Skin Cancer Classification: A comparative analysis of various Deep Learning Models with Feature Engineering techniques

Sameek Bhattacharya
sameekbhattacharya@my.unt.edu
University of North Texas
Denton, TX, USA

Harold Lucero
haroldlucero@my.unt.edu
University of North Texas
Denton, TX, USA

Aryan Patel
aryanpatel@my.unt.edu
University of North Texas
Denton, TX, USA

Abstract

This paper presents a deep learning approach to classify skin lesions into benign or malignant categories. Leveraging recognized convolutional neural networks such as ResNet50, DenseNet121, VGG16 and InceptionV3, we explore the impact of feature engineering methods such as CLAHE and polynomial transformations, and subsequently perform fine-tuning of these models. Experimenting with these techniques and their combinations, we evaluate which preprocessing method and architecture offer the best classification performance on a skin lesion image dataset.

In this study, DenseNet121 is shown to achieve a harmonic mean F1-score of 89%, indicating consistency between precision and recall, while also achieving an accuracy of 89%. These results highlight the great potential of deep networks, especially when combined with feature engineering techniques, opening the possibility to explore in the future methods of self-optimization of preprocessing and architectures to further improve performance.

Keywords: Deep Learning, Skin Cancer Classification, Neural Networks, Binary Classification, ResNet50, DenseNet121, VGG16, InceptionV3, Comparative Analysis

1 Introduction

Skin cancer is one of the most common forms of cancer worldwide, with increasing incidence rates attributed to factors such as rising UV exposure, aging populations, and lifestyle changes. Early detection is crucial, as it significantly improves treatment outcomes and survival rates. However, differentiating between benign and malignant skin lesions can be challenging for dermatologists due to the wide variety of appearances these lesions can exhibit. This is where the application of machine learning becomes invaluable.

In our project, we focus on two classes of skin lesions: benign and malignant. This distinction is critical because malignant lesions are cancerous, while benign lesions are not, making treatment optional and primarily cosmetic. Although these two classes differ greatly in their danger levels, they can appear strikingly similar in some cases, making it difficult to tell them apart. By developing a model to differentiate between these two classes, we can enhance early detection and reduce the risk of misdiagnosis.

We will be using advanced convolutional neural network (CNN) models, namely ResNet50, DenseNet121, VGG16, and Inception V3. We chose these architectures because each has unique strengths, allowing us to explore various aspects of image representation. Furthermore, to improve our model's accuracy, we will employ different feature engineering techniques, such as Contrast Limited Adaptive Histogram Equalization (CLAHE) and polynomial transformation. By systematically applying these techniques CLAHE, polynomial transformations, and a combination of both, we aim to enhance the quality of the input images. We will also determine which feature engineering technique is best suited for the task, along with the most effective architecture.

The organization of this paper is structured to detail the development and evaluation of the models. We begin with a review of related work to contextualize our project within the existing literature. The methodology section follows, outlining the problem formulation and detailing the architecture of each model, as well as the metrics used to evaluate them. We then describe the fine-tuning process and the overall framework. The evaluation section covers the experimental setup, the dataset used, the feature engineering techniques used, and present the results and analysis, including visual results and comparison between each implementation. Finally, we conclude the article and propose future lines of work to further improve the models' capabilities.

2 Related Work

This section details 3 works taken as reference for the development of this project.

2.1 Enhancing Skin Cancer Classification Accuracy Using Computer Vision Techniques [1]

This study concentrates on boosting the accuracy of skin cancer detection by leveraging computer vision approaches. It proposes two main strategies: one utilizing deep learning through Convolutional Neural Networks (CNNs) for automatic feature extraction and classification, and another combining manual feature extraction with conventional machine learning classifiers such as Support Vector Machines

(SVM), Naïve Bayes, K-Nearest Neighbor (KNN), and Decision Trees. The evaluation criteria included metrics like accuracy, precision, recall, and F1 score. The results indicated that the CNN-based deep learning model outperformed traditional machine learning methods. Nonetheless, the study acknowledged challenges such as inconsistent image quality and potential dataset biases, noting that models trained on specific datasets might struggle to generalize across broader, more diverse populations.

