and report generation determine whether a result can be traced, reviewed and managed over time. Regulatory experience with AI-enabled dermatology devices also shows that a functioning application and a clinically validated medical system are different evidentiary categories [19]. This study addresses the architecture-selection and system-integration aspects of these gaps. Five CNN architectures—MobileNetV2, EfficientNetB0, ResNet50, DenseNet121 and InceptionV3—were evaluated on the seven-class project task using the same reported metric set. ResNet50 achieved the highest initial accuracy and F1-score and was selected for further development. It was then retrained for a further 30 epochs, after which test accuracy increased from 72.06% to 80.24% and F1-score from 74.00% to 80.85%. The selected classifier was

integrated behind an image-suitability Gatekeeper and embedded in a web-based decision-support prototype with secure access, patient records, diagnostic history and report generation. The contribution is therefore fourfold. Experimentally, the project provides a comparative evaluation of five established CNN architectures on the same seven-class task. At model level, it documents the selection and 30-epoch retraining of ResNet50 and the resulting change in performance. At system level, it integrates classification with image-suitability screening, authentication, role-based access and clinical-record functions. Practically, it demonstrates how an image classifier can be placed within a traceable decision-support workflow rather than exposed as an isolated prediction service. The study does not claim a new CNN architecture, autonomous diagnosis, dermatologist-level performance, external generalisation or clinical effectiveness.

The remainder of the paper is organised as follows. Section 2 reviews deep learning for skin-lesion classification, CNN architectures, public datasets, image-suitability assessment and web-based decision support, and then states the research gap. Section 3 describes the dataset, preprocessing, model comparison, ResNet50 retraining, evaluation metrics, Gatekeeper workflow and system architecture. Section 4 presents and discusses the experimental and system results, including limitations and recommendations. Section 5 concludes the study.

2. Literature Review

2.1 Deep Learning for Skin-Lesion Classification

Deep learning became central to automated dermatological image analysis after CNNs demonstrated that visual features could be learned directly from large labelled image collections. Esteva et al. showed the potential of deep neural networks for skin-cancer classification on curated images [20], while the ISIC challenge series created a common benchmark environment for lesion segmentation and classification [8]. Subsequent reviews have confirmed rapid growth in CNN-based dermatology research, while also noting that experimental success is often difficult to translate into real-world evidence [3], [4].

Recent work illustrates both the strength and the variability of reported results. Gautam et al. reported high accuracy using a deep neural network pipeline for early skin-disease identification [21]. Wang et al. proposed multiscale feature fusion using DenseNet-121 and an improved VGG-16, reporting 91.24% test accuracy on ISIC2018 [22]. Wu et al. compared several CNN architectures on a clinical facial-skin dataset and found that performance differed materially between backbones [6]. Riaz et al. similarly reviewed transfer-learning and federated-learning studies and found a broad reported accuracy range that depended strongly on dataset and methodology [5]. These studies support the use of deep learning, but they do not establish a single architecture as universally optimal.

2.2 CNN Architectures and Transfer Learning

The five architectures evaluated in this project represent different design strategies. ResNet uses residual connections to make deeper networks easier to optimise [23]. MobileNetV2 uses inverted residual blocks and linear bottlenecks to reduce computation for resource-constrained inference [24]. EfficientNet scales network depth, width and input resolution jointly to improve the accuracy-efficiency trade-off [25]. DenseNet encourages feature reuse through dense layer connectivity [26], while InceptionV3 combines multiple receptive-field sizes through factorised convolutional modules [27]. Because these networks differ in depth, connectivity, parameter efficiency and representation, their ranking can change with the dataset and training procedure.

Transfer learning is the practical mechanism that connects these architectures to medical imaging. Reviews of skin-cancer classification show frequent use of ResNet, DenseNet, MobileNet and Inception families as pretrained backbones [5]. Wu et al. also demonstrated the usefulness of transfer learning when adapting CNNs to clinical facial-skin

images [6]. In the present project, transfer learning was explicitly used for the final ResNet50 implementation by replacing the original ImageNet output layer with a seven-class output layer. The study does not propose a new network; its model contribution is evidence-based selection and further training of an existing architecture.

2.3 Datasets, Preprocessing and Generalisation

HAM10000 is one of the most widely used datasets for multi-class lesion classification. It contains 10,015 dermatoscopic images across seven diagnostic categories and was released through the ISIC ecosystem [7]. The dataset is substantially imbalanced, with melanocytic nevi dominating the collection and dermatofibroma and vascular lesions represented by far fewer images. This imbalance makes aggregate accuracy potentially misleading if minority classes are not examined carefully.

Preprocessing and augmentation are common responses to limited data and class imbalance. Typical pipelines resize images to the network input dimension, normalise pixel values and apply geometric or photometric transformations. Alwakid et al. combined enhancement, segmentation and augmentation on HAM10000 and reported performance in the mid-80% range for their evaluated CNN variants [28]. Wang et al. used hair removal and multiple augmentation operations before multiscale classification [22]. Such procedures can reduce overfitting, but augmentation cannot create demographic diversity that is absent from the source data.

Representativeness therefore remains a distinct issue. Wen et al. found that publicly available skin-cancer image datasets contain limited reporting of ethnicity and skin type and substantially under-represent darker skin [9]. Daneshjou et al. similarly identified transparency and bias concerns in dermatology AI datasets [10], and later showed measurable performance disparities on a curated dataset containing diverse skin tones [11]. A model trained and tested only on HAM10000 should therefore be evaluated externally before its results are generalised to another population or acquisition environment.

