TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

KHOA TOÁN KINH TẾ

DSEB ESSAY REPORT

Topic:

Advancing Multiclass Skin Lesion Diagnosis with Deep Learning Techniques

Members:

1. Nhan Yến Trang - 11219290
2. Nguyễn Lê Bình - 11219262
3. Phạm Xuân Lộc - 11219278
4. Đỗ Ngọc Thiện - 11219289
5. Nguyễn Thành Long - 11213549

Instructor: Mr Nguyen Thanh Tuan

Hà Nội, 2024

TABLE OF CONTENT

ABSTRACT

In this study, we developed and evaluated a deep learning model to classify skin lesions using the SIIM-ISIC Melanoma Classification Challenge dataset. Utilizing advanced convolutional neural networks (CNNs) and vision transformers (ViTs), we aimed to improve the accuracy and efficiency of melanoma detection. The dataset comprised 33,126 high-resolution clinical images with detailed annotations. We implemented various techniques such as data augmentation, batch normalization, and learning rate reduction to enhance model performance. Our Inception-ResNet-V2 model achieved the highest accuracy, demonstrating its potential as a reliable diagnostic tool for melanoma and other skin conditions.

INTRODUCTION

Skin cancer is a pervasive global health issue, affecting millions of individuals each year. Among the different types of skin cancer, melanoma is particularly concerning due to its high mortality rates. Despite representing only 1% of skin cancers, melanoma accounts for over 80% of skin cancer deaths **[1]**. In addition, an estimated 57,000 people died of melanoma in 2020, according to GLOBOCAN, resulting in age-standardized mortality of 0.7/100,000 for men and 0.4/100,000 for women worldwide **[2]**. In light of these numbers it is imperative to develop effective diagnostic methods for early detection and treatment of melanoma.

Traditional diagnostic methods, relying on visual assessment and biopsies, have limitations such as subjectivity and invasiveness. However, advancements in technology offer promising solutions, particularly deep learning models like convolutional neural networks (CNNs), which can improve the accuracy and efficiency of melanoma detection by learning intricate patterns from medical images.

Vision transformers (ViT) are another emerging approach showing promise in melanoma diagnosis, capable of extracting crucial information from dermoscopy images. Leveraging these advancements, our project aims to develop a robust deep learning model using the SIIM-ISIC Melanoma Classification Challenge dataset. By distinguishing between benign and malignant lesions, we seek to streamline the diagnostic workflow, reducing unnecessary biopsies and improving patient outcomes.

Beyond melanoma, our project aims to classify a broader range of skin lesions, including normal moles and various skin conditions. This comprehensive approach empowers physicians to make accurate diagnoses and prescribe appropriate treatments, enhancing patient care and optimizing medical resource allocation. We are committed to ongoing research and refinement to

ensure the model's effectiveness and its positive impact on individuals affected by skin conditions.

CONTENTS

1. Literature review

Deep learning, a sophisticated subset of artificial intelligence, has exhibited remarkable efficacy in various image recognition tasks. Convolutional Neural Networks (CNNs) are particularly adept at analyzing medical images due to their capacity to learn and extract intricate features from expansive datasets. In the context of skin cancer classification, several studies have employed CNNs to distinguish between benign and malignant cutaneous lesions with commendable accuracy.

One seminal study utilized a ResNet-152 model, which was pretrained on the ImageNet dataset and subsequently fine-tuned using a diverse collection of clinical images from multiple sources, including the Asan, MED-NODE, Edinburgh, and Hallym datasets [3]. The ResNet-152 model demonstrated superior classification performance, achieving Area Under the Curve (AUC) values of up to 0.96 for Basal Cell Carcinoma (BCC) and melanoma on the Asan dataset [3]. These results underscore the model's capability to perform on par with, and occasionally surpass, the diagnostic accuracy of seasoned dermatologists. However, the study also highlighted the YY/imperative for more diverse datasets to enhance the model's generalizability and performance across different populations and imaging conditions [3].

In addition to CNNs, Vision Transformers (ViT) have emerged as a formidable alternative for image classification tasks. The ViT architecture, inspired by its success in natural language processing, leverages self-attention mechanisms to meticulously focus on salient image features while suppressing extraneous information [4]. An advanced ViT model, designated SkinTrans, was proposed and evaluated on the HAM10000 dataset and a clinical dataset [4]. This model integrated multi-scale and overlapping sliding windows along with contrastive learning techniques to amplify feature extraction and classification accuracy. The SkinTrans model achieved impressive outcomes, with an accuracy of 94.3% on the HAM10000 dataset and 94.1% on the clinical dataset, demonstrating its robustness and efficacy in classifying dermatological images [4].

The synthesis of CNN and ViT models has shown substantial promise in mitigating the limitations of traditional diagnostic methodologies. These deep learning models can process vast amounts of data, discerning subtle patterns and features that may elude human observers [5]. This capability not only augments diagnostic accuracy but also diminishes the subjectivity and variability inherent in visual examinations conducted by dermatologists [5]. EfficientNet models have also been employed for skin cancer classification due to their ability to balance accuracy and computational efficiency, scaling depth, width, and resolution effectively.

Another pivotal study by Haenssle et al. focused on developing a CNN-based system for automated melanoma diagnosis [6]. This system was trained on a large dataset of dermoscopic images and demonstrated high sensitivity and specificity in melanoma detection, outperforming traditional diagnostic methods. The study emphasized the potential of deep learning systems to provide reliable, non-invasive diagnostic tools that can aid dermatologists and improve early detection rates of melanoma.

Kumar et al. explored the application of deep learning and transfer learning techniques for multi-class skin cancer classification [7]. Various CNN architectures, including VGG16, ResNet50, and InceptionV3, were pre-trained on ImageNet and fine-tuned on a dataset of skin lesion images. Moreover, by employing DenseNet-121, the study showed significant improvements in detecting various types of skin cancer, further validating its use in medical image analysis. The study demonstrated that transfer learning significantly improves the classification accuracy of skin cancer images. Addressing class imbalance through data augmentation, the researchers achieved a balanced and robust model for skin cancer classification [7].

