Ensemble Classification of Skin Cancer Using Pre-trained Deep Learning Models: AlexNet and VGG16

By
Kehinde Adetola Ogundana

***Abstract*—This research addresses skin cancer classification by leveraging an ensemble of deep learning models, specifically AlexNet and VGG16, known for their feature extraction capabilities. The ensemble demonstrated superior accuracy and predictive capabilities, significantly outperforming individual models and the baseline results. Evaluation metrics such as accuracy, F1, recall, precision, and AUC indicated notable improvements, with the ensemble achieving an precision of 0.90 and an AUC of 0.77. These findings suggest that ensemble methods can effectively enhance classification accuracy in skin cancer detection.**

Keywords—Deep Learning, AlexNet, VGG16, Ensemble Model

I. Introduction

Skin cancer, particularly melanoma, poses a significant global health risk, highlighting the need for accurate and timely diagnosis for effective treatment. Traditional diagnostic methods are often subjective and error-prone, driving interest in advanced technologies like deep learning, specifically Convolutional Neural Networks (CNNs), for skin cancer classification [1], [2].

This research aims to enhance classification accuracy by combining the strengths of AlexNet and VGG16 into an ensemble model. Leveraging the complementary features of both models, the hypothesis is that the ensemble approach can outperform individual models. The study utilizes the International Skin Imaging Collaboration (ISIC) dataset and applies data augmentation techniques to improve accuracy [3]. CNN models such as AlexNet and VGG16, pretrained on a large number of images from ImageNet, have proven to be effective for this task [4], [5].

II. RELATED WORKS

Ensemble methods have shown promise in skin cancer classification and other medical image analysis tasks. Brinker et al. [6] developed an ensemble of convolutional neural networks (CNNs) combining architectures like ResNet, DenseNet, and InceptionV3. Their approach achieved an accuracy of 86.6% on the ISIC 2017 dataset, surpassing individual models but requiring significant computational resources. Similarly, Esteva et al. [7] used an ensemble of pretrained models to classify skin lesions, achieving a high accuracy of 94.4%. However, challenges include obtaining large, annotated datasets and the complexity of ensemble models, which limit practical deployment in clinical settings. Shen et al. [8] employed an ensemble approach for brain tumor segmentation, achieving state-of-the-art results, yet facing challenges in computational resources and model interpretability.

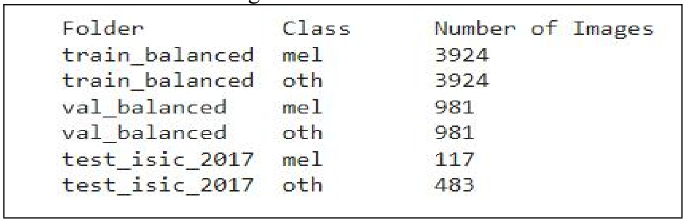
The aim of this research is to leverage AlexNet and VGG16, popular deep learning models, to enhance skin cancer classification. AlexNet's moderate complexity and VGG16's deep architecture offer complementary strengths. By combining these models into an ensemble and using data augmentation, the hypothesis is that the ensemble will generate performance comparable to recent deep learning models but with less computational resource usage. Using the ISIC dataset curated by Cassidy et al. [9], experimental results show significant improvements, with the ensemble model achieving an accuracy of 0.80 and an AUC of 0.77. These findings suggest that ensemble methods can be valuable tools in skin cancer diagnosis, potentially aiding dermatologists in making more accurate and timely decisions, despite challenges in computational resources and model interpretability.

III. DATASET

The dataset used for this research is from the International Skin Imaging Collaboration (ISIC) and is a curated, balanced subset prepared by B. Cassidy et al[9]. The researchers initially sourced 72,297 images from ISIC 2016-2020, then removed 14,310 duplicates, resulting in 57,987 images (4,905 melanoma, 52,082 other conditions).

