In the name of Allah

Course : Data Science

Final project

Abolfazl Golgolnia , SID : 98222082
Aryan Neizehbaz , SID : 400222112
Alireza Mottaghi , SID : 400222089

Introduction :

Skin cancer is a type of cancer that affects the skin cells. There are three main types of skin cancer: basal cell carcinoma, squamous cell carcinoma, and melanoma. Basal cell carcinoma and squamous cell carcinoma are the most common types and are usually not life-threatening, but melanoma can be aggressive and spread to other parts of the body.

Risk factors for skin cancer include excessive sun exposure, sunburns, fair skin, family history, and a weakened immune system. Symptoms can include a new or changing mole, a sore that doesn't heal, or a bump or spot that's growing.

Diagnosis is typically made through a skin biopsy, in which a sample of skin is removed and examined under a microscope. Treatment options depend on the type and stage of the cancer and may include surgery, radiation therapy, or chemotherapy. Prevention measures include avoiding sun exposure, wearing protective clothing and using a broad-spectrum sunscreen with an SPF of at least 30. Early detection is crucial for successful treatment, so it's important to have regular skin checks and be aware of any changes to your skin.

Dataset Overview :

HAM10000 ("Human Against Machine with 10000 training images") dataset - a large collection of multi-source dermatoscopic images of pigmented lesions

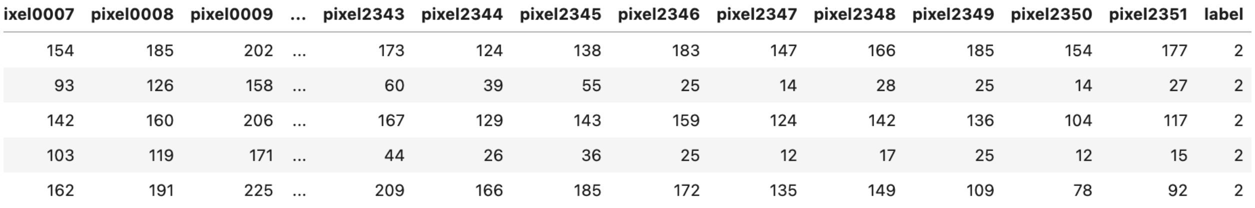
The dermatoscopic images are collected from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images.
It has 7 different classes of skin cancer which are listed below :

- Melanocytic nevi
- Melanoma
- Benign keratosis-like lesions
- Basal cell carcinoma
- Actinic keratoses
- Vascular lesions
- Dermatofibroma

EDA (Exploratory Data Analysis) :

At first, we go to two data sets and try to display general information from them.

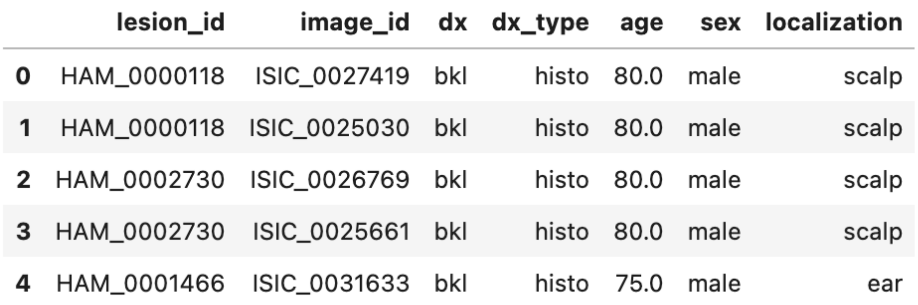
- **Importing Datasets :**
  - **hmnist_28_28_RGB :**
    This dataset contains the pixels of the images that the models are trained on.



ixel0007	pixel0008	pixel0009	...	pixel2343	pixel2344	pixel2345	pixel2346	pixel2347	pixel2348	pixel2349	pixel2350	pixel2351	label
154	185	202	...	173	124	138	183	147	166	185	154	177	2
93	126	158	...	60	39	55	25	14	28	25	14	27	2
142	160	206	...	167	129	143	159	124	142	136	104	117	2
103	119	171	...	44	26	36	25	12	17	25	12	15	2
162	191	225	...	209	166	185	172	135	149	109	78	92	2

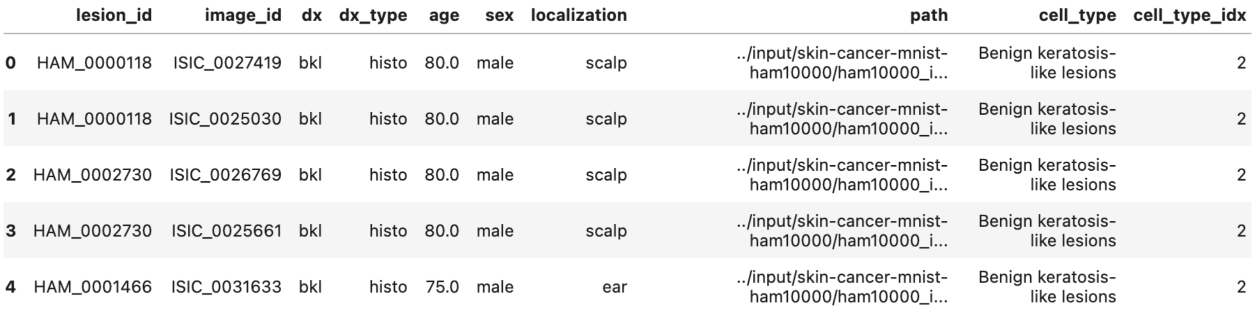
  - **HAM10000_metadata :**
    In the metadata, lesion_id and image_id are included to link the tables. The dataset contains lesions with multiple images that can be tracked by the lesion_id column in the HAM10000_metadata file.
    The dx and dx_type columns are placed for labeling and how to identify, and their values are as follows:
    - **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)**

More than 50% of lesions are confirmed through histopathology (histo), the ground truth for the rest of the cases is either follow-up examination (follow_up), expert consensus (consensus), or confirmation by in-vivo confocal microscopy (confocal).



	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

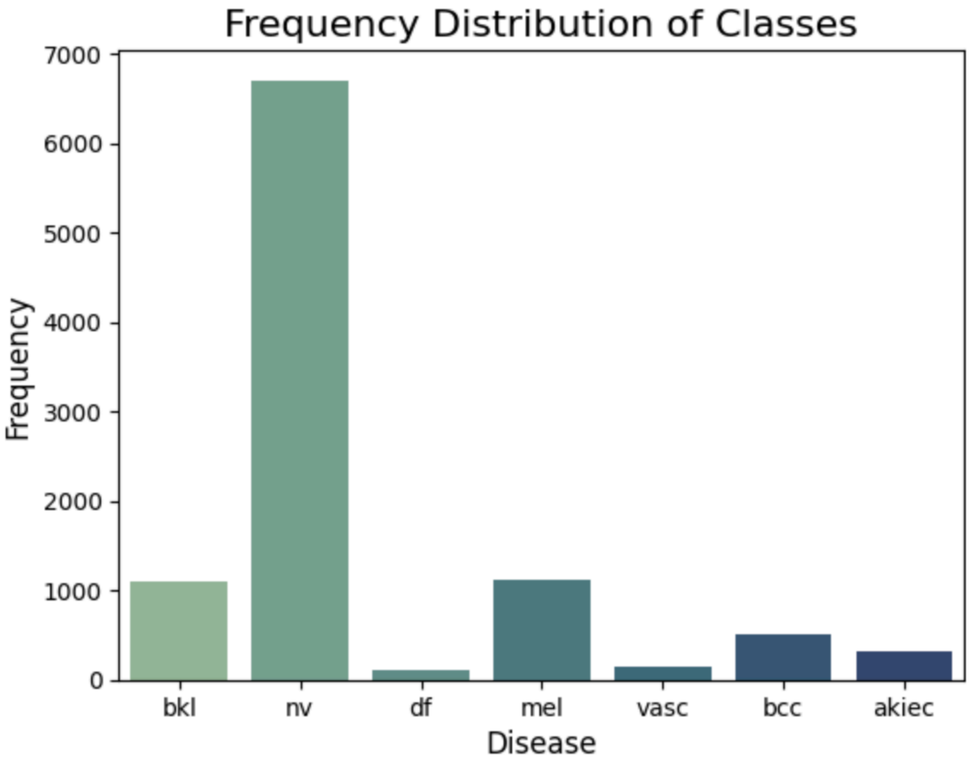
Next, we Create dictionary for displaying more human-friendly labels.
After that,, we merge images from both folders into one dictionary. Then, creating new columns for better understanding of features.



