Dataset Description

We have 2 datasets in this project.

Dataset 1

This dataset contains 12 Clinical Attributes and 22 Histopathological Attributes. The names and id numbers of the patients were recently removed from the database. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

Dataset 2

The dataset comprised a retrospective convenience sample across all images of Fitzpatrick I-VI but was also designed to allow direct comparison between Fitzpatrick I-II and Fitzpatrick V-VI by matching diagnostic category, age within 10 years, gender, and date of photograph within 3 years. The images included in the DDI dataset were retrospectively selected from reviewing pathology reports in Stanford Clinics from 2010-2020. There are 656 images representing 570 unique patients. Each image label was expertly curated: skin tone was labeled based on in-person evaluation at the clinic visit cross-referenced against demographic photos and review of the clinical images by two board certified dermatologists.

Problem

To build a model to classify the patients data into different categories based on the 34 attributes We want to analyse the dataset we considered for this. There are some clinical attributes and histopathological attributes which are recorded from the patients. There are some missing values in the dataset, and the values are very sensitive i.e there are many diseases that gives the same symptoms diagnosis gives the same result. This database contains 34 attributes, 33 of which are linear valued and one of them is nominal

Research Problem

1. Can we predict the skin disorder by only using the clinical attributes(data)? What is the influence or contribution of histopathological attributes on the output?

2. Can we identify the the type of skin disorder using a image provided? Can we predict the other parameters like skin tone and severity of the disease.

3. Is there any possiblity of finetuning the model using both the image and attributes from the dataset 1 to make better predictions of the skin disorder.

Data cleaning

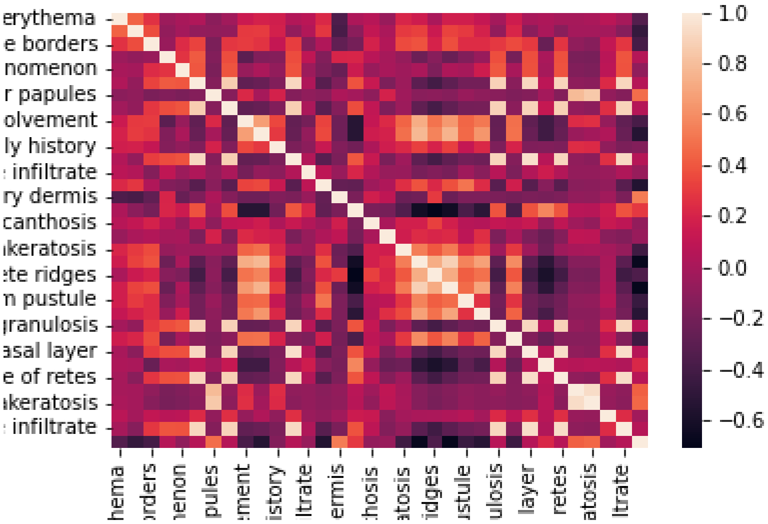
1. There are 8 missing values in the dataset 1, they are all in the 'Age' attribute. The dataset contains limited number of examples and each value is important for the analysis. So we cannot afford to lose the data and drop any rows. So to fill the data we are using regression models to predict the missing values.

We are creating a deep copy of the dataframe and removed the rows where there are missing values. We trained the model with the new dataframe and obtained the results by feeding the known data in the original dataframe to find the age. 2. For the dataset 2, the data is not in the correct structure for a model to be trained. So, the main idea is to split the data into the different directories that are marked by their respective labels. There are some missing values in the other output variables which were handled by filling them with mean values.

