SKIN CANCER CLASSIFICATION USING CNN AND ANN MODELS: LEVERAGING DEEP LEARNING AND METADATA FOR EARLY DETECTION

Keshav Jindal
Student# 1006277903
keshav.jindal@mail.utoronto.ca

Abhishek Arora
Student# 0997398427
abhishek.arora@mail.utoronto.ca

Youssef Fikry
Student# 1006682626
youssef.fikry@mail.utoronto.ca

Pasut Aranchaiya
Student# 1011835935
pasut.aranchaiya@mail.utoronto.ca

ABSTRACT

Skin cancer is the most common form of cancer globally, with over 5.4 million cases diagnosed annually in the United States. Melanoma, the deadliest form of skin cancer, accounted for 287,723 new cases and 60,712 deaths worldwide in 2018. Early detection significantly increases survival rates, with a 5-year survival rate of over 98.4%. This project aims to develop a deep learning model, leveraging Convolutional Neural Networks (CNNs), to predict skin cancer from dermoscopic images combined with patient metadata such as age and medical history. Prior research shows CNNs outperform classical models, achieving 89% accuracy compared to the 86% accuracy of Decision Trees and other traditional approaches. Our model architecture integrates CNNs with metadata to improve diagnostic accuracy. Ethical concerns, such as bias in datasets and data privacy, are also addressed to ensure fairness and transparency in predictions. This project has the potential to assist dermatologists in early and more accurate skin cancer diagnosis, ultimately reducing global mortality rates. —-Total Pages: 8

1 INTRODUCTION

Skin cancer is the most common form of cancer globally, with over 5.4 million cases [1] diagnosed annually in the United States alone. Worldwide, melanoma, the deadliest form of skin cancer, accounted for approximately 287723 new cancer cases and 60712 deaths in 2018. [2] Early detection of skin cancer significantly increases survival rates— with a 5-year survival chance of over 98.4% [3]. However, visual diagnosis by dermatologists can be subjective and may vary based on experience, leading to missed or incorrect diagnoses. A 2020 study found that CNN deep learning models outperformed 86.6% of dermatologists[4] Given the importance of early detection, an automated and objective tool that can assist in identifying skin cancer is not only valuable but also potentially lifesaving.

The goal of this project is to develop a deep learning model that predicts skin cancer from dermoscopic images combined with patient metadata, such as age, gender, geography and medical history. Convolutional Neural Networks (CNNs), have been demonstrated to be highly effective in such image classification tasks. In 2010, University of Toronto researchers developed a deep convolutional network to classify 1.2 million images in the ImageNet challenge. They achieved a breakthrough performance, getting a top 1 and top 5 error rate of 37.5% and 17% [5], which is considerably better than previous state of the art systems at that time. Our model will incorporate CNN's for our skin cancer classifier. We will also consider factors like patient age, medical history etc which are also known to influence cancer risk. A detailed description of our project architecture is outlined below.

This project is interesting because it addresses a real-world healthcare challenge where diagnostic limitations of clinicians and dermatologists can be supplemented through technology. By leveraging deep learning, we aim to build a scalable solution that can aid dermatologists in accurately diagnosing skin cancer in early stages, save resources through automation, and reduce mortality rates on a global scale.

