Project Report

Performance Evaluation of Machine Learning and Convolutional Neural Network Based Skin Lesion Classification Techniques Using Dermoscopic Images

ECE 740 A02: Advanced Topics in Signal and Image Processing

Term: Fall 2024

Submitted by **Md Touhidul Haque**

Contents

1 Introduction

Skin diseases, including various types of skin cancer, pose a significant public health challenge, affecting millions globally and impacting both physical health and quality of life. Among these, melanoma is particularly fatal if not detected and treated early. According to the World Health Organization, skin cancer accounts for one-third of all diagnosed cancers worldwide [1], [2]. In the United States alone, approximately 5.4 million new skin cancer cases are reported annually, with melanomas contributing to over 10,000 deaths each year [3].

Traditional diagnostic methods, such as physical examination and biopsy, are time-consuming, invasive, and often subject to variability in interpretation [4]. While dermoscopy offers a non-invasive method to enhance diagnostic accuracy through high-resolution visualization of deeper skin structures [5], its effectiveness depends heavily on the expertise of the dermatologist. Studies show that diagnostic accuracy ranges from 62% to 80%, with lower accuracy observed among less experienced dermatologists [6], [7]. Moreover, the requirement for extensive training and limited exposure to the full diversity of skin cancer cases further complicates accurate diagnosis.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have shown significant promise in addressing these challenges by providing automated, scalable, and reliable diagnostic tools. Deep convolutional neural networks (CNNs), in particular, have demonstrated exceptional performance in image classification tasks, including skin cancer classification, by detecting complex patterns and learning domain-specific features from dermoscopic images [8]–[11]. A key advantage of CNNs is their ability to automatically learn "deep features" from the data, eliminating the need for manual "feature engineering" by machine learning experts [12]. However, they require large datasets for effective training and optimal performance.

This project aims to develop and evaluate classical ML and deep learning-based techniques for multi-class skin lesion classification. Specifically, using the HAM10000 dataset [13], we replicate the work presented in the journal paper [14], where we conduct a comparative analysis of three classical ML models namely Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) with a customized CNN-based deep learning technique. Throughout the report, we address this journal paper as the reference paper.

The primary objectives of this course project are multifaceted. Specifically, this project aims to:

- Design, train, and evaluate a customized CNN-based multi-class skin lesion classification technique as outlined in the reference paper.
- Design, train, and test three classical ML models (DT, RF, and SVM) as baseline approaches for the multi-class skin lesion classification task.
- Perform a comparative analysis of the performance between the CNN-based technique and the ML-based models.
- Extract key insights into the strengths and limitations of ML-based and deep learning-based approaches for skin lesion classification.
- Critically review the reference paper, highlighting its strong and weak aspects.

The remainder of this report is organized as follows: Section 2 provides a detailed description of the dataset. Section 3 elaborates on the methods employed in this project, including preprocessing, feature extraction, classification models, fine-tuning, and performance metrics. Sections 4 and 5 present the results and discussions, respectively. Finally, Section 6 concludes the report, and Section 7 includes a link to the GitHub repository hosting the project.

2 Dataset

The HAM10000 (Human Against Machine with 10000 training images) dataset [13] is a comprehensive collection of 10015 dermoscopic images of pigmented skin lesions and is publicly available. It has been curated from multiple sources, representing diverse populations and acquisition modalities, to

serve as a benchmark training set for academic machine learning research.
Each image is stored in JPEG format with a resolution of 600×450 pixels. The images are manually cropped and centered around the lesion, with adjustments for contrast and color reproduction. Metadata accompanying the images includes patient age, sex, lesion ID (unique identifier for each lesion), image ID, diagnostic type (dx type), anatomical location of the lesion, and diagnostic category.
The dataset is classified into seven diagnostic categories:

i. *Actinic Keratoses and Intraepithelial Carcinoma (akiec)*: Noninvasive squamous cell carcinoma (327 images).

ii. *Basal Cell Carcinoma (bcc)*: A type of epithelial skin cancer with low metastatic potential (514 images).

iii. *Benign Keratosis-like Lesions (bkl)*: Includes seborrheic keratoses, lichen-planus-like keratoses, and solar lentigo (1099 images).

iv. *Dermatofibroma (df)*: Benign growths or inflammatory responses to minor trauma (115 images).

v. *Melanoma (mel)*: Malignant melanocytic tumors treatable by early surgical intervention (1113 images).

vi. *Melanocytic Nevi (nv)*: Benign melanocytic neoplasms appearing in various morphologies (6705 images).

vii. *Vascular Lesions (vasc)*: Includes cherry angiomas, angiokeratomas, and pyogenic granulomas (142 images).

A sample image of skin lesions from each class is shown in the Fig. 1. This dataset is highly imbalanced, with lesion class *df* containing only 115 images compared to 6705 images in class *nv*. To address this imbalance, data augmentation techniques are applied, and two distinct approaches are used to create two balanced dataset for training the ML-based and CNN-based models:

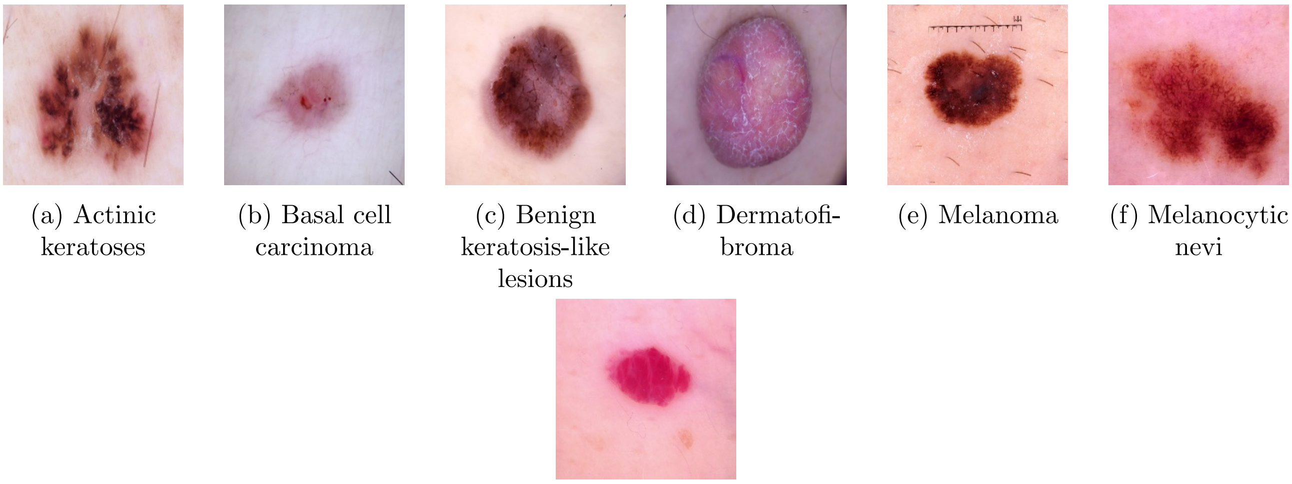
- *Balanced dataset-A for ML-based techniques:* A balanced dataset is created by randomly selecting 100 images per class (total 700 images). Data augmentation technique (through horizontal flip) is then applied to generate 200 images per class, resulting in a dataset containing 1400 images.

