FINAL REPORT

Cairo Cristante
Student# 1008301348
cairo.cristante@mail.utoronto.ca

Anastasia Dimov
Student# 1006880038
anastasia.dimov@mail.utoronto.ca

Miranda Su
Student# 1007039973
miranda.su@mail.utoronto.ca

Kate Bowen
Student# 1008215539
k.bowen@mail.utoronto.ca

ABSTRACT

Skin cancer is a dangerous disease that exhibits a significantly reduced survival rate upon metastasis. However, early detection dramatically enhances the treatability of skin cancer. This project aims to address the challenge of early detection by developing accessible machine-learning tools. This report outlines existing solutions and the process Group 49 made to develop CNN and Autoencoder-based models in addition to a fine-tuned MobileNet-based model. The report details these developed models' data processing, architecture and results. —-Total Pages: 9

1 INTRODUCTION

This project aims to develop a deep-learning model for skin lesion classification, involving a comparison between three models: a vanilla Convolutional Neural Network, a Convolutional Autoencoder, and a MobileNet for feature extraction.

Motivated by the urgency of early skin cancer detection and accurate diagnoses, the team sought to address this critical need. Melanoma, for instance, is a type of skin cancer that can become fatal in less than six weeks. While skin cancer is highly treatable when detected early, the survival rate drops substantially once it spreads, and treatment becomes strenuous betterhealth (2023). Moreover, the team was interested in exploring the performance of different convolutional models on the same problem.

All models were trained and tested on skin lesion images from the International Skin Imaging Collaboration (ISIC) gallery, encompassing over 25 skin conditions. For this project, the scope was narrowed down to seven specific skin conditions: melanoma, nevus, base cell carcinoma, seborrheic keratosis, pigmented benign keratosis, actinic keratosis, and squamous cell carcinoma ISIC.

The importance of this project stems from its potential to reduce skin cancer mortality. The models can serve as a self-surveillance tool for patients, allowing timely interventions, and aiding dermatologists in decision support to improve diagnostic accuracy. Given the intricacies of skin lesion analysis, deep learning is a suitable approach, capable of learning critical characteristics and patterns from extensive datasets. Convolutional Neural Networks (CNNs), Convolutional Autoencoders (CAEs), and MobileNet collectively all excel at capturing nuanced features from images and automatically modeling essential patterns, which are important qualities for accurate skin lesion classification.

2 BACKGROUND AND RELATED WORK

The following discusses research papers and current industry architecture used to classify and aid in diagnosing skin cancer and skin lesions. This background work will serve to influence the direction the team takes in developing the project in terms of CNN construction, data sourcing/prepossessing and expansion to wider consumer bases.

2.1 University of Espírito Santo - The impact of patient clinical information on automated skin cancer detection

The University of Espírito Santo conducted a study comparing training a CNN solely on skin lesion image data versus the effect of subsequently training the CNN on both image data and clinical data Pacheco & Krohling (2020). The researchers set up two scenarios with and without the clinical data and compared the accuracy of a series of well know CNN architectures when fed the varying data Pacheco & Krohling (2020). The architecture considered included various versions of ResNet, GoogleNet, VGGNet and MobileNet. The scenario, which includes the clinical data, has feature-reducing layers, which use a ceiling operator to reduce the number of image parameters to allow for the input of clinical data, which is much smaller in comparison to the image data. The article finds an increase in prediction accuracy in four of the six skin lesion classes considered, with the other two classes seeing little to no change Pacheco & Krohling (2020).

2.2 Infared Thermography

Researchers are exploring the effect that the use of infrared images has on the accuracy of skin cancer classification machine learning models Magalhaes et al. (2021). They used various AI strategies such as support vector machines, ANN and CNNs testing a dataset of skin lesion images with inferred thermography expansion and without Magalhaes et al. (2021). The study found promising results for the use of infrared images when it comes to melanoma, nevi and non-melanoma skin cancer classification Magalhaes et al. (2021). However, the models struggled to classify between benign and malignant skin lesions. The researchers stress the importance of additional supplementary information regarding the patient to aid in the model's accuracy, furthering the conclusions made in Section 2.1 Magalhaes et al. (2021).