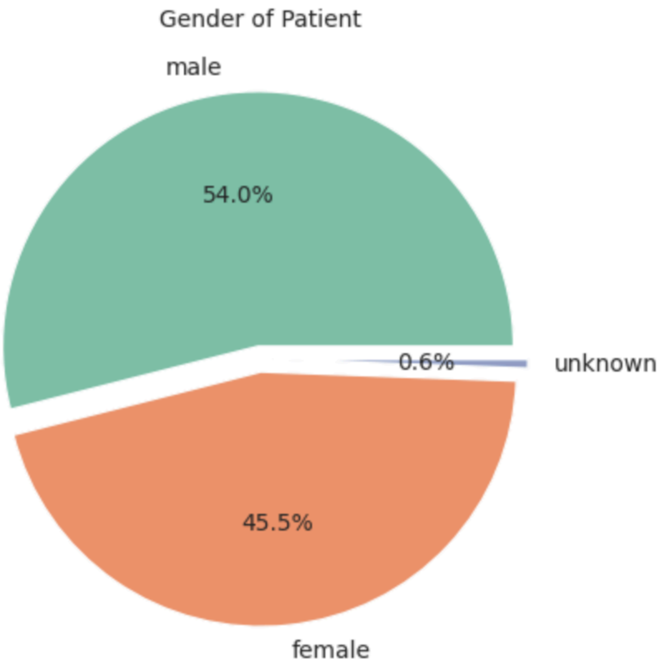
	lesion_id	image_id	dx	dx_type	age	sex	localization	path	cell_type	cell_type_idx
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	../input/skin-cancer-mnist-ham10000/ham10000_i...	Benign keratosis-like lesions	2

Next, we will add a series of columns to the metadata and based on them we will make a visualization.

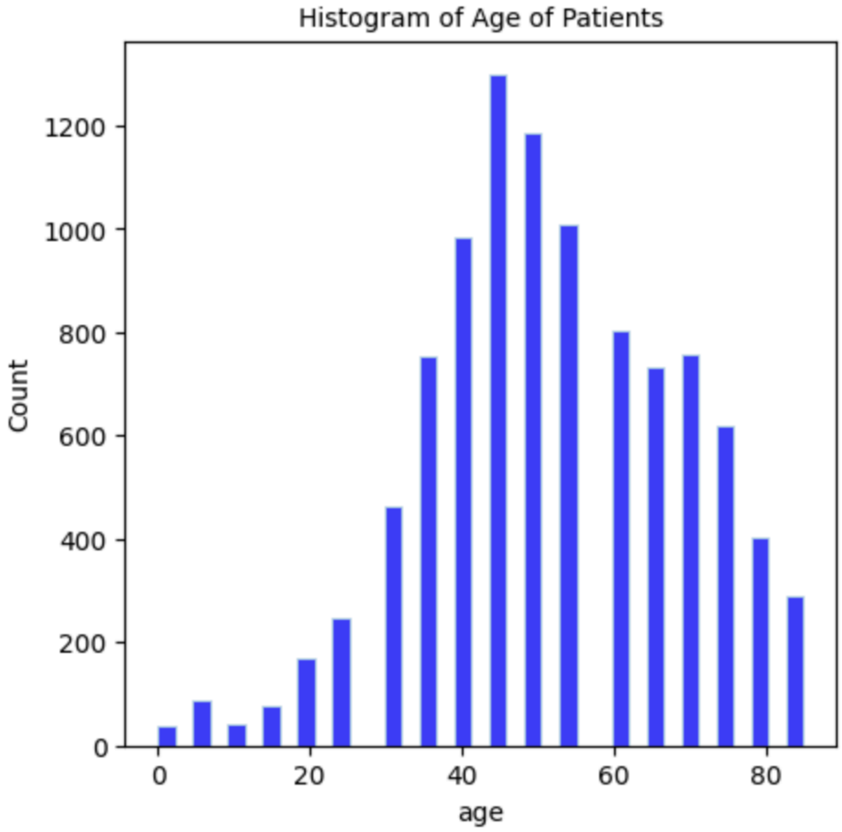
- **Visualization :**
    - Frequency Distribution of Classes :
        - There are vast number of cases of Melanocytic nevi as compared to others.
        - Melanoma and Benign keratosis-like lesions are quite less wide spread as compared to Melanocytic nevi.
        - Other cell type viruses subsequently affected less in numbers.



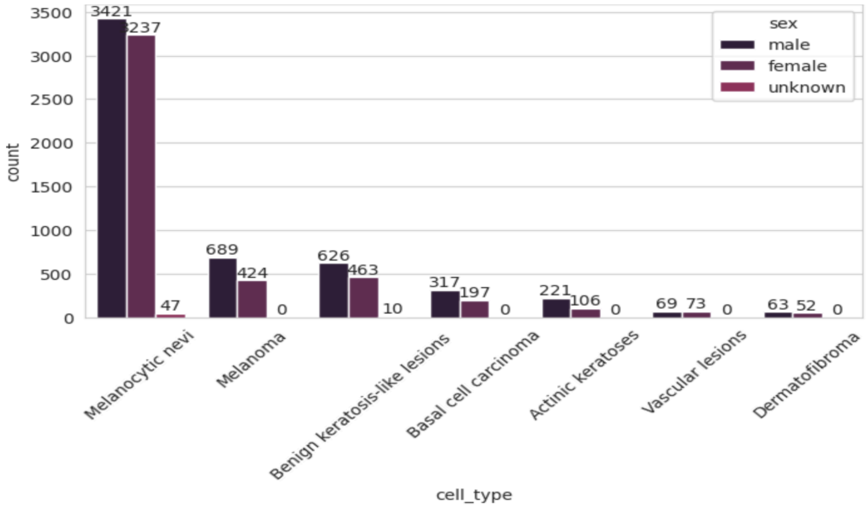
- Distribution of Disease over Gender :
    - It seems majority of the male being affected from skin cancer symptoms.
    - Though we can say there is not apar difference when considering being affected genderwise.



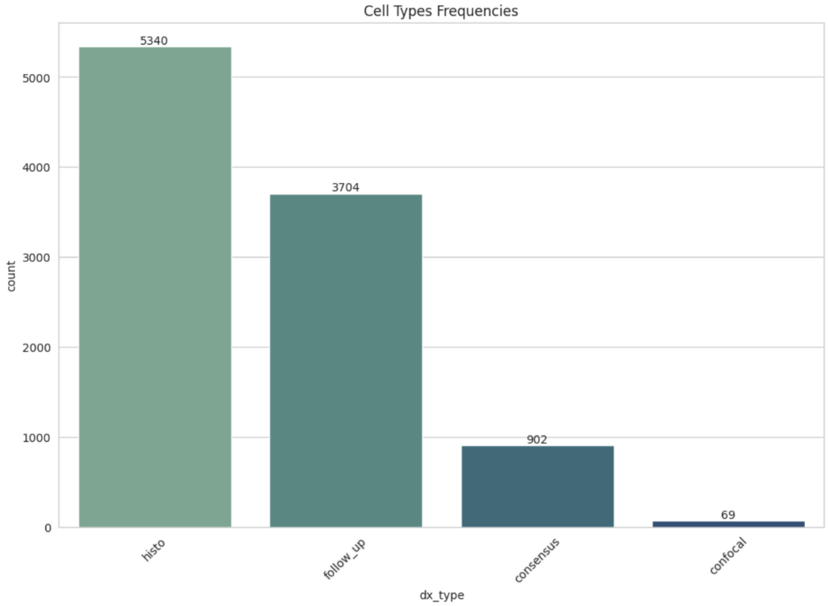
- Histogram of Age of Patients :
    - It seems most of the affected people are around mid 40s.



