A Deep Learning Analysis of Skin Lesion Classification

Sarah Morrison
Cornell University
sam583@cornell.edu

Rahul Nathan
Cornell University
rn333@cornell.edu

Paige Brown
Cornell University
pb584@cornell.edu

December 12, 2023

Abstract

The final project in AML provided our team with an opportunity to apply the deep learning models covered in class to real-world datasets as well as explore new techniques independently. The chosen dataset for this project was chosen from the collection of open source datasets provided by the International Skin Imaging Collaboration (ISIC). We implemented three distinct deep learning models: Sequential CNN, DenseNet, and LatentNet. Among these, the DenseNet model performed the best with an accuracy of 77.68%. This report will delve into the motivation behind the project, review relevant literature, provide insights into the different modeling approaches employed, present results derived from Exploratory Data Analysis (EDA) and model performance evaluations, and discuss limitations encountered during the project along with considerations for future work.

1 Introduction

The motivation for this project can be broken down into four main groups including prevalence, accessibility barriers, and accuracy limitations. In regards to prevalence, skin cancer is the most common cancer globally. Approximately 9500 Americans are diagnosed with some form of skin cancer daily, with over 20 Americans dying every day from melanoma specifically. Furthermore, Accessing a dermatologist for a skin biopsy or body scan can be challenging and often involves extremely long wait times of several months. This is a concern in regards to skin cancer (ex. melanoma) as early detection is crucial to increasing patient survival rates. In addition to these accessibility barriers, even when patients are able to access medical care regarding a skin abnormality, skin cancer is often misdiagnosed. This is especially true when the physician does not have specific dermatology training (ex. Primary Care Providers). Thus for this project we aim to combat the lack of access and accuracy by completing an analysis of how machine learning models can be leveraged to predict an individual's risk for skin cancer. For this analysis we aim open source labeled data from the International Skin Imaging Collaboration to develop three deep learning models that can be used as a screening tool for seven common skin disorders, some of which are cancerous (ex. melanoma). We will also specifically be aiming to combat data size limitations and potential algorithmic bias within the data by utilizing transfer learning techniques, complex deep learning model architecture, and novel LatentNet model architecture.

2 Related Works

2.1 Over Detection of Melanoma Suspect Lesions by a CE Smartphone Application

This is an interesting study that aims to prospectively investigate the diagnostic accuracy of a CE-certified mobile device application, SkinVision, in early melanoma screening. For context, a CE certification indicates that a medical device, software application, etc. meets the European requirements for safety, health, environmental, and consumer protection. This study observed that the SkinVision application classified an extremely high number

of lesions as high risk when compared to dermatologists which the researchers report would have led to a clinically harmful number of unnecessary biopsies/incisions. Further analysis observed that the confidence in the app was low amongst both patients and dermatologists. This paper specifically highlights the complex risk involved in trying to deploy a medical screening application that successfully achieves a high prediction accuracy without simultaneously allowing a high threshold for false positive predictions.

2.2 Automated Bias Reduction in Deep Learning Based Melanoma Diagnosis using a Semi Supervised Algorithm

In this paper researchers study the algorithmic racial bias that is often found during skin imaging classification problems specifically when training deep learning algorithms on large sets of publicly available data such as data sets from the International Skin Imaging Collaboration (ISIC) which is being used in our current analysis. This paper highlights that these large datasets used to train skin imaging datasets often have a lack of diversity and over represent light skin with minimal body hair, which consequently results in the model being less transferable to the more diverse global population. Thus, these researchers introduce a new model architecture called LatentNet to address these racial bias concerns. The LatentNet model architecture leverages a variational auto-encoder to detect over represented skin imaging features and reduce their weights during the training of the deep learning model in an attempt to increase the system's transfer-ability across different skin color demographics. When deployed on four unique skin color groups the LatentNet model performed significantly better than a baseline deep convolutional neural network demonstrating it's increased transferability.