Exploratory Data Analysis

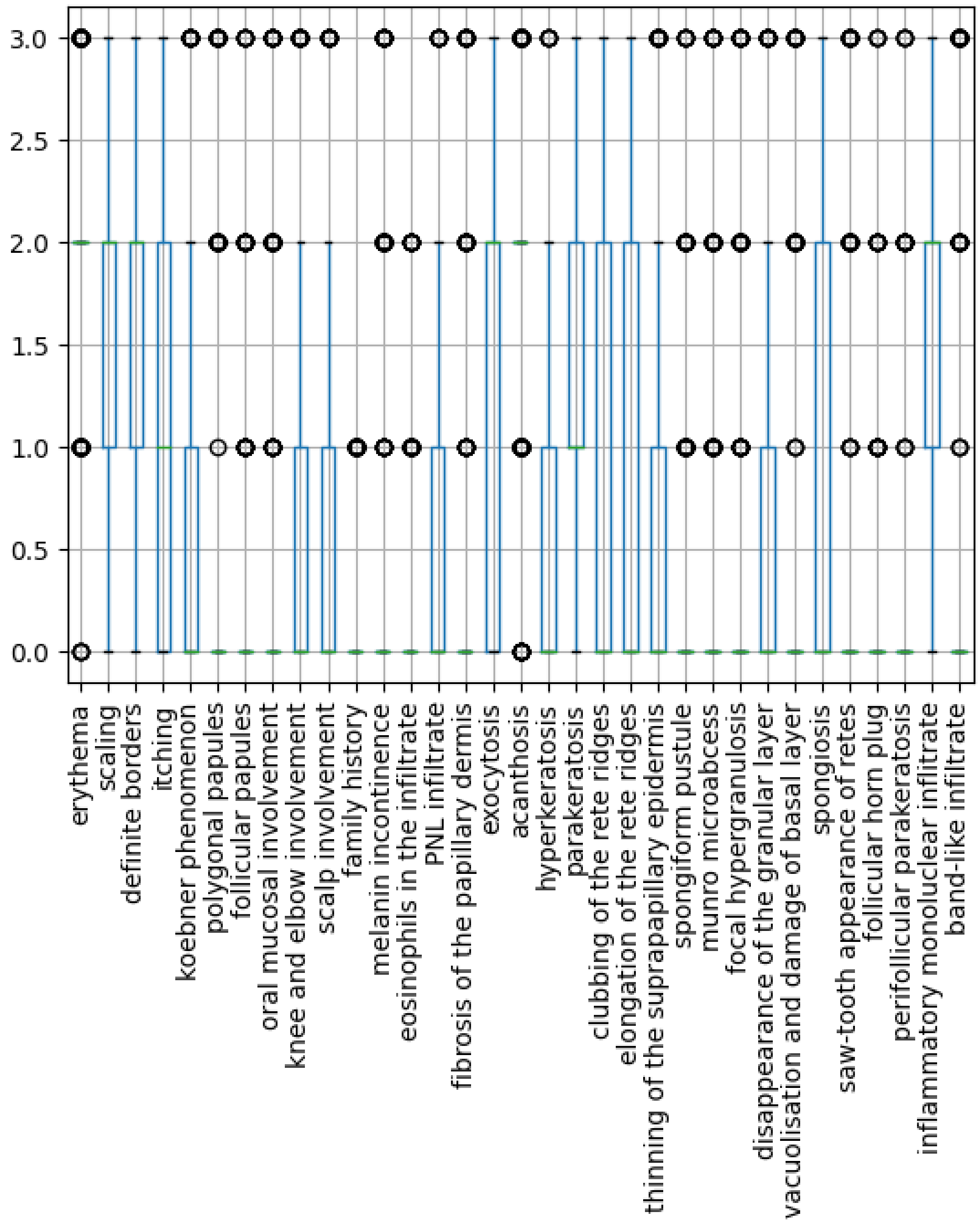
Dataset 1

Dataset 1 contains 365 rows and 35 attributes.
This is a multivariant data set.
34 attributes of the dataset are linear(int64) and 'Age' attribute is nominal(Object).
Correlation between the columns was also found (Pleasa refer to dataset1.py) for the results.



This is an observation done by the researchers, we may find the correaltion between the attributes but according to the analysis I have done till now, we need each and every attribute to train the further models.
Even when cleaning the data we trained the Linear Regression model with all the attributes.
The above heatmap is to just show the correlation between the attributes.

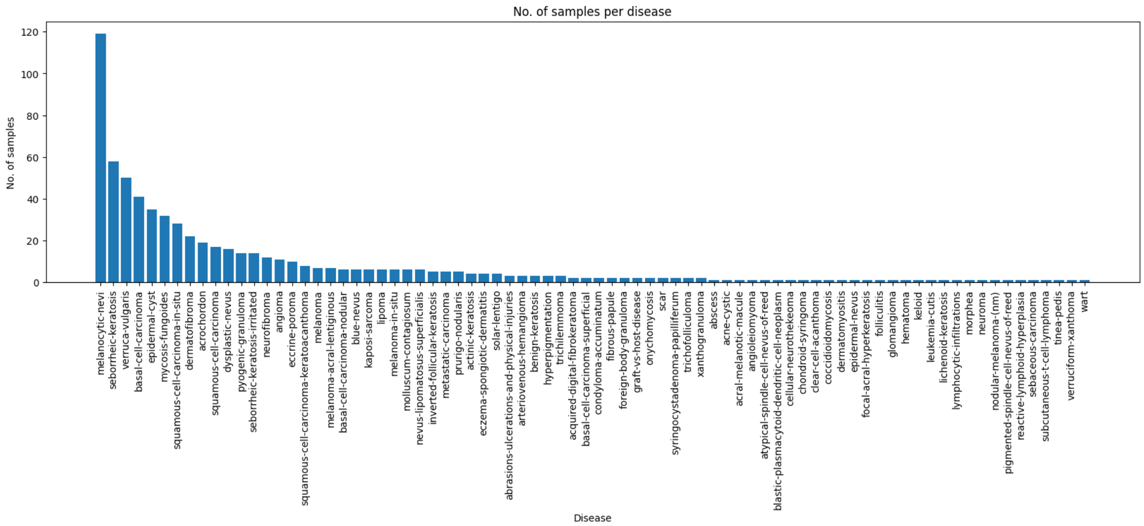
Coming the the outliers in the dataset. There are significantly less outliers in the dataset. We reffered to the guidlines for removing the outliers, and we came to conclusion that removing outliers for this dataset would be redundant and would result in a loss of valuable data, as the dataset contains a limited number of rows for the model to train upon. Moreover, upon careful observation, the outliers are formed for the catagorical columns which might be due to the presence of 0 class in majority. Hence we cannot simply eliminate the other classes considered them as outliers. Please refer to following guidelines Click here.



Dataset 2

Dataset 2 contains 656 image samples. Assigned to one of the 78 classes of skin disorders.

For image data it is usefull to check the no of samples per class.



From the plot we can see that **melanocytic-nevi** has the highest number of samples and **wart** has the lowest number of samples. This will effect out models learning. The model will be able to detect the melanocytic-nevi more accurately than other disorders.

