Progress Report for 6.862 project

Shaoling Han | shaoling@mit.edu

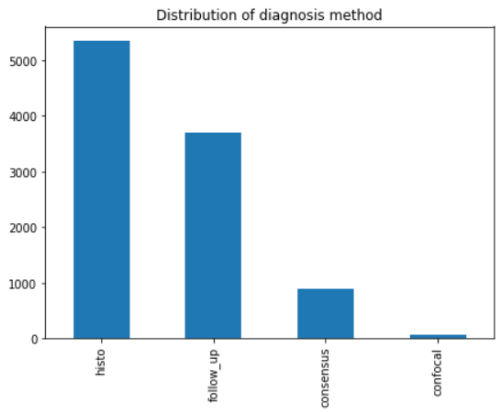
- **Introduction**

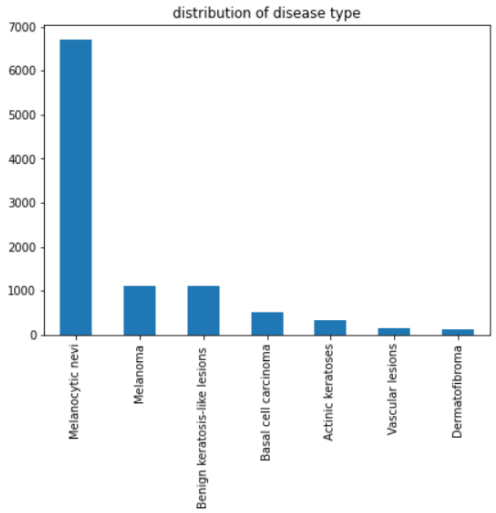
  Dermatosis is a major health burden for people nowadays. Early detection, distinguishing and diagnosis of skin cancers and benign dermatosis are of high importance. The main method of dermatosis diagnosis is based on observation of skin. This project focuses on applying machine learning methods to help diagnosis /detection of skin disease based on medical skin images and potentially other medical information.There are some previous study on this issue, mainly by using various architectures of CNN to classify the images. Amirreza Rezvantalab *et.al.* select 4 popular architectures (DenseNet 201, ResNet 152, Inception v3, InceptionResNet v2) and compae their performance on the combination of HAM10000 and PH2 dataset(1). Research from Aryan Mobiny *et.al.* also combine CNN with Bayesian methods(2). But these study use only the image data. In this project, I'll try to build some simple CNN models based on previous research and explore potential approaches to utilize medical information other than images. I will also try some non-parametric models as well.

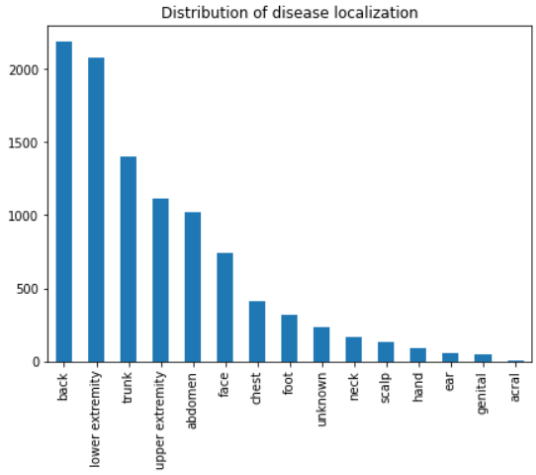
- **Data**

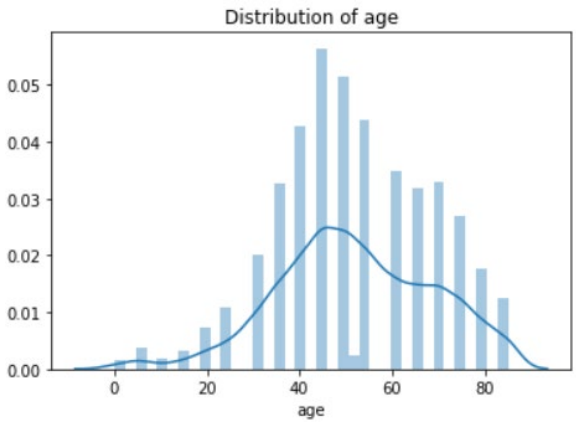
  The HAM10000 datasets ("Human Against Machine with 10000 training images") contains 10015 dermatoscopic images (600×450, about 270KB in size) and labels of these images, including patients' ID, image ID, ground truth disease classification within one of these 7 categories: Actinic keratoses and intraepithelial carcinoma / Bowen's disease, basal cell carcinoma, benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses), dermatofibroma, melanoma, melanocytic nevi and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage), method of obtaining ground truth (histopathology/follow-up/expert consensus/confocal microscopy), age, gender and location of dermatosis(3). There are 9987 patients involved, so some patients contribute more than one images to this dataset. 57 images are with missing age information. These NAs are simply replaced with the average of age. And to fit in CNN models, the images are resized to 64×48 to keep the original width/length ratio, with both RGB and grayness channel. The values of pixels are divided by 255 to keep the values within 0 to 1.