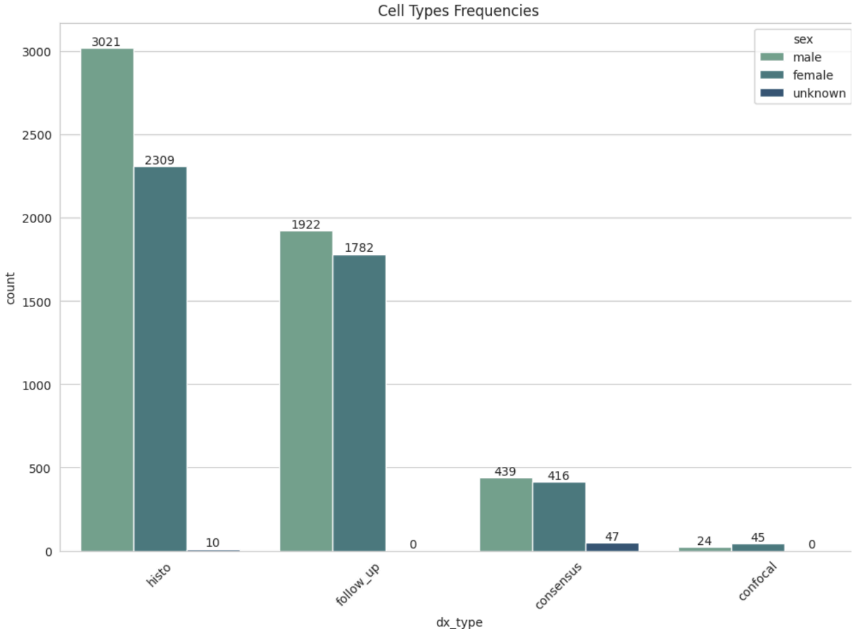
- Gender vs Cell type :
    - It seems for every skin type cancer, majority of males are victims.



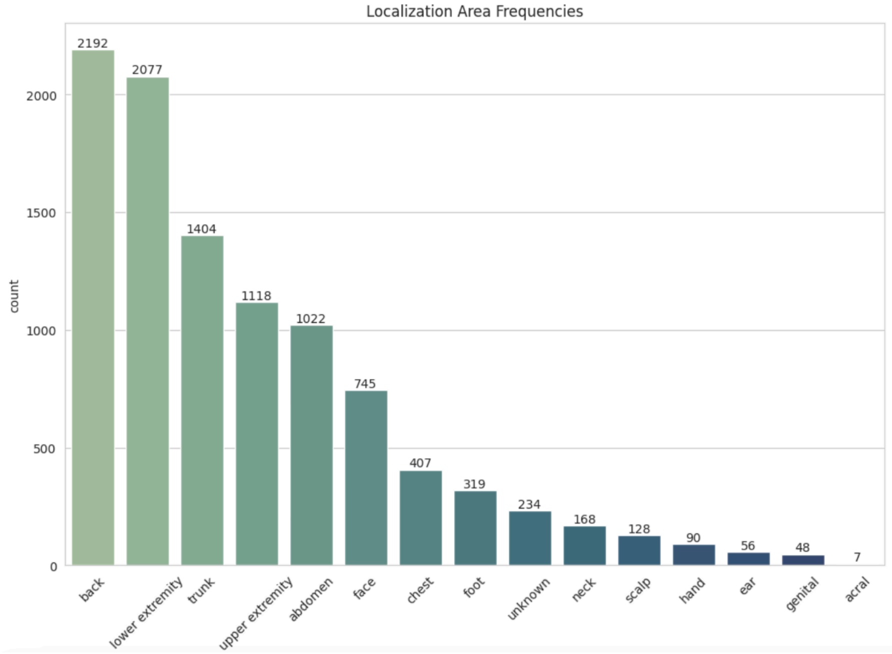
- Cell type :



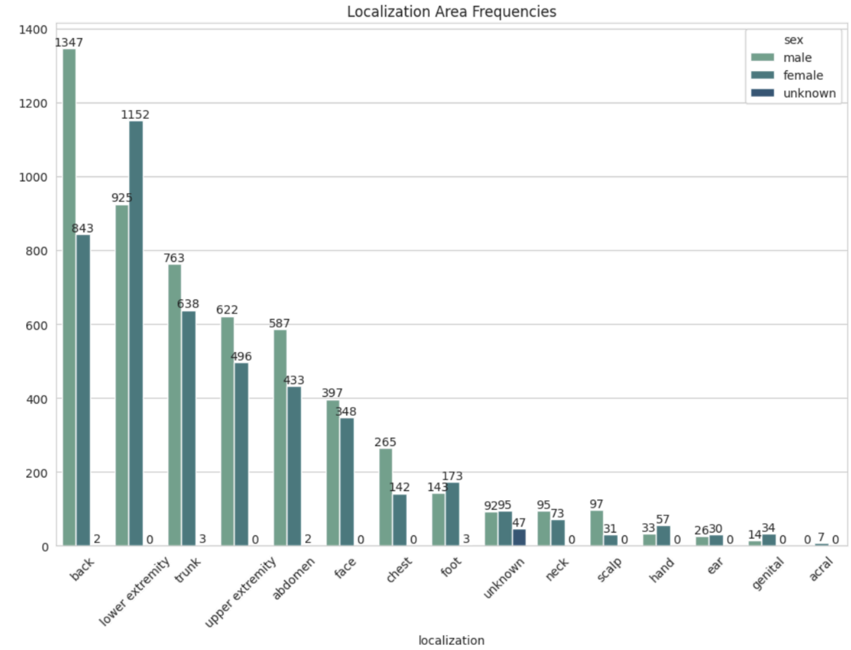
- Gender wise Cell Type Distribution :



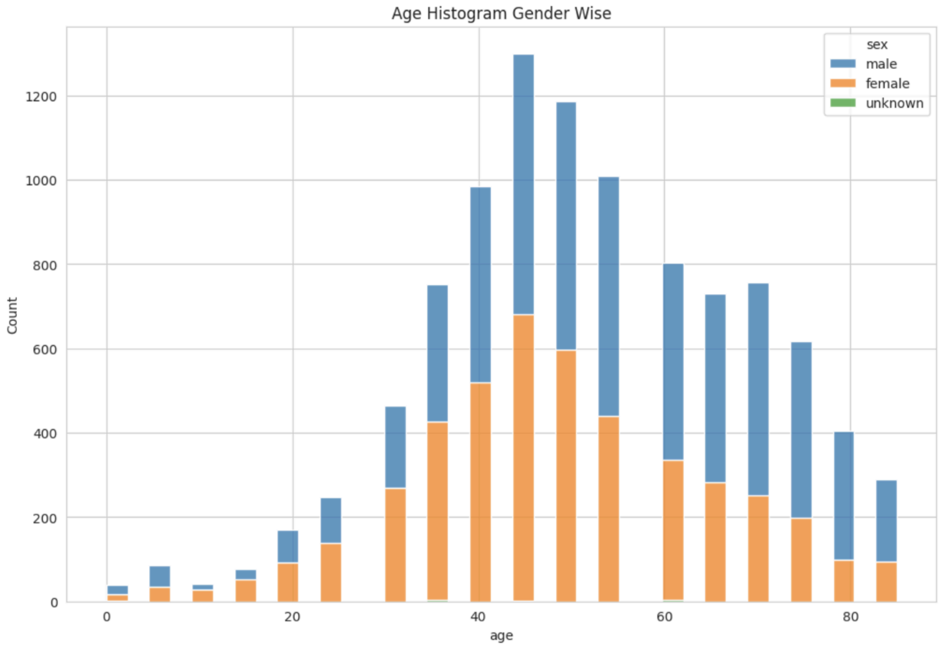
- Localization area :
    - It seems most of the area affected is related to particularly back, lower extremity or trunk etc.
    - The significance we take out of it as the areas where the part gets sweaty easily.



