Executive Summary for analysis on Skin Cancer MNIST dataset

Sushant Baskota

MNIST Skin Cancer Dataset:

The dataset consists of 10000 images of skin cancer lesions and the filename so the images are kept track in a csv file which also has the cell types, localization of the lesion in the body, age and sex of the patient, and also how the condition was diagnosed. The best thing for working in this dataset would be a classification problem and classifying the cancer cell type. However, the dataset is not a well distributed dataset. It consists a high number of data for some cell types while a very small amount for others.

```
cell_type
Actinic keratoses                     327
Basal cell carcinoma                  514
Benign keratosis-like lesions        1099
Dermatofibroma                        115
Melanocytic nevi                     6705
Melanoma                             1113
Vascular lesions                      142
Name: image_id, dtype: int64
```

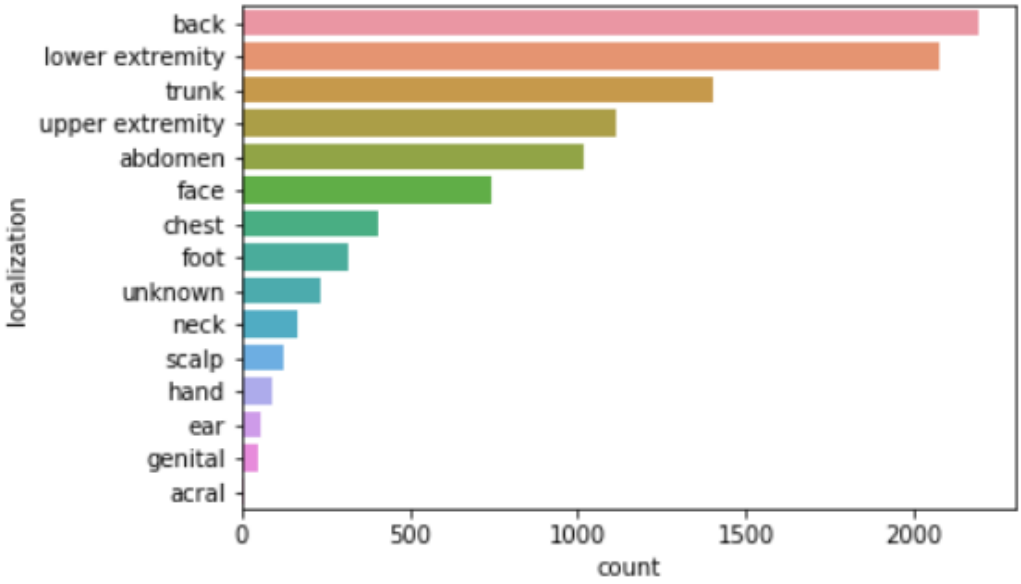
	age
count	9958.000000
mean	51.863828
std	16.968614
min	0.000000
25%	40.000000
50%	50.000000
75%	65.000000
max	85.000000

Melanocytic nevi has almost 65% of the data. This might affect any analysis on this dataset heavily.

Age Distribution: The dataset doesn't have numerical data other than age. But the age distribution can be seen alongside. Mean age seems to be 51 for the dataset.

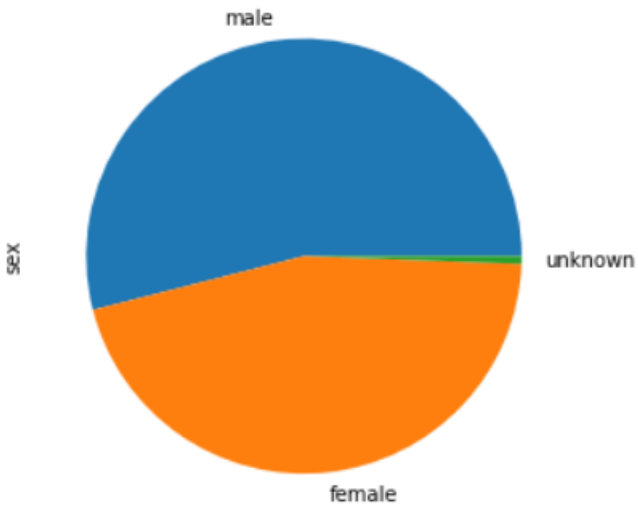
It is very important to note than any of these analysis and graphs have no significance outside this dataset and cannot be extrapolated to make age, localization predictions about skin cancer. They just describe the dataset. The learning model, however, should be able to predict the cell types.

