JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024

Research Article
https://doi.org/10.56979/801/2024

Neural Network Based Skin Cancer Classification from Clinical Images: Accuracy and Robustness Analysis

Rana Arbab Haider[1], Khadija Zafar[2], Sana Basharat[3*], and Muhammad Faheem Khan[4]

[1]Department of Computer Science, Muhammad Nawaz Sharif University of Engineering & Technology, Multan, Pakistan.
[2]Department of Computer Science, University of Agriculture, Faisalabad, Pakistan.
[3]Department of Computer Science, University of Management and Technology, Lahore, Pakistan.
[4]Department of Computer Science, TIMES Institute, Multan, 60000, Pakistan.
[*]Corresponding Author: Sana Basharat. Email: Sana.basharattt@gmail.com

Received: April 11, 2024 **Accepted:** September 21, 2024

Abstract: This study aims to investigate the use of neural networks in classifying skin cancer from clinical images, with a specific emphasis on evaluating accuracy and robustness. Skin cancer, melanoma included, poses a significant worldwide health challenge, and early detection is vital for enhancing patient prognoses. Conventional diagnostic approaches heavily depend on clinical expertise, which can be subjective and inconsistent. The progress in deep learning, especially convolutional neural networks (CNNs), presents a promising alternative by automating skin lesion classification with high accuracy. The research involves developing a neural network model using a varied set of clinical skin images, enabling it to distinguish between benign and malignant lesions. Multiple architectures are tested, and their effectiveness is assessed using standard metrics such as accuracy, precision, recall, and F1-score. Beyond measuring overall accuracy, the study emphasizes robustness by evaluating the model in challenging conditions, including variations in illumination, obstructions, and diverse skin tones. Results indicate that neural networks can achieve superior accuracy in skin cancer classification, often outperforming traditional diagnostic techniques. However, robustness remains a crucial area for enhancement, particularly in real-world applications where image quality and patient diversity can fluctuate significantly. By examining the strengths and weaknesses of neural network-based models, this research underscores the potential of AI in clinical diagnostics while highlighting the necessity for further improvements in model generalization to ensure reliable implementation in healthcare settings.

Keywords: Cancer; Classification; Skin Cancer; Neural Networks.

1. Introduction

The increasing incidence of skin cancer, especially melanoma, has necessitated the development of more precise and efficient diagnostic methods. Conventional approaches to skin cancer diagnosis often involve clinical examinations and histopathological evaluations, which can be lengthy processes and potentially delay treatment. In recent times, neural networks have emerged as potent instruments for automating skin cancer diagnosis from clinical images, offering the prospect of swifter and more accurate detection. These networks, particularly deep learning models like convolutional neural networks (CNNs), are engineered to handle intricate visual data, making them ideal for analyzing dermatological images [1].

Studies in this field have indicated that neural networks can perform on par with, and sometimes surpass, dermatologists in categorizing skin lesions as benign or malignant. Research has shown high levels of accuracy, with some models achieving sensitivity and specificity rates comparable to or exceeding those of seasoned clinicians [2]. Nevertheless, despite these encouraging outcomes, there remains a need to evaluate the resilience of these models across diverse datasets and under various clinical circumstances. Factors such as differences in lighting, skin tone, and lesion presentation can introduce biases that affect the generalizability of neural network models [3]. Consequently, it is essential to conduct a comprehensive assessment of both accuracy and robustness when evaluating neural network-based skin cancer classification systems.

Alongside accuracy, robustness is a crucial metric in determining the clinical feasibility of these models. A robust model must not only excel on its training data but also maintain its accuracy when exposed to new, unfamiliar data. This can be particularly challenging in skin cancer classification, where variations in image quality and patient demographics can significantly impact model performance. Scientists have investigated various techniques to enhance robustness, including data augmentation, domain adaptation, and adversarial training [4]. These methods aim to improve the model's ability to generalize across diverse patient populations and imaging conditions, ensuring that the neural network remains dependable in real-world clinical environments.