Despite these promising advancements, challenges persist in the domain of deep learning-based skin cancer classification. The processing latency for high-resolution images and the necessity for extensive clinical validation remain critical areas for refinement [4]. Moreover, these models must be trained on diverse and representative datasets to ensure their applicability in real-world clinical settings. Future research endeavors should prioritize optimizing these models for expedited processing and validating their performance in large-scale clinical trials to establish their reliability and efficacy in routine medical practice [4].

In conclusion, the advancements in deep learning techniques, particularly the utilization of CNNs and ViTs, have significantly contributed to the field of dermatological oncology. These models offer a non-invasive, efficient, and highly accurate method for the early detection and diagnosis of melanoma and other skin cancers, ultimately improving patient prognoses and alleviating the burden on healthcare systems.

Table 1. *List of review papers written on skin lesion analysis.*

Paper	Year	Scope
Arshed et al. [3]	2023	Explored Vision Transformers (ViT) and pre-trained CNN models for multi-class skin cancer classification
Rezvantalab et al. [4]	2021	Evaluated state-of-the-art CNN architectures for classifying eight types of skin diseases using dermoscopy images
Adegun and Viriri [5]	2021	Reviewed deep learning techniques for skin lesion analysis and melanoma cancer detection

Haenssle et al. [6]	2018	Developed a CNN-based system for automated melanoma diagnosis
Kumar et al. [7]	2022	Reviewed deep learning and transfer learning techniques for multi-class skin cancer classification

2. Dataset Description

The SIIM-ISIC Melanoma Classification dataset is a comprehensive collection of medical images and associated metadata that has been curated specifically for melanoma classification. The dataset was generated by the International Skin Imaging Collaboration (ISIC) and includes images from various sources, such as the Hospital Clinic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School

The data used in the SIIM-ISIC 2020 competition comprised 33,126 individual cases. Each case included an image of a skin lesion and descriptive details. The dataset consists of high-resolution clinical images captured using different imaging modalities, including dermoscopy and clinical photography. These images showcase a diverse range of melanoma lesions, representing different sizes, shapes, colors, and stages of development. Each image has been meticulously annotated by trained dermatologists, providing precise bounding boxes and pixel-level masks to accurately delineate the melanoma lesions.

Beyond the images themselves, the dataset also provides detailed descriptions for each one. This metadata encompasses patient demographics, clinical history, lesion characteristics, and other relevant information. The dataset includes the following information for each patient: their unique patient ID, gender, age, the location on the body where the lesion is found (anatomic site), the diagnosis of the lesion (such as 'nevus', 'melanoma', 'seborrheic keratosis', etc.), whether the lesion is benign or malignant, and a target variable where '**1**' represents melanoma and '**0**' represents other types of lesions.

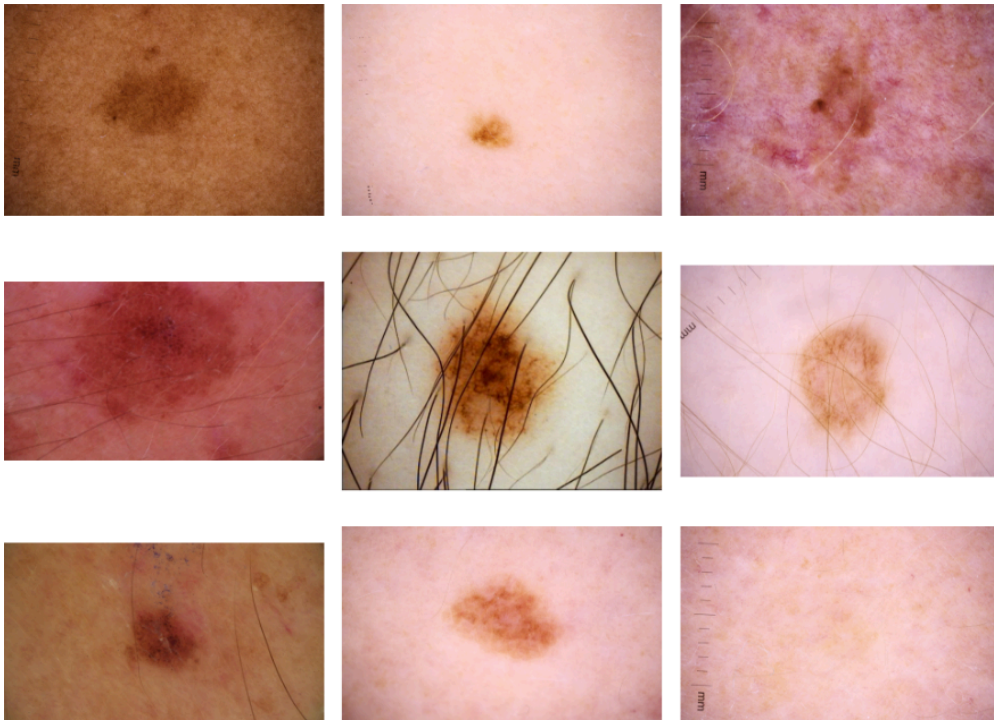

Figure 1: *Displaying random images from the dataset*

The diagnosis column in the dataset initially contained a variety of terms describing different skin conditions. To standardize these terms for consistent analysis, we categorized them into four main groups as follows:

- UNK (Unknown): This category includes diagnoses that are either unknown or do not fit neatly into other categories. Specifically, it encompasses terms like unknown, cafe-au-lait macule, and atypical melanocytic proliferation.
- NV (Nevus): This category is reserved for instances labeled as nevus, commonly referred to as moles.
- MEL (Melanoma): This category includes instances diagnosed as melanoma, a dangerous form of skin cancer characterized by the uncontrolled growth of melanocytes.
- BKL (Benign Keratosis-like Lesions): This category covers several benign skin conditions that resemble keratosis but are non-cancerous. It includes terms such as seborrheic keratosis, lentigo NOS, lichenoid keratosis, and solar lentigo. Although these lesions are generally harmless, they can sometimes be mistaken for more serious conditions if not carefully examined.