2 Illustration

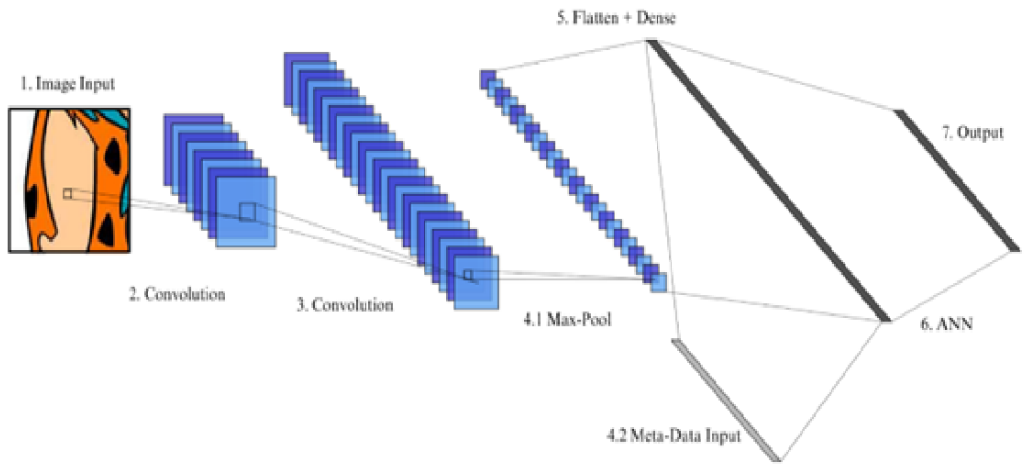

Figure 1: A detailed model showing the CNN and ANN layer architecture

3 Background & Related Work

3.1 Comparing Classical Models to Neural Networks

Skin cancer classification has advanced significantly, with researchers exploring a range of modeling approaches to improve prediction accuracy. In a study by Khan et al. (2021), classical machine learning models were compared to Convolutional Neural Networks (CNNs) using the ISIC 2018 dataset. [6] The results showed that Decision Trees achieved an accuracy of 86% with a recall of 0.70, K-Nearest Neighbors (KNN) reached 81% accuracy with a recall of 0.68, and Logistic Regression obtained an accuracy of 88% and a recall of 0.70. In contrast, the CNN outperformed all classical models, achieving an accuracy of 89% with a recall of 0.70. These findings highlight the superiority of deep learning models like CNNs for complex image classification tasks, making them the more suitable approach in this field.

3.2 Comparing Neural Networks

After confirming that deep learning outperforms classical models, researchers shifted their focus to determining which neural network architecture is most effective for skin cancer classification. Dildar et al. (2021) conducted a study comparing several neural networks, including Artificial Neural Networks (ANNs), CNNs, Kohonen Self-Organizing Neural Network (KNN), and Generative Adversarial Networks (GANs). [8]

Their findings revealed that while an ANN with backpropagation achieved an accuracy of 86.66% on a small dataset of 90 dermoscopic images, the CNN trained on the larger ISIC dataset outperformed the others with an accuracy of 98%. KNN achieved 93.15% accuracy, while GANs performed the least well, with 86.1% accuracy. As datasets and types of classification are different across these results, the brief conclusion is that CNN which achieved the highest accuracy performs better than other neural network models

3.3 Improving CNN Architectures

With CNNs emerging as the top choice for skin cancer classification, researchers have continued to experiment with various CNN architectures to boost performance and reduce biases stemming from differences in dermatologists' expertise. Wu et al. (2022) compared several CNN variants, evaluating their advantages and limitations. [9] For example, Inception-v2, a modified CNN, was designed to enhance sensitivity through sonification, but its performance could be inconsistent depending on the variability in pathologists' diagnoses.

Beyond architectural improvements, Wu et al. (2022) also highlighted the need to address challenges in the datasets themselves. Skin cancer datasets often suffer from limitations such as a small number of patients, insufficient diversity in patient demographics, and noise introduced during image capture. To counteract these issues, various data processing techniques have been explored to improve the robustness of models. These insights suggest that our project must carefully select the model architecture and apply appropriate data preprocessing techniques to achieve the best possible results.

3.4 Ethnic Bias

As skin cancer classification methods evolved, A thorough consideration on each option is needed to find the most optimal solution. Mahbod et al. (2020) proposed an ensemble approach that combined several CNN architectures, including ResNet-50 and DenseNet-121. [7] By leveraging the strengths of multiple networks, they were able to enhance classification accuracy. Their method achieved impressive results on public datasets like ISIC 2018, showing how ensemble learning can address the complexity of medical imaging and improve performance across diverse data.

3.5 An Optimized Approach

Another notable approach that utilizes the potential of choosing the correct model and method comes from the winning team of the ISIC 2019 competition. To start, they included additional data from outside the competition, resulting in more variation in the datasets. The datasets can be further improved by adding color constancy method for normalization (Yap et al., 2018), as well as image augmentations like random flipping and rotation. Seeing the result, Gessert et al. (2019) stated that this strategy led to improvement in performance [21]. Their method included fully connected neural networks for metadata and ensemble of EfficientNets for dermoscopy images, which resulted in substantial performance gains. The result of their project suggested that EfficientNets are capable of handling large input sizes, and ensembling these models yielded the best performance, achieving an accuracy of 0.925 (Gessert et al., 2019).

3.6 Takeaway

Our key takeaways from the related works are: CNNs outperform classical models in skin cancer classification, ensemble methods and advanced CNN architectures enhance performance, addressing ethnic bias and dataset limitations is crucial, and effective model selection and data pre-processing are essential for achieving the best results.

4 Data processing

This section outlines the steps involved in preparing the HAM10000 dataset for our deep learning model. These steps include data sourcing, cleaning, resizing, normalizing, and partitioning, all of which help ensure that the dataset is well-suited for training, validation, and testing.