As neural networks continue to advance and improve, their application in skin cancer classification shows great potential for revolutionizing dermatological practice. However, for these models to gain widespread clinical acceptance, it is vital to address concerns regarding their accuracy, robustness, and generalizability. By conducting rigorous evaluations across multiple datasets and clinical settings, researchers can ensure that neural network-based diagnostic tools are both effective and reliable in supporting skin cancer detection and treatment. Timely identification of skin cancer is vital for successful treatment and improved patient prognosis, particularly for aggressive forms such as melanoma. Conventional diagnostic approaches, which often involve visual examination followed by tissue sampling, can be subjective and inconsistent based on the physician's expertise. In recent times, machine learning methodologies have emerged as promising tools for automating skin cancer detection, offering a more impartial and efficient alternative. Among these techniques, deep learning models, specifically convolutional neural networks (CNNs), have demonstrated significant success in categorizing skin cancer from dermoscopic and clinical images [1].

Machine learning algorithms can recognize patterns and characteristics in images that may be undetectable to human observers. These models are educated using extensive datasets of labeled skin images, enabling them to differentiate between benign and malignant lesions. A notable advantage of machine learning approaches is their ability to scale, as they can swiftly analyze large volumes of data with minimal human oversight. For example, studies have shown that machine learning models can achieve dermatologist-level accuracy in identifying melanoma and other skin cancers [5].

Despite the accomplishments of these models, obstacles persist in guaranteeing their precision and applicability across diverse populations. Skin cancer manifestations can vary depending on factors such as skin color, lesion type, and imaging conditions. These variations can impact the performance of machine learning models if not properly addressed during training. Methods like data augmentation, which artificially expands the training dataset by creating modified versions of the images, have been employed to enhance model robustness (Codella et al., 2018). Furthermore, efforts are underway to develop more diverse datasets that encompass a wide range of skin types and cancer presentations, to mitigate biases that could otherwise distort model predictions [6].

As machine learning technology advances, its application in skin cancer detection shows considerable promise for enhancing clinical workflows. Automated systems based on machine learning can support dermatologists in making quicker, more accurate diagnoses, ultimately leading to earlier interventions and better patient outcomes. However, ongoing research is crucial to address limitations related to dataset diversity, model interpretability, and real-world implementation to ensure these tools are dependable and effective in clinical environments.

2. Literature Review

Among the most prevalent forms of cancer, skin cancer stands out, with melanoma being its most dangerous variant. The chances of survival significantly improve with early identification, and in recent times, artificial intelligence (AI), particularly neural networks, has emerged as a promising tool for classifying skin cancer. Convolutional neural networks (CNNs), a specific type of neural network, have exhibited remarkable accuracy in identifying melanoma and other skin cancers from clinical images, including dermoscopic ones. These CNNs show promise in automating skin cancer detection, offering instantaneous, non-invasive diagnostic tools that can enhance dermatologists' expertise.

A primary advantage of employing CNNs for skin cancer classification is their capacity to learn hierarchical representations of image data. Through multiple layers, CNNs extract image features, beginning with basic patterns like edges and progressing to more intricate features such as shapes and textures. Research has indicated that CNN models trained on extensive skin image datasets can sometimes surpass dermatologists in performance. For example, Esteva et al. (2017) developed a CNN using over 129,000 clinical images and showed that the model could perform comparably to dermatologists in differentiating between benign and malignant lesions. The model's accuracy matched that of human experts, achieving an area under the receiver operating characteristic curve (AUC) of 0.96 [1] [7].

The durability of neural network-based skin cancer classification systems is another crucial factor to evaluate. Durability refers to the model's ability to sustain high performance despite variations in image quality, lighting conditions, or diverse skin types. A study conducted by Tschandl et al. (2019) emphasized the difficulties in generalizing neural networks across various populations and imaging conditions. Although CNNs have demonstrated high accuracy on carefully curated datasets, their performance may decline when applied to images from different clinics or devices. This constraint suggests that CNN models need training on diverse and representative datasets to ensure durability in real-world scenarios. Furthermore, additional validation of these models across various demographic groups is necessary, as most datasets used for training neural networks often favor certain skin types, potentially resulting in classification bias (Tschandl et al., 2019). Although Convolutional Neural Networks (CNNs) show promise in detecting skin cancer, challenges persist regarding their tendency to overfit and their ability to generalize. Overfitting occurs when a model excels with training data but struggles with new, unseen information. To address these issues, researchers have implemented strategies such as data augmentation, transfer learning, and cross-validation. These techniques aim to enhance the robustness and generalizability of CNN models. For example, data augmentation artificially expands the training dataset by applying random transformations to images, helping the model learn invariant features [8]. Transfer learning has also proven effective in improving skin cancer classification. This approach involves fine-tuning a model pre-trained on a large dataset using a smaller, specialized dataset. Networks like ResNet, VGGNet, and Inception, initially trained on ImageNet, have been adapted for skin cancer datasets, resulting in improved classification accuracy (Brinker et al., 2019). This method allows the network to apply knowledge from other domains to skin cancer detection, reducing the need for extensive, annotated medical datasets, which are often scarce.