2.2 The Role of Feature Engineering in Enhancing Machine Learning Performance [2]

This paper explores the vital role feature engineering plays in improving the performance of various machine learning models. The authors applied a range of feature engineering techniques including feature selection, feature extraction, and feature transformation across different datasets and assessed their impact on model performance. The findings revealed that selecting the appropriate feature engineering methods significantly enhanced both the accuracy and efficiency of machine learning algorithms. However, the paper also highlights that the success of specific techniques can vary depending on the nature of the dataset and the models employed.

2.3 Optimized CNN Models and Multi-Criteria Decision-Making for Skin Cancer Classification [3]

This research targets the early detection of skin cancer by developing and optimizing CNN-based models. It uses architectures such as AlexNet, Inception V3, MobileNet V2, and ResNet50 for feature extraction, with feature reduction performed using two versions of the Grey Wolf Optimizer (GWO) algorithm. Skin lesion images were categorized into four classes using six different machine learning classifiers, producing a total of 51 models, which were subsequently ranked through the multi-criteria decision-making method, RAPS. The highest classification accuracy of 94.5% was achieved by combining AlexNet with GWO. Additionally, feature reduction led to improved classification accuracy and decreased training time. A noted limitation of the study is its reliance on a specific dataset; broader validation on larger and more varied dermoscopic datasets is needed to enhance model generalizability.

3 Methodology

In this section, we first detail how was the study done. Next, we provide an overview of the models by detailing its architecture and the metrics used to evaluate them.

3.1 Feature Engineering Techniques

This section details the techniques used for image processing such as CLAHE and Polynomial Transformation.

CLAHE [4] : Contrast Limited Adaptive Histogram Equalization is an image processing technique used to enhance the contrast of an image, especially in cases where lighting conditions are poor or uneven.

It divides the image into small tiles (or regions) and applies histogram equalization separately to each tile. This enhances local contrast without over-amplifying noise (AHE). To prevent over-amplification of noise, CLAHE limits the contrast enhancement by clipping the histogram at a predefined threshold (clip limit) and redistributing the intensity values. This ensures that details in bright and dark areas are preserved without causing artifacts (CL).

The relevance of CLAHE for Skin Cancer Images is detailed below:

- Enhances the visibility of lesion boundaries and fine details.
- Improves contrast without over-enhancing noise, making it useful for medical images.
- Helps in better feature extraction for machine learning and deep learning models.

Polynomial Feature Transformation [5]: Feature extracted through different steps are flattened and fed to ML models for classification. The extracted features do not always fit a straight line and that's where polynomial feature transformation adds value. Once all the feature extraction steps are complete, the final features are recorded in a 1D vector for each data point. Then the features are combined by multiplying each other and itself for 'n' times. The polynomial value of 'n' is a hyperparameter that can be tuned for every problem. In our case, we will start the training process with a value of n = 2, which refers to a polynomial degree of 2.

Once the polynomial features are calculated, they will be appended with the existing features extracted from the previous steps and passed to the machine learning model for training.

The relevance of polynomial feature transformation for Skin Cancer Images is detailed below:

- Combining extracted features adds information on how each feature is related to the other. For e.g., mean intensity of an image and luminance noise of an image individually provides information, but when combined through dot product gives insight if images with higher noise is expected to have high mean intensity or not.
- This adds non-linearity to the features which will allow the machine learning model to learning complex non-linear relations in the data.

3.2 Deep Learning Models

Within this section the fine-tuned models and the number of parameters used are detailed.

ResNet50 Architecture.

