INDIANA UNIVERSITY
BLOOMINGTON

ENGR-E516 Engineering Cloud Computing

MIDTERM PROJECT REPORT

SKIN LESION CLASSIFICATION FOR SKIN CANCER

Sakshi Rathi

Sreshta Reddy N

Shruti Houji

October 30th, 2023

1. **Project Title: "Skin Lesion Classification For Skin Cancer"**

Skin lesions, can be life-threatening if not detected early. The goal of this project is to utilize machine learning and image processing techniques to analyze dermoscopic images of skin lesions and classify them as benign or malignant. By deploying this solution on AWS, we aim to provide a scalable and efficient system that can assist dermatologists in early diagnosis and potentially save lives. AWS, as a leading cloud platform, provides several advantages over traditional hosting methods, including scalability, affordability, and adaptability. By harnessing AWS's robust suite of services, we envision a system that not only efficiently classifies uploaded skin images but also scales dynamically based on demand, ensuring uninterrupted, swift, and reliable service for end-users.

The project's life cycle will encompass several phases, including data acquisition, preprocessing, feature extraction, model training, and hyperparameter optimization. AWS's cloud computing capabilities will be instrumental at each stage, from storing large datasets in S3 to employing powerful EC2 instances for training complex neural networks. Once the model reaches satisfactory accuracy, it will be deployed as a web service using AWS Lambda and API Gateway. Continuous monitoring, facilitated by AWS CloudWatch, will be set in place to oversee the system's performance, promptly identify potential anomalies, and ensure optimal service uptime and reliability.

2. **Project Goal:**

a) Model Development: Develop a machine learning model aimed at classifying skin lesions to detect potential skin cancers.

b) Data Management: Gather relevant skin lesion data, preprocess it for optimal model performance, carry out feature engineering, and select the best model for training. The process will also involve tuning hyperparameters to enhance the model's predictive accuracy.

c) User Interface: Design and implement a user-friendly website interface where users can effortlessly input skin images and promptly receive a prediction regarding the possibility of it being cancerous.

d) Deployment: Host the developed machine learning model on a cloud platform, specifically AWS, to benefit from its scalability, cost-effectiveness, and flexible infrastructure.

e) Monitoring & Logging: Integrate cloud-based monitoring and logging services. This will help in continuously tracking the model's performance, user interactions, and identifying any potential issues or anomalies that might arise during its operation.

f) Documentation: Compile comprehensive documentation detailing the project's architecture, model details, operational instructions, and other relevant aspects. This will ensure that future maintenance, updates, or modifications can be seamlessly undertaken by other developers or team members.

3. Related work and gap analysis:

We have done research on Skin Lesion Images for Skin Cancer classification based on the following papers:

1. "Deep Semantic Segmentation and Multi-Class Skin Lesion Classification Based on Convolutional Neural Network" by Muhammad Almas Anjum, Javaria Amin, Muhammad Sharif , (Senior Member, Ieee), Habib Ullah Khan, (Member, Ieee), Muhammad Sheraz Arshad Malik, And Seifedine Kadry et al. (2020) focused on localization, segmentation, and classification of the skin lesion in early stages (3 phases) which is evaluated on the top MICCAI ISIC challenging 2017, 2018 and 2019 datasets based on CNN model. Phase I utilizes a tinyYOLOv2 model with ONNX and squeeze Net as the backbone, effectively localizing various types of skin lesions. Phase II employs a 13-layer 3D-semantic segmentation model to segment the lesions, with a pixel classification layer ensuring accuracy. Finally, in Phase III, deep features extracted using a ResNet-18 model are optimized using the ant colony optimization (ACO) method and passed to classifiers like optimized (O)-SVM and O-NB.
2. "An approach for multiclass skin lesion classification based on ensemble learning" by Zillur Rahman, Md. Sabir Hossain, Md. Rabiul Islam, Md. Mynul Hasan, Rubaiyat Alim Hridhee et al. (2021) proposed a weighted ensemble model using 5 deep neural networks like ResNeXt, SeResNeXt, ResNet, Xception, and DenseNet based on the dataset ISIC 2019 and 18730 dermoscopy images from official Human Against Machine using grid search method with recall score of 94%.
3. "Skin Lesions Classification into Eight Classes" for ISIC 2019 Using Deep Convolutional Neural Network and Transfer learning" by Mohamed A. Kassem, Khalid M. Hosny, and Mohamed M. Fouad et al. (2019) presented pre trained model with GoogleNet and transfer learning to classify into 8 types and one class as unknown images which achieved an accuracy of 94.92%.
4. "WonDerM: Skin Lesion Classification with Fine-tuned Neural Networks" by Yeong Chan Lee, Sang-Hyuk Jung, and Hong-Hee Won et al. (2018) described a neural network fine-tuned with segmentation task data and classified into 7 different types and achieved an accuracy of 0.899 and 0.785 in the validation and test sets respectively.
5. "Skin Lesion Classification Using Hybrid Deep Neural Networks" by Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Rupert Ecker, Isabella Ellinger et al. (2019) stated that CNN is the best among all the classical methods, which uses optimized deep features from a number of well established CNN's along with pre trained deep models are used with 83.83% for melanoma classification and of 97.55% for seborrheic keratosis classification.