2.4 Image Suitability and Web-Based Decision Support

Image suitability becomes important once a classifier accepts user uploads. Jalaboi et al. developed an explainable image-quality assessment approach for teledermatology and identified common acquisition problems such as poor framing, lighting, blur, low resolution and distance [12]. Jiang et al. found that many patient-submitted images were not judged sufficiently useful for decision-making [13]. Han et al. demonstrated that a state-of-the-art classifier can degrade when applied to patient-driven image sources [14]. These findings support a separate screening stage that can reject or warn on unsuitable images before disease classification.

Research on image-quality assistance also shows that screening should be evaluated as its own task. Shih and Oh review current AI approaches to dermatological image-quality assessment and emphasise the need for better data and prospective validation [15]. Vodrahalli et al. reported that an AI support tool could improve telemedicine photo quality by giving users feedback at acquisition time [16]. The implication for this study is limited but practical: image-suitability screening can be integrated before classification, but its effect on downstream disease-classification accuracy must be measured separately rather than assumed.

Web and mobile deployment has also been demonstrated previously. Akyeramfo-Sam et al. developed a browser-based CNN system for skin-disease prediction [17], while Maduranga and Nandasena implemented a MobileNet-based mobile system using HAM10000 [18]. These projects establish technical feasibility. The present study extends the engineering scope by linking prediction to authentication, role-based access, patient records, diagnostic history and reporting. Those features improve traceability and workflow structure, but do not

by themselves establish clinical effectiveness. The distinction is consistent with the broader evidence expectations surrounding authorised AI-enabled dermatology devices [19].

2.5 Synthesis and Research Gap

The literature establishes three points. First, CNN-based lesion classification is technically feasible and can achieve strong benchmark performance [5], [8], [20]. Second, transfer learning makes established vision architectures practical for dermatology datasets that are much smaller than general-purpose image collections [5], [6]. Third, web and mobile interfaces can make a trained classifier accessible without specialised desktop software [17], [18].

The open problem is not simply whether a CNN can classify a dermatoscopic image. Architecture performance varies by dataset and protocol; external validation and demographic representativeness remain limited; uncontrolled user-uploaded images introduce an input-suitability problem; and many research prototypes stop at prediction rather than integrating record keeping and access control [3], [9]–[16]. This study addresses a defined subset of that gap by comparing five CNN architectures on one seven-class task, further training the selected ResNet50 model, integrating image-suitability screening, and embedding the resulting inference pipeline in a web-based decision-support workflow. External generalisation, local clinical validation and dermatologist comparison remain outside the evidence of the present study.

3. Methodology

3.1 Research Design

The project combined comparative deep-learning experimentation with web-system development. The experimental sequence was: preparation of the skin-lesion dataset, evaluation of five CNN architectures, selection of the strongest model according to the reported project metrics, additional 30-epoch training of ResNet50, and integration of the selected classifier with a separate image-suitability Gatekeeper. The system-development sequence connected this inference pipeline to secure access, patient records, diagnostic history and report generation.

The study is experimental and engineering-oriented rather than clinical. Model performance was evaluated on project image data, while the application was assessed functionally through its implemented features. No prospective patient study, dermatologist reader study or clinical outcome evaluation was performed.

3.2 Dataset and Classification Task

The disease-classification task used HAM10000, a publicly available collection of 10,015 dermatoscopic images across seven diagnostic categories [7]. The seven implemented output classes are actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc). These seven classes are used consistently in this manuscript because they match the implemented output layer.

Separate training, validation and testing partitions were used; the exact final split counts were not recorded consistently and are therefore not reported. The complete HAM10000 class counts are shown in Table 1 to describe the source dataset and its imbalance.

Table 1. HAM10000 skin-lesion categories used by the classifier

Class code	Diagnostic category	Images
akiec	Actinic keratoses	327
bcc	Basal cell carcinoma	514
bkl	Benign keratosis-like lesions	1,099
df	Dermatofibroma	115

Class code	Diagnostic category	Images
mel	Melanoma	1,113
nv	Melanocytic nevi	6,705
vasc	Vascular lesions	142
	Total	**10,015**

The distribution is strongly imbalanced: melanocytic nevi account for about two thirds of the dataset, while dermatofibroma and vascular lesions are much less frequent. This imbalance is relevant when interpreting aggregate metrics.

3.3 Image Preprocessing and Augmentation

The inference implementation converts uploaded images to RGB, resizes them to 224 × 224 pixels, converts them to tensor form and applies normalisation before prediction. The training methodology also describes image augmentation using rotation, flipping, scaling and brightness adjustment. These operations are consistent with common practice in skin-lesion classification because they expose the model to variation in orientation and acquisition.

The exact numerical ranges and probabilities for each augmentation operation, and all normalisation constants, were not recorded consistently. During inference, the final ResNet50 implementation also evaluates the original image and a horizontally flipped version and averages the resulting probabilities, providing a simple form of test-time augmentation.

3.4 CNN Models and Transfer Learning

Five architectures were evaluated in the project: MobileNetV2, EfficientNetB0, ResNet50, DenseNet121 and InceptionV3. Their roles were comparative rather than novel: the aim was to determine which established backbone provided the strongest result for the project dataset. All five were assessed using the same reported output metrics—test accuracy, precision, recall, F1-score and training time.