To facilitate effective model training and evaluation, we partitioned the dataset into training, validation, and test sets. The training set comprises 80% of the dataset, amounting to 26,160 samples, which represent the bulk of the data used to train the model. From the remaining 20%, we allocated 25% to the test set and the remaining 75% to the validation set. Consequently, the

validation set contains 4,905 samples used to tune model parameters and prevent overfitting, while the test set includes 1,636 samples used to evaluate the final model performance.

In summary, the SIIM-ISIC Melanoma Classification dataset provides a valuable resource for researchers and clinicians working on melanoma classification and related fields. With its diverse image data, accurate annotations, comprehensive metadata, and commitment to quality, this dataset facilitates advancements in melanoma detection and improves patient care by enabling the development of more accurate and personalized diagnostic tools and interventions.

3. Methods and Algorithms

3.1. Techniques

3.1.1. Data Augmentation

In this study, two primary augmentation techniques were employed using TensorFlow: random flipping and random zooming.

- Random flipping involves flipping images horizontally and vertically, helping the model learn orientation-invariant features, enhancing its ability to recognize objects regardless of their orientation.
- Random zooming entails zooming in and out of images within a specified range, allowing the model to learn features at different scales, especially objects in varying contexts.

These augmentation methods not only increase the variability of the training data, thereby preventing overfitting, but also act as a form of regularization, ensuring the model learns more general features applicable to unseen data. As a result, these techniques contribute to the development of more accurate, robust, and generalizable machine learning models.

3.1.2. Batch Normalization

Batch Normalization standardizes the inputs to a layer for each mini-batch, therefore, stabilizes the learning process and reduces the number of training epochs.

Batch Normalization speeds up learning, makes the network more robust to the choice of initial weights, reduces the risk of vanishing/exploding gradients, and provides a bit of regularization and noise resistance, often eliminating the need for Dropout.

3.1.3. Learning Rate reduction

Learning rate reduction, particularly through the ReduceLROnPlateau method, lowers the learning rate when performance metrics, like validation loss, stagnate after a set number of epochs, helping the model escape local minima and ensures stable convergence.

The ReduceLROnPlateau callback monitors validation accuracy. If no improvement is seen for several epochs, the learning rate is reduced by a factor, enabling more precise weight updates. This is crucial for distinguishing between different types of skin lesions. A minimum learning rate is set to prevent the rate from becoming too small, which could halt learning. This, combined with additional callbacks, like EarlyStopping and ModelCheckpoint, ensure efficient convergence, prevent overfitting, and preserve optimal parameters.

Using ReduceLROnPlateau, the model adapts better to the complexities of skin cancer classification, achieving a more stable learning process. This adaptability is essential for high accuracy in medical image classification, where precision is critical for patient outcomes.

3.1.4. Early Stopping

Early stopping prevents overfitting while maintaining accuracy. Its core principle involves halting the training process before the model becomes prone to overfitting.

We use a parameter called **early_stop**, which is an instance of the **EarlyStopping** callback from the Keras library, to stop the training process early if the performance of the model on the validation set does not improve after a certain number of epochs

During the model training phase, we continuously monitor the performance on a distinct validation dataset. If there is no improvement in the model's validation performance after a specified number of epochs, a parameter often referred to as **'patience'**, the training process is halted. This juncture is typically reached when the model has effectively captured the underlying patterns in the training data but has not yet begun to overfit.

3.1.5. Model Checkpoint

A parameter named **model_chkpt** is also used to save the weights and parameters of a model during training at specific intervals. It allows for the preservation of the best-performing version of the model. This ensures that the model is monitored, and if needed, the saved checkpoints can be loaded to restore the model to its previous state.

Model Checkpointing operates by intermittently storing the state of the model during the training phase, including the model's architecture, the learned parameters, and the state of the optimizer. This is particularly beneficial in a competitive environment where computational resources and time are constrained.

The "optimal" model is typically characterized as the one that exhibits superior performance on a validation set. By preserving the state of the model at its peak performance, we ensure the most effective version of the model for predictions, even if the model's performance deteriorates in subsequent training epochs.

3.2. Algorithms

3.2.1. Neural Network Architectures

Convolutional Neural Network (CNN)

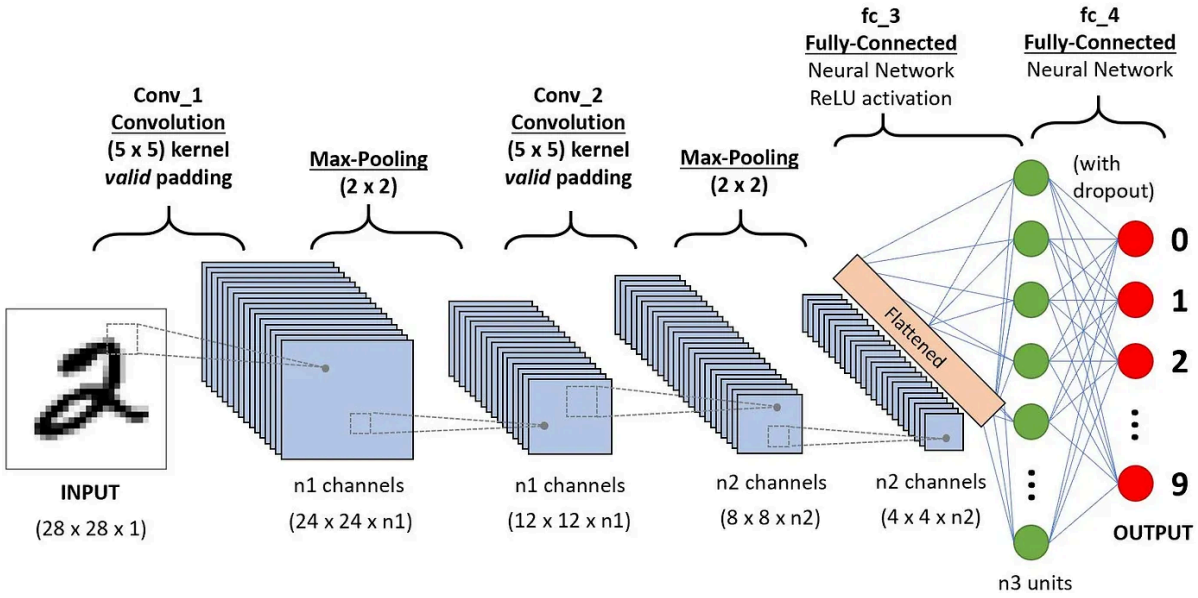

***Figure 2**: Architecture of the CNNs applied to digit recognition*

Convolutional Neural Networks (CNNs) excel in image processing and recognition. They start with an input layer representing an image as a pixel array. Convolutional layers apply filters to detect features, producing feature maps. ReLU activation adds non-linearity to learn complex patterns.