From all the above papers, we can conclude that the ensemble model excels in recall, transfer learning in accuracy, and CNNs in overall effectiveness. However, there remains a potential gap in addressing the interpretability and explain ability of these models, which is crucial for gaining trust among medical professionals and ensuring the models' clinical relevance. So far, we have implemented a deep neural network model, and, in the future, we plan to implement a CNN model. Hence, scalability and ease of deployment are factors that need further consideration in real-world applications.

4. Proposed tasks

The project aims to develop a robust skin lesion classification system that has 9 distinguishing categories. The model will use deep learning/machine learning techniques for classification and be deployed on Amazon Web Services (AWS) to enable real-time classification and accessibility.

The image classification part of the project would follow this:

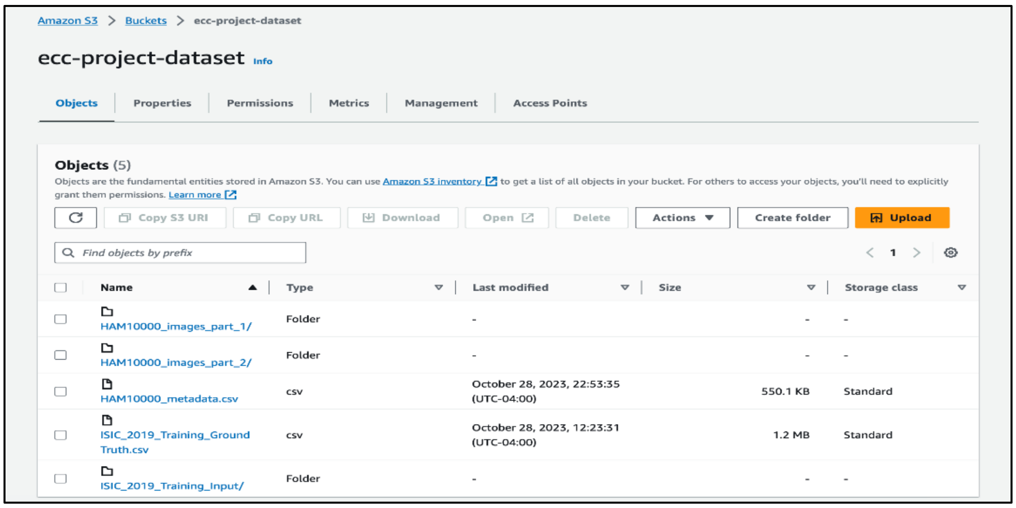
1. Data pre-processing: Preprocess the dataset by resizing, normalizing, and augmenting the images to ensure uniformity and increase model robustness.
2. Model Development and Evaluation : A suitable deep learning architecture for image classification will be chosen and a model will be trained using the preprocessed dataset. We can also use the SageMaker's GPU-powered instances for efficiency. The optimization of the model will be done by hyperparameters tuning through automatic or manual tuning. The trained model will be then deployed to a web server for monitoring and inference. The model's performance will be assessed on the trained model's performance on a validation dataset.

The AWS aspect involves leveraging Amazon Web Services for various stages of the project.