- *Balanced dataset-B for CNN-based technique:* Deep learning models such as CNN need large dataset for training the model. For this reason, a larger balanced dataset are created with 2000 images per class (except for class *nv*) using data augmentation technique (thorugh horizontal flips, random rotations, shearing transformations, contrast, saturation, and hue adjustments, and addition of Gaussian noise), resulting in a dataset containing 14,000 images.

Each of the above balanced dataset is split into an 80-20 ratio for training and testing purposes following the reference paper. However, due to performing data augmentation before data splitting, the training dataset may contain images whose augmented versions may be included in the testing dataset which may result in overly optimistic performance of the ML and CNN models. The HAM10000 dataset also contains an independent testing set (ISIC2018_Task3_Test_Images) of 1512 lesion images which are used in this course project to get the more realistic performance of the ML and CNN models.

3 Methodology

The proposed work is primarily based on the methodology presented in the reference paper, where a convolutional neural network (CNN)-based skin lesion classification technique has been developed and a performance comparison is made with several well-known machine learning (ML) models. The methodological steps of the proposed project are shown in Fig. 2. All models is implemented in Python using *Keras*, *Tensorflow*, *OpenCV*, *Imutils*, and *cv2Numpy* libraries. The key steps of the proposed methodology are outlined below:



(a) Actinic keratoses

(b) Basal cell carcinoma

(c) Benign keratosis-like lesions

(d) Dermatofibroma

(e) Melanoma

(f) Melanocytic nevi

(g) Vascular lesions

Figure 1: Sample images for the seven skin lesion categories from the HAM10000 dataset.

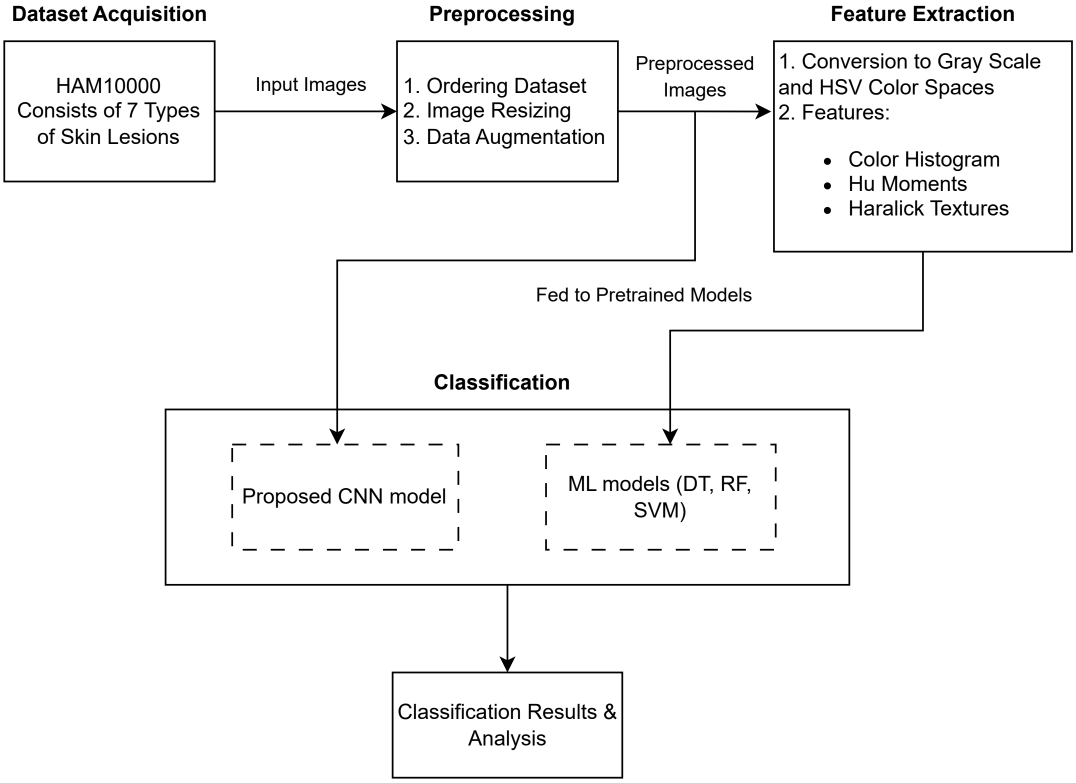


Figure 2: Overall methodology of the proposed skin lesion classification technique

3.1 Image Preprocessing

The image preprocessing module takes the HAM1000 dataset as input and conducts the following tasks-

- *Dataset Organization:* Lesion images are systematically categorized based on unique identifiers (lesion ID and diagnostic labels), ensuring clear organization for analysis.
- *Image Resizing:* Lesion images are resized to a resolution of 220×220 pixels for ML models and 96×96 pixels for CNN model, reducing memory requirements and computational overhead.
- *Pixel Normalization:* Pixel values are normalized to a range of $0 - 1$, and class labels were one-hot encoded to facilitate multi-class classification.

- *Data Augmentation:* The HAM10000 dataset is highly imbalanced. To address this challenge, data augmentation techniques are applied, including horizontal flips, random rotations, shearing transformations, contrast, saturation, and hue adjustments, and addition of Gaussian noise.

