Skin Lesion Detection and Classification Using Deep Learning

Singh, Akshita Ravinder | singh.akshi@northeastern.edu
Dewangan, Sachin | @northeastern.edu
Singh, Utkarsh | singh.utka@northeastern.edu

ABSTRACT

This project explores the application of deep learning techniques for skin lesion detection and classification. By utilizing convolutional neural networks (CNNs) trained on medical images, we aim to develop automated systems capable of identifying various types of skin abnormalities. Such advancements hold promise for improving early detection and treatment planning for conditions like skin cancer. We'll delve into the methods used in training these networks, discuss challenges, and explore future directions for research in this area.

INTRODUCTION

The objective is to address the limitation of small and homogeneous datasets in training neural networks for automated diagnosis of pigmented skin lesions by releasing the HAM10000 dataset, which comprises diverse dermatoscopic images from different populations and acquisition modalities, thereby facilitating more robust machine learning models for accurate diagnosis.

BACKGROUND

The paper about "Deep Learning for Medical Image Analysis" explores the impact of deep learning on medical image analysis, emphasizing its role in tasks like segmentation, cancer detection, and diagnosis. The resurgence of deep convolutional neural networks, coupled with computational resources, has significantly advanced diagnostic precision. The paper surveys research on convolutional neural networks, pretrained models, and generative adversarial networks in medical imaging. It highlights the challenge of obtaining large medical image datasets and emphasizes the use of transfer learning to address this issue. The paper concludes by noting trends, such as modifying pretrained models and utilizing GANs to enhance segmentation accuracy, to overcome challenges in medical image analysis.

The paper "Intelligent fusion-assisted skin lesion localization and classification for smart healthcare" presents a cutting-edge fusion approach for skin lesion segmentation and classification. Utilizing a hybrid CNN framework and novel fusion techniques, high accuracy is achieved: 92.06% to 92.70% for segmentation and 86.4% to 87.02% for classification across multiple datasets, surpassing existing methods. This underscores the efficacy of the fusion and selection techniques employed for enhanced skin lesion analysis.

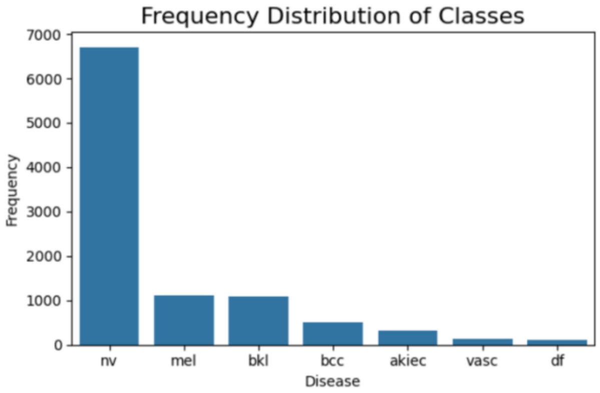
DATASET

The HAM10000 dataset, publicly available since 2018, comprises 10,000 high-resolution images labeled with seven types of skin lesions. Accompanying metadata includes diagnosis, confirmation method, patient demographics, and lesion location. Approximately 50% of the dataset has histopathological confirmation, while the rest is validated through follow-up exams, expert consensus, or confocal microscopy. This dataset serves as a crucial resource for developing machine learning models to classify skin lesions accurately, aiding in dermatological research and medical diagnosis.