Pooling layers reduce feature map size, lowering computational load and preventing overfitting. Fully connected layers then flatten the data for prediction, with the final output layer providing the result.

CNNs are trained via backpropagation, updating parameters to minimize loss. This process enables CNNs to effectively learn spatial features, making them ideal for image classification, object detection, and segmentation.

Residual Network (ResNet)

ResNet addresses the vanishing/exploding gradient problem with residual blocks and skip connections, which link layer activations to subsequent layers, enhancing information flow and learning.

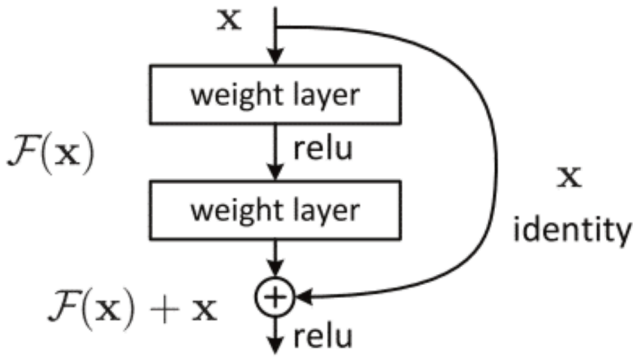

Figure 3: *Skip connection or Shortcut*

ResNet's skip connections help the network learn identity functions, ensuring higher layers perform at least as well as lower layers. This ability to learn intricate patterns is particularly beneficial for the SIIM-ISIC Melanoma Classification competition.

The deep architecture of ResNet allows it to learn both local features (e.g., lesion color and texture) and global features (e.g., lesion shape and symmetry), making it highly effective for skin cancer classification.

Inception ResNet v2

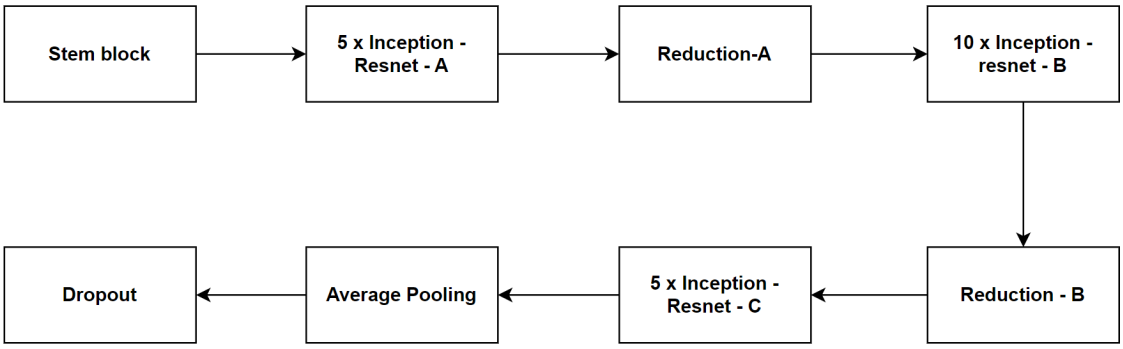

Figure 4: *The basic architecture of Inception-Resnet-v2.*

The Inception-ResNet-v2 network starts with a stem block that reduces spatial dimensions and extracts basic features from the input image. It then incorporates five Inception-ResNet-A modules, combining parallel convolutions with residual connections for efficient complex feature extraction.

The Reduction-A block follows, downscaling feature maps to decrease spatial dimensions and increase depth, reducing computational complexity. Next, ten Inception-ResNet-B modules refine features, capturing intricate patterns, followed by the Reduction-B block, which further reduces spatial dimensions.

Five Inception-ResNet-C modules then continue refining features, capturing high-level details. Average pooling reduces each feature map to a single value, lowering dimensionality while retaining critical information. The dropout layer prevents overfitting by randomly setting a fraction of input units to zero during training, enhancing generalization. These stages lead to the final classification layer, making the network effective for complex image analysis.

Dense Convolutional Network (DenseNet)

In 2017, Huang et al. [8] proposed DenseNet, a convolutional neural network with densely connected structure, and DenseNet structure is shown here:

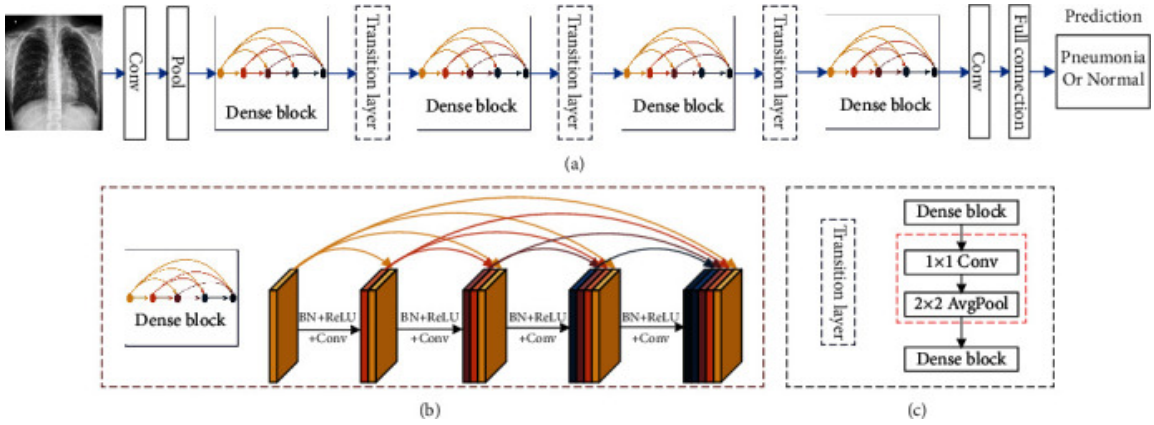

Figure 5: *DenseNet structure.*

- DenseNet's basic structure includes dense blocks, transition layers, convolutional layers, and fully connected layers.
- DenseBlocks consist of densely connected units with nonlinear mapping functions (BN, ReLU, Conv). Inputs merge with previous outputs, enabling feature reuse and mitigating gradient vanishing, resulting in a scalable DenseNet model.
- Transition layers, located between dense blocks, use 1×1 convolution and 2×2 average pooling to compress inputs and reduce dimensionality, preventing overfitting. The fully connected layer integrates and classifies feature information, minimizing the impact of feature location on classification.

Vision Transformer (ViT)

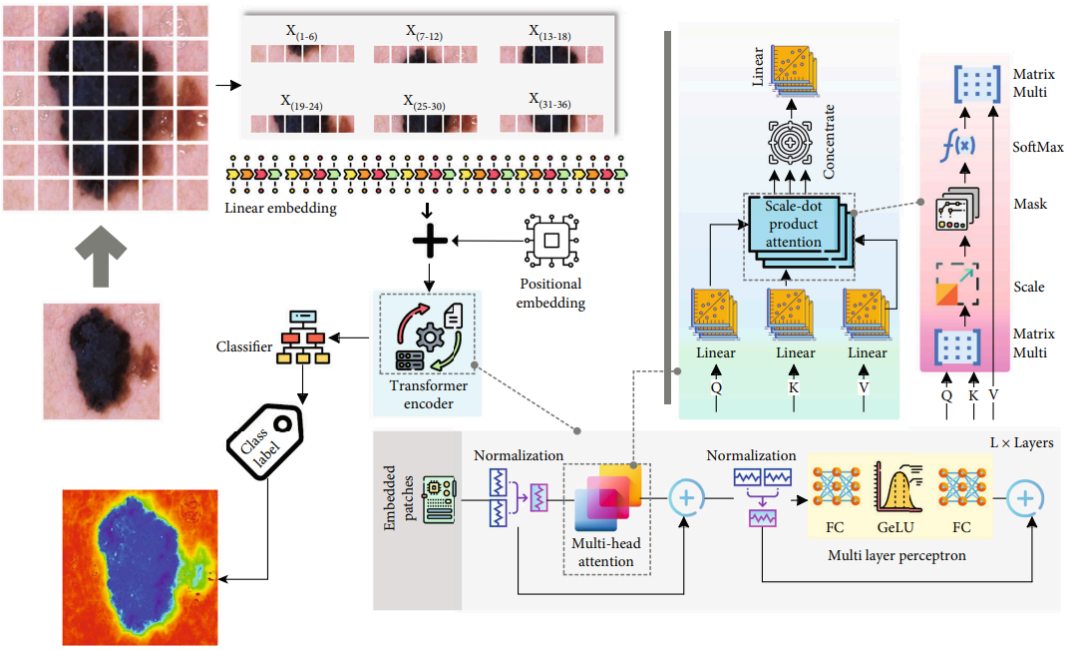

Figure 6: *Vision transformer-based skin cancer classification model.*

The skin lesion image is divided into patches, converted into vectors with a linear embedding layer, and positional embedding is added. Multi-head attention focuses on various parts of the image simultaneously using Query (Q), Key (K), and Value (V) matrices. Normalization is applied for stability, followed by processing through a multi-layer perceptron with GELU activation.

In Scale-dot Product Attention, Queries, Keys, and Values are derived from input vectors. The dot product of Queries and Keys is computed, scaled, and transformed into probabilities using Softmax, which are then used for a weighted sum of the Values.

The resulting vector is fed into a classifier to determine the image label. ViT-B/16 is used for detailed skin cancer prediction, while ViT-B/32 is suitable for resource constraints or faster training times.

3.2.2. Adam Optimize

Adam, short for Adaptive Moment Estimation, is a popular deep learning optimization algorithm. It combines the benefits of AdaGrad and RMSProp, adjusting learning rates efficiently for each parameter.

Adam enhances gradient descent by using adaptive learning rates, effectively handling varying gradient scales and data sparsity. It initializes two moment vectors to zero, then calculates the exponentially decaying average of past gradients and squared gradients, applying bias correction to account for initial biases.

Finally, parameters are updated using these corrected estimates. Adam's adaptive learning rates and bias correction ensure efficient and reliable performance, making it ideal for large-scale, complex models. Its computational efficiency and robustness to noisy and sparse gradients make it widely used in deep learning.

3.2.3. Softmax Activation

Softmax activation is crucial for neural network classification tasks. It converts raw model outputs (logits) into probabilities, ensuring the sum equals 1. This is essential for multi-class classification, where each class needs an assigned probability.

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

The softmax function transforms a vector of logits into probabilities by exponentiating each logit, summing these values, and then dividing each exponentiated logit by the sum.

Key properties of softmax include producing probabilistic outputs, ensuring normalization, and emphasizing differences between logits. It is widely used in the final layer of neural networks for classification tasks and in attention mechanisms to assign weights to inputs.

3.3. Confusion matrix

The confusion matrix serves as a fundamental tool for evaluating the performance of machine learning classification models, applicable when the output involves two or more categories. It is essentially a table that displays the intersections of the predicted and actual classes.