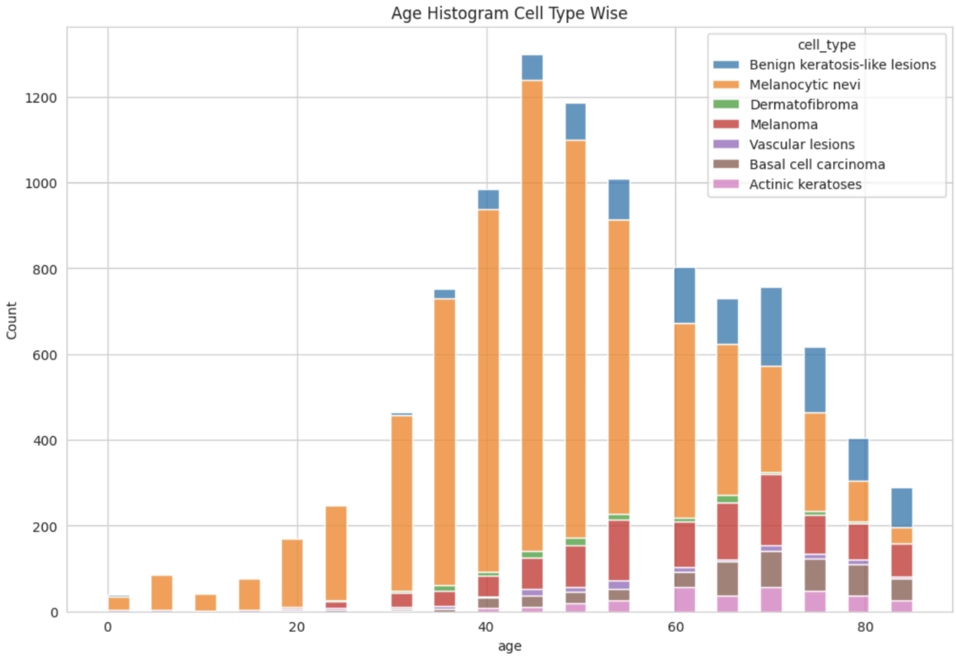
- Genderwise localization areas :



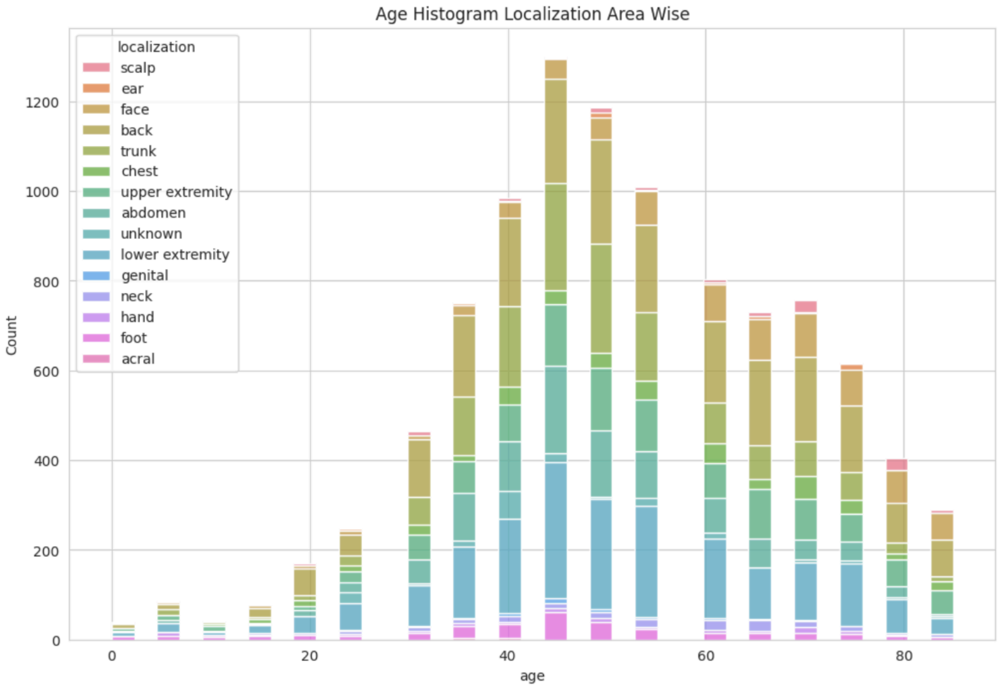
- Gender wise Age distribution :



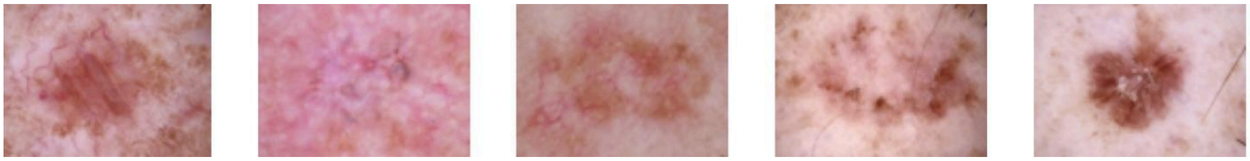
- Cell Type vs Age :



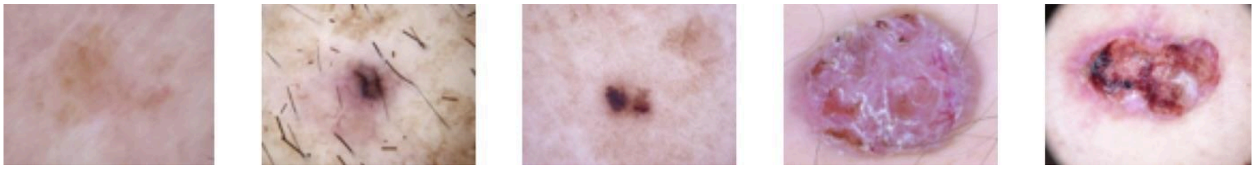
- Localization Area vs Age :



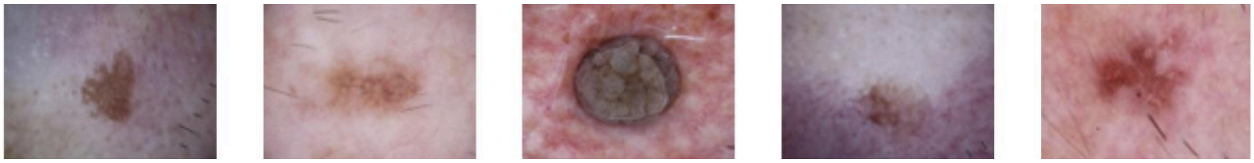
- **Pictorial representation of Images of dataset :**
  - Actinic keratoses :



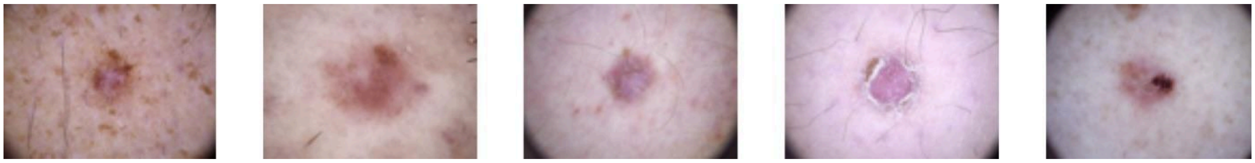
  - Basal cell carcinoma :



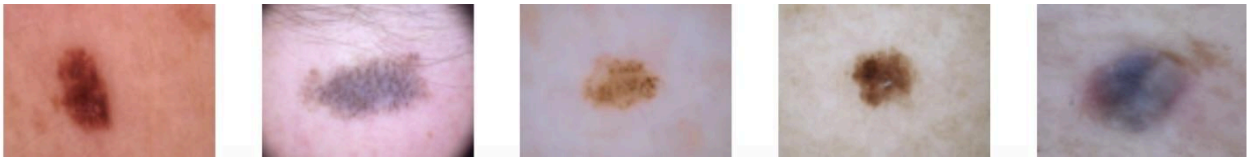
- Benign keratosis-like lesions :



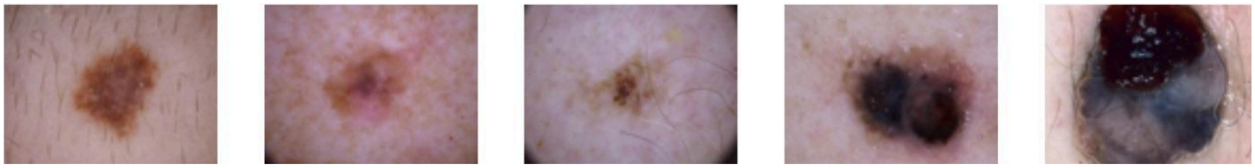
- Dermatofibroma :