The interpretability of neural networks is crucial for their adoption in clinical settings. Dermatologists require explanations for AI-driven decisions, particularly for critical diagnoses like skin cancer. While CNNs are often considered "black boxes" due to their complex structures, techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) have been developed to provide visual explanations for model predictions. Grad-CAM identifies the image regions most influential in the model's decision-making process, offering a level of interpretability essential for gaining clinicians' trust [9].

In summary, CNN-based skin cancer classification from clinical images shows great potential, demonstrating high accuracy and the ability to complement human expertise. However, for these models to be fully integrated into clinical practice, issues related to robustness, generalizability, and interpretability must be resolved. Future research should concentrate on diversifying training datasets, developing techniques to improve model robustness, and ensuring that AI-powered diagnostic tools are transparent and explainable to healthcare professionals.

Skin cancer detection has become a significant focus in medicine, as early and precise diagnosis is vital for improving patient outcomes. Machine learning, especially in image processing, has emerged as a valuable tool for enhancing the accuracy and efficiency of skin cancer detection. Scientists have employed various machine learning methods to classify skin lesions, including melanoma, basal cell carcinoma, and squamous cell carcinoma, utilizing both dermoscopic and clinical images. Skin cancer detection is an area where machine learning algorithms excel, as they can process extensive datasets and recognize subtle patterns that humans might miss. Among these algorithms, convolutional neural networks (CNNs), a type of deep learning model, have demonstrated exceptional results in image-related tasks, including identifying skin cancer. A study by Esteva et al. (2017) showcased the effectiveness of CNNs in this field by training a model using over 129,000 skin lesion images. The model's performance rivaled that of skilled dermatologists, achieving accuracy levels that matched or surpassed human experts in detecting melanoma and other skin cancer varieties [10]

While CNNs have shown remarkable success, researchers have also investigated traditional machine learning techniques for skin cancer detection, such as support vector machines (SVM) and random forests. These methods typically require experts to manually extract relevant features from images, like texture, color, or shape. Although these conventional approaches have shown some success in differentiating between malignant and benign lesions, they generally fall short of CNNs in performance. The superiority of CNNs stems from their ability to automatically learn hierarchical representations of input data, eliminating the need for manual feature engineering [11]. This capability allows CNNs to learn directly from raw image data, contributing to their enhanced performance in skin cancer detection tasks. Data augmentation is a common technique used in machine learning to address the limited availability of labeled skin cancer images. By applying transformations such as rotation, scaling, and flipping, researchers can increase the diversity of training data, helping models generalize better to unseen examples. In a study by [12], the authors used data augmentation to enhance the performance of their CNN model in detecting melanoma. Their findings showed that data augmentation significantly improved the model’s ability to classify skin lesions, achieving an accuracy of over 95% in some cases, which is comparable to the performance of dermatologists [13]. Transfer learning is a crucial method widely employed in skin cancer identification. This technique involves fine-tuning a model, initially trained on a vast dataset like ImageNet, using a smaller set of skin cancer images. This approach enables models to leverage knowledge from general image classification tasks and apply it to more specialized tasks, such as differentiating various skin cancer types. A study by Brinker et al. (2019) showcased the efficacy of transfer learning by fine-tuning a pre-trained CNN on a smaller set of dermoscopic images. The resulting model exhibited high accuracy in identifying melanoma and other skin cancers, underscoring the potential of transfer learning to address the issue of limited medical datasets [14].