2.3 Dermatologist Level Classification of Skin cancer using Neural Networks

This paper provides broad insights into an image classification analysis pipeline for the purposes of skin cancer detection. In this paper the authors attempt to classify two skin types of cancers by training a convolutional neural network (CNN) model on over 125,000 skin lesion images. The authors are specifically working on the binary classification of keratinocyte carcinomas VS. benign seborrheic keratoses and malignant melanomas VS. benign nevi. The model's performance in this study was tested against 21 certified dermatologists, and performed on par with all dermatologists indicating the models high competency levels.

2.4 Development of a Convolutional Neural Network to detect abnormal aortic aneurysms

In this paper the authors seek to develop a machine learning model that can be used to screen for the presence of infra-renal abdominal aortic aneurysms. To achieve this goal the authors developed a CNN model that they trained using computed tomography angiography (CTA) scans. One specific stage of model development that the authors highlighted was their integration of transfer learning, specifically by using the ImageNet database to optimize their model and overcome data limitations. The model was validated using both prediction accuracy and Area Under the Receiver Operating Curve (AUROC), which were revealed to be 99.1% and 0.99 respectively. The results indicate that the screening of abdominal aortic aneurysms can be done effectively using a CNN model supplemented with transfer learning.

3 Methods

3.1 Exploratory Data Analysis Approach

For the exploratory data analysis, we wanted to see if there was correlation between features, observe if there were some skin conditions that were more prevalent than others and with which groups, and more. Before doing this, we wanted to ensure that we could see the images that fell within each category of skin conditions. To do this, we grouped the images by their classifications and then displayed them within the notebook. This allowed us to see the similarities visually between images within the same category.

Next, we created a bar graph to see the amount of individuals that had conditions from each cell type according

to the metadata files. Thus, we were able to see how many people had conditions of Melanocytic nevi, Melanoma, Benign keratosis-like lesions, Basal cell carcinoma, Actinic keratoses, Vascular lesions, and Dermatofibroma. Following that, we also did a bar graph to visualize the distribution of ages among the dataset to see if a certain range of ages was more common in the dataset than others. This showed us that most individuals in the dataset had age ranges from around 40 - 60.

After this, we made a new kind of bar graph to visualize the distribution of skin conditions (cell type according to the dataset) by sex. It showed that the distribution of these conditions was relatively similar between male and female participants. We did a similar graph, but this time comparing sex and localization of the skin conditions on the body. By plotting the correlation matrix, we saw there was a relatively small correlation between age and cell type. Because of this, we created bar graphs similar to the previous ones to see the distribution of age as it compares to cell type and localization.

Additionally, using the RGB file from the dataset, we plotted the average image from across the dataset. This was to see if there was any information that could be drawn from seeing the average skin condition across all images. We also wanted to visualize the average image before normalizing all the images with the averages in the preprocessing step.

3.2 Data Processing Approach

First, we wanted to be able to map the column 'dx', which describes the type of condition associated with the image, to something more readable in the future. To do this, we created a couple new columns to map the condition codes to their readable names, along with a number for their categorical classification.

Next, we wanted to check to see if there are any data points with missing values. This showed us that there were 57 null values in the age feature, so we decided to replace the missing values in the age feature with the average age across all data points.

Following this, we resized the images to ensure that the processing time would be shorter with the models in the future. After splitting the dataset into training and test sections, we one hot encoded categorical variables so that they could be interpreted by the model. Finally, we converted the training and testing images to Numpy arrays in order to take the average and standard deviation. This permitted us to normalize all of the images for increased accuracy and decreased training time.

3.3 Sequential CNN Modelling Approach

The first modeling approach was a simple sequential CNN in order to obtain a baseline for this task. The sequential CNN was selected for its simplicity and effectiveness in image classification tasks. In addition, this model was discussed in the class and we also had practice implementing this in Homework #5. The architecture for this CNN consisted of three convolutional layers, each followed by ReLU activations, and max-pooling layers. Also, a kernel size of 3 and a padding of 1 were used for the convolutional layer and a stride of 2 was used for the pooling layers. This network structure is extremely common in image recognition tasks since it allows the model to learn hierarchical features from input images and captures spatial patterns at a variety of levels.

