JOURNAL OF
EMERGING INVESTIGATORS

Article

Transfer Learning with Convolutional Neural Network-Based Models for Skin Cancer Classification

Ryan Lin[1], Sindhu Ghanta[2]
[1]Saratoga High School, Saratoga, CA
[2]AIClub, Mountain View, CA

SUMMARY

Skin cancer is a common and potentially deadly form of cancer. According to the Skin Cancer Foundation, 1 in 5 Americans will develop skin cancer by the age of 70. If caught early, it can often be treated with minimal surgical intervention and a high likelihood of full recovery. However, diagnosis using current methods requires a physician, is time-consuming, and is expensive. This study's purpose was to develop an automated approach for early detection for skin cancer. We hypothesized that convolutional neural network-based models using transfer learning could accurately differentiate between benign and malignant moles using natural images of human skin. To test this hypothesis, we developed a skin cancer detection system using four types of deep learning model architectures MobileNetV2, ResNet50V2, EfficientNetV2B0, and VGG16. We tested our models with a publicly available dataset from International Skin Imaging Collaboration. Through training, evaluation, and hyper-parameter tuning across four different models, the best-performing model, VGG16, achieved an Area Under the Curve (AUC) score of 0.95 and a test accuracy of 84.7%. We deployed the model as a publicly available service using Representational State Transfer Application Programming Interface (REST API). Skin images can be submitted to the API endpoint for skin cancer prediction. Our findings suggest that deep learning models can play a vital role in accessible and automated skin cancer screening, empowering healthcare professionals to make informed decisions and potentially improving outcomes for patients with limited access to healthcare resources.

INTRODUCTION

Skin cancer is one of the most common and potentially most deadly forms of cancer (1-3). According to the Skin Cancer Foundation, 1 in 5 Americans will develop skin cancer by the age of 70 (1). According to the estimates provided by the American Cancer Society for the year 2023, the incidence of new melanoma skin cancer cases is expected to be approximately 97,700, with higher numbers in men compared to women (4). Early detection of melanoma is crucial for effective treatment and improved prognosis (5). Therefore, identifying the warning signs of skin cancer, such as changes in size, shape, or color of moles or skin lesions, as well as the appearance of new growths on the skin, plays a vital role in timely intervention (6). If caught early, skin cancer can often be treated with minimal surgical intervention and a high likelihood of a full recovery (3, 7).

Current methods for detecting and classifying skin cancer, such as visual inspection by a dermatologist, can be time-consuming and costly (1, 2, 8). Access to healthcare professionals for skin cancer detection and treatment may be limited, particularly in developing countries (1). Also, the rate of correct diagnosis by expert dermatologists using images is estimated at 75-84% (9, 10). A lack of knowledge and awareness, misdiagnosis, fear, stigma, or inconvenience can all act as limitations that prevent a patient from seeking medical care (1).

Interest in deep neural networks in recent years has resurged due to several factors. These include the rise in availability of large, high quality, labeled datasets; advancements in parallel computing capabilities; and the development of accessible software platforms, such as PyTorch and Tensorflow, that facilitate GPU-based computations (11). As a result of these advances, deep learning techniques have found significant applications in the realm of computer-aided diagnosis in healthcare (12). These include analyses of medical images to determine whether they indicate the presence or absence of a pre-defined disease (13). Applications of this technique can be seen across various medical disciplines, including dermatology for skin disease identification, ophthalmology for recognizing conditions like diabetic retinopathy and glaucoma, and oncology for classifying pathological images for cancers, such as breast and brain cancer (12).

Dermatologists detect skin cancer beginning with an initial clinical screening, which might be followed up by a biopsy and histopathological examination. However, the time of a dermatologist is limited and expensive (13). In an effort to save time, cost, and effort, researchers have used machine learning algorithms, such as support vector machine or random forest, to detect melanoma from images of the skin to rule out healthy cases in an automated manner (14). Further advances in deep learning for classification of skin cancer images can assist the dermatologists by improving the diagnosis. Transfer learning techniques can reduce training time and improve performance and accuracy (15-17). Specifically in medical images and the domain of skin cancer, researchers have experimented with several different types of model architectures. Artificial neural networks (ANN), convolutional neural networks (CNN), and generative adversarial neural networks (GAN) have been successfully used to automate and improve the accuracy of skin cancer classification (18). We leverage the advances in CNN literature in the context of the International Skin Imaging Collaboration (ISIC) dataset, to find the best performing model and deploy it on the cloud so that it can be accessed by anyone with an internet connection.