The preprocessing module outputs are passed to the feature extraction module to extract handcrafted features for ML-based models. Deep learning models such as CNN can automatically extract "deep features" from the images. Therefore, outputs of the preprocessing module are directly fed to the CNN model.

3.2 Feature Extraction

The feature extraction module is only used for ML-based models. Images are first converted to *HSV* color space and gray-scale format to extract the following handcrafted global features:

- *Color:* Quantified using a color histogram in the HSV color space, with 32 bins per channel, resulting in a feature size of $F_{\text{color}} = 32 \times 32 \times 32 = 32,768$.
- *Shape:* Extracted using 7 invariant Hu moments from the gray-scale image, with a feature size of $F_{\text{shape}} = 7$.
- *Texture:* Captured using Haralick texture features comprising 6 texture properties (contrast, dissimilarity, homogeneity, energy, correlation, ASM), averaged across 4 directions, yielding a feature size of $F_{\text{texture}} = 6$.

These handcrafted features are concatenated into a single feature vector:

$$F = \{F_{\text{color}}, F_{\text{shape}}, F_{\text{texture}}\},$$

with a total size of $32,781$. These features are then fed to the ML-based models.

3.3 ML-based Models

Three well-known classic ML models namely Decision Tree (DT), Random Forest (RF), and SVM are employed to classify the multi-class skin lesion images and compare their performances with the proposed CNN-based technique. Though seven ML models are employed in the reference paper, we chose top three performing models to keep the analysis concise and get better insights. Their hyperparameters used in the modelling are shown in the table 1.

Classifiers	**Hyperparameter**
Decision Tree (DT)	`---`
Random Forest (RF)	`n_estimators = 200, random_state = 0`
Support Vector Machine (SVM)	`Kernel = 'linear', c = 1, random_state = 0`

Table 1: Hyperparameter values for different machine learning models.

3.4 Proposed CNN-based Model

The proposed CNN architecture is shown in the Fig. 3. It is customized for classification of the described types of skin lesion images shown and consists of the following key components:

(i) *Input Layer:* Takes images of size 96 × 96 × 3 (RGB format).

(ii) *Convolutional Layers:* Three sets of convolutional layers with increasing filter counts: 32, 64, and 128 filters, each using 3×3 kernels. Each convolutional layer is followed by Batch Normalization to stabilize and accelerate training.

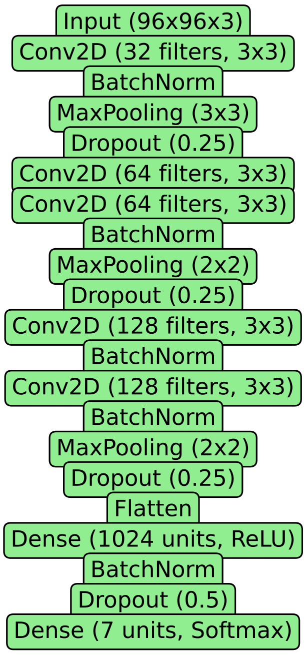


Figure 3: Architecture of the proposed CNN model

(iii) *Pooling Layers:*

- Max Pooling layers reduce spatial dimensions.
- 3×3 pooling after the first convolutional block.
- 2×2 pooling after the second and third convolutional blocks.

(iv) *Dropout Layers:* Dropout regularization is applied with rates of 0.25 after the convolutional blocks and 0.5 before the final classification layer to prevent over fitting.

(v) *L2 Regularization:* L2 regularization is added to the convolutional layers and Dense layers to further mitigate over fitting by penalizing large weights.

(vi) *Early Stopping:* Early stopping is implemented to halt training when the validation loss stops improving, preventing unnecessary epochs and over fitting.

(vii) *Flatten Layer:* Converts the feature map into a one-dimensional vector.

(viii) *Fully Connected Layers:*

- A Dense layer with 1024 units using ReLU activation and L2 regularization.
- A final Dense layer with 7 units and softmax activation for classification into 7 skin lesion classes.

This architecture combines feature extraction through convolutional and pooling layers with dense layers for classification, ensuring both spatial feature learning and robust generalization. Though not mentioned in the reference paper, we adopt $L2$ regularization and early stopping in the model which would enhance the model's ability to generalize and prevents over fitting. The CNN model is optimized using the Adam optimizer and categorical cross-entropy loss function. The k-fold ($k = 10$) cross-validation is used in the CNN model to reduce overfitting, and maximize the utilization of data by training and validating the model on multiple data splits. Training is conducted over maximum of 150 epochs with a batch size of 32, and the learning rate is dynamically adjusted between 0.001 and 0.00001. Early stopping is employed to monitor validation loss and terminate training early if no further improvement is observed. The hyperparameters used in the CNN model is given in the table 2.

Hyperparameter	Value
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Epochs	150
Batch Size	32
Learning Rate	0.001–0.00001 (dynamically adjusted)
k-Fold Cross Validation	k = 10
L2 Regularization	Weight Decay Factor (e.g., 0.001)
Early Stopping	Patience = 5 epochs, monitored on validation loss

Table 2: Hyperparameters used for the CNN model.

The key differences between this customized CNN model and popular image classification CNN-based models (e.g., AlexNet [15], VGG [16], ResNet [17], and Inception [18]) lie in the architecture's design, customization, and focus on simplicity and application-specific requirements. Those key differences are presented in the table 3.

	Proposed CNN Model	**Popular CNN Models (e.g., AlexNet, VGG, ResNet, Inception)**
Input Size	$96 \times 96 \times 3$ (small)	$224 \times 224 \times 3$ (large)
Architecture Depth	Shallow (fewer layers, 3 convolutional blocks)	Deep (more layers, can have 100+ layers)
Feature Learning	Automatic feature learning from the entire image	Automatic hierarchical feature learning
Number of Parameters	Lightweight (fewer parameters)	Heavy (millions of parameters)
Target Dataset	HAM10000 (small, domain-specific dataset)	Large-scale datasets (e.g., ImageNet)
Regularization	Dropout, L2 regularization, early stopping	Dropout, batch normalization, skip connections (e.g., ResNet)
Computational Requirements	Low (suitable for limited resources)	High (requires GPUs/TPUs)
Output Classes	7 classes (skin lesions)	Up to 1000 classes (general-purpose classification)

Table 3: Comparison of Custom CNN Model and Popular Image Classification CNNs.

