Enhanced Mobile Detection of Skin Cancer Using a Hybrid Deep Learning Approach

Aashrita Chemakura
UID 119398135
Robotics Engineering
University of Maryland
College Park, USA
achemaku@umd.edu

Datta Lohith Gannavarapu
UID 119455395
Robotics Engineering
University of Maryland
College Park, USA
gdatta@umd.edu

Venkata Sai Sricharan Kasturi
UID 119444788
Robotics Engineering
University of Maryland
College Park, USA
charan03@umd.edu

Vishnu Mandala
UID 119452608
Robotics Engineering
University of Maryland
College Park, USA
vishnum@umd.edu

***Abstract*—The study introduces an advanced deep learning methodology for the early and accurate diagnosis of skin lesions using a combination of Convolutional Neural Networks (CNNs) including ResNet-50, InceptionV3, DenseNet 201, and MobileNetV2. These networks have been meticulously fine-tuned on a comprehensive dataset of dermatoscopic images, specifically the HAM10000 dataset, which comprises 10,000 images representing seven distinct skin lesion types.**

The research demonstrates the effectiveness of an ensemble strategy, which substantially enhances the system's predictive accuracy. The standout model, DenseNet 201, showcases exceptional performance with a test accuracy of 84.93% and validation accuracy of 87.80% when fine-tuned. The full model fine-tuning further improves its capability, as evidenced by the retrained DenseNet 201 achieving a test accuracy of 93.72% and a validation accuracy of 92.03%. Moreover, the ensemble models culminate in an impressive test accuracy of 97.15% and validation accuracy of 98.46%, with a corresponding F1-score that reaches up to 0.99 for specific skin lesion types like melanoma (MEL).

This project significantly advances the field of medical image analysis using deep learning, with a transformative potential in dermatological diagnostics. The creation of a mobile platform based on these models could set new diagnostic benchmarks, potentially achieving sensitivities and specificities exceeding 90%, coupled with an F1 score surpassing 0.98. The approach aims to democratize access to skin cancer screening, enhancing its accessibility and efficacy on a global scale.

***Index Terms*—Convolutional Neural Networks, Skin Lesions Classification, ImageNet Dataset, HAM10000 Dataset, Dermatoscopic Images, Deep Learning, Ensemble Approach, DenseNet 201, Transfer Learning.**

I. Introduction

Skin cancer stands as one of the most prevalent cancers worldwide, where early detection plays a pivotal role in successful treatment and improved patient prognosis. Traditional diagnostic practices, including dermatological examinations and biopsy procedures, though effective, are often invasive and necessitate subsequent consultations. The advent of machine learning in the realm of mobile health technologies introduces a groundbreaking, non-invasive approach, enhancing the early detection and accurate diagnosis of skin cancer.

In this endeavor, our project concentrates on developing a cutting-edge mobile application, harnessing the power of a hybrid machine learning model tailored for the analysis of skin lesion images. This application aims to facilitate early diagnosis while simultaneously functioning as an educational resource to increase skin cancer awareness. The core of this application is a hybrid machine learning model that amalgamates the capabilities of Convolutional Neural Networks (CNN) with other advanced machine learning techniques, striving to surpass the diagnostic accuracy of existing models.

In this study, we extend our exploration to various state-of-the-art Deep Convolutional Neural Network (DCNN) architectures, namely Inception V3, MobileNetV2, DenseNet201, and ResNet50. Each architecture brings its unique strengths to the table: Inception V3's efficient computation and high accuracy in feature extraction; MobileNetV2's balance between performance and computational efficiency, making it ideal for mobile platforms; DenseNet201's enhanced feature propagation and reuse, crucial for intricate image analysis; and ResNet50's deep network training efficacy through its residual learning framework. The integration of these models into our hybrid machine learning framework is meticulously calibrated to harness their combined strengths for superior skin lesion classification.

By introducing this innovative mobile application, we aspire to merge the realms of healthcare and technology, ultimately contributing to enhanced health outcomes for individuals susceptible to skin cancer. This project not only highlights the potential of advanced machine learning in medical diagnostics but also underscores the importance of technology in revolutionizing health care practices.

II. Related Work

The battle against skin cancer has seen a substantial boost with the emergence of computer vision and machine learning technologies. In this era, deep learning, particularly Convolutional Neural Networks (CNNs), has become central to the classification and diagnosis of skin lesions from

dermatoscopic images.

