Jagadesh Varma Nadimpalli

Professor Dr. Krishna Bathula

Analytics Capstone Project 23917 CS 668

30 April 2024

Vision Transformer and EfficientNetV2 for Skin Cancer Diagnosis

Skin cancer diagnosis is a critical task in healthcare, with early detection significantly impacting patient outcomes. In this project, I explored the efficacy of two state-of-the-art architectures, Vision Transformer (ViT8x8) and EfficientNetV2, for skin cancer diagnosis. The Vision Transformer model, originally proposed for natural image classification, is adapted to handle dermatological images through fine-tuning and transfer learning techniques. Similarly, EfficientNetV2, known for its superior efficiency and scalability, is tailored to the task of skin cancer diagnosis. Through extensive experimentation, we assess the performance of both models in terms of accuracy, sensitivity, and specificity. Furthermore, we analyze the interpretability of the model’s predictions to gain insights into their decision-making processes. Our results demonstrate the potential of leveraging advanced deep learning architectures to improve the accuracy and efficiency of skin cancer diagnosis, thus contributing to the advancement of computer-aided diagnostic systems in dermatology.

**Introduction**

Skin cancer is the prevailing type of cancer in the United States, impacting one in every five people during their lives. Every day, around 9,500 Americans are diagnosed with skin cancer, making it more common than any other kind of cancer altogether. Nonmelanoma skin lesions, such as basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), affect more than 3 million Americans each year, with BCC occurrences increasing by 145% and SCC by 263% between 1976-1984 and 2000-2010.

Melanoma, the deadliest type of skin cancer, is also on the rise, with over one million Americans affected. In 2022, an estimated 197,700 additional cases are projected, evenly split between invasive and noninvasive forms. Melanoma is now the sixth most frequent cancer among men and women in the United States. Although worldwide melanoma rates have doubled since 1982, recent patterns suggest a drop in younger people under 30, while rates among those over 80 have risen. Early identification is critical since it considerably boosts survival chances. For localized melanoma, the five-year survival rate is 99%, but this lowers to 68% and 30% when it spreads to lymph or other organs, respectively. The yearly cost of treating skin cancer in the United States exceeds $8 billion, emphasizing the economic and health importance of successful skin cancer prevention and early detection measures.

The HAM10000 image dataset is used for skin cancer diagnosis, and it comprises dermatoscopic images representing seven distinct classes of skin lesions commonly encountered in clinical practice: melanocytic nevi, melanoma, benign keratosis, basal cell carcinoma (bcc), actinic keratoses, vascular lesions, and dermatofibroma. Melanocytic nevi, benign growths of melanocytes, are often harmless but may exhibit atypical features. Melanoma, the most aggressive skin cancer type, requires early detection for effective treatment. Benign keratosis presents as waxy bumps and can resemble malignant lesions. Basal cell carcinoma is the most common skin cancer, while actinic keratoses are precancerous lesions resulting from sun exposure. Vascular lesions encompass various skin abnormalities related to blood vessels, and dermatofibroma is a benign fibrous tumor. Understanding these distinct classes is crucial for accurate skin cancer diagnosis and classification.

**Literature Review**

In the past, research has extensively explored skin cancer classification, identifying seven categories. Evaluation using the VGG-16 and VGG-19 models revealed varying performance metrics, with the latter showcasing superior precision accuracy, recall, and F1 score. In recent years, numerous convolutional neural network (CNN) architectures have emerged, aiming to enhance image classification tasks.

Among these, the Efficient-Net model has garnered attention for its remarkable effectiveness and efficiency, achieving an impressive 84.4% accuracy in the ImageNet challenge with just 66 million parameters. By scaling width, depth, and resolution proportionally, Efficient-Net offers a promising solution to the computational load challenges encountered by earlier models. Moreover, transfer learning has been widely adopted in the skin cancer classification literature to address the scarcity of labeled datasets. This approach involves reusing pre-trained models, such as Dense-Net, Res-Net, Mobile-Net, Google-Net, VGG19, and Alex-Net, trained on a source task and adapting them to the target task.

