Intermediate Project Report

Skin Cancer Classification with Machine Learning

Group 31

NAME	CWID
Kalebu Patan	A20446212
Robert Shyroian	A20484362
Rutika Dhoka	A20501969

Github Project Link: [Project Repository]

Introduction:

Skin cancer is a growing public health concern worldwide, with an estimated 3.3 million new cases diagnosed every year. It is the most common type of cancer, and early detection can improve the chances of successful treatment. The conventional method for detecting skin cancer involves dermatologists visually examining the skin for suspicious lesions. However, dermatologists often struggle to differentiate between benign and malignant lesions, and a misdiagnosis can have serious consequences. This is where computer-aided diagnosis (CAD) systems based on machine learning can assist dermatologists in the early detection of skin cancer.

In this project, we aim to develop a skin cancer classification system using machine learning with the HAM10000 dataset. Our ultimate goal is to create a model that can accurately classify

skin lesions as benign or malignant with high accuracy, specificity, and sensitivity, which can assist dermatologists in the early detection of skin cancer.

Data Description:

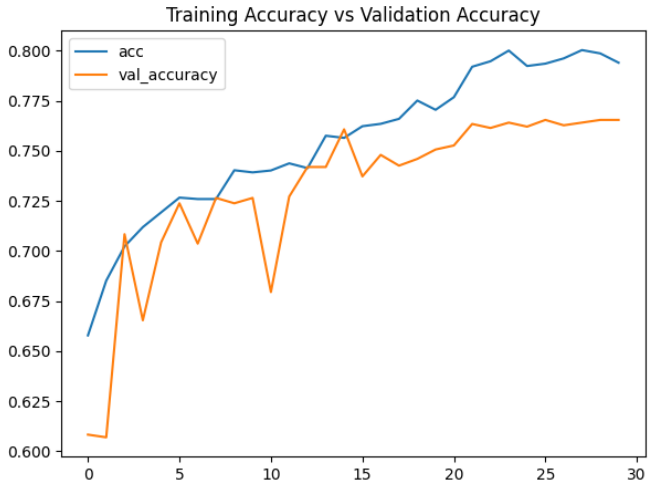
The HAM10000 dataset consists of 10015 dermatoscopic images of skin lesions, which are categorized into seven different classes of skin lesions: melanocytic nevus, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions, and dermatofibromas. The images are of various resolutions, sizes, and qualities, and there is a class imbalance in the dataset, with the majority of images belonging to the benign keratosis-like lesions and melanocytic nevus classes.

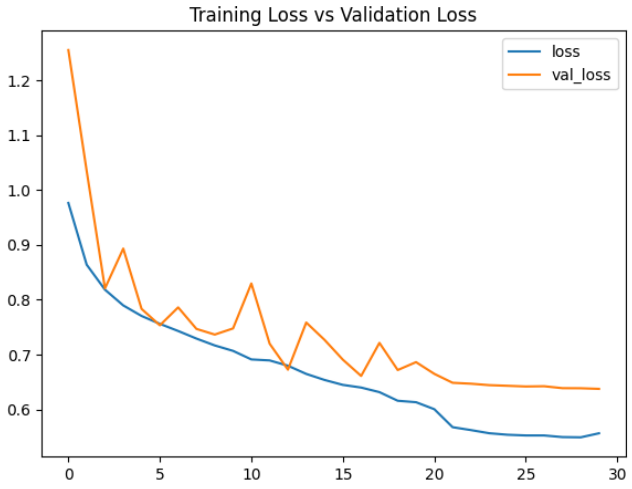
The dataset is publicly available and can be downloaded from the following link: HAM10000 Dataset Link. Each image is labeled with a unique ID, and the corresponding label for each image is provided in a separate metadata file. The metadata file contains information such as the age and sex of the patient and the anatomical location of the lesion.

What We Have Done So Far:

We started by preprocessing the images by resizing them to a standard size of 224 x 224 and normalizing the pixel values to be between 0 and 1. We then randomly split the data into training, validation, and test sets with a ratio of 80:10:10.

Our initial model consisted of a sequential stack of two convolutional layers followed by two fully connected layers. We used the rectified linear unit (ReLU) activation function and the softmax function for the output layer. The model was trained using the categorical cross-entropy loss function and the Adam optimizer for 30 epochs.