To address class imbalance (melanoma to others ratio of 1:10.62), they created a balanced dataset. The training set has 7,848 images (3,924 melanoma, 3,924 others), and the validation set has 1,962 images (981 each class). Thus, the balanced dataset totals 9,810 images with a 1:1 ratio.

The study uses these curated balanced datasets, with performance evaluated on the separate ISIC 2017 test set (117 melanoma, 483 others), which is imbalanced for real-world testing. Including the test images, the total is 10,412 images, and the size is 171MB. The folder structure and image distribution are as follows:



```
Folder            Class      Number of Images
train_balanced    mel        3924
train_balanced    oth        3924
val_balanced      mel        981
val_balanced      oth        981
test_isic_2017    mel        117
test_isic_2017    oth        483
```

Fig. 1. Number of classes & Images in the dataset

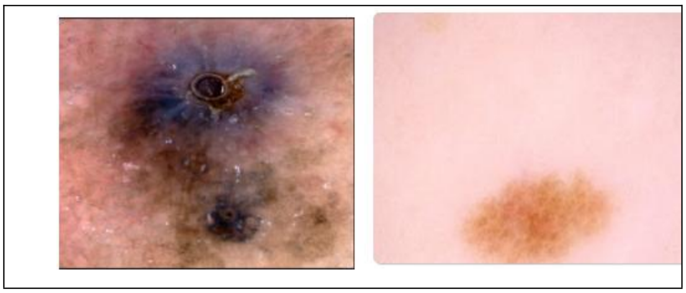


Fig. 2. Image Sample: Left :Melanoma, Right:Non-Melanoma

IV. METHODOLOGY

Overview of Deep Neural Network Architecture

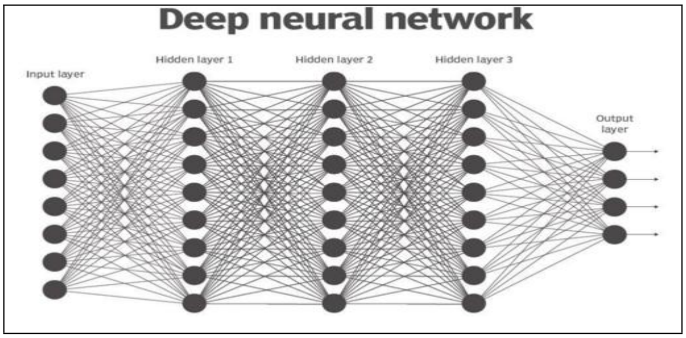


Fig. 3 A typical Deep Neural Network Architecture[10]

A typical Deep Neural Network (DNN) architecture for binary classification consists of an input layer, multiple hidden layers, and an output layer[10],[12]. The input layer receives features from the dataset, where each feature corresponds to a neuron. The hidden layers, often comprising fully connected layers or dense layers, process the input through weighted connections and activation functions like ReLU (Rectified Linear Unit). These layers learn complex patterns and representations from the data. The final output layer consists of a single neuron with a sigmoid activation function, which outputs a probability value between 0 and 1, representing the two classes. During training, the network adjusts its weights using optimization algorithms such as Adam or SGD and loss functions like binary cross-entropy to minimize the prediction error. This architecture allows the DNN to effectively classify data into one of the two categories, making it suitable for tasks such as spam detection, medical diagnosis, and sentiment analysis.

Chosen Deep Learning Algorithm -AlexNet and VGG16

A. AlexNet Architecture

The AlexNet model utilizes a pretrained feature extraction component comprising five convolutional layers and max-pooling layers[13]. Its custom classifier comprises three fully connected layers, with the first two having 4096 neurons each and ReLU activations, and the final layer outputting the number of classes. Dropout layers are integrated for regularization. Overall, the model consists of 8 layers, including the additional fully connected layers. This architecture leverages the pretrained feature extraction to ensure robust feature extraction based on a large dataset. The customized classifier enhances non-linearity with ReLU activation functions, crucial for handling complex patterns in the data. By combining pretrained feature extraction with tailored classification layers, this model architecture enhances performance on the specific task.