ResNet (Residual Network) [6] is a deep learning architecture introduced by Microsoft Research in 2015 to tackle the

problem of training very deep neural networks. As networks get deeper, they often suffer from vanishing gradients, making it hard for them to learn effectively. ResNet solves this by using skip connections (also called residual connections) that bypass one or more layers. Instead of learning the full output, each block learns the residual (or difference) from the input. Mathematically, if a few stacked layers fit a mapping H(x), ResNet lets them fit F(x) = H(x) - x. Thus the original function becomes H(x) = F(x) + x.

This small change allows much deeper networks like ResNet-50, ResNet-101, or ResNet-152 to be trained successfully without performance degradation. Each "residual block" usually has two or three convolutional layers, batch normalization, and ReLU activations. ResNet-50, for example, is a 50-layer deep network that stacks many such blocks. Because of its strong performance and stability in training, ResNet has become a foundational architecture in computer vision tasks like image classification, detection, and segmentation. Its design is simple yet powerful, making it highly adaptable for many variations and improvements.

DenseNet121 Architecture.

This architecture is especially efficient in terms of parameters and takes advantage of dense connections between layers to improve feature and gradient propagation, which is useful for complex tasks such as medical diagnosis from images [7].

For this specific case, we loaded a pre-trained version of the model, modifying the last layer of the model to adapt it to this benign and malignant classification problem, replacing the final layer by a new linear layer with two outputs, corresponding to the two classes, resulting in a Deep Neural Network with 6,955,906 parameters.

VGG-16 Architecture.

This is a deep convolutional neural network that uses only convolutional layers of size 3x3 and max pooling layers, followed by several fully connected layers at the end [8].

For this specific case, we loaded a pre-trained version of the model, modifying the last layer of the model to adapt it to this benign and malignant classification problem, replacing the final layer by a new Linear layer with two outputs, corresponding the two classes, resulting in a Deep neural network with 35,601,738 parameters.

Inception V3 Architecture.

The Inception model originated from GoogleLeNet, which won the ImageNet Scale Visual Recognition Challenge in 2014. InceptionV2/V3 are improved versions with better performance and training stability, introduced in 2015 in this paper [9] The core principle behind InceptionV3 is that instead of choosing a single filter size, InceptionV3 processes input through multiple parallel filter sizes, plus max pooling. This allows it to capture information at multiple scales.

For this project, we used a pre-trained InceptionV3 base for transfer learning, then added global average pooling, Dense and softmax for 2 class classification, resulting in a network with 21,806,882 parameters.

3.3 Baseline & Evaluation Metrics

For this purpose, we will use a stratified model as baseline, which assigns class labels randomly, but maintaining the actual proportion of classes present in the dataset.

In our case, we are tackling a binary skin cancer classification task (benign vs malignant), and although the classes are not perfectly balanced, the difference is not extreme, as one of the classes has approximately 120 more images than the other. Thus, we expect the accuracy of the stratified baseline model to be slightly above 50%, reflecting that small bias toward the most represented class.

The effectiveness of the methodology for detecting benign and malignant skin lesions will be evaluated using Precision, Recall, and F1-score, metrics to be obtained through the classification report.

Precision:

Measures the quality of the classifier when predicting whether a lesion is benign or malignant. It is calculated using Equation 1:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall:

Represents the percentage of cases where the model correctly identifies a lesion as benign or malignant. It is calculated using Equation 2:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

F1-score:

Measures the balance between precision and recall, reflecting the overall quality of the classifier. It is calculated using Equation 3:

$$F_1 - score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Accuracy:

It is defined as the proportion of correct predictions in both classes with respect to the total number of examples provided. It is calculated using the following Equation 4:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Where:

TP = True Positive **TN** = True Negative
FP = False Positive **FN** = False Negative

3.4 Fine-Tuning the models

When fine-tuning the models the following parameters were selected: For the loss function we selected Cross Entropy Loss because it is the standard choice for multiclass classification problems as the function allows counteracting the existing imbalance of 120 additional images of the majority class by penalizing errors on the minority class more heavily. Overall, the loss guarantees a well-behaved gradient aligned with the evaluation metrics. As for the optimizer, we used Adam with an initial learning rate of 1×10^{-4}.