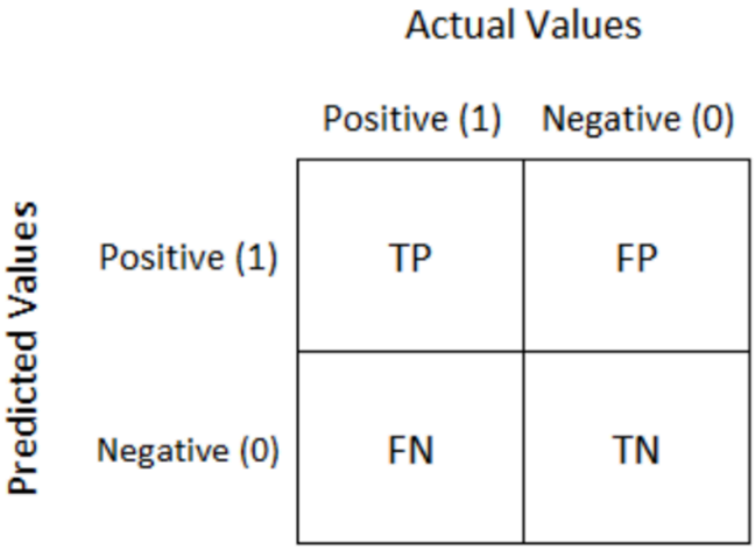

Figure 7: *Confusion Matrix illustration*

Defined as a matrix, it facilitates the performance assessment of a classification algorithm on a test dataset whose true values are known. It proves invaluable for quantifying Recall, Precision, Accuracy, and the AUC-ROC curve. To comprehend these metrics, it's essential to introduce four fundamental variables:

- True Positive (TP): The count of positive instances accurately identified by the model.
- True Negative (TN): The count of negative instances accurately identified by the model.
- False Positive (FP): The count of negative instances incorrectly identified as positive by the model, also known as a Type I error.
- False Negative (FN): The count of positive instances incorrectly identified as negative by the model, also referred to as a Type II error.

Accuracy represents the ratio of correctly predicted observations to the total observations:

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

Precision, or the positive predictive value, calculates the ratio of correctly predicted positive observations to the total predicted positive observations:

$$Precision = \frac{TP}{FP + TP}$$

Recall measures the ratio of correctly predicted positive observations to all observations in the actual class:

$$Recall = \frac{TP}{TP + FN}$$

The F1 Score represents the harmonic mean of Precision and Recall, balancing both metrics:

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Specificity measures the proportion of actual negative instances accurately identified by the model, reflecting the fraction of true negative outcomes from all genuine negative cases.

$$Specificity = \frac{TN}{TN + FP}$$

The ROC Curve plots the true positive rate against the false positive rate at various threshold levels, with the AUC representing the area under the ROC Curve. This area measures the model's classification quality; the larger the area, the better the model's performance. An AUC of 1 implies perfect distinction between positive and negative classes, 0 implies complete misclassification, and 0.5 indicates an inability to distinguish between classes.

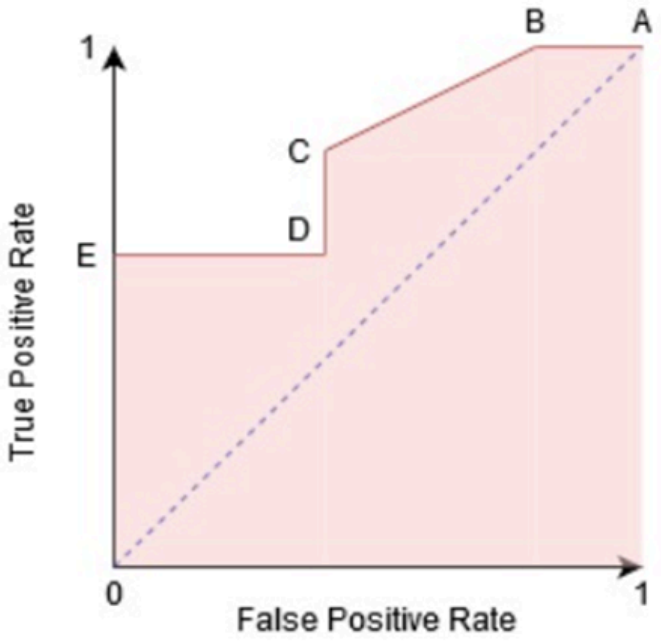

Figure 8. *The ROC curve*

4. Project Analysis

4.1. Data Augmentation

In this project, we employed Data Augmentation techniques to enhance the diversity and size of our training dataset, aiming to improve the generalization capability of our deep learning model. Specifically, we used the Sequential class from the tensorflow.keras library to apply random transformations to the images, including horizontal flips, vertical flips, and zooming in and out.

Using random flips helps the model become more flexible in recognizing features regardless of the orientation of the object. Zooming in and out between -10% and 10% allows the model to learn how to identify objects at various scales, enhancing its ability to generalize. By applying these transformations before further processing, we ensure that the training data is rich and diverse.

We chose to focus on these specific transformations rather than others, such as color adjustments, because our primary concern is the shape and structure of the objects within the images. In medical imaging, where the form and structure of features (e.g., moles or lesions) are crucial for diagnosis, maintaining the integrity of these details is essential. Color adjustments can introduce variability that is irrelevant to the model's performance in this context.

Following these transformations, we implemented an important step called hair removal. This process cleans the images by reducing noise and allowing the model to focus on significant features. The hair removal procedure involves converting the image to grayscale to highlight details like hair, applying a Blackhat morphological transformation to emphasize these areas, thresholding to create a binary image, and using inpainting to replace the hair pixels with surrounding pixels. This reconstruction creates a cleaner image with less noise.