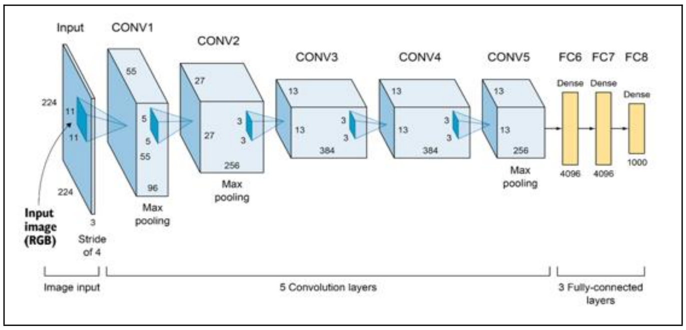


Fig. 4: Alexnet

B. VGG16 Architecture

The VGG16 model features a pretrained section for feature extraction with 13 convolutional layers, followed by adaptive average pooling to a fixed size of (7, 7). Its customized classifier comprises three fully connected layers, each with 4096 neurons and ReLU activations, culminating in an output layer for class prediction. Dropout layers are integrated for regularization. Overall, the model encompasses 18 layers, including the fully connected layers.

This architecture leverages pretrained feature extraction with 13 convolutional layers organized into five blocks, each followed by max-pooling layers, essential for discerning intricate features from input images. Post-feature extraction, an adaptive average pooling layer prepares the features for the custom classifier. The classifier consists of fully connected layers with ReLU activations, dropout layers for regularization, and an output layer determining class prediction. This integration of robust feature extraction and a tailored classifier enhances the model's performance in capturing nonlinear relationships within the data, optimizing its effectiveness in the classification task.

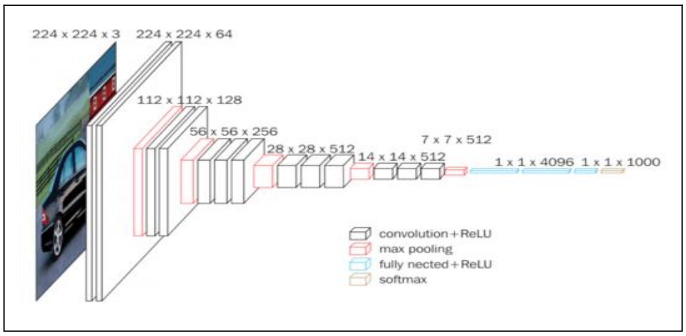


Fig 5: VGG16

C. Data augmentation

For this study, data augmentation techniques were applied to the training and validation datasets, while only resizing, transforming to tensors, and normalizing were performed on the testing datasets to improve model performance.

Random Horizontal and Vertical Flips: This randomly flips images horizontally and vertically to increase the diversity of the dataset.

Color Jitter: This adjusts the brightness, contrast, and saturation of the images to simulate different lighting conditions and enhance model robustness.

Random Rotation: This applies random rotations up to 10 degrees to the images, helping the model become invariant to orientation changes.

Random Resized Crop: This crops the images randomly and resizes them to 224x224 pixels, ensuring that the model can handle various scales and perspectives.

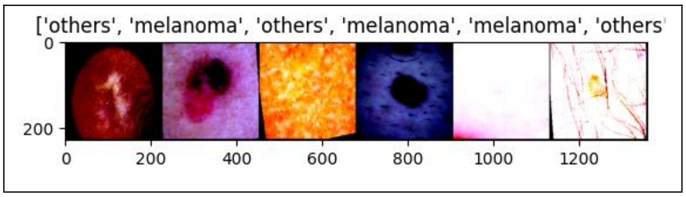

['others', 'melanoma', 'others', 'melanoma', 'melanoma', 'others'
0
200
0
200
400
600
800
1000
1200

