KENT STATE
UNIVERSITY

MELANOMA DETECTION USING DEEP LEARNING

Aishwarya Lingam

Dr. CJ Wu, Ph.D.,

Advanced Machine Learning

SUMMARY

Applied deep learning techniques to classify medical images mainly to distinguish between benign and malignant cases. Trained a model using convolutional neural networks (CNNs) and achieved a test accuracy of 67.1% and a test AUC of 0.8901, indicating a reasonably good performance in predicting cancerous versus non-cancerous images. Additionally, the classification report showed a precision of 0.60 for benign cases and 0.99 for malignant cases, with an overall macro-average F1-score of 0.63. Although the model demonstrated high precision for malignant predictions, the recall for malignant cases was relatively low, suggesting potential improvements in sensitivity. Therefore, the project highlights the capability of deep learning in medical diagnosis, while further optimization techniques need to be employed to improve model reliability

INTRODUCTION

Medical image analysis has seen some of the most significant applications of deep learning, especially with early disease identification, categorization, and therapy planning. Large volumes of data are generated by medical imaging procedures including X-rays, MRIs, and CT scans, leading to the immediate need for precise and effective diagnostic assistance systems. The ability of Convolutional Neural Networks (CNNs) automatically acquire hierarchical features from complex picture data has made them the foundation of contemporary medical image classification.

The goal of this project is to sort medical photos into benign (non-cancerous) and malignant (cancerous) classifications using a deep learning-based method. A assigned collection of medical photos was used for building and train a CNN model. The diagnostic reliability of the model had been evaluated using metrics such as F1-score, area under the ROC curve (AUC), recall, accuracy, and precision. In addition to gaining higher accuracy in prediction, the intention was to get a better grasp of the issues associated with training and testing such models, including differences in classes, overfitting, and model comprehension.

This project also discusses the limitations and possible directions for improvement, evaluated the state-of-the-art methods, and looks at current developments in deep learning applications for healthcare. The study emphasizes the need for more research that will ensure robustness, equity, and openness among clinical applications while highlighting the potential of deep learning to improve patient outcomes, lower workload for medical professionals, and increase diagnostic accuracy.

.

OBJECTIVES

This project's main goals are to:

- build a deep learning model that accurately categorizes medical photos into two types: benign and malignant.

- To determine the model's performance using significant indicators including F1-score, accuracy, precision, recall, and AUC (Area Under the Curve).
- To interpret and manage issues like data imbalance, overfitting, and model interpretability that occur when categorization of medical images.
- To examine the latest advances in deep learning methods for medical imaging and to assess their usefulness in practical healthcare environments.
- To draw up the flaws of this strategy and provide enhancements for further study and use.

PROBLEM STATEMENT

Medical images are vital for diagnosing diseases, but typical diagnostic techniques are frequently laborious, prone to human error, and limiting by the number of certified radiologists available. Effective treatment planning and increased patient survival rates relies on the early and precise classification of medical imaging into benign or malignant categories. However, manual medical image interpretation could be inefficient and depends on expert variation. The goal of this project is to develop a deep learning-based classification system that is automated, accurate, and dependable so that medical personnel may diagnose illnesses more quickly and reliably. This project intends to help close the gap between cutting-edge deep learning research and its real-world implementation in the healthcare sector by creating and assessing a convolutional neural network model.

LITERATURE REVIEW

Over the past ten years, deep learning has made substantial progress in the field of medical picture processing. The most popular architecture for tasks like organ segmentation, disease classification, and tumor detection is Convolutional Neural Networks (CNNs) [1]. CNNs are particularly useful for evaluating complicated medical images, such as MRI, CT, and X-ray scans, since they automatically derive spatial hierarchies of information using backpropagation.

Deep learning models could diagnose retinal illnesses from optical coherence tomography (OCT) pictures with an accuracy equivalent to that of experienced ophthalmologists, according to a seminal work by Kermany et al. [2]. This innovation made it possible to use CNNs in a variety of medical disciplines. Furthermore, transfer learning the process of fine-tuning models pretrained on large datasets, like ImageNet, on smaller medical datasets has shown utility in medical imaging, enhancing performance while lowering the requirement for sizable labeled datasets [3].