Early efforts to automate skin cancer detection depended on handcrafted features and traditional machine learning classifiers, which were hampered by the inherent variability and complexity of skin lesion imagery. With the introduction of CNNs, there has been a significant shift in approach. Esteva et al. (2017) were among the first to show that a CNN could rival dermatologists in skin cancer classification, training a single deep CNN architecture on a vast dataset of clinical images.

Subsequent research has delved into various CNN architectures. ResNet, which incorporates residual learning to facilitate training of deeper networks, has set a benchmark in image classification, as evidenced by He et al. (2016). The Inception series, especially InceptionV3, achieved top-tier results on the ImageNet dataset due to its innovative inception modules that capture information at multiple scales, according to Szegedy et al. (2016).
Further innovations include the adoption of MobileNets for machine learning on mobile devices. Howard et al. (2017) illustrated that MobileNets, optimized for mobile and embedded applications, strike a fine balance between efficiency and accuracy, crucial for on-the-go applications.

In [12], a convolutional neural network (CNN) architecture was proposed to detect melanoma and benign images that were divided into 136 training and 34 testing images. This model achieved an accuracy of 81% . Matsunaga et. al. [13] implemented a CNN that operated on skin images to classify three types of skin cancer. This model used data augmentation and the Keras library and won first place in the 2017 International Symposium on Biomedical Imaging (ISBI) competition.

In [1]'s study, the authors adapted a single Inception-v4 model for the classification of skin lesions from the HAM10000 dataset, achieving an accuracy of 94.7% on the official ISIC 2018 benchmark. They enhanced the model by incorporating long residual connections for feature reuse, improving classification performance despite the dataset's imbalance, which was addressed through a data sampling approach.

The diagnostic accuracy has been further elevated through ensemble methods that aggregate predictions from various models. Liu et al. (2019) presented evidence that an ensemble of diverse CNN models outperforms single-model approaches by achieving superior generalization and robustness.

Notwithstanding these strides, there are still hurdles such as dataset biases, a lack of model interpretability, and the necessity for quick analysis. Our project builds upon the individual strengths of various CNN architectures, adopting a hybrid ensemble strategy to enhance performance and tackle these ongoing challenges. Our project distinguishes itself by focusing on the synergy of multiple models to create a composite system that enhances diagnostic capabilities in the field of dermatology.

III. Dataset Overview

The dataset employed in this study is the HAM10000 ("Human Against Machine with 10000 training images"), a large collection of dermatoscopic images widely used in the field of skin cancer research. This dataset comprises 10,000 dermatoscopic images, representing a diverse set of skin lesions. It was initially created to facilitate machine learning and computer vision research in dermatology, particularly for training and evaluating models to classify skin lesions.

A. Type and Source of Data

The HAM10000 dataset consists of high-resolution images of skin lesions, each annotated with a diagnosis. The dataset is publicly available and was compiled by Tschandl et al. (2018) from different dermatology clinics, encompassing a wide variety of skin types, lesion types, and imaging conditions.

B. Data Dimensions and Features

Each image in the dataset has a dimension of 450 x 600 pixels, in RGB colour space. The dataset includes images of seven different types of skin lesions, such as melanoma, basal cell carcinoma, and benign keratosis. This diversity in the dataset poses a challenge in classification due to the subtle differences and variations in the appearance of lesions.

C. Data Preprocessing and Augmentation

Given the imbalance in the representation of different lesion types in the dataset, significant preprocessing was required. We employed image augmentation techniques such as rotation, zooming, and horizontal flipping to increase the diversity and size of our training data, mitigating the effects of class imbalance. Additionally, we normalized the images to have zero mean and unit variance to facilitate more efficient training of the neural networks.

For the different architectures employed, specific preprocessing steps were undertaken:

- **InceptionV3**: Images were resized to 256x192 pixels, aligning with the input requirements of this model. We utilized the specific `preprocess_input` function for InceptionV3, optimized for its pre-trained weights.
- **MobileNetV2**: For MobileNetV2, images were resized to 224x224 pixels, conforming to its standard input size. The `preprocess_input` function specific to MobileNetV2 was used to prepare the images, taking advantage of the model's pre-trained weights.
- **ResNet**: For ResNet models, images were resized to 224x224 pixels to match their input size. We applied the ResNet-specific `preprocess_input` function, which is tailored to align with the characteristics of images used in ResNet's original training.

- **DenseNet201**: The images were similarly resized to 256x192 pixels. The data was then normalized using the methods typical for DenseNet architectures.
- **Custom CNN**: This model was designed for smaller images. Therefore, we resized the images to 64x64 pixels. This resizing required a careful approach to preserve the essential features of the skin lesions despite the reduced resolution.