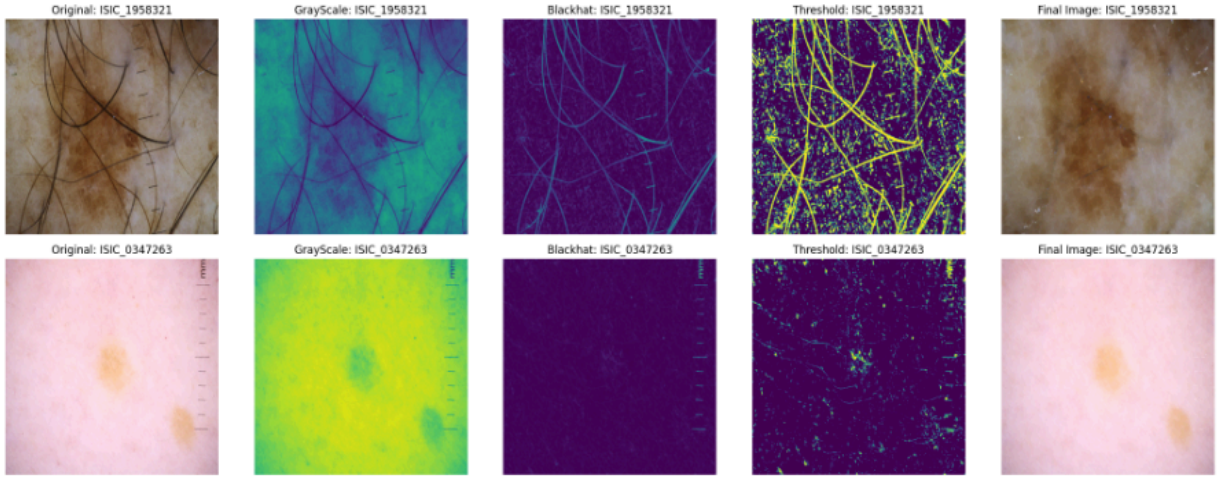

Figure 9: *Hair removal process*

The combination of Data Augmentation and hair removal has resulted in a training dataset that is both diverse and clean, helping to minimize overfitting and improve the model's accuracy on new test data. By removing hair, the images become clearer, enabling the model to concentrate more effectively on essential features. These techniques have significantly contributed to the overall performance of the model, making it more robust and capable of accurately identifying important characteristics.

4.2. Model Architecture

In this project, we utilized several models to classify skin lesions into four categories, with the Inception-ResNet-V2 model achieving the highest performance. Initially, we set up the Inception-ResNet-V2 model with pre-trained weights on the ImageNet dataset. This pre-training helped the model recognize fundamental features in images, providing a strong foundation for learning the specific characteristics of our skin lesion dataset.

We removed the original top classification layer of the model and added custom layers tailored to our classification task. These custom layers included a Global Average Pooling layer to reduce the dimensionality of the feature maps without losing essential information. Next, we added a Dense layer with 128 units, using ReLU activation and L2 regularization to prevent overfitting. We also included a Dropout layer with a 0.5 dropout rate to randomly disable some nodes during training, enhancing the model's generalization capability. Finally, we added a

Dense layer with the number of units matching the number of classes and used softmax activation to produce probability outputs for each class.

After constructing the model, we compiled it using the Adam optimizer with a low learning rate to update the weights while employing the categorical_crossentropy loss function for multi-class classification and monitored accuracy during training. We set up several callbacks to improve the model's performance. These included ReduceLROnPlateau to decrease the learning rate when validation accuracy plateaued, EarlyStopping to halt training when there is no performance improvements after several epochs, and ModelCheckpoint to save the best version.

During training, we used data generators to augment the training data with techniques such as flipping, rotating, shifting, and zooming images. Additionally, we computed class weights to handle class imbalance in the training data, ensuring that the model does not become biased towards more frequent classes. The model was trained on the augmented training dataset and validated on a separate validation dataset over multiple epochs.

Once training was complete, we evaluated the model on the test dataset to determine its overall performance. We used metrics such as accuracy, recall, precision, and F1 score to assess the model. The results showed that the Inception-ResNet-V2 model achieved high performance in classifying skin lesions, demonstrating its practical application in assisting medical diagnoses. By using this model, we developed a robust system for classifying skin lesions, improving diagnostic accuracy and supporting doctors in making quick and precise treatment decisions.

5. Result and Discussion

Overall, the performance of the models varied significantly, with CNN-based models generally outperforming Vision Transformer models.

The Inception-ResNet-V2 model achieved the highest performance with an accuracy of 0.87, precision of 0.94, an F1-score of 0.90, and a micro-average ROC AUC of 0.98. This superior performance can be attributed to its advanced architecture, which combines inception modules and residual connections, allowing it to capture complex patterns and features more effectively.

Table 2. *Performance Comparison of Deep Learning Models.*

Model	Accuracy	Precision	F1-score	Micro-average ROC AUC
ResNet 50	0.84	0.95	0.89	0.96
DenseNet 121	0.80	0.95	0.87	0.96
Inception ResNet V2	0.87	0.94	0.90	0.98
VIT - B16	0.76	0.94	0.83	0.93
VIT - B32	0.60	0.96	0.73	0.88

On the other hand, the Vision Transformer models, particularly VIT-B32, showed the lowest performance with an accuracy of 0.60, precision of 0.96, an F1-score of 0.73, and a micro-average ROC AUC of 0.88. The lower performance of these models can be explained by their reliance on large amounts of data for effective training and their inherent differences in processing images compared to CNNs. Transformers excel in tasks with sequential data and may not capture spatial hierarchies as effectively as CNN-based models in image classification tasks, especially with limited data.

Looking at Figure 10, the confusion matrix for the Inception-ResNet-V2 model highlights its strengths and weaknesses in classifying different classes. The model performs exceptionally well in classifying class 2 and class 3 with accuracies of 0.87 and 0.88, respectively. However, it struggles more with class 0 and class 1, showing higher misclassification rates. Specifically, class 0 is often misclassified as class 1, indicating a potential overlap in features between these two classes. The ROC curves further illustrate this, with class 2 achieving an AUC of 0.98, while class 1 has a significantly lower AUC of 0.76, highlighting the need for further optimization for class 1.