Improvements in deep learning have prompted several novel ways to diagnose skin cancer, which is one of the most prevalent varieties of cancer globally. Chaturvedi et al. found that utilizing pre-trained Mobile-Net models fine-tuned on dermatoscopic images from the HAM10000 dataset resulted in significant accuracies of up to 83.1% across seven skin cancer classifications.

This study highlights the ability of deep learning models to aid dermatologists in early-stage cancer detection, considerably improving decision-making in clinical settings. Similarly, another landmark work used the Noisy Student (EfficientNet-L2) architecture to successfully handle essential difficulties such as unequal class distribution and the incorporation of new metadata to improve classification performance. This model pushed state-of-the-art approaches and the winners of the ISIC 2019 competition, demonstrating convolutional neural networks' fast-expanding capabilities in medical imaging.

Another innovative approach for identifying melanoma and other skin illnesses was developed using a simplified strategy that included image pre-processing, feature extraction with Densenet-121, and classification using a fully connected neural network. Using the ISIC dataset, our system achieved an impressive accuracy rate of 95%, outperforming existing state-of-the-art approaches (Study, no date). This technology improves the precision of skin disease diagnosis and provides a practical answer for healthcare workers by enhancing diagnostic processes in terms of time and dependability. By incorporating these complex CNN models into user-friendly apps, the study offers prospective paths for expanding the accessibility of effective skin cancer screening, contributing significantly to healthcare technology advancement. Deep learning techniques in clinical applications are growing, with models designed expressly to improve diagnosis accuracy and minimize subjectivity when evaluating skin lesions.

Furthermore, building on these advances, this study seeks to harness the potential of EfficientNetV2, which is regarded as one of the most powerful and efficient CNNs, as well as Vision Transformer and Swin Transformer. Both Vision Transformer and Swin Transformer, which are noted for their capacity to capture spatial associations in images, provide greater scalability and flexibility than typical CNNs. These characteristics are predicted to contribute significantly to the accuracy and efficiency of skin cancer identification and categorization in this study.

**Methodology**

*Data Analysis Techniques*

The HAM10000 dataset was found to contain a diverse array of skin lesions, with melanocytic nevi constituting the largest category, consisting of 6,705 images, followed by benign keratosis with 1,099 images, dermatofibroma with 115 images, melanoma with 1,113 images, vascular lesions with 142 images, basal cell carcinoma with 514 images, and actinic keratoses with 327 images. A crosstab function was used to count between the diagnostic (dx) and localization columns from the HAM10000 meta dataset. This method provided useful insights into the distribution and location of distinct skin lesion types, allowing for a more complete knowledge of the dataset for research objectives.

A graph of numbers and a bar

Description automatically generated with medium confidenceA table with numbers and a number

Description automatically generated with medium confidence

Figure 1: Explored Distribution Patterns of Lesions in the HAM10000 Dataset: Diagnosis and Localization Analysis

*Image Preprocessing*

The proposed methodology employs a Digital Hair removal Algorithm for denoising skin images as part of preprocessing. Initially, the algorithm converts the input image into grayscale to simplify processing. Subsequently, it utilizes a morphological operation known as Black Hat Transform to highlight regions containing noise, such as hair strands. By employing an appropriately sized rectangular kernel, the algorithm effectively identifies and isolates noisy areas. Following this, a thresholding technique is applied to create a binary mask, distinguishing noise from the background.

A diagram of a process

Description automatically generated

Figure 2: Digital Hair Removal Algorithm

Finally, the inpainting process is performed using the in-plant function, where the noisy regions identified by the binary mask are reconstructed based on surrounding pixel information, resulting in a denoised image. This method offers an efficient means of mitigating noise traces commonly present in skin images, enhancing their quality for subsequent analysis and diagnosis. The results after applying the Digital Hair Removal algorithm are shown below.

Close-up of hair loss and hair loss on skin

Description automatically generated

Figure 3: Comparison of Original image with Denoised image

*Image Augmentation*

To address the inherent class imbalance within the dataset and ensure robust training, we employed a systematic approach to balance the dataset through targeted image augmentation. Initially, we categorized images by their respective diagnostic labels using a DataFrame that grouped indices based on these labels. This grouping facilitated targeted augmentation strategies for each class, depending on their representation within the dataset.

