Exploring Vision-Language and Deep Learning Models for Skin Cancer Classification

Zheyuan Xiao
Department of Computer Science
The University of Texas at Austin

Abstract

Skin cancer is one of the most prevalent forms of cancer, with early detection being critical for effective treatment. This paper presents an exploration of various machine learning techniques, including Random Forest, Custom CNNs, ResNet with transfer learning, and advanced vision-language models like CLIP and BLIP, for the classification of skin lesions using the HAM10000 dataset. Our findings highlight the superior performance of ResNet (85.16% accuracy) and the competitive results achieved by vision-language models, particularly CLIP (84.96% accuracy), demonstrating their potential for medical imaging applications. The results emphasize the value of integrating multimodal approaches for robust and scalable diagnostics in skin lesion classification. All code and resources for this study are publicly available at https://github.com/zxia545/SkinLesionClassifier.

1 Introduction

Skin cancer is one of the most prevalent and potentially life-threatening forms of cancer, accounting for a significant proportion of global cancer cases. According to the World Health Organization (WHO), nearly

[*] The implementation code and additional resources are publicly available at: https://github.com/zxia545/SkinLesionClassifier.

one-third of all diagnosed cancers are attributed to skin malignancies (Urban et al. (2021)). The primary types include melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC), with melanoma being the most aggressive form, often fatal if not identified and treated early (Ahmed et al. (2023)). Factors such as increased exposure to ultraviolet (UV) radiation, environmental changes, and shifting lifestyle habits have contributed to a marked rise in the incidence of skin cancer worldwide (Hekler et al. (2019)).

Accurate and timely diagnosis is crucial for improving patient outcomes and survival rates. However, traditional diagnostic methods, which depend on clinical examination and histopathological evaluation, are resource-intensive and require highly skilled dermatologists, making them less accessible in under-resourced regions (Arooj et al. (2022)). To address these challenges, computational techniques such as machine learning and other data-driven methodologies have emerged as promising tools to provide scalable and cost-effective diagnostic solutions (Surendren and Sumitha (2023)).

Machine learning methods, including Random Forest classifiers, have demonstrated considerable success in skin cancer classification. These models are particularly valued for their ability to handle high-dimensional datasets and effectively distinguish between benign and malignant lesions (Javaid et al. (2021)). On the other hand, deep learning methods, especially Convolutional Neural Networks (CNNs), have transformed the field of medical image classification. CNNs excel in automatically learning complex features from dermoscopic images and have consistently achieved diagnostic accuracies that rival or even surpass those of experienced dermatologists (Kumari and Rattan (2023); Arooj et al. (2022)).

Within the domain of CNN architectures, ResNet (Residual Networks) has emerged as a noteworthy innovation. By introducing skip connections, ResNet addresses the challenges associated with vanishing gradients in deep networks, enabling efficient training and extraction of hierarchical features (He et al. (2015)). Research involving ResNet architectures has reported substantial improvements in the accuracy of dermoscopic image classification (Magdy et al. (2023)).

In recent years, emerging multimodal models, such as CLIP (Contrastive Language–Image Pre-training) (Radford et al. (2021)) and BLIP (Bootstrapped Language–Image Pre-training) (Li et al. (2022)), have introduced innovative methodologies for combining visual and textual data. These models provide an opportunity to integrate image analysis with contextual information, enhancing diagnostic interpretability. While their application in dermatology is still developing, their ability to leverage both dermoscopic images and textual metadata holds the potential to revolutionize workflows in skin cancer diagnostics (Hekler et al. (2019)).

In this study we evaluates the ability of Machine Learning Models skin lesions classification ability, including Random Forest, custom CNN architectures, and we also apply transfer learning based on ResNet. All the training and evaluations are through HAM10000 dataset (Tschandl et al. (2018)). Additionally, we also explores the potential of vision language models such as CLIP and BLIP that fine-tuned specifically for skin lesion classification. It highlights innovative methodologies to address critical challenges in the detection and classification of skin cancer.

2 Research Background and Related Work

The classification of skin cancer has been a focal point of medical research in recent years, with significant advancements in machine learning (ML) and deep learning (DL) technologies contributing to improvements in diagnostic accuracy and efficiency.

2.1 Machine Learning Approaches

Classical machine learning techniques such as Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN) have been commonly employed for skin cancer detection. These models rely on handcrafted features, such as color histograms and texture descriptors, to differentiate between benign and malignant lesions. For instance, Almutairi and Khan (2023) demonstrated that Random Forest classifiers were particularly effective in this domain, achieving strong performance based on extracted features.

Also the use of ensemble learning techniques has further improved the diagnostic capabilities of classical ML methods. For example, Ramya and Sathiyabhama (2023) successfully combined Random Forest with enhanced genetic algorithms, resulting in improved classification accuracy.

2.2 Deep Learning Advancements

Deep learning has revolutionized skin cancer diagnosis as it be able to automatic extract feature from image data that bypassing the need for manual feature engineering. Convolutional Neural Networks (CNNs), in particular, have emerged as the most effective architecture for this task. Architectures like ResNet, VGG, and DenseNet have consistently delivered high levels of accuracy in skin lesion classification. For example, Bechelli and Delhommelle (2022) reported that ResNet50 achieved classification accuracies of up to 88% when fine-tuned for skin tumor identification.

Recent innovations in deep learning also include the integration of attention mechanisms, which allow networks to focus on the most critical regions of an image. Aggarwal et al. (2019) proposed an attention-guided CNN model, achieving a 12% improvement in classification accuracy over traditional CNNs by highlighting lesion-specific features within dermoscopic images.

2.3 Hybrid and Ensemble Models

Hybrid approaches, such as those evaluated by Ahmed et al. (2023), integrate CNNs with classical models like SVM and k-NN to improve overall classification performance. In their study, hybrid models achieved an impressive accuracy of 99.5% for image-based tasks, demonstrating their potential for robust diagnostic applications.

Ensemble learning has also gained traction in the field. By aggregating predictions from multiple classifiers, ensemble methods enhanced reliability and robustness across diverse datasets. For instance, Avanija et al. (2023) developed an ensemble approach that achieved 86% accuracy by combining decisions from various models and ensure consistent performance across different lesion types.

2.4 Challenges and Future Directions

2.4.1 Challenge: Class Imbalance

One of the major challenges in skin cancer classification is the class imbalance within datasets, where malignant lesions are often significantly underrepresented compared to benign lesions. This imbalance could lead to biased model predictions and reduced sensitivity in detecting critical cases. To mitigate this issue, advanced techniques such as data augmentation, transfer learning, and synthetic data generation using Generative Adversarial Networks (GANs) have been explored. These approaches improved the diversity of training data such that enhanced the models' ability to generalize to underrepresented classes (Munir et al. (2019)).