2.3 SkinVision

SkinVision Inc. (2021) is a smartphone app that monitors moles, makes a body map, and assesses the skin cancer risk based on moles' appearance and skin type. It utilizes a fractal theory algorithm that characterizes a mole's risk of skin cancer based on size, shape, colour, growth rate, and other aspects. This provides a more accurate diagnosis since it takes into account the fractal patterns of different skin lesions. In a study conducted by Maier et al, SkinVision results were compared to clinically analyzed results for accuracy Maier et al. (2015). This study analyzed 144 lesions, and the app could make an accurate risk assessment on an 81% accuracy basis Maier et al. (2015). Clinically analysis of those lesions yields a high accuracy.

2.4 Aysa from VisualDx

Aysa is an AI-powered skin rash app. Users can take or upload images of their rash or other skin condition and enter any symptoms. The app will then provide possible matches on the skin condition and guidance for the next steps. It does not provide a medical diagnosis. Aysa is powered by VisualDx, a diagnostic clinical decision support system currently used in hospitals and clinics. The company has a selection of patented medical data and images Graedon (2018).

2.5 DermAssist from Google

Google has developed a tool called DermAssist Der (2021), that people can use to predict possible skin conditions that they might have. The user takes a picture of their skin and inputs any other related symptoms that they might have. All the information it gives the user after the screening process is dermatologist-reviewed. Google has said that the tool is not intended to give medical advice or be used instead of seeing a professional. Its intended use is to help users seeking care make informed decisions about their health. So far, Google has published several papers studying the model to ensure that it is fit for use Peggy Bui (2021).

3 Ethical Considerations

Over the course of this project, a few different ethics questions have been discussed, mainly concerning how the data used to train the models were sourced to maintain patient privacy and how the model might be used in the future. When deciding what data to use it was important that the metadata and images did not contain any information that could be used to identify the patients. It was also ensured that the data come from a trusted source where the patients were made aware of how their information would be used. Additionally, if the models are released to the public through GitHub or other means, a disclaimer will be attached explaining to users that the models cannot diagnose skin conditions and that the task should be left up to a healthcare professional.

4 Data Processing

The following details the datasets chosen to train, test and demonstrate the model's abilities. It additionally details the steps to clean the metadata and pre-process the image data before being fed into the models.

4.1 Data Selection

The team will be used the International Skin Imaging Collaboration (ISIC) to train our model. The dataset combines various datasets from different universities, hospitals, and AI competitions organized into collections Collections.

For the initial model construction, the team has selected the HAM10000 collection due to its more extensive distribution of lesion diagnoses compared to other collections, such as the 2019 Challenge and its relatively small storage size compared to the 2020 Challenge collection.

Additionally, all images within HAM10000 are sourced from the ViDIR Group from the Department of Dermatology at the Medical University of Vienna Medical University of Vienna. Since all images come from the same source, the other collections from different sources can be used to get a more genuine test of the model's overall accuracy. We will use the HIBA Collection to test the model accuracy for the initial models, as all of its data is sourced from Hospital Italiano de Buenos Aires de Buenos Aires.

Lastly, to demonstrate the model on data that has not been seen by the model thus far, we will be using the PROVe-AI collection from the ISIC, which sources its data from the Memorial Sloan Kettering Cancer Center web.