Specifically, four pre-trained models architectures -MobileNetV2 (19), ResNet50V2 (20), EfficientNetV2B0 (21), and VGG16 (22) were used for developing and testing skin cancer prediction performance. Our experiments resulted in VGG16 having the best model performance. Evaluation and comparison of the performance of different models are done using metrics such as precision, recall, Receiver Operating Characteristic (ROC) curves, confusion matrix, and accuracy. Our model that used the VGG16 pre-trained architecture achieved a test accuracy of 84.7%, with a recall of 0.94 and an Area Under the Curve (AUC) of 0.95.

We have deployed the model in the AWS cloud as a REST API, so that it is available over the internet to anyone who wants to use it for making predictions on natural images of the human skin containing moles.

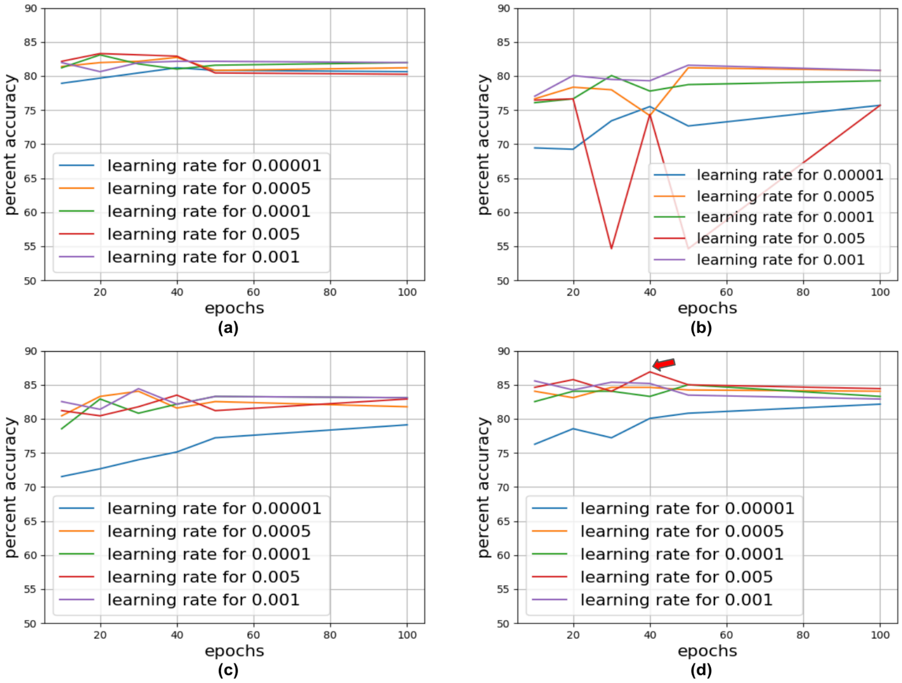
RESULTS

We aimed to achieve high accuracy in skin cancer prediction through the experiments. We hypothesized that CNN-based models using transfer learning could accurately differentiate between benign and malignant moles using natural images of human skin. To test this hypothesis, we employed transfer learning, which leverages the knowledge learned from pre-trained models to expedite the training process and improve overall performance.

We briefly describe the neural network architecture used in the experiments of this study. MobileNetV2 consists of 53 layers and has approximately 3.4 million parameters (19) and is an excellent choice for resource-constrained environments like mobile devices or edge devices. ResNet50V2 on the other hand has 50 layers and approximately 25.6 million parameters (20). It is a very popular model in the computer vision literature and has demonstrated excellent performance on several types of image classification applications. EfficientNetV2B0 has 153 layers and approximately 5.3 million parameters (21), which is designed to achieve high accuracy while being computationally efficient. VGG16 on the other hand has only 16 layers. However, it has approximately 138 million parameters (22), which is very high compared to the other models such as MobileNetV2, ResNet50V2 and EfficientNetV2B0. It is known for its effectiveness in image classification tasks. It is widely used as a baseline model in computer vision applications.

