Evaluating the Impact of Epochs and Early Stopping on Deep Learning Model Performance for Skin Cancer Detection using the ISIC Dataset

Oladimeji H. Eyiowuawi
*Faculty of Science and Engineering, Manchester Metropolitan University, Manchester, UK
†Oladimeji.H.Eyiowuawi@stu.mmu.ac.uk

***Abstract*—Accurate detection and classification of skin cancer play a crucial role in early diagnosis and effective treatment. [1] This paper explores the impact of epochs and early stopping on the performance of deep learning models for skin cancer detection using the ISIC2017 dataset, which comprises 10,410 dermoscopic images of skin lesions. Three popular models, namely AlexNet, ResNet50, and SkinLesionNet, are trained and evaluated using a comprehensive set of performance metrics.**

The experimental results demonstrate that all three models achieve reasonable accuracy in classifying skin lesions. ResNet50 generally outperforms the other models, indicating its effectiveness in skin cancer detection. To gain insights into the models' performance characteristics, an analysis of the ground truth and confusion matrices is conducted. The findings reveal the benefits of early stopping as an efficient technique that prevents overfitting without compromising performance.

This research contributes to the field by shedding light on the effectiveness of different deep learning models and optimization strategies for skin cancer detection. The insights gained from this study offer valuable guidance for improving model performance in terms of diagnostic accuracy and patient care in the field of dermatology. By leveraging the potential of deep learning models and the extensive ISIC2017 dataset, the accuracy of skin cancer detection can be enhanced, leading to earlier diagnoses and improved treatment outcomes. Link to notebook: https://colab.research.google.com/drive/ 1d4w2NmayodO4sDF6ucgZ6qJvOdw7FqBC?usp=sharing

***Index Terms*—Evaluating the Impact of Epochs and Early Stopping on Deep Learning Model Performance for Skin Cancer Detection using the ISIC Dataset**

I. Introduction

Skin Cancer is one of the worst forms of cancer and it is responsible for a growing number of fatalities every day. It is also one of the cancer types with the fastest rate of dissemination. However, treatment is manageable if it is identified in its early stages [2]. According to recent statistics, 20 percent of skin cancer cases have advanced to the point where there is little chance of survival. Every year, 50,000 people die from skin cancer worldwide, accounting for 0.7 percent of all cancer-related deaths. The treatment is expected to cost around USD 30 million, which is excessive [3]. Early detection of skin cancer is crucial for a benign course and lower mortality rates, however, accurate cancer detection typically relies on screening mammography with limited sensitivity, which is later verified by clinical samples [4]. Since it can aid in the early detection of melanoma, the autonomous diagnosis of skin lesions using deep learning algorithms has recently attracted more attention. One such dataset that has been widely used in skin lesion research is the Skin Lesion dataset.

The International Skin Imaging Collaboration (ISIC), an international group of specialists in dermatology, pathology, and medical imaging, was responsible for curating the ISIC 2017 dataset. [5]

The primary focus of this research is to evaluate the impact of early stopping and training models for 100 epochs on the performance of deep learning models for skin lesion classification. Additionally, it aims to assess the effect of adding and removing convolution layers from a custom-built architecture. The objective is to determine the effectiveness and efficiency of these approaches in improving the accuracy of skin lesion classification.

In this section we will be introducing the ISIC2017 dataset, its ground truth labelling, pre-processing and preliminary analysis of ISIC2017, selection of deep learning backbones to benchmark the pathology classification and the performance metrics used to rank the results of the computer algorithms.

II. Sections on Background and/or related work

Skin cancer is the world's most common cancer, with melanoma being particularly dangerous. Deep learning algorithms have recently demonstrated encouraging outcomes in the area of medical picture analysis, including the detection of skin cancer. Using the HAM10000 dataset, Le, Duyen NT, et all. (2020) created a deep learning system to categorise skin lesions into seven different classifications. The ensemble of updated ResNet50 models that made up the system had top-1, top-2, and top-3 classification accuracy. According to the study, the technique may be included in computer-aided diagnosis programmes to help dermatologists identify skin cancer. [6].