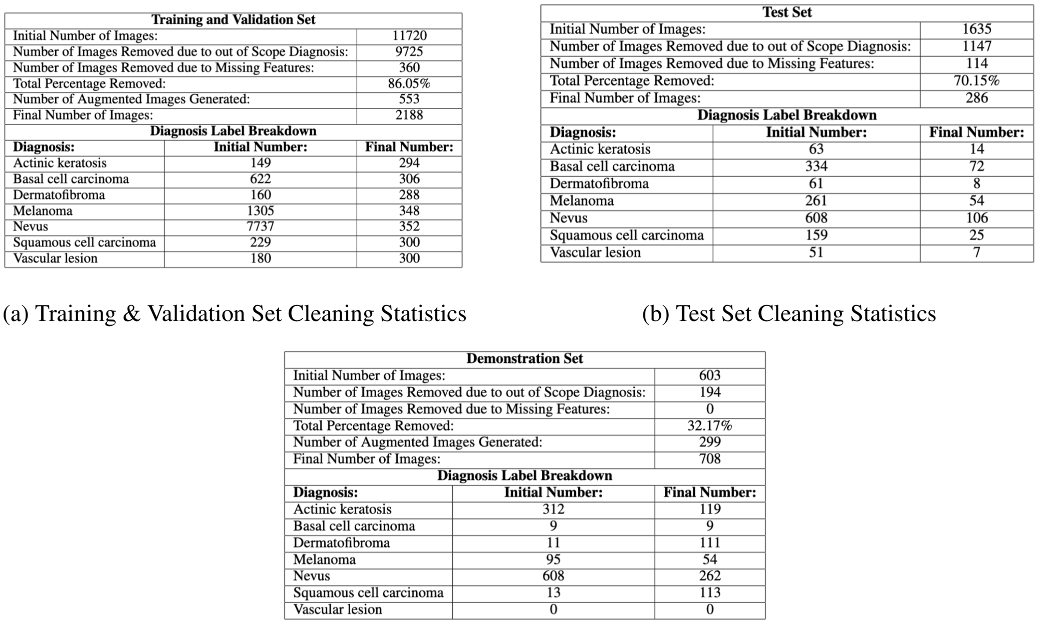
4.2 Metadata and Image Pre-Processing

The approach the group took was to first process and clean the metadata associated with all of the images, then pass this information into the model using a Pytorch Dataset class which pulls the images files from the folder location based on the image id's found in the metadata CSV file. All data go through the same processing steps, except for the data augmentation and replication steps.

1. All columns with a majority missing data, contained data that was not useful to the problem (such as the attribution section) or columns containing features that were not consistent among all of the ISIC collections mentioned were outright removed and not considered
2. Images/data points containing missing data OR diagnosis that were out of the project scope were also removed from consideration as the model would not be able to consider these
3. Images with a nevus, melanoma or basal cell carcinoma diagnosis labels were removed in varying amounts such to eliminate a class imbalance, biasing these classes
4. The age column was normalized using Min-Max Normalization to eliminate the magnitude effect on the model
5. All categorical features (Site Location & Diagnosis) were converted into one-hot encoding representation

6. Augmented images for the classes other than nevus, melanoma or basal cell carcinoma were generated with varying degrees of rotation and metadata was kept the same as the original image
7. All metadata was saved into separate CSV files and loaded into the model using a custom Pytorch Dataset class found in the “utils.py” file
8. For both the CNN & Autoencoder, the images were simply resized to be of dimensions 450x450 pixels.
9. For the MobileNetV2 models, the images first had their height resized to 256 pixels. Next, the image was cropped to the centre such that the new image’s size was 224x224 pixels. Lastly, the image values were normalized with each RGB channel being normalized with means of [0.485, 0.456, 0.406] and standard deviations of [0.229, 0.224, 0.225], respectively.

The results of the cleaning for all of the datasets used for our models and in this report can be found in Figure 1.



Training and Validation Set		
Initial Number of Images:		11720
Number of Images Removed due to out of Scope Diagnosis:		9725
Number of Images Removed due to Missing Features:		360
Total Percentage Removed:		86.05%
Number of Augmented Images Generated:		553
Final Number of Images:		2188
Diagnosis Label Breakdown		
Diagnosis:	**Initial Number:**	**Final Number:**
Actinic keratosis	149	294
Basal cell carcinoma	622	306
Dermatofibroma	160	288
Melanoma	1305	348
Nevus	7737	352
Squamous cell carcinoma	229	300
Vascular lesion	180	300

(a) Training & Validation Set Cleaning Statistics

Test Set		
Initial Number of Images:		1635
Number of Images Removed due to out of Scope Diagnosis:		1147
Number of Images Removed due to Missing Features:		114
Total Percentage Removed:		70.15%
Final Number of Images:		286
Diagnosis Label Breakdown		
Diagnosis:	**Initial Number:**	**Final Number:**
Actinic keratosis	63	14
Basal cell carcinoma	334	72
Dermatofibroma	61	8
Melanoma	261	54
Nevus	608	106
Squamous cell carcinoma	159	25
Vascular lesion	51	7

(b) Test Set Cleaning Statistics

Demonstration Set		
Initial Number of Images:		603
Number of Images Removed due to out of Scope Diagnosis:		194
Number of Images Removed due to Missing Features:		0
Total Percentage Removed:		32.17%
Number of Augmented Images Generated:		299
Final Number of Images:		708
Diagnosis Label Breakdown		
Diagnosis:	**Initial Number:**	**Final Number:**
Actinic keratosis	312	119
Basal cell carcinoma	9	9
Dermatofibroma	11	111
Melanoma	95	54
Nevus	608	262
Squamous cell carcinoma	13	113
Vascular lesion	0	0