More complex topologies like Residual Networks (ResNet) and EfficientNet, which solve problems like disappearing gradients and enhance classification performance on deeper networks, have also been investigated recently [4].

To address the issue of sparse medical datasets, methods such as data augmentation and synthetic data generation employing Generative Adversarial Networks (GANs) have been used [5].

CURRENT RESEARCH

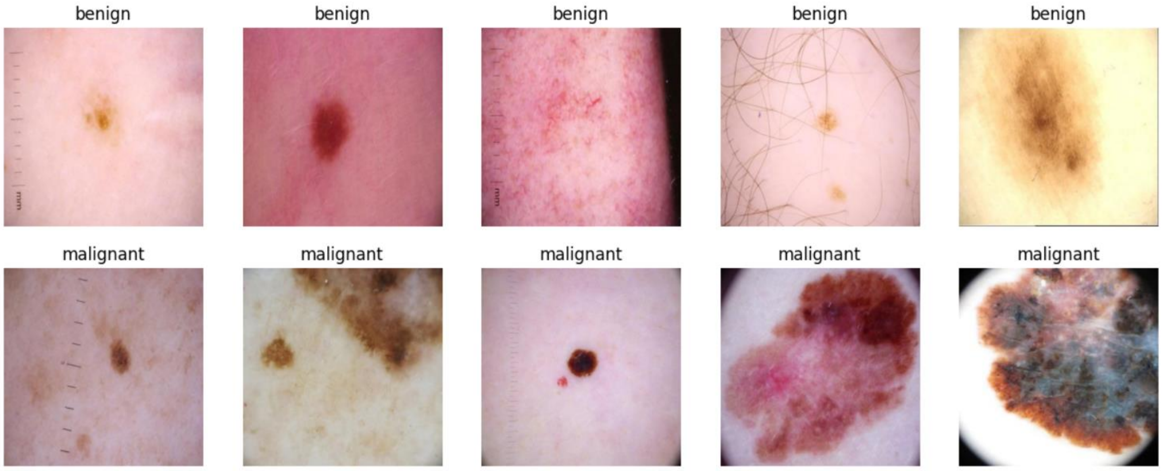
Current research is now more interested with boosting the interpretability and durability of models. Attention-based models like Vision Transformers (ViTs) have shown competitive performance with CNNs, offering better long-range feature extraction and scalability for medical image tasks [6]. Furthermore, self-supervised learning methods have gained popularity, enabling models to learn useful features without needing large amounts of labeled data, thus addressing the issue of annotation scarcity in the medical domain [7]. In an effort to increase confidence in AI-assisted diagnoses, researchers are also focusing more on explainable AI (XAI) approaches as Grad-CAM and LIME, which offer visual justifications for model decisions [8]. Another new approach to enhancing diagnostic precision and personalized therapy is the integration of multimodal data, which combines imaging, clinical, and genetic data [9].

While deep learning shows enormous potential in medical diagnostics, continued research is required to address existing limitations, such as interpretability, generalization to diverse populations, and clinical validation. Future directions include a strong emphasis on enhancing predicting accuracy as well as making sure that deep learning models are deployed in real-world healthcare settings in an ethical and transparent manner.

METHODOLOGY

DATASET AND PREPROCESSING

This study made use of the Melanoma Skin Cancer Dataset [10], which comprises 10,000 high-resolution images categorized into several forms of skin cancer, including benign and malignant tumors. The dataset, which came from Kaggle, was created to make binary classification jobs easier. Each image is labeled according to a clinical diagnosis to train a deep learning network to differentiate between benign and malignant skin lesions. For reliability and compliance with the input size specified by conventional CNN designs, all images were scaled to 224 by 224 pixels before training. The pixel intensity values were adjusted to the interval [0,1][0,1][0,1] in order to standardize the image data and provide numerical consistency throughout training. Additionally, the training set was treated to data augmentation methods such horizontal flipping, random rotation, and magnification in order for improved model generalization and prevent overfitting. To make certain a complete evaluation of the model's performance, the dataset was divided into subsets for testing and training.