Although machine learning has made significant progress in skin cancer detection, several obstacles must be overcome before these models can be broadly implemented in clinical environments. One primary challenge is the variation in image quality, which can impact model performance. Skin cancer images can differ greatly in terms of lighting, resolution, and the presence of artifacts like hair or shadows. [15] stressed the importance of training models on diverse datasets to ensure their ability to generalize effectively to images from various clinical settings. Their research revealed that while machine learning models performed well on curated datasets, their effectiveness decreased when tested on images from different sources, emphasizing the need for robust skin cancer detection models [16].

An additional challenge is the interpretability of machine learning models. Despite CNNs demonstrating impressive performance in skin cancer detection, their "black box" nature makes it difficult for healthcare professionals to comprehend how the model reaches a specific diagnosis. Researchers have developed techniques such as class activation mapping (CAM) to visualize the most significant parts of an image for a model's decision. For instance, Grad-CAM generates heatmaps that highlight the areas of the image the model deems most relevant for its classification. [17] Introduced Grad-CAM as a method to enhance the transparency of CNN models, assisting clinicians in interpreting model predictions and fostering trust in AI-driven diagnostic tools [18] [19].

In summary, machine learning has made remarkable advancements in skin cancer detection, with CNNs at the forefront due to their capacity to automatically learn from extensive image datasets. Techniques like data augmentation and transfer learning have further enhanced these models' performance, making them comparable to human experts. However, challenges related to data variability, robustness, and interpretability must be resolved to ensure the safe and effective integration of machine learning models into clinical practice.

3. Methodology

The methodology of this research is given as

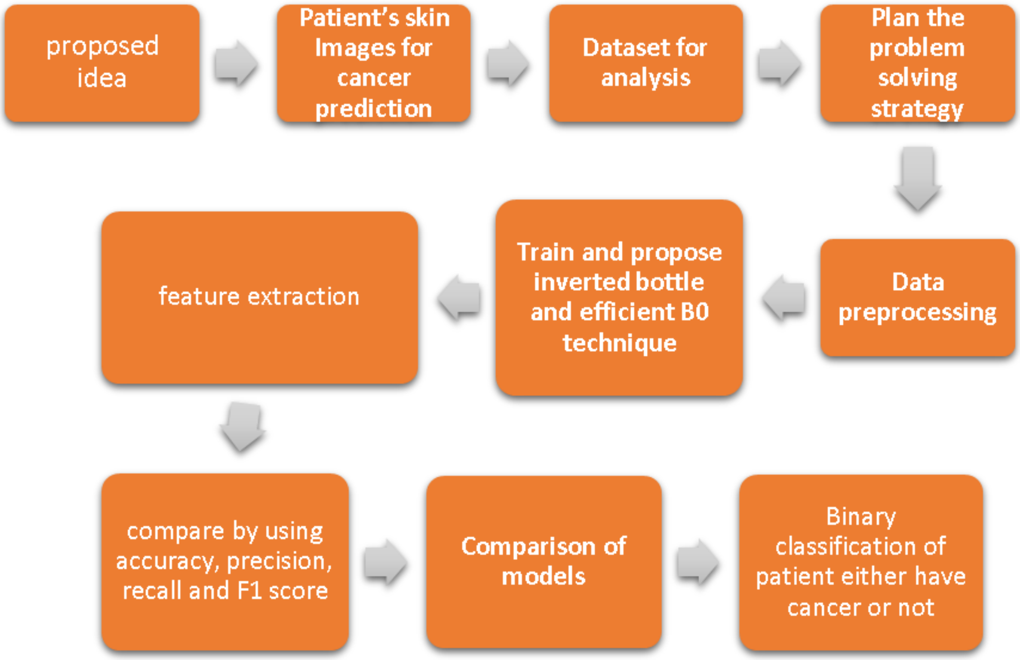

proposed idea
Patient's skin Images for cancer prediction
Dataset for analysis
Plan the problem solving strategy
Data preprocessing
Train and propose inverted bottle and efficient B0 technique
feature extraction
compare by using accuracy, precision, recall and F1 score
Comparison of models
Binary classification of patient either have cancer or not

Figure 1. Block diagram

The diagram illustrates a workflow for creating a machine learning system to classify skin cancer. This process begins with an initial concept and moves through several crucial stages. At the outset, patient skin images are gathered to serve as the foundation for cancer prediction. These images are subsequently compiled into a dataset for further examination. After dataset creation, a problem-solving approach is developed to guide the cancer prediction process. With the approach established, data preprocessing commences. This step involves refining and converting the raw data to ensure its suitability for analysis. This stage may include tasks such as data normalization, image augmentation, or addressing missing information. The dataset is accessible at no cost through SIM ISIC MELANOMA Classification 2020.