2.4.2 Multimodal Learning with CLIP and BLIP

Recent advancements in multimodal learning have introduced models such as CLIP (Contrastive Language–Image Pre-training) and BLIP (Bootstrapped Language–Image Pre-training), which able to combine visual and textual data to generate richer representations for classification tasks. CLIP, introduced by Radford et al. (2021), aligns images and textual descriptions in a shared latent space, enabling zero-shot classification and generalizable representations. Studies such as Khandelwal et al. (2021) highlight CLIP's potential that improved diagnostic accuracy by leveraging textual metadata alongside dermoscopic images.

BLIP extends CLIP's capabilities by focusing on fine-tuned pretraining for specific tasks, including medical diagnostics. It combines image encoders with language models to effectively integrate visual-textual data. Li et al. (2023) and Chen et al. (2023) demonstrated BLIP's utility in medical imaging, showcasing its ability to analyze dermoscopic images alongside textual metadata such as lesion descriptions or clinical notes.

2.5 Possible Future Directions

Both CLIP and BLIP represent promising advancements in the field of skin cancer classification. While CLIP excels in aligning visual and textual data for generalization across diverse datasets. BLIP focuses on task-specific fine-tuning, making it highly possible to be effective for medical applications.

3 Dataset Overview

The dataset used in this study originates from the HAM10000 (Human Against Machine with 10000 training images) dataset, originally published by Tschandl et al. (2018) and available on Harvard Dataverse. We used a copy of this dataset, which structured images into folders by class. It was obtained from Kaggle at Mohammad (2024). The dataset consists of 10,015 dermoscopic images of skin lesions, each with a resolution of 600×450 pixels, and is categorized into seven classes:

- **Melanoma (MEL)**: A malignant tumor of melanocytes.
- **Melanocytic Nevi (NV)**: Benign proliferations of melanocytes.
- **Benign Keratosis-like lesions (BKL)**: A mixture of seborrheic keratoses, lichen planus-like keratoses, and solar lentigo.
- **Basal Cell Carcinoma (BCC)**: A common form of skin cancer arising in basal cells.
- **Vascular lesions (VASC)**: Blood vessel-related lesions, such as hemangiomas.
- **Actinic Keratoses and intraepithelial Carcinoma (AKIEC)**: Pre-cancerous or cancerous lesions caused by sun damage.
- **Dermatofibroma (DF)**: Benign fibrous histiocytomas of the skin.

The dataset is imbalanced, with the majority of images belonging to the **NV** class (6,705 images) and the minority being from the **DF** class (115 images) and **VASC** class (142 images). Figure 1 shows the frequency distribution of images across these seven classes, while Figure 2 presents sample images from each class.

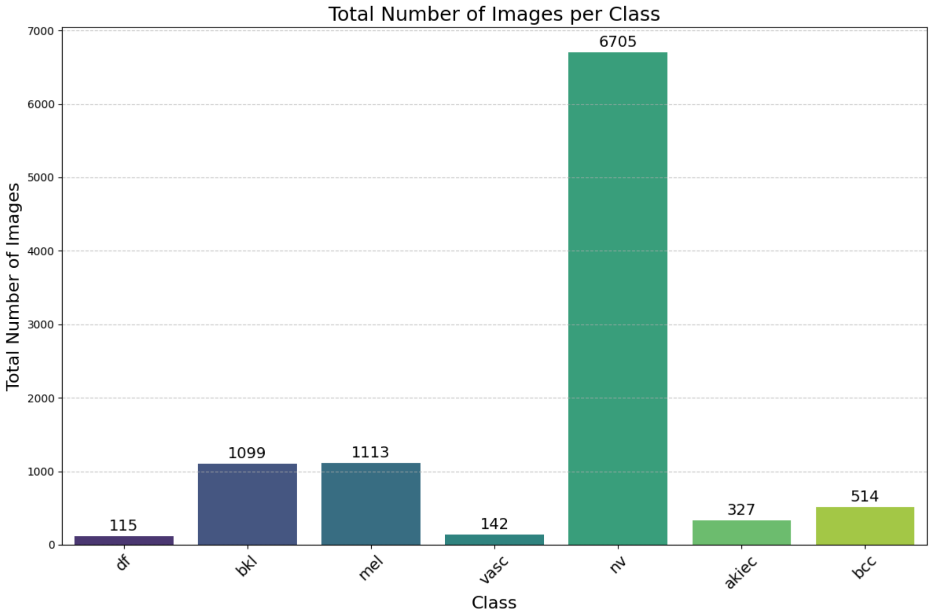


Figure 1: Frequency of images per class in the dataset.

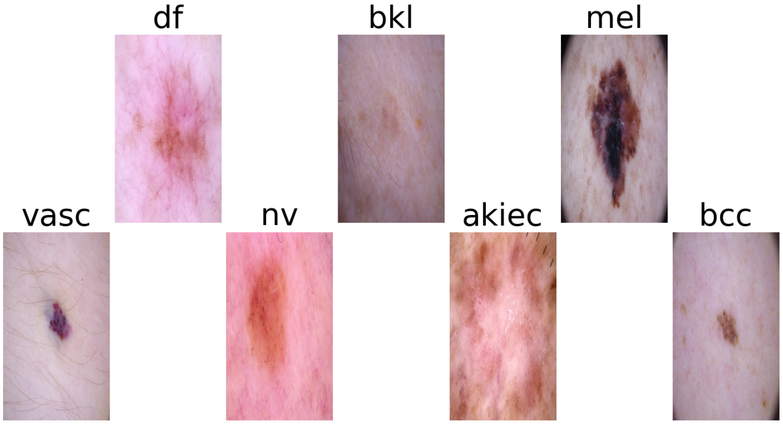


Figure 2: Representative images from each class in the dataset.

3.1 Dataset Splits

The dataset was divided into three subsets: **training** (75%), **validation** (10%), and **test** (15%)by simply sampling a fixed percentage of images from each class. Specifically, 75% of the images in each class were assigned to the training set, 10% to the validation set, and 15% to the test set.

This straightforward sampling method is commonly used in machine learning workflows when proportional representation by class is required (Joseph (2022)). The training set, comprising 75% of the data, provides a large pool of examples to expose the model to diverse patterns and characteristics. The validation set, 10% of the data, is used for hyperparameter tuning and monitoring overfitting during training, while the test set, 15% of the data, serves as the final, unbiased evaluation to assess the model's generalization on unseen data (Nguyen et al. (2021)).