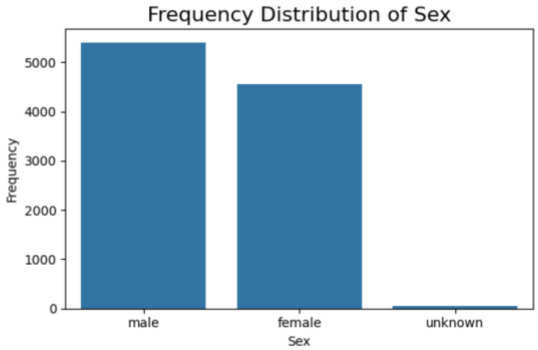
APPROACH

Exploratory Data Analysis -

The class distribution in this graph reveals a significant imbalance. Melanocytic nevi (nv) are the most frequent class, followed by Melanoma (mel), Benign keratosis-like lesions (bkl), and others. This imbalance can bias machine learning models towards the majority class (nv) and underestimate the presence of Melanoma (mel), the class of primary interest. To achieve balanced representation and improve the model's performance across all classes, we can employ data sampling techniques. Here, two main approaches exist: oversampling and under sampling. Oversampling increases the number of minority class (e.g., Melanoma) data points. This can be achieved through data augmentation techniques, which generate synthetic variations of existing minority class data. Under sampling, on the other hand, reduces the number of majority class (e.g., Melanocytic nevi) data points.

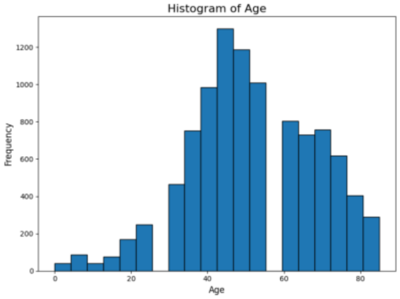


The second graph reveals a balanced distribution of male and female data points within the dataset. This balance is crucial for mitigating gender bias in the model's performance. When a dataset is skewed towards one gender, the model can become overly reliant on patterns specific to that gender. This can lead to inaccurate predictions for the underrepresented gender.



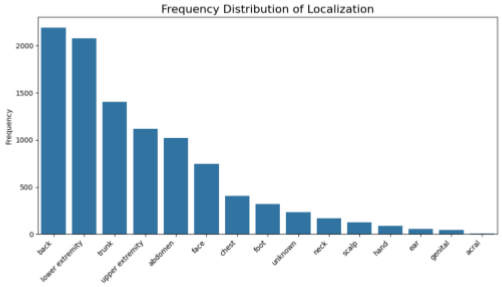
A balanced distribution, like the one observed here, allows the model to learn from a broader range of data and make fairer predictions for both males and females.

The age distribution histogram provides valuable insights into the demographics of the patients within the dataset. It reveals a concentration of data points between 40 and 60 years old, suggesting a higher prevalence of skin conditions in this age group with the body part ’back’ being the most common place for the disease to occur. This finding aligns with established medical knowledge, as several skin diseases become more common with age due to factors like decreased skin elasticity, weakened immune system, sun exposure etc.,



Understanding this age distribution can inform the development and application of the model. For instance, the model might require adjustments to optimize performance for patients in this specific age range.

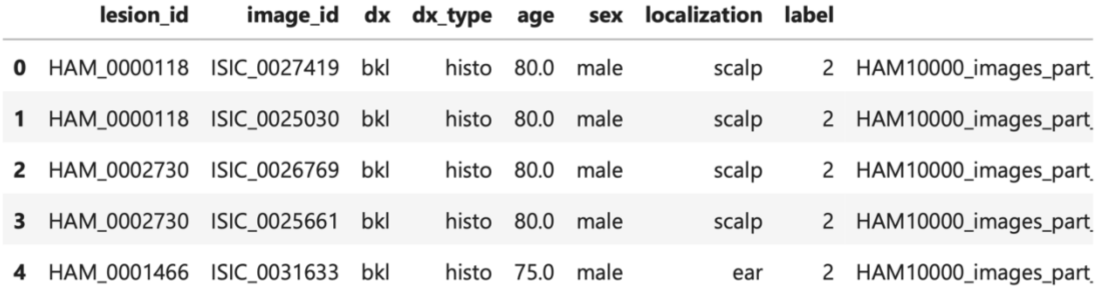
The frequency distribution plot provides insights into the distribution of cancer occurrences across various anatomical regions. It reveals that a substantial number of cases were detected on the back and lower extremities, with more than 2000 instances recorded. Following this, the trunk and upper extremities



exhibited notable frequencies of occurrence. This distribution pattern highlights the varying prevalence of cancer across different body regions, offering valuable information for further analysis and decision-making in medical research and clinical practice.