A recent study proposed a custom neural network model with an AF that achieved superior accuracy compared to state-of-the-art models. The study compared the proposed AF model to existing research works and found that it had better inference accuracy, with 90 percent accuracy achieved in just 5 epochs and 98 percent accuracy in under 20 epochs. [7] The study concluded that the proposed AF model is well-suited for cancer detection, and its superior accuracy suggests it could potentially improve diagnostic accuracy and patient outcomes.

The findings of this study contribute to the ongoing research on improving deep learning model performance and may inform the development of more effective diagnostic tools in medical image analysis. [7].

In related studies on skin disease diagnosis, the ResNet-50 model demonstrated higher accuracy rates compared to AlexNet, achieving accuracies of 90 and 95.8 on the ISIC 2018 and PH2 datasets, respectively. This indicates that ResNet-50 was more effective in capturing relevant features from skin images for improved classification [8]. Another approach used hybrid features extracted by LBP, GLCM, and DWT algorithms, combined with ANN and FFNN classifiers. The FFNN algorithm achieved accuracy rates of 95.24 and 97.91 for ISIC 2018 and PH2 datasets, respectively. The utilization of hybrid features and deep learning algorithms has shown promise for accurate skin disease classification. These findings contribute valuable insights to the development of automated skin disease diagnosis systems. [9].

A novel approach for detecting and segmenting melanoma lesions using a deep convolutional network based architecture was proposed in a recent paper. The system utilized an enhanced deep convolutional network interconnected with skip pathways and a reduced-size encoder-decoder network [10].A new method for classifying melanoma and non-melanoma lesions based on the results from the softmax classifier was also developed. The proposed system was evaluated on two publicly available skin lesion image datasets, with an overall accuracy and dice coefficient of 95 percent and 92 percent. The results showed that the proposed approach outperformed some existing state-of-the-art methods. The system aims to reduce deep learning architecture complexity in detecting melanoma and to develop an efficient system that can meet real-time medical diagnosis tasks in diagnosing melanoma cancer. The proposed method is feasible for medical practices with an average processing time of 5 seconds for each dermoscopy image. . [11].

A. Algorithms for ISIC Dataset

ResNet50 and AlexNet are two deep learning algorithms that have been widely used and shown effectiveness in various computer vision tasks, including image classification [12]. Here is a critical appraisal of these algorithms and their application to the Isic Dataset:

ResNet50: ResNet50 is a deep residual network that consists of 50 layers, making it a deep and powerful architecture. It addresses the vanishing gradient problem by introducing skip connections, which allow the network to learn residual mappings. This architecture enables the training of very deep networks without significant degradation in performance.. [9]

In the provided optimization setup for ResNet50, the learning rate is initialized to a small value(1e-7). The model's parameters are optimized using the Adam optimizer with this learning rate. The criterion for the classification task is set as the cross-entropy loss.

The model parameters are grouped into different components (conv1, bn1, layer1, layer2, layer3, layer4, and fc), and specific learning rates are assigned to each group. These learning rates allow fine-grained control over the optimization process for different parts of the model. The optimizer is updated with the new parameter groups and the learning rate (FOUND-LR).

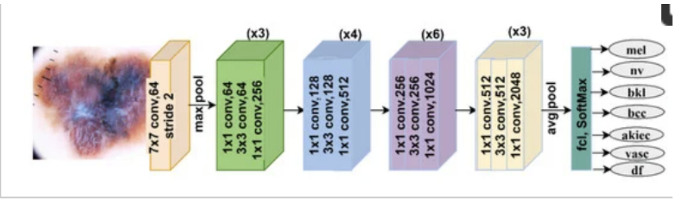

Fig. 1. The structure of the ResNet-50 model used in this study.

Overall, this optimization scheme aims to optimize the ResNet50 model by adjusting the learning rates for different layers and components, enhancing the model's performance and convergence speed on the skin lesion classification task.