The dataset used for experiments across all four model architectures was kept fixed in order to have a fair comparison of their performance. The dataset was downloaded from a publicly available dataset of images of skin lesions from ISIC (23) and split into disjoint train, validation and test subsets. For each model, 2637 images were used for training and 660 images were used for validating the performance of the model. The images contained both malignant and benign samples. The hyper-parameters of each model varied between 0.00001 and 0.001 for learning rate and 10 and 100 for epochs. For MobileNetV2, the accuracy of the model on the validation dataset varied between 78.93 and 83.30 peaking at a learning rate of 0.005 after 20 epochs (**Figure 1**). ResNet50V2's accuracy spanned between 54.65 and 81.59 reaching the highest value at a learning rate of 0.0001 over 50 epochs (**Figure 1**). EfficientNetV2B0 saw a range of 71.54 and 84.44 in accuracy, with the best performance at a learning rate of 0.001 and 30 epochs (**Figure 1**). Lastly, in the case of VGG16, accuracy varied between 76.28 and 86.9 (**Figure 1**), with optimal results at a learning rate of 0.005 and 40 epochs.

The best model from each of the four pre-trained



(a)

(b)

(c)

(d)

Figure 1: Hyper-parameter tuning results for four model architectures. (a) MobileNetV2 (b) ResNet50V2 (c) EfficientNetV2B0 and (d) VGG16 on validation data. X-axis represents the hyper-parameter epochs. y-axis represents the performance of the model. Different learning rates are represented by varying colors along with their legends in the graph. The red arrow indicates the best validation accuracy (86.907%).

	MobileNetV2	ResNet50V2	EfficientNetV2B0	VGG16
Accuracy	0.8394	0.8136	0.8288	0.8470
Precision	0.7957	0.7580	0.8582	0.7741
Recall	0.8700	0.8667	0.7467	0.9367
F1-score	0.8312	0.8087	0.7986	0.8477
AUC	0.9148	0.9136	0.9204	0.9497

Table 1: Test results from the models chosen from each of the four pre-trained architectures. Best model was chosen based on validation accuracy and was used for making predictions on the test dataset to generate the results in the table. The last column is green as it corresponds to the model with highest accuracy. Precision refers to the ratio of true positive to all positive predictions, where positive in this case is the occurrence of cancer. Recall is the ratio of true positives to all actual positives. F1-core is the harmonic mean of precision and recall, balancing the trade-off between them. AUC is the area under the curve for the ROC and measures the ability of the model to distinguish between classes.

architectures was selected based on its validation accuracy and used to evaluate its performance on the test data (**Table 1**). The confusion matrix along with the ROC curves are reported for each model on the test data (**Figure 2, Figure 3**). The threshold values used for calculating the confusion matrix is equal to 0.5 since it is a binary classifier.

A VGG16 model predicted skin cancer with the highest test accuracy. The VGG16 model was trained with a learning rate of 0.005 and 40 epochs. (test accuracy: 84.7%, AUC: 0.9497) (**Table 1**). As can be seen from the confusion matrix, in addition to having the best performance VGG16 has a better precision in detecting the malignant images, which is a desired characteristic in medical applications. Error in detecting a malignant image is a worse outcome compared to error in detecting a benign image since a mis-prediction of a benign image can be corrected when a physician examines the image down the pipeline. However, if a malignant image is missed and hence not examined further by a physician, it can lead to a bad outcome for the patient.