Table 1: Number of Images per Class in Each Split

Class	**Train**	**Validation**	**Test**
DF	86	11	18
BKL	824	109	166
MEL	834	111	168
VASC	106	14	22
NV	5028	670	1007
AKIEC	245	32	50
BCC	385	51	78

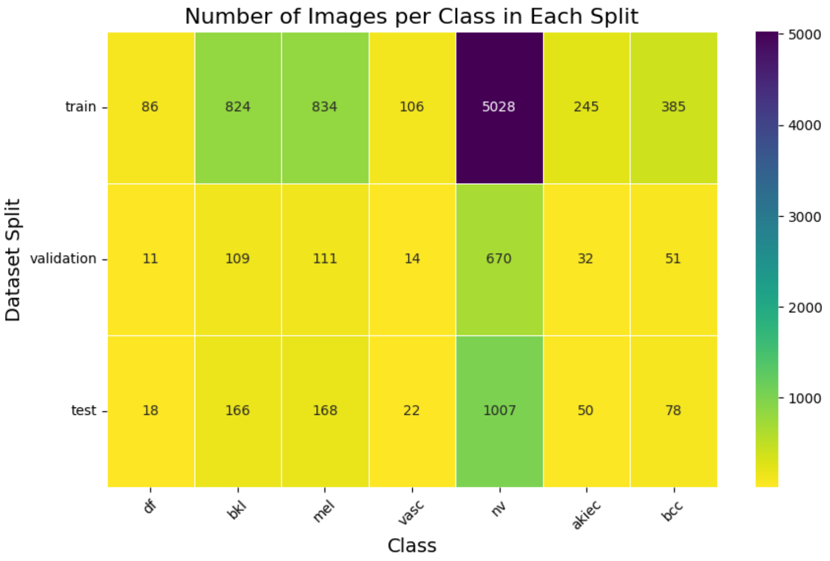


Figure 3: Heatmap showing the distribution of images across classes in each split.

This fixed-percentage sampling ensures a consistent number of samples from each class across splits. Figure 3 provides a heatmap of the distribution of images for each class in the training, validation, and test sets, illustrating the proportional representation achieved through the splitting process.

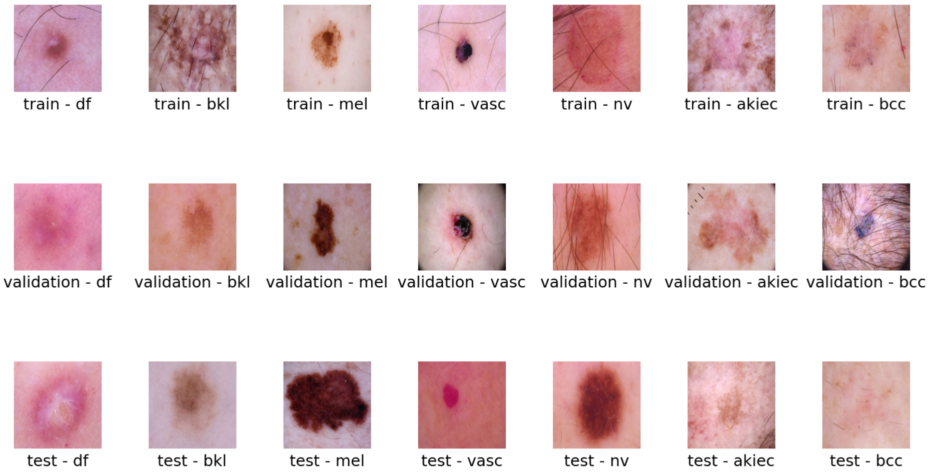


Figure 4: Representative images from the training, validation, and test splits for each class.

In addition, Figure 4 showcases representative images from each class across the training, validation, and test sets. These visual examples highlight the diversity within and between the splits, demonstrating the variety of features and patterns present in the dataset.

3.2 Pre-Processing and Feature Augmentation

In our endeavor to achieve robust skin cancer classification using deep learning techniques, data pre-processing plays a pivotal role in preparing the dataset for optimal model training. We apply a series of transformations and augmentations to enhance generalization and model robustness against varying conditions in dermoscopic images. This section details the mathematical formulations of these transformations and the overall augmentation pipeline.

3.2.1 Image Resizing

The original images in the dataset have varying resolutions. To ensure compatibility with deep learning architectures, particularly Convolutional Neural Networks (CNNs), all images are resized to 224×224 pixels. This standardization simplifies model training and aligns with the input requirements of pre-trained models such as ResNet and BLIP, which expect inputs of consistent dimensions for efficient feature extraction.

Formally, each image $\mathbf{I}$ is resized using a scaling function $\mathcal{S}$:

$$\mathbf{I}_{\text{resized}} = \mathcal{S}(\mathbf{I}, 224, 224) \tag{1}$$

3.2.2 Geometric Transformations

To increase the diversity of training data and improve model generalization, we apply several geometric augmentations. These augmentations simulate real-world variances in lesion presentation.

1. **Random Rotation**: Each image is rotated by an angle θ sampled uniformly from $[-90°, 90°]$.

$$\mathbf{I}_{\text{rotated}} = \mathcal{R}(\mathbf{I}_{\text{resized}}, \theta), \quad \theta \sim \mathcal{U}(-90°, 90°) \tag{2}$$

2. **Random Horizontal Flip**: With probability $p = 0.5$, we apply a horizontal flip to account for symmetry.

$$\mathbf{I}_{\text{flipped}} = \begin{cases} \mathcal{F}_h(\mathbf{I}_{\text{rotated}}), & \text{with probability } p = 0.5 \\ \mathbf{I}_{\text{rotated}}, & \text{otherwise} \end{cases} \tag{3}$$

3. **Random Affine Transformation**: We apply random translations to the image within a range of ±10% of the image dimensions.

$$\mathbf{I}_{\text{affine}} = \mathcal{A}(\mathbf{I}_{\text{flipped}}, t_x, t_y), \quad t_x, t_y \sim \mathcal{U}(-0.1, 0.1) \tag{4}$$

where t_x and t_y are the horizontal and vertical translation factors, respectively.

4. **Random Resized Crop**: We randomly crop a region of the image and resize it back to 224×224 pixels. The scale of the crop is sampled from $[0.8, 1.0]$.

$$\mathbf{I}_{\text{crop}} = \mathcal{C}(\mathbf{I}_{\text{affine}}, s), \quad s \sim \mathcal{U}(0.8, 1.0) \tag{5}$$

where $\mathcal{C}$ represents the cropping and resizing operation.