Plots from the data

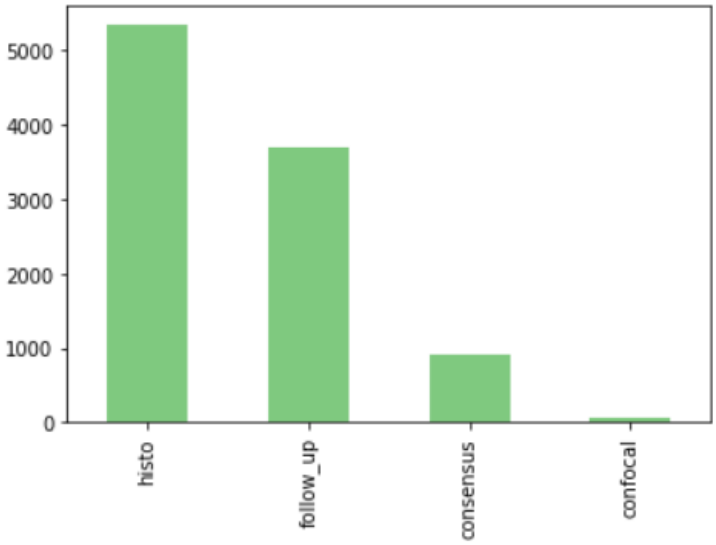


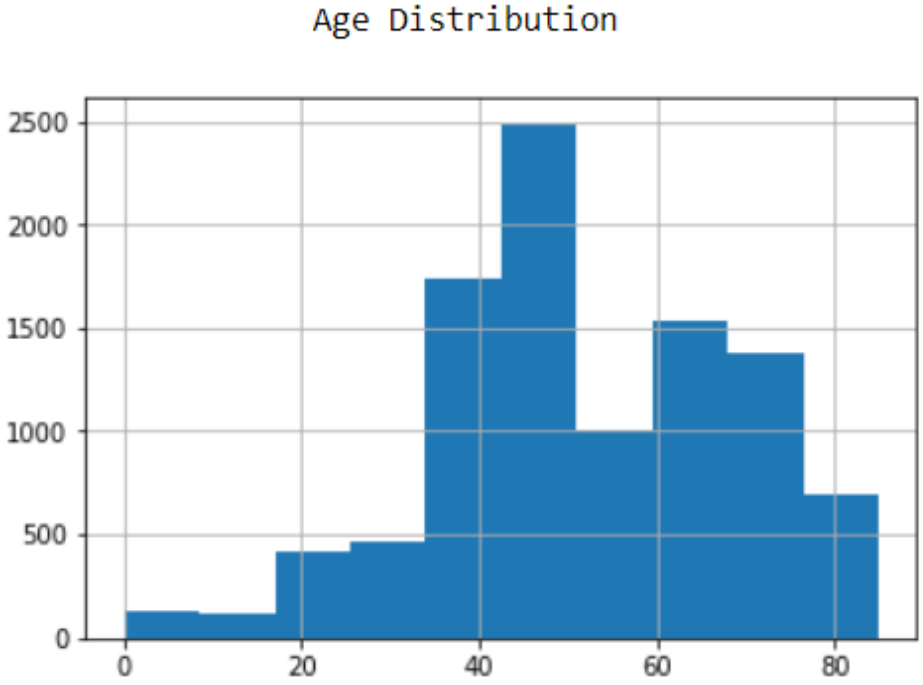
This shows that in the dataset, most of the skin conditions occurred in the back and the least number in the acral area. As mentioned above, this just describes the dataset. But the reason that most of the conditions come from the back are because it is easier to collect the pictures from such area than it is from complicated areas like genital area.

Sex Distribution



Diagnosis method





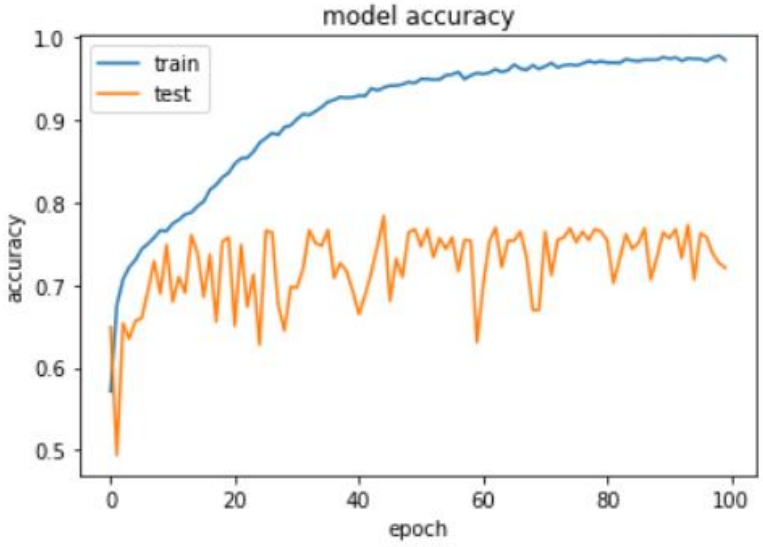
The graph alongside gives how the skin conditions were diagnosed. The highest number of the data seems to have been from histopathology and most others from follow ups.