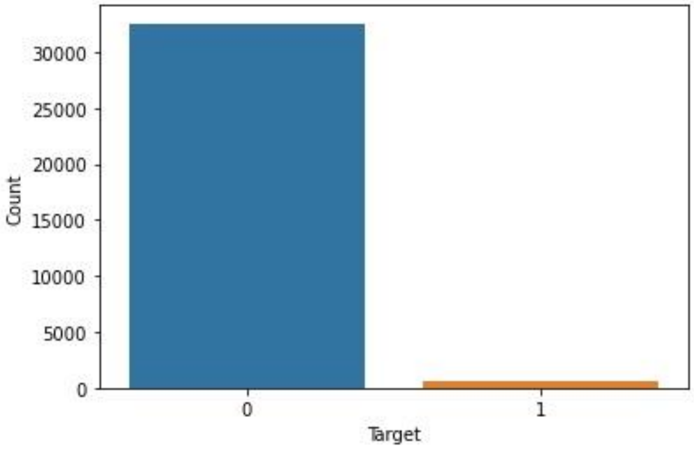


Figure 2. Dataset Description

Following preprocessing, the system moves to feature extraction, where crucial characteristics are identified from the processed images. These attributes play a vital role in educating the model to distinguish between malignant and benign skin lesions. The subsequent phase involves model training, utilizing an inverted bottle architecture in conjunction with a sophisticated Bayesian Optimization (BO) method to enhance the model's efficacy. Once trained, the model is employed to categorize images, yielding a binary outcome indicating the presence or absence of skin cancer in a patient. Subsequently, various models undergo comparison to determine which one delivers superior results in cancer detection.

The final stage involves evaluating the models' performance using metrics such as accuracy, precision, recall, and F1 score. These measurements are crucial in assessing the overall effectiveness of the model in accurately identifying skin cancer cases. The process flowchart thus presents a methodical approach to developing and refining a machine learning model for skin cancer detection, with emphasis on data processing, model optimization, and performance assessment.

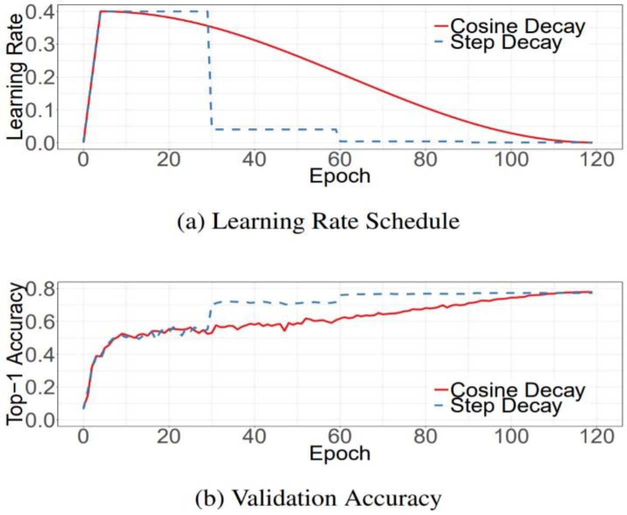
4. Results

For this project, we employed a collective strategy, training multiple models and combining their probability rankings to produce the final prediction. We opted for pre-trained EfficientNet variants B4, B5, and B7, rather than B0, due to their superior performance in the ImageNet competition. The models were trained across nine categories, with original images trimmed to 768x768 and 512x512 pixels to minimize random noise and eliminate black borders.