To avoid this we need to apply image transormations while training in order to increase the number of samples of the classes.

Data

Dataset 1

Cleaned Data: https://github.com/CS418/group-project-team-teradata/blob/main/cleaned_dataframe.csv

Original Data: https://github.com/CS418/group-project-team-teradata/blob/main/data/dermatology_data.data

Dataset 2

Cleaned Data: https://github.com/CS418/group-project-team-teradata/tree/main/data/ddi_images

Original Data: https://github.com/CS418/group-project-team-teradata/tree/main/data/image_data

ML/Stats

Dataset 1

The Machine Learning model used for this dataset is Classification. We used,
Logistic Regression(Can also use for classification)
Random Forest Classifier

Support Vector Machines(Linear)
for our dataset. While using the Logistic Regression we came across a problem of scaling the. For implementation of Logistic Regression we prepared a pipeline. In the pipeline itself we incorporated Standard Scaler and Logistic Regression. Data which is not scaled enters the pipeline and data gets scaled and trians the Logistic Regression Model. Trained the models with all attributes, just the clinical attributes and just the histopathological attributes.

We also considered another Machine Learning model for this dataset, i.e. Clustering. We used,
K-Means clustering
MiniBatchKMeans
Affinity Propagation
GaussianMixture.

Dataset 2

The Machine Learning model used for this dataset is CNN. There are 3 architectures which were used in the training process. ResNet, EfficientNet, VGGNet. EfficientNet and VGGNet are shallow networks so they fail to extract deeper features from the image. ResNet on the other hand has a residual block which aids the learning process. Here the dataset is skewed towards only some classes (Please refer the EDA section). To eliminate the skewness we introduced zoom_range of 0.2 for in the ImageDataGenerator. But this will not be of a much use because the skin disease is area specific and random zooming may result in the loss in data of the spot where the skin disease exists.

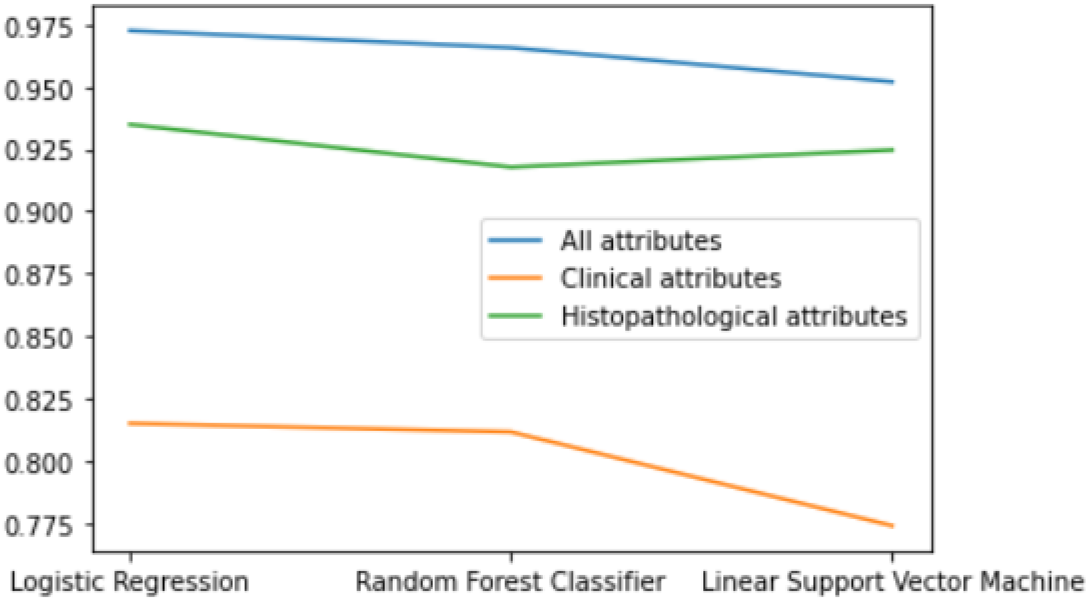
Results

Dataset 1

Classification :

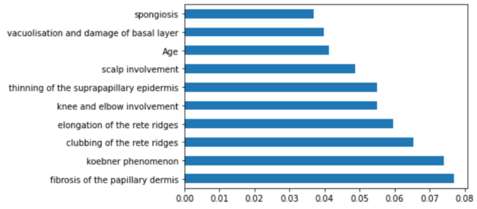
1. Trained with all attributes
   We trained the models with all the attributes and found out that LogisticRegression model performed better than Random Forest Classifier and Linear Support Vector Machine.
   Here I am not specifying any specific values, since they might be different in the current python notebook in github.
2. Trained with Clinical attributes
   We isolated the clinical attributes and trained the models with the new dataframe.
   The performance of the models dropped compared to the performance when trained with all the attributes but the ranking of the models according to the performance remained same (most of the time), i.e. LogisticRegression model performed better than Random Forest Classifier and Linear Support Vector Machine
3. Trained with Histopathological attributes

We isolated the histopathological attributes and trained the models with the new dataframe. The performance of the models dropped compared to the performance when trained with all the attributes but the ranking of the models according to the performance remained same (most of the time), i.e. LogisticRegression model performed better than Random Forest Classifier and Linear Support Vector Machine

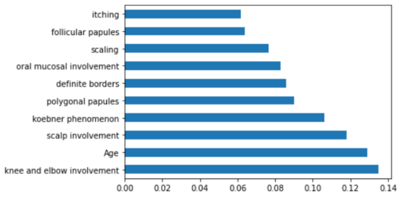


The performance of models is best when trained with all the attributes and worst when trained with just the clinical attributes.