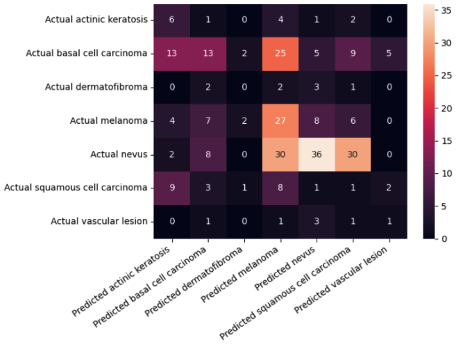
(c) Demonstration Set Cleaning Statistics

Figure 1: Data Metadata Cleaning Statistics and Results

5 BASELINE MODEL

The baseline model was changed from the progress report as we realized that the data did not accurately reflect what was found in the model due to differences in the data processing. As such, the baseline model was changed to reflect the data pre-processing found in the MobileNetV2 Transformations.

The baseline model uses a simple SVM trained and tested on the same collections used for the primary models (HAM10000 & HIBA, respectively). The model uses $C = 1$ and a Radial Basis Function Kernel. It has an overall accuracy of brought 29%, which is rather low but makes sense for this simple model. The baseline model results statistics can be found in Figure 2.



(a) Baseline Model Confusion Matrix

Baseline Model Statistics			
Class	Precision	Recall	F-1 Score
Actinic keratosis	0.18	0.43	0.25
Basal cell carcinoma	0.37	0.18	0.24
Dermatofibroma	0.00	0.00	0.00
Melanoma	0.28	0.50	0.36
Nevus	0.63	0.34	0.44
Squamous cell carcinoma	0.02	0.04	0.03
Vascular lesion	0.12	0.14	0.13
Accuracy:			0.29

(b) Test Set Cleaning Statistics

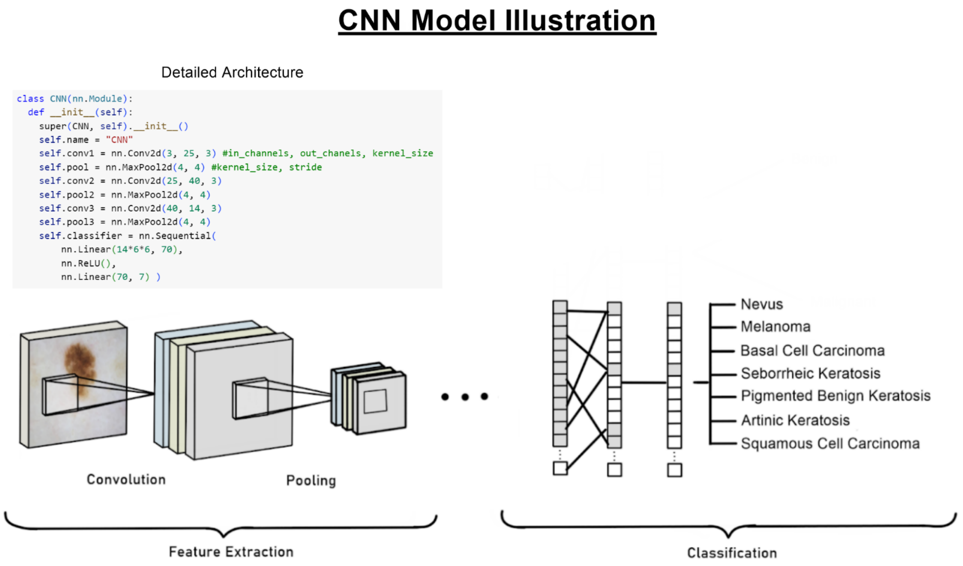
Figure 2: Baseline Model Results

6 Architecture

The team utilized three models for the classification of skin lesions. Two CNN models and one AutoEncoder model. Among the two CNN models, the team fully constructed one, and the other was a modified version of MobileNetV2 to fit the project parameters. All the models utilized weight decay, skip connections, dropout, and early stopping to mitigate overfitting and yield better results.