The pie chart below gives the distribution of data by sex. It is very balanced in this case. Male and Females have almost same amount of data.

The Histogram shows the age distribution in the dataset. It looks like the most data comes from age range 40 to 60.

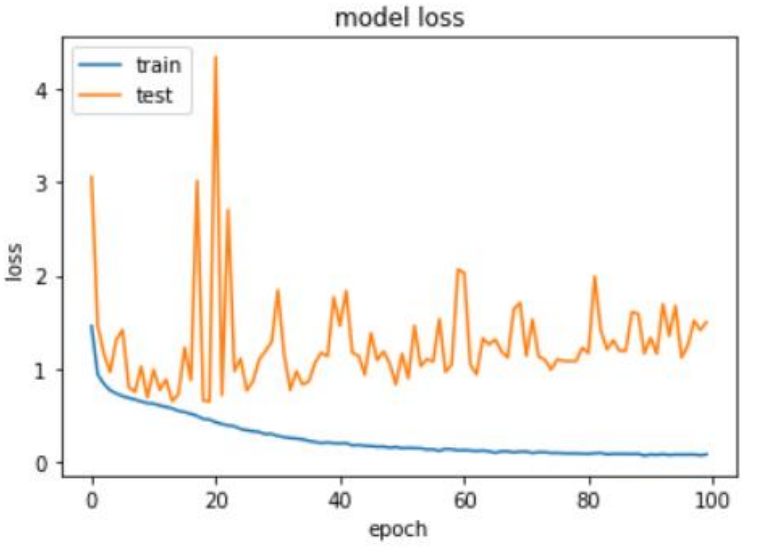
Machine Learning

For training purpose, using the whole image would be very resource intensive and it took almost four hours to run a Convolutional Neural Net on a supercomputer, so, I used the csv file in the dataset that has numerical value for each image. There are total 28x28x3 features 28x28 being total pixels and 3 being color value of each of those pixels in RGB in this case. So, you could imagine a 3D matrix or tensor that is fed to the neural net. These values are sent to the neural net instead of a high-resolution picture which could have 100x100x3 or higher number of features based on the size of the image. Taking a numerical data from the start really helps the neural net to train faster although there might be a significant drawback in the accuracy.



Why convolutional neural network?

Convolutional neural networks are known for their efficiency in pattern recognition. These have convoluted layers that are a bit different than normal hidden layers in a Multi-Layer Perceptron and use filters that emphasizes the patterns in the image (in this case). For example, if we have a 2x2 filter, we have a 2x2 matrix that is multiplied/applied to the 2x2 pixels of the pictures. This helps in leaving off the pixels with less information and emphasizing the pixels with most information. For this dataset, the neural network model used is a custom-built neural network with 5 convoluted layers. This neural net also uses Pooling (MaxPooling to be specific). In this case, MaxPooling takes A X A pixel matrix in the image and finds the max values of those pixels and creates a single pixel. So, this downsizes the images while still keeping the most important feature intact. This helps in speeding up the training process.



Since we have a data with that is not well distributed, some classifications were easy to learn while others were not. So, the training was done for almost 100 epochs and the trained model was saved to a file ‘cancer.model’. This helps us in just predicting the accuracy of classification without having to train the model again.