Class-specific augmentation multipliers were then applied to achieve a more uniform distribution of classes. For instance, the least represented classes were more heavily augmented than those with a higher initial count. Importantly, images from underrepresented classes were duplicated by factors up to 57 times, depending on their original counts, to ensure that all classes had comparable images. This approach aimed to ensure the model's fairness by mitigating bias towards more frequently occurring classes and enhancing the generalizability of the trained models.

Following the balancing process, the augmented dataset encompassed a significantly more significant number of images, ensuring that each class was equally represented. This was crucial for training deep learning models sensitive to imbalanced data, which can skew their predictive accuracy and reduce their clinical utility.

To further enhance the dataset and introduce variability that closely mimics real-world conditions, we implemented a dynamic image augmentation pipeline. This pipeline was designed to randomly apply transformations to each image during training, thereby preventing the model from overfitting and improving its ability to adapt to new, unseen images. The transformations included random rotations, width and height shifts, shear transformations, zooms, brightness adjustments, channel shifts, and both horizontal and vertical flips. The parameters for these transformations were randomly selected within specified ranges to ensure diverse variations in the augmented images, enhancing the model's adaptability to real-world scenarios.

Using the “ImageDataGenerator” from Keras library, these transformations were applied on the fly during model training. This method preserved the original images and generated augmented images in real-time, significantly enhancing the dataset's variability without requiring additional storage space. The augmented dataset was split into training, validation, and test sets for the training and validation of the models. We utilized a stratified approach to maintain equivalent class distributions across these subsets, critical for unbiased model evaluation. The data was first split into a training set (75% of the data) and a temporary set (the remaining 25%). The temporary set was further divided equally into validation and test sets. This stratified split ensured that each set represented the overall dataset's characteristics, providing a reliable basis for training and subsequently assessing model performance.

*EfficientNetV2 Model*

This study uses two advanced computer vision algorithms to diagnose skin cancer. At the core of EfficientNetV2's architecture lies its efficient scaling method, which harmoniously adjusts the depth, width, and resolution of the network to optimize performance while maintaining computational efficiency. The architecture comprises stacked blocks of convolutional layers, activation functions, and normalization layers, each progressively extracting higher-level features from input data. Alongside these, EfficientNetV2 may integrate squeeze-and-excitation (SE) blocks to enhance feature recalibration, allowing the network to focus on salient features and disregard irrelevant ones. Moreover, some versions of EfficientNetV2 incorporate attention mechanisms to capture long-range dependencies within the data, further enhancing its ability to model complex patterns and towards state-of-the-art performance on image classification tasks.

The architecture of EfficientNetV2 begins with convolutional layers (Conv 3x3) that extract features from the input data using 3x3 filters. This is followed by fused Mobile-Net convolutional blocks (Fused MB Conv 3x3) comprising four layers, each employing 3x3 convolutions. Additionally, the architecture incorporates Mobile-Net convolutional blocks (MB Conv 3x3) with three layers, further enhancing feature extraction capabilities.

A diagram of a network

Description automatically generated A diagram of a diagram of a triangle and a triangle

Description automatically generated

Figure 4: Architectural Overview of EfficientNetV2 and Structure Comparison of MBConv and Fused-MBConv

A convolutional layer (Conv 1x1) is included for dimensionality reduction or feature transformation, followed by a fully connected layer that aggregates extracted features and maps them to the output classes or predictions. Through the strategic combination of these components, EfficientNetV2 achieves a balance between computational efficiency and performance, making it suitable for a wide range of image classification tasks.

I opted for the EfficientNetV2 B2 model since it outperformed the other Efficient-Net and EfficientNetV2 models. The EfficientNetV2 B2 model is notable for its improved accuracy and efficiency, establishing a convincing balance that enables robust image categorization jobs. I used an expanded dataset to increase the model's performance and adjust it to the intricacies of skin cancer diagnosis. This method not only helped us simulate a more diverse collection of skin photos, allowing for differences in lighting, skin tones, and lesion sizes, but it also dramatically decreased the danger of overfitting by increasing the model's variability during training. EfficientNetV2 B2 typically uses an input resolution of 260x260 pixels, has an increased depth of approximately 55 layers, and scales up the number of channels in layers.