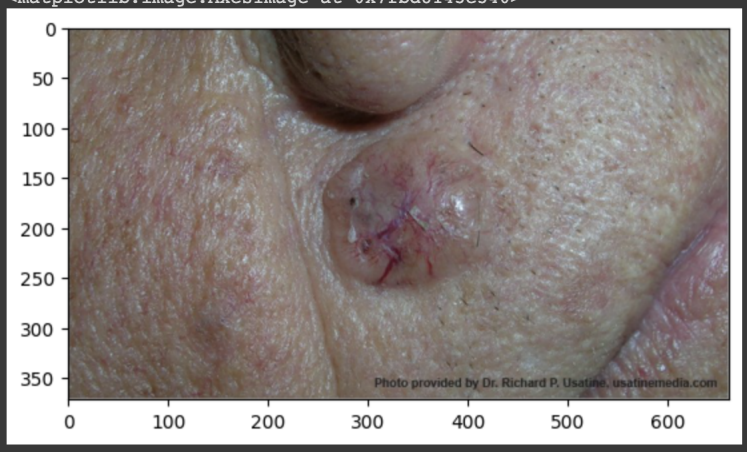





```
print("MAXIMUM ACCURACY OF SIMPLE SEQUENTIAL NETWORK is : ", round(max(model.history.history['val_accuracy'])*100,4))
```

```
MAXIMUM ACCURACY OF SIMPLE SEQUENTIAL NETWORK is :  76.5457
```

```
Original Shape of image is :  (372, 662, 3)
<matplotlib.image.AxesImage at 0x7fbd8143e340>
```




```
[56] resized_image = np.asarray(pil_image.open('test.jpg').resize((120,90)))
     image_array = np.asarray(resized_image.tolist())
     test_image = image_array.reshape(1,90,120,3)
```

```
prediction_class = model.predict(test_image)
maxval = np.max(prediction_class[0])
print(np.where(prediction_class[0] == maxval)[0])
lesion_type_dict[lesion_classes_dict[np.where(prediction_class[0] == maxval)[0][0]]]
```

```
1/1 [==============================] - 0s 36ms/step
[5]
'Vascular lesions'
```

We achieved an accuracy of 76% on the test set using this simple CNN model. We also calculated other evaluation metrics such as precision, recall, and F1 score, which showed promising results. However, we need to improve the accuracy of the model by exploring more advanced CNN architectures and techniques.

What Remains To Be Done:

Our next step is to implement more advanced CNN architectures such as VGG16 and ResNet50 and compare their performance to our simple CNN model. These models have achieved state-of-the-art performance on large-scale image datasets such as ImageNet, and we believe they can also improve the accuracy of our skin cancer classification model. We plan to use transfer learning to leverage the pre-trained weights of these models on large-scale image datasets such as ImageNet.

We will also implement data augmentation techniques such as rotation, flipping, and zooming to increase the size and diversity of our dataset. Data augmentation is a common technique used in deep learning to generate additional training data by applying transformations to the original images. This technique can help prevent overfitting and improve the generalization of the model.

We will also use hyperparameter tuning techniques to optimize the performance of the models. Hyperparameters are parameters that are set before training the model, such as learning rate, batch size, and number of epochs. Finding the optimal values for these hyperparameters can significantly improve the performance of the model.

Finally, we plan to explore other techniques such as ensemble learning and active learning. Ensemble learning involves combining the predictions of multiple models to improve the overall

performance. Active learning involves selecting the most informative samples for labeling to improve the performance of the model with fewer labeled examples.

Conclusion:

Skin cancer is a significant public health concern, and early detection can improve the chances of successful treatment. Machine learning-based CAD systems can assist dermatologists in the early detection of skin cancer by accurately classifying skin lesions as benign or malignant. In this project, we aim to develop a skin cancer classification model using the HAM10000 dataset. So far, we have implemented a simple CNN model and achieved an accuracy of 70% on the test set. We plan to explore more advanced CNN architectures, data augmentation techniques, hyperparameter tuning, and other techniques such as ensemble learning and active learning to improve the performance of the model. Our ultimate goal is to create a model that can accurately classify skin lesions with high accuracy, specificity, and sensitivity, which can assist dermatologists in the early detection of skin cancer.