- ResNet50 was trained in 15 epochs.
- DenseNet121 was trained in 15 epochs.
- VGG-16 was trained in 18 epochs.
- InceptionV3 was trained in 10 epochs

For all models, the best trained model was saved based on the validation metrics.

4 Results

This section details the resources used for the implementation of this project, the dataset, the experiments performed and the results obtained by training each of the models.

4.1 Experimental Setup

The training of the models was carried out on the Google Colab platform using an NVIDIA T4 GPU with 16 GB of memory. For light debugging tasks and local tests, a computer with an Intel Core i7-8550U processor at 1.80 GHz and 12 GB of RAM was used.

The entire implementation was developed in Python, using PyTorch for model definition and training and OpenCV for image preprocessing and augmentation. All training ran on Google Colab, and we tested the model with inference tests on a separate set of images.

4.2 Dataset

The International Skin Imaging Collaboration (ISIC) Archive is the largest publicly available collection of dermoscopic images, aimed at advancing research in skin cancer detection. As of 2025, it hosts over 1.16 million images, with more than 500,000 publicly accessible, covering a wide range of skin lesion types and associated metadata. Key datasets include HAM10000, containing 10,015 images of common pigmented skin lesions, and the ISIC Challenges datasets (2016–2020), which provided curated images for lesion segmentation, attribute detection, and disease classification tasks. Other notable collections, like the Hospital Italiano de Buenos Aires dataset, add diversity for AI research. The archive supports standardized metadata and data dictionaries to ease integration and analysis. Widely used by researchers, clinicians, and educators, ISIC is a vital resource for improving the accuracy and accessibility of skin cancer diagnostics [10].

The specific filters used to download the dataset are shown below.

- **Benign**
  - Benign Epidermal Proliferations
  - Pigmented Benign Keratosis
  - Lichen Planus-like Keratosis
- **Malignant**
  - Malignant Epidermal Proliferations
  - Squamous Cell Carcinoma, NOS
  - Squamous Cell Carcinoma, Invasive
  - Squamous Cell Carcinoma in situ

Subsequently, the images are distributed into training, validation, and test datasets, following the distribution shown below.

- **Train Folder (Total = 2768)**
  - Benign : 1324
  - Malignant : 1444
- **Val Folder (Total = 664)**
  - Benign : 339
  - Malignant : 325
- **Test Folder (Total = 80)**
  - Benign : 40
  - Malignant : 40

4.3 Experiments

In the conducted experiments, each model was initially trained using images enhanced with CLAHE. Subsequently, the models were trained using images processed exclusively with polynomial transformation. Finally, both preprocessing techniques, CLAHE and polynomial transformation were applied jointly to train all models. The resulting images after each preprocessing step are presented in Fig. 1.

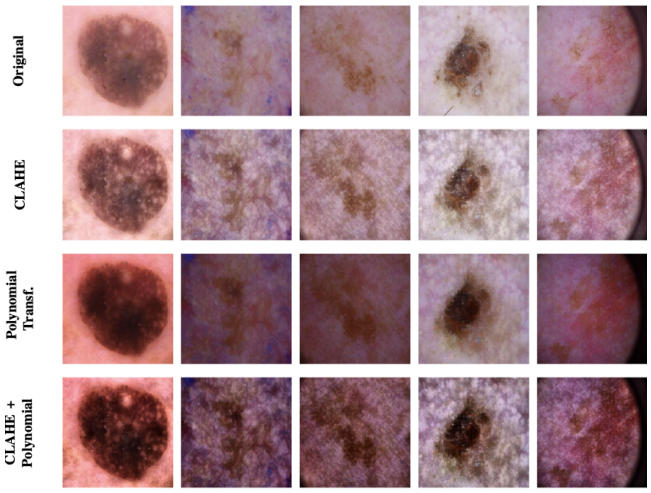

Figure 1. Visual comparison of different Feature Engineering methods applied to the same input image.