Data Preprocessing

Our initial step involved integrating image data with the corresponding cancer types in our dataset. To achieve this, we first encoded the 'dx' column, which represents different types of cancer, into numerical labels ranging from 0 to 6, reflecting the seven cancer types in our study. Subsequently, we linked each row in the dataset with its associated image by appending the image path to a new column. This process was facilitated by mapping unique identifiers extracted from image filenames to their respective file paths. The images were sourced from two folders, 'HAM10000_images_part_1' and 'HAM10000_images_part_2', with each folder containing a portion of the dataset's images. By iteratively mapping image identifiers to file paths in both folders, we seamlessly integrated the images with the dataset, enabling further analysis and model development.



	lesion_id	image_id	dx	dx_type	age	sex	localization	label	
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	2	HAM10000_images_part_
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	2	HAM10000_images_part_
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	2	HAM10000_images_part_
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	2	HAM10000_images_part_
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	2	HAM10000_images_part_

Following the encoding process, we proceeded to create a new column named 'image_array,' which served as a repository for the image data. This column allowed us to store the images in array format within the dataset. To achieve this, we applied a function row-wise to read and resize each image using OpenCV. The function read_and_resize_image() took the image path as input and resized the image to the specified dimensions, converting it from the default BGR to RGB format. Finally, the resulting RGB image array was stored in the 'image_array' column for further analysis and model development.

Following the integration of image data into the dataset, our subsequent action involved partitioning the data into training and testing sets in an 80-20 ratio. Subsequently, we employed the Adaptive Synthetic Sampling (ADASYN) technique to address class imbalance, ensuring a more representative distribution of classes in the training data. Upon balancing the dataset, the resampled training set comprised 37,565 samples, with each sample represented as a 32x32x3 array (X) and its corresponding label encoded as a categorical vector (y) with dimensions (37565, 7), reflecting the seven classes of cancer types under consideration. This preprocessing step enhanced the robustness and generalizability of our model by mitigating the effects of class imbalance and ensuring adequate representation of all classes during model training.

MODEL IMPLEMENTATION

CNN

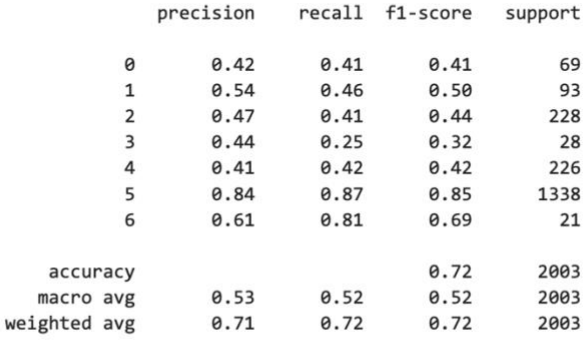
For the Convolutional Neural Network (CNN) model, we designed a sequential architecture that leverages the power of convolutional layers to extract hierarchical features from the input images. The sequential model is composed of several layers, including convolutional, max-pooling, batch normalization, dropout regularization, and fully connected layers.

In the model architecture, we started with two convolutional layers with 32 filters each, followed by a max-pooling layer to down sample the feature maps. Batch normalization was applied after each convolutional layer to improve the convergence and stability of the network. We then added two more sets of convolutional layers with 64 and 128 filters, respectively, each followed by max-pooling and batch normalization.

To prevent overfitting, dropout regularization was incorporated before the fully connected layers. This technique randomly drops a fraction of neurons during training, forcing the network to learn more robust and generalized features. The fully connected layers consist of densely connected neurons with rectified linear unit (ReLU) activation functions, facilitating non-linear transformations and enabling the network to learn complex patterns in the data.