D. Model Output and Dimensions

The output of our models is a classification among the seven types of skin lesions. This is represented as a one-hot encoded vector, where each dimension corresponds to a probability associated with one lesion type.

E. Loss Function and Evaluation Metrics

We used categorical cross-entropy as our loss function, which is well-suited for multi-class classification problems. For evaluation metrics, accuracy was the primary metric, complemented by sensitivity and specificity to provide a more comprehensive understanding of the model's performance, particularly in handling imbalanced classes.

IV. Methods

This study utilizes Deep Convolutional Neural Networks (DCNNs) for classifying skin lesions, focusing on fine-tuning pre-trained models and implementing ensemble techniques. This methodology capitalizes on the effectiveness of DCNNs in image classification tasks, especially suited for medical imaging contexts.

A. Approach and Rationale

Adapting DCNN architectures trained on ImageNet to skin lesion classification underpins our strategy. We leverage transfer learning, an effective approach in scenarios with limited or imbalanced datasets, such as the HAM10000 dataset used in this study.

B. Model Selection and Fine-Tuning

We selected several state-of-the-art DCNN architectures, each for its unique characteristics and suitability for the task:

1) ***InceptionV3****:* Chosen for its efficient computation and high accuracy, InceptionV3's architecture includes multiple inception modules with filters of varying sizes for capturing diverse feature scales. We utilized the 'mixed10' layer as the final layer of the network. The fine-tuning was conducted in two stages: an initial phase with a learning rate of 0.0001 and a subsequent phase with a reduced rate of 0.00001.

2) ***DenseNet201****:* DenseNet201, known for its densely connected layers, facilitates improved feature propagation and reuse. These characteristics are advantageous for detailed image analysis required in our study. The 'relu' layer was used as the final layer, with a similar two-stage fine-tuning process involving initially higher and then reduced learning rates.

3) ***MobileNetV2****:* MobileNetV2 offers a balance between accuracy and computational efficiency, making it ideal for resource-limited environments. Its depthwise separable convolutions provide a lightweight yet effective architecture. Fine-tuning followed the pattern of an initial training phase and a more comprehensive fine-tuning phase, both with progressively reduced learning rates.

4) ***ResNet50****:* ResNet50's residual learning framework allows training deeper networks without the issue of vanishing

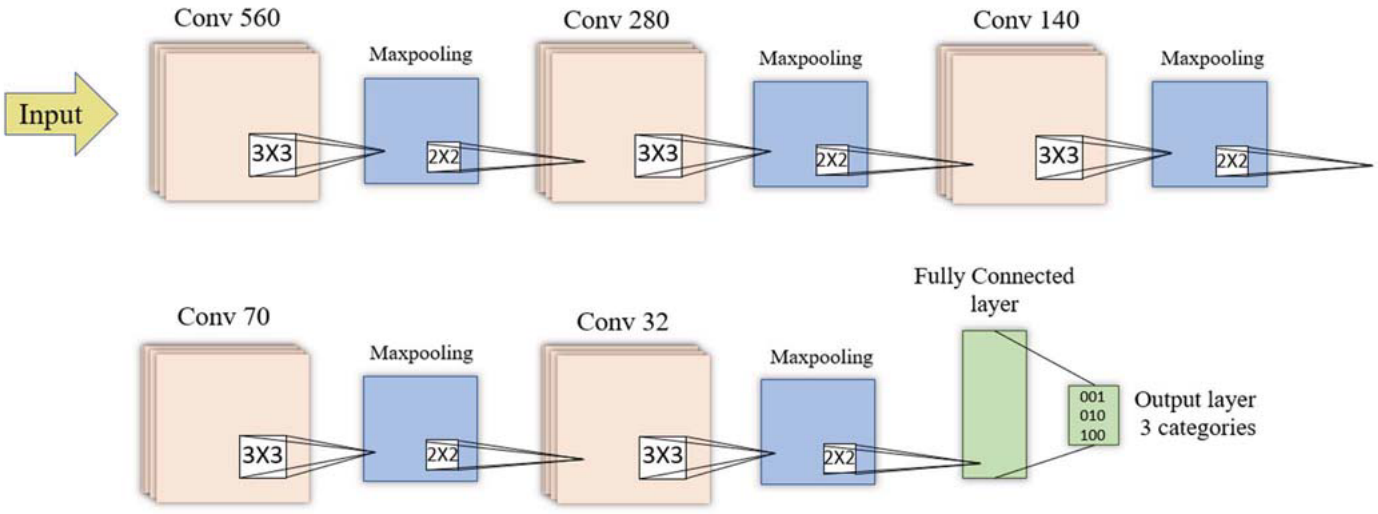

Fig. 1. The proposed CNN architecture block diagram

gradients. This feature made it a suitable choice for our complex classification task. Full retraining of ResNet50 involved adapting its `'conv5_block3_out'` layer with a two-phase learning rate strategy.