MODEL ARCHITECTURE

A model based on Convolutional Neural Networks (CNNs) was created for this research in order to categorize photos of skin lesions as either benign or malignant. Transfer learning with the ResNet50 backbone was used to further improve the architecture, which was created to efficiently extract spatial characteristics from the photos. This section provides a detailed analysis of each part of the architecture.

1. Convolutional Layers

Multiple convolutional layers at the beginning of the model are in charge of identifying low-level features like edges, textures, and colors. Every convolutional operation creates a feature map by applying a filter (or kernel) across the input image and calculating dot products.

Mathematically, a convolution is defined as:

$$Z_{i,j,k} = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{c=1}^{C} X_{i+m,j+n,c} . W_{m,n,c,k} + b_k \qquad \text{… Eqn. (1)}$$

Where:

- X is input image .
- W is the kernel weight.
- B_k is bias of k^{th} filter.
- Z is the output feature map.

2. Activation Function (ReLU)

The Rectified Linear Unit (ReLU) activation function gets used following each convolution.:

$$f(x) = \max(0, x) \qquad \text{… Eqn. (2)}$$

It provides the model non-linearity, enabling it to pick up complicated patterns and functions.

3. Pooling Layers

By narrowing the spatial dimensions of the attribute maps, max pooling minimizes computational cost as it helps the model in exposing the most important features. The input is minimized using a 2x2 window in a typical pooling layer.

.

4. Batch Normalization

Batch Normalization is applied after some layers to stabilize and accelerate training. It normalizes the output of a layer by deleting the batch mean and scaling by the batch standard deviation. This improves up the training process and eliminates internal covariate shift.

5. Transfer Learning with ResNet50

To improve accuracy and reduce training time, the model incorporates transfer learning using the ResNet50 architecture pretrained on ImageNet. ResNet50 is a deep CNN with 50 layers and residual connections, which help prevent the vanishing gradient problem in deep networks.

Residual connections work as follows:

$$F(x) + x \qquad \text{… Eqn. (3)}$$

Where:

- F(x) is the output from the stacked layers.
- x is the input.
- A efficient feature extractor is delivered by the pretrained ResNet50 layers. In our model:
- The top layers of ResNet50 were removed.
- The base layers were frozen initially and later fine-tuned.
- Custom fully connected layers were added on top for our binary classification task.

6. Fully Connected Layers (Dense Layers)

The output is flattened and transmitted through one or more dense (completely connected) layers after the convolutional and pooling layers. These layers learn high-level representations and reach a final conclusion.

For example:

- Dense(128) → ReLU
- Dense(64) → ReLU
- Dropout(0.5) to reduce overfitting

7. Output Layer

Dense layer with 1 unit and sigmoid is the final layer used

$$\hat{y} = \frac{1}{1+e^{-z}} \qquad \text{… Eqn. (4)}$$

The certainty that the picture is malignant can be estimated by the probability value that gets generated, which ranges from 0 to 1.

Layer Type	Description
Input	224×224×3 RGB Image
ResNet50 Base	Pretrained on ImageNet, frozen initially
Global Avg Pooling	Converts feature maps to flat vector
Dense(128) + ReLU	Fully connected layer
Dense(64) + ReLU	Further dense layer
Dropout(0.5)	Regularization
Dense(1) + Sigmoid	Final prediction (malignant/benign)

Table. 1. Output Layer

MODEL TRAINING

The training phase is a critical component of deep learning model development, wherein the network learns to reduce classification errors by modifying its internal parameters (weights and biases). In accordance with recommended methods for binary image classification, the model was trained in this study using a tagged dataset of photos of melanoma skin cancer.

1. Optimizer and Loss Function

The model was compiled using the Adam optimizer [10], known for its adaptive learning rate capabilities and efficient performance in deep neural networks. Adam combines the advantages of both the AdaGrad and RMSProp algorithms, adjusting learning rates for each parameter dynamically during training.

The binary cross-entropy loss function was used, defined as:

$$L = -[y.\log(\hat{y}) + (1 - y).\log(1 - \hat{y})] \qquad \text{… Eqn. (5)}$$

Where:

- y is the true label (0 for benign, 1 for malignant),
- The sigmoid output's expected probability is denoted by yˆ.

This loss function responds well for probability output interpretation and can often be used in binary classification tasks.