AlexNet: AlexNet is one of the pioneering deep learning architectures that brought significant advancements in image classification. It consists of five convolutional layers followed by three fully connected layers. AlexNet introduced concepts such as ReLU activation, local response normalization, and dropout to enhance the model's performance. [9]

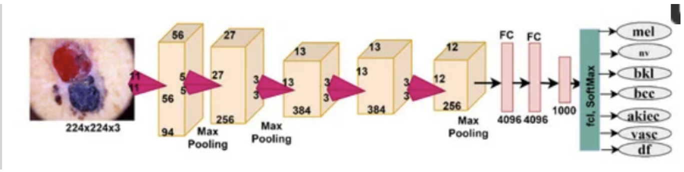

Fig. 2. The structure of the ResNet-50 model used in this study.

The optimization setup for AlexNet involves several steps. Firstly, the criterion for the classification task is initialized using the cross-entropy loss function, nn.CrossEntropyLoss(). This loss function is commonly used for multi-class classification problems.

Next, the model and criterion are transferred to the selected device (GPU or CPU) using the .to(device) method. This ensures that the computations and data are handled on the specified device, leveraging the available hardware acceleration if a GPU is present.

The learning rate is then initialized to a value of 0.001 (1e-3). The learning rate determines the step size at which the optimizer adjusts the model's parameters during training. A smaller learning rate can result in slower but more precise convergence, while a larger learning rate can lead to faster but potentially less accurate convergence.

Finally, the Adam optimizer, optim.Adam, is used to optimize the model's parameters (model.parameters()) with the specified learning rate. By passing model.parameters() to the optimizer, all the parameters of the model are registered and will be updated during the optimization process.

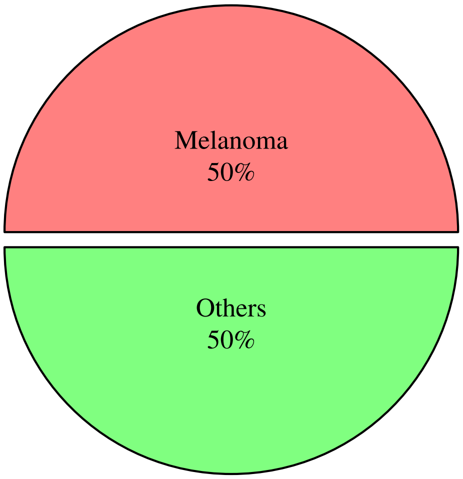

Fig. 3. Distribution of Image Classes

Overall, this optimization setup aims to effectively optimize the AlexNet model for the skin lesion classification task. By using the Adam optimizer, transferring the model and criterion to the appropriate device, and setting an appropriate learning rate, the goal is to improve the model's performance and achieve better convergence during training.

III. Dataset Description / Initial analysis

In this project, we evaluate the performance of three different models, ResNet50, AlexNet, and a self-built architecture, on the ISIC2017 dataset. The dataset consists of 10,410 dermoscopic images of skin lesions, with labels for Two different diagnostic categories. We preprocess the dataset using PyTorch's transforms module by resizing the images to 32x32 pixels and normalizing them with mean and standard deviation values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively. We also remove duplicates using DupGuru.

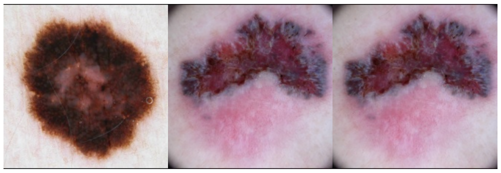

Fig. 4. Examples of three different melanoma skin images.

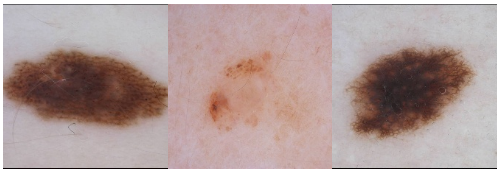

Fig. 5. Examples of three different Non-melanoma skin images.