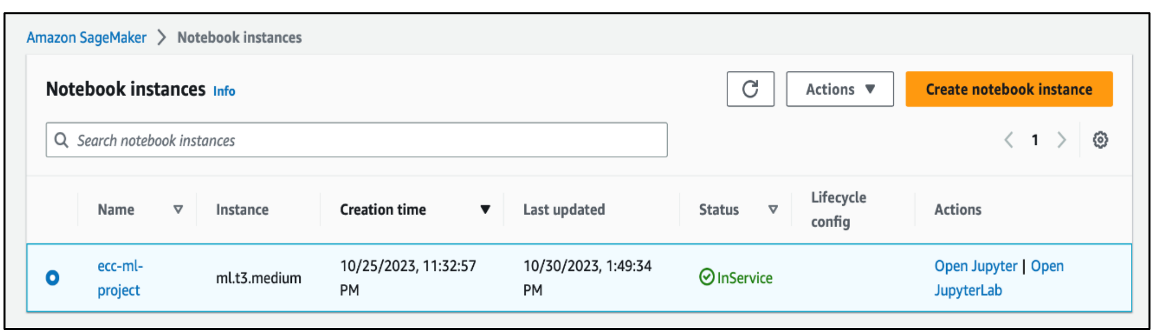
- Amazon S3: It'll serve as a primary data storage solution for the project. This dataset stored will include both the training and validation images. The model checkpoints and logs are also stored in S3 for easy access and versioning.
- AWS Lambda: The model can be deployed as a Lambda function. It can also be used for predicting the image using the model and returns the classification results in real-time. For the web application, the various backend tasks will be handled by lambda.
- Amazon SageMaker: It'll be used to train the machine learning/deep learning model on GPU-powered instances, enabling faster training times. The built-in algorithms or custom scripts can be used for model development. The automatic model tuning can be used for hyperparameter optimization. In addition, sagemaker will be used for deployment of the trained model and creating a scalable and secure API for real-time inference.
- AWS QuickSight: For data visualization and monitoring purposes, to analyze the performance of the models and to create an interactive dashboard. This will be useful for the user to monitor and to keep track of any anomalies.

5. Progress on Proposed Tasks

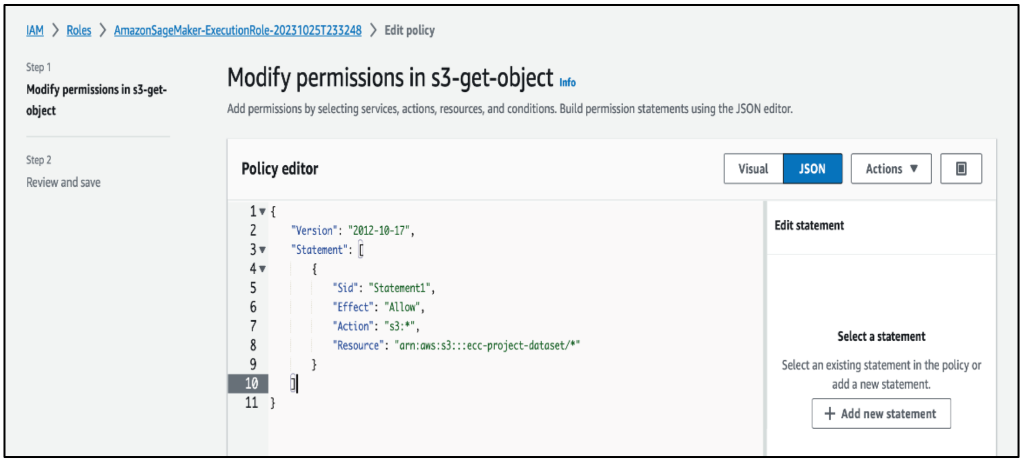
- Getting started with AWS: Setting up an AWS account. Initializing an s3 bucket "ecc-project" where we can upload our data and all the required files.
- Data Upload: This csv file named: HAM10000_metadata.csv is uploaded in the s3 bucket name ecc-project-dataset which contains features like age, sex, lesion_id and image id. Input folder of images of part1 and part2 containing images of the skin.
- Dataset link :(https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/code).