To accommodate GPU memory limitations, we downsized the input images to 380x380 and 448x448 pixels. Ideally, we would have used the original cropped dimensions, but current constraints made this unfeasible. Our learning rate strategy incorporated Cosine Decay with initial values of 3e-5 and 1e-5, coupled with a single warmup epoch. This warmup approach gradually increased the learning rate from zero to the target value during the initial training stages, enhancing stability. The Cosine Decay and warmup combination allowed for a smooth reduction in learning rate, avoiding sudden changes that could disrupt training. We favored Cosine Decay over exponential or step decay methods, as it slows the learning rate at the beginning and end while decreasing linearly in the middle, thus improving the training process.

We selected the Adam optimizer for its ability to handle sparse gradients in noisy data, combining the advantages of RMSProp and AdaGrad. Adam's adaptive learning rate capabilities made it particularly well-suited for our dataset. In our ensemble approach, each model underwent 15 epochs of training. Memory constraints necessitated a training and validation batch size of 8 for EfficientNet-B4, and 4 for EfficientNet-B5 and B7. Without these limitations, a batch size of 64 would have been preferable to enhance gradient estimation and speed up convergence.

As depicted in Table 1, all EfficientNet models exhibited comparable training and validation accuracy. Nevertheless, the inference results presented in the same table revealed that the EfficientNet B5 model yielded the highest accuracy on the dataset. To generate the final prediction, we utilized a straightforward method of averaging the probability rankings from all three models. Prior to this averaging process, we normalized the probability predictions to a range of [0, 1] to ensure uniformity across the models.



(a) Learning Rate Schedule

(b) Validation Accuracy

Figure 3. Validation Accuracy

Table 1. Accuracy Comparison

Model No	Backbone	Image Input Size	Resize	Batch Size	Training Accuracy	Validation Accuracy
2	B4	768	380	8	81%	91%
10	B5	512	448	4	85%	93%
16	B7	768	380	4	83%	92%
Ensemble					83%	92%

The performance of three EfficientNet models (B4, B5, and B7) with varying configurations is presented in the table. EfficientNet-B4, utilizing a 768x768 pixel input resized to 380x380 and a batch size of 8, obtained 81% training accuracy and 91% validation accuracy. EfficientNet-B5, employing a 512x512 input resized to 448x448 with a batch size of 4, achieved 85% training accuracy and 93% validation accuracy. EfficientNet-B7, using a 768x768 pixel input resized to 380x380 and a batch size of 4, reached 83% training accuracy and 92% validation accuracy. The combined performance of these models resulted in an overall training accuracy of 83% and validation accuracy of 92%, showcasing consistent performance across the models.

The EfficientNet B4 model exhibits good generalization on the unseen validation dataset. Extended training over more epochs could potentially yield higher accuracy. The validation loss shows initial fluctuations in the early epochs but stabilizes towards the end of the training process, as illustrated in Figure 22. Both training and validation losses show a consistent downward trend throughout, indicating that increasing the number of training epochs might further improve the model's performance. The section on Limitations, Future Extensions, and Improvements provides additional information on potential enhancements.