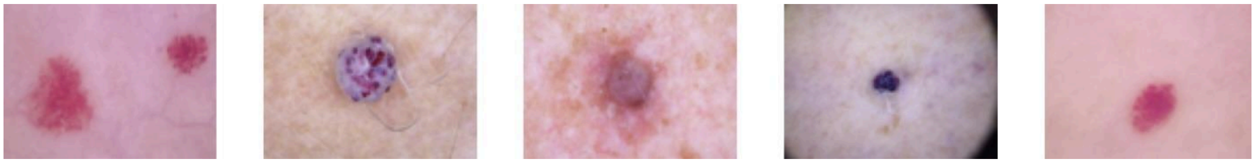
- Melanocytic nevi :



- Melanoma :

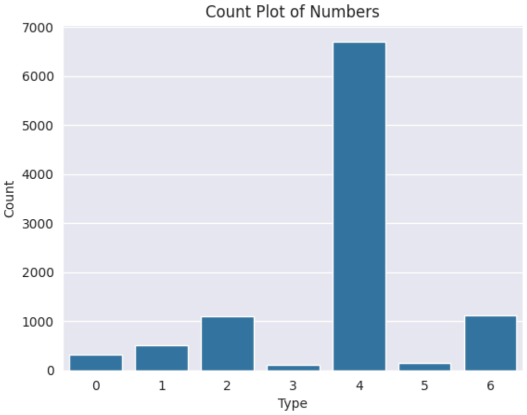


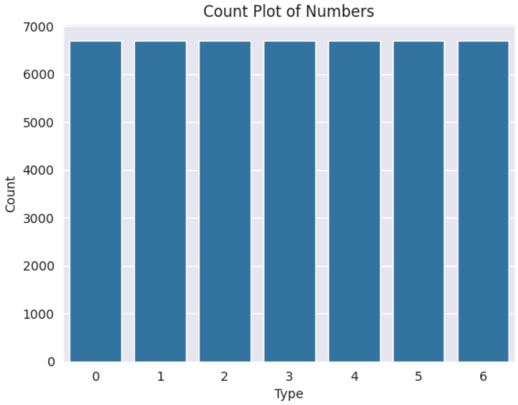
- Vascular lesions :



Data Preparation and Augmentation

The dataset undergoes a process of oversampling to address class imbalance, ensuring that all classes in the dataset are equally represented. This step is crucial for training a model that can accurately recognize and classify different types of skin lesions. After balancing the classes, the data is reshaped into a format appropriate for input into a convolutional neural network, specifically adjusting it to match the dimensions of 28x28 pixels with 3 color channels.





Random Forest Model

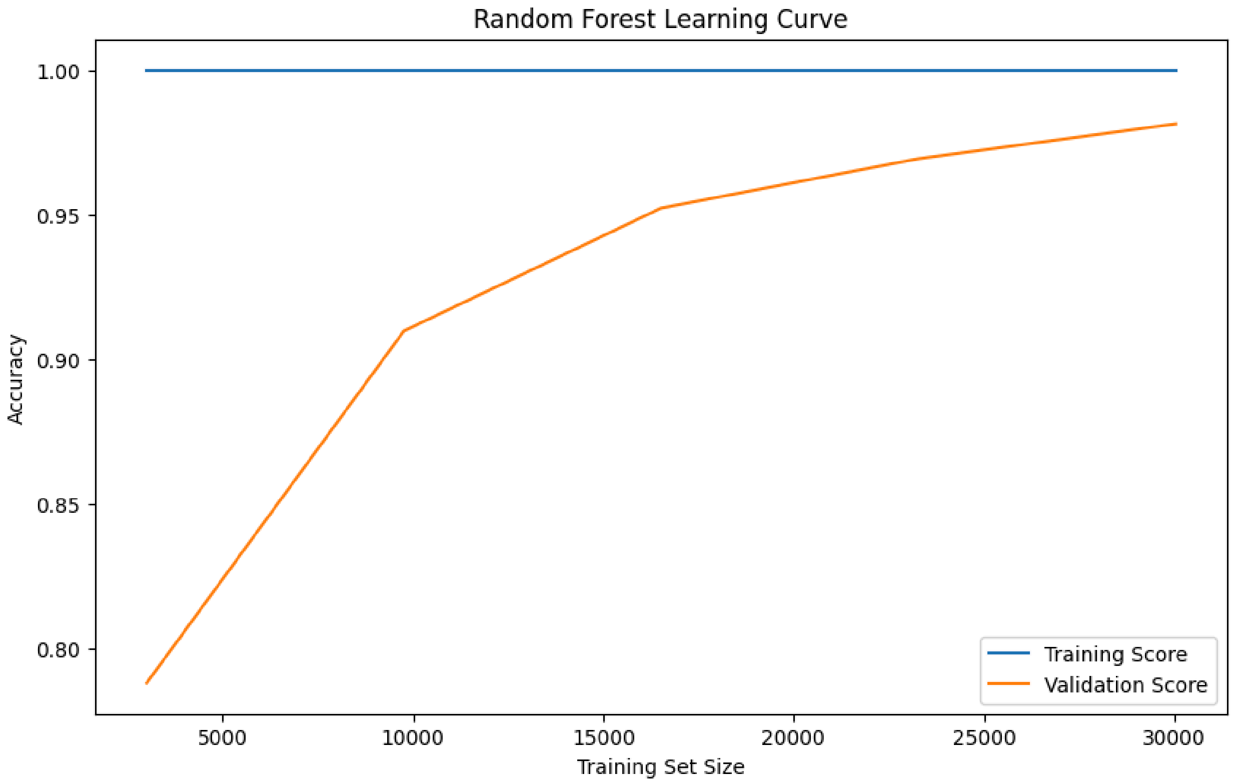
In the initial phase, the data was preprocessed using a column transformer that applied StandardScaler to standardize features by removing the mean and scaling to unit variance, and PCA (Principal Component Analysis) to reduce the dimensionality to 30 principal components.

A Random Forest classifier with a fixed random state was then trained on the preprocessed training data. The model's performance was evaluated on the test set, and the accuracy was reported. Furthermore, a classification report was generated, providing a detailed account of the model's performance across various metrics such as precision, recall, and F1-score.

To optimize the model, a RandomizedSearchCV was performed with a specified distribution for the number of trees (n_estimators) and the number of features to consider when looking for the best split (max_features). The search was conducted over a 5-fold cross-validation to identify the best hyperparameters for accuracy.

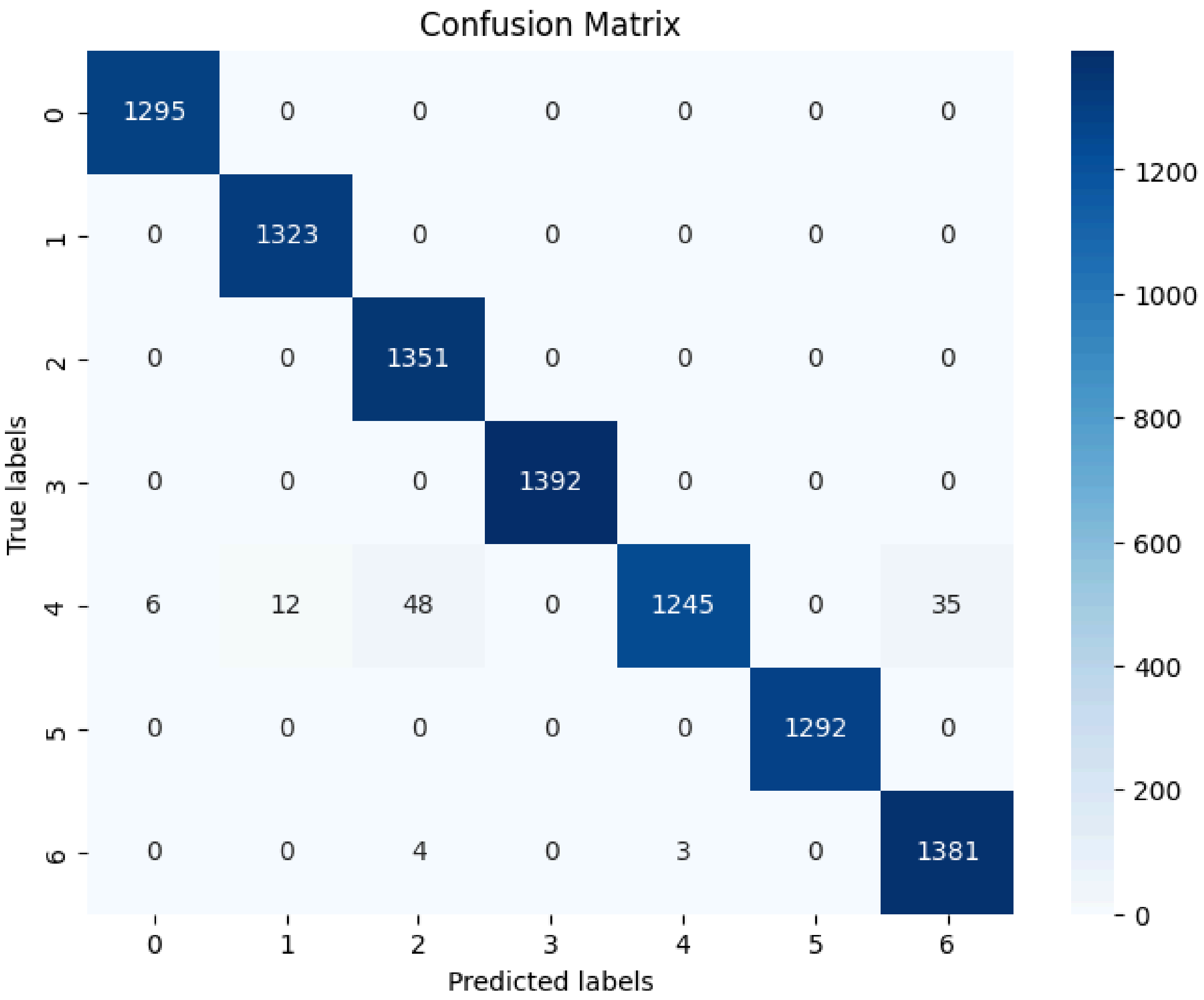
The best hyperparameters were reported along with the best accuracy score obtained through the randomized search. This indicated the effectiveness of the hyperparameter tuning process.