These transformations are sequentially applied to each image to enhance the model's ability to generalize across diverse clinical conditions.

3.2.3 Color Transformations

To simulate variations in lighting and skin tone, we apply color transformations using the `ColorJitter` function. This introduces random variations in brightness, contrast, saturation, and hue.

$$\mathbf{I}_{\text{color}} = \mathcal{J}(\mathbf{I}_{\text{crop}}, \delta_b, \delta_c, \delta_s, \delta_h) \tag{6}$$

where δ_b, δ_c, δ_s, and δ_h are random factors sampled from the ranges:

$$\delta_b, \delta_c, \delta_s \sim \mathcal{U}(0.8, 1.2), \quad \delta_h \sim \mathcal{U}(-0.05, 0.05) \tag{7}$$

These adjustments help the model learn robust features that are invariant to changes in lighting conditions.

3.2.4 Normalization

After color transformations, we convert the image to a tensor and normalize it to standardize the pixel values. Normalization is performed using the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the training dataset, computed over each channel (R, G, B).

$$\mathbf{I}_{\text{norm}} = \frac{\mathbf{I}_{\text{tensor}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{8}$$

This operation ensures that the pixel values have zero mean and unit variance, which facilitates faster convergence during training.

3.2.5 Noise Addition

To enhance the model's robustness against noisy inputs, we add Gaussian noise to the images with a probability of $p = 0.5$. The noise $\boldsymbol{\eta}$ is sampled from a normal distribution with zero mean and variance σ^2.

$$\mathbf{I}_{\text{noisy}} = \text{clamp}(\mathbf{I}_{\text{norm}} + \boldsymbol{\eta}, 0, 1), \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2) \tag{9}$$

In our implementation, we set $\sigma = 0.01$. The `clamp` function ensures that the pixel values remain within the valid range $[0, 1]$.

3.2.6 Algorithm for Pre-Processing and Feature Augmentation

The pre-processing and augmentation pipeline is formalized in Algorithm 1. This algorithm sequentially applies the transformations to each image in the dataset.

Algorithm 1 Pre-Processing and Augmentation Pipeline

```
Require: Dataset $\mathcal{D} = \{\mathbf{I}_i\}_{i=1}^{N}$
Ensure: Augmented dataset $\mathcal{D}'$
for all $\mathbf{I} \in \mathcal{D}$ do
    Resize image:
        $\mathbf{I} \leftarrow \mathcal{S}(\mathbf{I}, 224, 224)$
    Random rotation:
        $\theta \sim \mathcal{U}(-90^\circ, 90^\circ)$
        $\mathbf{I} \leftarrow \mathcal{R}(\mathbf{I}, \theta)$
    Random horizontal flip (with $p = 0.5$):
        $\mathbf{I} \leftarrow \mathcal{F}_h(\mathbf{I})$
    Random affine transform:
        $t_x, t_y \sim \mathcal{U}(-0.1, 0.1)$
        $\mathbf{I} \leftarrow \mathcal{A}(\mathbf{I}, t_x, t_y)$
    Random resized crop:
        $s \sim \mathcal{U}(0.8, 1.0)$
        $\mathbf{I} \leftarrow \mathcal{C}(\mathbf{I}, s)$
    Color jitter:
        $\delta_b, \delta_c, \delta_s \sim \mathcal{U}(0.8, 1.2)$, $\delta_h \sim \mathcal{U}(-0.05, 0.05)$
        $\mathbf{I} \leftarrow \mathcal{J}(\mathbf{I}, \delta_b, \delta_c, \delta_s, \delta_h)$
    Convert to tensor:
        $\mathbf{I} \leftarrow \mathrm{ToTensor}(\mathbf{I})$
    if With probability $p = 0.5$ then
        Add Gaussian noise:
            $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.01$
            $\mathbf{I} \leftarrow \mathrm{clamp}(\mathbf{I} + \boldsymbol{\eta}, 0, 1)$
    end if
    Normalize image:
        $\mathbf{I} \leftarrow \dfrac{\mathbf{I} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$
    Add augmented image $\mathbf{I}$ to $\mathcal{D}'$
end for
return $\mathcal{D}'$
```

This pipeline ensures the creation of a diverse and standardized dataset, allowing the model to generalize better to unseen data by simulating real-world variations and imperfections.

4 Methodology

4.1 Random Forest (RF)

Random Forest (RF) is an traditional machine learning algorithm that constructs multiple decision trees during training and aggregates their outputs to predict new data input. This approach has proven particularly effective for image classification tasks, as it can manage noisy datasets and complex feature interactions (Lowe and Kulkarni (2015)).

The RF algorithm generates each decision tree using bootstrap sampling, and for any given input $\mathbf{x}$, the final classification is determined by majority voting among the trees:

$$\hat{Y}_{\text{RF-classification}}(\mathbf{x}) = \arg\max_{k \in \mathcal{K}} \sum_{i=1}^{n} \delta(Y_i(\mathbf{x}) - k), \tag{10}$$

where $Y_i(\mathbf{x})$ represents the prediction from the i-th tree, $\mathcal{K}$ denotes the set of all possible classes, and δ is the Kronecker delta function, which equals 1 if the predicted label matches class k, and 0 otherwise. The $\arg\max$ operation identifies the class receiving the highest number of votes across the trees.

In this study, the RF classifier was tailored for skin lesion classification using a combination of Histogram of Oriented Gradients (HOG) to extract features. We used HOG to captures essential edge and gradient-based structures to offer robust texture and shape representations (Vo et al. (2015), Zhou et al. (2020)).

To mitigate computational overhead from the high-dimensional feature vectors, Principal Component Analysis (PCA) was employed for dimensionality reduction (Jolliffe and Cadima (2016)). We used PCA to help retain the most relevant features while improving computational efficiency and reducing the risk of overfitting.

As the dataset is imbalance, we apply the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. (2002)) to create synthetic samples for underrepresented classes. In order to improve the balance of training set. SMOTE interpolates between existing samples to produce new examples. Thereby addressing biases toward majority classes and improving sensitivity for rare lesion types.

4.2 Custom Convolutional Neural Network (CNN) with Residual Blocks

Our custom CNN architecture was inspired by ResNet (He et al. (2015)). As model was developed to handle the challenges of skin lesion classification, we believe Residual blocks will be helpful for our problem. As it mitigate the vanishing gradient problem by introducing skip connections, allowing identity mappings to be learned alongside complex transformations. This can efficient training of deeper networks which is crucial due to the diverse and complex patterns found in dermoscopic images (Zhang et al. (2021), Saini and Rawat (2022)).