Transfer learning was explicitly used for the final ResNet50 classifier. The original ImageNet-oriented output layer was replaced with a seven-output classification layer corresponding to the classes in Table 1. A complete common hyperparameter record (optimizer, learning rate, batch size, scheduler, random seed and freeze/unfreeze settings) was not retained for all five model runs.

3.5 Model Selection and 30-Epoch ResNet50 Retraining

The initial comparison identified ResNet50 as the strongest model on the project task because it achieved the highest recorded test accuracy and F1-score. ResNet50 was then selected for a further training stage of 30 epochs.

3.6 Evaluation Metrics

The project reported test accuracy, precision, recall, F1-score and training time. Accuracy measures the proportion of test examples classified correctly. Precision expresses how often predictions assigned to a class are correct, recall expresses how much of the relevant class is recovered, and F1-score combines precision and recall through their harmonic mean. Training time was recorded to show the computational cost of the compared architectures.

Precision, recall and F1-score are weighted averages across the seven classes; weighted recall is therefore identical to accuracy, as seen in Table 2. ROC-AUC, specificity, per-class metrics and a seven-class confusion matrix were not computed in this study and are therefore not reported.

3.7 Gatekeeper Image-Suitability Screening

A separate Gatekeeper component was placed before the ResNet50 classifier to screen whether an uploaded image was appropriate for lesion classification. The Gatekeeper has three practical

outcomes: a valid lesion image is allowed to proceed; a clearly invalid or unsuitable image is rejected and the user is asked to submit a clearer lesion image; and an uncertain image may proceed with a warning attached to the resulting decision-support record.

The Gatekeeper was intended to address unrelated photographs, screenshots, blurred images, low-light images, full-body photographs and other inputs that a fixed seven-class classifier should not be forced to label. Although the component was functionally integrated, no independent quantitative evaluation of Gatekeeper performance was conducted. The paper therefore describes image-suitability screening as an implemented system feature and does not claim that it measurably increased ResNet50 classification accuracy.

3.8 Web Application Architecture

The prototype integrates the AI pipeline into a web application designed for healthcare-user workflows. The frontend uses Next.js and TypeScript. The backend is Python-based and coordinates image upload, preprocessing, model inference, authentication and data handling.

The implemented workflow supports secure login, email-based one-time-password verification, role-based access control, patient registration, image upload, disease prediction with a confidence score, diagnostic history and report generation. The system provides three user roles and a super-administrator function for managing system access. These features provide traceability around a prediction, but they do not constitute evidence of clinical effectiveness.

3.9 Ethical and Clinical Scope

The model-development work used public research images rather than a prospectively recruited patient cohort. No locally collected patient images were used to establish the classification results reported in this paper, and clinical validation against conventional diagnostic practice was not completed.

The present system is therefore a research prototype and decision-support demonstration. Any future evaluation using identifiable patient data or prospective clinical use should be conducted under appropriate institutional ethics review, consent procedures, privacy safeguards and data-governance requirements. The system is not presented as an autonomous diagnostic tool or as a replacement for qualified clinical judgement.

4. Results and Discussion

4.1 Comparative Performance of the Five CNN Architectures

Table 2 and Figure 1 report the five-model comparison. ResNet50 achieved the highest initial test accuracy (72.06%) and F1-score (74.00%). EfficientNetB0 was the next strongest model on accuracy (68.00%) and F1-score (70.93%). MobileNetV2 and EfficientNetB0 reported slightly higher precision than ResNet50, showing that the selected model did not lead every metric.

Table 2. Comparative performance of the five CNN architectures

Architecture	Accuracy	Precision	Recall	F1-score	Training time
MobileNetV2	67.60%	77.99%	67.60%	70.71%	2,235.83 s
EfficientNetB0	68.00%	77.82%	68.00%	70.93%	1,895.59 s
ResNet50	**72.06%**	**77.71%**	**72.06%**	**74.00%**	**11,183.51 s**
DenseNet121	63.34%	76.83%	63.34%	67.06%	12,376.54 s
InceptionV3	63.21%	73.49%	63.21%	66.79%	24,190.24 s

The margin between ResNet50 and EfficientNetB0 was 4.06 percentage points in accuracy and 3.07 percentage points in F1-score. The overall accuracy spread from ResNet50 to InceptionV3

was 8.85 percentage points. These differences show that architecture choice mattered within the project. However, the result should be interpreted as a within-project comparison rather than evidence that ResNet50 is universally superior, because architecture ranking can change with training strategy, dataset composition and preprocessing.

Training cost produced a different ranking. EfficientNetB0 had the lowest recorded training time, whereas InceptionV3 required the most time and achieved the lowest accuracy. ResNet50 required substantially more training time than the lightweight models to obtain its accuracy advantage. This creates an accuracy-efficiency trade-off: ResNet50 was selected because the project prioritised predictive performance, while EfficientNetB0 and MobileNetV2 remained attractive alternatives where computational efficiency is more important.