Finally, a learning curve was plotted to visualize the model's performance over increasing sizes of the training set. The curve depicted the relationship between the training and validation scores, indicating how well the model learned from the data. The Random Forest model displayed a consistent improvement in validation accuracy as the training set size increased, though the training score remained relatively constant.



The confusion matrix, a fundamental tool in classification tasks, was plotted using Seaborn's heatmap visualization to display the number of correct and incorrect predictions made by the model. Each cell in the matrix was annotated with the absolute number of predictions, providing clear insight into the instances of true positives, false positives, true negatives, and false negatives.

With the axes labeled for 'Predicted labels' and 'True labels', the confusion matrix offered a straightforward interpretation of the model's predictive accuracy. The 'Blues' color map used in the heatmap provided a gradient that helped in distinguishing between the different values, with lighter colors indicating lower values and darker colors corresponding to higher values.



True labels \ Predicted labels	0	1	2	3	4	5	6
0	1295	0	0	0	0	0	0
1	0	1323	0	0	0	0	0
2	0	0	1351	0	0	0	0
3	0	0	0	1392	0	0	0
4	6	12	48	0	1245	0	35
5	0	0	0	0	0	1292	0
6	0	0	4	0	3	0	1381

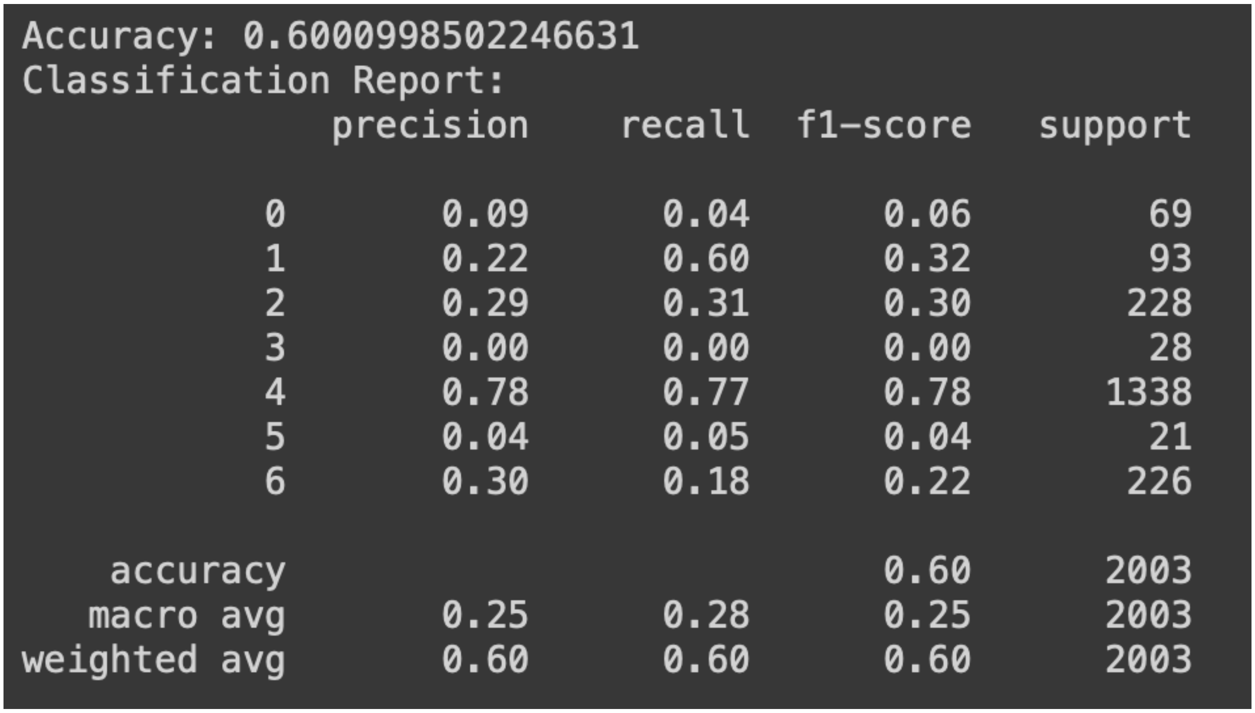
SVM

In preparation for the SVM application, the dataset consisting of RGB images was partitioned into training and test sets. The training set was employed to fit the LinearSVC model, which involved learning from feature vectors representing the images and their associated binary labels denoting the presence of cancer.

The performance of the trained model was quantitatively assessed using the test dataset. An accuracy score was calculated to determine the proportion of correct predictions made by the model. To provide a more granular analysis of performance, a classification report was produced, detailing precision, recall, and

F1-scores for each class, along with the support numbers that indicate the frequency of each class within the test set.

The derived accuracy indicates the model's general classification capability, and the additional metrics in the classification report give insight into the model's strengths and weaknesses in predicting each class. High precision is critical to minimize false positives, which is particularly vital in the medical field to avoid unnecessary treatments. Conversely, high recall ensures that actual instances of cancer are not overlooked, emphasizing the model's ability to serve as a reliable diagnostic tool.

