Automated Detection of Skin Cancer from Dermoscopy Images

Department of Computer Science and Engineering
Project Group ID: 25HR10

***Abstract*—Skin cancer is one of the most common cancers worldwide, and early detection significantly improves patient outcomes. Dermoscopic images provide a non-invasive way to analyze suspicious lesions, but manual diagnosis is time-consuming and subjective. This project focuses on building reliable deep-learning based classification models to distinguish between benign and malignant lesions. We evaluate multiple approaches including a custom CNN, a 7-class CNN, VGG16 transfer learning, ResNet50 fine-tuning, and VGG16 feature extraction with SVM. The HAM10000 dataset is used for training and validation with strict lesion-level splitting to prevent patient-level leakage.**

***Index Terms*—Skin Cancer Detection, Deep Learning, VGG16, ResNet50, CNN, Transfer Learning, HAM10000.**

I. Roles and Responsibilities

Name	Roll No.	Responsibilities
Priyanshi Agrawal	2301CS90	Coordinated project timeline; ensured consistency across all notebooks; supervised preprocessing pipeline; managed lesion-level splitting and dataset integrity verification; final QA and integration of all model outputs.
Shiksha Raginee	2301AI13	Handled dataset reconstruction from Part 1 and Part 2 folders; removed duplicates; validated metadata; implemented oversampling strategies; organized train/val/test CSVs; maintained reproducibility.
Preeti Kumari	2301AI17	Implemented all model architectures: Binary CNN, 7-Class CNN, VGG16 transfer learning, ResNet50 fine-tuning, and SVM using VGG16 features; conducted hyperparameter tuning and training in separate notebooks; GPU optimization.
CH. Rincy Chelciya	2301CS40	Prepared report structure; designed methodology diagrams; wrote algorithms ; consolidated literature review; documented system architecture and prepared slide deck.

TABLE I
Key Responsibilities of Group Members

II. Introduction

Skin cancer is a rapidly increasing global health concern, and accurate early diagnosis plays a major role in reducing mortality. Traditional diagnosis depends heavily on dermatologist expertise, which is not always universally accessible. Deep learning models trained on dermoscopic images have shown potential for achieving dermatologist-level performance. This work focuses on the HAM10000 dermoscopy dataset and builds a complete detection pipeline: preprocessing, lesion-level splitting, data augmentation, CNN-based classification, transfer learning with VGG16 and ResNet50, and hybrid feature-based classification with SVM.

III. Literature Review

Deep learning methods have transformed medical image analysis. Esteva et al.[5] demonstrated dermatologist-level classification performance using a large CNN trained on dermoscopic images. LeCun et al.[6] established CNNs as powerful feature learners for visual tasks. Transfer learning approaches using ImageNet-pretrained networks such as VGG16[2] and ResNet[3] have proven effective in medical imaging where labeled data is limited. Data augmentation improves generalization in image classification[4]. SVMs remain strong shallow classifiers for high-dimensional embeddings[7]. The HAM10000 dataset[1] provides multi-source dermoscopic images for benchmarking. Despite progress, challenges persist including class imbalance, dataset bias, and patient-level leakage, which we address explicitly in our methodology.

IV. Methodology

A. Dataset Description

We use the HAM10000 dataset consisting of 10,015 dermoscopic images belonging to seven diagnostic categories. The dataset originally appears in two folders; however, many images are duplicated across Part 1 and Part 2 with identical filenames. A strict deduplication step was applied.

B. Preprocessing Pipeline

- Deduplication across both folders based on identical filenames.
- Merging images into a single directory.
- Metadata cleaning and validation.
- Lesion-level splitting using `GroupShuffleSplit` to prevent patient leakage.
- Oversampling (row-level) to balance minority classes.
- Saving finalized splits as CSV files for reproducibility.

C. Model Architectures

1) Binary CNN (Benign vs. Malignant): A lightweight CNN with three convolution blocks:

- Conv(32) $\rightarrow$ MaxPool $\rightarrow$ BatchNorm
- Conv(64) $\rightarrow$ MaxPool $\rightarrow$ BatchNorm
- Conv(128) $\rightarrow$ MaxPool $\rightarrow$ BatchNorm
- Dense(128) $\rightarrow$ Dropout(0.5) $\rightarrow$ Dense(2)

2) 7-Class CNN (From Scratch): Same structure as the binary CNN but final layer expanded to 7 classes and Dense(256) before output.