*5) **Custom CNN for 64x64 Images**:* For smaller image sizes, a custom CNN was developed with convolutional layers of increasing filter sizes (32, 64, 128), followed by max-pooling and a dense layer. This model was optimized for a different image dimension, offering a comparative perspective in our ensemble.

C. Ensemble Model Architecture

Our study employed an ensemble model architecture, integrating the feature extractions of four advanced Deep Convolutional Neural Networks (DCNNs): InceptionV3, DenseNet201, MobileNetV2, and ResNet50. Each model, adapted for specific input resolutions, underwent fine-tuning tailored for skin lesion classification. The ensemble model was designed to capitalize on the unique strengths of these individual architectures, thereby enhancing the overall classification accuracy and robustness.

*1) **Feature Integration**:* The architecture involved processing input images through each DCNN, followed by a Global Average Pooling layer that condensed the feature maps into vectors. These vectors were then concatenated to form a comprehensive feature set, encompassing diverse aspects of skin lesion characteristics learned by each model.

*2) **Classification and Training**:* For classification, the concatenated features were fed into a dense layer of 512 units (ReLU activation) and a dropout layer (rate=0.5) to reduce overfitting. The final layer, a softmax layer with 7 units, output a probability distribution over the lesion classes. The ensemble model was compiled with the Adam optimizer, using categorical cross-entropy as the loss function and accuracy as the primary performance metric.

*3) **Model Evaluation**:* The performance of the ensemble model was rigorously evaluated on validation and test datasets, focusing on metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation aimed to assess the effectiveness of the ensemble approach in skin lesion classification.

D. Data Augmentation

Data augmentation techniques including rotation, shifting, shearing, zooming, and flipping were applied to all models. This not only combated overfitting but also ensured that the models learned to recognize lesions in varied orientations and scales.

V. Experiments

Our experiments were meticulously designed to evaluate the effectiveness of Deep Convolutional Neural Networks (DCNNs) in classifying skin lesions, with a focus on understanding the impact of architectural nuances, fine-tuning methodologies, and various hyperparameters.

A. Model Performance Comparison

In our primary set of experiments, we assessed the performance of pre-trained DCNN architectures — Inception V3, DenseNet201, MobileNetV2, and ResNet50 — on the HAM10000 dataset. Each model was subjected to a structured two-stage fine-tuning process: initial fine-tuning of specific layers, followed by an extensive fine-tuning of the entire network. The efficacy of these models was quantified through metrics such as accuracy, sensitivity, and specificity, offering a comprehensive view of their diagnostic capabilities.

B. Ablation Study

To delineate the significance of each architectural element, we conducted an ablation study. This involved selectively removing or modifying components (like specific layers or distinct training techniques) within each DCNN model. The objective was to pinpoint the elements that were pivotal to the models' accuracy and overall performance, thereby elucidating the most impactful strategies in our architecture design.

C. Effect of Hyperparameters on Model

*1) **Addressing Class Imbalance**:* The dataset presented a notable challenge of class imbalance. Initially, class weights were employed to mitigate this imbalance. However, this approach adversely affected the model's accuracy and loss metrics. To overcome this, we adopted the Synthetic Minority Over-sampling Technique (SMOTE) for oversampling the underrepresented classes. Furthermore, we enhanced data augmentation with brightness adjustments and rescaling, introducing more variability into the training data. These modifications led to a significant improvement in the model's performance metrics and effectively reduced overfitting.