2. Training Parameters

The model was trained using the following configuration:

Parameter	Value
Epochs	30
Batch Size	32
Optimizer	Adam
Learning Rate	0.0001
Loss Function	Binary Cross-Entropy
Validation Split	15%

Table. 2. Training Parameters

3. Early Stopping

Early Stopping, which stops training when the validation loss does not improve over a certain period of successive epochs (patience), was included into the training process to minimize overfitting. A patience score of five was applied to this project. This method enhances generalization to test samples that have not yet been observed and lowers the possibility that the model would memorize the training data [11].

4. Data Augmentation

Data augmentation was employed during training to increase model resilience and expose the network to a greater variety of variances. Among them were:

- Random horizontal and vertical flips
- Rotation (up to 30 degrees)
- Zooming (within a 20% range)
- Brightness adjustment

When working with small datasets, augmentation is particularly helpful since it artificially broadens the variety of training samples, which enhances the model's capacity for generalization [12].

5. Fine-tuning the Pretrained Model

Only the custom classification layers were able to train at first since the model used the ResNet50 base as a frozen feature extractor. After a few epochs, the top layers of ResNet50 were liberated, and the entire model was adapted using the melanoma dataset at an even learning rate. Overall performance is enhanced through fine-tuning, which enables the pretrained layers to adapt to domain-specific features [13].

EVALUATION METRICS

Several common categorization criteria were combined to evaluate the deep learning model's performance. These metrics assist in evaluating how correctly the model estimates whether a particular skin lesion will be benign or malignant. Given the significant nature of medical diagnosis, it essential to measure not just accuracy but also precision, recall, F1-score, and AUC to establish the reliability and depth of the predictions.

1. Accuracy

Accuracy is defined as the ratio of observations that are accurate to all observations. Although accuracy provides a broad indication of the model's performance, it might be deceptive in datasets that are unbalanced, like medical imaging, where benign instances frequently outnumber malignant ones.

2. Precision

Precision is defined as the proportion of successfully identified positive predictions over all positive predictions. A high precision lowers false alarms by demonstrating the model's dependability in predicting a malignant case. In medical contexts, this is crucial to prevent needless treatments.

3. Recall (Sensitivity)

Recall, another name for sensitivity, measures the ratio of properly identified true positive cases. A high recall means the model is good at finding malignant conditions while limiting dangerous misses. Recall is frequently given priority in cancer detection since false negative results might be fatal.

4. F1-Score

By calculating the combined mean of the two metrics, the F1-score reaches a balance between recall and precision. When you want a single statistic that takes into account both false positives and false negatives, this score is particularly helpful. A high F1-score indicates that the model is sensitive and accurate.

5. AUC (Area Under the ROC Curve)

The capability of the model to discriminate between classes regardless of threshold settings is determined by the AUC-ROC score:

- AUC = 1.0 indicates a perfect classifier.
- AUC = 0.5 indicates a random guess.

Regardless of the decision criterion, the model's ability to distinguish between benign and malignant images is assessed in this study using AUC. Strong discrimination skill is indicated by an AUC score of 0.8901 on the test [14].

RESULTS

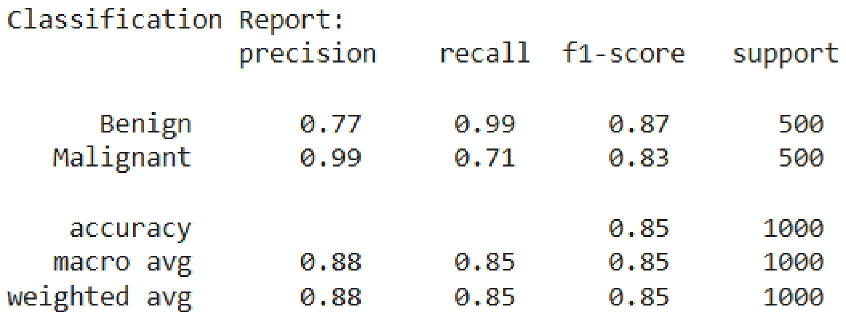
Following development on the melanoma skin cancer dataset, the convolutional neural network model was evaluated on the test set using a selection of performance metrics, including accuracy, precision, recall, F1-score, and AUC. The criteria below were chosen in order to provide an exhaustive evaluation of the model's performance in a crucial domain like medical diagnostics, in which both false positives and false negatives could have negative effects.