The decision to implement a sequential CNN was motivated by its simplicity and ease of obtaining a baseline for this task. For the sequential CNN's training process, the Adam optimizer with a learning rate of 1e-3 was chosen for its adaptive learning rate capabilities and efficient weight updates. The cross-entropy loss function was utilized to measure the model's performance during training and it is a common approach for multi-class classification tasks. Lastly, the weights of the model were initialized randomly, enabling the model to learn from scratch based on the characteristics of the given dataset.

3.4 DenseNet Modelling Approach

The second modeling approach was to implement DenseNet, which is widely used for image classification. The decision to use DenseNet was motivated by its ability to capture intricate patterns through densely connected layers, promoting feature reuse and gradient flow throughout the network. DenseNet121 was initialized with pre-trained weights from ImageNet to leverage knowledge gained from a diverse and large-scale dataset with over 14 million images. For the task at hand, the last fully connected layer was adapted for the target dataset, allowing the model to adapt to the unique characteristics

of skin lesion images while retaining general knowledge acquired from ImageNet. Similar to the sequential CNN, the Adam optimizer with a learning rate of 1e-3 was employed for optimization, and the cross-entropy loss function was used to measure the model's performance.

3.5 LatentNet Modelling Approach

The final modeling approach was the development of a simplified LatentNet model architecture which was motivated by the paper "Automated Bias Reduction in Deep Learning Based Melanoma Diagnosis using a Semi Supervised Algorithm", a description of which can be found in the related works section. At a high level, the purpose of this approach is to develop a model that combats the algorithmic bias that often exists when skin cancer classifiers are trained using photos of predominantly light skin tones. The research paper this model is inspired by was able to show that using a LatentNet model architecture significantly improves the model's ability to generalize across different skin colors, thus we wanted to go beyond the development of the traditional deep learning models and attempt to develop a simplified version of this exploratory LatentNet approach. To create a simplified version of the model architecture described in this research paper we decided to combine a variational autoencoder (VAE) with the architecture of a convolutional neural network for classifying the skin lesion images. For context, a VAE is a type of artificial neural network used for generative tasks, which learns to encode input data into a compressed representation and then decode it back to the original format. The first step was to define the VAE encoder which takes an input image and processes it through convolutional, pooling, and dense layers to produce the mean and log variance. The next step was to define a sampling layer which generates latent vectors from the mean and log variance produced by the encoder. Next a decoder attempts to reconstruct the original image using the latent vector. The loss function for the VAE is defined in terms of both the reconstruction loss and the divergence loss. The reconstruction loss is the mean squared error between the reconstructed image and the original image. Yet, a VAE alone is not enough for the completion of image classification and thus we have combined the VAE with a convolutional neural network in order to perform the broader image classification task. In order to do this we have defined a modified sequential CNN architecture that classifies an image based on the output of the VAE encoder. The architecture of the modified CNN is similar to the previously mentioned sequential CNN approach. The primary evaluation metric for this exploratory model architecture was prediction accuracy.

4 Results

4.1 Exploratory Data Analysis

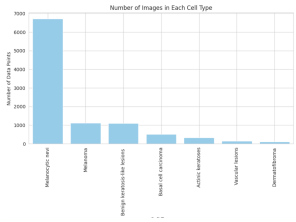

Figure 1: Frequency of Skin Disease in Dataset

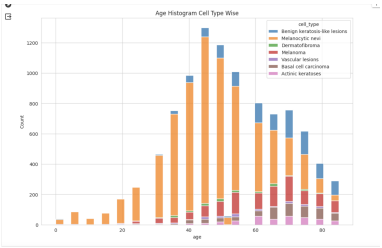

Figure 2: Distribution of Cell Type with Age

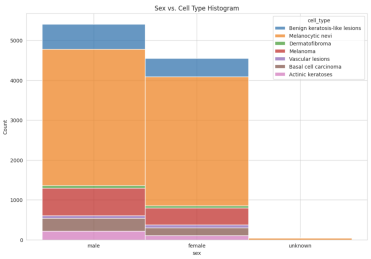

Figure 3: Distribution of Cell Type with Sex

4.2 Deep Learning Results Overview

Model	Accuracy
Sequential CNN	70.95%
DenseNet	77.68%
LatentNet	69.58%

Table 1: Model Accuracy Scores

4.3 Sequential CNN Modelling

After training and testing the sequential CNN model that we used as a baseline, it resulted in a validation accuracy of 70.95%. In addition, a confusion matrix was created in order to visualize which data points were incorrectly classified and the differences between the true and predicted outputs. This is shown in Figure 4:

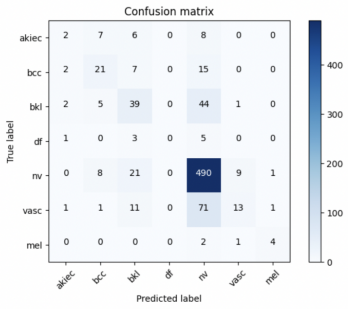

Figure 4: Sequential CNN Confusion Matrix

Some insights that can be generated from this confusion matrix above is that there is an imbalance in the validation data set because we see a total of 529 true data points for nv (Melanocytic Nevi) and 9 true data points for df (Dermatofibroma). Apart from that, one observation to note is that bkl (Benign Keratosis-like Lesions) is misclassified as nv almost 48% of the time. In addition, bcc (Basal Cell Carcinoma) is misclassified as nv 33% of the time. Therefore, some of these categories may overlap, resulting in this misclassification.

4.4 DenseNet Modelling

After training and testing the DenseNet model, it resulted in a validation accuracy of 77.68%. The confusion matrix for the outputs of this model is shown in Figure 5:

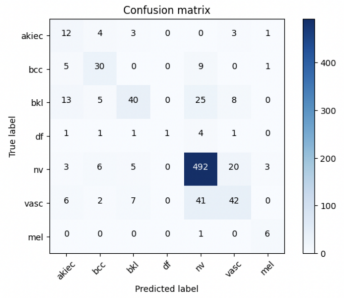

Figure 5: DenseNet Confusion Matrix

Some insights that can be generated from this confusion matrix in comparison to the one outputted by the sequential CNN model is that there are fewer misclassifications. One example is that bcc is now misclassified as nv only 20% of the time rather than 33% of the time. Also, bkl is misclassified as nv 27% of the time rather than 48% of the time. Potential reasonings for why DenseNet works better than a simple sequential CNN will be discussed in the next section of this report.

4.5 LatentNet Modelling

The LatentNet model was run with a limited set of ten epochs due to the significant computational load and frequent iterations associated with the new model development. When run under these constraints, the testing accuracy was calculated to be 69.58%.

5 Discussion

In regards to the exploratory data analysis, the results of the exploratory data analysis showed that Melanocytic nevi was by far the most common skin condition in the dataset. This can be seen in Figure 1. There are more than 6500 instances of Melanocytic nevi recorded in the dataset, and it was also the most common condition among both male and females, and most age groups.

However, as ages increased, there seemed to be a trend of other skin conditions becoming proportionally more common in addition to Melanocytic nevi, as seen in Figure 2. Whereas Melanocytic nevi was by far the most common condition in most age groups, others like benign keratosis-like lesions became more apparent in older age groups. When observing the frequency of cell-types in males versus females in Figure 3, the rate of appearance of the conditions in each group was relatively similar. As mentioned above, we observed a small correlation between age and cell type.