Figure 6: Pictures of transformed image augmentation

D. Loss Function

Cross-EntropyLoss

Cross-entropy loss is a pivotal metric in classification tasks within deep learning, providing a measure of the disparity between predicted probabilities and true labels[14]. Its equation,

$$\mathrm{CE} = -\sum_i y_i \log(p_i)$$

encapsulates this evaluation process succinctly. In this equation, y_i represents the true label of the ith sample, usually encoded as a one-hot vector, while p_i denotes the predicted probability assigned to the true class. The logarithm of these probabilities is taken to penalize confidently incorrect predictions more severely, and the negative sign ensures that the loss increases as the predicted probabilities diverge from the true labels. By minimizing the cross-entropy loss during model training, classifiers are guided towards making more confident and accurate predictions, thereby enhancing overall classification performance and model generalization.

V. EXPERIMENTAL RESULT AND DISCUSSION

In this study, deep learning experiment was conducted to classify skin cancer using the ISIC 2017 dataset. Two pre-trained models, AlexNet and VGG16 was utilized, and created an ensemble model to improve classification performance. The results was compared with baseline models reported in previous research by B. Cassidy et all[9], which did not use transfer learning.

Experimental settings are as follows:-

Optimization: Used Stochastic Gradient Descent (SGD) optimizer with learning rate scheduling through ReduceLROnPlateau to adapt the learning rate based on validation performance.

Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is a widely used optimization algorithm for training machine learning models, particularly in the context of deep learning. In the realm of binary classification tasks, SGD aims to minimize the loss function by iteratively updating the model parameters based on the gradients computed from a randomly selected subset of training data (mini-batch)[11]. The key idea is to find the optimal set of parameters that best separate the two classes by adjusting them in the direction that minimizes the loss. Mathematically, the update rule for the parameters θ t each iteration ***t*** is given by:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta \mathcal{L}(\theta_t)$$

Where η is the learning rate, $L(\theta t)$ is the loss function, and $\nabla\theta$ represents the gradient of the loss with respect to the parameters. This process continues until convergence or a predetermined number of iterations. SGD offers an efficient and scalable approach to optimizing models for binary classification tasks by iteratively updating parameters to minimize the loss function.

Other parameters used:

Learning Rate: The learning rate was set to 0.01, determining the step size for each iteration of parameter updates. This value balances the trade-off between convergence speed and stability during training.

Epochs: Due to computational constraints, the number of epochs was limited to 10. This controls the number of complete passes through the training dataset, balancing thoroughness of training with available resources.

Batch Size: A batch size of 64 was chosen to optimize memory usage and computational efficiency. This consideration was especially important given the hardware limitations of a laptop or the Google Colab computing environment.

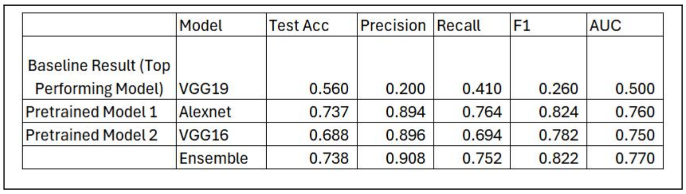
Training & Testing Platform: The models was trained and evaluated using PyTorch within the Google Colab environment. PyTorch (version 1.13.1) and CUDA (version 11.8) were employed, leveraging the T4 GPU with 16GB RAM to enhance training performance.

Evaluation Metrics: For classification performance, the following metrics were used: accuracy, precision, AUC, confusion matrix, F1 score, recall, and prediction accuracy (actual vs. true label). The results of these metrics will be discussed in the next section.