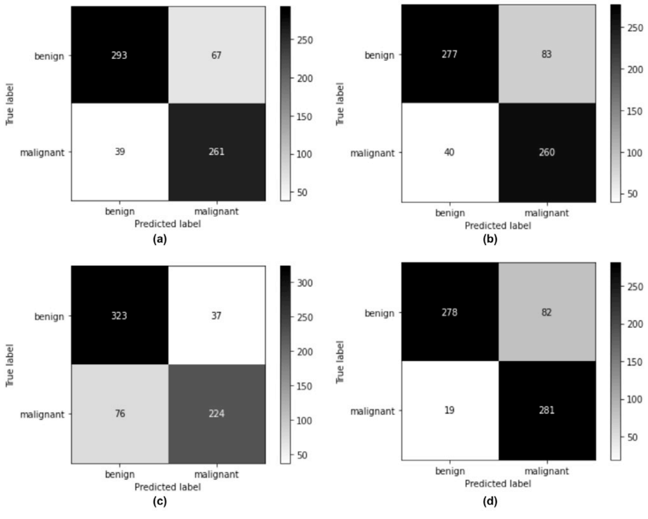


Figure 2. Confusion matrix for four model architectures. (a) MobileNetV2 (b) ResNet50V2 (c) EfficientNetV2B0 and (d) VGG16 on test data composed of 360 images of benign moles and 300 images of malignant moles. Confusion matrix was calculated for a threshold value of 0.5.

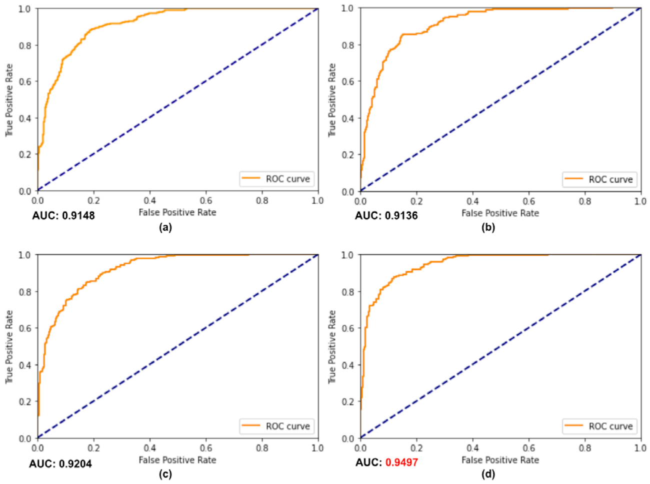


Figure 3. ROC curve for four model architectures. (a) MobileNetV2 (b) ResNet50V2 (c) EfficientNetV2B0 and (d) VGG16 on test data composed of 360 images of benign moles and 300 images of malignant moles. The AUC value for the VGG16 model is red as it is the highest (0.9497).

Low learning rates, like 0.00001, led to poor performance in all neural network architectures, likely due to inadequate parameter adjustments. Higher learning rates yielded better results across various epoch values. Increasing epochs on the other hand did not enhance performance, showing that the models reached the local optimum and iterating more over the data does not lead to any further improvement in model performance.

To make our model accessible to a broader audience, we integrated it into a publicly available service utilizing a Representational State Transfer Application Programming Interface (REST API) on Amazon Web Services. We stored the tensorflow model in the h5 format in the simple storage service. We used AWS lambda to store the code for inference and the API Gateway to create the REST API interface. This API-based service enables users to interact with the model programmatically and submit skin lesion images for prediction. One can submit the skin images to the API endpoint, and the model returns the corresponding prediction results. The url of the endpoint is: https://askai.aiclub.world/c2810e82-bfb5-4157-876f-1a5fa4805e83.

An image can be provided to this endpoint in its base64 encoded form. The response is a json with the key values status code, body and headers. Body of the response has a key called predicted_label, which contains the model prediction.

DISCUSSION

This study aimed to create an automated skin cancer detection system using convolutional neural networks and transfer learning. The hypothesis centered on the ability of these models to distinguish between benign and malignant skin lesions effectively. Previous attempts at automating the process of screening based on images of skin have utilized euclidean and fractal geometric perspective to train a fuzzy inference engine (24). Researchers have also looked into

https://doi.org/10.59720/23-204

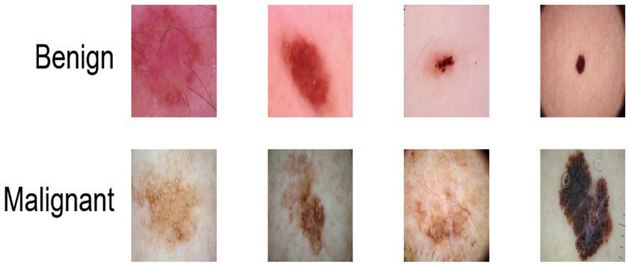


Figure 4. Sample images in the dataset. Top row: images of benign moles. Bottom row: images of malignant moles. Each picture is from a different individual.

image enhancement techniques that would assist in detecting skin cancer (25). However, these techniques use a different dataset and do not report any quantitative metrics of model evaluation on its ability to correctly identify cancerous lesions. Four deep learning architectures MobileNetV2 (19), ResNet50V2 (20), EfficientNetV2B0 (21), and VGG16 (22) were employed and tested against the ISIC dataset (23). The comprehensive process of training, evaluating and fine-tuning these modes led to the VGG16 emerging as the architecture with the best predictive performance, achieving an AUC score of 0.95 and a test accuracy of 84.7%.