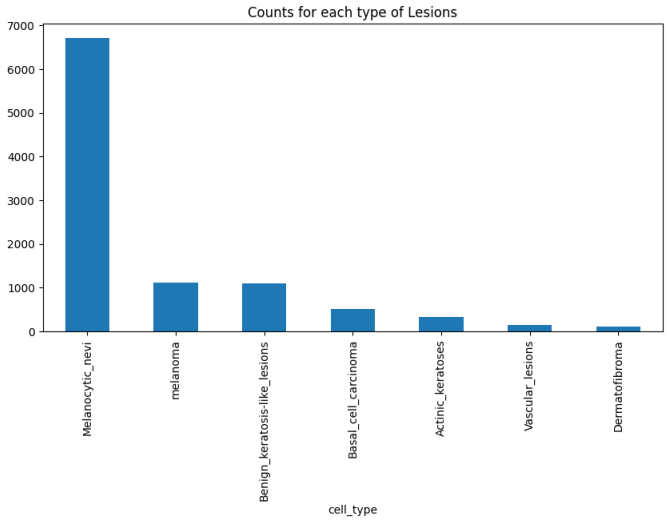

Fig. 2. Class Imbalance

*2) **Fine-Tuning Model Layers**:* Our experiments with fine-tuning revealed that training the entire model, as opposed to only the top layers, yielded superior results. This comprehensive fine-tuning approach allowed for more nuanced

adjustments to the models, aligning them more closely with the specific characteristics of our dataset.

*3) **Learning Rate Adjustments**:* We observed that even minor variations in the learning rate resulted in substantial changes in loss metrics. This sensitivity highlighted the importance of carefully selecting and adjusting the learning rate to ensure optimal convergence of the models.

*4) **Impact of L2 Regularization**:* While exploring regularization techniques, the use of L2 regularization was found to reduce the model's accuracy. Despite its effectiveness in controlling overfitting, it appeared to constrain the model's ability to learn complex patterns in the data.

*5) **Data Augmentation and Overfitting**:* To counteract the overfitting observed across all models, we augmented the training data, incorporating techniques such as rotations, shifts, and flips. This strategy expanded the diversity of the training dataset, helping the models generalize better to new, unseen data.

*6) **Implementation of Dropout**:* Dropout layers were introduced as a regularization strategy to further mitigate overfitting. By randomly disabling a fraction of neurons during training, dropout prevented the models from becoming overly reliant on specific features, leading to a more generalized and optimal solution.

D. Comparison with Published Methods of InceptionV3

Our findings were juxtaposed with those from analogous studies in the domain, particularly those employing DCNNs for skin lesion analysis. This comparative analysis served as a benchmark, providing a contextual backdrop to gauge the progress and efficacy of our proposed methodologies.

Fig. 3. Comparison of Model Performance

- The first plot corresponds to a model trained on an imbalanced dataset without brightness adjustments or rescaling augmentations. Additionally, L2 regularization was applied to this model. The outcome was overfitting, characterized by a significant gap between the training and validation performance.

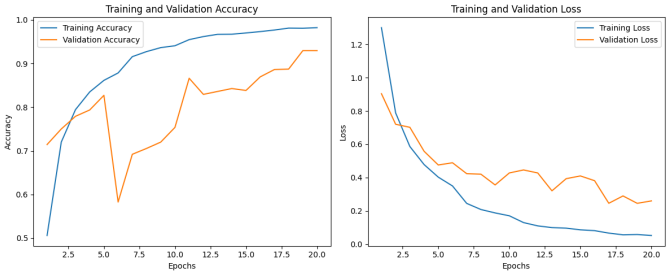

Fig. 4. Training Loss and Accuracy after SMOTE

- On the other hand, the second plot represents a model trained after applying SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance. Unlike the first model, L2 regularization was not used, and brightness and other augmentations were applied during training. This resulted in a different training trajectory, with improvements in model generalization and reduced overfitting tendencies.

E. Common Failure Modes

Analysis of the MobileNet model's training and validation suggests overfitting, as evidenced by higher training accuracy alongside lower validation accuracy. The model's performance on the training set improved with each epoch, indicating effective learning. However, the validation accuracy demonstrated significant fluctuations and failed to converge with the training accuracy, implying the model may not generalize well to unseen data. Additionally, the validation loss experienced increases at certain points, reinforcing concerns about the model's ability to perform consistently on new data. These indicators highlight the necessity for refining the model to enhance its generalizability.

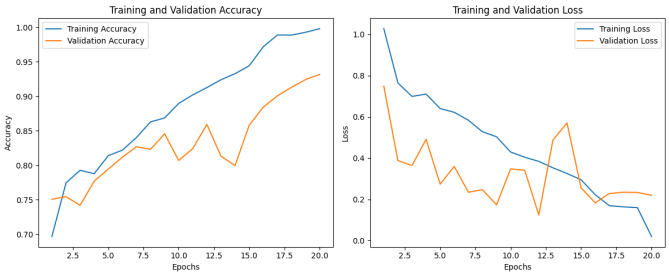

Fig. 5. Performance over 20 epochs showing increasing training accuracy and fluctuating validation accuracy and loss, indicative of overfitting.