Results:

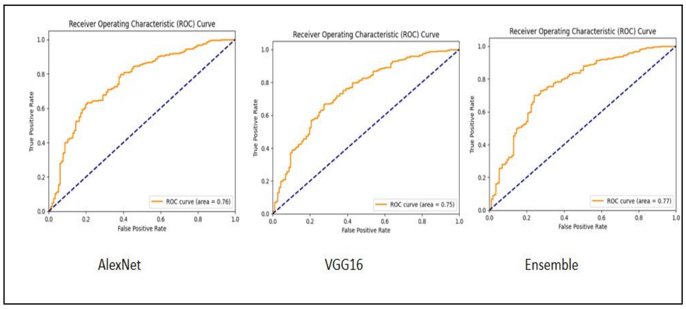
The table below shows the performance metrics of our models compared to baseline results from previous research:

Performance Metrics Results: Discussion and Analysis



	Model	Test Acc	Precision	Recall	F1	AUC
Baseline Result (Top Performing Model)	VGG19	0.560	0.200	0.410	0.260	0.500
Pretrained Model 1	Alexnet	0.737	0.894	0.764	0.824	0.760
Pretrained Model 2	VGG16	0.688	0.896	0.694	0.782	0.750
	Ensemble	0.738	0.908	0.752	0.822	0.770

Fig. 7. Metrics Results Table



AlexNet VGG16 Ensemble

Fig. 8. AUC Curve

Baseline Result vs Pretrained Models:

Compared to the top-performing baseline model by B. Cassidy et al. [9], both AlexNet, VGG16, and the ensemble exhibit significantly improved performance across all metrics. The incorporation of transfer learning from pretrained models plays a pivotal role in achieving these enhanced results, especially considering that the baseline model did not utilize any transfer learning.

AlexNet

AlexNet's high precision(0.894) indicates that the model made fewer false positive errors. Its recall of 0.764 suggests relatively fewer false negatives. The F1 score of 0.824 and AUC of 0.760 illustrate strong performance, effectively balancing precision and recall. It test accuracy(0.737) is slightly lower than that of the ensemble.

VGG16

VGG16 also outperformed the baseline but lagged slightly behind AlexNet in terms of recall, test accuracy and F1 score. However, its precision was slightly higher(0.896), suggesting it was very effective at minimizing false positives. The lower recall(0.694) compared to AlexNet indicates more false negatives, affecting the F1 score(0.782). Despite this, the overall performance was still strong, as evidenced by the high AUC of 0.750.

Ensemble Model vs. Individual Models:

The most notable performance is from the ensemble method, which combines the strengths of both AlexNet and VGG16. The ensemble achieved the highest precision (0.908) among all models, indicating the fewest false positives. Although its recall was slightly lower than AlexNet's, it was higher than VGG16's, resulting in a strong F1 score of 0.822. It also achieves the highest test accuracy (0.738) and highest AUC (0.770) among all models suggests the ensemble had the best balance of true positive and false positive rates, making it the most reliable and balanced model.

Training and Validation Accuracy and loss Curves-AlexNet

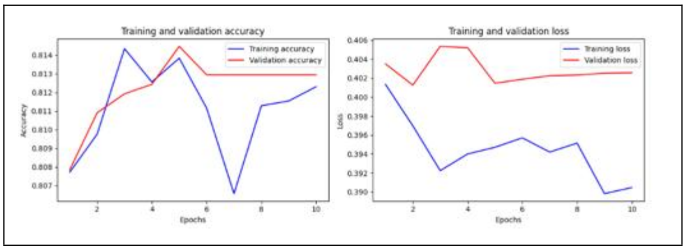


Fig.9 AlexNet

Validation Loss and Accuracy: The validation loss fluctuates slightly but remains close to the training loss. The validation accuracy also stabilizes around 0.813.

Training Loss and Accuracy: The training loss decreases consistently over epochs.The training accuracy increases gradually and stabilizes around 0.812.

The training and validation curves follow similar trends, indicating that the model is not significantly overfitting.

The validation accuracy is close to the training accuracy, suggesting that the model generalizes well to unseen data.

Overall, the model seems to be performing reasonably well without severe overfitting or underfitting.