The hyper-parameters settings of the VGG16 model that led to the best test performance were a learning rate of 0.005 and 40 epochs. In contrast, the diagnosis provided by a dermatologist has an accuracy of 75-84% (9, 10). Even though this model performs marginally better than human diagnosis, it should be used only as an assistive tool to a dermatologist to help speed up the process of analysis and diagnosis.

We observed that a low value of learning rate such as 0.00001 yielded the worst performance across all the neural network architectures. This suggests that such a low value of learning rate caused the models to make insufficient adjustments of the corresponding model's parameters, limiting its ability to learn from the data effectively and resulting in suboptimal performance. A higher learning rate performed much better for all values of epochs. We also observed that changing the epoch values to higher numbers did not improve the model's performance. This suggests that the model found the local minima and could not further enhance the algorithm's performance. A low value of epochs was enough for the models to achieve a local optimal solution.

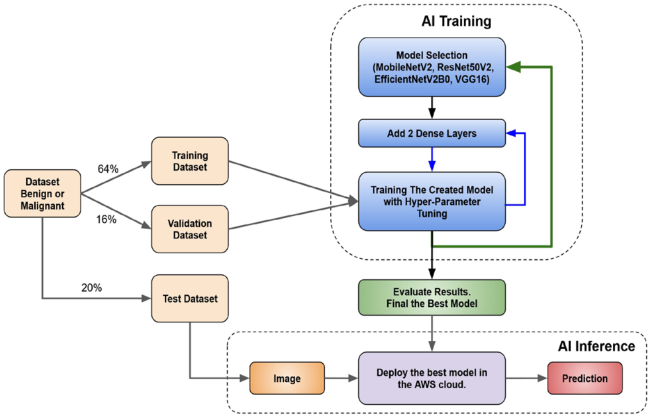

AI Training
Model Selection (MobileNetV2, ResNet50V2, EfficientNetV2B0, VGG16)
Add 2 Dense Layers
Training The Created Model with Hyper-Parameter Tuning
Dataset Benign or Malignant
64%
16%
20%
Training Dataset
Validation Dataset
Test Dataset
Evaluate Results. Final the Best Model
AI Inference
Image
Deploy the best model in the AWS cloud.
Prediction

Figure 5. Project flowchart. The experimental procedure included training, evaluating, comparing the performance of the models with hyper-parameter tuning, and model inference. The dataset was divided into train, validation and test sets. The training and validation set were used to train different models and evaluate the best model and hyper-parameter set. After the best model was chosen, it was evaluated on the test set to report the final results.

The performance of deep neural networks is often influenced by the size of the dataset used for training. In this study, a larger dataset with more examples of malignant skin lesion images could potentially enhance the model's accuracy and generalization. Additionally, the research was constrained to four specific models, but further investigation into alternative algorithms could yield valuable insights and potentially even better results. Expanding the dataset size and exploring additional algorithms are crucial factors to consider when interpreting and improving the outcomes of this study.

Another possibility for future research is inclusion of early-stage cancer images into the dataset that would enable early skin cancer detection. One can also deploy this model on mobile applications which would enable convenient screening and access to remote areas which may lack medical facilities. It could be used as a preliminary screening tool, rapidly sifting through large volumes of cases and flagging those of concern. In this context, the balance between precision and recall becomes extremely important. On one hand, a higher recall might lead to more false positives, which could cause undue stress and further medical tests for patients. On the other hand, higher precision could increase the risk of false negatives, potentially missing genuine cases. In medical diagnosis, perhaps leaning towards slightly higher recall is judicious, as it ensures fewer missed diagnosis. This would mean a few false alarms, which could be removed by subsequent medical review by the professional. However, a medical professional should always be involved to determine the final outcome.