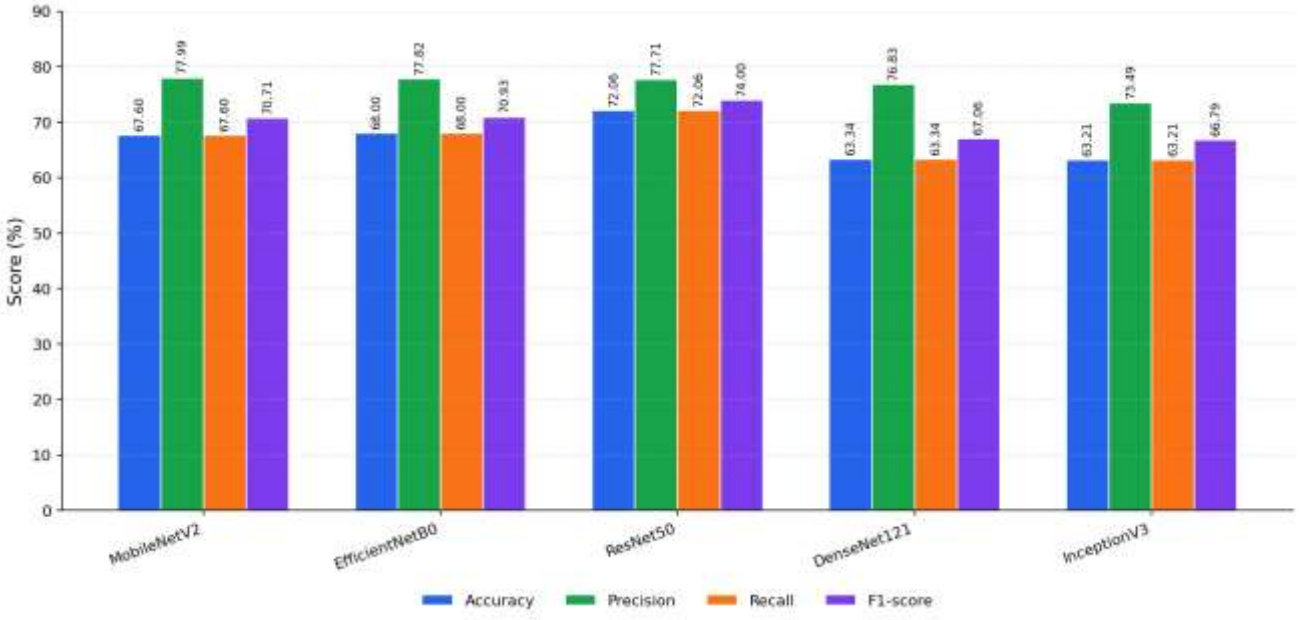


Figure 1. Comparative classification performance of MobileNetV2, EfficientNetB0, ResNet50, DenseNet121 and InceptionV3.

4.2 ResNet50 Performance After 30-Epoch Retraining

After model selection, ResNet50 underwent the additional 30-epoch training stage documented in the project results. Table 3 and Figure 2 compare the initial ResNet50 result with the final retrained result. Test accuracy increased from 72.06% to 80.24%, a gain of 8.18 percentage points, while F1-score increased from 74.00% to 80.85%, a gain of 6.85 percentage points.

Table 3. ResNet50 performance before and after the additional 30-epoch training stage

Metric	Initial	After 30 epochs	Difference
Test accuracy	72.06%	**80.24%**	+8.18 percentage points
F1-score	74.00%	**80.85%**	+6.85 percentage points

The final 80.24% accuracy should not be inserted into the original five-model comparison as though every architecture received the same additional training. Only ResNet50 underwent the 30-epoch retraining stage. Table 2 therefore answers the model-selection question, while Table 3 answers a different question: how the selected model changed after further training. The project did not repeat the same extra stage for the other four architectures, so the post-training result cannot establish how they would have ranked under equivalent fine-tuning.