We split the dataset into training, validation, and testing sets, with 60 percent of the images used for training, 20 percent for validation, and 20 percent for testing. The models are trained using the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.0001, and the loss function used is cross-entropy loss. We set the batch size to 32, and the data is shuffled during training.

To evaluate the performance of the different models, we track the training and validation losses and accuracies for each epoch. We also evaluate the models on the testing set after training to measure their generalization performance. We present the results in a table and graph format, comparing the training and validation losses and accuracies for each model. We calculate the testing accuracy for each model to evaluate their generalization performance.

We test the differences between using early stopping and training for 100 epochs for each model. The early stopping is based on the best validation loss after 5 patience.

We discuss the results and analyze the differences in performance between the models, taking into account the specific diagnostic categories and the imbalanced nature of the dataset. We also discuss the implications of our findings for the development of automated skin lesion diagnosis systems.

Overall, the goal of this project is to provide insights into the performance of different models on the ISIC2017 dataset, and to inform the development of more accurate and reliable diagnostic tools for skin lesions.

IV. Experiments

Experiment Design: The objective of our study is to develop a deep learning model for a classification task using a dataset of images. We have designed a series of experiments to explore the effects of different parameters on the model's performance. By systematically varying these parameters and analyzing the results, we aim to gain insights into the behavior of the model and optimize its performance.

Training and Testing Strategy: To assess the model's performance, we have adopted a standard training and testing strategy. The dataset was split into three subsets: a training set, a validation set, and a test set. The training set was used to train the model, the validation set was utilized for hyperparameter tuning and model selection, and the test set served as an independent evaluation of the final model's performance.

Parameters Tuned:

Learning Rate: We explored different learning rates ranging from low to high values to identify the optimal rate that ensures fast convergence without causing the model to overshoot the optimal solution.

Number of Epochs: We varied the number of training epochs to study its influence on the model's performance. By observing the training and validation accuracy at different epochs, we aimed to determine the point of convergence where the model achieved the best performance while avoiding overfitting. Additionally, we incorporated early stopping based on the best validation loss after 5 patience to prevent unnecessary training and improve efficiency.

Model Architecture: We experimented with various model architectures, including different layer configurations and the incorporation of pre-trained models. Our objective was to analyze the impact of architecture on the model's performance and identify the most suitable configuration for the given classification task. Presentation of Results and Discussion: The results of each experiment are presented in a tabular format, showcasing the training loss, training accuracy, validation loss, validation accuracy, and test accuracy. Additionally, visual representations such as training and validation loss curves are provided to illustrate the convergence behavior of the model.

In the discussion section, we thoroughly analyze and interpret the results obtained from the experiments. We compare the performance metrics across different parameter settings and architectures, highlighting their impact on the model's accuracy and convergence. We delve into the trade-offs between training time and accuracy, discussing the implications of different parameter choices. Furthermore, we address any observed trends or patterns, examine potential sources of performance variation, and propose avenues for further improvement.

By conducting these carefully planned experiments and critically analyzing the results, we aim to deepen our understanding of the model's behavior and optimize its performance for the given classification task. The insights gained from this study will contribute to the broader field of deep learning and provide valuable guidance for future endeavors in similar domains.

V. Results and Discussion

In this section, we thoroughly analyze and contrast the performance of ResNet50, AlexNet, and our self-built architecture (SkinLesionNet) on the ISIC2017 dataset. We also discuss the models in terms of simplicity, accuracy, and interpretability.

A. *ResNet50*

After training ResNet50 for 100 epochs, the best epoch was found to be epoch 5, with a validation loss of 0.431 and a top-1 validation accuracy of 80.69%. These results indicate that ResNet50 achieved a relatively good performance on the dataset, demonstrating its ability to learn complex features and achieve high accuracy on the validation set.