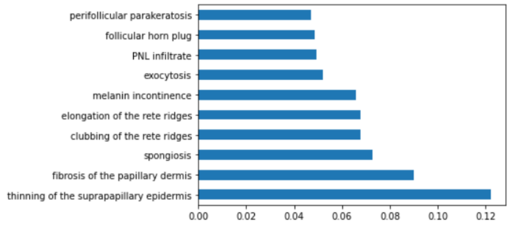
Further we analyzed the 10 most important features when trained with all features. Most of the time in those features we observed that most of the times 3 of them are clinical attributes and 7 of them are histopathological attributes.



Further we analyzed the 10 most important features when trained with clinical features.



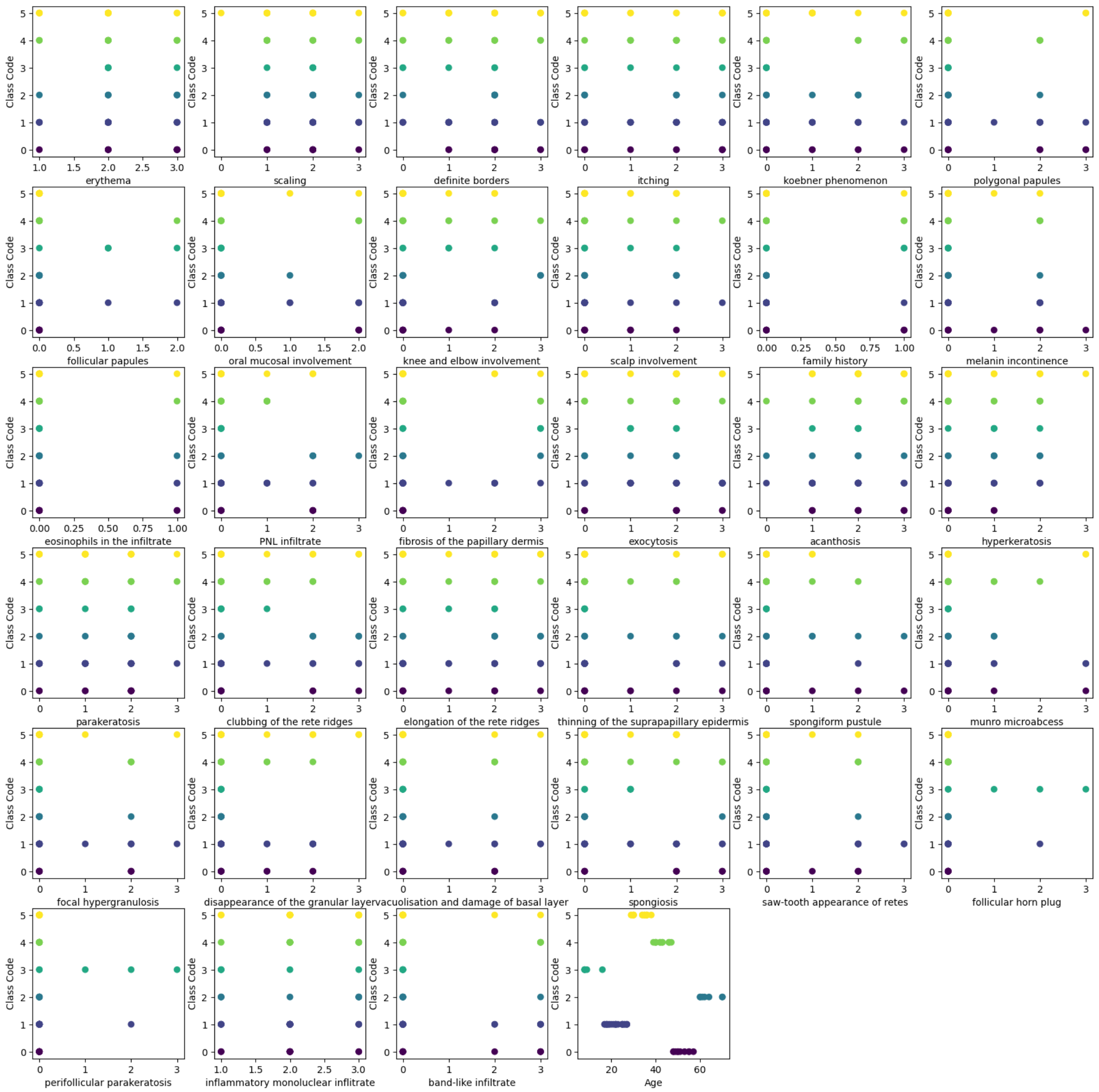
Further we analyzed the 10 most important features when trained with histopathological features.