In regards to evaluating the performance of the implemented deep learning models, it became evident that DenseNet outperformed the Sequential CNN in skin lesion classification. A potential explanation for this can be due to the utilization of transfer learning and densely connected model layers. This approach involves leveraging a pre-trained model on a large and diverse dataset such as ImageNet.DenseNet's incorporation of pre-trained weights from ImageNet facilitates the model in capturing intricate patterns and features from a wide array of images. The diverse nature of ImageNet allows the model to gain a robust understanding of general image features. As a result, when fine-tuned for the specific task of skin lesion classification, DenseNet generalizes extremely well to the nuances of the provided images. One noteworthy advantage of transfer learning is that it reduces overfitting. The model's exposure to a vast amount of data during its pre-training phase allows it to learn generic features, preventing it from memorizing specific characteristics of the training dataset. Ultimately, this contributes to better performance on unseen data, which is shown in the Results section. In contrast, the Sequential CNN, while effective, lacks the inherent advantages of pre-trained weights and the broad knowledge base that transfer learning provides. Its performance is more reliant on learning features solely from the task-specific dataset, potentially limiting its ability to identify complex patterns present in these images.

In regards to the exploratory LatentNet approach, the prediction accuracy of this model was just slightly lower than the baseline sequential CNN model approach, and notably lower than the optimized DenseNet model approach. One possible reason is that the LatentNet model integrated a fairly basic CNN model, which was similar to the sequential CNN architecture and a variational autoencoder (VAE), to perform its classification. Thus perhaps the performance could have been improved via integrating the VAE with a more complex model architecture. Regardless, it's important to highlight that the goal of this exploratory LatentNet approach was not to maximize the prediction accuracy on this particular dataset but to develop an algorithm that could overcome the racial biases of this dataset and generalize better across different skin color demographics, using the aforementioned research paper as the foundational proof of concept. An interesting next step would be to collect data from a more diverse population and complete a comparative analysis between our simplified LatentNet model's performance and the baseline sequential CNN performance, given that they have similar model architecture. This would provide interesting insight into the transferability of each model.

For next steps, one task we are particularly interested in pursuing is increasing the model transparency for each of our deep learning models via the development of saliency maps. Deep learning models are frequently thought of as black box systems, but given that these models are being deployed to combat a high risk medical issue it is crucial that the explainabililty of each model be increased. The specific saliency maps technique provides a clear visualization of which image areas are most influential in the predictions made by the deep learning model, which consequently helps users to understand how the model is making decisions.

6 References

[1] Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., & Thrun, S. (2017, January). Dermatologist-level classification of skin cancer with Deep Neural Networks. Nature.
https://pubmed.ncbi.nlm.nih.gov/28117445/

[2] Jahn, A., Kunz, M., Huber, S., Kostner, L., Cerminara, S., & Navarini, A. (2022, August). Over-

detection of melanoma-suspect lesions by a CE-certified smartphone app: Performance in comparison to dermatologists, 2d and 3D convolutional neural networks in a prospective data set of 1204 pigmented skin lesions involving patients' perception. Cancers.
https://pubmed.ncbi.nlm.nih.gov/35954491/

[3] Das, S. (2021, January 1). Automated bias reduction in deep learning based melanoma diagnosis using a semi-supervised algorithm. medRxiv.
https://www.medrxiv.org/content/10.1101/2021.01.13.21249774v1.full-text

[4] Camara, J., Tomihama, R., Pop, A., Shedd, M., Dobrowski, B., & Knox, C. (2022, May). Development of a convolutional neural network to detect abdominal aortic aneurysms. Journal of vascular surgery cases and innovative techniques.
https://pubmed.ncbi.nlm.nih.gov/35692515/

[5] Ruiz, P. (2018). Understanding and Visualizing DenseNets. Towards Data Science.
https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

[6] Mallick, D. (2021). Using Sequential Module to Build a Neural Network. Medium.
https://medium.com/writeasilearn/using-sequential-module-to-build-a-neural-network-a34ca3f37203

[7] Suephy C. Chen, M. (2001, December 1). A comparison of dermatologists' and primary care physicians' accuracy in diagnosing melanoma. Archives of Dermatology.
https://jamanetwork.com/journals/jamadermatology/fullarticle/478587

[8] Liddy, C. (2020, June). How long are Canadians waiting to Access Specialty Care? retrospective study from a Primary Care Perspective. Canadian family physician Medecin de famille canadien.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7292524/

[9] Duniphin, D. D. (2023, January 1). Limited access to Dermatology Specialty Care: Barriers and teledermatology. Dermatology practical & conceptual.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9946088/