4.1 Data Sourcing

Our data will be sourced from the popular data science platform Kaggle. Specifically, "Skin Cancer MNIST: HAM10000" [10] - which can be summarized into an image and patient attributes below:

lesion_id	image_id	age	sex	localization

The HAM10000 dataset is a large collection of labeled dermatoscopic images aimed at supporting skin lesion classification research. It contains 10,015 images of pigmented skin lesions, which are divided into the seven different categories below:

- Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec)
- Basal cell carcinoma (bcc)
- Benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl)
- Dermatofibroma (df)
- Melanoma (mel)
- Melanocytic nevi (nv)
- Vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc)

Fortunately, this dataset is pre-sourced, allowing us to focus on developing and optimizing our model without the need for data concatenation.

4.2 Data Cleaning

Given that the HAM10000 dataset was specifically created for machine learning, it is already evenly distributed across attributes and requires minimal low-level data cleaning. Although the dataset includes 57 patients with null age values, removing their rows and the corresponding images does not significantly affect our data stratification or overall quantity.

4.3 Data Resizing

To ensure consistency in input dimensions, all images from the HAM10000 dataset will be resized to 224x224 pixels. This size is widely used in CNN architectures and the related works seen above for computational efficiency. [11]

4.4 Data Normalizing

All image pixel values will be normalized to a range of [0, 1] by dividing by 255 (RGB max). This normalization ensures consistent pixel intensity values across images and prevents our model from assigning disproportionate weight to higher pixel values.

4.5 Data Partitioning

The dataset will be split into training, validation, and test sets, with 80% allocated for training, 10% for validation, and 10% for testing. Each allocation will be stratified.

5 Architecture

Our proposed skin cancer classification network leverages a hybrid architecture that integrates both Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) to process and combine image data with patient metadata. This architecture is designed to capture both visual and non-visual information relevant to skin cancer diagnosis.

5.1 CNN for Image Feature Extraction

The CNN component of the network is responsible for extracting high-level features from the input image of the skin lesion. We will experiment with architectures like ResNet [12], EfficientNet [13], and DenseNet [14] due to their success in medical imaging tasks.

The output from the CNN is a feature vector representing the image at a high level of abstraction, which serves as input to the subsequent decision-making stages of the network.

5.2 Metadata Integration

In addition to image data, our model incorporates patient-specific metadata, including age, gender, and medical history. To integrate this metadata effectively, we concatenate it directly with the CNN-extracted image features. This allows the fully connected layers (ANN) to process both the visual and non-visual information simultaneously, ensuring that the metadata provides context for the image features.

We opted for direct concatenation over late fusion due to the inconclusiveness of the metadata, which may be insufficient without the contextual images.

5.3 Fully Connected Layers

After the concatenation of the image feature vector and the metadata, the resulting combined feature vector is passed into a series of fully connected layers. These layers function as classifiers, where the model learns to correlate the combined information and make a final prediction about the skin lesion's classification.

6 Baseline Model

A research named "Towards Skin Cancer Classification Using Machine Learning and Deep Learning Algorithms: A Comparison" mentioned before in Background & Related Work will be used to decide the baseline model for comparing our model. The result of this research which includes performance of multiple models will ensure us that CNN should perform better than these baseline models. Therefore, Decision Tree and Logistic Regression are selected as they provide slightly worse predictions than CNN based on accuracy and recall (Khan et al., 2021). As a brief explanation to these models, Decision Tree is a tree-like structure that splits data into subsets based on feature values at each node, with the goal to create subsets that ideally contain only one class of the outputs for prediction, and Logistic Regression is a model which uses sigmoid function to map input features to the probabilities of being in a class. We will also use Vanilla CNN, which we expect to produce better results as our model gets more complex. Lastly, We will use the model from the winner of ISIC 2020 competition as our upper bound.

7 Ethical Considerations

The use of AI in skin cancer diagnosis raises several ethical concerns. First, there is the risk of algorithmic bias due to unrepresentative training data. [19] AI systems may perform poorly on underrepresented demographic groups in the dataset, leading to disparities in diagnosis accuracy. Studies have shown that AI models trained on biased datasets can lead to unequal treatment. [20]

Another concern is over-reliance on AI by both patients and clinicians. While AI can assist in diagnosis, it is not immune to errors. The black-box nature of deep learning algorithms [causality vs correlation] means it can be difficult to understand how a particular decision was made, raising questions about accountability when errors occur. This highlights the importance of maintaining human oversight and using AI as a tool to support, rather than replace, clinical judgment.

