

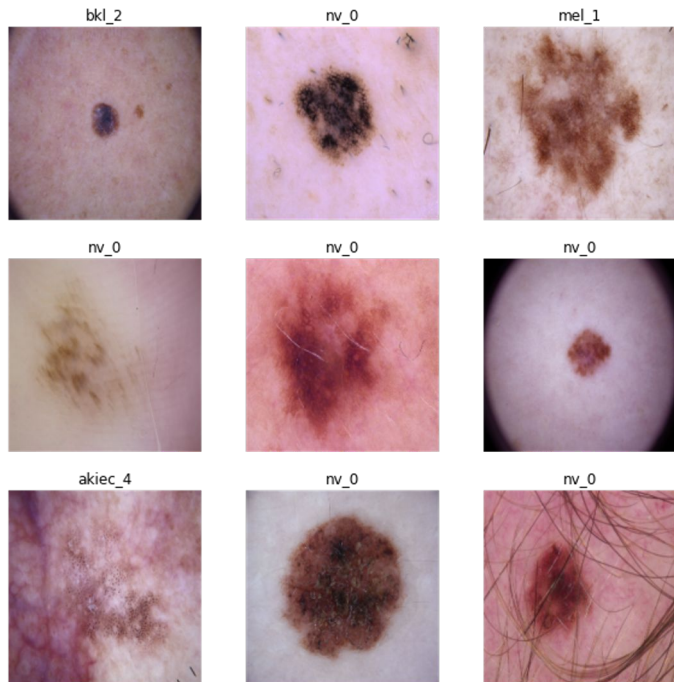
More than a mole?

Digging into classification of skin lesions
Using Spark

Goal of project

- **Deployment of neural networks to featurize images for machine learning.**
we explored the ability of using PySpark's deep learning, and Tensorflow's deep convolutional neural networks, especially ResNet50, to featurize dermoscopic images of common pigment skin lesions.
- **Using spark to develop a pipeline for efficient machine learning for image classification**
We tried 3 Spark ML algorithms (Logistic Regress, Random Forest, Gradient Boosted Trees) to examine multi-class image classification on the Spark cluster.

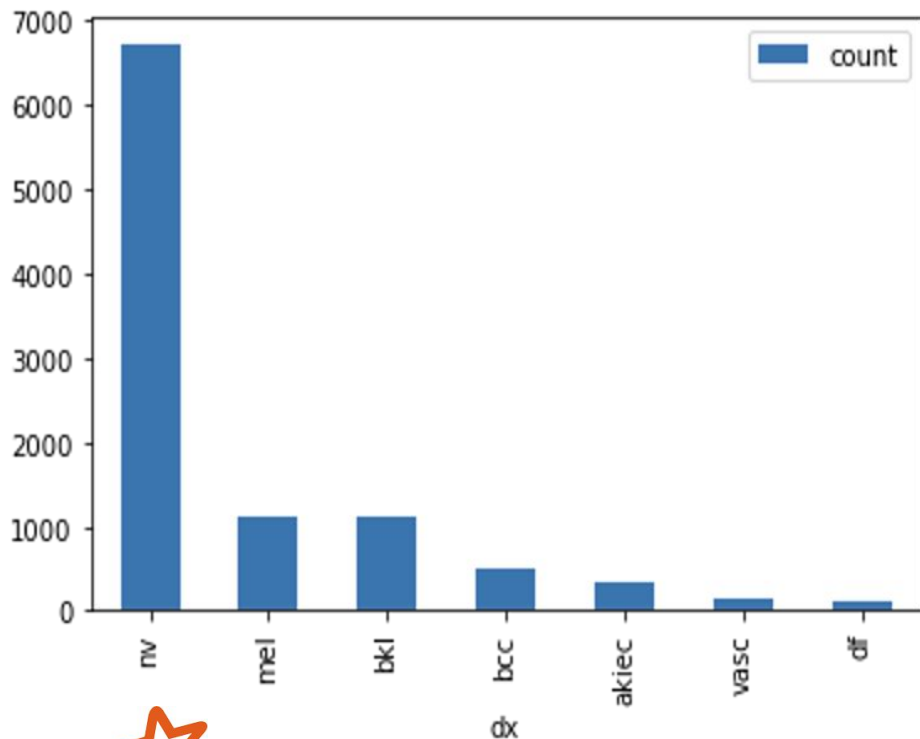
Dataset



Skin Lesion classifications(7 class)

- melanocytic nevi (nv)
- benign keratosis-like lesions (bkl)
- Bowen's disease (akiec)
- basal cell carcinoma (bcc)
- dermatofibroma (df)
- melanoma (mel)
- vascular lesions (vasc)