As illustrated in Figure 5, within each residual block, the output $\mathbf{y}$ is computed as:

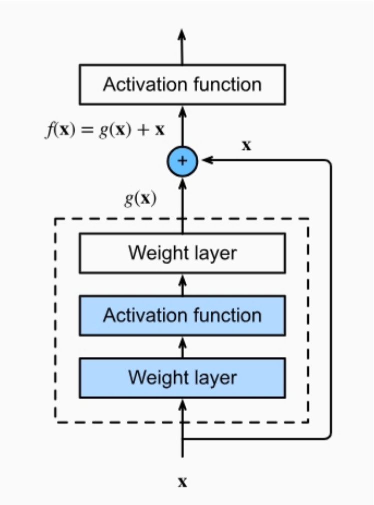


Figure 5: Residual Block. The skip connection adds the input $\mathbf{x}$ to the residual mapping $\mathcal{F}(\mathbf{x})$, ensuring efficient gradient flow.

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x}, \tag{11}$$

where $\mathbf{x}$ is the input, $\mathcal{F}(\mathbf{x})$ represents the residual mapping (consisting of two convolutional layers, batch normalization, and ReLU activations), and $\mathbf{y}$ is the final output. If the input and output dimensions differ due to changes in filter size or stride, a 1×1 convolution is applied to align dimensions (Lan et al. (2019), Zhang et al. (2021)).

4.3 Transfer Learning with ResNet

We also study on transfer learning to enhance performance that we employed the ResNet-50 architecture pre-trained on ImageNet (He et al. (2015)). Transfer learning leverages pre-trained weights to enable the network to learn domain-specific features with minimal labeled data. This significantly reduces training time and data to improve model generalization, particularly in medical imaging tasks (Liu et al. (2019)).

For this study, the ResNet-50 model was fine-tuned by replacing the final fully connected layer with a new layer tailored for the seven lesion classes. The network retained its convolutional layers, which were capable of extracting hierarchical features.

4.4 Fine-Tuning Vision-Language Models

Two advanced vision-language models, CLIP and BLIP, were fine-tuned for skin lesion classification.

4.4.1 Fine-Tuning CLIP

CLIP, developed by OpenAI, aligns visual and textual data in a shared latent space, facilitating zero-shot learning and robust transfer to downstream tasks (Radford et al. (2021)). In this study, the vision encoder (ViT-B/32) was fine-tuned while the text encoder remained frozen, preserving its pre-trained capabilities. A classification head was added to predict lesion classes.

4.4.2 Fine-Tuning BLIP

BLIP extends the functionality of CLIP by focusing on fine-grained image-text alignment (Li et al. (2023)). Same as CLIP, only the vision encoder was fine-tuned, with the textual components frozen. A lightweight classification head mapped visual embeddings to lesion classes.

5 Results and Evaluation

5.1 Performance Metrics

To evaluate the performance of our models, we applied widely used classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's performance across the whole dataset. The formulas used for these metrics are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. These metrics collectively assess not only the overall accuracy of the models but also their capability to correctly identify each class, particularly in imbalanced datasets like HAM10000.

5.2 Test Results

The performance of the models on the test split is summarized in Table 2. The evaluation metrics include accuracy, precision, recall, and F1-score.

Table 2: Performance Metrics on the Test Split

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.6740	0.5520	0.6740	0.5755
ResNet Transfer Learning	0.8516	0.8612	0.8516	0.8485
Custom CNN	0.7614	0.7324	0.7614	0.7433
Fine-Tuned CLIP	0.8496	0.8487	0.8496	0.8464
Fine-Tuned BLIP	0.8323	0.8202	0.8323	0.8203

In order to visualize the classification performance, confusion matrices for each model are provided in Figures 6, 7, and 8. These matrices offer detailed insights into the wrong classifications and true classifications across different skin lesion classes.

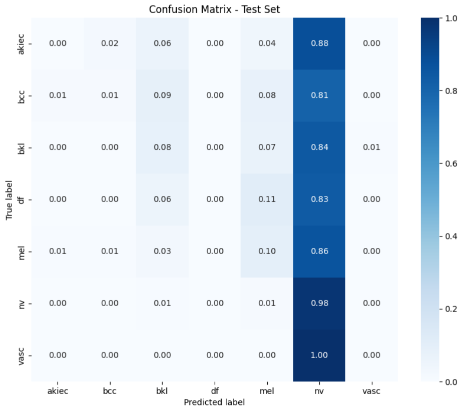
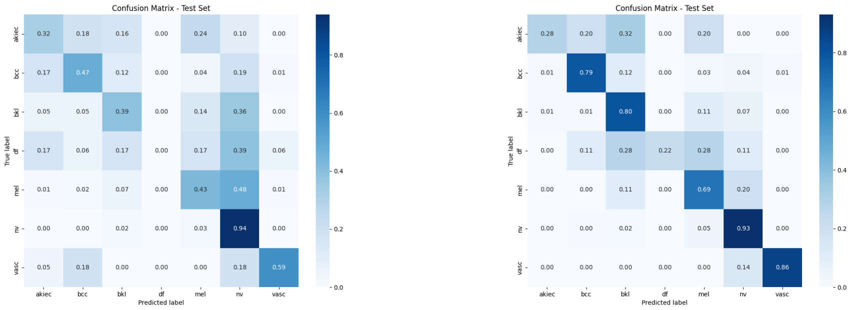


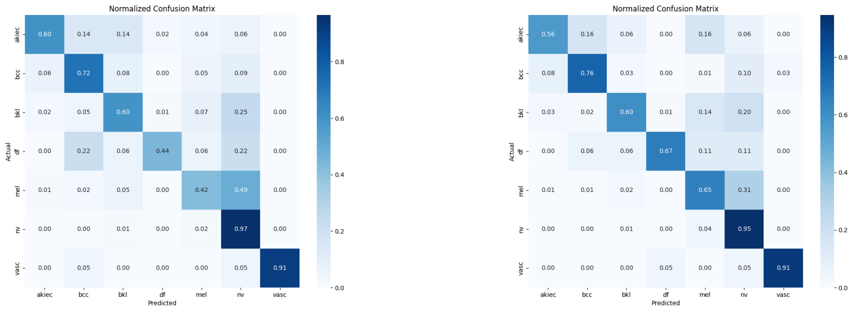
Figure 6: Confusion Matrix for Random Forest Model.



(a) Custom CNN

(b) ResNet Transfer Learning

Figure 7: Confusion Matrices for Custom CNN and ResNet Transfer Learning Models.