6.1 CNN

A graphic of the CNN model illustration is shown below, the specific architecture of the layers can be seen on the upright side of figure **??**. The architecture is composed of three convolution layers were used with three max pooling layers. A kernel size of 3*3 was used in the convolution layers. All pooling layers have the size of 4*4, which reduces dimensions more aggressively than 2*2 pooling layers, allowing quicker computations. Then, two fully connected linear layers with a ReLU activation function between the layers were used. The fully connected layer takes in The overall number of parameters is 220541.

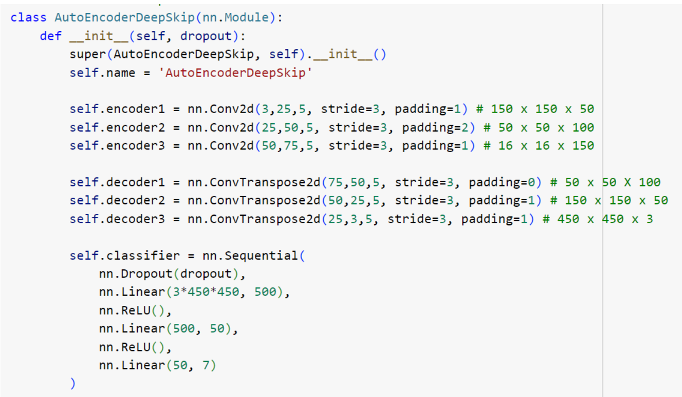


```
class CNN(nn.Module):
  def __init__(self):
    super(CNN, self).__init__()
    self.name = "CNN"
    self.conv1 = nn.Conv2d(3, 25, 3) #in_channels, out_chanels, kernel_size
    self.pool = nn.MaxPool2d(4, 4) #kernel_size, stride
    self.conv2 = nn.Conv2d(25, 40, 3)
    self.pool2 = nn.MaxPool2d(4, 4)
    self.conv3 = nn.Conv2d(40, 14, 3)
    self.pool3 = nn.MaxPool2d(4, 4)
    self.classifier = nn.Sequential(
        nn.Linear(14*6*6, 70),
        nn.ReLU(),
        nn.Linear(70, 7) )
```

Figure 3: CNN Architecture

6.2 AUTOENCODER

The AutoEncoder was also fully constructed by the team with the following architecture found in Figure **??**. It is composed of a three-layer convolutional encoder and three layers transpose convolutional decoder, with each layer used a kernel of size 5. It uses a ReLU activation function between layers, decreasing and increasing dimensionality using a stride of three in each layer. The decoder is then passed into a sigmoid function to bring outputs between 0 and 1 before being passed into a three-layer linear classifier. Overall the model consists of over 304 Million parameters but performed similarly on less than half the amount with smaller versions.



```
class AutoEncoderDeepSkip(nn.Module):
    def __init__(self, dropout):
        super(AutoEncoderDeepSkip, self).__init__()
        self.name = 'AutoEncoderDeepSkip'

        self.encoder1 = nn.Conv2d(3,25,5, stride=3, padding=1) # 150 x 150 x 50
        self.encoder2 = nn.Conv2d(25,50,5, stride=3, padding=2) # 50 x 50 x 100
        self.encoder3 = nn.Conv2d(50,75,5, stride=3, padding=1) # 16 x 16 x 150

        self.decoder1 = nn.ConvTranspose2d(75,50,5, stride=3, padding=0) # 50 x 50 X 100
        self.decoder2 = nn.ConvTranspose2d(50,25,5, stride=3, padding=1) # 150 x 150 x 50
        self.decoder3 = nn.ConvTranspose2d(25,3,5, stride=3, padding=1) # 450 x 450 x 3

        self.classifier = nn.Sequential(
            nn.Dropout(dropout),
            nn.Linear(3*450*450, 500),
            nn.ReLU(),
            nn.Linear(500, 50),
            nn.ReLU(),
            nn.Linear(50, 7)
        )
```

Figure 4: Autoencoder Architecture

6.3 MOBILENETV2

MobileNetV2 was also utilized and the pre-made architecture was kept. A dropout layer was added to the forward pass. MobileNetV2 was selected because it performs well in skin cancer diagnosis tasks and functions well on mobile devices, where this technology would likely end up being used Balaha & Hassan (2022).

7 RESULTS

The following section outlines the performance of each of the models prepared in addition to providing both quantitative and qualitative analysis.