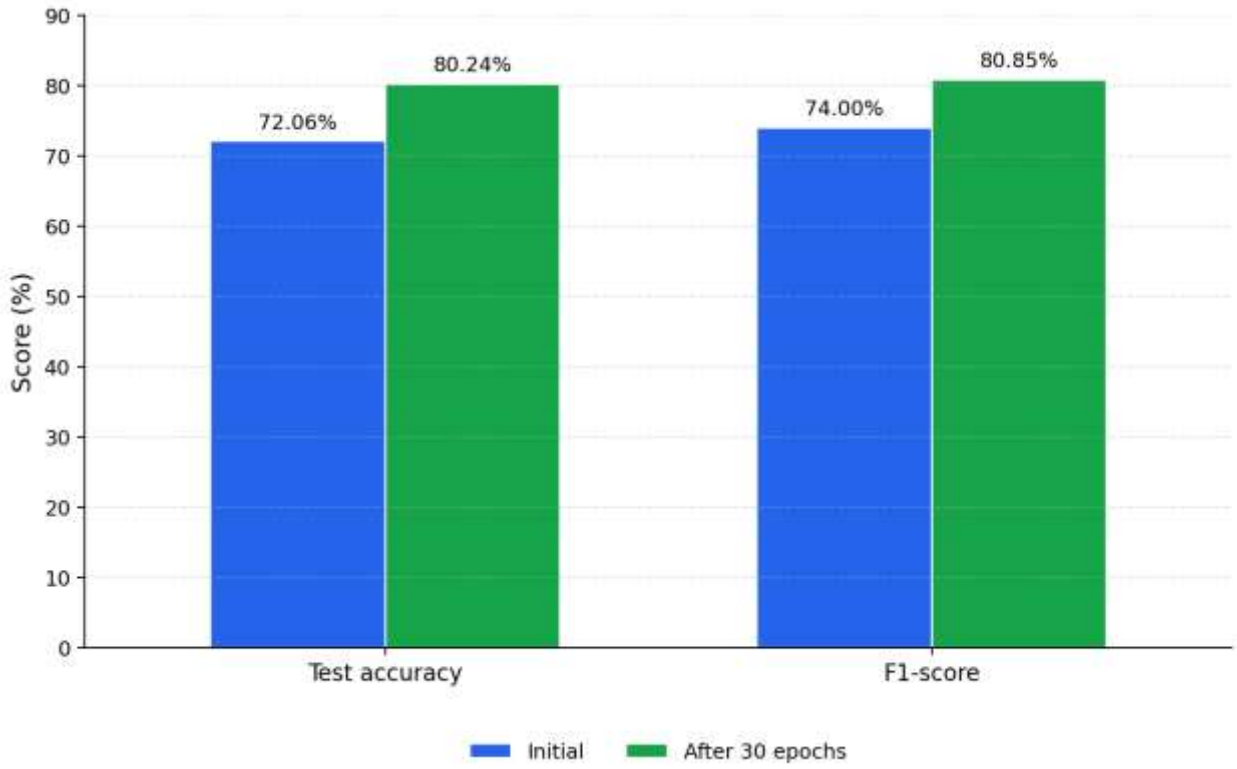


Figure 2. ResNet50 test accuracy and F1-score before and after the additional 30-epoch training stage.

4.3 Image-Suitability Screening and System Functionality

The Gatekeeper addresses a deployment failure mode that the classifier metrics do not capture: whether the uploaded image should be classified at all. The implemented workflow screens image suitability before disease prediction (Figure 3). Suitable images proceed to ResNet50, clearly unsuitable images are rejected, and uncertain inputs may continue with a warning. This structure prevents the application from treating every upload as a valid dermatological image, although the study did not quantify the Gatekeeper's independent classification performance or its downstream effect on ResNet50 accuracy.

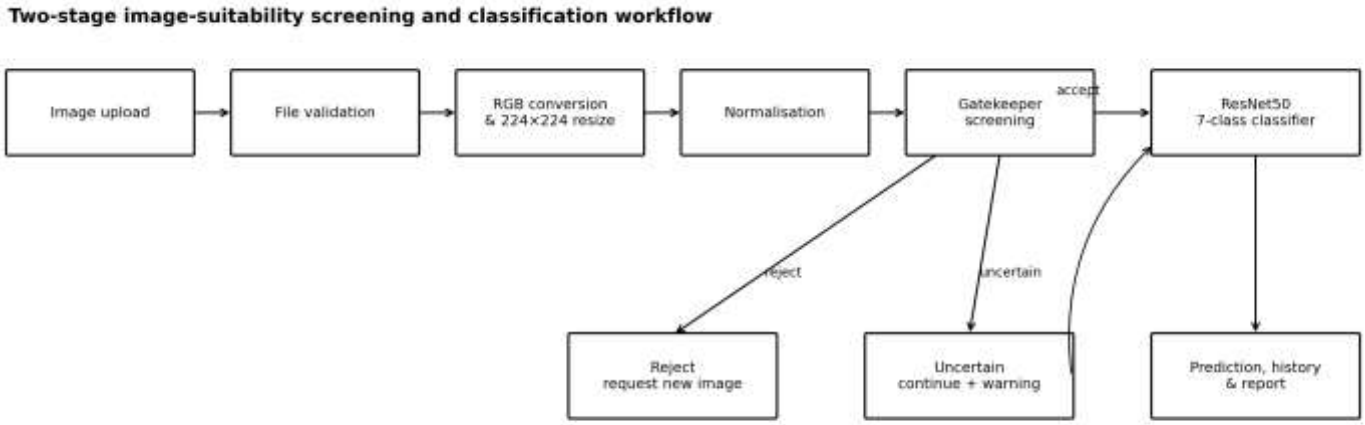


Figure 3. Two-stage image-suitability screening and seven-class classification workflow.

The application also demonstrates how model output can be connected to a persistent workflow. Table 4 summarizes the implemented system functions, and Figures 4 and 5 show the system architecture and selected interfaces. Secure authentication, email OTP verification, role-based access, patient records, diagnostic history and report generation were implemented alongside image upload and prediction. Clinical validation was not completed.

Table 4. Implemented system functions and evidence status

Function	Status	Evidence
Secure login	Implemented	System implementation and findings
Email one-time-password verification	Implemented	Authentication workflow
Role-based access control	Implemented	Three-role access model
Patient registration and records	Implemented	Patient-management workflow
Image upload and seven-class prediction	Implemented	ResNet50 inference workflow
Diagnostic history	Implemented	Stored scan/prediction records
Report generation	Implemented	Generated decision-support reports
Gatekeeper suitability screening	Implemented	Pre-classification screening workflow
Clinical validation against conventional diagnosis	**Not completed**	**Identified as incomplete**