(a) Fine-tuned BLIP

(b) Fine-tuned CLIP

Figure 8: Confusion Matrices for Fine-tuned BLIP and CLIP Models.

Among the evaluated models, **ResNet with Transfer Learning** was the top performer, achieving the highest accuracy (0.8516) and F1-score (0.8485). This shows that using pre-trained weights from ImageNet helped ResNet adapt well to skin lesion images, enabling strong feature extraction and classification. The high precision (0.8612) also indicates its ability to reduce false positives, which is critical for medical diagnostics. The confusion matrix (Figure 7(b)) shows it performed well across all classes, especially with high recall for the **NV** and **BCC** classes. However, there were still misclassifications between similar classes like **MEL** and **BKL**, and between **AKIEC** and **BCC**. This highlights the difficulty in distinguishing between certain lesion types, even for advanced models.

The **Fine-Tuned CLIP** model performed well, with an accuracy of 0.8496 and an F1-score of 0.8464, which is close to the ResNet model's performance. The confusion matrix for CLIP (Figure 8(b)) shows that it identified the **NV** and **VASC** classes accurately. CLIP's use of both visual and textual data helps it capture complex features. However, there were still some confusions between classes like **AKIEC** and **MEL**, likely because they look similar in certain images.

The **Fine-Tuned BLIP** model had slightly lower performance, with an accuracy of 0.8323 and an F1-score of 0.8203. The confusion matrix (Figure 8(a)) shows that BLIP has high recall for classes like **NV** and **BCC**, which helps reduce false negatives. However, this comes with a cost—more frequent misclassifications for rare classes like **DF**. This suggests that while BLIP is good at identifying true positives, especially for common classes, it struggles with maintaining precision for less common categories.

The **Custom CNN** model had moderate performance, with an accuracy of 0.7614 and an F1-score of 0.7433. The confusion matrix (Figure 7(a)) shows that this model struggles particularly with differentiating between classes such as **MEL** and **BKL**, as well as **AKIEC** and **BCC**. Additionally, the recall for the **DF** class was notably low, even reaching 0. This seems to be influenced by the model placing disproportionately high emphasis on the **NV** class, which may have skewed its ability to generalize across less frequent classes. The lack of pre-trained weights limits the model's ability to generalize, particularly on a relatively small dataset. Despite these challenges, the model achieved reasonable recall values for certain classes, indicating that it can effectively identify most true positives, though its feature extraction is not as strong as models using transfer learning.

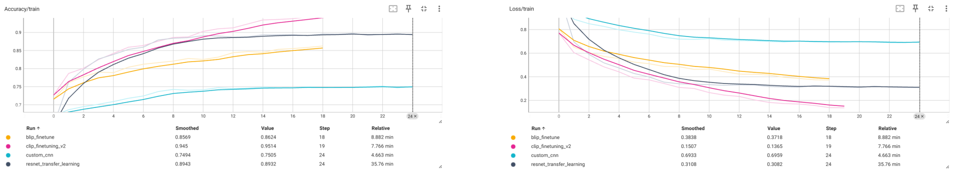
Finally, the **Random Forest** model exhibited the lowest performance, with an accuracy of 0.6740 and F1-score of 0.5755. The confusion matrix for Random Forest (Figure 6) clearly illustrates its limitations, showing particularly poor recall for the **VASC**, **AKIEC**, and **DF** classes—all of which have recall values of zero. Moreover, the recall for most other classes, apart from **NV**, is also close to zero. This means the model is largely unable to correctly identify instances of these classes. Its simpler structure and reliance on handcrafted features severely limit its ability to capture the complex visual patterns present in dermoscopic images.

For instance, while **NV**, being the most common class, is recognized to some extent, it is still often confused with other types like **MEL** and **BKL**. This demonstrates the Random Forest model's inability to effectively capture the nuanced differences between lesion types. These shortcomings highlight the superiority of deep learning-based approaches, which can automatically learn complex representations from raw images and achieve far better performance, particularly in distinguishing between visually similar classes.

6 Performance of Vision-Language Models and Convolutional Neural Networks

In this section, we present the performance metrics of different models, including training and validation accuracies and losses, as shown in

Figures 9 and 10.



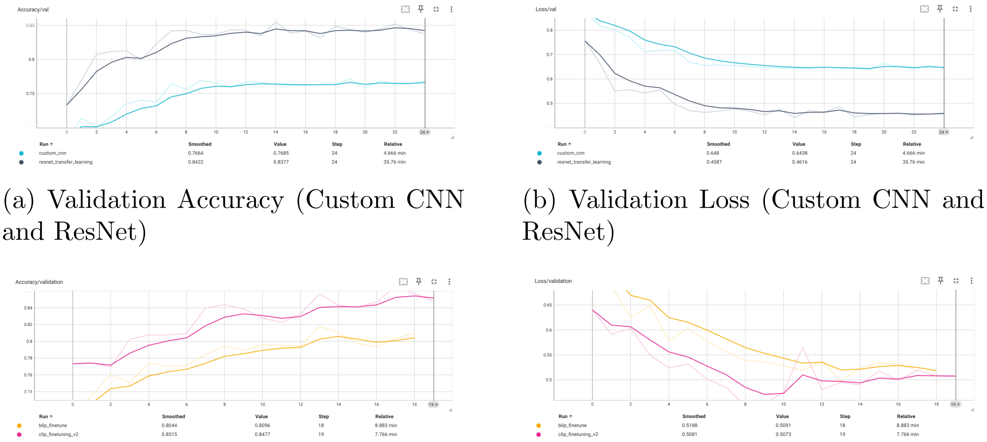
	Run ↑	Smoothed	Value	Step	Relative
●	blip_finetune	0.8569	0.8624	18	8.882 min
●	clip_finetuning_v2	0.945	0.9514	19	7.766 min
●	custom_cnn	0.7494	0.7505	24	4.663 min
●	resnet_transfer_learning	0.8943	0.8932	24	35.76 min

(a) Training Accuracy

	Run ↑	Smoothed	Value	Step	Relative
●	blip_finetune	0.3838	0.3718	18	8.882 min
●	clip_finetuning_v2	0.1507	0.1365	19	7.766 min
●	custom_cnn	0.6933	0.6959	24	4.663 min
●	resnet_transfer_learning	0.3108	0.3082	24	35.76 min

(b) Training Loss

Figure 9: Training Metrics for Different Models: Training accuracy and training loss curves for BLIP, CLIP, Custom CNN, and ResNet Transfer Learning.