  Here is some EDA. The majority of the disease outcome is "Melanocytic nevi". The sex and age are pretty balanced among diseases. Localization seems to be related with disease type.









- **Method**

  Baseline method (RGB CNN): a convolutional neural network, with RGB 64×48 images as input, and disease type as outcome. The loss function is categorical cross entropy, which is widely used for multiclassification (there are 7 possible disease types).

  Alternative methods: 1. Gray CNN: instead of RGB images, grayscale images are used here. The other components are the same as RGB CNN. 2. Autoencoder kNN: Train a convolutional NN which takes RGB images as input, extract major features and embed into a latent space, then use information in latent space to reconstruct the original images. The loss function is MSE here. The idea is that for similar original images, they should also be close in the latent embedding space. With the embedding in the latent space, k-nearest-neighbor is used to predict the most likely disease type.

- **Results**

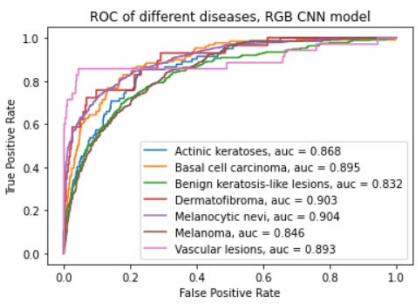
  Tensorflow 2.0 is used for CNN and scikit-learn is used for kNN. The train/test split ratio is 3:1.

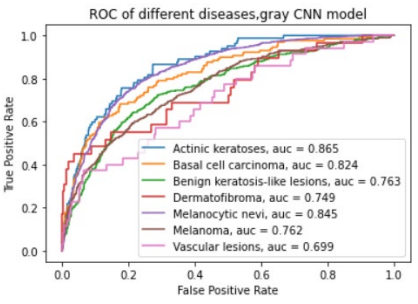
  RGB CNN: the model consists of "2 Conv2D layers, batch noralization, maxpooling" for 3 times, and 3 dense layers after flatten layer. The number of filters is 64,64,128,128,128,128, and all use "same" padding and "relu" activation. The dense layers have 128,64,7 units, and the final layer use "softmax" activation. Adam is used for compiling. The model is trained with batch size 32 and at most 100 epochs.