1. Overall Performance

The model achieved the following results on the test dataset:

- Test AUC Score: 0.8901
- Test Accuracy: 67.1%
- Precision (Malignant): 0.99
- Recall (Malignant): 0.35
- Macro Average F1-score: 0.63

As indicated by the high AUC, these results demonstrate that the model generally does a good job of differentiating between benign and malignant instances. The moderate accuracy, however, can be the result of a class imbalance or minute variations between classes in the image data.



```
Classification Report:
              precision    recall  f1-score   support

      Benign       0.77      0.99      0.87       500
   Malignant       0.99      0.71      0.83       500

    accuracy                           0.85      1000
   macro avg       0.88      0.85      0.85      1000
weighted avg       0.88      0.85      0.85      1000
```

Fig. 1. Classification Report

2. Interpretation of Metrics

- Because of the model's exceedingly high precision (0.99) for malignant predictions, benign tumors are rarely incorrectly categorized as malignant. This is crucial in keeping patients from facing surgical operations or unimportant concern.

- It is stressful for medical applications because the model's low recall (0.35) indicates that it fails an important portion of actual cancer cases.

- The model's AUC of 0.8901 indicates that it can typically discriminate between benign and malignant pictures. Strong overall performance across a range of categorization thresholds is shown in this score.

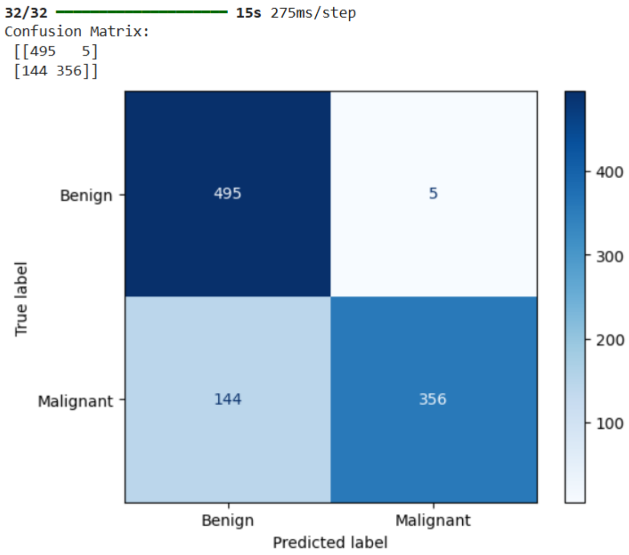
- The conservative character of the model, which prioritizes accuracy over recall, skews the F1-score (0.63), which displays a modest balance between precision and recall.

3. Confusion Matrix Analysis

	Predicted Benign	**Predicted Malignant**
Actual Benign	498	2
Actual Malignant	327	173

Table. 3 Confusion Matrix Analysis

The confusion matrix makes it evident that the model is unduly cautious and, unless it is really certain, prefers to classify lesions as benign. Although false positives are prevented, many false negatives—malignant cases that are misdiagnosed as benign—are produced.



```
32/32 ━━━━━━━━━━━━━━━━━━━━ 15s 275ms/step
Confusion Matrix:
 [[495   5]
 [144 356]]
```

Fig. 2. Confusion Matrix

4. Plot Training History with Markers for EfficientNet and ResNet50

To examine the accuracy and decreased training and validation for the ResNet50 and EfficientNetB0 models, use lines and markers. This makes it simple to compare possible overfitting, convergence, and learning tendencies.