3.5 Evaluation Metrics

Four well-known metrics: *Accuracy*, *Precision*, *Recall*, and *F1-score*, are used to evaluate the performance of different ML and CNN-based skin lesion classification techniques.

4 Result and Analysis

We evaluate the performance of ML models and the proposed CNN model using two distinct testing datasets. The first test dataset is taken from the balanced dataset-A and -B (by 80-20 split) generated for ML and CNN models respectively, as per the reference paper. The second is the independent testing dataset provided in the HAM10000 dataset.

4.1 Evaluation of the ML models

Tables 4 and 5 present the test accuracies along with the weighted average of precision, recall and F1-score of ML models for both testing datasets. Results show that accuracy drops significantly when models are tested on the independent dataset. This discrepancy likely arises from the dataset augmentation process, where the balanced dataset is augmented prior to splitting into training and testing sets. This can lead to potential data leakage, as the original image may end up in the training set while its transformed version is included in the testing set, or vice versa. Due to some degree of similarity of extracted features between an image and its transformed version, the ML models perform well on the balanced dataset but fail to generalize effectively to completely unseen data, as evidenced by their lower accuracy on the independent dataset.

We also observe that the highest performance is achieved by the RF model, followed by SVM, with the lowest performance observed for the DT model. This can be attributed to RF's ability to manage feature variability through ensemble learning, and its efficiency in handling multi-class problems. In contrast, SVM faces challenges with multi-class problems, and decision trees are more prone to overfitting, which adversely impacts their performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Decision Tree	0.7107	0.7137	0.7107	0.7099
Random Forest	0.9179	0.9255	0.9179	0.9187
SVM	0.7536	0.7624	0.7536	0.7550

Table 4: Performance metrics for the ML models using the testing set of the balanced dataset-A

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Decision Tree	0.3686	0.5777	0.3686	0.4287
Random Forest	0.5142	0.6803	0.5142	0.5620
SVM	0.3739	0.5953	0.3739	0.4386

Table 5: Performance metrics for the ML models using the independent testing dataset

4.2 Evaluation of the CNN model

The training and validation accuracy curves along with the corresponding loss curves of the CNN model, averaged across all folds, are shown in the Fig. 4 and 5, respectively. From the loss vs epoch curves, it is evident that both training and validation losses decrease rapidly during the initial epochs and gradually stabilize around epoch 60. Notably, the validation loss remains consistently higher than the training loss, indicating a potential gap between training and generalization performance. The accuracy vs epoch curves demonstrate a similar trend, with both training and validation accuracies increasing sharply during the early epochs and leveling off around epoch 60. The training accuracy approaches approximately 98%, while the validation accuracy stabilizes at around 86%. This suggests that the model achieves good generalization, with the testing accuracy expected to be close to the observed validation accuracy.

Table 8 presents the test accuracy as well as the weighted average of precision, recall and F1-score of the CNN model for the two testing datasets discussed earlier. The testing accuracy for the balanced dataset-B (86.34%) aligns closely with the accuracy reported in the reference paper for this dataset.

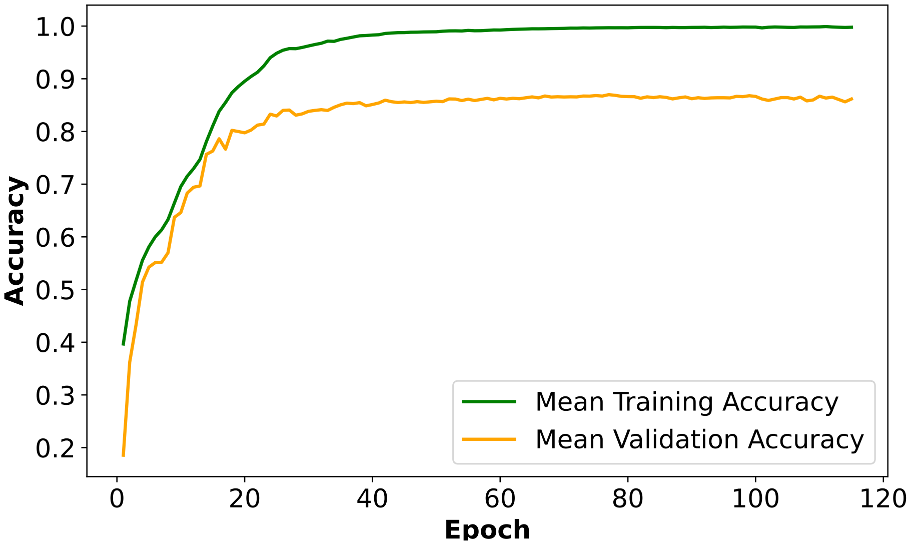


Figure 4: Plot of the accuracy vs epoch for training and validation of the CNN model

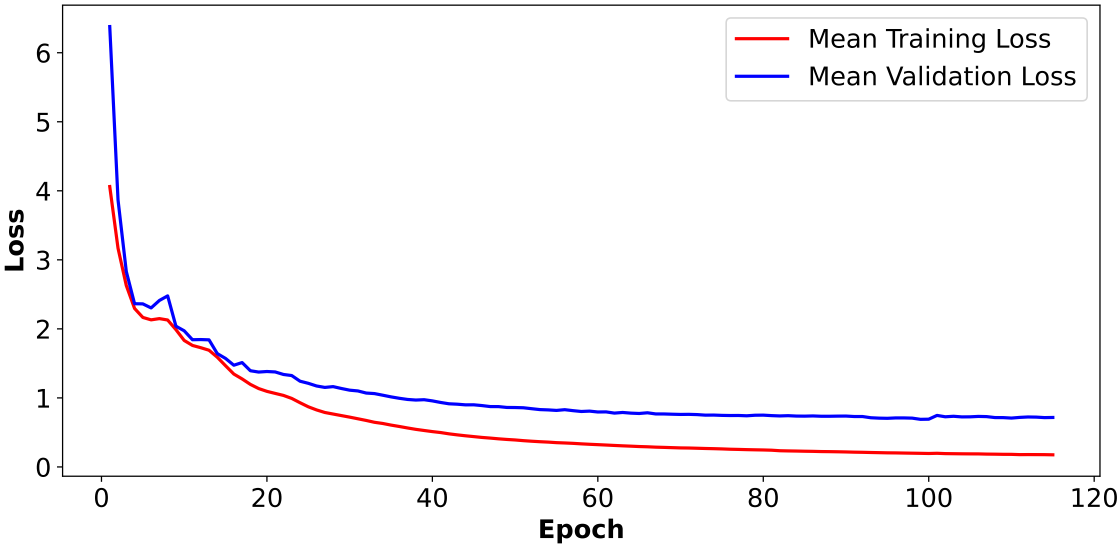


Figure 5: Plot of the loss vs epoch for training and validation of the CNN model

However, when tested on a completely new independent dataset, the accuracy drops significantly to 66.59%. This decline can likely be attributed to the same issue observed with the ML models: dataset augmentation performed prior to train-test splitting, leading to potential data leakage. As a result, the CNN model shows overly optimistic performance on the balanced dataset but struggles to generalize effectively to unseen data.