Training and Validation Accuracy and loss Curves-VGG16

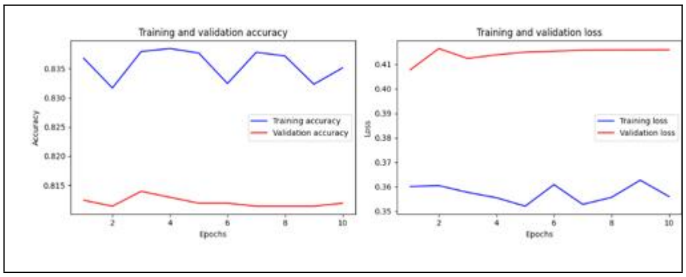


Fig. 10. VGG16

Validation Loss and Accuracy: The validation loss fluctuates slightly but remains close to the training loss.The validation accuracy also stabilizes around 0.814.

Training Loss and Accuracy: The training loss decreases consistently over epochs.The training accuracy increases gradually and stabilizes around 0.835.

Similar to AlexNet, the training and validation curves follow similar trends.The validation accuracy is close to the training accuracy, indicating reasonable generalization.

The model does not exhibit severe overfitting or underfitting.

Confusion Matrices Insights:

The confusion matrices provide detailed insights into the performance of each model—AlexNet, VGG16, and the Ensemble model—by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

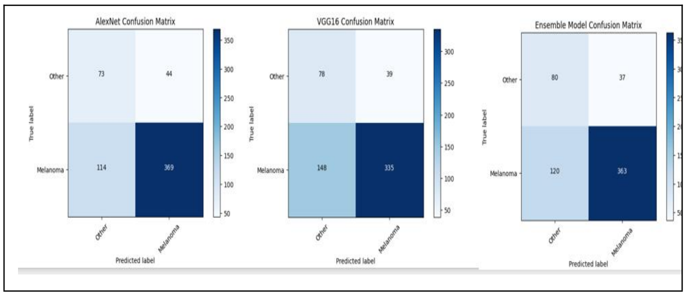


Fig.11 Confusion Matrices

AlexNet demonstrates the highest accuracy in predicting melanoma cases, with 369 true positives indicating strong ability to correctly identify melanoma cases, but also the highest false positive rate among the models, incorrectly classifying 44 other cases as melanoma.

VGG16, on the other hand, shows a slightly lower true positive rate(335) compared to AlexNet but maintains a lower false positive rate(39), suggesting a more balanced performance.

The ensemble model combines predictions from both AlexNet and VGG16, leveraging their strengths, resulting in a more robust performance. It achieves a comparable true positive rate to AlexNet with fewer false positives(363), making it the most effective model overall in accurately distinguishing melanoma cases while minimizing misclassification. It has the highest true negatives (80) and lowest false positives (37), indicating the best performance in correctly identifying non-melanoma cases.

Therefore, based on the analysis of the confusion matrices,

Alexnet: Excels in detecting melanoma cases but has a higher rate of false positives and false negatives.

VGG16: Better at identifying non-melanoma cases compared to Alexnet but misses more melanoma cases.

Ensemble model: Combines the strengths of both Alexnet and VGG16, achieving the highest accuracy in identifying non-melanoma cases and maintaining strong performance in detecting melanoma.

Predictions:

Here , in this section, the images for Predicted vs True labels are shown for all the models:

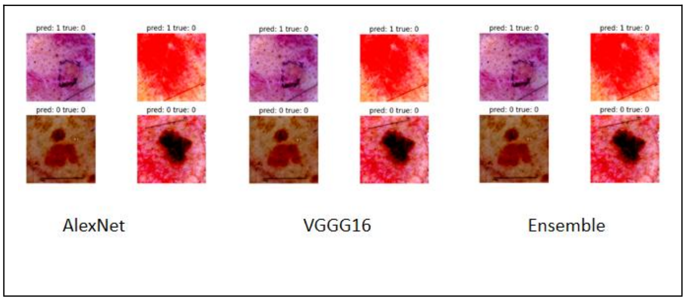


Fig 12. Prediction Images (Predicted vs True)

Interestingly , all three models performed identically on this set of images predictions, they correctly predicted the last two images but made incorrect predictions for the first two. This highlights that while ensemble methods can enhance accuracy, they are not immune to the limitations of their constituent models.