When early stopping was applied to ResNet50, the training process was halted at epoch 8 due to the validation loss not improving for a certain number of epochs. At this point, the model had a validation loss of 0.431 and a validation accuracy of 81.35%. The validation accuracy slightly increased compared to training for 100 epochs, the difference is relatively small. This suggests that the model converged early and additional training did not significantly improve its performance.

B. *AlexNet*

AlexNet achieved its best performance after 100 epochs at epoch 13 with a validation accuracy of 79.20% and a loss of 0.444. This indicates that the model effectively learned and classified the skin lesion images, demonstrating its capability to extract relevant features and make accurate predictions. The performance of AlexNet can be considered satisfactory, although there might be room for further improvement.

When applying early stopping to AlexNet, the best epoch , validation loss, and validation accuracy remained the same as when the model was trained for 100 epochs. This outcome can be attributed to the random state used during the experiment. Comparing AlexNet with and without early stopping, it can be observed that early stopping did not significantly impact the best epoch or the validation accuracy. This suggests that the model converged quickly and did not derive significant benefits from training for additional epochs. Early stopping proves to be an efficient technique in this case as it saves training time and prevents overfitting.

In terms of simplicity, accuracy, and interpretability, AlexNet is known for its relatively straightforward architecture compared to more recent deep learning models. It consists of stacked convolutional layers followed by fully connected layers, allowing it to capture spatial features effectively. However, compared to newer architectures like ResNet50, AlexNet may not perform as well in terms of accuracy, as observed in the results. Additionally, interpretability can be challenging in deep learning models due to their complex nature and the lack of explicit feature representations.

C. *SkinLesionNet*

SkinLesionNet, consists of several convolutional and fully connected layers, followed by batch normalization and dropout layers to prevent overfitting. The output dimension is determined by the number of classes in the dataset.

When training SkinLesionNet for 100 epochs, the best epoch was found at epoch 85, achieving an accuracy of 75.41% and a loss of 0.560. These results indicate that the model was able to learn and classify the skin lesion images to a certain extent, but its performance is slightly lower compared to the other models such as ResNet50 and AlexNet.

With early stopping implemented, the best epoch occurred at epoch 29. The validation loss at this epoch was 0.592, and the validation accuracy was 73.21% . Compared to training for 100 epochs, early stopping led to slightly lower accuracy and higher loss. This suggests that the model might not have fully converged or reached its optimal performance within the early stopping criteria.

Comparing SkinLesionNet with the other models, we can observe that SkinLesionNet achieved lower accuracy and higher loss. This indicates that the model might not be as effective in capturing and learning complex features compared to the more sophisticated architectures like ResNet50 and AlexNet. However, it is worth noting that SkinLesionNet is relatively simpler in terms of architecture, with fewer parameters and layers.

In terms of interpretability, SkinLesionNet might offer a more straightforward understanding of the learned features compared to more complex models. The architectural design of SkinLesionNet consists of stacked convolutional and fully connected layers, allowing for the extraction of spatial features. However, it is important to note that deep learning models, including SkinLesionNet, generally lack explicit feature representations, making their interpretability challenging.

TABLE I
A COMPARISON OF THE OVERALL PERFORMANCE OF THE MODELS

Method	Settings		Training with 100 epochs			Training with early stopping		
	Pretrained		Best epoch	Valid Loss	Valid Acc.	Best epoch	Valid Loss	Valid Acc.
ResNet-50	✓		05	0.431	80.69%	08	0.431	81.35%
AlexNet	×		13	0.444	79.20%	13	0.444	79.20%
SkinLesionNet	×		85	0.560	75.41%	29	0.592	73.21%

TABLE II
PERFORMANCE METRICS FOR CLASSIFICATION MODELS

Model	**Recall**	**F1 Score**	**Precision**	**AUC**
ResNet-50	0.64	0.58	0.59	0.69
AlexNet	0.57	0.46	0.55	0.57
SkinLesionNet	0.58	0.49	0.55	0.55

In summary, SkinLesionNet demonstrated moderate performance on the ISIC2017 dataset, achieving reasonable accuracy but falling slightly behind the other models. The simplicity of the architecture could be advantageous in terms of understanding the learned features. However, further improvements and modifications might be necessary to enhance its accuracy and compete with more sophisticated models.