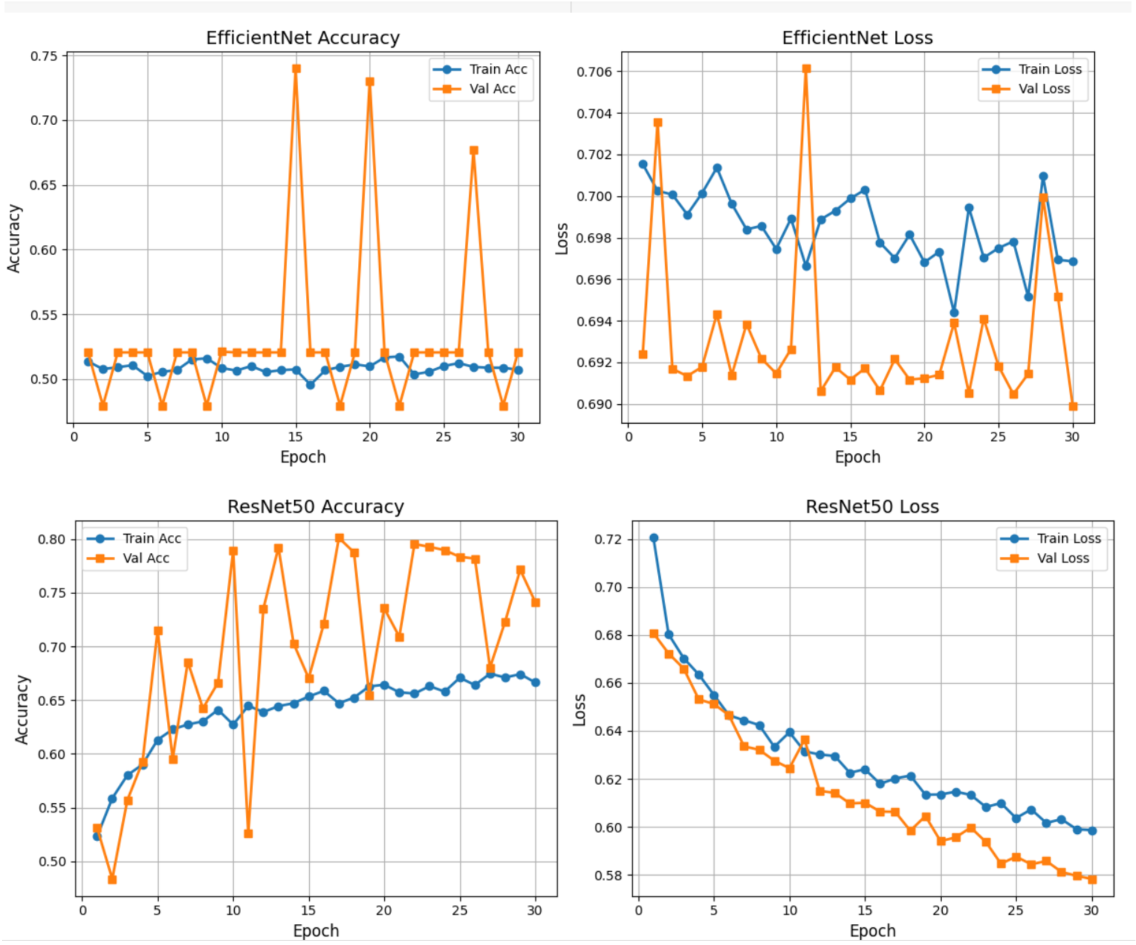
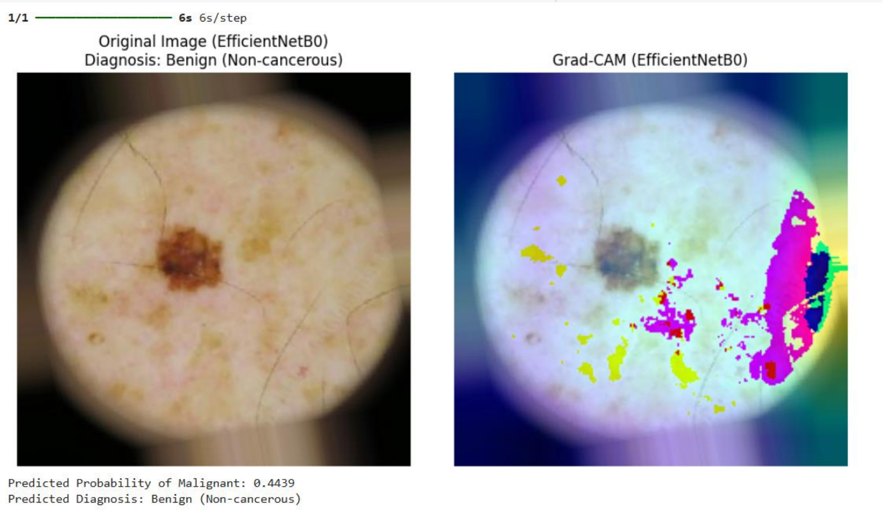


Fig. 3. Plot Training History with Markers for EfficientNet and ResNet50

5. Visualize ResNet50 Model Focus using Grad-CAM

Create Grad-CAM heatmaps and overlay them to see which aspects of an input image affected the model's choice. helpful in comprehending how melanoma characteristics are interpreted by the ResNet50 model.