*Vision Transformer*

The Vision Transformer (ViT) introduces a paradigm shift in image classification by departing from traditional convolutional neural network (CNN) architectures and adopting the transformer architecture initially developed for Computer Vision tasks. Unlike CNNs, which rely on convolutional layers for feature extraction, ViT replaces these layers with self-attention mechanisms, enabling the model to process images as sequences of patches and capture global dependencies and relationships within them.

A diagram of a transformer

Description automatically generated

Figure 5: Vision Transformer Architecture

This unique approach allows ViT to effectively model spatial relationships in images, facilitating robust feature extraction and classification. The architecture comprises patch embedding layers that convert input images into sequences of fixed-size patches, transformer encoder blocks containing self-attention mechanisms and feedforward neural networks, positional encoding to incorporate spatial information, and a classification head for producing final predictions.

Using the mentioned architecture, I have adapted the standard Vision Transformer (ViT) architecture, originally designed for larger 224x224 pixel images using 16x16 pixel patches, which can be better for addressing the constraints of smaller 32x32 pixel images. Recognizing the need for more granular feature extraction in smaller images, developed a modified version, the vit8x8 model, which utilizes 8x8 pixel patches. This approach allows for a higher resolution of attention within the smaller input dimensions, enabling more detailed analysis and better suitability for datasets with limited image resolution. The vit8x8 model begins with a custom Patches layer that extracts 8x8 patches from the input images.

Each patch is flattened and linearly transformed into a 64-dimensional embedding vector. These embeddings are then processed through four sequential transformer blocks, each comprising layer normalization, a multi-head self-attention mechanism with four heads, and a dropout rate of 0.1 to mitigate overfitting. Residual connections follow each attention and multi-layer perceptron (MLP) module, which includes two dense layers with ReLU activation to introduce additional non-linearity, enhancing the model's ability to learn complex patterns from the data.

The architecture culminates in a global average pooling layer that aggregates information across the entire image, followed by a dense layer of 1024 units with ReLU activation to synthesize the extracted features into a representation suitable for classification. The model outputs through a softmax layer that categorizes each image into seven classes. This customized approach not only effectively addresses the challenges posed by smaller image sizes but also maintains the flexibility and robustness of the transformer architecture, making it exceptionally well-suited for detailed and computationally efficient image analysis in contexts where image detail and processing resources are limited.

Next, I created a Swin Transformer model which employs a novel architecture that integrates the principles of transformers specifically adapted for vision tasks on images of size 32x32. Initially, the model inputs pass through a Conv2D layer with a 4x4 kernel and a stride of 4, which serves to down-sample the input while maintaining spatial hierarchies; this effectively compresses each 32x32 input image into an 8x8 grid of 96-dimensional embeddings.

A diagram of a computer

Description automatically generated

Figure 6: Swin Transformer Architecture

Following this initial transformation, the model utilizes two consecutive Swin Transformer Block instances, each configured with three attention heads and a dimensionality (dim) of 96. These blocks operate on 4x4 window partitions of the input grid, enabling localized attention within these sub-regions. The first block applies a cyclic shift of size 2 to these windows before computing self-attention, which helps the model capture relationships between adjacent windows by overlapping their contexts.

This shifting is reversed in the subsequent layer, where the windows are processed without a shift. Each Swin Transformer block comprises a layer normalization followed by multi-head attention and a two-layer MLP with GELU activation, another layer normalization, and residual connections after both the attention and MLP stages to facilitate deeper stacking of blocks without vanishing gradients. After processing through the transformer blocks, the model aggregates the features using global average pooling, followed by a dense layer of 1024 neurons with RELU activation, and finally, a softmax output layer that classifies the inputs into one of seven categories. This architecture leverages the strengths of localized attention within patches and across shifted contexts, improving the model's ability to manage hierarchically structured data efficiently.

**Analysis**

Extensive analysis is being conducted on the meta-dataset to identify valuable patterns for skin cancer lesions. The following visualizations explain these patterns.