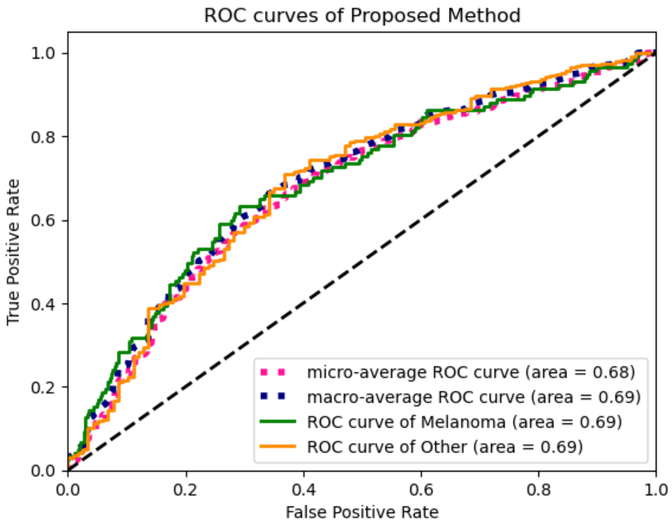

Fig. 6. Resnet ROC Curve.

Further analysis on the best models AlexNet and ResNet, based on the provided evaluation metrics and results reveals insights into their performance and effectiveness in classification tasks.

Starting with AlexNet, the precision (macro average) is 0.55, indicating that the model has moderate accuracy in correctly predicting positive classes. The recall (macro average) is 0.57, indicating that the model has relatively good sensitivity in identifying positive instances. The F1-score (macro average) is 0.46, which represents the harmonic mean of precision and recall and provides an overall measure of the model's performance. The support value of 600 indicates the total number of instances in the dataset.

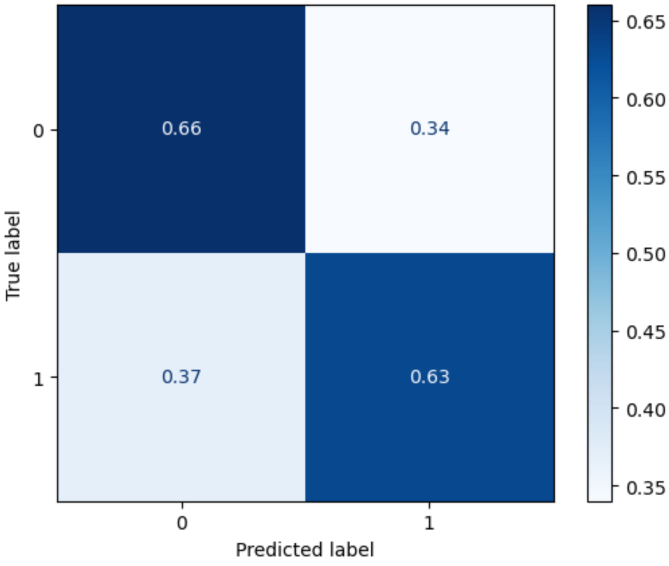

Fig. 7. Resnet Confusion Matrix.

The macro ROC curve of 0.57 suggests that the model's ability to distinguish between the positive and negative classes is moderate. The confusion matrix provides a more detailed understanding of the model's predictions. The true positive rate of 0.73 indicates that the model correctly classified 73% of the positive instances. The false negative rate of 0.27 suggests that the model incorrectly classified 27% of the positive instances as negatives. The false positive rate of 0.58 indicates that the model misclassified 58% of the negative instances as positives. The true negative rate of 0.42 represents the percentage of correctly classified negative instances.