The results in Table 1 reveal key differences in model performance across classes and preprocessing techniques.

Table 1. Precision, Recall, and F1-Score for each model and preprocessing method across benign and malignant classes.

Model	Preprocessing	Class	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	CLAHE	Benign	92.00	85.00	88.00
		Malignant	86.00	93.00	89.00
	Polynomial	Benign	88.00	85.00	86.00
		Malignant	85.00	88.00	86.00
	CLAHE + Polynomial	Benign	91.00	85.00	88.00
		Malignant	85.00	91.00	88.00
DenseNet121	CLAHE	Benign	89.00	89.00	89.00
		Malignant	88.00	89.00	89.00
	Polynomial	Benign	90.00	87.00	88.00
		Malignant	87.00	90.00	88.00
	CLAHE + Polynomial	Benign	86.00	90.00	88.00
		Malignant	89.00	85.00	87.00
VGG16	CLAHE	Benign	86.00	92.00	89.00
		Malignant	91.00	84.00	87.00
	Polynomial	Benign	88.00	83.00	86.00
		Malignant	84.00	89.00	86.00
	CLAHE + Polynomial	Benign	87.00	85.00	86.00
		Malignant	84.00	87.00	86.00
InceptionV3	CLAHE	Benign	70.00	88.00	78.00
		Malignant	83.00	62.00	71.00
	Polynomial	Benign	85.00	85.00	85.00
		Malignant	85.00	85.00	85.00
	CLAHE + Polynomial	Benign	67.00	97.00	80.00
		Malignant	95.00	53.00	68.00

- ResNet50 Performance: With CLAHE, ResNet50 achieves strong malignant detection, reporting a precision of 86%, recall of 93%, and F1-score of 89%. Polynomial transformation yields balanced results with F1-scores of 86% for both benign and malignant classes. However, using polynomial transformation degrades performance a little bit, dropping F1-scores to 86% for both classes..
- DenseNet121 Performance: DenseNet121 consistently delivers high performance. With CLAHE alone, it reaches an F1-score of 89% for both classes. The Polynomial transform maintains similar effectiveness with a F1-score of 88%, while the combination of CLAHE and Polynomial remains competitive, achieving F1-scores of 88% ffor benign and 87% for malignant.
- VGG16 Performance: VGG16 performs best with CLAHE, where benign lesions are detected with an F1-score of 89% and malignant ones with 87%. Polynomial transformation and the combined approach yield stable results around 86% F1-score across both classes.
- InceptionV3 Performance: InceptionV3 shows variability depending on preprocessing. Polynomial transformation offers the most balanced results, with 85% F1-score for both classes. CLAHE combined with Polynomial enhances benign recall to 97% but causes a significant drop in malignant recall of 53%, resulting in an imbalanced F1-score of 80% for benign and 68% for malignant.

On the other hand, Table 2 summarizes the overall accuracy achieved by each combination of model and preprocessing technique evaluated.

Table 2. Accuracy (%) of each model across different preprocessing techniques.

Model	Preprocessing	Accuracy (%)
ResNet50	CLAHE	89.00
ResNet50	Polynomial	86.00
ResNet50	CLAHE + Polynomial	88.00
DenseNet121	CLAHE	89.00
DenseNet121	Polynomial	88.00
DenseNet121	CLAHE + Polynomial	88.00
VGG16	CLAHE	88.00
VGG16	Polynomial	86.00
VGG16	CLAHE + Polynomial	86.00
InceptionV3	CLAHE	75.00
InceptionV3	Polynomial	85.00
InceptionV3	CLAHE + Polynomial	75.00

The accuracy results presented in Table 2 provide an overview of how each model performs under the different preprocessing strategies.

