Automated Skin Cancer Detection

Aditya Prasad
(2022036)
Upal Majumder
(2021214)

1. Problem Statement:

Skin cancer, particularly malignant melanoma, poses a significant health threat with increasing instances reported annually. Detecting skin lesions, a precursor to skin cancer, is challenging due to artifacts, low contrast, and visual similarities with benign features like moles or scars. Manual inspection lacks precision, often leading to misdiagnosis and adverse outcomes, including fatalities. The imperative for accurate and efficient detection is evident, as early identification greatly improves survival rates. However, existing methods face hurdles in effectively differentiating between malignant and benign lesions, necessitating automated approaches for enhanced reliability.

2.Literature Review:

Most of the research in skin cancer detection and classification is done on applying Deep Learning algorithms or applying Deep learning model results on classical Machine learning models. [1] proposes a deep learning model which uses multiple CNN architectures like AlexNet, VGGNet etc. to train their model and come up with accuracy of 92.70%. [2] proposes a MobileNet(CNN architecture) model in which after edge based filtering, an accuracy of 88.70% was achieved and on data augmentation, an accuracy of 97.58% was achieved. [3] proposes an ANN based network which achieves an accuracy of 95.43%.
Apart from that, many papers([4], [5], [6]) have been developed which mainly focus on Deep learning models.

Our aim in this project is to come up with methods, which emphasizes on classical ML based models and assess how they perform as compared to the DL model

3. Dataset:

The dataset used for this project is the HAM10000 dataset. There are a total of 10015 data samples in this data. The dataset consists of two parts.:-

1) **HAM10000_metadata.csv**: It consists of tabular data like patient id, class id, sex, place of cancer etc. and the class dx which signifies the skin cancer disease the patient is suffering from, with outliers being a part of this dataset. A total of 6 features have been assessed and encoded(in our project) for the analysis of the data. Below image gives a rough idea of the dataset:-

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

2) **HAM10000_images folder:** This folder contains 10015 gfg image files representing different forms of skin cancer suffered from patients, with the image size being There are a total of 7 classes of skin cancer diseases the patients suffer from. They are as follows:-
    - Bkl: Benign Keratosis lesions
    - Nv: Melanocytic nevi
    - Mel: Melanoma
    - Bcc: Basal cell carcinoma
    - Akeic : Actinic kerotoses
    - Vasc: Vascular lesions
    - Df: Dermatofibra

Originally, HAM10000 was built with the motivation of training neural networks after it was noted that the training of these models were hampered by the small size and lack of diversity of available dermatoscopic images. This dataset is currently available at the Harvard Dataverse page: [Harvard dataverse]

4. Proposed Architecture:

The proposed architecture is as follows:-

1) Preprocessing of data
2) Data augmentation(can be a part of step 1)
3) Feature extraction
4) Model 0 training
5) Model 1 training(if required)

6) Activation function in the end(if required)
7) Testing the model

For the first step, the image data is combined with the tabular data. Furthermore, the image data is augmented and they are rotated, cropped, translated to produce more possible datasets and assisting in training the model.
After data augmentation, the aim is to convert the (m*n*3) array into a flattened list and expand the columns based on the pixels. One option can also be to grayscale the images to reduce the number of columns. Once the images are augmented, all the features are to be scaled and in case of missing data, they are filled with the corresponding median values of the column. The column data is then normalized using Standard Scaler to make sure the data points have a balanced scale, aiming to reduce sensitivity in feature scales. The scaled values are then passed through PCA to reduce the dimensionality of the dataset.

After the feature extraction stage, the next step is to train on the models. The following models have been proposed:-

1) Random Forest Classifier
2) SVM
3) K Nearest Neighbours
4) Convolutional Neural Networks(very useful tool in image data noting the fact that these use filters which are convoluted with the image data, which produces excess hidden layers for the model to be trained on)

Another possible proposal(if within scope) is to send the results of one model(like ResNet, out of scope for now) into another one of the models proposed above and send the output of this model to an activation function like Logistic regression, and then determine the possible output.

The evaluation metrics for the model is undecided for now.

5.Data Visualisation:

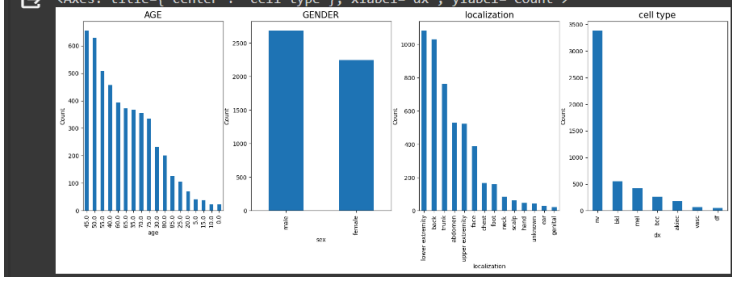
Visualizing different classes of skin cancer

```
fig.savefig('category_samples.png', dpi=300)
```

Distribution of tabular data and the class in in the dataset

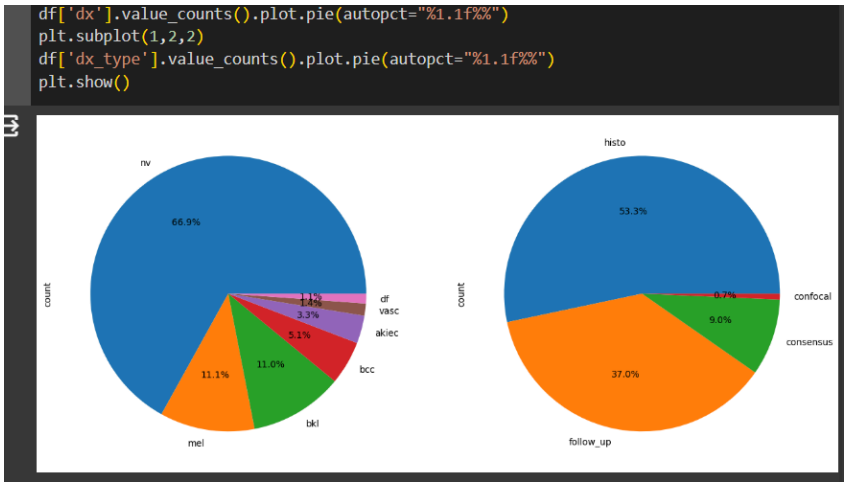
```
plt.title("cell type",fontsize=15)
plt.ylabel("Count")
filtered_df['dx'].value_counts().plot.bar()
```

```
<Axes: title={'center': 'cell type'}, xlabel='dx', ylabel='Count'>
```



Pie chart showcasing the distribution of classes in the dataset

```
df['dx'].value_counts().plot.pie(autopct="%1.1f%%")
plt.subplot(1,2,2)
df['dx_type'].value_counts().plot.pie(autopct="%1.1f%%")
plt.show()
```



6. Next Steps

The next steps involve the working on the models proposed, assess the testing accuracy and the test metrics, and check if there is scope for improvement.