- AWS Sagemaker: To get started with AWS Sagemaker, we need to create a notebook instance that we can use to process our data and model it. We create an instance of t3.medium, (since it is free and easily available). However, for working with deep learning models on our data we need to create an instance of p3.large or p4.large, which needs to requested to AWS via the support system, we are waiting for their response.



- AWS Policy Sagemaker: To access data from the s3 bucket, we have to set permissions for the user to allow sagemaker to use the s3 objects. We have updated the policy which allows to use any action (s3:*) on all objects within the ecc-project-dataset S3 bucket:



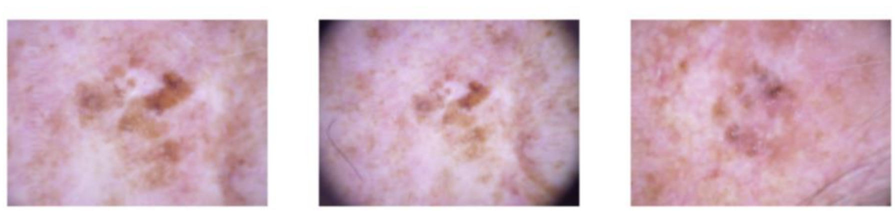
- Data Pre-processing: We first pre-process and perform data analysis on our meta data: HAM10000_metadata.csv containing the features listed below.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear

Now, we map the image id from the dataframe to the images in the image folder part1 and part2, we create a new path column containing the full path of s3//s3-bucket//foldername//image-name.

	lesion_id	image_id	dx	dx_type	age	sex	localization	path
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	s3://ecc-project-dataset/HAM10000_images_part_1/ISIC_0027419.jpg
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	s3://ecc-project-dataset/HAM10000_images_part_1/ISIC_0025030.jpg
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	s3://ecc-project-dataset/HAM10000_images_part_1/ISIC_0026769.jpg
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	s3://ecc-project-dataset/HAM10000_images_part_1/ISIC_0025661.jpg
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	s3://ecc-project-dataset/HAM10000_images_part_2/ISIC_0031633.jpg

Let's display a few images from our data and see how our data actually looks like.

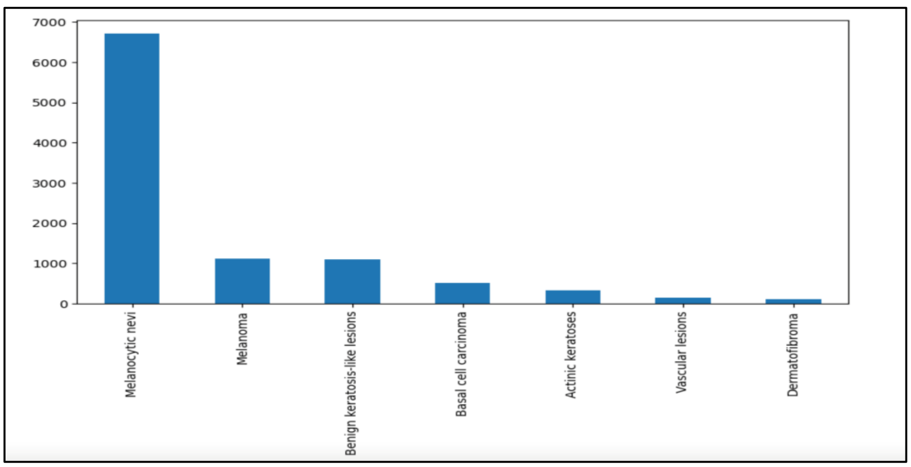


Creating a dictionary of the labels: Since we have abbreviated labels in our data, we map them to the full form of the skin cancer names.