- ResNet50 achieves its highest accuracy of 89% when using CLAHE, followed by Polynomial transformation preprocessing with 86%. Then, when both techniques are combined, the model achieves an accuracy of 88%.
- DenseNet121 shows excellent stability, maintaining high accuracy across all preprocessing variants. Both CLAHE and CLAHE + Polynomial transformation yield an accuracy of 89% and 88%, respectively, suggesting that this architecture is less sensitive to preprocessing changes.
- VGG16 also maintains consistent accuracy, scoring 88% with CLAHE and 86% with both Polynomial transformation and the combined approach. The results suggest moderate gains with CLAHE alone.
- InceptionV3, in contrast, exhibits more sensitivity to preprocessing methods. Its best performance is observed with Polynomial transforms of 85%, while both CLAHE alone and the combined method result in a lower accuracy of 75%, pointing to potential limitations in how InceptionV3 handles enhanced or compounded image features.

4.4 Inference

Once the metrics obtained by each model trained under the different combinations of Feature Engineering techniques have been evaluated, the inference stage is carried out using images from the test dataset. For this purpose, different hypothetical scenarios have been considered in which, in a clinical context, physicians could apply these models for the detection of skin cancer in their patients.

The first scenario involves individual patient care, where a single image is analyzed and the result is delivered immediately. For this case, the ResNet50 and InceptionV3 models have been used. The inference is performed on a point image, as shown in Fig. 2.

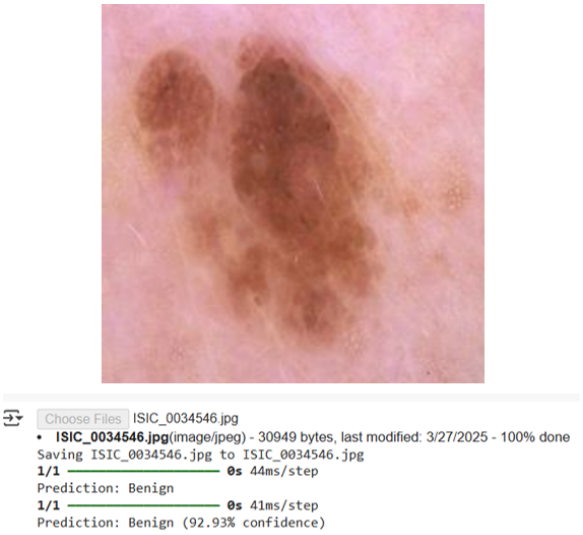

Figure 2. ResNet50 model giving the prediction of a test image.

In this case, the user can provide an image for the model to perform inference. The model then classifies the image as benign with a confidence score of 92.93%.

In a second scenario, the massive evaluation of multiple images at the same time is considered, for example, in community screening or patient follow-up studies. In this case, the use of DenseNet121 and VGG16 on a batch of images is proposed, generating joint predictions. If the models disagree, a comparative analysis is proposed to determine a break-even point or diagnostic consensus. This implementation is illustrated in Fig. 3.

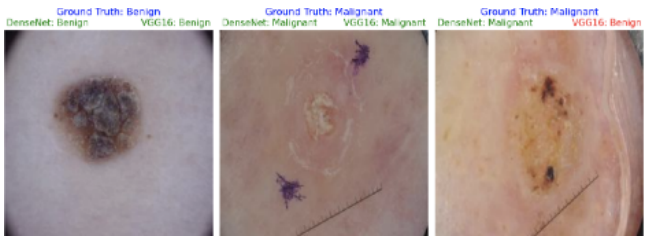

Figure 3. Inference on a batch of pictures with VGG-16 and DenseNet121.

As shown in Fig. 3, the first box corresponds to an image with benign ground truth, where both models made correct predictions. The second image presents a malignant case,

also correctly classified by both models. Finally, the third image corresponds to a malignant case in which DenseNet121 makes a correct prediction, while VGG16 misclassifies it as benign. This type of discrepancies could be useful to prioritize certain images and refer them to a more detailed medical review.