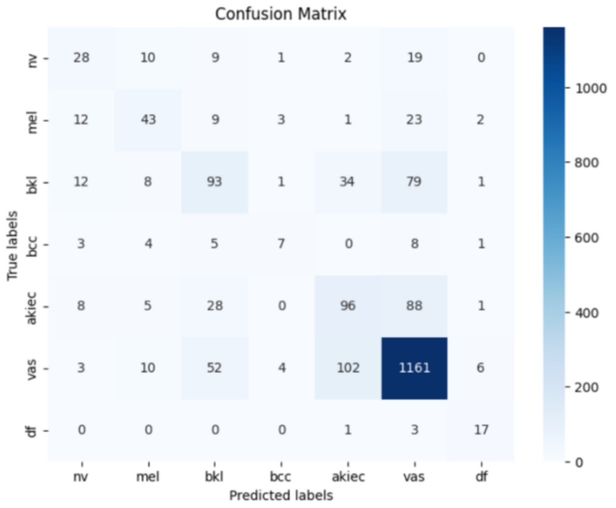
During training, we utilized the Adam optimizer with a learning rate of 0.001 to minimize the categorical cross-entropy loss function, which measures the dissimilarity between the predicted and actual class distributions. The model was trained for 25 epochs with a batch size of 8, allowing it to iteratively update its parameters and improve its performance over successive epochs.

Upon evaluation on the test set, the CNN model achieved an accuracy of approximately 72%, indicating its ability to correctly classify images into their respective cancer types. Additionally, precision, recall, and F1-score metrics were computed for each class, providing a comprehensive assessment of the model's performance across different categories.



```
              precision    recall  f1-score   support

           0       0.42      0.41      0.41        69
           1       0.54      0.46      0.50        93
           2       0.47      0.41      0.44       228
           3       0.44      0.25      0.32        28
           4       0.41      0.42      0.42       226
           5       0.84      0.87      0.85      1338
           6       0.61      0.81      0.69        21

    accuracy                           0.72      2003
   macro avg       0.53      0.52      0.52      2003
weighted avg       0.71      0.72      0.72      2003
```

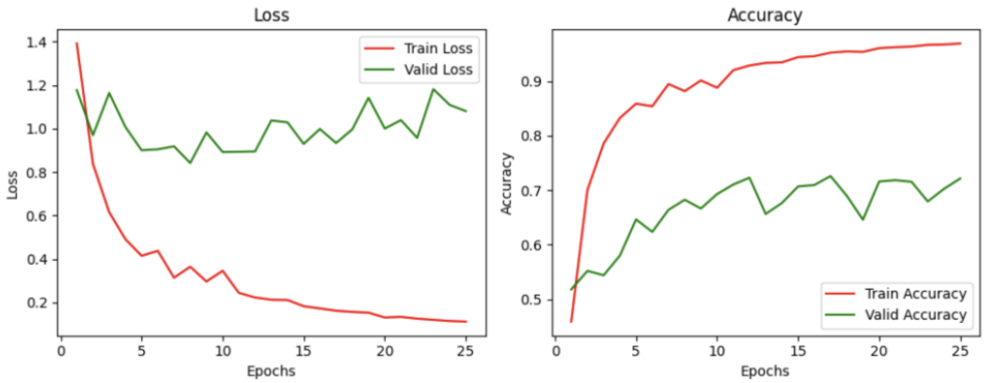
To gain further insights into the classification results, we generated a confusion matrix heatmap, which visualizes the model's predictions compared to the ground truth labels. This visualization helps identify any areas of confusion or misclassification, guiding potential adjustments or enhancements to the model architecture or training process.



Confusion Matrix

True labels \ Predicted labels	nv	mel	bkl	bcc	akiec	vas	df
nv	28	10	9	1	2	19	0
mel	12	43	9	3	1	23	2
bkl	12	8	93	1	34	79	1
bcc	3	4	5	7	0	8	1
akiec	8	5	28	0	96	88	1
vas	3	10	52	4	102	1161	6
df	0	0	0	0	1	3	17

The plots illustrate training and validation accuracy, along with training and validation loss, across epochs. They offer insights into model learning and convergence, aiding in performance assessment and improvement strategies.



Overall, the CNN model demonstrates promising performance in classifying skin cancer types from medical images, showcasing its potential utility in clinical settings for aiding in diagnosis and treatment decision-making. With further optimization and refinement, the model could potentially contribute to improved patient outcomes and healthcare delivery.