The below histogram exhibits a right-skewed distribution, where most of the people are concentrated on the right side of the graph, indicating a higher frequency of higher age values. The distribution reaches its peak around the age of 40-50 years, suggesting that this age range person might have a higher tendency to be diagnosed with skin cancer. As the age increases beyond the peak, the frequency gradually decreases, with a long tail extending towards the higher age values.

A graph with red lines

Description automatically generated

Figure 7: Frequency distribution of skin cancer lesions based on age

The right-skewed nature of the distribution implies that people who are diagnosed with skin cancer are more likely to be 35+ years of age.

The visualization shows a bar graph that depicts the distribution of lesion sites across various body areas, categorized by gender. The y-axis depicts the number or frequency of lesions, while the x-axis shows numerous body sites. Lesions are commonly found in the back, lower extremities, trunk, and upper extremities, with men having a much greater incidence than females. Despite comparatively lower numbers, males outweigh females in physical parts such as the belly, face, chest, and foot. However, the numbers in the neck, scalp, hand, ear, genital, and acral areas are low, with no significant difference between males and females. A graph of different colored bars

Description automatically generated

Figure 8: Frequency of Lesion Locations by Sex

The "unknown" category, which includes situations when the sex is uncertain, has relatively low counts across all locales. This data sheds light on the occurrence and distribution of lesions across various body areas, indicating potential changes based on sex or gender.

The violin plot presents the age distribution across different categories, likely representing various medical conditions or diseases. The benign keratoses like-lesion (bkl) category exhibits the most comprehensive distribution, from around 20 to 90 years, with the highest density between 60 and 80 years old. The melanocytic nevi(nv) category also has a broad age range, concentrated slightly higher around 70-75 years. Categories like "df," "mel," and "bcc" have narrower distributions, primarily focused between 50 and 80 years, with peak densities around 65-75 years.

A graph showing different colored shapes

Description automatically generated with medium confidence

Figure 9: Age Distribution for various skin cancer diagnosis

The vascular lesions category is broadly distributed from 30 to 90 years, peaking between 65 and 75 years. Notably, the actinic keratoses(akeic) category displays the narrowest distribution, concentrated mainly between 60 and 80 years, with the highest density around 70 years. These observations suggest that certain conditions may be more prevalent or diagnosed within specific age ranges. In contrast, others have a broader age distribution, potentially indicating differences in risk factors, disease progression, or diagnostic patterns across the categories.

**Results**

In this study, I utilized advanced convolutional neural network architectures, specifically EfficientNet and EfficientNetV2, and transformer-based models, such as Vision Transformers (ViT8x8) and Swin Transformers, to develop models capable of diagnosing skin cancer through image analysis. The models were trained and evaluated using a comprehensive approach involving multiple sessions and phases, focusing on improving model performance through data preprocessing and augmentation.

Initial training is applied on all 15 variants of the EfficientNet and EfficientNetV2 models using denoised images over 10 epochs. During this session, we tracked performance metrics, including accuracy, precision, recall, and loss, for both training and validation datasets. The EfficientNetV2 models, particularly B0, B1, and B2, demonstrated superior performance across these metrics, which led to their selection for further training.

The commitment to enhancing model robustness led us to further train the top three EfficientNet models (V2B0, V2B1, and V2B2) on an expanded dataset of 46,543 augmented images. This process, spanning 10 epochs, significantly improved the models' accuracy and stability in loss metrics. The augmentation process played a crucial role in this, demonstrating its effectiveness in our approach.

In this final session, the EfficientNetV2B2 model underwent an extended training of 50 epochs using the augmented images. Post this extensive training, we also evaluated the performance of two transformer-based models, ViT8x8 and Swin Transformer. The models were chosen to compare their capabilities with convolutional architectures in handling complex image data.

The evaluation process was designed to provide an unbiased measure of the models real-world efficacy and generalizability. I first used a hold-out set of test images, ensuring the models were assessed on unseen data. Performance metrics, including accuracy, precision, recall, and detailed breakdowns via confusion matrices, were recorded. The robustness and reliability of our models, particularly the EfficientNetV2B2, were demonstrated. Further testing involved using external test images from the ISIC 2018 dataset, validating the models in a real-world scenario and confirming their generalizability beyond the curated training and validation datasets.