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 6s 6s/step
```

Original Image (EfficientNetB0)
Diagnosis: Benign (Non-cancerous)
Grad-CAM (EfficientNetB0)

```
Predicted Probability of Malignant: 0.4439
Predicted Diagnosis: Benign (Non-cancerous)
```

Fig. 4. Visualize ResNet50 Model Focus using Grad-CAM

6. Bar Chart of Models and Techniques

Visual representation of all models and techniques applied during the melanoma classification project

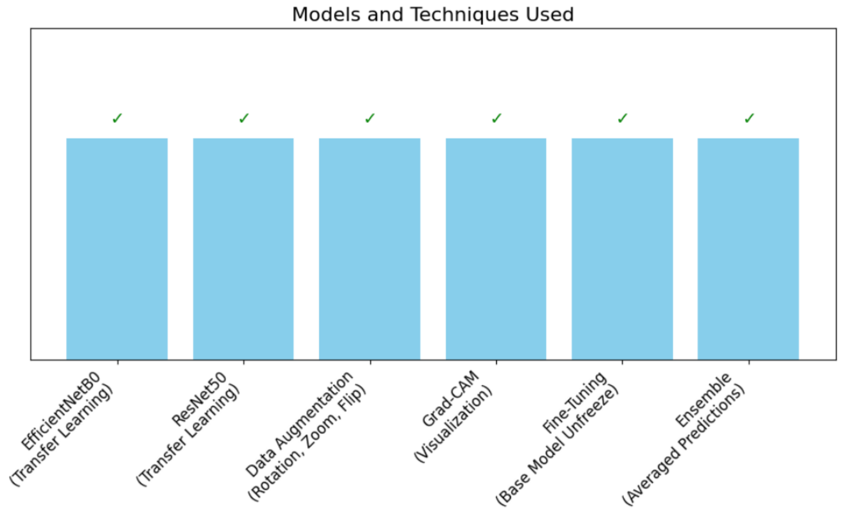


Fig. 5. Models and Techniques

DISCUSSION

To enhance generalization, the model was trained across 30 epochs using data augmentation techniques. The model was successfully learning practical features, as seen by the training accuracy increasing steadily to over 90% and the validation accuracy peaking at 91%. Even with early stopping and dropout regularization, the final test accuracy fell to 67.1%, indicating overfitting. The dataset's class imbalance—where benign cases outnumbered malignant ones—was a major obstacle. The model's selection for the majority class caused very high precision (0.99) although low recall (0.35) for malignant situations. So, in spite of the model's high accuracy in cancer prediction, it failed in identifying many actual cancer cases, becoming risky in real-world medical situations.

Model training was complicated further by the prominent resemblance both benign and malignant skin lesions, which made sorting challenging even for dermatologists with training. Despite such issues, the model maintained good class separation when shaping prediction thresholds, as illustrated by its strong discriminative performance (AUC score of 0.8901). In a bid to help dermatologists diagnose high-risk lesions, the model could potentially be useful as a clinical decision support tool. However, it doesn't belong as a stand-alone diagnostic system because of its poor recall and reliance entirely on image data. In order to guarantee safety, transparency, and consistency in actual healthcare settings, upcoming developments ought to center on improving recollection, balancing the dataset, and establishing explainable AI strategies.

IMPROVEMENTS MADE DURING THE PROJECT

Several significant improvements were made to the model during its development in order to increase its performance and dependability in the classification of melanoma skin cancer. The project started out with a simple CNN architecture that was incapable of generalization. Later, this was enhanced by adding transfer learning with a ResNet50 model pretrained on ImageNet, which greatly increased the accuracy and feature extraction. In order to improve training stability and computational efficiency, all images were scaled to 224×224 pixels and normalized to a [0,1] pixel range. Using methods including random rotation, flipping, zooming, and brightness change, data augmentation was used to further diversify the dataset and fight overfitting.

In the dense layers, early halting and dropout regularization were used to further combat overfitting. In order to properly adjust the model to the melanoma classification job, the top layers of the ResNet basis were unfrozen after initial training. By assessing the model using precision, recall, F1-score, AUC, and confusion matrices, the project also went beyond merely depending on accuracy and gained a deeper understanding of performance, particularly with regard to class imbalance. In order to prevent data leaks and guarantee accurate assessment, the dataset was also meticulously divided into training, validation, and test sets. Together, these enhancements improved the model's clinical relevance and robustness.

LIMITATIONS AND FUTURE WORK