MATERIALS AND METHODS

Dataset

The dataset was downloaded from a publicly available dataset of images of skin lesions from ISIC (23) (**Figure 4**). The dataset had a total of 3297 images (1800 images were benign and 1497 were malignant).

The data was split so that 80% of the dataset became the training set and 20% of the dataset became the test set. Out of the training dataset, 80% of the data was used for training the model and 20% was used for validation.

Algorithms

Four available pre-trained models (MobileNetV2, ResNet50V2, EfficientNetV2B0, and VGG16) were selected from the TensorFlow Keras library by looking at their sizes, parameters, and accuracy (19-22). We focused on evaluating a spectrum of model architectures that encompassed both compact and extensive network sizes. This included the lightweight MobileNetV2 model, which has a small footprint of just 14 MB, as well as significantly larger models, with the largest being 528 MB. Additionally, we selected two intermediate-sized models—EfficientNetV2B0 at 29 MB and ResNet50V2 at 98 MB—to provide a comprehensive size range analysis. The chosen models were all within an accuracy bracket of 71.3% to 78.7% as reported on the ImageNet challenge, ensuring a balanced comparison of

https://doi.org/10.59720/23-204

efficiency and performance for selection of pre-trained networks. This diversity of characteristics ensures a comprehensive evaluation of the models capabilities in the context of skin cancer detection, allowing us to cover a broad spectrum of deep learning architectures.

The architecture of each pre-trained model was customized by adding a single dense layer with 100 neurons. The goal of the training phase of the experiment is to come up with a model that has the optimal predictive performance. In order to accomplish this, several experiments were conducted with different hyper-parameter values.

To identify the optimal values of hyper-parameters, typically an iterative process of experimentation and analysis is conducted. The process involves training the neural network with different combinations of learning rates and epochs and evaluating the model's performance on a validation dataset.

During the training process, the hyper-parameter learning rate was varied between the values - 0.00001, 0.0005, 0.0001, 0.005, 0.001) and epochs were varied between the values -10, 20, 30, 40, 50, 100. The results of training were evaluated on the validation set. This process is repeated on several pre-trained models for all combinations of learning rates and epochs. Once the best model is selected, model performance on the test set is evaluated. (**Figure 5**)

Evaluation of the model was done by using metrics such as accuracy, precision and recall to measure the model's ability to correctly classify images of skin lesions. A ROC-AUC analysis was also performed on the test dataset. Accuracy is calculated by dividing the number of predictions that are correct by the total predictions. It tells us the percentage of the correct classifications that the model makes. Precision is calculated by dividing the true positives by the sum of true positives and false positives. The precision tells us, out of all the times the model predicted positive, the percent of positive predictions that were correct. Recall is calculated by dividing the true positives by the sum of true positives and false negatives. The recall tells us, out of all the data that is positive, the percent that the model predicted correctly as positive.

Software

A standard desktop or laptop computer with internet was used to access the google colab account, where tensor flow packages were used for conducting experiments. Keras pre-trained deep learning models were used to download the model architectures along with their imagenet weights. Standard python packages such as matplotlib and numpy were used to create the figures. AWS Lambda and AI Gateway were used to create a REST API deployment.