5 Discussion

The models demonstrated strong performance in classifying skin lesions as benign or malignant, particularly with effective feature engineering. DenseNet121 was the most consistent performer, achieving high F1-scores and accuracies, while ResNet50 and VGG16 excelled with CLAHE preprocessing, highlighting the importance of contrast enhancement. InceptionV3 showed variability, performing best with polynomial transformations but struggling when combining feature techniques.

A noted issue was the inconsistent performance when combining CLAHE and polynomial transformations, especially with InceptionV3, suggesting that excessive manipulation may introduce noise. While the dataset was relatively balanced, a modest class imbalance may have affected recall for malignant cases.

Our findings support previous research indicating that deep learning outperforms traditional methods in skin cancer classification when feature engineering is applied. CLAHE notably improved performance for DenseNet121 and VGG16, but model generalizability remained sensitive to preprocessing choices.

Future work could focus on multi-class classification for specific skin cancer types and optimizing models for real-time deployment to enhance early detection accessibility.

6 Conclusion

In this project, we focused on implementing and evaluating deep learning models for the detection of skin cancer using dermoscopic images. Specifically, we employed four well-known convolutional neural network architectures: ResNet50, DenseNet121, VGG16, and InceptionV3. The primary goal of this work was not only to assess the performance of these models but also to explore the impact of combining them with feature engineering techniques to enhance classification accuracy.

For feature engineering, we applied two methods: Contrast Limited Adaptive Histogram Equalization (CLAHE) for image enhancement and a Polynomial Feature Transformation technique, which involved squaring the extracted features. CLAHE was used to improve the contrast and highlight important structures within the images, while the polynomial transformation aimed to introduce non-linear interactions between features.

Among the different model and feature engineering combinations tested, ResNet50 combined with CLAHE achieved the highest classification accuracy of 89%. This result highlights the effectiveness of using CLAHE as a preprocessing step to boost model performance by enhancing the visual quality of dermoscopic images, making critical features more distinguishable for the network.

However, an interesting and important observation emerged when we combined Inception V3 with both CLAHE and polynomial feature transformation. In this case, the model's accuracy significantly dropped to 75% as well as with only CLAHE. This outcome suggests that while certain feature engineering techniques can individually enhance performance, combining multiple methods without careful validation can lead to feature redundancy, noise amplification, or overfitting, ultimately degrading model accuracy.

This finding serves as a crucial takeaway for future work: it is essential to carefully evaluate the interaction effects of different preprocessing and feature engineering techniques. Blindly stacking multiple transformations may not always result in improved performance and could, in fact, have the opposite effect. Overall, this project emphasizes the importance of not only choosing appropriate deep learning models but also selecting and combining feature engineering strategies thoughtfully, based on empirical validation rather than assumptions.

References

[1] Magdy, A., Hussein, H., Abdel-Kader, R. F., & Abd El Salam, K. (2023). Performance enhancement of skin cancer classification using computer vision. IEEE Access, 11, 72120-72133.

[2] Abdelouahed, S. M., Abla, R., Asmae, E., & Abdellah, A. (2024, May). Harnessing feature engineering to improve machine learning: A review of different data processing techniques. In 2024 International Conference on Intelligent Systems and Computer Vision (ISCV) (pp. 1-6). IEEE.

[3] Saleh, N., Hassan, M. A., & Salaheldin, A. M. (2024). Skin cancer classification based on an optimized convolutional neural network and multicriteria decision-making. Scientific Reports, 14(1), 17323.

[4] Musa, P., Al Rafi, F., & Lamsani, M. (2018, October). A Review: Contrast-Limited Adaptive Histogram Equalization (CLAHE) methods to help the application of face recognition. In 2018 third international conference on informatics and computing (ICIC) (pp. 1-6). IEEE.

[5] Chen, X., & Tung, Y. K. (2003). Investigation of polynomial normal transform. Structural Safety, 25(4), 423-445.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks. https://doi.org/10.48550/arXiv.1608.06993

[8] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

[9] Szegedy, C., Vanhoucke, V., Loffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. https://arxiv.org/abs/1512.00567

[10] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Medical image analysis, 75, 102305.