7.1 QUANTITATIVE RESULTS

Overall, the model that yielded the best results was the MobileNetV2 model. This model achieved a final training accuracy of 71.64%, validation accuracy of 70.43%, and test accuracy of 40.21% with limited overfitting of the data, as shown in the accuracy graph in Figure53. Additionally, this model produced the best example of the 'ideal diagonal' on its confusion matrix, where most of the predictions made by the model fall on the line connecting the top left corner to the bottom right corner shown in Figure 5. This pattern signifies that the model is beginning to learn through either recall or precise predictions. Unfortunately, none of the models achieved good enough training and validation accuracies, and often times the recall and precision measurements were very close, therefore the recall and precision measurements were not seriously taken into account for model evaluation.

The CNN model achieved the best results out of the built-from-scratch models with a training accuracy of 66.63%, validation accuracy of 56.72%, and test accuracy of 36.71%. Notably, this model's training was stopped after 17.5 epochs due to extreme overfitting of data whereas the others were trained for approximately 400 epochs. Additionally, the confusion matrix in Figure 6 shows that the model could predict some of each of the diagnoses.

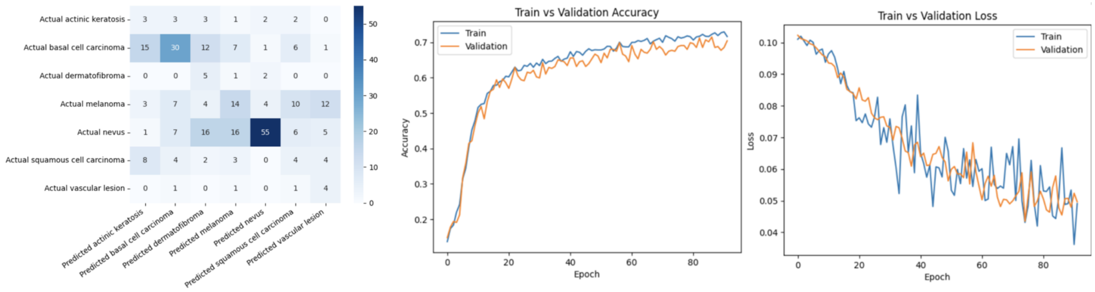


Figure 5: MobileNetV2 Results: Confusion Matrix, Accuracy Graph, Loss Graph

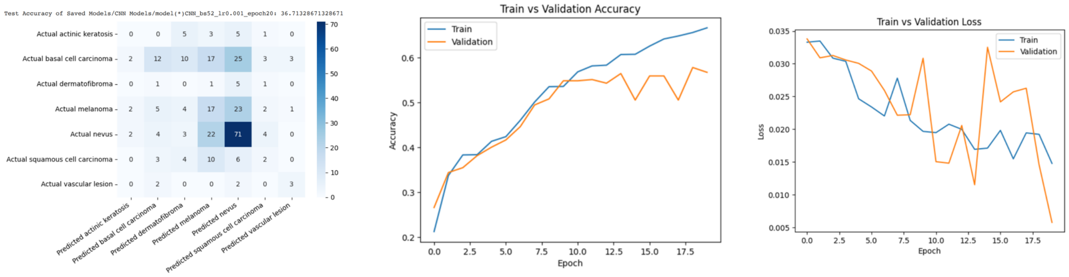


Figure 6: CNN Results: Confusion Matrix, Accuracy Graph, Loss Graph

The autoencoder model had a training accuracy of 42.57%, a validation accuracy of 40.59%, and a test accuracy of 27.62%. The disparity between the validation accuracy and testing accuracy suggests that the model relies on recall rather than precision since the testing data was from an entirely different dataset from a separate source. This model struggled with overfitting, as shown in the accuracy graph in Figure 7. However, the place where the trendline would be placed can clearly be seen, meaning that this model has potential but the hurdles of higher accuracies and less overfitting would have to be jumped.