	Run ↑	Smoothed	Value	Step	Relative
●	custom_cnn	0.7664	0.7685	24	4.666 min
●	resnet_transfer_learning	0.8422	0.8377	24	35.76 min

(a) Validation Accuracy (Custom CNN and ResNet)

	Run ↑	Smoothed	Value	Step	Relative
●	custom_cnn	0.648	0.6438	24	4.666 min
●	resnet_transfer_learning	0.4587	0.4616	24	35.76 min

(b) Validation Loss (Custom CNN and ResNet)

	Run ↑	Smoothed	Value	Step	Relative
●	blip_finetune	0.8044	0.8096	18	8.883 min
●	clip_finetuning_v2	0.8515	0.8477	19	7.766 min

(c) Validation Accuracy (CLIP and BLIP)

	Run ↑	Smoothed	Value	Step	Relative
●	blip_finetune	0.5188	0.5091	18	8.883 min
●	clip_finetuning_v2	0.5081	0.5073	19	7.766 min

(d) Validation Loss (CLIP and BLIP)

Figure 10: Validation Metrics for Vision-Language Models and Convolutional Neural Networks: Validation accuracy and validation loss curves for Custom CNN, ResNet Transfer Learning, CLIP, and BLIP.

6.1 Training Metrics Analysis

Figure 9 shows the training accuracy and training loss for BLIP, CLIP, Custom CNN, and ResNet Transfer Learning. It is evident that both vision-language models (BLIP and CLIP) reach a higher training accuracy compared to the CNN-based models (Custom CNN and ResNet).

Specifically, the CLIP model quickly converges with a smooth reduction in training loss, reaching an accuracy close to 95%. This indicates that the model benefits from its pre-trained multimodal representations, which accelerate convergence.

ResNet Transfer Learning also performs well, achieving an accuracy of around 89% while keeping a relatively low training loss. This demonstrates the strength of transfer learning in utilizing pre-trained weights from ImageNet for skin lesion classification. In contrast, Custom CNN has the lowest training accuracy and exhibits a much slower convergence rate, highlighting the difficulty of training a deep CNN from scratch on a relatively small dataset.

6.2 Validation Metrics Analysis

Figure 10 compares the validation metrics between CNN-based models (Custom CNN and ResNet) and vision-language models (CLIP and BLIP).

CNN-Based Models (Custom CNN and ResNet) The validation metrics (Figures 10(a) and 10(b)) show that ResNet achieves the highest validation accuracy among CNN-based models, reaching approximately 84%. The validation loss curve for ResNet also decreases smoothly, suggesting that the model generalizes well on the validation dataset split. Custom CNN, however, performs worse in both validation accuracy and loss, reinforcing the observation that the model's lack of pre-trained features affects its generalization ability.

Vision-Language Models (CLIP and BLIP) The CLIP model demonstrates the best validation accuracy, as seen in Figure 10(c), reaching around 85%. Its validation loss (Figure 10(d)) remains consistently lower compared to BLIP, indicating that it is not overfitting as much as BLIP. BLIP shows slightly lower performance, which may indicate that its architecture is not as well-suited for distinguishing fine-grained features in dermoscopic images as CLIP. Still, BLIP maintains a competitive validation accuracy, proving that its multimodal capabilities provide robust sensitivity.

7 Discussion

This study investigates the use of advanced machine learning models, including Random Forest, Custom CNNs, ResNet with transfer learning, and vision-language models (CLIP and BLIP), for classifying skin lesions in the HAM10000 dataset. Among these, the integration of vision-language models marks a significant contribution, as their application to medical image classification tasks remains a relatively unexplored area.

Our findings reveal that ResNet with transfer learning achieved the highest overall accuracy **85.16%** and F1-score **84.85%**, demonstrating the effectiveness of pre-trained models in adapting to domain-specific tasks. ResNet's ability to extract hierarchical features through its deep architecture made it particularly suited for the diverse and complex visual patterns in dermoscopic images. Vision-language models, particularly CLIP, also delivered competitive performance, with an accuracy of **84.96%** and F1-score of **84.64%**. CLIP's ability to leverage both visual and textual data enabled robust feature extraction and better generalization across lesion types.

BLIP, while slightly behind CLIP, achieved promising results, highlighting the potential of fine-tuned multimodal pre-training for medical image tasks. These models have shown that vision-language frameworks, traditionally used for broader computer vision tasks, can be effectively adapted for highly specialized domains like medical imaging. Custom CNNs, while achieving reasonable performance, were limited by their lack of pre-trained features, which hindered their ability to generalize on a relatively small and imbalanced dataset. Random Forest, as expected, performed the worst due to its inability to handle the complex feature space of dermoscopic images effectively.

The key contribution of this work is the exploration and application of vision-language models like CLIP and BLIP to a medical image classification problem. This study demonstrates that these models can rival and, in some cases, complement traditional deep learning architectures in terms of classification performance. Additionally, the results suggest that vision-language models may provide new pathways for incorporating multimodal data, such as clinical notes or textual metadata, into skin lesion diagnostics.

8 Limitations

This study has several limitations. First, the class imbalance in the HAM10000 dataset, particularly the underrepresentation of certain classes such as Dermatofibroma (DF), was not fully addressed beyond basic data augmentation. This likely impacted the models' sensitivity to minority classes. Second, the dataset size, though significant for medical imaging, is relatively small for training complex models like BLIP and CLIP, limiting their full potential. Lastly, no specific optimization for computational efficiency was performed, making some models resource-intensive and less practical for deployment in low-resource settings.

9 Future Work

Future work can explore the integration of vision-language models with medical-specific large language models to enhance performance by utilizing both clinical metadata and image features. Expanding the dataset to include more multimodal data or using advanced techniques to address class imbalance, such as synthetic data generation, could further improve model generalization. Additionally, optimizing the architecture of vision-language models for medical imaging tasks would enhance their practical applicability, particularly in resource-limited environments.

10 Conclusion

This study highlights the potential of advanced machine learning and vision-language models, such as ResNet, CLIP, and BLIP, in skin lesion classification. ResNet with transfer learning demonstrated the highest accuracy, while CLIP offered competitive performance by leveraging multimodal data. These findings underscore the promise of vision-language models in medical diagnostics, paving the way for future research integrating expanded datasets and optimized architectures to address challenges like class imbalance and dataset size.