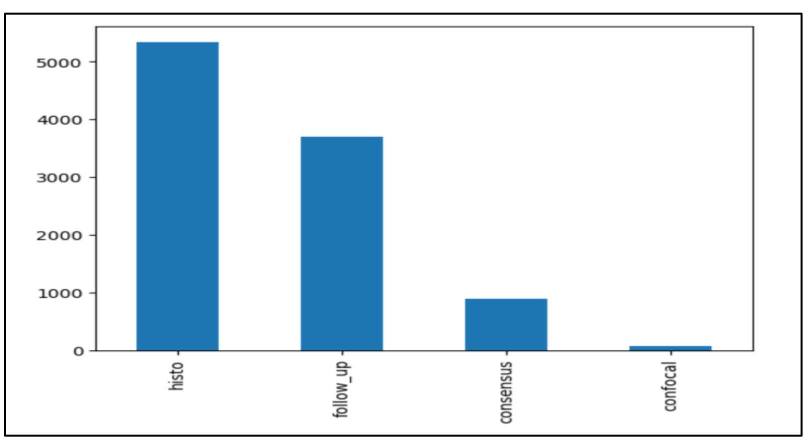
```
lesion_type_dict = {
    'nv': 'Melanocytic nevi',
    'mel': 'Melanoma',
    'bkl': 'Benign keratosis-like lesions',
    'bcc': 'Basal cell carcinoma',
    'akiec': 'Actinic keratoses',
    'vasc': 'Vascular lesions',
    'df': 'Dermatofibroma'
}
```

- Exploratory Data Analysis:

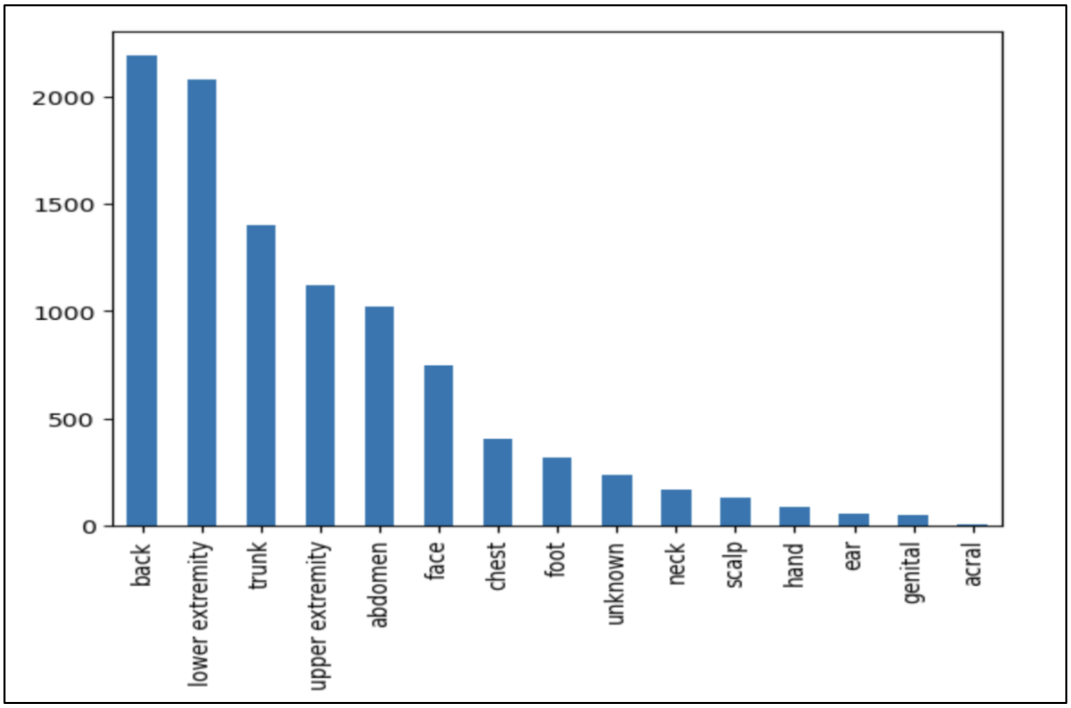
a) Distribution of the 7 classes of cancer across the data. As we can see our dataset is imbalanced with Melanocytic class have the maximum data. And the actinic, vascular and dermatofibroma have the least amount of data. So, while modeling our data, we'll have to make sure our data is balanced enough for us to predict new data after that.