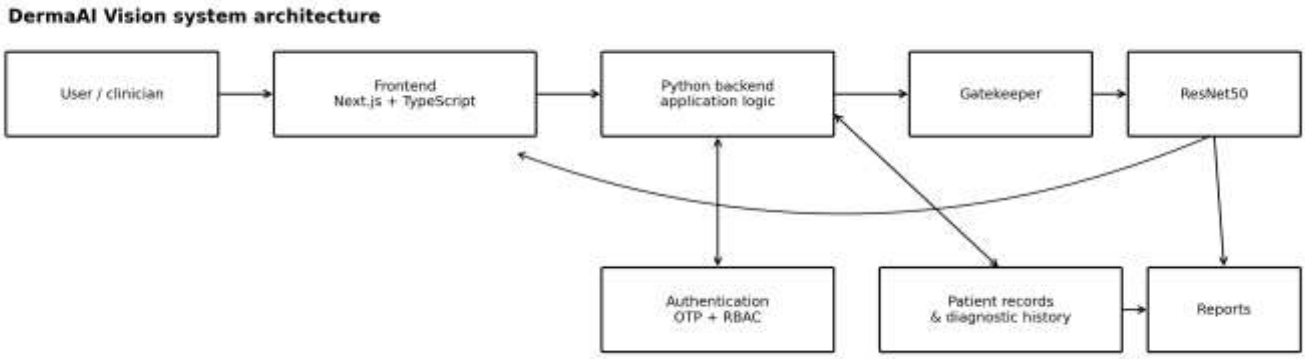


Figure 4. Architectural view of the DermaAI Vision prototype.

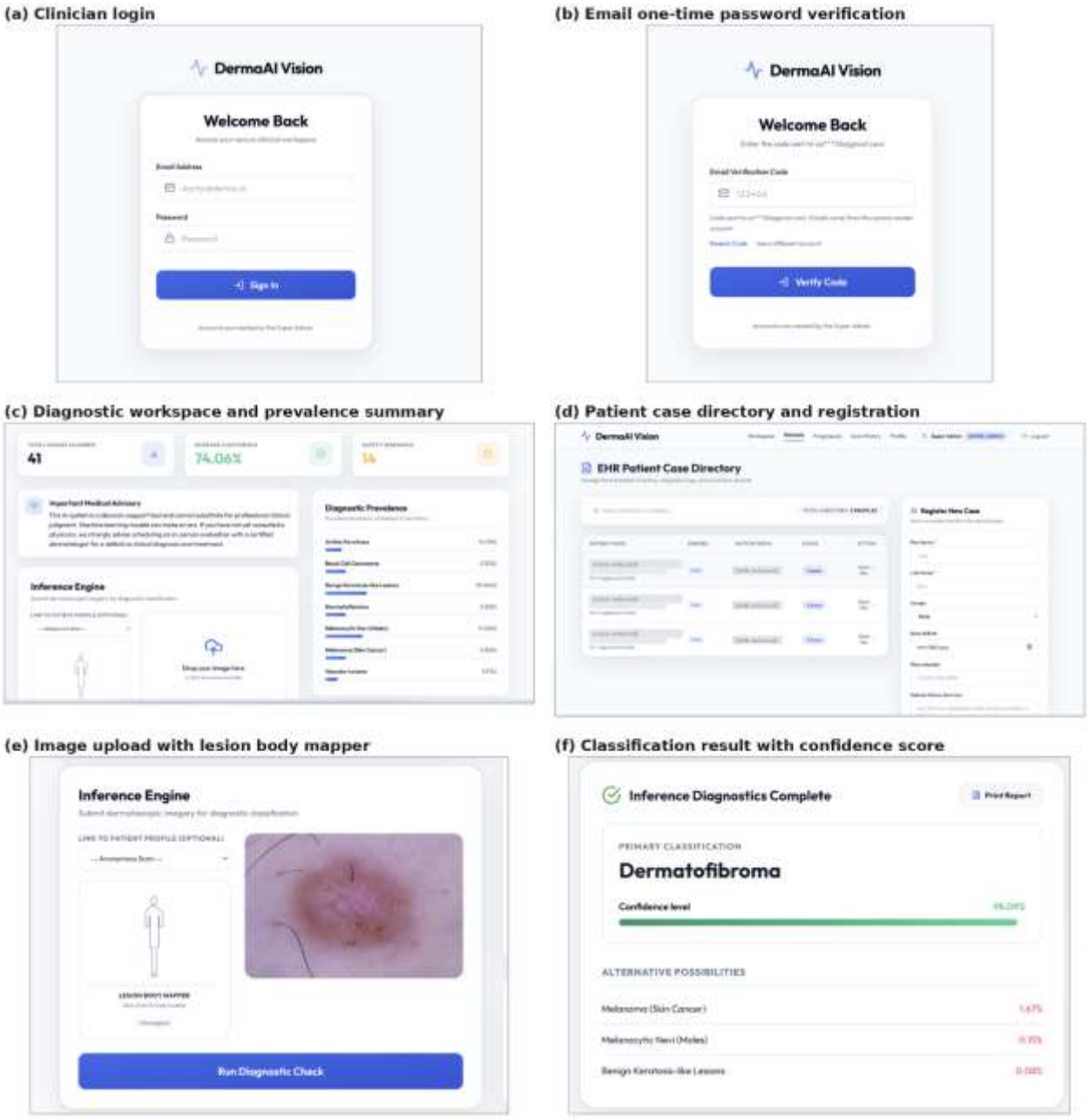


Figure 5. Selected interfaces of the DermaAI Vision prototype, including authentication, patient management, image upload and prediction output.