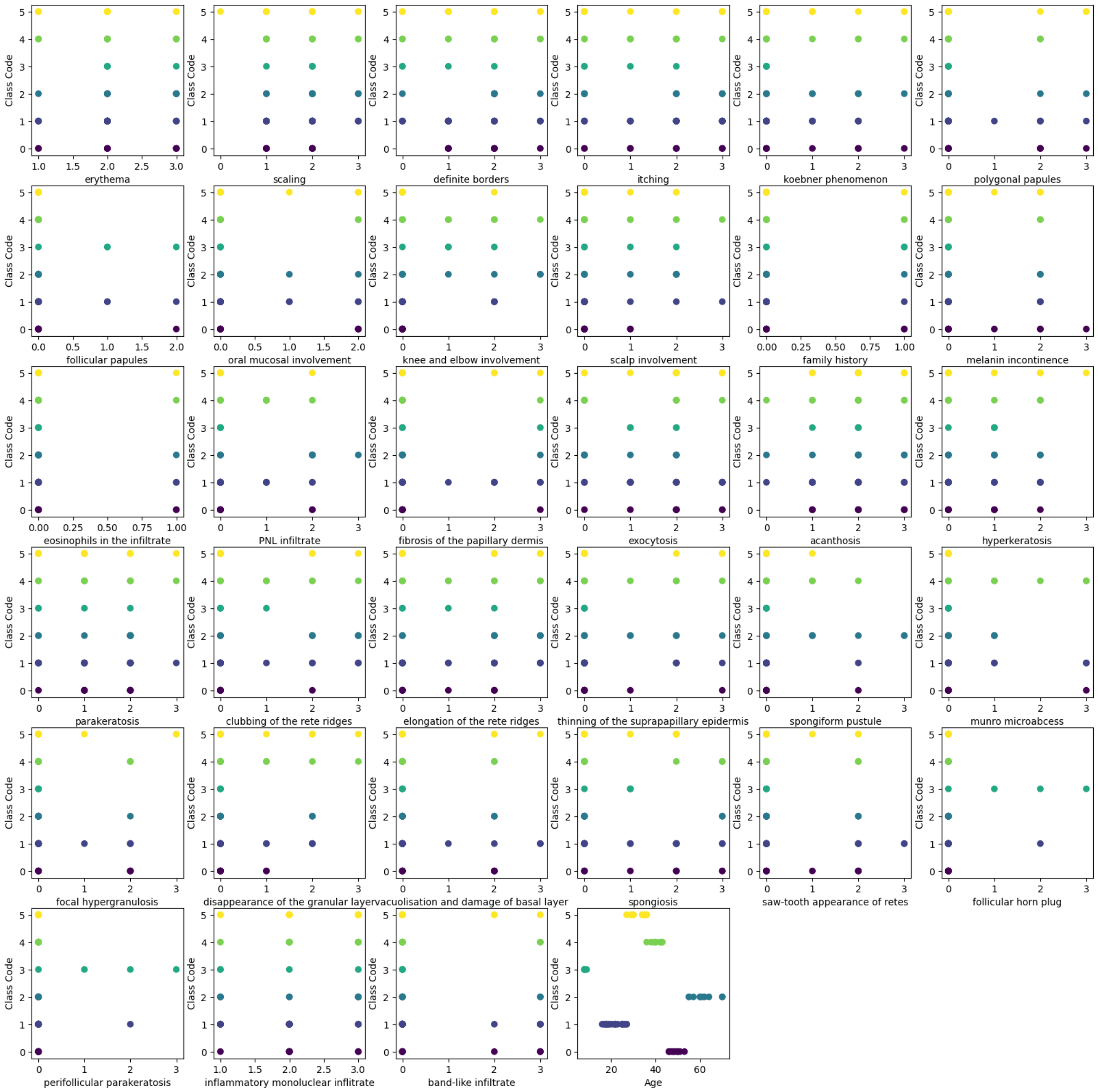
Clustering

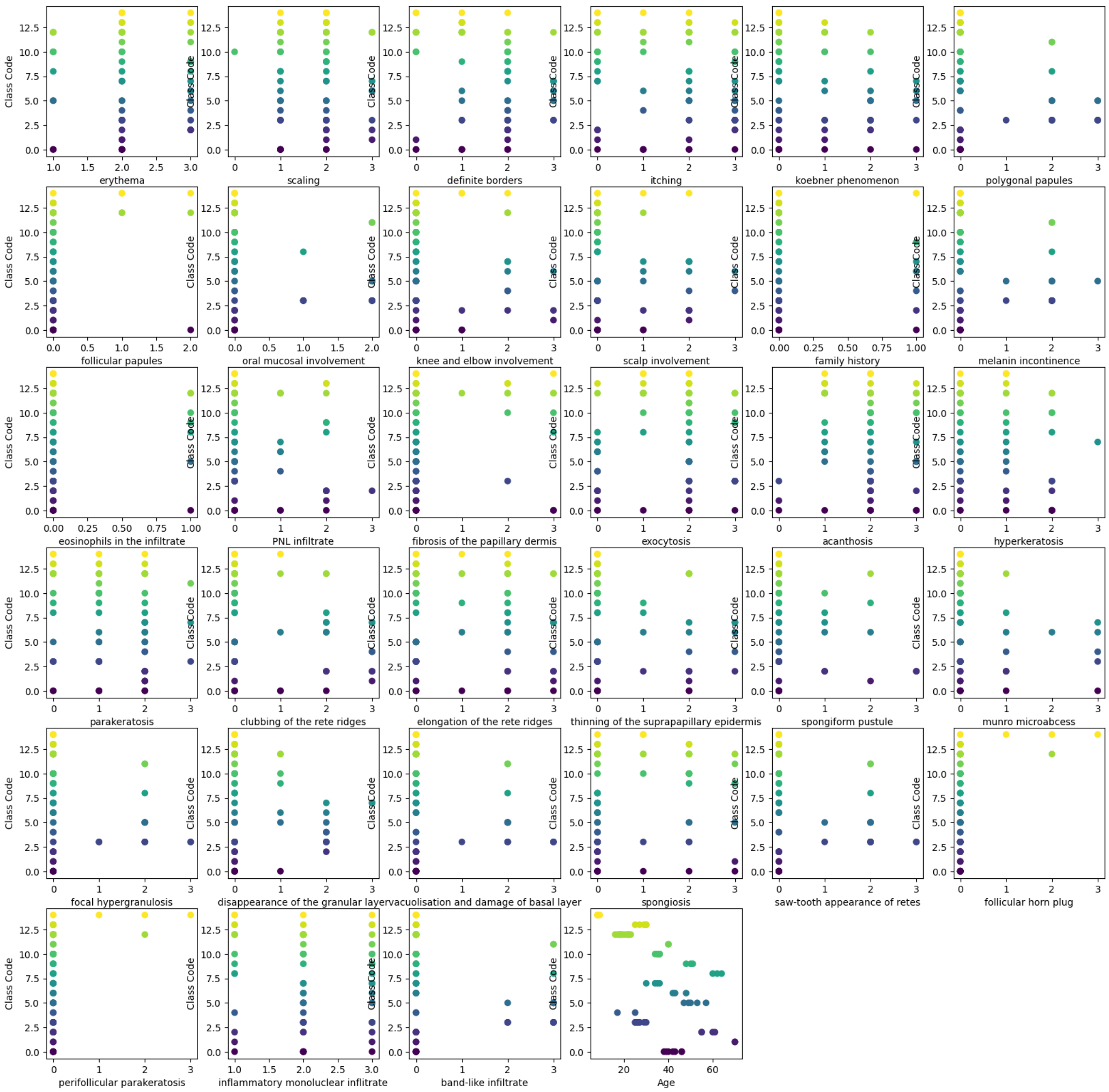
We also wanted to explore the unsupervised learning and compare the model performance to that of classification (supervised approach). As the number of rows in the datset are limited the clustering the models performed poorly when compared to the classification models. Below are the clusters given by the Kmeans, MiniBatchKMeans, AffinityPropogation, GaussianMixture. KMeans and MiniBatchKMeans gave best results while training and from the plots we can see that they form very similar clusters.