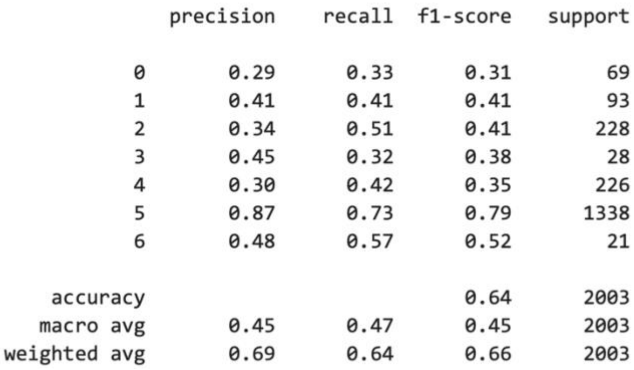
VGG16

The VGG16 model, a powerful convolutional neural network (CNN) architecture, was employed to develop a classification model for medical image data. Leveraging the pre-trained VGG16 model as a feature extractor, we added custom layers for classification atop its base. The model's architecture included additional fully connected layers with ReLU activation functions, followed by a final softmax layer for multi-class classification. The base VGG16 layers were frozen to retain their learned features during training, thereby preventing their modification. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, with accuracy as the evaluation metric.

During training, the model underwent 20 epochs on the resampled training set, resulting in an accuracy of approximately 89.96%. This training accuracy demonstrates the model's ability to learn from the augmented data generated through the ADASYN balancing technique. Subsequently, the model was evaluated on the testing set, achieving a test accuracy of 63.60%.

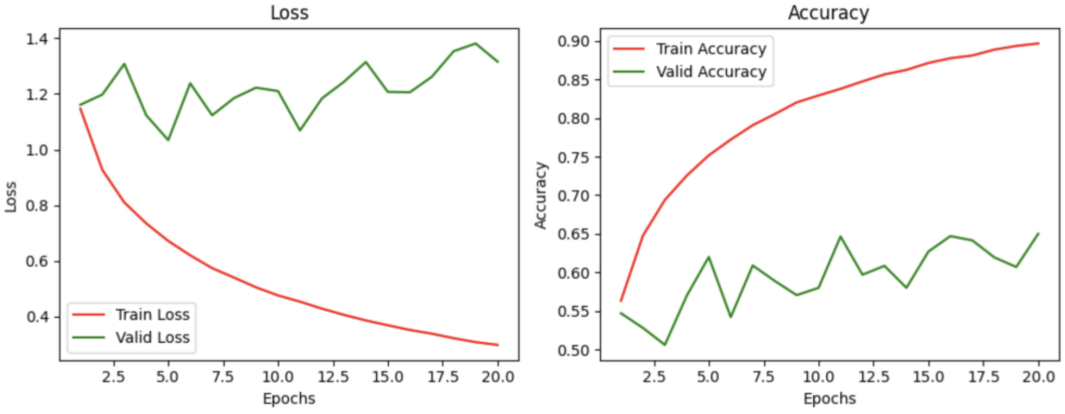
This performance indicates the model's capability to generalize well to unseen data, despite being trained on imbalanced medical image data.

Furthermore, precision, recall, and F1-score metrics were calculated for each class to evaluate the model's performance across different cancer types. These metrics provide insights into the model's ability to correctly classify instances of each class while considering both false positives and false negatives. The precision-recall trade-off varies across different classes, reflecting the model's strengths and weaknesses in distinguishing between cancer types.