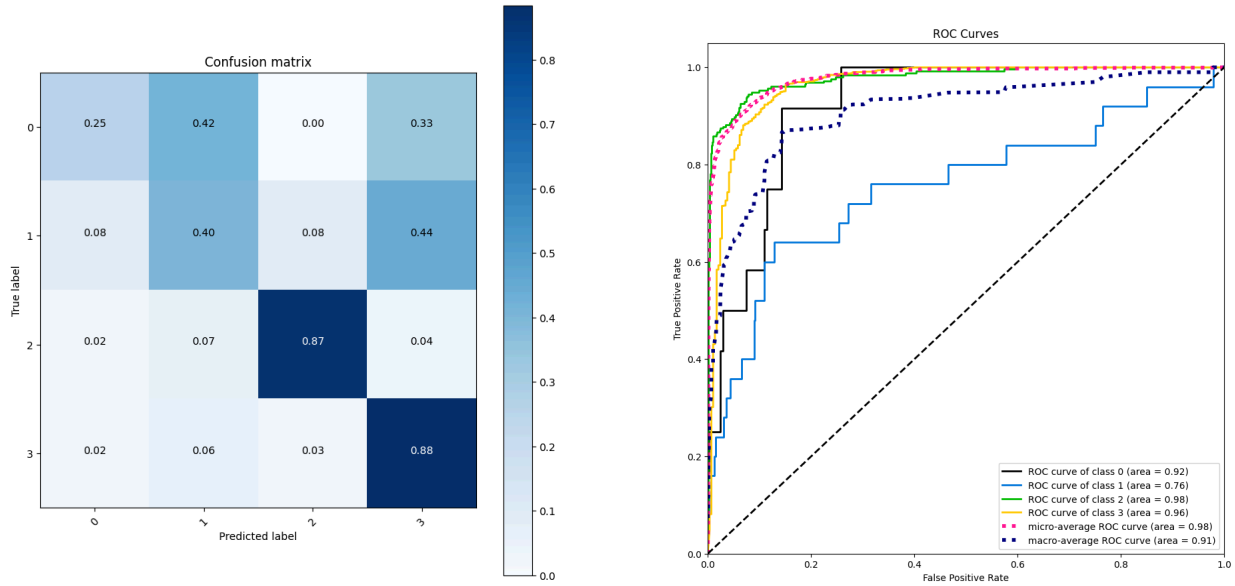

Figure 10: *Confusion matrix and ROC curve of Inception ResNet V2*

6. Conclusion and Future work

In this study, we successfully developed a robust deep learning model for the classification of skin lesions, utilizing the SIIM-ISIC Melanoma Classification Challenge dataset. Our research focused on leveraging state-of-the-art convolutional neural networks (CNNs) and vision transformers (ViTs) to enhance diagnostic accuracy and efficiency. Among the various models

tested, the Inception-ResNet-V2 model achieved the highest performance, with an accuracy of 87%, precision of 94%, an F1-score of 90%, and a micro-average ROC AUC of 0.98. This success underscores the model's potential as a valuable diagnostic tool in dermatology, capable of aiding dermatologists in the early detection and classification of melanoma and other skin conditions. By reducing the reliance on invasive biopsy procedures and providing accurate diagnostic support, our model has the potential to significantly improve patient outcomes and streamline clinical workflows.

Despite the promising results, several areas require further exploration and improvement to fully realize the potential of our deep learning model for skin lesion classification. Future work will focus on expanding the dataset with more diverse images to improve the model's generalizability across different populations and imaging conditions. This includes collecting additional images from various demographics, skin types, and imaging conditions, as well as including more examples of rare skin conditions to help the model accurately classify less common cases. Implementing more sophisticated data augmentation techniques, such as elastic transformations and color space augmentations, can further diversify the training data and prevent overfitting. Additionally, exploring advanced regularization methods, such as dropout with spatial constraints or adversarial training, will enhance the model's robustness and performance.

Integrating genetic information, patient history, and other clinical data with image data can provide a more comprehensive diagnostic approach, leading to more personalized and accurate diagnoses. Utilizing natural language processing (NLP) techniques to analyze patient records and clinical notes can complement the image-based diagnosis, offering a holistic view of the patient's condition. Optimizing the model to reduce processing latency, especially for high-resolution images, is crucial for its practical application in clinical settings. Developing lightweight versions of the model that can be deployed on mobile devices or edge computing platforms will make the technology more accessible and versatile.

Conducting extensive clinical trials to validate the model's performance in real-world settings is necessary to ensure its reliability and acceptance among healthcare professionals. Ensuring that the model adheres to ethical guidelines, including transparency, fairness, and accountability, is vital. This involves addressing potential biases in the dataset and ensuring equitable performance across different patient groups. Meeting regulatory standards for medical devices and AI in healthcare is crucial for the deployment of the model in clinical settings. This will involve collaboration with regulatory bodies to achieve necessary certifications. By addressing these areas, we aim to develop a more accurate, reliable, and comprehensive tool for the early detection and classification of skin lesions, ultimately enhancing its impact on patient care and contributing to advancements in dermatological diagnostics.

REFERENCE

1. National Cancer Institute Melanoma of the Skin-Cancer Stat Facts. [(accessed on 10 May 2021)]; [CrossRef]

2. Sung H., Ferlay J., Siegel R.L., Laversanne M., Soerjomataram I., Jemal A., Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. 2021;71:209–249. [CrossRef]

3. Arshed, Muhammad Asad & Mumtaz, Shahzad & Ibrahim, Muhammad & Ahmed, Saeed & Tahir, Muhammad & Shafi, Muhammad. (2023). Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models. [CrossRef].

4. Rezvantalab, Amirreza & Safigholi, Habib & Karimijeshni, Somayeh. (2018). Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms. [CrossRef]

5. Adegun, Adekanmi & Viriri, Serestina. (2021). Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. [CrossRef]

6. Haenssle, Holger & Fink, Christine & Schneiderbauer, R & Toberer, Ferdinand & Buhl, Timo & Blum, A & Kalloo, A & Hassen, A & Thomas, Luc & Enk, A & Uhlmann, Lorenz. (2018). Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. [CrossRef]

7. Kumar, Vinayshekhar & Kumar, Sujay & Saboo, Varun. (2016). Dermatological disease detection using image processing and machine learning. [CrossRef]

8. Wang, D., Huang, C., Bao, S. *et al.* Study on the prognosis predictive model of COVID-19 patients based on CT radiomics. *Sci Rep* 11, 11591 (2021). [CrossRef]

The source code could be found at: Link