Individual classification performances for the two testing datasets are presented in Tables 7 and 9. The highest performance is observed for the *vasc* class in the balanced dataset-B and for the *nv* class in the independent testing dataset. Conversely, the lowest performance is observed for the *bkl* class in the balanced dataset-B and for the *mel* class in the independent testing dataset. The superior performance of the *nv* class in the independent testing dataset can be attributed to the substantial number of images available for this class, eliminating the need for augmentation. This abundance of data provided the model with sufficient variety, enabling it to learn generalized features and achieve high accuracy in detecting this class on unseen dataset. This observation suggests that if the CNN model were trained on sufficiently large datasets for other classes, its performance might improve for those classes as well. The lowest performance for the *bkl* and *mel* classes could be attributed to their high intra-class variability. Data augmentation might inadvertently bias the model towards specific intra-class variations, making it difficult for the model to learn generalized features for accurate classification.

Metric	Value
Accuracy	0.8634
Precision	0.8617
Recall	0.8634
F1 Score	0.8622

Table 6: Performance metrics for the CNN model across all folds using the testing set of the balanced dataset-B

Class	Precision	Recall	F1 Score	Support
akiec	0.8720	0.9025	0.8870	400.0
bcc	0.8894	0.9050	0.8971	400.0
bkl	0.7732	0.7500	0.7614	400.0
df	0.9540	0.9850	0.9692	400.0
mel	0.7995	0.7475	0.7726	400.0
nv	0.8379	0.8400	0.8390	400.0
vasc	0.9702	0.9775	0.9738	400.0
Macro Avg	0.8709	0.8725	0.8715	2800.0
Weighted Avg	0.8709	0.8725	0.8715	2800.0

Table 7: Performance metrics for individual skin lesion classes for the CNN model using the testing set of the balanced dataset-B (last fold)

Metric	Value
Accuracy	0.6659
Precision	0.7111
Recall	0.6659
F1 Score	0.6827

Table 8: Performance metrics for the CNN model across all folds using the independent testing dataset

Class	Precision	Recall	F1 Score	Support
akiec	0.2917	0.4884	0.3652	43.0
bcc	0.4286	0.4839	0.4545	93.0
bkl	0.4578	0.4747	0.4661	217.0
df	0.6857	0.5455	0.6076	44.0
mel	0.3414	0.4971	0.4048	171.0
nv	0.8852	0.7731	0.8254	908.0
vasc	0.6250	0.5714	0.5970	35.0
Macro Avg	0.5308	0.5477	0.5315	1511.0
Weighted Avg	0.7055	0.6618	0.6786	1511.0

Table 9: Performance metrics for individual skin lesion classes for the CNN model using the independent testing dataset (last fold)

ROC curves of the individual classes for CNN model (last fold) for the two testing datasets are depicted in the Fig. 6 and 7. The ROC curves for the testing set of the balanced dataset-B indicate excellent performance, with AUC values ranging from 0.96 to 1.00, suggesting the model is highly effective in distinguishing between classes. In contrast, the ROC curves for the independent testing dataset show a decline in performance compared to the previous case, with AUC values ranging from 0.83 to 0.98. Notably, the *bkl* and *mel* classes show lower AUC values, indicating potential issues with feature representation or class overlap. The highest AUC value is observed for the *vasc* class for both of the testing sets.

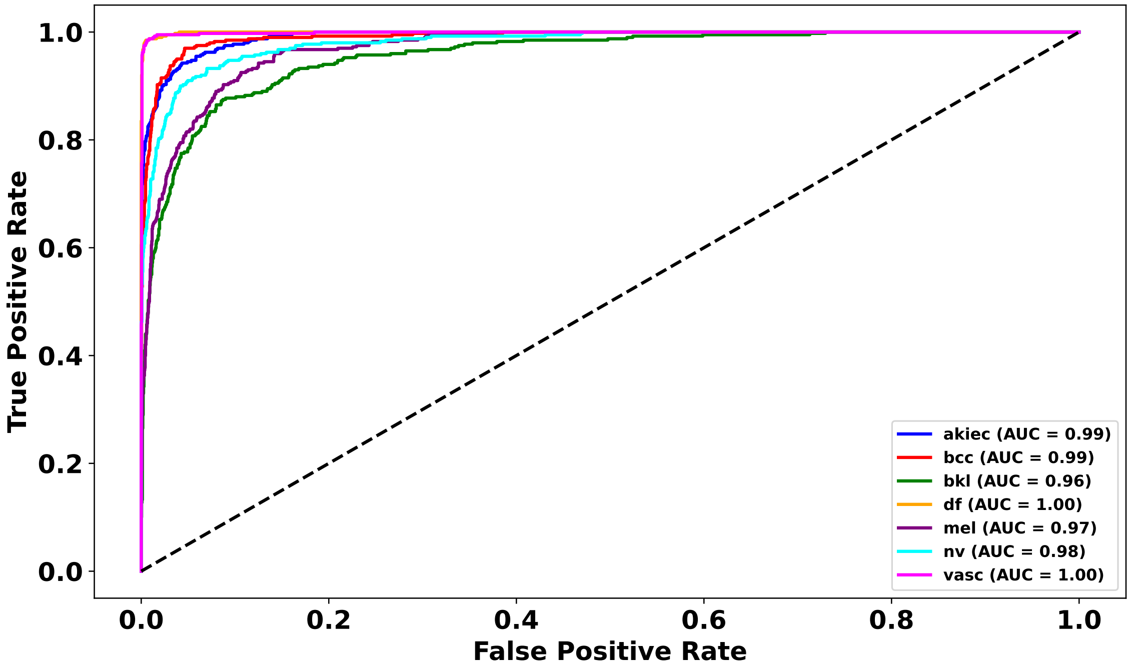


Figure 6: ROC curves for the individual classes using the testing set of the balanced dataset-B (last fold)