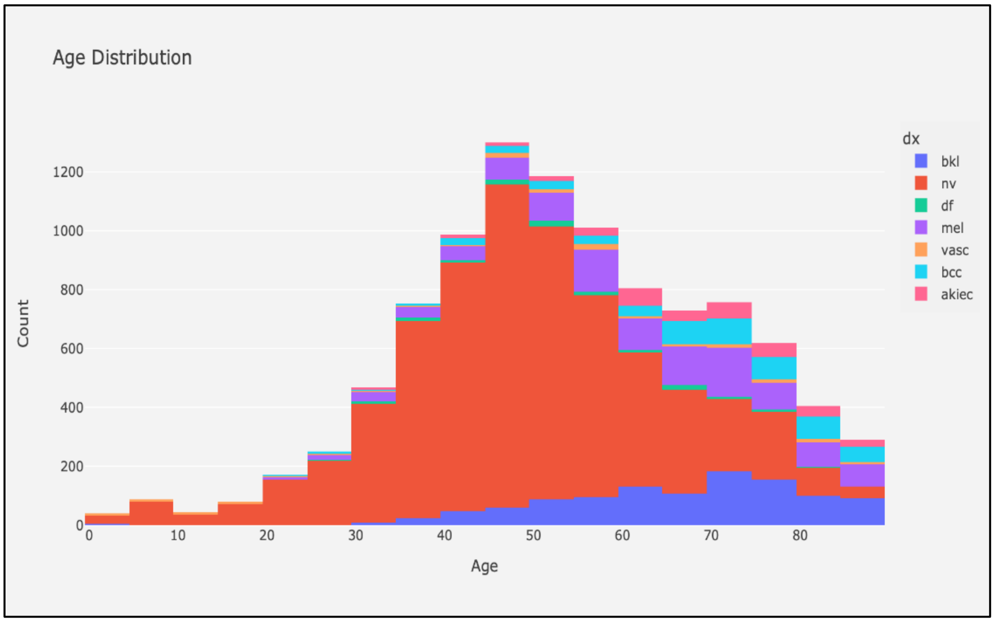
b) The data also has 4 different types of lesions like Histopathology, Confocal, follow up, and consensus. We can the distribution of the lesion types in our data, with histopathology having the most distribution.



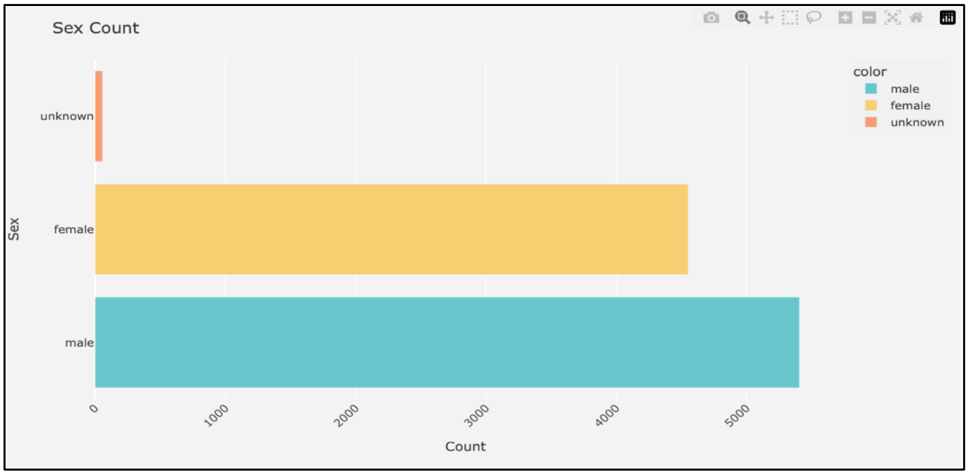
c) Other location of the lesions like back, abdomen, face is visualized.



d) Histogram for age frequency: This plot shows the total count for the range of 5 ages starting from the age 0 to 90 for all the different types of lesions.



e) Bar plot for sex count: This plot shows the total count of two sex: female and male along with the unknown count if that person is not in both the categories.