Discussion and Conclusion

The experimental results demonstrate that the ensemble model outperformed the individual AlexNet and VGG16 models in terms of precision and AUC. The ensemble model achieved a higher AUC (0.77) compared to AlexNet (0.76) and VGG16 (0.75), indicating better overall performance. The confusion matrices also show that the ensemble model had fewer false negatives and false positives compared to the individual models, making it more reliable for skin cancer classification.

This research demonstrated that fine-tuning pre-trained models with data augmentation and combining their predictions through ensemble methods can significantly enhance the performance of skin cancer classification tasks. The ensemble model, in particular, showed promising results, outperforming individual models and providing a balanced performance across various metrics.

This report highlights the promise of utilizing pre-trained models and ensemble techniques in medical image classification tasks, ultimately leading to more precise and dependable diagnostic tools.

Key Findings

Ensemble Model Superiority: The ensemble of AlexNet and VGG16 achieved the highest AUC and overall better performance metrics compared to individual models.

Model Robustness: Data augmentation techniques improved the robustness and generalization capability of the models.

Computational Efficiency: AlexNet provided a good balance between complexity and performance, suitable for environments with moderate computational resources, whereas VGG16, though more complex, delivered superior individual model performance.

Future Work

In respect of this research, the availability of computational resources significantly impacted this project. To address this challenge, allocating additional computational power would allow for testing more advanced models. Specifically, exploring efficient architectures like ResNet, GoogLeNet, and other state-of-the-art models could enhance skin cancer detection accuracy. Additionally, incorporating thses state-of-the-art models into ensemble methods—such as weighted averaging or model stacking—would further improve classification performance.

The quality and diversity of training and testing data significantly impact model generalization. Utilizing larger and more diverse skin cancer datasets would enhance the model's ability to handle variations in skin lesions. Collecting data from different populations, skin types, and clinical scenarios would contribute to robustness.

References

[1] N. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, K. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 ISIC," I*EEE Journal of Biomedical and Health Informatics,* vol. 23, no. 2, pp. 501-512, March 2019.

[2] J. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature,* vol. 542, no. 7639, pp. 115-118, Feb. 2017.

[3] N. Mishra, S. K. Sahu, and S. V. Shrivastava, "Deep convolutional neural network based analysis of X-ray images for the detection of COVID-19," *Pattern Recognition Letters,* vol. 139, pp. 180-187, Dec. 2020.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM,* vol. 60, no. 6, pp. 84-90, June 2017.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in P*roc. Int. Conf. Learning Representations* (ICLR), San Diego, CA, USA, May 2015, pp. 1-14.

6] A. M. Brinker, D. Hekler, S. Utikal, N. Grabe, F. H. Schadendorf, C. Klode, A. Berking, C. Steeb, and A. Enk, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *European Journal of Cancer*, vol. 118, pp. 91-96, June 2019.

[7] J. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, Feb. 2017.

[8] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proc. Int. Conf. Information Processing in Medical Imaging* (IPMI), Boone, NC, USA, June 2015, pp. 588-599.

[9] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M.H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Medical Image Analysis*, vol. 75, p. 102305, 2022. doi: https://doi.org/10.1016/j.media.2021.102305.

[10] A. Duval, "Explainable AI, the key to open 'black boxes,'" *Towards Data Science,* Jun. 30, 2021. [Online]. Available: https://towardsdatascience.com/explainable-ai-the-key-to-open-black-boxes-1234567890. [Accessed: May 24, 2024].

[11] M. Stojiljkovic, “Stochastic gradient descent algorithm with python ´and numpy,” Apr 2024. [Online]. Available: https://realpython.com/gradient-descent-algorithm-python/.[Accessed: May 24, 2024].

[12] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data,* 8(1), 53.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems,* 25, 1097-1105.

[14] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.