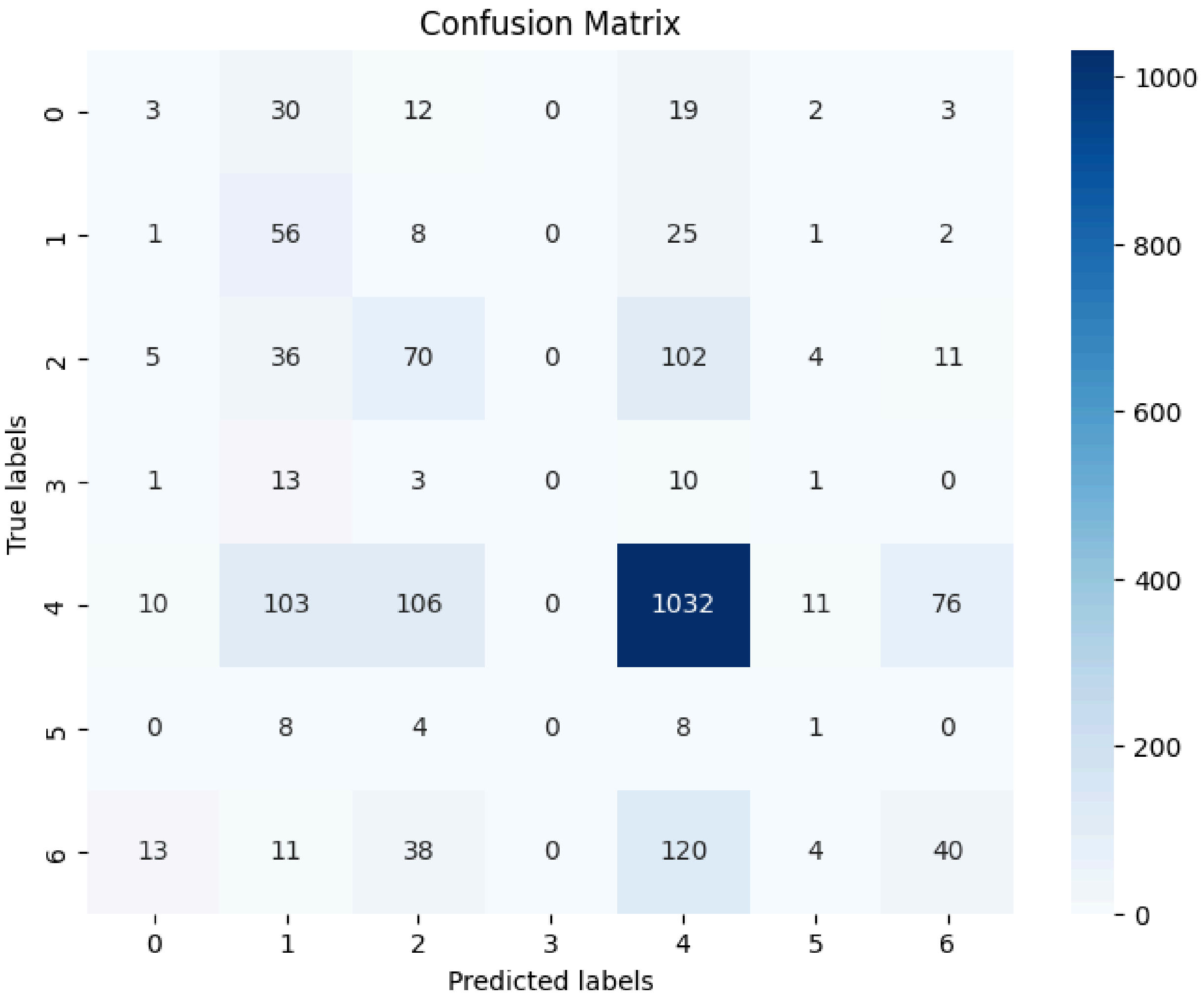


```
Accuracy: 0.6000998502246631
Classification Report:
              precision    recall  f1-score   support

           0       0.09      0.04      0.06        69
           1       0.22      0.60      0.32        93
           2       0.29      0.31      0.30       228
           3       0.00      0.00      0.00        28
           4       0.78      0.77      0.78      1338
           5       0.04      0.05      0.04        21
           6       0.30      0.18      0.22       226

    accuracy                           0.60      2003
   macro avg       0.25      0.28      0.25      2003
weighted avg       0.60      0.60      0.60      2003
```

- Confusion matrix



CNN Model Architecture

The model architecture is constructed using a sequential approach. It begins with a normalization step, where the pixel values of the images are scaled to a range between 0 and 1, which aids in the model's learning process.

The core of the model consists of several convolutional layers, each followed by a 'ReLU' activation function. These layers are designed to extract various features from the input images, with the number of filters increasing in deeper layers to capture more complex patterns. The model also employs varying kernel sizes throughout the architecture.

To mitigate the risk of overfitting, a combination of Max Pooling and Dropout techniques is utilized. Max Pooling is used to reduce the spatial dimensions of the output from previous layers, decreasing the number of parameters and computation in the network, thereby also reducing overfitting. The Dropout layer randomly sets a fraction of input units to 0 at each update during training time, which helps prevent overfitting.

The model transitions from convolutional layers to fully connected layers using a Flatten layer. This is followed by dense layers, including a final output layer tailored for a multi-class classification problem, indicated by its multiple units (one for each class) and a 'softmax' activation function. This setup suggests the model is aimed at classifying images into one of several predefined categories.

Model architecture

Training Configuration

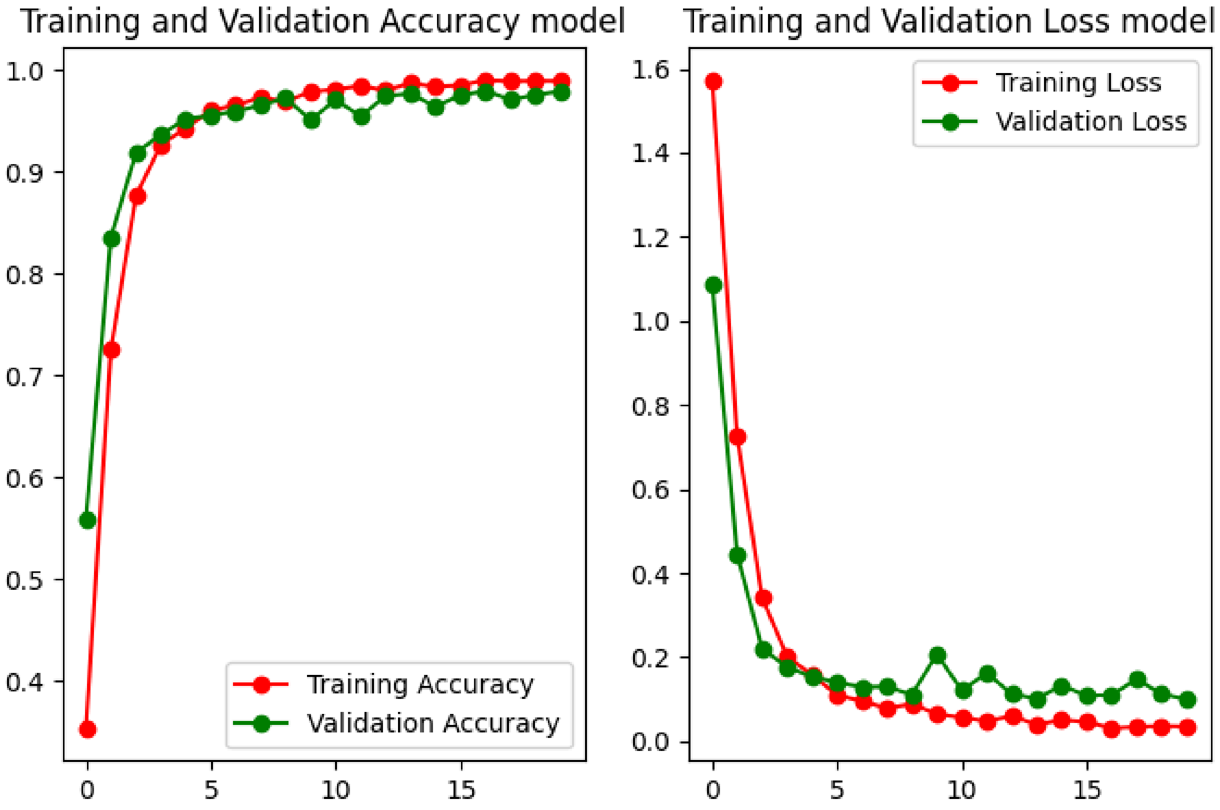
The model employs a sophisticated training regimen, incorporating techniques such as dynamic learning rate adjustments and early stopping. The learning rate adjustment mechanism reduces the learning rate when the validation accuracy plateaus, enhancing the training process's efficiency. Early stopping is used as a form of regularization, terminating training if the validation accuracy does not improve for a series of epochs, preventing overfitting and unnecessary computations.

The model is compiled with a specific optimizer and loss function, both chosen to be suitable for a multi-class classification task. The optimizer is responsible for updating the model weights, while the loss function quantifies the difference between the predicted outputs and the actual labels.

The training of the model is carried out over several epochs with a defined batch size, using separate sets of data for training and validation. This approach allows for the evaluation of the model's performance and generalization capabilities on unseen data.

Conclusion

This CNN model represents a comprehensive approach to classifying skin cancer images, utilizing advanced neural network techniques and training strategies. The model's architecture, combined with its training regimen, is designed to effectively learn from the dataset's complex patterns while ensuring robustness and preventing overfitting. The application of such a model has significant implications in the field of dermatology, particularly in early and accurate detection of various skin cancers.