VI. Mobile Application 'MyDerma'

The architecture of the MyDerma application is depicted in Fig. 6, which outlines the integration of a user-interactive front-end with a backend deep learning pipeline. Upon user initiation, the front-end facilitates image selection for

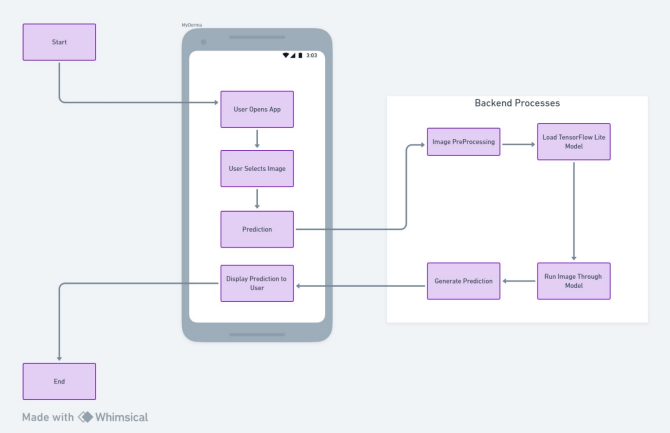

Fig. 6. Architectural workflow of the MyDerma mobile application

subsequent analysis. The image is then processed to conform to the neural network's input specifications, including resizing and normalization as part of the preprocessing routine.

The backend is powered by TensorFlow Lite, a model selected for its efficiency on mobile devices (TensorFlow Team, 2021). This DCNN model conducts feature extraction and classification directly on the device, leveraging optimized computational pathways for expedited inference.

The output from the DCNN is post-processed into a user-intuitive format and displayed within the application's interface. This end-to-end workflow, from image selection to diagnostic prediction, is engineered to exploit the device's computational prowess, ensuring prompt and accurate skin lesion classification within the mobile application context.

A. System Architecture of Mobile Application

1) ***Front-end User Experience****:* The MyDerma application's front-end is meticulously crafted to prioritize user engagement and accessibility. Developed using Android Studio 2023 Hedgehog, the premier Integrated Development Environment for Android, this interface ensures users can effortlessly initiate the diagnostic process through a streamlined image selection process. The chosen image is promptly subjected to pre-processing operations, including resizing to a uniform scale and normalizing pixel values, thereby standardizing inputs for the subsequent analytical phase. These initial steps are critical in transforming raw image data into a form amenable to systematic analysis.

2) ***Back-end Computational Workflow****:* The back-end of the MyDerma application is powered by TensorFlow Lite, chosen for its ability to perform efficient on-device machine learning inference. This framework supports the application's core functionality, eliminating the delay typically introduced by cloud-based processing and emphasizing the swift delivery of results. Through the utilization of Android Studio 2023 Hedgehog, the integration of TensorFlow Lite is streamlined, ensuring that the application is constructed using the latest and most effective development tools for Android, thus enhancing the app's performance and responsiveness.

3) ***Diagnostic Synthesis and Display****:* The diagnostic output is the culmination of the application's processing pipeline. Post-analysis, the DCNN's output undergoes a transformation into an intelligible format, subsequently displayed within the application's interface—a process made seamless by the utilization of Flutter 3.19.0. Flutter's robust cross-platform capabilities enable a cohesive and dynamic presentation of results, enhancing the user's interpretability of the diagnostic information. The choice of Flutter as the development framework allows for a flexible and responsive design, ensuring that the diagnostic results are presented in a clear, concise, and visually appealing manner.

VII. Experimental Results

A. Results and Analysis

In our study, we evaluated the performance of various machine learning models on a classification task.
Table I summarizes the key performance metrics:

- The 'Retrained ResNet50' model achieved the highest test accuracy at 96.84%, demonstrating excellent classification performance.
- The 'Ensemble Models' approach yielded an impressive test accuracy of 97.15% and a validation accuracy of 98.46%, showcasing the effectiveness of model ensembling techniques.
- Models like 'Retrained DenseNet201' and 'Fine-tuned DenseNet201' achieved high validation accuracies of 92.03% and 87.80%, respectively, with relatively lower test losses.