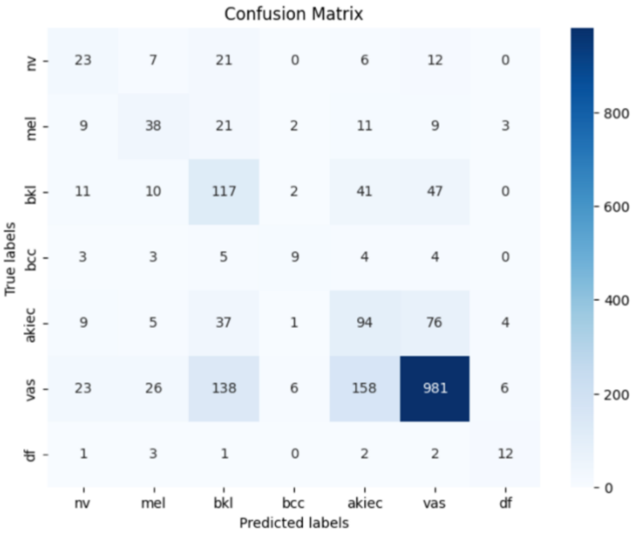


```
              precision    recall  f1-score   support

           0       0.29      0.33      0.31        69
           1       0.41      0.41      0.41        93
           2       0.34      0.51      0.41       228
           3       0.45      0.32      0.38        28
           4       0.30      0.42      0.35       226
           5       0.87      0.73      0.79      1338
           6       0.48      0.57      0.52        21

    accuracy                           0.64      2003
   macro avg       0.45      0.47      0.45      2003
weighted avg       0.69      0.64      0.66      2003
```

The provided plots depict the training and validation performance of the model over epochs. In terms of accuracy, the model demonstrates a consistent improvement in training accuracy throughout the epochs, indicating its ability to learn from the training data. However, validation accuracy plateaus or even diminishes after approximately the 6th epoch, suggesting a potential issue of overfitting, where the model performs well on the training data but fails to generalize to unseen data. Regarding loss, the training loss steadily decreases over epochs, indicating the model's ability to minimize prediction errors on the training data. Conversely, the validation loss begins to rise after the 6th epoch, indicating overfitting as the model becomes increasingly tailored to the training data at the expense of generalization. These observations highlight the need for further optimization or regularization techniques to mitigate overfitting and enhance model performance on unseen data.



In addition to the model's performance metrics, a confusion matrix heatmap was generated to visualize the classification results in detail. This heatmap provides a comprehensive overview of the model's performance by illustrating the number of true positive, true negative, false positive, and false negative predictions for each class. By visualizing the confusion matrix, we gain insights into the model's ability to correctly classify instances of each cancer type and identify any potential areas of confusion or misclassification. This heatmap serves as a valuable tool for evaluating the model's strengths and weaknesses across different classes, enabling further analysis and refinement to improve its performance.



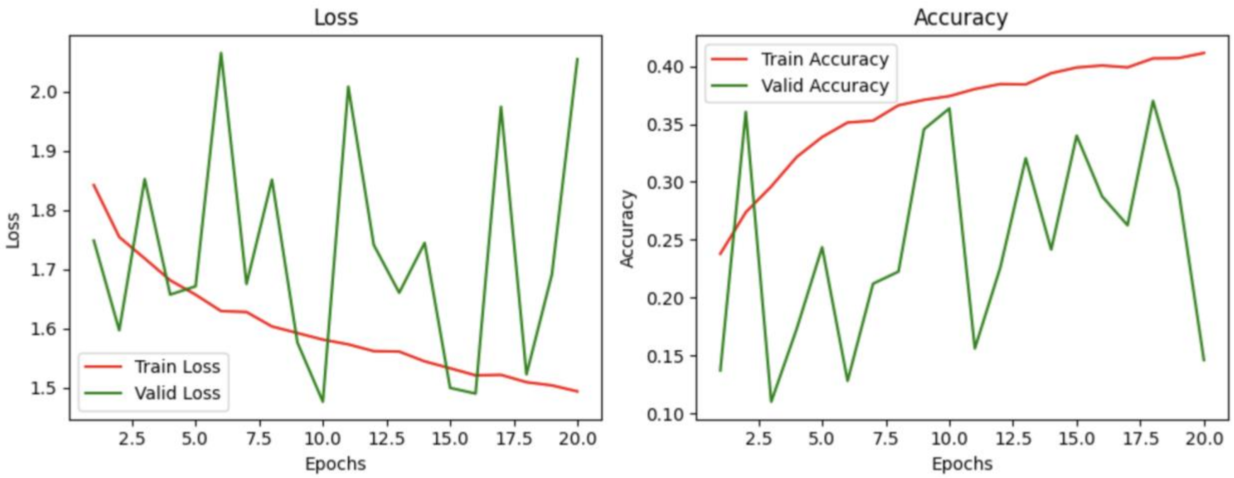
	nv	mel	bkl	bcc	akiec	vas	df
nv	23	7	21	0	6	12	0
mel	9	38	21	2	11	9	3
bkl	11	10	117	2	41	47	0
bcc	3	3	5	9	4	4	0
akiec	9	5	37	1	94	76	4
vas	23	26	138	6	158	981	6
df	1	3	1	0	2	2	12