Finally, privacy concerns must be considered, especially when handling sensitive patient data. Our deep learning model will rely on large datasets of medical images and patient metadata. If these are not properly encrypted or secured, there is a risk of data breaches. Also, medical patients should give informed consent for using their data and clearly understand how it is being used.

8 PROJECT PLAN

The team has set 1 meeting time per week, on Sunday afternoon from 4-5 PM. A backup meeting time is on Wednesday evenings 6-7 PM. To ensure seamless communication, we've made a whatsapp group where all members are active. We've set an internal 3 strike policy should team members be unable to finish their assigned task (extenuating circumstances etc.). We will be using zoom as our means of video communication, and will adopt technologies like chat gpt to improve our productivity. We will have a shared github colab link, along with a shared folder on google drive with our research and other deliverables. The team has set an internal deadline of 1-2 days before each deliverable to plan for contingencies.

8.1 Project Proposal		
Responsibilities & Weights	**Assignee**	**Deadline**
Introduction (4)	Keshav	10/3/2024
Illustration (4)	Youssef	10/3/2024
Background & Related Work (4)	Sean	10/3/2024
Data Processing (4)	Abhishek	10/3/2024
Architecture (2)	Youssef	10/3/2024
Baseline Model (2)	Sean	10/3/2024
Ethical Considerations (2)	Keshav	10/3/2024
Project Plan (4)	Keshav & Sean	10/3/2024
Risk Register (4)	Keshav & Abhishek	10/3/2024
Link to Github Colab (1)	All members	10/3/2024
References (1)	All members	10/3/2024
Structure, grammar (8)	Youssef & Abhishek	10/4/2024
8.2 Project Progress Report		
Responsibilities & Weights	**Assignee**	**Deadline**
Brief Project Introduction (5)	Keshav	10/30/2024
Individual Contributions and Responsibilities (5)	Youssef	10/30/2024
Notable Contributions (10):Data Processing,Baseline Model,Primary Model	Abhishek & Sean	10/30/2024
Document Formatting	All members	10/30/2024
8.3 Project Presentation		
Responsibilities & Weights	**Assignee**	**Deadline**
Rough Draft of Video (Written)	All members	11/25/2024
Problem Definition (4)	Abhishek	11/28/2024
Data Processing (4)	Youssef	11/28/2024
Model Description (4)	Keshav	11/28/2024
Results & Impact (4)	Sean	11/28/2024
Conclusion	Keshav	11/28/2024
Presentation Flow (5), Demonstration (10)	All members	11/28/2024
Video Editing and Upload	Abhishek & Keshav	11/29/2024
8.4 Final Report		
Responsibilities & Weights	**Assignee**	**Deadline**
Introduction (2)	Keshav	11/26/2024
Illustration (2)	Youssef	11/26/2024
Background & Related Work (2)	Sean	11/26/2024
Data Processing (4)	Abhishek	11/26/2024
Architecture (4)	Sean	11/26/2024
Baseline Model (4)	Keshav	11/26/2024
Quantitative Results (4)	Abhishek	11/26/2024
Qualitative Results (4)	Youssef	11/26/2024
Model Evaluation on New Data (10)	Keshav & Youssef	11/26/2024
Discussion of Project (8)	Sean & Abhishek	11/26/2024
Ethical Considerations (2)	Keshav	11/26/2024
Explain Project Difficulty/Usefulness (6)	Keshav & Sean	11/27/2024
Structure, Grammar, and Mechanics (8)	Abhishek & Youssef	11/27/2024
Contribution Summary Table	All members	11/27/2024

9 RISK REGISTER

We've identified risk factors which can lead to potential problems during the course of this project.

9.1 DATA IMBALANCE IN TRAINING DATASET

- Risk: The data set may be imbalanced, with certain skin types, cancers and demographics being under-represented. This can lead to biased predictions, and the model can fixate on unproductive features, rather than finding the underlying causal relationships leading to skin cancer. For example, let's consider a skewed data set of 85% males and 15% females. If all of the female patients in the dataset are cancer positive, then the model would fixate on that correlation and make inaccurate real world predictions.
- Mitigation: The team will ensure a balanced representation across all groups and demographics. The model performance across all demographics will be measured individually to identify any disparities with accuracy,

9.2 Team member drops out

- Risk: A team member may drop out of the course in the middle of the term
- Mitigation: Clear communication will be maintained, and specific responsibilities will be documented so that remaining team members re-distribute the extra work efficiently.