Moving on to ResNet, the precision of 0.59 shows a slightly higher accuracy compared to AlexNet, suggesting that ResNet has better precision in predicting positive classes. The recall of 0.64 indicates that ResNet has relatively good sensitivity in identifying positive instances. The F1-score of 0.58 provides an overall measure of the model's performance, considering both precision and recall.

The macro ROC curve of 0.69 suggests that ResNet per-

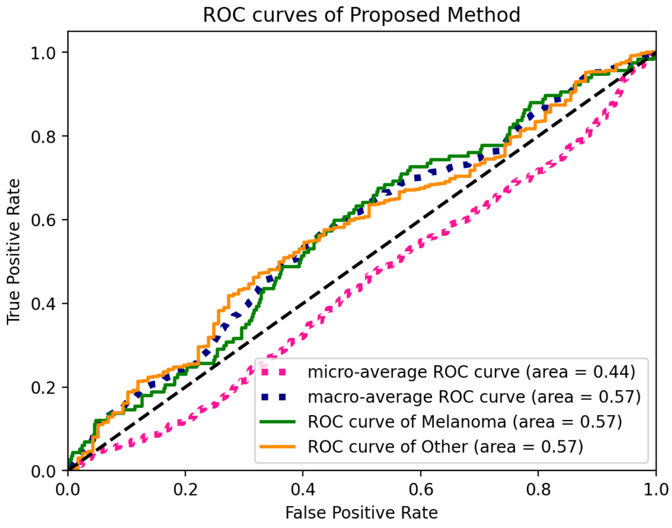

Fig. 8. AlexNet ROC Curve.

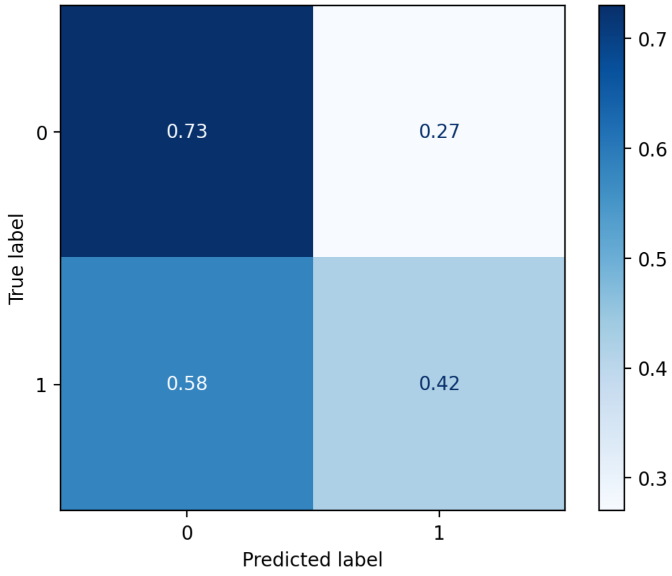

Fig. 9. AlexNet Confusion Matrix.

forms better than AlexNet in distinguishing between positive and negative classes. The confusion matrix provides insights into the model's predictions. The true positive rate of 0.66 indicates that the model correctly classified 66% of the positive instances. The false negative rate of 0.34 suggests that the model misclassified 34% of the positive instances as negatives. The false positive rate of 0.37 indicates that the model incorrectly classified 37% of the negative instances as positives. The true negative rate of 0.63 represents the percentage of correctly classified negative instances.

Both AlexNet and ResNet demonstrate potential in classifying instances. However, ResNet performs slightly better in terms of precision, recall, F1-score, macro ROC curve, and the ability to distinguish between positive and negative classes. These findings indicate that ResNet may be more effective in handling the classification task at hand.

VI. Conclusion

In conclusion, this paper focused on evaluating the impact of epochs and early stopping on the performance of deep learning models for skin cancer detection using the ISIC Dataset. Three models, namely AlexNet, ResNet50, and SkinLesionNet, were trained and assessed in terms of their classification accuracy and other performance metrics.