KMeans clusters

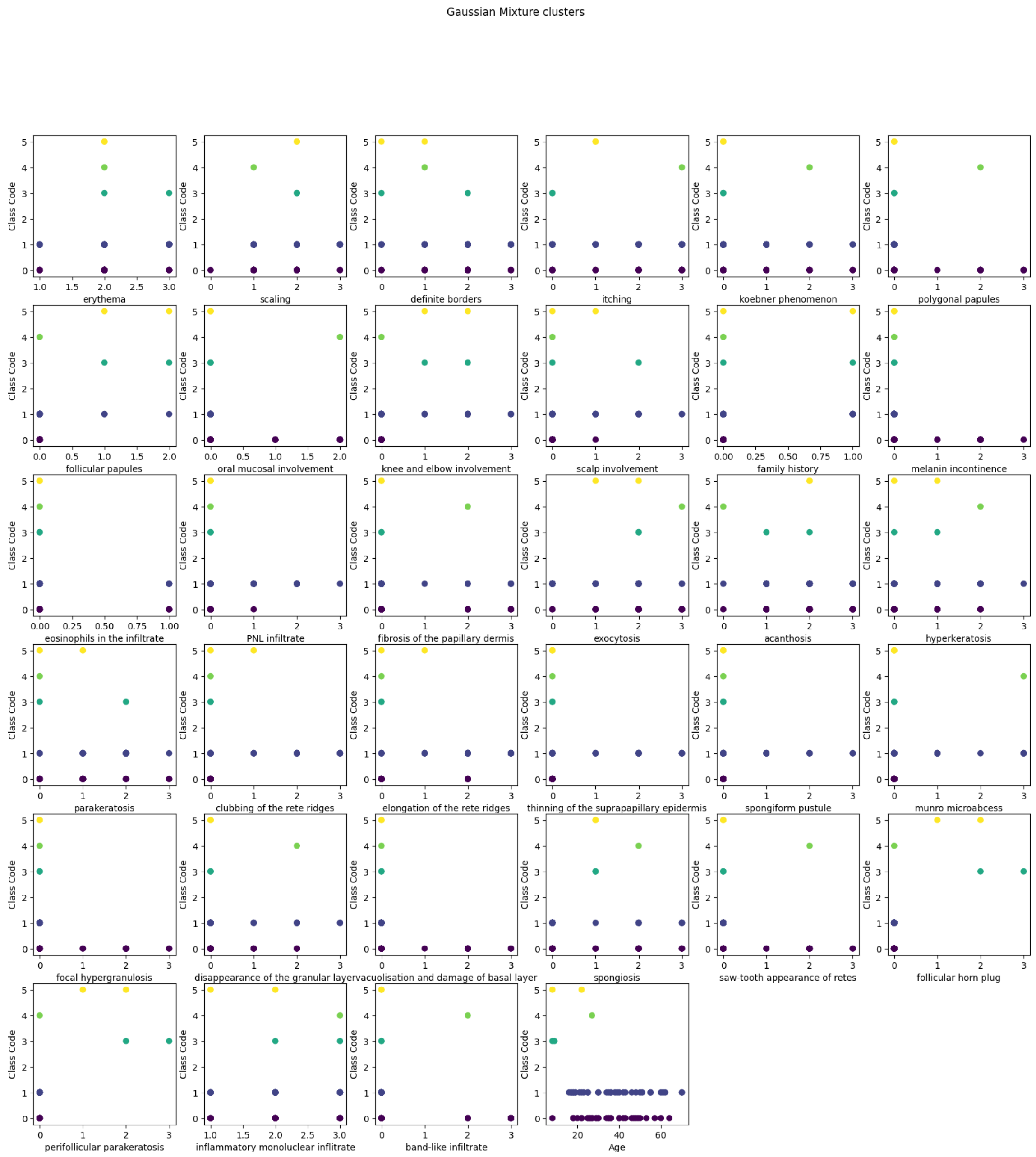


MiniBatchKMeans clusters



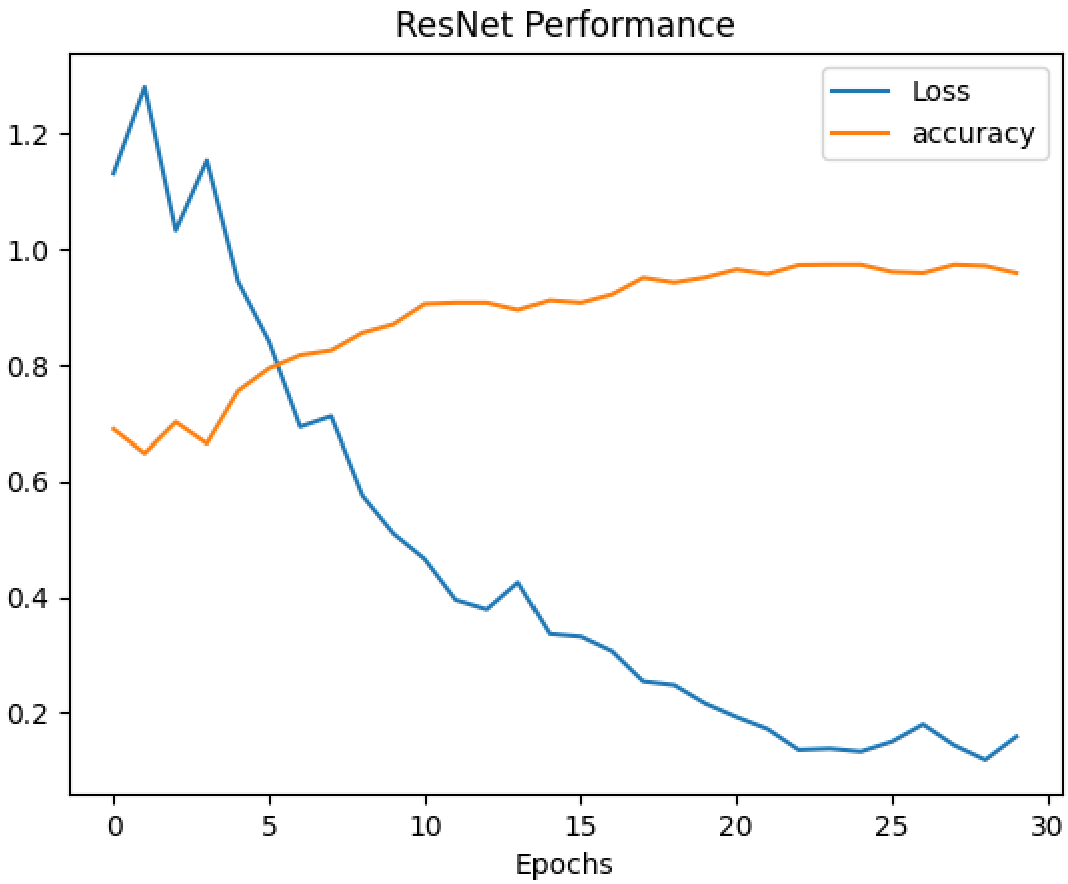


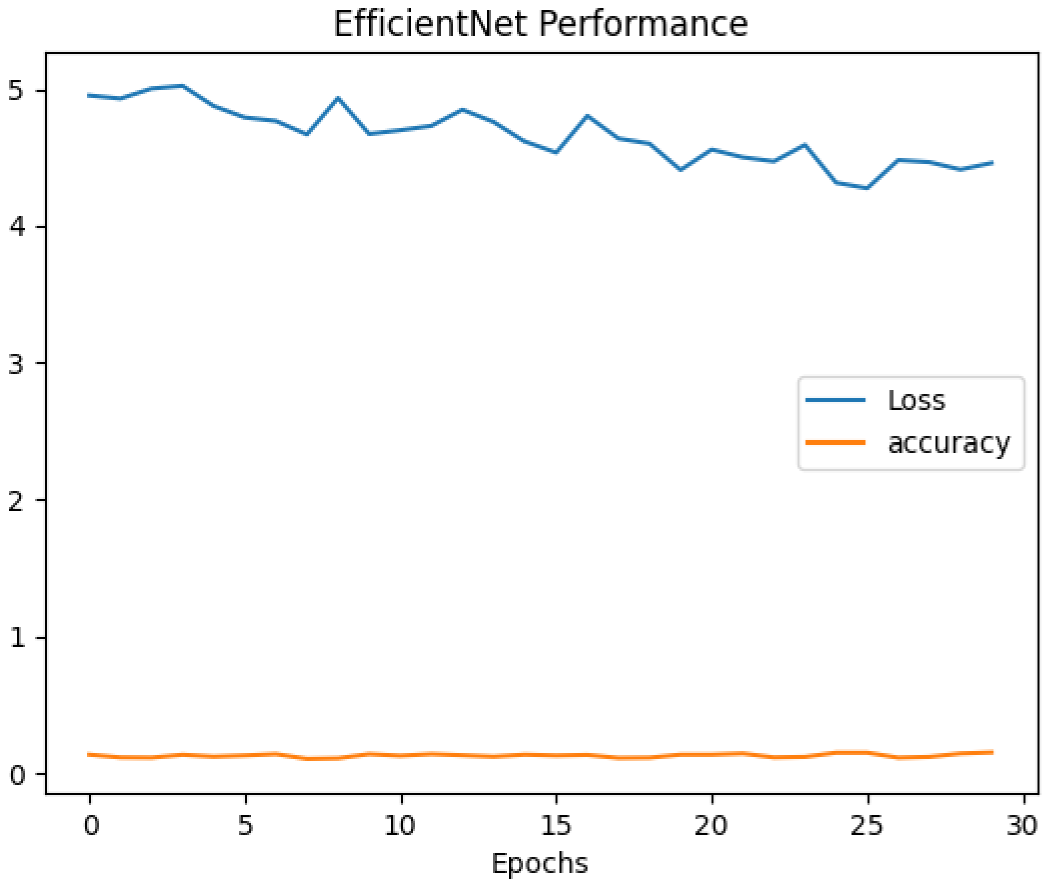
AffinityPropagation clusters

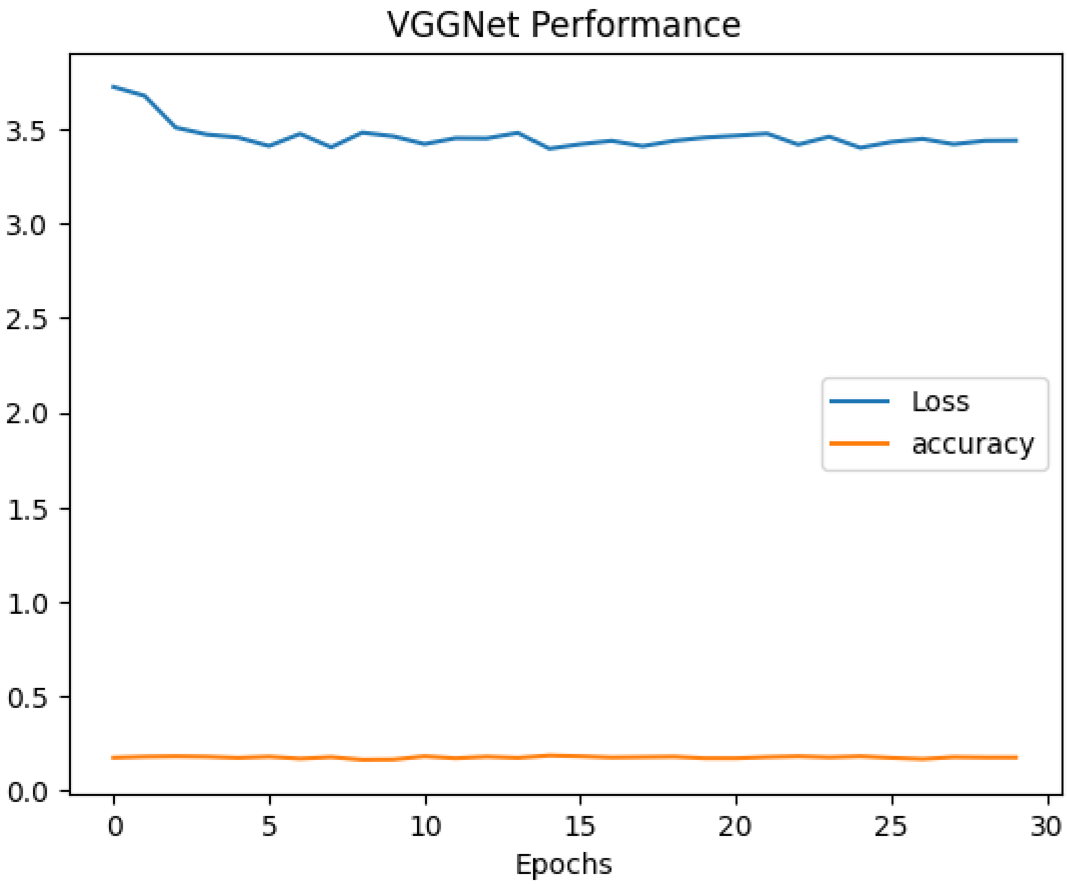


Dataset 2

As mentioned in the ML/Stats section Resnet perfomed best with an accuracy of 85%. Below are the plots of model performance. The metrics considered are the accuracy and loss.


ResNet Performance


EfficientNet Performance



References

https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
https://www.datacamp.com/blog/classification-machine-learning
https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/