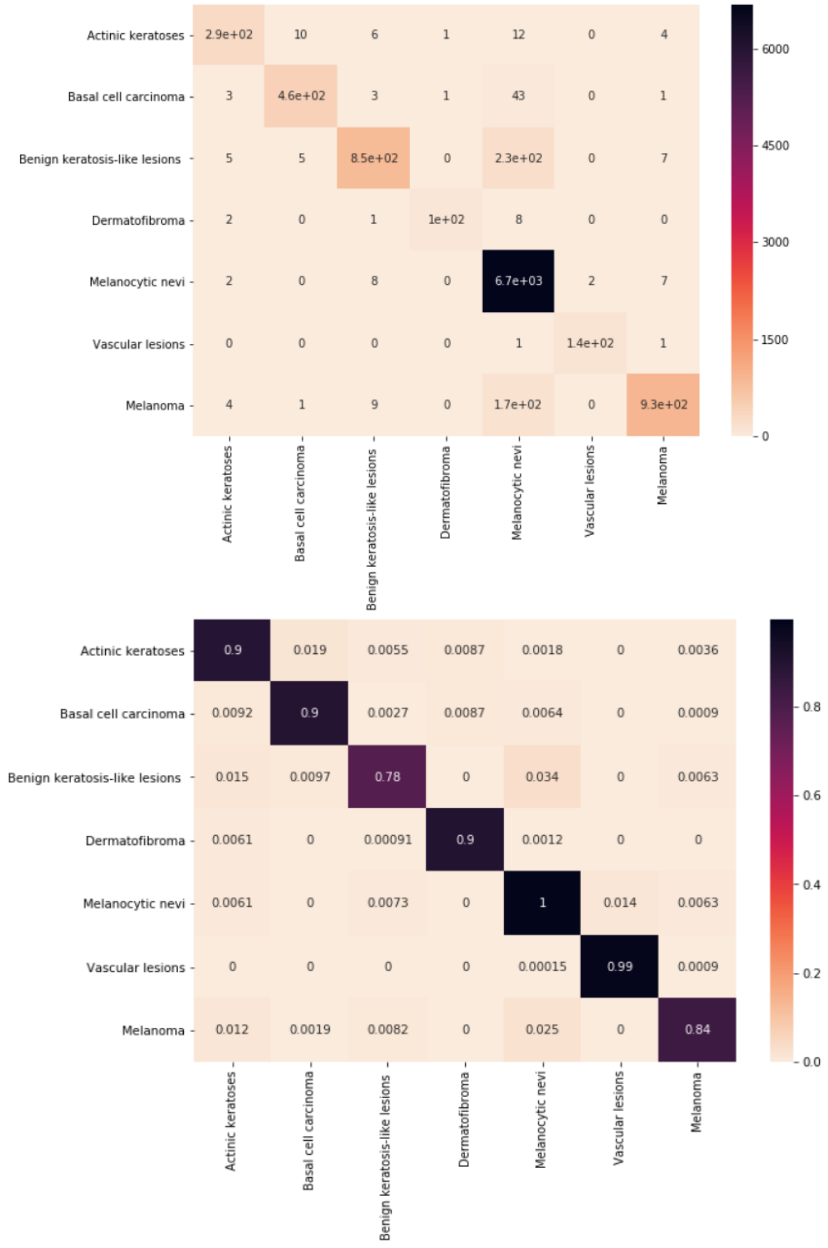
Accuracy on test set: 75%
Accuracy on whole set: 94.58%

This is not a good enough prediction because we can see that training prediction has been increas ing while the validation prediction is not. This is because of overfitting and also because of the imbalance of the dataset.

Overfitting: This occurs when the model conforms to the training set so much that it doesn't pre dict right on the test set. This often occurs if we make the problem more difficult that it is. While a deep convolutional neural network could be making the problem difficult than it should be but we cannot say that because the dataset is poorly distributed into classes for this classification pro blem.

Confusion Matrix

The confusion matrices alongside were plot after the model was traied. The topmost matrix is not normalized which is why it shows a huge number in melanocytic nevi because there is a huge number of those in the dataset. For the sake of accuracy, the matrix was normalized as seen in the bottom matrix. We can interpret from that matrix that it found difficult to predict the Benign keratosis-like lesions more than any other cell types. The model seems pretty certain about the predictions of Melanocytic nevi but that is because data has more than 50% of those cell types which causes the imbalance.



Future Predictions/Learning on this dataset

It is very evident that dataset is not well distributed categorically. So, for a better accuracy at both the train and test data, without overfitting, the data can be trained such that each class has similar number of data. This can be done by randomly picking a certain number from classes that have high number of data and taking similar number of data from other classes and going through the whole data such that the model can be as confident about other cells as it is about Melanocytic nevi. Also if I had access to a good amount of time and high processing resource, I would be using the whole picture in its high resolution which would be much better than taking 28 x 28 into consideration because the higher the features the more intricate details the neural net can pick up and the accuracy is higher as well. The overfitting problem can also be solved by this method.

References

https://towardsdatascience.com/a-guide-to-pandas-and-matplotlib-for-data-exploration-56fad95f951c
https://www.learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras/
https://seaborn.pydata.org/tutorial/distributions.html