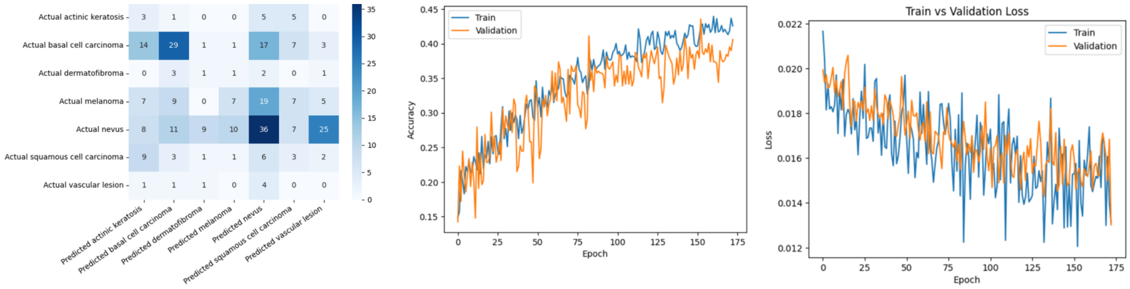


Figure 7: Autoencoder Results: Confusion Matrix, Accuracy Graph, Loss Graph

Across all three models, the loss curves seemed to sporadically decrease despite the optimization of the learning rate as much as possible. This trend also suggests that all the models were still overfitting the data even after much tuning.

7.2 Qualitative Results

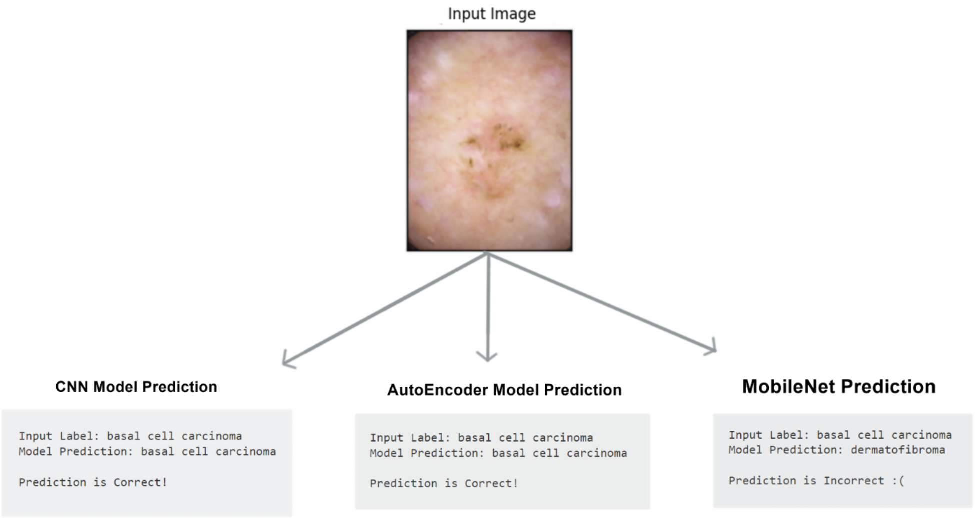
The models did not perform as well as hoped. However, decent training and validation results were able to be produced by the MobileNetV2 and CNN models meaning that the models perform well when only using the dataset they were trained on. As discussed in the previous subsection, all of the models' accuracies drop considerably between the validation and testing phases partly because the testing set is comprised of data that the models haven't seen before. But mostly because the training and validation set are split from the HAM10000 dataset and the testing set is from the HIBA dataset. So two completely different datasets were used instead of all sets being split from the same original dataset, making the testing set more difficult for the models to classify.

Generally, the most difficult diagnosis for all of the models to predict were: actinic keratosis, dermatofibrosis, squamous cell carcinoma, and vascular lesion. There was no clear trend between models to suggest why they all struggled with these diagnoses. However, the MobileNet and Autoencoder matrices were similar in the fact that both models tended to predict the same incorrect label over and over again but the labels that they predicted were different between models.

7.3 Evaluation of the Models

Evaluation of the models happened in two stages. First, the dataset was split into three sets, training, validation, and testing. The testing set isolated images into a CSV file which were not used for training or validation, making them unique and appropriate for testing. All the models were then tested for accuracy using this test set.

To further evaluate each model, the team produced functions that would take, never seen before by the model, images and tested the models. A collection, PROVe-ai from the HAM10000 and placed in a separate folder. The images from this collection were never used for any of the previous training and testing of the model. Those images were then taken and used to evaluate the model. The figure below illustrates how the three models predict a new unique image.



```
Input Label: basal cell carcinoma
Model Prediction: basal cell carcinoma

Prediction is Correct!
```

```
Input Label: basal cell carcinoma
Model Prediction: basal cell carcinoma

Prediction is Correct!
```

```
Input Label: basal cell carcinoma
Model Prediction: dermatofibroma

Prediction is Incorrect :(
```

Figure 8: Prediction Example of Each Model

Furthermore, each model was then tested on the whole collection to determine the testing accuracy on new input data. The testing accuracy of each model and matrices were obtained and compared to the testing accuracy from the original HAM10000 collection used for the training. The following testing accuracies were obtained: 32.17% for the CNN model, 15.94% for the AutoEncoder model, and 26.80% for the MobileNet model.