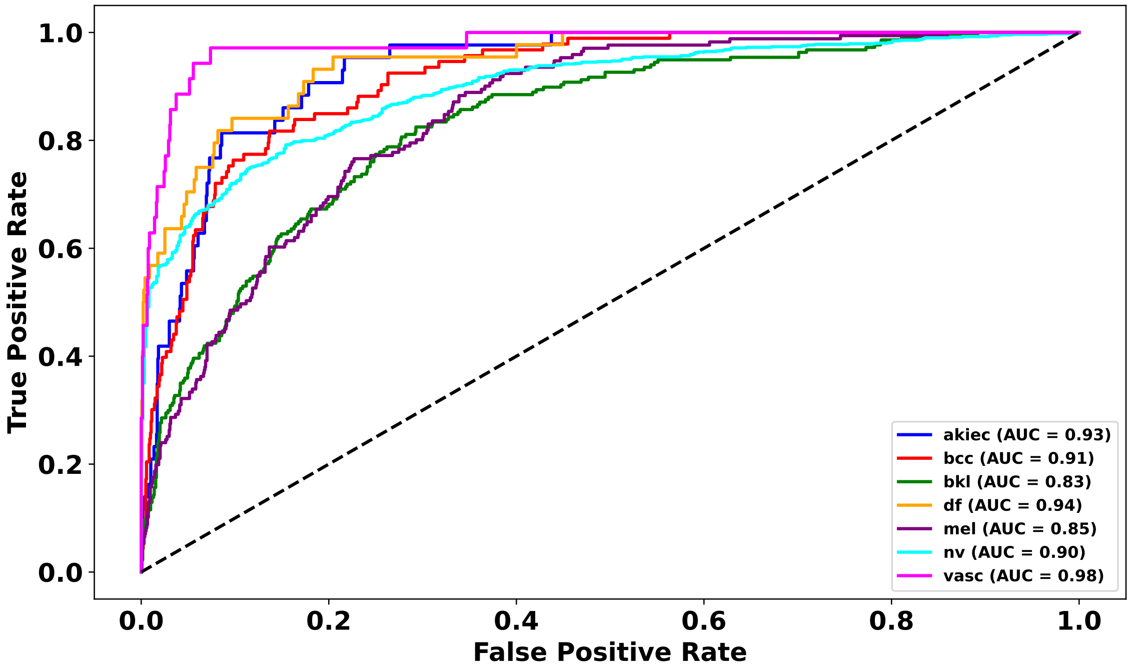


Figure 7: ROC curves for the individual classes using the independent testing dataset (last fold)

The confusion matrices for the CNN model (last fold) for the two testing datasets are depicted in the Fig. 8 and 9. The confusion matrix for testing from balanced dataset demonstrates a good classification accuracy, with the diagonal values nearing or exceeding 0.75 for all classes, indicating stronger model performance. Notable performance is observed for the *akiec* and *df* classes, achieving accuracies of 0.90 and 0.98, respectively, while off-diagonal values are significantly reduced, highlighting fewer misclassifications. However, moderate classification performance of the CNN model is observed in the confusion matrix for the independent testing dataset. In this case, the *nv* class shows the highest correct prediction rate (0.77), whereas other classes like *mel* and *bcc* exhibit more significant misclassification.

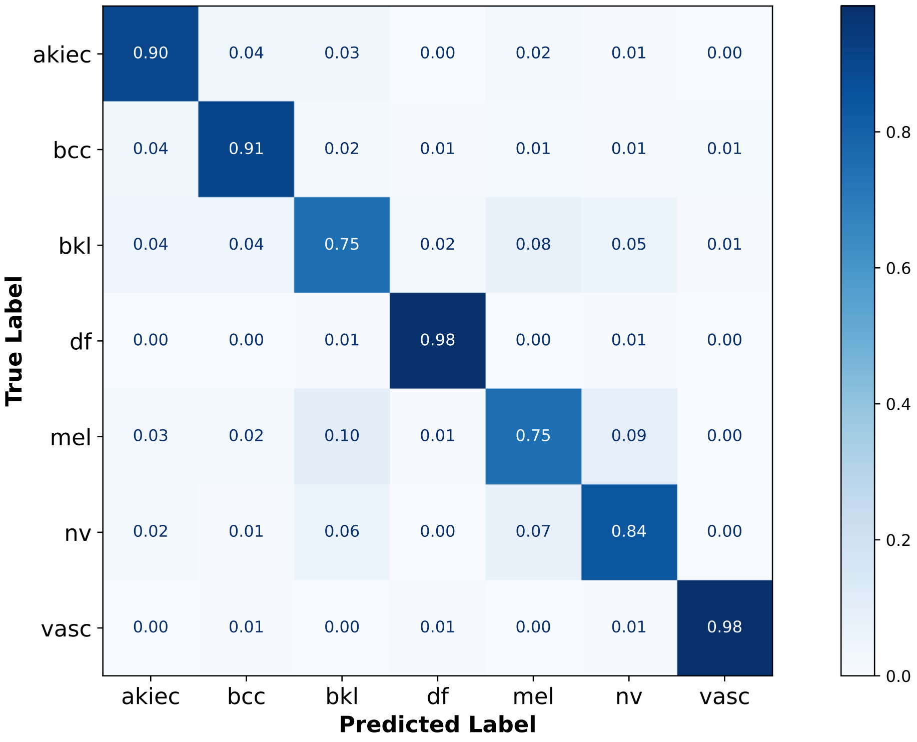


Figure 8: Confusion matrix using the testing set of the balanced dataset-B (last fold)

The performance comparison of the CNN and ML-based models (DT, RF, and SVM) in terms of Accuracy, Precision, Recall, and F1 Score for the two testing datasets is illustrated in Fig. 10 and Fig. 11. For the testing set derived from the balanced dataset, the RF model demonstrates superior performance compared to the proposed CNN model. However, the independent testing set highlights the clear advantage of the CNN model, outperforming all other ML models in terms of performance metrics. This observation suggests that the proposed CNN model effectively captures more generalized features, thereby indicating its potential to deliver better performance on unseen datasets.