EDA



Total image number: 10015

After removing null and unknown

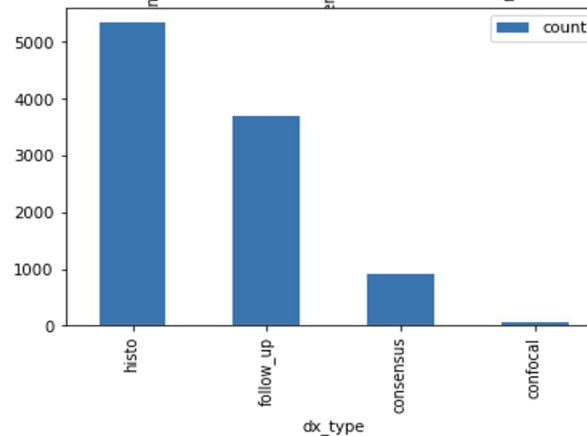
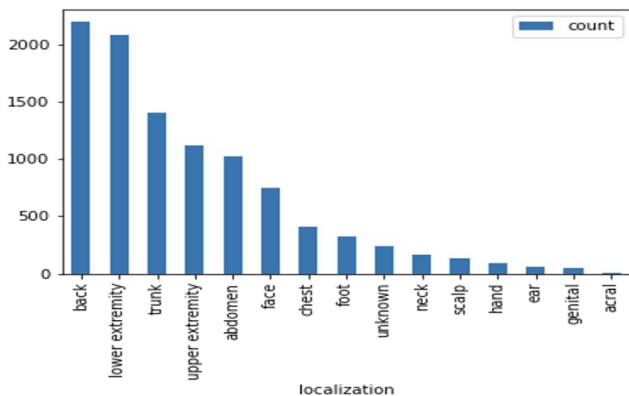
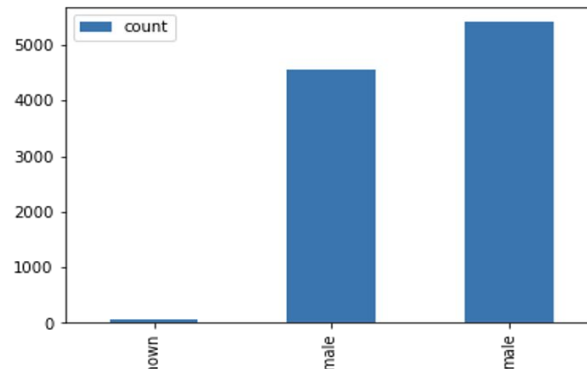
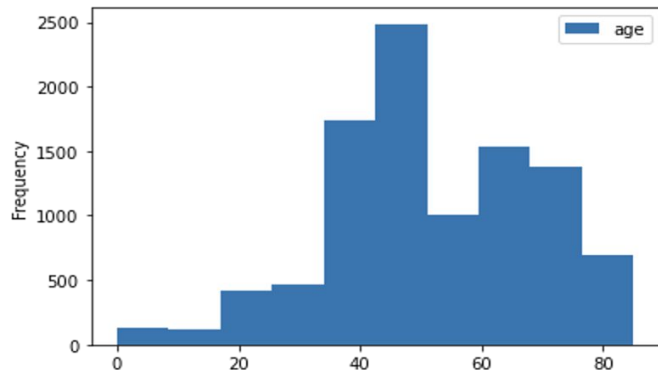
Total image number: 9948