Received: August 9, 2023
Accepted: November 1, 2023
Published: June 28, 2024

REFERENCES

1. The Skin Cancer Foundation. "Skin Cancer Facts & Statistics." Skin Cancer Foundation, www.skincancer.org/skin-cancer-information/skin-cancer-facts/. Accessed 9 Jan. 2023.
2. Takiddin, Abdulrahman, et al. "Artificial Intelligence for Skin Cancer Detection: Scoping Review." *Journal of Medical Internet Research*, vol. 23, no. 11, 2021, doi:10.2196/22934.
3. Javaid, Arslan, et al. "Skin Cancer Classification Using Image Processing and Machine Learning." 2021 International Bhurban Conference on Applied *Sciences and Technologies (IBCAST)*, 2021, doi:10.1109/ibcast51254.2021.9393198.
4. "Melanoma Skin Cancer Statistics." *American Cancer Society*. www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html. Accessed 21 July 2023.
5. Brinker, Titus Josef, et al. "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review." *Journal of Medical Internet Research*, vol. 20, no. 10, 2018, doi:10.2196/11936.
6. Jaworek-Korjakowska, Joanna. "Computer-Aided Diagnosis of Micro-Malignant Melanoma Lesions Applying Support Vector Machines." *BioMed Research International*, vol. 2016, 2016, pp. 1–8, doi:10.1155/2016/4381972.
7. Jerant, A. F., et al. "Early Detection and Treatment of Skin Cancer." American Family Physician, vol. 62, no. 2, 2000, pp. 357-368, 375-376, 381-382.
8. Adegun, Adekanmi, and Serestina Viriri. "Deep Learning Techniques for Skin Lesion Analysis and Melanoma Cancer Detection: A Survey of State-of-the-Art." *Artificial Intelligence Review*, vol. 54, no. 2, 2020, pp. 811–841, doi:10.1007/s10462-020-09865-y.
9. Ozkan, Ilker Ali, and Murat Koklu. "Skin Lesion Classification Using Machine Learning Algorithms." *International Journal of Intelligent Systems and Applications in Engineering*, vol. 4, no. 5, 2017, pp. 285–289, doi:10.18201/ijisae.2017534420.
10. Jain, Shivangi, et al. "Computer Aided Melanoma Skin Cancer Detection Using Image Processing." *Procedia Computer Science*, vol. 48, 2015, pp. 735–740, doi:10.1016/j.procs.2015.04.209.
11. Sengupta, Saptarshi, et al. "A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends." *Knowledge-Based Systems*, vol. 194, 2020, p. 105596, doi:10.1016/j.knosys.2020.105596.
12. Liu, Xiaoqing, et al. "Advances in Deep Learning-Based Medical Image Analysis." *Health Data Science*, vol. 2021, 2021, doi:10.34133/2021/8786793.
13. Ameri, A. "A Deep Learning Approach to Skin Cancer Detection in Dermoscopy Images." *Journal of Biomedical Physics and Engineering*, vol. 10, no. 6, 2020, doi:10.31661/jbpe.v0i0.2004-1107.
14. Murugan, A., et al. "Diagnosis of Skin Cancer Using Machine Learning Techniques." *Microprocessors and Microsystems*, vol. 81, 2021, p. 103727, doi:10.1016/j.micpro.2020.103727.
15. Weiss, Karl, et al. "A Survey of Transfer Learning." *Journal of Big Data*, vol. 3, no. 1, 2016, doi:10.1186/s40537-016-0043-6.
16. Krishna, Sajja, et al. "Deep Learning and Transfer Learning Approaches for Image Classification." *International Journal of Recent Technology and Engineering (IJRTE)* 7.5S4 (2019): 427-432.
17. Kim, Hee E., et al. "Transfer Learning for Medical Image Classification: A Literature Review." *BMC Medical Imaging*, vol. 22, no. 1, 2022, doi:10.1186/s12880-022-00793-7.
18. Dildar, Mehwish, et al. "Skin Cancer Detection: A Review Using Deep Learning Techniques." *International Journal of Environmental Research and Public Health*, vol. 18, no.

https://doi.org/10.59720/23-204

10, 2021, p. 5479, doi:10.3390/ijerph18105479.

19. Sandler, Mark, et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, doi:10.1109/cvpr.2018.00474.
20. He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi:10.1109/cvpr.2016.90.
21. Tan, Mingxing, et al. “Efficientnetv2: Smaller Models and Faster Training.” International Conference on Machine Learning. PMLR, 2021.
22. Simonyan, Karen, et al. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv preprint arXiv:1409.1556 (2014).
23. 23.“International Skin Imaging Collaboration.” ISIC, www.isic-archive.com/#!/topWithHeader/wideContentTop/main. Accessed 30 July 2023.
24. Blackledge, Jonathan, et al. “Moletest: A Web-based Skin Cancer Screening System.” (2011).
25. Sao, Pratibha, et al. “The Diagnosis and Analysis on the Skin Cancer Detection Using Pattern Recognition Technology.” *ICONIC Research and Engineering Journals (IRE)*, vol. 2, no. 6, 2018, pp. 81-86.