5 Discussion

Upon reproducing the work presented in the reference paper, we derive the following key insights regarding multi-class image classification using ML and deep learning models such as CNNs:

- *Data Leakage during Preprocessing:* Data leakage can occur during preprocessing, particularly when performing data augmentation to balance the dataset or for other purposes. To prevent this, it is crucial to create a separate test dataset before applying data augmentation. Failure to do so may lead to data leakage, resulting in an overestimation of the model's performance.
- *Limitations of Using Training Accuracy:* The reference paper reports training accuracy (95.18%) as the primary performance metric, which can be misleading. Model performance should always be evaluated on new, unseen test datasets to ensure a reliable measure. As demonstrated in

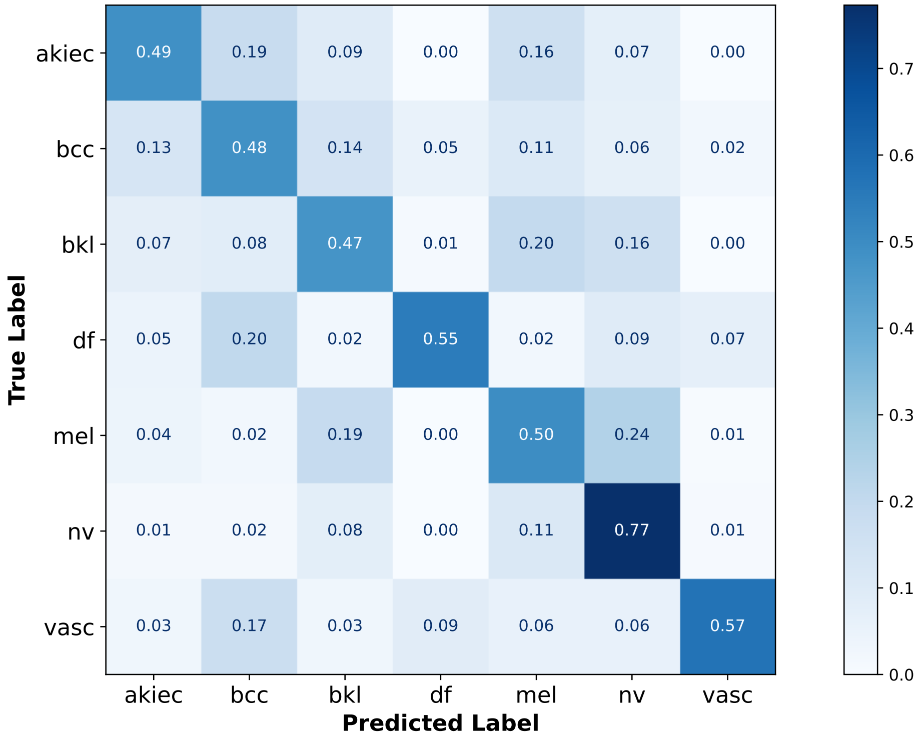


Figure 9: Confusion matrix using the independent testing dataset (last fold)

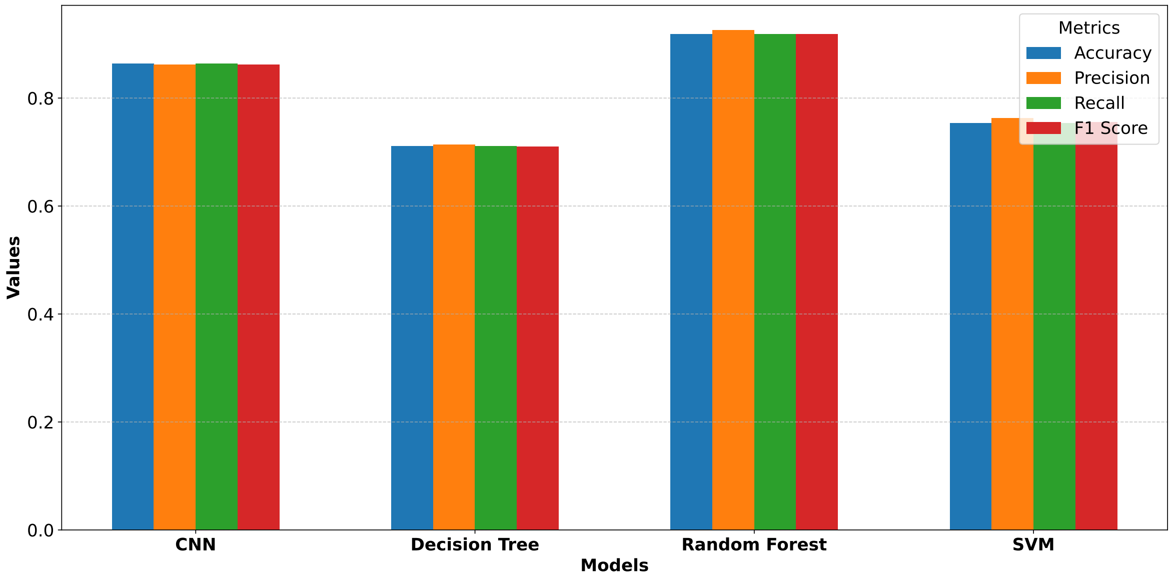


Figure 10: Performance comparison of the CNN and ML models using the testing set of the balanced dataset

our evaluation, the proposed model's performance on a new, independent testing dataset is significantly lower (66.50%), highlighting the importance of reliable testing.

- *Incorporating Regularization and Early Stopping:* The proposed CNN model in the reference paper does not include regularization techniques or early stopping. Based on our reproduction of this work, we recommend incorporating these methods to reduce overfitting, achieve better generalization, and improve training efficiency, especially when dealing with deep CNN architectures with many layers and high numbers of epochs.