9.3 Inability to meet internal deadlines

- Risk: A team member may miss internal deadlines due to external commitments
- Mitigation: Regular check-ins and progress updates will be scheduled to monitor the project timeline. If someone falls behind, tasks will be reassigned. We will have a 3 strike system, before the issue is communicated with the professor.

9.4 Hardware limitations - if the model takes too long to train

- Risk: Training deep learning models is resource-intensive, and limited access to GPUs or cloud computing resources may delay model training and testing.
- Mitigation: The team will use external cloud computing platforms like AWS, or Azure to access additional computational power. If delays persist, the model's complexity will be reduced, or more efficient architectures will be employed.

9.5 Unexpected data quality issues

- Risk: The quality of the data may be compromised due to issues like missing values, corrupted images, or inconsistent metadata.
- Mitigation: The team will conduct an early audit of the dataset to identify and address any issues, including cleaning the data, filling missing values where appropriate, or removing unusable data. The standardization of data and balanced representation across all demographics will be ensured.

10 Google Colab Link

https://colab.research.google.com/drive/1KCFxgy9dRUaq4mT5ytLKJ6Fz9WFWkvL8?usp=sharing

References

[1] "Basal & squamous cell skin cancer statistics," *American Cancer Society*, https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/key-statistics.html (accessed Oct. 4, 2024).

[2] Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries - Bray - 2018 - *CA: A Cancer Journal for Clinicians* - Wiley Online Library, https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21492 (accessed Oct. 4, 2024).

[3] "Melanoma survival rates," *Melanoma Research Alliance*, https://www.curemelanoma.org/about-melanoma/melanoma-staging/melanoma-survival-rates (accessed Oct. 4, 2024).

[4] T. Mazhar et al., "The role of machine learning and deep learning approaches for the detection of skin cancer," *Healthcare (Basel, Switzerland)*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9914395/ (accessed Oct. 4, 2024).

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, NeurIPS, https://proceedings.neurips.cc/paper_2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (accessed Oct. 4, 2024).

[6] S. H. Khan, H. N. Abbas, and M. J. Khan, "Towards skin cancer classification using machine learning and deep learning algorithms: A comparison," *IEEE Access*, vol. 9, pp. 15457-15465, 2021.

[7] Mahbod, A., Schaefer, G., Ellinger, I., Pitiot, A., Wang, C., and Ecker, R., "Fusing fine-tuned deep features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 84, pp. 101-111, 2020.

[8] M. Dildar, M. Akram, T. Fatima, and N. Raza, "Skin cancer detection: A review using deep learning techniques," *International Journal of Environmental Research and Public Health*, vol. 18, no. 10, pp. 5479-5490, May 2021.

[9] X. Wu, Y. Li, Q. Liu, and W. Jiang, "A comparative study of convolutional neural network variants for skin cancer classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1402-1412, Apr. 2022.

[10] T. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000.

[11] Simonyan, K., and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[12] 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.

[13] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML).*

[14] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. CVPR.

[15] Samirsalman. (n.d.). Samirsalman/skincancer-machine_learning_project: Skin cancer diagnosis using different machine learning methodologies and Techniques. GitHub. https://github.com/samirsalman/SkinCancer-Machine_Learning_Project

[16] prashantverma13. (2023, July 20). Skin cancer classification: CNN approach. Kaggle. https://www.kaggle.com/code/prashantverma13/skin-cancer-classification-cnn-approach/notebook

[17] Siim-ISIC melanoma classification. Kaggle. (n.d.). https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/175412

[18] A. Kaushal, R. Altman, and C. Langlotz, ”Health care AI systems are biased,” *Scientific American*, https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/ (accessed Oct. 4, 2024).

[19] O. Z. B. C. S;, ”Dissecting racial bias in an algorithm used to manage the health of populations,” *Science (New York, N.Y.)*, https://pubmed.ncbi.nlm.nih.gov/31649194/ (accessed Oct. 4, 2024).

[20] ”Dissecting racial bias in an algorithm used to manage the health of populations,” *Science (New York, N.Y.)*, https://science.org/doi/10.1126/science.aax2342 (accessed Oct. 4, 2024).

[21] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, ”Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata,” *MethodsX*, vol. 7, pp. 100864, 2020. Available: https://www.sciencedirect.com/science/article/pii/S2215016120300832 (accessed Oct. 4, 2024).

[22] OpenAI, ”ChatGPT,” OpenAI, Oct. 4, 2024. [Online]. Available: https://chat.openai.com/