8 Discussion

The project's difficulty stems from the complex nature of classifying skin lesions from images. Unlike conventional image classification, differentiating between various skin cancer and disorders requires a deep understanding of dermatological patterns, size, and structural variation. Despite scoping down to classifying 7 skin conditions, and ensuring an adequate amount of images for each class, different skin condition still presents unique visual characteristics that can be subtle. The similarities between certain skin conditions can lead to misdiagnoses even for experienced dermatologists. Expecting a model to be able to accurately predict even a few conditions provided the amount of time and processing resources the team had access to is a tall order.

The best results achieved were from the CNN model when looking at both testing methods. However, all models did not perform well on new data as the maximum achieved testing accuracy when evaluating the model was 32.17%. MobileNetV2, the pre-trained model, resulted in a more balanced accuracy result and was able to mitigate overfitting the most. If the team were to continue to work on this project, the next steps would be to provide each of the models with the metadata as an addition to the images and experiment more with other pre-trained models to see if another one is better suited for the task.

This project provided a great deal of insight into the difficulty of training a deep learning model to be able to predict complex outcomes. The most important lesson learned was that building off of and fine-tuning existing pretrained models is much easier and more effective than writing our own feature extraction models. While this did help with the overall understanding of the project, it would have saved about a week and a half of almost constant training and tuning that yielded little results to just start with a few pre-trained models and tune and compare from there. Additionally, the team learned the importance of having a balanced dataset when the first couple of models that were trained had a 70% testing accuracy but only because 70% of the data was comprised of one diagnosis.

9 Link to Code

The following is a shared link to the Google Drive Folder where all the code and images the group used to train and test the model are located.

Link: https://drive.google.com/drive/folders/1EtVoysTu9pil246Q2g5Gxc0Dx348Qfhh?usp=sharing

References

Prove-ai. URL https://api.isic-archive.com/collections/218/.

A whole new way to help identify your skin conditions. 2021. URL https://health.google/consumers/dermassist/.

Hossam Magdy Balaha and Asmaa El-Sayed Hassan. Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm. *Neural Computing and Applications*, 35(1):815–853, 2022. doi: 10.1007/s00521-022-07762-9.

betterhealth. Melanoma, 2023. URL https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/melanoma.

ISIC Collections. Isic collections. URL https://api.isic-archive.com/collections/.

Hospital Italiano de Buenos Aires. Hiba. URL https://api.isic-archive.com/collections/175/.

Terry Graedon. How aysa, an ai app, helped a mother with her child's rash. 2018. URL https://www.peoplespharmacy.com/articles/how-aysa-an-ai-app-helped-a-mother-with-her-childs-rash.

SkinVision Inc. Getting started, 2021. URL https://www.skinvision.com/getting-started/#explore_skinvision.

ISIC. Isic archive. URL https://gallery.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D.

Carolina Magalhaes, João Manuel R.S. Tavares, Joaquim Mendes, and Ricardo Vardasca. Comparison of machine learning strategies for infrared thermography of skin cancer. *Biomedical Signal Processing and Control*, 69:102872, 2021. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2021.102872. URL `https://www.sciencedirect.com/science/article/pii/S1746809421004699`.

T. Maier, D. Kulichova, K. Schotten, R. Astrid, T. Ruzicka, C. Berking, and A. Udrea. Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. *Journal of the European Academy of Dermatology and Venereology*, 29(4):663–667, 2015. doi: https://doi.org/10.1111/jdv.12648. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jdv.12648`.

Department of Dermatology Medical University of Vienna. Ham10000. URL `https://api.isic-archive.com/collections/212/`.

Andre G.C. Pacheco and Renato A. Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.103545. URL `https://www.sciencedirect.com/science/article/pii/S0010482519304019`.

Yuan Liu Peggy Bui. Using ai to help find answers to common skin conditions. 2021. URL `https://blog.google/technology/health/ai-dermatology-preview-io-2021/`.