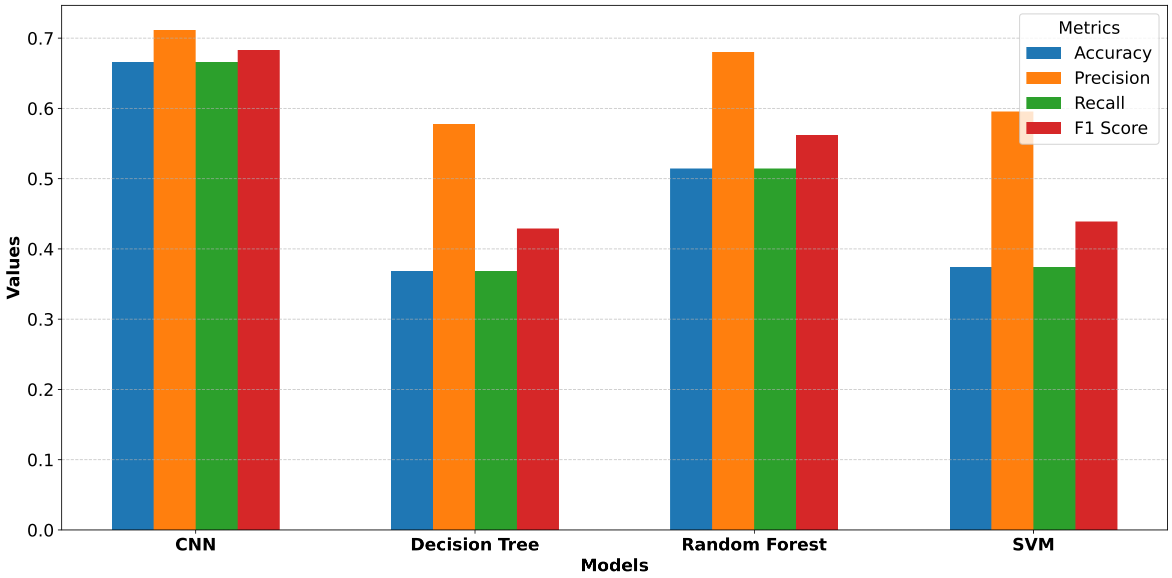


Figure 11: Performance comparison of the CNN and ML models the independent testing dataset

- *Performance Indices:* In case of weighted average of precision, recall, and F1 score for multi-class classification task, we would typically get accuracy and recall equal which have been seen in our results. However, in the reference paper, values of precision and recall are different for CNN model and most of the ML models. No explanation or justification for this unusual result is presented in the reference paper.

The customized CNN model proposed in the reference paper generally outperforms classical ML models. Its design requires minimal preprocessing and incorporates fewer layers compared to other popular CNN architectures, making it computationally efficient. However, our observations indicate that the model's performance on completely new and unseen datasets falls short of the results achieved during the training and validation phases. This discrepancy may be attributed to potential data leakage during the augmentation of dataset followed by splitting into training and testing sets, leading to an overestimation of performance in the reference paper. Splitting the dataset into training and testing sets before applying any kind of augmentation would eliminate the possibility of data leakage, and this would also ensure reliable performance measurement of the model.
In summary, careful preprocessing, rigorous performance evaluation, and appropriate model design choices are critical for reliable and generalizable outcomes in multi-class image classification tasks.

6 Conclusion

Skin cancer remains a pressing global health concern, with incidence rates rising significantly in recent decades. Deep CNNs have demonstrated remarkable potential for medical image classification, including skin cancer diagnosis, owing to their ability to automatically extract deep features from data. In this course project, we have reproduced the work presented in the reference paper by developing and evaluating the proposed CNN model alongside three popular ML-based models for multi-class skin lesion classification. Our findings reveal that the proposed CNN model generally outperforms the classical ML models in this task. However, its performance on new, unseen datasets does not align with the results reported in the reference paper. This discrepancy is likely due to data leakage caused by performing dataset augmentation prior to splitting into training and testing sets, leading to an overestimation of the model's performance. To obtain accurate performance measures, it is essential to evaluate ML and deep learning models on independent, unseen test data. Furthermore, achieving reliable and generalizable outcomes in multi-class image classification tasks utilizing ML and deep learning models, such as CNN, requires careful preprocessing of dataset, well-considered model design, and rigorous performance evaluation practices.

7 Dataset and Source Code Repository

The dataset, source code, saved models, and results can be accessed from the Github repository.

References

[1] B. C. Furriel, B. D. Oliveira, R. Prôa, *et al.*, "Artificial intelligence for skin cancer detection and classification for clinical environment: A systematic review," *Frontiers in Medicine*, vol. 10, p. 1 305 954, 2024.

[2] S. S. Chaturvedi, J. V. Tembhurne, and T. Diwan, "A multi-class skin cancer classification using deep convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 28 477–28 498, 2020.

[3] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012," *JAMA dermatology*, vol. 151, no. 10, pp. 1081–1086, 2015.

[4] F. Grignaffini, F. Barbuto, L. Piazzo, *et al.*, "Machine learning approaches for skin cancer classification from dermoscopic images: A systematic review," *Algorithms*, vol. 15, no. 11, p. 438, 2022.

[5] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, vol. 159, no. 3, pp. 669–676, 2008.

[6] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The lancet oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[7] C. Morton and R. Mackie, "Clinical accuracy of the diagnosis of cutaneous malignant melanoma," *British Journal of Dermatology*, vol. 138, no. 2, pp. 283–287, 1998.

[8] T. G. Debelee, "Skin lesion classification and detection using machine learning techniques: A systematic review," *Diagnostics*, vol. 13, no. 19, p. 3147, 2023.

[9] A. R. Lopez, X. Giro-i-Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED international conference on biomedical engineering (BioMed)*, IEEE, 2017, pp. 49–54.

[10] R. Baig, M. Bibi, A. Hamid, S. Kausar, and S. Khalid, "Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images-a review," *Current medical imaging*, vol. 16, no. 5, pp. 513–533, 2020.

[11] A. K. Nugroho, R. Wardoyo, M. E. Wibowo, and H. Soebono, "Image dermoscopy skin lesion classification using deep learning method: Systematic literature review," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 2, pp. 1042–1049, 2024.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[13] P. Tschandl, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, version V4, 2018. DOI: 10.7910/DVN/DBW86T. [Online]. Available: https://doi.org/10.7910/DVN/DBW86T.

[14] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 18 134, 2022.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[16] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.