4.4 Interpretation and Comparison with Previous Research

The initial comparison supports ResNet50 as the strongest of the five evaluated architectures for this project because accuracy and F1-score agreed on the selection. The result should not be reduced to the idea that a deeper network is automatically better. InceptionV3 and DenseNet121 required more recorded training time yet achieved lower accuracy, while EfficientNetB0 was much faster and remained close to ResNet50. The practical implication is that model choice depends on the objective: the present project prioritised classification performance, whereas a mobile or frequently retrained deployment might favour a lighter network.

The 30-epoch retraining produced the largest improvement recorded in the study. The evidence establishes that the selected ResNet50 performed better after this additional stage, but it does not isolate why. Longer exposure to the training data, fine-tuning decisions and other

training factors may all contribute. Because the project did not include an ablation in which each factor was varied separately, the appropriate conclusion is an observed performance improvement rather than a causal explanation of that improvement.

The final 80.24% accuracy is lower than several published figures, including the 91.24% reported by Wang et al. on ISIC2018 [22] and the mid-80% results reported by Alwakid et al. on HAM10000-based experiments [28]. Direct ranking would be misleading because preprocessing, augmentation, split construction, model training and evaluation metrics differ between studies. The ISIC challenge itself showed that models with similar benchmark performance can differ in generalisation [8]. The meaningful comparison is therefore qualitative: the present project achieved moderate-to-strong seven-class performance while leaving clear room for improvement in evaluation depth and external validation.

The system contribution is complementary to the classifier result. Prior web and mobile systems established that skin-image predictions can be delivered through accessible interfaces [17], [18]. This project adds image-suitability screening, authenticated users, role-based access, patient records, history and reporting. Those features make the output traceable and reviewable, but they should not be described as proof of clinical benefit. Authorised AI-enabled dermatology devices require evidence from clinical studies that goes beyond technical implementation [19].

4.5 Limitations and Recommendations

First, the disease-classification results come from a single public dataset. HAM10000 does not represent the full range of skin tones, imaging devices or acquisition conditions expected in the intended Somali setting, and no independent external dataset or locally collected Somali clinical images were used. The reported performance therefore describes the project dataset and should not be assumed to generalise directly to the target population [9]–[11].

Second, the source dataset is highly imbalanced. The study reports aggregate accuracy, precision, recall and F1-score, but a complete per-class evaluation and a seven-class confusion matrix were not produced. This limits interpretation of performance on minority categories. Future evaluation should report per-class precision, recall, F1-score and confusion-matrix counts, with particular attention to clinically important errors.

Third, reproducibility is limited because the experimental configuration was not fully recorded. The model families, 224 × 224 input, major preprocessing steps, comparative metrics and the final 30-epoch ResNet50 retraining are documented, but not every split count, optimizer setting, learning rate, batch size, random seed and augmentation parameter across the five-model comparison. A future rerun should use a fixed documented protocol and retain machine-readable training logs.

Fourth, the Gatekeeper was functionally integrated but not independently quantified, and no controlled comparison measured classification with versus without screening. The application also received functional rather than clinical evaluation: no dermatologist comparison, reader study, prospective patient assessment or clinician usability study was performed. Future work should evaluate Gatekeeper performance on realistic poor and irrelevant images, test the complete pipeline on a partition unseen by all models, validate the classifier on external and local data, and conduct clinician-centred studies under appropriate ethical oversight.

Further model development should also consider class-balancing strategies, richer local data, lesion-specific preprocessing and explainability methods such as Grad-CAM. These additions should be tested through controlled experiments so that any observed improvement can be attributed to a defined methodological change rather than to several changes introduced simultaneously.

5. Conclusion

This study presented a comparative and system-oriented decision-support investigation of skin-lesion classification. Five CNN architectures—MobileNetV2, EfficientNetB0, ResNet50, DenseNet121 and InceptionV3—were evaluated on the seven-class HAM10000 task using the project's recorded metrics. ResNet50 produced the strongest initial accuracy (72.06%) and F1-score (74.00%) and was selected for a further 30-epoch training stage. After retraining, test accuracy reached 80.24% and F1-score reached 80.85%, improvements of 8.18 and 6.85 percentage points. The selected classifier was then integrated with a Gatekeeper image-suitability screen and a web-based workflow supporting authentication, email OTP verification, role-based access, patient records, diagnostic history and report generation. The evidence supports comparative model selection, further development of ResNet50 and functional integration into a research prototype. It does not establish external generalisation, dermatologist-level performance or clinical effectiveness. No external dataset, locally collected Somali clinical dataset, prospective patient evaluation or reader study was completed. The most important next steps are therefore reproducible rerunning of the experiments with complete logs, class-specific evaluation, independent Gatekeeper testing, external and local validation, and clinician-centred assessment. DermaAI Vision should remain positioned as a research-stage web-based decision-support prototype rather than an autonomous diagnostic system until those steps are completed.

Conflict of Interest

The authors confirm that there is no conflict of interest related to this manuscript.

References

[1] L. Deng, S. Fan, H. Zhao, J. Song, W. Li, and X. Xu, "Global, regional and national burden of skin and subcutaneous diseases: a systematic analysis of the Global Burden of Disease Study 2021," International Health, vol. 18, no. 2, pp. 183–196, 2026, doi: 10.1093/inthealth/ihaf070.