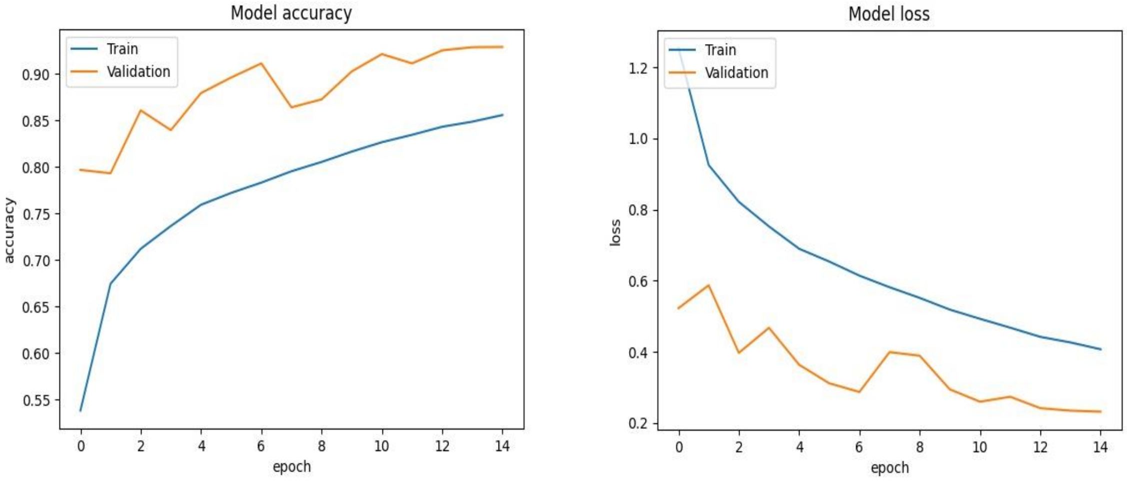


Figure 4. Model Accuracy and Loss

This project primarily concentrated on enhancing model accuracy rather than optimizing it for deployment. When preparing a model for serving, three key aspects must be considered: the size of the model, its speed in making predictions, and its throughput. The EfficientNet B4, B5, and B7 models have substantial unoptimized weights of 1.64 GiB, 2.56 GiB, and 2.78 GiB, respectively. These large sizes necessitate GPU usage for loading and are not suitable for deployment scenarios.

To tackle this issue, the original model weights were transformed into the Open Neural Network Exchange (ONNX) format. This format provides benefits such as interoperability and flexibility across different hardware platforms. Additionally, this conversion resulted in a 36% reduction in model size, which led to a notable improvement in inference time. The optimized ONNX model was subsequently integrated into a CAD system designed to assist dermatologists. This system processes skin lesion images and patient demographic information as inputs and generates probabilities for nine distinct classes.

5. Conclusion

To summarize, we employed a combination of EfficientNet architectures (B4, B5, and B7) to enhance accuracy and robustness in a multi-class classification problem. The neural networks were trained using a Cosine Decay learning rate schedule and the Adam optimization algorithm. To address GPU limitations, we resized images and utilized smaller batch sizes. Among the individual models, EfficientNet B5 demonstrated the highest accuracy. The ensemble approach further improved prediction performance by mitigating error variance. For practical implementation, we converted the model weights to ONNX format, which resulted in a 36% reduction in size and improved inference speed. These optimized models were then successfully incorporated into a computer-aided diagnosis (CAD) system, designed to assist dermatologists in categorizing skin lesions across nine distinct classes.

References

1. Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., & Esser, S. (2019). Skin cancer classification using convolutional neural networks: Systematic review. *Journal of the European Academy of Dermatology and Venereology*, *33*(3), 424-430. https://doi.org/10.1111/jdv.15218
2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056
3. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287-1289. https://doi.org/10.1126/science.aaw4399
4. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229-1234. https://doi.org/10.1038/s41591-020-0942-0
5. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the International Skin Imaging Collaboration (ISIC) 2017. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 501-512. https://doi.org/10.1109/JBHI.2018.2885977
6. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056
7. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., & Thomas, L. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836-1842. https://doi.org/10.1093/annonc/mdy166
8. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., & Kittler, H. (2019). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229-1234. https://doi.org/10.1038/s41591-020-0942-0
9. Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., & von Kalle, C. (2019). Deep learning outperforms dermatologists in the classification of cutaneous melanoma: A multi-center, multi-reader, blinded validation study. *European Journal of Cancer*, *113*, 47-54. https://doi.org/10.1016/j.ejca.2019.04.023
10. Codella, N. C., Nguyen, Q. B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., & Smith, J. R. (2018). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, *61*(4/5), 5-1. https://doi.org/10.1147/JRD.2017.2768203
11. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056
12. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74
13. Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., & Halpern, A. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, *20*(7), 938-947. https://doi.org/10.1016/S1470-2045(19)30333-X
14. Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., & von Kalle, C. (2019). Deep learning outperforms dermatologists in the classification of cutaneous melanoma: A multi-center, multi-reader, blinded validation study. *European Journal of Cancer*, *113*, 47-54. https://doi.org/10.1016/j.ejca.2019.04.023
15. Codella, N. C., Nguyen, Q. B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., & Smith, J. R. (2018). Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, *61*(4/5), 5-1. https://doi.org/10.1147/JRD.2017.2768203
16. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

17. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., & Thomas, L. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836-1842. https://doi.org/10.1093/annonc/mdy166
18. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74
19. Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H. ... & Halpern, A. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, *20*(7), 938-947. https://doi.org/10.1016/S1470-2045(19)30333-X