TABLE I
Comparison of Model Performance Metrics

Model	Test Accuracy	Validation Accuracy	Test Loss	# Params
Baseline Model	76.64%	76.98%	0.63	2,124,839
Fine-tuned Inception V3	79.84%	78.04%	0.67	22,855,463
Fine-tuned DenseNet201	84.93%	87.80%	0.66	19,309,127
Retrained Inception V3	90.92%	91.36%	0.27	22,855,463
Retrained DenseNet201	93.72%	92.03%	0.12	19,309,127
Retrained ResNet50	96.84%	97.51%	0.27	27,798,407
Retrained MobileNet V2	89.66%	90.51%	0.24	3,576,903
Ensemble Models	97.15%	98.46%	0.08	27,798,407

These results provide insights into the comparative performance of different models, helping guide the selection of the most suitable model for the task at hand.

	precision	recall	f1-score	support
AKIEC	0.97	0.97	0.97	146
BCC	0.98	0.98	0.98	46
BKL	0.99	0.99	0.99	1081
DF	1.00	1.00	1.00	20
NV	0.94	0.98	0.96	123
MEL	1.00	0.97	0.98	66
VASC	1.00	1.00	1.00	15

Fig. 7. Classification Report with precision, recall, and f1-scores, showing high accuracy

*1) **Classification Report Findings**:* The classification report revealed near-perfect precision and recall for most classes, with classes DF and VASC achieving flawless metrics. The NV class exhibited slightly lower precision, indicating a propensity for false positives. Despite class imbalances, the model's performance across classes was robust, as reflected by the f1-scores.

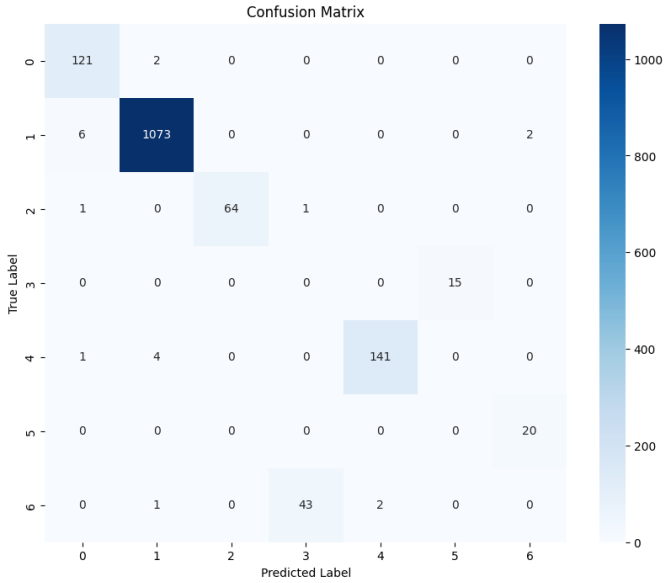

Fig. 8. Confusion Matrix highlighting the model's accurate predictions

*2) **Confusion Matrix Findings**:* The confusion matrix demonstrated high predictive accuracy for the classification model, with classes 1 and 4 showing particularly strong performance. Misclassifications were primarily between classes 5 and 6, suggesting a targeted area for model refinement.

*3) **Mobile Application**:* In practical applications, the deployed model has demonstrated commendable performance when exposed to new, real-world data and uploaded data from other datasets. This finding is significant as it indicates the model's robustness and ability to generalize beyond the controlled conditions of the initial training environment.

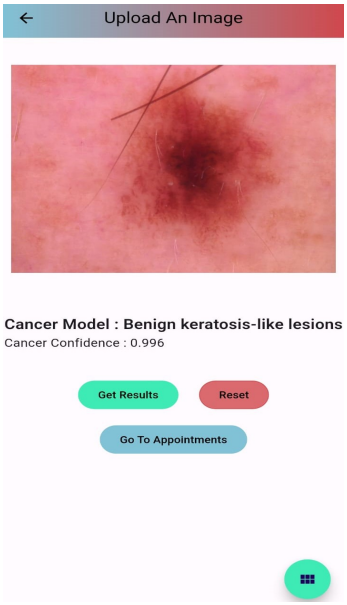

Fig. 9. Result obtained from Mobile Application

VIII. Limitations

Our study presents notable progress in classifying skin lesions with DCNNs, yet it's important to recognize its inherent limitations for accurate result interpretation and future research guidance.

A. Dependency on Data Quality and Diversity

Our models' performance is heavily reliant on the quality and diversity of the dataset. The HAM10000 dataset, while comprehensive, may not encompass all variations of skin lesions, especially rare types. This limitation could impact the models' ability to generalize to a broader range of real-world dermatological cases.

B. Mobile Application Deployment:

While the deployment of our classification model in a mobile application context had yielded a high accuracy rate, there were instances of incorrect classifications. These misclassifications were significant in practical scenarios where diagnostic precision is critical and could have been influenced by factors such as image quality variations from mobile device cameras and lighting conditions, as well as the limited computational capabilities of mobile platforms.