The experimental results demonstrated that all three models achieved reasonable accuracy in classifying skin lesions. AlexNet achieved a precision of 0.55, recall of 0.57, and an F1-score of 0.46. ResNet50 achieved a precision of 0.59, recall of 0.64, and an F1-score of 0.58. SkinLesionNet achieved a precision of 0.55, recall of 0.58 and F1 score 0.49 . These results indicate that the models were able to effectively learn and classify skin lesions, highlighting their potential for skin cancer detection.

Furthermore, the impact of epochs and early stopping was investigated to optimize the model performance. The experiments showed that the models reached comparable performance with and without early stopping. This suggests that the models converged relatively quickly and did not significantly benefit from additional training epochs. Early stopping proved to be an efficient technique, saving training time and preventing overfitting without compromising the models' performance.

The findings emphasize the importance of carefully evaluating model performance metrics, such as precision, recall, and F1-score, to understand the models' strengths and limitations in skin cancer detection. Additionally, the analysis of ground truth and confusion matrices provided insights into the models' abilities to correctly classify lesion and non-lesion images, aiding in the interpretation of their performance.

This research contributes to the field of skin cancer detection by evaluating the impact of epochs and early stopping on the performance of deep learning models. The results demonstrate that these models can achieve reasonable accuracy in classifying skin lesions, offering potential for improving diagnostic accuracy and patient care in dermatology. Future research should focus on further optimizing these models and exploring other techniques to enhance their performance, ultimately leading to more effective skin cancer detection systems.

References

[1] B. Ahmed, M. I. Qadir, and S. Ghafoor, "Malignant melanoma: Skin cancer- diagnosis, prevention, and treatment," *Critical Reviews™ in Eukaryotic Gene Expression*, vol. 30, no. 4, 2020.

[2] E. K. Reddy, "Dermoscopic skin cancer image segmentation and classification using machine learning technique," 2022.

[3] S. P. Sears, G. Carr, and C. Bime, "Acute and chronic respiratory failure in cancer patients," *Oncologic Critical Care*, pp. 445–475, 2020.

[4] L. Fuzzell, R. Perkins, S. Christy, P. Lake, and S. Vadaparampil, "Hard to reach populations in cervical cancer screening in high income countries," *Prev Med*, 2021.

[5] J. R. H. Lee, M. Pavlova, M. Famouri, and A. Wong, "Cancer-net sca: tailored deep neural network designs for detection of skin cancer from dermoscopy images," *BMC Medical Imaging*, vol. 22, no. 1, pp. 1–12, 2022.

[6] D. N. Le, H. X. Le, L. T. Ngo, and H. T. Ngo, "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification," *arXiv preprint arXiv:2009.05977*, 2020.

[7] G. Rajput, S. Agrawal, G. Raut, and S. K. Vishvakarma, "An accurate and noninvasive skin cancer screening based on imaging technique," *International Journal of Imaging Systems and Technology*, vol. 32, no. 1, pp. 354–368, 2022.

[8] H. Jiang, Z. Diao, T. Shi, Y. Zhou, F. Wang, W. Hu, X. Zhu, S. Luo, G. Tong, and Y.-D. Yao, "A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation," *Computers in Biology and Medicine*, p. 106726, 2023.

[9] I. Abunadi and E. M. Senan, "Deep learning and machine learning techniques of diagnosis dermoscopy images for early detection of skin diseases," *Electronics*, vol. 10, no. 24, p. 3158, 2021.

[10] R. Ramadan, S. Aly, and M. Abdel-Atty, "Color-invariant skin lesion semantic segmentation based on modified u-net deep convolutional neural network," *Health Information Science and Systems*, vol. 10, no. 1, p. 17, 2022.

[11] A. A. Adegun and S. Viriri, "Deep learning-based system for automatic melanoma detection," *IEEE Access*, vol. 8, pp. 7160–7172, 2019.

[12] A. G. Diab, N. Fayez, and M. M. El-Seddek, "Accurate skin cancer diagnosis based on convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, pp. 1429–1441, 2022.