Fifteen versions of Efficient-Net and EfficientNetV2 are used in multiple model training sessions. The first training session involves the training of all models for ten epochs, using a dataset that has 10015 images. The outcomes of this session can be observed in the figure 10 provided.

A table of numbers and letters

Description automatically generated

Figure 10: Metrics after training Efficient-Net models for 10 epochs using original images.

Clearly, EfficientNetV2 functions perform better than Efficient-Net. I decided to train the best EfficientNetV2 models on an augmented dataset to choose the best one and compare it with Vision Transformer and Swin Transformer.

I carried out the next session with EfficientNetV2 B0, B1, and B2 versions using the augmented dataset, which is generated using image augmentation techniques like image transformation on the original images to generate unbiased frequency for each type of diagnosis.

A graph with blue and orange lines

Description automatically generated

Figure 11: Line plots representing training and validation accuracies of three EfficientNetV2 models on augmented images.

After performing the model building, the EfficientNetV2B2 function illustrated considerable training accuracy and stability. EfficientNetV2B2 is compared with ViT8x8 and Swin Transformer models because it has better compound scaling than the other two models in the second session.

The final session of model training feature models to be trained for fifty epochs with a reduced learning rate (=0.001) callback parameter. The previous EfficientNetV2 history or fitting is not considered, and new variables are created for the final model for the final EfficientNetV2 model. Although the EfficientNetV2 model is a convolutional model, the layer architectures and the hyperparameters are the same after using the pre-trained function and defined functions. This approach maintains the novelty of the comparison between these models.

A group of graphs with numbers and symbols

Description automatically generated with medium confidence

Figure 12: Training and Validation metrics trained over 50 epochs for EfficientNetV2B2, ViT8x8 and Swin Transformer

A graph of numbers and a line

Description automatically generated with medium confidenceA diagram of a number

Description automatically generated with medium confidenceA graph with numbers and squares

Description automatically generated with medium confidenceAccording to the results of the model testing, the EfficientNetB2 model demonstrated the highest accuracy of 98.5%, followed by the Swin transformed model with 94.64% accuracy and the Vit8x8 model with 91.27% accuracy. As per the findings of this research, EfficientNetV2B2 can be considered as the most suitable model for skin cancer diagnosis. The following presentations of the confusion matrices will help us understand the predicted incorrect labels for each type of skin cancer diagnosis among the 5818 images tested with these models.

Figure 13: Visual representation of confusion matrices for EfficienNetV2B2, Swin Transformer and Vision Transformer (ViT8x8) models.

The confusion matrices for three different models used for skin cancer diagnosis are visualized in the accompanying images. In the second phase of our testing process, we evaluated the performance of the developed models on an external dataset comprised of images from the ISIC 2018 challenge. This step was crucial for assessing the real-world applicability of the models by testing their generalization capabilities on data that was not part of the initial training or internal testing sets.

The EfficientNetV2B2 model, a standout performer, demonstrated a promising accuracy of 70.41% on the external dataset. It not only achieved the highest precision with the nv class (Nevus) at 0.85 and the highest recall within the same class at 0.87, but also delivered an impressive F1-score of 0.86. Despite some challenges in accurately classifying 'akiec' (Actinic Keratosis) and distinguishing bcc (Basal Cell Carcinoma) and bkl (Benign Keratosis-like Lesions), the model's moderate effectiveness in these areas is a testament to its reliability.

The ViT8x8 model exhibited an overall accuracy of 63.5%. It performed best with the nv class like EfficientNetV2B2 but with a slightly lower precision of 0.89 and recall of 0.75. Challenges were evident in classifying smaller or less represented classes, such as mel (Melanoma), where it only achieved a precision of 0.34 and a high recall of 0.53, indicating a tendency to over-diagnose this class. The performance was weakest in detecting df (Dermatofibroma), with the model failing to correctly identify any cases (precision and recall of 0.00).

The Swin Transformer model reached an accuracy of 65.04%. It showed a balanced performance with relatively consistent precision and recall across several classes compared to ViT8x8. However, it struggled particularly with df, like ViT8x8, managing only a precision of 0.07 and recall of 0.06. Notably, it demonstrated decent performance in detecting mel, achieving a higher recall of 0.51 compared to ViT8x8 and a competitive precision of 0.33. In the nv class, it maintained a robust precision of 0.86 and recall of 0.80.