Training and Validation Accuracy Plot:

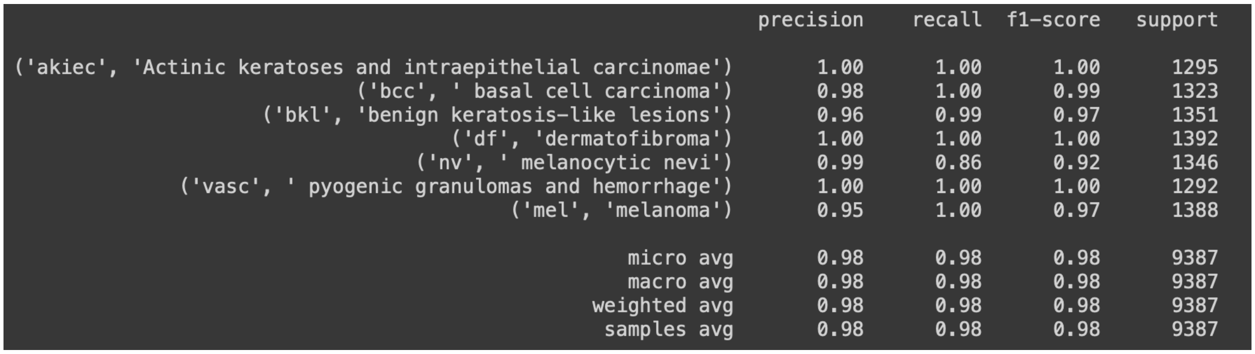
- **Training Accuracy (Red)**: This line represents the accuracy of the model on the training dataset. It increases sharply within the first few epochs and then plateaus, indicating that the model quickly learned to classify the training data and then made only marginal improvements.
- **Validation Accuracy (Green)**: This line represents the accuracy on the validation dataset. It follows a similar trend to the training accuracy, rapidly increasing and then leveling off. This close tracking with the training accuracy suggests that the model is generalizing well and is not overfitting significantly to the training data.

Both accuracies converge to a high value close to 1 (or 100%), which indicates a high level of performance of the CNN model on both the training and validation datasets.

Training and Validation Loss Plot:

- **Training Loss (Red)**: This line represents the model's loss (such as cross-entropy loss) on the training dataset. The loss decreases sharply within the first few epochs, which means the model's predictions are getting closer to the true labels quickly.
- **Validation Loss (Green)**: This line shows the loss on the validation dataset. Similar to the training loss, it decreases rapidly at first and then stabilizes. The validation loss appears to remain consistent after the sharp decline, suggesting the model is not learning any incorrect patterns from the training data.

The loss values settling down and the accuracy stabilizing at high values are indicative of a well-fitting model. The plots don't show signs of overfitting, as both training and validation lines are close together without a significant gap.



```
                                                               precision    recall  f1-score   support

('akiec', 'Actinic keratoses and intraepithelial carcinomae')       1.00      1.00      1.00      1295
                             ('bcc', ' basal cell carcinoma')       0.98      1.00      0.99      1323
                     ('bkl', 'benign keratosis-like lesions')       0.96      0.99      0.97      1351
                                     ('df', 'dermatofibroma')       1.00      1.00      1.00      1392
                                  ('nv', ' melanocytic nevi')       0.99      0.86      0.92      1346
              ('vasc', ' pyogenic granulomas and hemorrhage')       1.00      1.00      1.00      1292
                                          ('mel', 'melanoma')       0.95      1.00      0.97      1388

                                                    micro avg       0.98      0.98      0.98      9387
                                                    macro avg       0.98      0.98      0.98      9387
                                                 weighted avg       0.98      0.98      0.98      9387
                                                  samples avg       0.98      0.98      0.98      9387
```

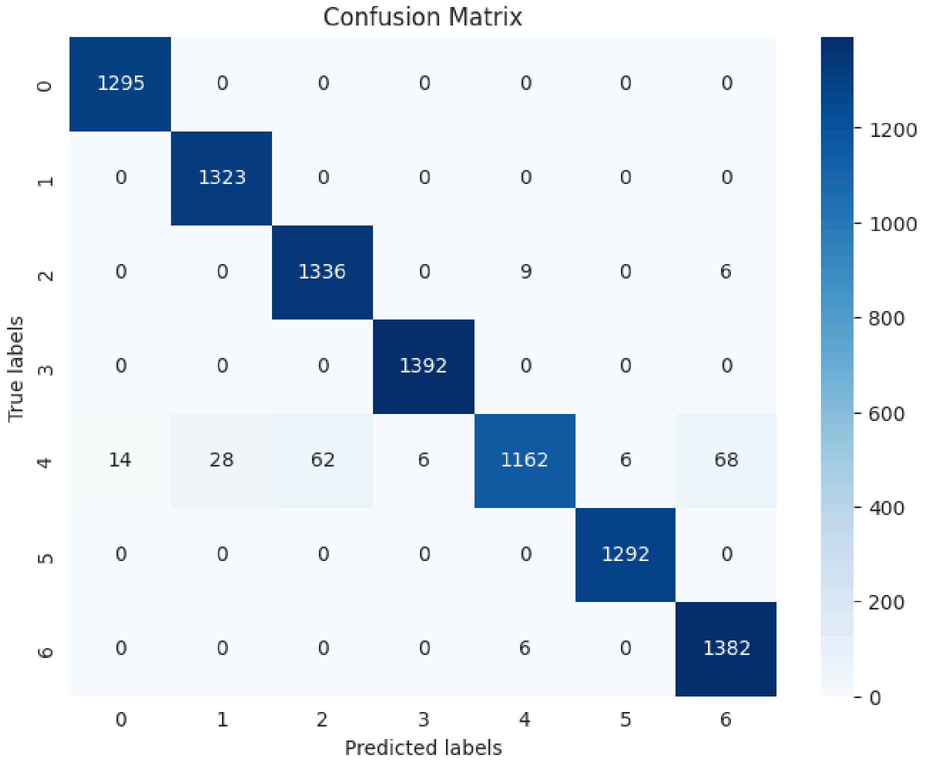
- **Actinic Keratoses and Intraepithelial Carcinomae (akiec)**: The model perfectly classifies this condition with a precision and recall of 1.00, which means all predictions for this class were correct (no false positives or false negatives).
- **Basal Cell Carcinoma (bcc)**: The model has a precision of 0.98 and a recall of 1.00, indicating high accuracy, with very few false positives and no false negatives.
- **Benign Keratosis-like Lesions (bkl)**: The precision is 0.96, and the recall is 0.99, which means the model is slightly more likely to incorrectly predict other conditions as bkl (false positives), but it rarely misses actual bkl cases (high recall).
- **Dermatofibroma (df)**: The model achieves perfect precision and recall for dermatofibroma, similar to akiec.
- **Melanocytic Nevi (nv)**: The model shows a precision of 0.99 but a lower recall of 0.86, indicating that while its predictions for nv are very accurate, it tends to miss more actual cases of nv (false negatives).
- **Pyogenic Granulomas and Hemorrhage (vasc)**: This condition is perfectly classified by the model with precision and recall both at 1.00.
- **Melanoma (mel)**: The model's precision is 0.95, and the recall is 1.00, indicating high accuracy with very few false positives.

The support column indicates the number of true instances for each class in the dataset used to generate this report.

The averages provided at the bottom of the report (micro, macro, weighted, and samples) are all 0.98, which shows excellent overall performance across all classes.

- **Micro Average**: This is the overall accuracy of the model, computed from the total true positives, false negatives, and false positives.
- **Macro Average**: This is the average precision, recall, and f1-score of all classes, treating all classes equally regardless of their support.
- **Weighted Average**: This takes the support of each class into account when computing the average, giving more weight to classes with more instances.
- **Samples Average**: This applies only to multi-label classification tasks and treats every instance-label pair as a binary classification.

  - Confusion matrix



	0	1	2	3	4	5	6
0	1295	0	0	0	0	0	0
1	0	1323	0	0	0	0	0
2	0	0	1336	0	9	0	6
3	0	0	0	1392	0	0	0
4	14	28	62	6	1162	6	68
5	0	0	0	0	0	1292	0
6	0	0	0	0	6	0	1382