3) VGG16 Transfer Learning:

- Pretrained VGG16 with `include_top=False`
- Freeze all except last four layers
- Global Average Pooling
- Dense(256) → Dropout(0.5) → Dense(7)

4) ResNet50 Fine-Tuning:

- Pretrained ResNet50
- Freeze first 100 layers, fine-tune the rest
- Global Average Pooling → Dense(256) → Dropout → Dense(7)

5) VGG16 Feature Extractor + SVM:

- Use VGG16 backbone to extract 512-D GAP features
- Standardize using `StandardScaler`
- Train Linear SVM

V. Algorithms (Russell–Norvig Style)

A. Algorithm 1: Preprocessing and Splitting

Algorithm 1 Dataset Preprocessing

1: Load metadata and image files
2: Remove duplicate filenames across both folders
3: Build absolute filepaths
4: Encode class labels
5: Perform lesion-level train/val/test split
6: Oversample minority classes in training set
7: Save split CSV files

B. Algorithm 2: Model Training Framework

Algorithm 2 Training a Deep Learning Model

1: Load train/val CSVs
2: Select preprocessing function (rescale or pretrained)
3: Initialize model architecture
4: Compile with Adam optimizer
5: Train with early stopping and LR scheduler
6: Evaluate on validation and test sets
7: Save best model weights

VI. Results and Discussion

A. 7-Class CNN (Baseline Model)

The baseline CNN was trained on 29,491 oversampled training images, with 861 validation and 1,780 test images. The model reached its best validation accuracy of 67.13% in the 5th epoch, after which validation loss began to increase, indicating early overfitting.

On the test set, the model achieved an accuracy of 64.55% and a test loss of 1.0353. Performance varied across classes, with the model performing well on visually consistent lesions such as *nv* and *vasc*, while struggling with fine-grained and clinically similar categories such as *akiec*, *df*, and *mel*.

TABLE II
Classification Report (7-Class CNN)

Class	Precision	Recall	F1-score
akiec	0.32	0.20	0.25
bcc	0.46	0.50	0.48
bkl	0.37	0.60	0.46
df	0.15	0.67	0.25
mel	0.37	0.51	0.43
nv	0.92	0.71	0.80
vasc	0.63	0.85	0.72
Overall Accuracy		**64.55%**	

Key Observations: The CNN successfully learned broad lesion patterns but lacked the capacity to distinguish subtle structural and texture-based differences. This establishes a baseline for comparison, highlighting the need for stronger feature extractors such as VGG16 and ResNet50.

B. VGG16 Transfer Learning

The VGG16-based model was trained using ImageNet-initialized weights with the final four convolutional blocks unfrozen for fine-tuning. A total of 29,491 training samples were used after balanced oversampling, with 861 validation and 1,780 test images.

The model achieved a best validation accuracy of 73.29% in the second epoch and later showed mild overfitting, despite early stopping and learning-rate scheduling. On the test set, VGG16 reached an accuracy of 72.64% with a test loss of 0.909, achieving a significant improvement over the baseline CNN.

TABLE III
Classification Report (VGG16 Transfer Learning)

Class	Precision	Recall	F1-score
akiec	0.51	0.33	0.40
bcc	0.63	0.69	0.66
bkl	0.49	0.75	0.59
df	0.48	0.48	0.48
mel	0.40	0.62	0.49
nv	0.94	0.77	0.85
vasc	0.88	0.75	0.81
Overall Accuracy		**72.64%**	

Key Observations: VGG16 demonstrated stronger generalization than the baseline CNN owing to deeper feature hierarchies and pretrained filters. Improvements were especially visible in *bkl*, *bcc*, and *mel*, while *nv* remained the easiest class. Moderate confusion persisted among visually similar lesion types, suggesting room for further gains through deeper architectures (e.g., ResNet) or attention-based models.

C. Binary CNN (Benign vs. Malignant)

The custom binary CNN was trained on 29,491 oversampled images, validated on 861 samples, and tested on 1,780 images. The model achieved a best validation accuracy of 84.67% and a final test accuracy of **81.01%** with a test loss of 0.4903.

TABLE IV
CLASSIFICATION REPORT (BINARY CNN)

Class	Precision	Recall	F1-score
Benign	0.87	0.89	0.88
Malignant	0.54	0.49	0.51
Overall Accuracy		**81.01%**	

Key Observations: The binary CNN successfully captured coarse distinctions between benign and malignant lesions but struggled with malignant recall (49%), likely due to lesion variability. Despite oversampling, class imbalance and limited feature depth restricted generalization. This model provides a strong early baseline for binary classification performance.

D. ResNet50 Fine-Tuning (7-Class Classification)

The ResNet50 model was partially unfrozen (fine-tuning after the first 100 layers) and trained on 29,491 oversampled samples. The best validation accuracy reached 80.14% at epoch 4. On the test set, the model achieved a final accuracy of **77.53%** with a test loss of 1.0418.

TABLE V
CLASSIFICATION REPORT (RESNET50 FINE-TUNED)

Class	Precision	Recall	F1-score
akiec	0.67	0.30	0.41
bcc	0.73	0.46	0.57
bkl	0.49	0.82	0.61
df	0.64	0.33	0.44
mel	0.52	0.55	0.53
nv	0.93	0.87	0.89
vasc	0.93	0.65	0.76
Overall Accuracy		**77.53%**	

Key Observations: ResNet50 outperformed the baseline CNN and approached VGG16 performance, particularly in high-variance classes (*bkl*, *mel*). The model demonstrated strong feature extraction capabilities, especially for structurally consistent lesions such as *nv*. However, overfitting began after epoch 4 despite learning-rate scheduling. This suggests that deeper fine-tuning or augmentation strategies could further enhance accuracy.