C. Training Epochs:

Our model's training was constrained to 20-25 epochs due to computational resource limitations. A higher number of epochs might have allowed the model to converge more effectively, potentially enhancing its generalization on unseen data. The limited training might have curtailed the model from reaching its optimal performance.

D. Dataset Balancing Using SMOTE:

To mitigate class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was utilized, which expanded the training dataset size to 20GB. This increase in size introduced a computational overhead and could have impacted the training duration, rendering the model less scalable and more difficult to implement in environments with restricted storage or processing capabilities.

E. Performance in Clinical Settings

While our models achieved high accuracy in a controlled experimental setup, their performance in real-world clinical settings may vary. Factors such as different imaging equipment, lighting conditions, and patient demographics can affect the models' effectiveness.

F. Computational Requirements

The training and fine-tuning of deep learning models require significant computational resources, which may not be readily available in all research or clinical settings. This aspect can limit the widespread adoption and practical application of our approach.

IX. Conclusion

In conclusion, this project showcases the successful implementation and evaluation of deep learning models for skin lesion classification, leveraging the strengths of an ensemble approach to achieve high accuracy and effectiveness. This approach has significantly enhanced the precision of dermatological diagnoses and has paved the way for integrating these technologies into mobile platforms, increasing accessibility for early detection of skin cancer.

The ensemble models achieved an impressive test accuracy of 97.15% and a validation accuracy of 98.46%, highlighting the effectiveness of model ensembling techniques. Individual models like 'Retrained DenseNet201' and 'Fine-tuned DenseNet201' also performed exceptionally well, with high validation accuracies of 92.03% and 87.80%, respectively, and relatively lower test losses.

However, it's important to acknowledge the limitations of our current approach. The reliance on high-quality image data and the challenges in generalizing the model to diverse skin types are areas that require further research.

X. Future Work

1) **Model Generalization:** Improving the model's ability to generalize across diverse skin types and conditions, ensuring accuracy and reducing demographic biases.
2) **Data Quality and Diversity:** Expanding the dataset to include a broader spectrum of skin lesions, particularly from underrepresented groups.
3) **Integration of Patient Data:** Incorporating patient history and demographic information for more personalized diagnostics.
4) **User Interface Development:** Enhancing the mobile application interface for ease of use by both healthcare professionals and individuals.
5) **Clinical Trials:** Conducting extensive clinical trials to validate the system's reliability and efficacy in real-world medical environments.
6) **Interdisciplinary Collaboration:** Engaging with dermatologists and healthcare experts to ensure the practicality and clinical relevance of the technology.

Acknowledgment

We extend our heartfelt gratitude to all those who have contributed to the success of this project. This project wouldn't have been possible without the constant support and knowledge imparted by the course instructor of ENPM809K - Fundamentals for AI and Deep Learning Framework, Dr. George Zaki and the teaching assistant, Mr. Davit Soselia. Lastly, we are grateful to our colleagues and peers for their insightful feedback and encouragement throughout the duration of this project.

References

[1] T. Emara, H. M. Afify, F. H. Ismail and A. E. Hassanien, "A Modified Inception-v4 for Imbalanced Skin Cancer Classification Dataset," 2019 14th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2019, pp. 28-33, doi: 10.1109/ICCES48960.2019.9068110.

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

[3] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

[4] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. *CoRR, abs/1409.4842*.

[6] Huang, G., Liu, Z., Weinberger, K. Q., & Maaten, L. (2017). Densely connected convolutional networks. In *CVPR*.

[7] Kieffer, B., Babaie, M., Kalra, S., & Tizhoosh, H. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. *arXiv preprint arXiv:1710.05726*.

[8] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S. (2017).Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[9] A. Chemakura, D. L. Gannavarapu, V. S. S. Kasturi, V. Mandala https://drive.google.com/drive/folders/1_gAadz9rnbw40V_pu3BTeTQ045pHIrm7?usp=sharing, 2023

[10] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... Tschandl, P. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836-1842.

[11] Menegola, A., Tavares, J., Fornaciali, M., Li, L. M., Seixas, J. M. (2017). Deep learning for dermoscopic melanoma detection: A retrospective study. PloS One, 12(12), e0189505.

[12] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. Soroushmehr, M. Jafari, K. Ward, and K. Najarian, "Melanoma detection by analysis of clinical images using convolutional neural network," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016.

[13] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, "Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble," 2017.