Overall, the VGG16-based classification model demonstrates promising performance in classifying medical images, laying the foundation for further optimization and refinement to enhance its accuracy and robustness in real-world applications.

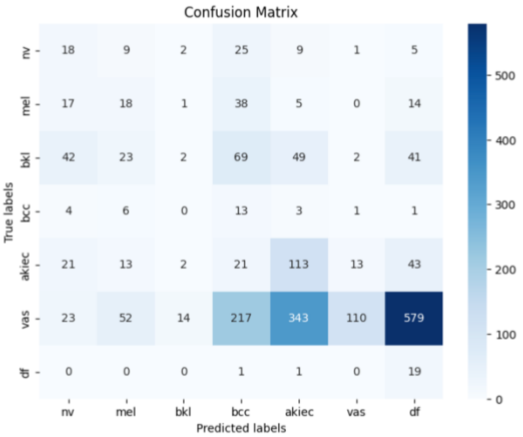
ResNet

The ResNet50 model was employed for image classification tasks. Initially, the pre-trained ResNet50 model without the classification layer was loaded. Custom top layers were added for classification, including a global average pooling layer and dense layers with ReLU activation. The base ResNet50 layers were frozen to retain their pre-trained weights. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, with accuracy as the evaluation metric.

During training, the model was fitted to the training data for 20 epochs, with validation data used for monitoring performance. The training process revealed an increase in both training and validation accuracy over epochs, indicating the model's learning process. However, fluctuations were observed in validation accuracy, suggesting potential overfitting or model instability.



The evaluation of the trained model on the test data yielded a test accuracy of 24.86%. This indicates the model's ability to generalize to unseen data. Additionally, a confusion matrix and classification report provide further insights into the model's performance across different classes.



	nv	mel	bkl	bcc	akiec	vas	df
nv	18	9	2	25	9	1	5
mel	17	18	1	38	5	0	14
bkl	42	23	2	69	49	2	41
bcc	4	6	0	13	3	1	1
akiec	21	13	2	21	113	13	43
vas	23	52	14	217	343	110	579
df	0	0	0	1	1	0	19

Overall, while the ResNet50 model demonstrates promising performance, further optimization and fine-tuning may be required to enhance its accuracy and robustness for practical applications.

CONCLUSION

In this project, we employed various CNN architectures to classify skin lesions, aiming to enhance early detection of skin abnormalities. Despite challenges such as class imbalance and dataset size, the traditional CNN model exhibited the best performance, achieving a test accuracy of 72%. Through 25 epochs of training, the model showed consistent improvement in accuracy and loss metrics, indicating effective learning and convergence. Precision, recall, and F1-score metrics further validate the model's ability to correctly classify skin lesion types. Our findings underscore the significant potential of CNNs in medical image analysis for improving patient outcomes, highlighting their efficacy in dermatological diagnosis and treatment planning.

BIBLIOGRAPHY

- https://ieeexplore.ieee.org/abstract/document/8716400 "Multi-Model Deep Neural Network based Features Extraction and Optimal Selection Approach for Skin Lesion Classification."

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10241570/ "Deep Learning for Medical Image Analysis"

- https://link.springer.com/article/10.1007/s00521-021-06490-w "Intelligent fusion-assisted skin lesion localization and classification for smart healthcare"

- https://www.activeloop.ai/resources/glossary/adaptive-synthetic-sampling-adasyn/#:~:text=Adaptive%20Synthetic%20Sampling%20(ADASYN)%20is,classification%20performance%20for%20underrep resented%20classes.

- https://pytorch.org/vision/main/models/resnet.html