E. VGG16 Feature Extraction + Linear SVM

In this experiment, the VGG16 convolutional base was used strictly as a fixed feature extractor, and a linear Support Vector Machine (SVM) was trained on the resulting feature embeddings. This model tested whether classical machine learning can effectively utilize deep feature representations for dermoscopic classification.

On the 7-class test set, the model achieved a final accuracy of **61.69%**.

Key Observations: Using VGG16 as a fixed feature extractor provided stable representations, but the linear SVM struggled with highly similar lesion types and produced limited recall for malignant categories. The method performed best for the *nv* class, which is abundant and visually consistent. However, its inability to adapt feature hierarchies (due to frozen weights) constrained performance. The results confirm that transfer learning with end-to-end fine-tuning (as in VGG16 and ResNet50) is better suited for complex dermoscopic patterns.

TABLE VI
CLASSIFICATION REPORT (VGG16 + SVM)

Class	Precision	Recall	F1-score
akiec	0.18	0.37	0.25
bcc	0.27	0.33	0.30
bkl	0.37	0.45	0.41
df	0.16	0.29	0.20
mel	0.32	0.45	0.37
nv	0.90	0.72	0.80
vasc	0.50	0.55	0.52
Overall Accuracy		**61.69%**	

F. Overall Model Comparison

Table VII summarizes the performance of all four models implemented in this study. The comparison highlights the progression from a simple CNN baseline to deeper transfer learning architectures and hybrid ML-Deep Learning approaches.

Model	Test Accuracy (%)	Key Notes
Binary CNN (Benign vs Malignant)	81.01	Strong binary performance; malignant recall still low
Custom CNN (7-Class Scratch)	64.55	Baseline model; limited generalization
VGG16 (Transfer Learning)	72.64	Strong pretrained features; good improvement
ResNet50 (Fine-Tuning)	**77.53**	Best 7-class model; deeper residual features
VGG16 + Linear SVM	61.69	Good features; linear classifier limits performance

TABLE VII
PERFORMANCE COMPARISON OF ALL MODELS IMPLEMENTED IN THIS STUDY

The results show that deeper pretrained architectures with end-to-end fine-tuning consistently outperform both classical machine-learning hybrids and shallow CNNs.

TABLE VIII
COMPARISON OF ALL MODELS USING ACCURACY, MACRO RECALL, AND MACRO F1-SCORE

Model	Accuracy (%)	Macro Recall	Macro F1
Custom CNN (Scratch)	64.55	0.49	0.48
VGG16 (Transfer Learning)	72.64	0.63	0.61
ResNet50 (Transfer Learning)	**77.53**	**0.57**	**0.60**
VGG16 + Linear SVM	61.69	0.45	0.41
Binary CNN (Benign/Malig.)	81.01	0.69	0.70

VII. CONCLUSION

This work investigated multiple deep learning and hybrid architectures for dermoscopic image classification using the HAM10000 dataset. The experimental results demonstrated a clear trend: models that leverage transfer learning from large-scale pretrained networks achieve substantially better performance than shallow CNNs trained from scratch. Among all models evaluated, ResNet50 achieved the highest test accuracy of 76.91%, indicating that deeper residual representations are well suited for capturing the subtle visual variations between different lesion types.

Despite these promising results, several limitations remain. The dataset exhibits significant intra-class variability and inter-class similarity, making certain lesions inherently difficult to

classify. The class imbalance, even after oversampling, influenced minority-class recall. Moreover, some models showed signs of overfitting, suggesting that additional regularization, more diverse augmentation, or larger balanced datasets could further improve performance.

For future work, several directions appear promising. First, the integration of attention mechanisms or vision transformers may enhance the model's ability to focus on medically relevant regions. Second, ensembling multiple pretrained models could yield more robust predictions. Third, incorporating metadata (age, lesion location, patient history) alongside image features may allow for more clinically meaningful predictions. Finally, exploring semi-supervised or self-supervised learning could reduce dependence on fully annotated medical datasets.

Overall, the study validates the effectiveness of deep transfer learning for skin lesion classification and provides a strong foundation for building more accurate and interpretable computer-aided diagnostic systems.

VIII. References

[1] 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, 2018.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[4] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv:1712.04621, 2017.

[5] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 1995.

[8] M. Lin, Q. Chen, and S. Yan, "Network in network," in *ICLR*, 2014.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE TKDE*, 2010.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2020.

[11] F. Chollet et al., "Keras," GitHub Repository, 2015.

[12] M. Abadi et al., "TensorFlow," 2016.

[13] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal*, 2000.

[14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *JMLR*, 2011.