From a quantitative perspective, all models showed strengths in classifying nv class lesions but had varied success across other types. EfficientNetV2B2 proved the most reliable overall, with consistently higher accuracy and F1 scores across multiple classes. ViT8x8 and Swin Transformer, while innovative in their approach using transformer architectures, demonstrated areas of potential overfitting or misclassification, particularly in classes with fewer training examples or greater complexity in visual features.

**Conclusion**

In conclusion, this research project has demonstrated the significant potential of advanced deep learning models, specifically EfficientNetV2 B2, Vision Transformer (ViT8x8), and Swin Transformer, in enhancing the diagnosis and classification of skin cancer. Through extensive experiments and evaluations, EfficientNetV2 B2 emerged as the better-performing model, achieving the highest accuracy (98%) and demonstrating exceptional capability in handling the varied complexities presented by the HAM10000 dataset. These findings underscore the value of incorporating cutting-edge neural network architectures in medical imaging, particularly dermatology, where precision and efficiency are crucial.

The external testing phase has been instrumental in highlighting each model's strengths and limitations, providing valuable insights into their practical utility for skin cancer diagnosis. The varied performance across different classes suggests further model tuning and combining model predictions to enhance diagnostic accuracy and reliability. By leveraging the distinct advantages of transformers and convolutional neural networks, this study not only pushes the boundaries of what is technologically feasible in skin cancer diagnostics but also sets a foundation for future research to build upon, aiming to refine these models further and extend their applicability to real-world clinical settings.

**Works Cited**

1. American Academy of Dermatology. "Skin Cancer Statistics." American Academy of Dermatology, www.aad.org/media/stats-skin-cancer. Accessed 30 Apr. 2024.
2. Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.
3. Tan, M., and Q. V. Le. "Efficientnetv2: Smaller models and faster training. arXiv 2021." arXiv preprint arXiv:2104.00298.
4. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
5. Ghosh, Pronab, et al. "SkinNet-16: A deep learning approach to identify benign and malignant skin lesions." Frontiers in Oncology 12 (2022): 931141.
6. Belaala, Abir, et al. "Enhancing Dermoscopy-Based Skin Lesions Diagnosis: A Fusion of Convolutional Neural Networks and Self-Training." 2023 International Conference on Decision Aid Sciences and Applications (DASA). IEEE, 2023.
7. Latha, D. U., et al. "Skin Cancer Classification using ConvNET." 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET). IEEE, 2023.
8. Hoshyar, Azadeh Noori, Adel Al-Jumaily, and Riza Sulaiman. "Review on automatic early skin cancer detection." 2011 International Conference on Computer Science and Service System (CSSS). IEEE, 2011.
9. Montaha, Sidratul, et al. "A shallow, deep learning approach to classify skin cancer using the down-scaling method to minimize time and space complexity." PloS one 17.8 (2022): e0269826.
10. Gessert, Nils, et al. "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting." IEEE Transactions on Biomedical Engineering 67.2 (2019): 495-503.
11. Khan, Muhammad Attique, et al. "Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification." 2019 international conference on computer and information sciences (ICCIS). IEEE, 2019.
12. Bi, Lei, et al. "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks." arXiv preprint arXiv:1703.04197 (2017).
13. Alaubaidy, Mateen Naser, Abdulqadir Ismail Abdullah, and Fatimah Mousa Alwan. "Evaluation of Modified VGG16 Learning Model for Classifying Skin Cancer Lesions." 2023 International Conference on Network, Multimedia and Information Technology (NMITCON). IEEE, 2023.
14. Bhardwaj, Paarth, Vaibhav Bhargava, and Shivendra Singh. "SkinSage-Lesion Diagnosis Using Deep Learning Techniques." 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2024.
15. Kaur, Gurjot, et al. "Early Detection of Melanoma using EfficientNet B4 Transfer Learning Model." 2023 Global Conference on Information Technologies and Communications (GCITC). IEEE, 2023.