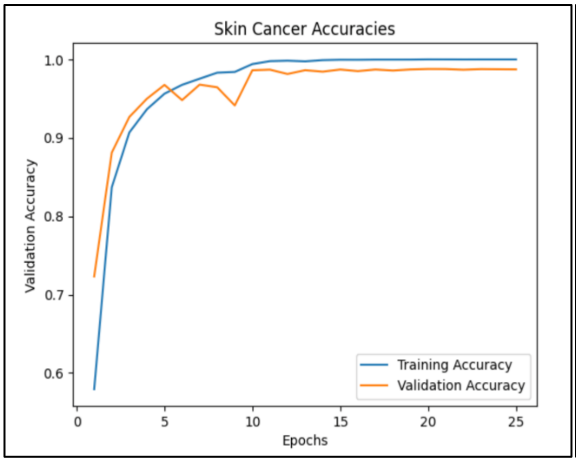
- Data Modeling: For modeling our data, we have different features in our data. We mainly use the image data, the pixel intensities to create a model that when given an input of an image, can predict the kind of skin cancer the skin type looks like. So, for this we have to use a machine learning algorithm that works on image data type. For this we use a deep neural network with several layers of convolution2D, MaxPooling, BatchNormalisation, and a softmax layer with 7 classes.

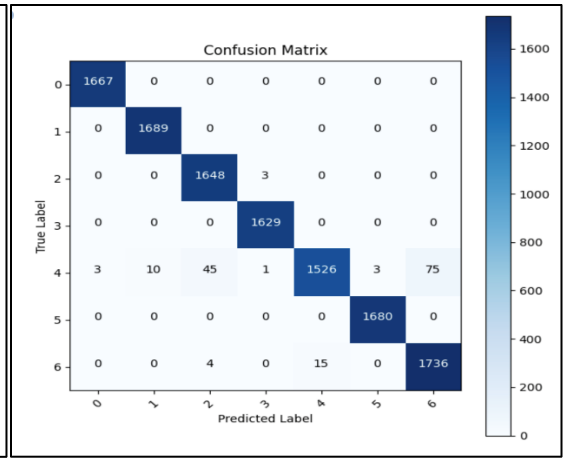
6. Preliminary Observations

We trained a model for 25 epochs with reduce learning rate on plateau custom learning rate function.

```
Epoch 20/25
276/276 [==============================] - 4s 13ms/step - loss: 0.0012 - accuracy: 0.9999 - val_loss: 0.0550 - val_accuracy: 0.9877 - lr: 6.2500e-05
Epoch 21/25
276/276 [==============================] - 4s 14ms/step - loss: 0.0013 - accuracy: 0.9999 - val_loss: 0.0566 - val_accuracy: 0.9876 - lr: 6.2500e-05
Epoch 22/25
275/276 [============================>.] - ETA: 0s - loss: 0.0011 - accuracy: 0.9999
Epoch 22: ReduceLROnPlateau reducing learning rate to 3.125000148429535e-05.
276/276 [==============================] - 4s 13ms/step - loss: 0.0012 - accuracy: 0.9999 - val_loss: 0.0598 - val_accuracy: 0.9868 - lr: 6.2500e-05
Epoch 23/25
276/276 [==============================] - 4s 13ms/step - loss: 9.4523e-04 - accuracy: 0.9999 - val_loss: 0.0578 - val_accuracy: 0.9876 - lr: 3.1250e-05
Epoch 24/25
276/276 [==============================] - ETA: 0s - loss: 0.0012 - accuracy: 0.9999
Epoch 24: ReduceLROnPlateau reducing learning rate to 1.5625000742147677e-05.
276/276 [==============================] - 4s 13ms/step - loss: 0.0012 - accuracy: 0.9999 - val_loss: 0.0570 - val_accuracy: 0.9874 - lr: 3.1250e-05
Epoch 25/25
276/276 [==============================] - 3s 13ms/step - loss: 0.0010 - accuracy: 0.9999 - val_loss: 0.0585 - val_accuracy: 0.9872 - lr: 1.5625e-05
```

We observe the train and validation accuracies vary starting from 0.7 to roughly 0.98. Also, a confusion matrix representing the true and predicted labels matrix for the test data.





References:

1. https://www.sciencedirect.com/science/article/pii/S2352914821001465
2. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9139937
3. AWS SageMaker Inbuilt Model Tuning
4. AWS API Gateway Integration with AWS Lambda and AWS SageMaker
5. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/code