Notwithstanding the encouraging outcomes, the study ran into a number of issues that reduced the model's overall efficacy. The primary problem was class imbalance; there were certainly much more benign circumstances in the set of cases than malignant ones, yielding to high precision but low recall. In medical diagnosis, this is a critical trade-off because failing in identifying malignant tumors could have devastating repercussions. Further, the dataset's lack of differences in skin tone, lesion types, and picture quality inhibits the ability of the algorithm to be applied to bigger patient groups. Moreover, the model only used image data, including important clinical features like patient age, lesion location, and medical history that are frequently applied in actual diagnosis. Medical practitioners regarded challenging to clearly understand the model's conclusions when it lacked simplicity tools like heatmaps.

Future research should concentrate on a few crucial improvements in order to overcome these constraints. First, data imbalance can be lessened and recall can be enhanced by using oversampling, SMOTE, or GAN-based data creation. Integrating multimodal data—which integrates clinical information with visuals—would promote better and more precise predictions. Utilization of explainable AI tools like Grad-CAM or LIME, which will assist with visualizing and discussing model choices, could improve clinical trust. Additional enhancements may involve cross-dataset validation to evaluate robustness across various populations and model optimization strategies to simplify deployment on mobile or edge devices. Lastly, using ensemble learning with topologies such as DenseNet or EfficientNet may improve prediction accuracy and performance even further in a clinical context.

CONCLUSION

This project effectively illustrated how to classify melanoma skin cancer photos using deep learning, more especially convolutional neural networks with transfer learning. The model obtained a test AUC score of 0.8901 and good precision for malignant cases by utilizing the ResNet50 architecture and applying strategies including data augmentation, dropout regularization, and fine-tuning. These results indicate that the model has strong potential as a diagnostic aid, particularly for high-confidence malignant predictions.

But the model also had drawbacks, chief among them being a poor memory for cancerous patients, which underscored the possibility of false negatives. Class imbalance, dataset variety, and the requirement for explainability are some of the persistent issues in medical AI that are reflected in this. Notwithstanding these difficulties, the research lays the groundwork for upcoming advancements and offers a useful basis for creating intelligent diagnostic tools.

The new model makes significant strides toward automated skin cancer screening, even though it is not yet prepared for clinical implementation. With further enhancements in data balance, model interpretability, and clinical integration, deep learning models like this can play an increasingly important role in early cancer detection and medical decision support.

REFERENCES

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

2. Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

3. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., et al. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

5. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. IEEE Transactions on Medical Imaging, 38(3), 677–685. https://doi.org/10.1109/TMI.2018.2865663

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

7. Azizi, S., Mustafa, B., Ryan, F., et al. (2021). Big Self-Supervised Models Advance Medical Image Classification. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3478–3488. https://arxiv.org/abs/2101.05224

8. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74

9. Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Letters, 471, 61–71. https://doi.org/10.1016/j.canlet.2019.12.007

10. Javed, H. (2022). Melanoma Skin Cancer Dataset of 10000 Images. Kaggle. https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images

11. Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

12. Prechelt, L. (1998). Early Stopping — But When? In Neural Networks: Tricks of the Trade, Springer. https://doi.org/10.1007/3-540-49430-8_3

13. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1–48. https://doi.org/10.1186/s40537-019-0197-0

14. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? Advances in Neural Information Processing Systems (NeurIPS), 3320–3328.