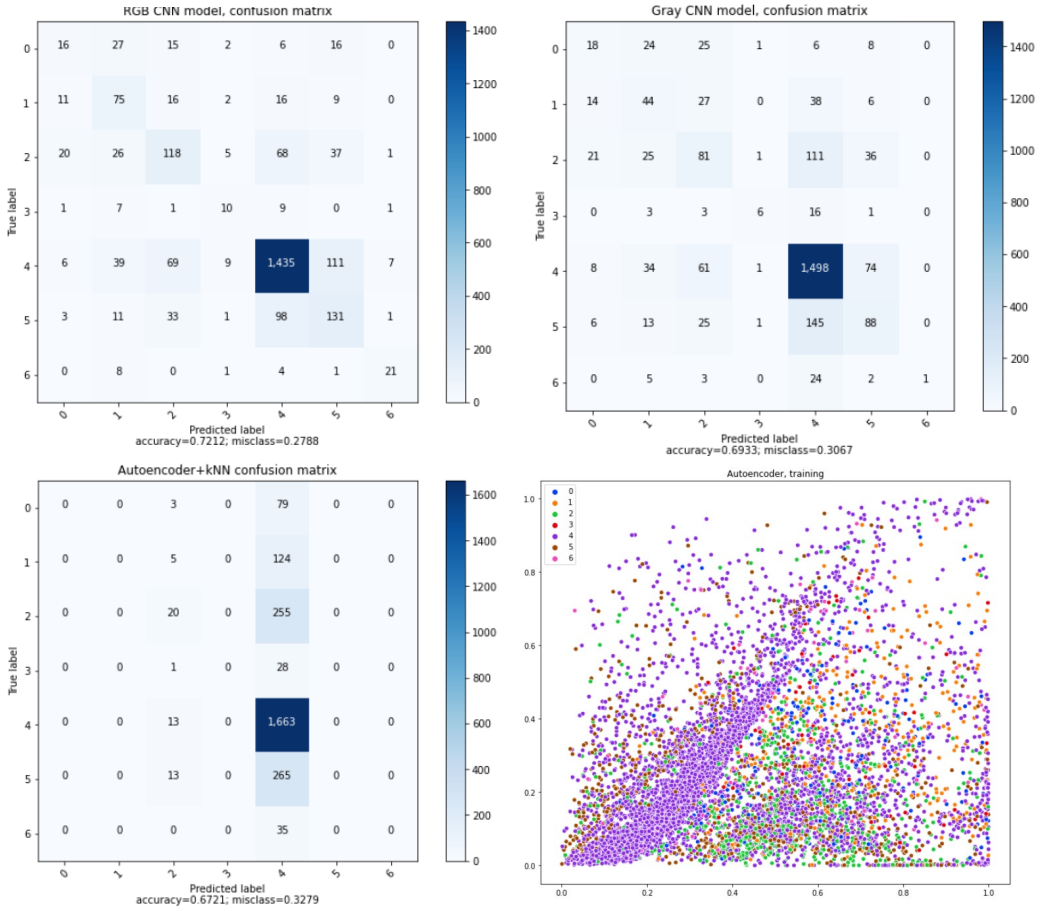
  Gray CNN: almost the same as RGB CNN, but with different number of channels in the input layer. Use the same hyperparameter and procedure as RGB CNN.

  Autoencoder+kNN: instead of having 7 units in the final dense layer, it has 2 units with activation "sigmoid" (to keep the values within 0 to 1) as the embedding in the latent space, and reverse the whole network to reconstruct the original 64×48 images. RMSprop is used in compiling. The batch size and epochs are the same as above. Retrieve the outcome in the latent space and use 5-fold crossvalidation to find the optimal k value.

  Using GPU in Google Colab, it takes about 10 mins to train each model. The RGB CNN model has 72% accuracy in validation data, while gray CNN gives 69%. K=200 is selected after cross validation in autoencoder-knn model, and it gives 67.2% accuracy.







From the ROC curve, the RGB CNN model is better than gray CNN model. And the autoencoder model gives lowest accuracy. The heatmaps of confusion matrix show the distribution of prediction. Because from EDA the dataset is unbalanced (most images are from Melanocytic nevi), it may affect performance of the models. The autoencoder model even gives prediction only on Melanocytic nevi and benign keratosis-like lesions. The scatter plot of the latent space coordinate also shows that it is not separating different disease well.

- **Discussion**

  Given that the dataset is unbalanced, it may help to 1. Upsample the minority; 2. Use data augmentation to make more images for the minority; 3. Use weighted loss function for regularization.

  The images can also be resized to different size, and different size contains different amount of information. It is expected that with less loss of information, the model can perform better. It is consistent with the results that RGB (more information) is better than grayscale. The baseline method performs as expected, giving acceptable accuracy. More CNN architectures will be used to test performance.

  The autoencoder model could be used to combine information from images (in the latent space) and other information like age, disease localization etc. by concatenating these vectors and use kNN. Yet the dimension of latent space will need careful tuning. Clustering is of less practical use in clinical setting, so I will focus on classification.

- **Plan**

  1. Try different resize ratio to see whether it influences model performance.
  2. Try different CNN architectures to achieve higher accuracy
  3. Try data augmentation/weighted loss function to deal with the unbalanced dataset.

  Due to the COVID19 issue the project didn't stick with plan, but I will adjust accordingly.

- ## Reference

1. Amirreza Rezvantalab HS, Somayeh Karimijeshni. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms 2018. Available from: https://arxiv.org/abs/1810.10348.
2. Mobiny A, Singh A, Van Nguyen H. Risk-Aware Machine Learning Classifier for Skin Lesion Diagnosis. Journal of clinical medicine. 2019;8(8).
3. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T.