References

Aggarwal, A., Das, N., & Sreedevi, I. (2019, November). Attention-guided deep convolutional neural networks for skin cancer classification. In *2019 ninth international conference on image processing theory, tools and applications (ipta)* (pp. 1–6). Istanbul, Turkey: IEEE. DOI: 10.1109/IPTA.2019.8936100

Ahmed, S. G., Zeng, F., Alrifaey, M., & Ahmadipour, M. (2023, October). Skin cancer classification utilizing a hybrid model of machine learning models trained on dermoscopic images. In *2023 3rd international conference on emerging smart technologies and applications (esmarta)* (pp. 1–7). Taiz, Yemen: IEEE. DOI: 10.1109/eSmarTA59349.2023.10293619

Almutairi, A., & Khan, R. U. (2023, June). Image-based classical features and machine learning analysis of skin cancer instances. *Applied Sciences*, *13*(13), 7712. DOI: 10.3390/app13137712

Arooj, S., Khan, M. F., Khan, M. A., Khan, M. S., & Taleb, N. (2022, February). Machine learning models for the classification of skin cancer. In *2022 international conference on business analytics for technology and security (icbats)* (pp. 1–8). Dubai, United Arab Emirates: IEEE. DOI: 10.1109/ICBATS54253.2022.9759054

Avanija, J., Chandra Mohan Reddy, C., Sri Chandan Reddy, C., Harshavardhan Reddy, D., Narasimhulu, T., & Hardhik, N. V. (2023, June). Skin cancer detection using ensemble learning. In *2023 international conference on sustainable computing and smart systems (icscss)* (pp. 184–189). Coimbatore, India: IEEE. DOI: 10.1109/ICSCSS57650.2023.10169747

Bechelli, S., & Delhommelle, J. (2022, February). Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. *Bioengineering*, *9*(3), 97. DOI: 10.3390/bioengineering9030097

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. DOI: 10.1613/jair.953

Chen, Q., Hu, X., Wang, Z., & Hong, Y. (2023, May). *Medblip: Bootstrapping language-image pre-training from 3d medical images and texts* (No. arXiv:2305.10799). arXiv. DOI: 10.48550/arXiv.2305.10799

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). *Deep residual learning for image recognition* (No. arXiv:1512.03385). arXiv. DOI: 10.48550/arXiv.1512.03385

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., ... Thiem, A. (2019, October). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, *120*, 114–121. DOI: 10.1016/j.ejca.2019.07.019

Javaid, A., Sadiq, M., & Akram, F. (2021, January). Skin cancer classification using image processing and machine learning. In *2021 international bhurban conference on applied sciences and technologies (ibcast)* (pp. 439–444). Islamabad, Pakistan: IEEE. DOI: 10.1109/IBCAST51254.2021.9393198

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*, 531 - 538. DOI: 10.1002/sam.11583

Khandelwal, A., Weihs, L., Mottaghi, R., & Kembhavi, A. (2021). Simple but effective: Clip embeddings for embodied ai. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14809-14818. DOI: 10.1109/CVPR52688.2022.01441

Kumari, A., & Rattan, D. P. (2023, November). Skin cancer detection and classification using deep learning methods. *International Journal of Electrical and Electronics Research*, *11*(4), 1072–1086. DOI: 10.37391/ijeer.110427

Lan, R., Zou, H., Pang, C., Zhong, Y., Liu, Z., & Luo, X. (2019). Image denoising via deep residual convolutional neural networks. *Signal, Image and Video Processing*, *15*, 1-8. DOI: 10.1007/S11760-019-01537-X

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, June). *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models* (No. arXiv:2301.12597). arXiv. DOI: 10.48550/arXiv.2301.12597

Li, J., Li, D., Xiong, C., & Hoi, S. (2022, February). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation* (No. arXiv:2201.12086). arXiv. DOI: 10.48550/arXiv.2201.12086

Liu, S., Tian, G., & Xu, Y. (2019). A novel scene classification model combining resnet based transfer learning and data augmentation with a filter. *Neurocomputing*, *338*, 191-206.

DOI: 10.1016/j.neucom.2019.01.090

Lowe, B., & Kulkarni, A. (2015). Multispectral image analysis using random forest. *International Journal of Soft Computing*, *6*, 1-14. DOI: 10.5121/IJSC.2015.6101

Magdy, A., Hussein, H., Abdel-Kader, R. F., & Salam, K. A. E. (2023). Performance enhancement of skin cancer classification using computer vision. *IEEE Access*, *11*, 72120–72133. DOI: 10.1109/ACCESS.2023.3294974

Mohammad, M. H. M. (2024). *Skin cancer image dataset.* Retrieved from https://www.kaggle.com/datasets/rauf41/skin-cancer-image-dataset

Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019, August). Cancer diagnosis using deep learning: A bibliographic review. *Cancers*, *11*(9), 1235. DOI: 10.3390/cancers11091235

Nguyen, Q., Ly, H., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V., . . . Pham, B. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, 1-15. DOI: 10.1155/2021/4832864

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021, February). *Learning transferable visual models from natural language supervision* (No. arXiv:2103.00020). arXiv. DOI: 10.48550/arXiv.2103.00020

Ramya, P., & Sathiyabhama, B. (2023). Skin cancer prediction using enhanced genetic algorithm with extreme learning machine. *Journal of Trends in Computer Science and Smart Technology*, *5*(1), 1–13.

Saini, S. S., & Rawat, P. (2022). Deep residual network for image recognition. *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 1-4. DOI: 10.1109/icdcece53908.2022.9792645

Surendren, D., & Sumitha, J. (2023, August). Machine learning algorithms for skin cancer diagnosis: Comparative analysis. In *2023 5th international conference on inventive research in computing applications (icirca)* (p. 608–613). Coimbatore, India: IEEE. Retrieved from https://ieeexplore.ieee.org/document/10220845/ DOI: 10.1109/ICIRCA57980.2023.10220845

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.

DOI: 10.1038/sdata.2018.161

Urban, K., Mehrmal, S., Uppal, P., Giesey, R. L., & Delost, G. R. (2021, January). The global burden of skin cancer: A longitudinal analysis from the global burden of disease study, 1990–2017. , *2*, 98–108. DOI: 10.1016/j.jdin.2020.10.013

Vo, T., Tran, D., & Ma, W. (2015). Tensor decomposition and application in image classification with histogram of oriented gradients. *Neurocomputing*, *165*, 38-45. DOI: 10.1016/j.neucom.2014.06.093

Zhang, J., Zhu, Y., Li, W., Fu, W., & Cao, L. (2021). Drnet: A deep neural network with multi-layer residual blocks improves image denoising. *IEEE Access*, *9*, 79936-79946. DOI: 10.1109/ACCESS.2021.3084951

Zhou, W., Gao, S., Zhang, L., & Lou, X. (2020). Histogram of oriented gradients feature extraction from raw bayer pattern images. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *67*, 946-950. DOI: 10.1109/TCSII.2020.2980557