[2] E. E. Freeman, "Global health dermatology: An emerging field addressing the access to care crisis," Indian Journal of Dermatology, Venereology and Leprology, vol. 90, pp. 3–4, 2024, doi: 10.25259/IJDVL_1361_2023.

[3] S. P. Choy et al., "Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease," npj Digital Medicine, vol. 6, art. no. 180, 2023, doi: 10.1038/s41746-023-00914-8.

[4] O. T. Jones et al., "Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review," The Lancet Digital Health, vol. 4, no. 6, pp. e466–e476, 2022, doi: 10.1016/S2589-7500(22)00023-1.

[5] S. Riaz, A. Naeem, H. Malik, R. A. Naqvi, and W.-K. Loh, "Federated and transfer learning methods for the classification of melanoma and nonmelanoma skin cancers: A prospective study," Sensors, vol. 23, no. 20, art. no. 8457, 2023, doi: 10.3390/s23208457.

[6] Z. Wu et al., "Studies on different CNN algorithms for face skin disease classification based on clinical images," IEEE Access, vol. 7, pp. 66505–66511, 2019, doi: 10.1109/ACCESS.2019.2918221.

[7] 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, art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[8] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.

[9] D. Wen et al., “Characteristics of publicly available skin cancer image datasets: a systematic review,” The Lancet Digital Health, vol. 4, no. 1, pp. e64–e74, 2022, doi: 10.1016/S2589-7500(21)00252-1.
[10] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, “Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review,” JAMA Dermatology, vol. 157, no. 11, pp. 1362–1369, 2021, doi: 10.1001/jamadermatol.2021.3129.
[11] R. Daneshjou et al., “Disparities in dermatology AI performance on a diverse, curated clinical image set,” Science Advances, vol. 8, no. 32, art. no. eabq6147, 2022, doi: 10.1126/sciadv.abq6147.
[12] R. Jalaboi, O. Winther, and A. Galimzianova, “Explainable image quality assessments in teledermatological photography,” Telemedicine and e-Health, vol. 29, no. 9, pp. 1342–1348, 2023, doi: 10.1089/tmj.2022.0405.
[13] S. W. Jiang et al., “Quality and perceived usefulness of patient-submitted store-and-forward teledermatology images,” JAMA Dermatology, vol. 158, no. 10, pp. 1183–1186, 2022, doi: 10.1001/jamadermatol.2022.2815.
[14] S. S. Han et al., “The degradation of performance of a state-of-the-art skin image classifier when applied to patient-driven internet search,” Scientific Reports, vol. 12, art. no. 16260, 2022, doi: 10.1038/s41598-022-20632-7.
[15] P. E. Shih and D. H. Oh, “Artificial intelligence for dermatological image quality assessment,” Current Dermatology Reports, vol. 14, art. no. 23, 2025, doi: 10.1007/s13671-025-00483-x.
[16] K. Vodrahalli, J. Ko, A. S. Chiou, R. Novoa, A. Abid, M. Phung, K. Yekrang, P. Petrone, J. Zou, and R. Daneshjou, “Development and clinical evaluation of an artificial intelligence support tool for improving telemedicine photo quality,” JAMA Dermatology, vol. 159, no. 5, pp. 496–503, 2023, doi: 10.1001/jamadermatol.2023.0091.
[17] S. Akyeramfo-Sam, A. Addo Philip, D. Yeboah, N. C. Nartey, and I. Kofi Nti, “A web-based skin disease diagnosis using convolutional neural networks,” International Journal of Information Technology and Computer Science, vol. 11, no. 11, pp. 54–60, 2019, doi: 10.5815/ijitcs.2019.11.06.
[18] M. W. P. Maduranga and D. Nandasena, “Mobile-based skin disease diagnosis system using convolutional neural networks (CNN),” International Journal of Image, Graphics and Signal Processing, vol. 14, no. 3, pp. 47–57, 2022, doi: 10.5815/ijigsp.2022.03.05.
[19] K. P. Venkatesh, K. T. Kadakia, and S. Gilbert, “Learnings from the first AI-enabled skin cancer device for primary care authorized by FDA,” npj Digital Medicine, vol. 7, art. no. 156, 2024, doi: 10.1038/s41746-024-01161-1.
[20] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, pp. 115–118, 2017, doi: 10.1038/nature21056.
[21] V. Gautam, N. K. Trivedi, A. Anand, and R. G. Tiwari, “Early skin disease identification using deep neural network,” Computer Systems Science and Engineering, vol. 44, no. 3, pp. 2259–2275, 2023, doi: 10.32604/csse.2023.026358.
[22] G. Wang, P. Yan, Q. Tang, L. Yang, and J. Chen, “Multiscale feature fusion for skin lesion classification,” BioMed Research International, vol. 2023, art. no. 5146543, 2023, doi: 10.1155/2023/5146543.
[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[25] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Machine Learning (ICML), PMLR, vol. 97, 2019, pp. 6105–6114.
[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception architecture for computer vision,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
[28] G. Alwakid, W. Gouda, M. Humayun, and N. U. Sama, “Melanoma detection using deep learning-based classifications,” Healthcare, vol. 10, no. 12, art. no. 2481, 2022, doi: 10.3390/healthcare10122481.