Benign and malignant skin lesions

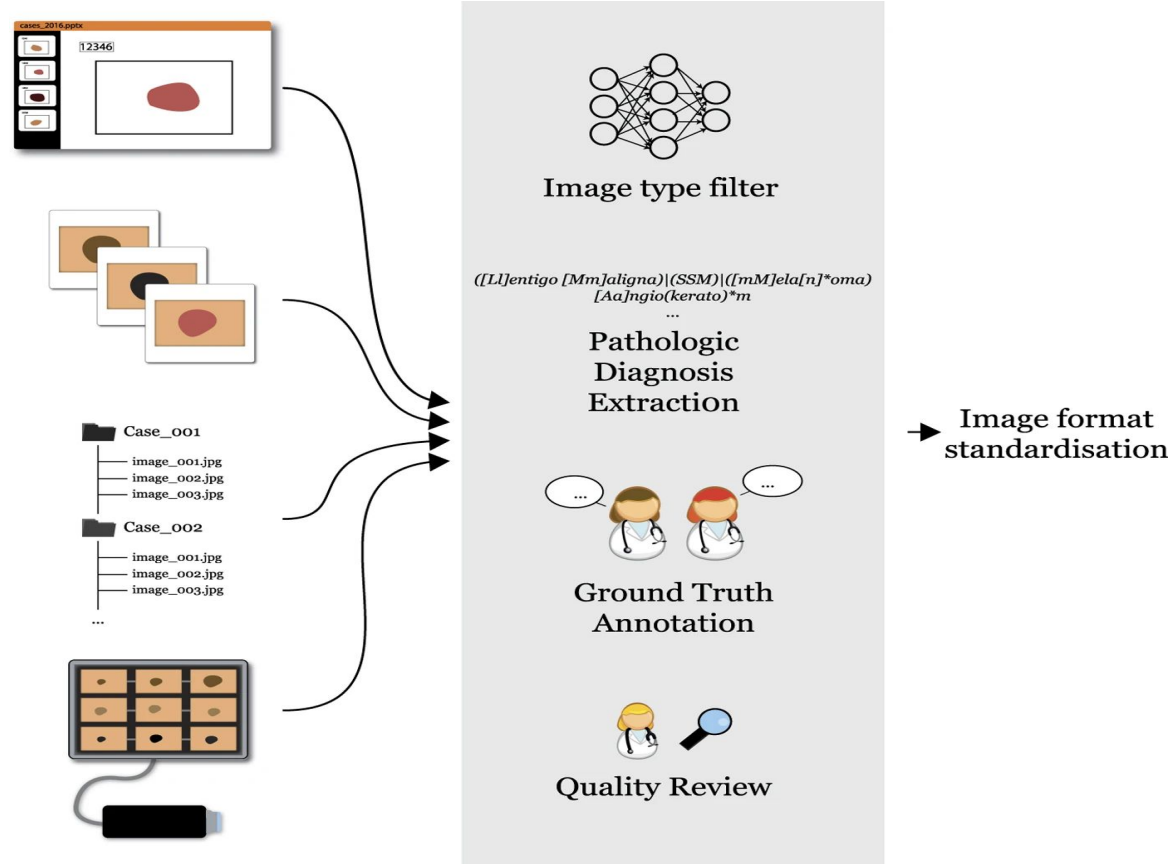
-imbalanced

-mostly benign

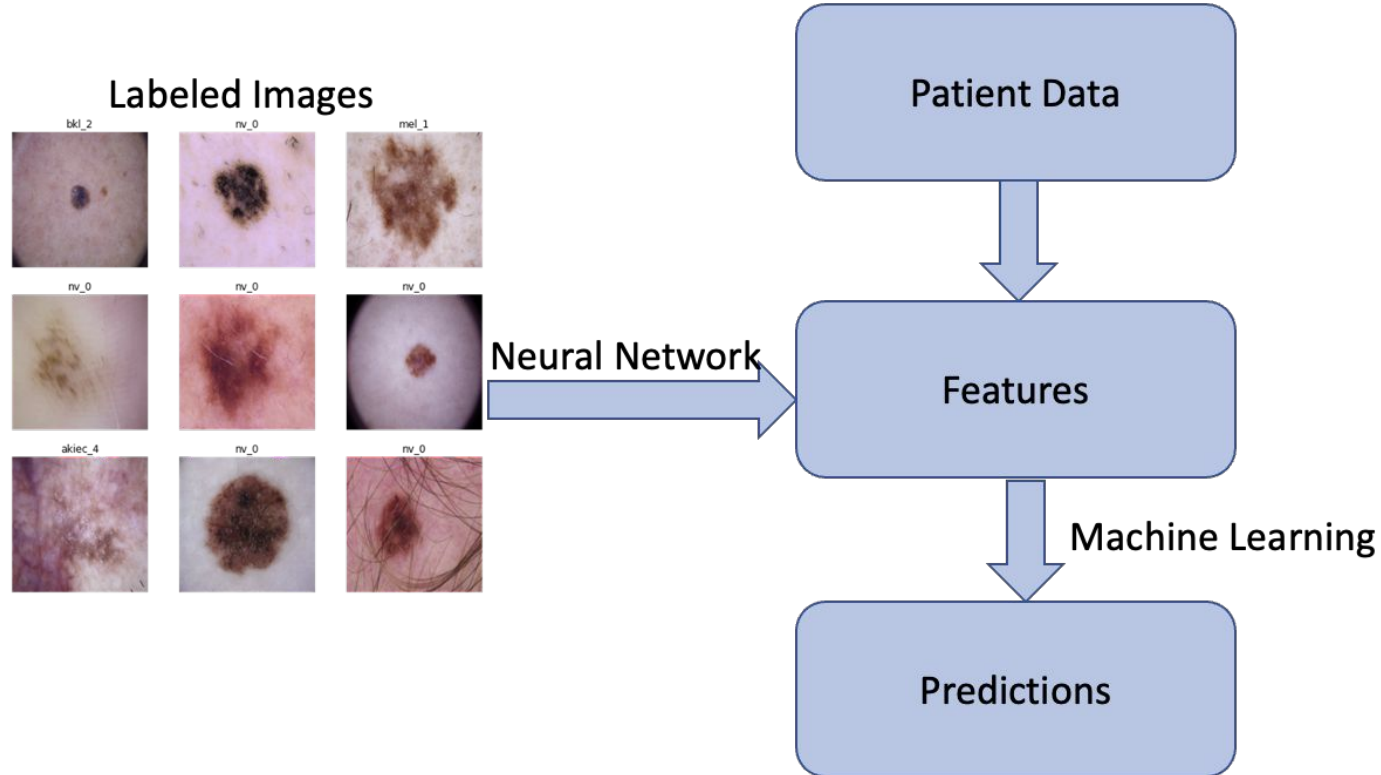
EDA (patient information)



Procedure

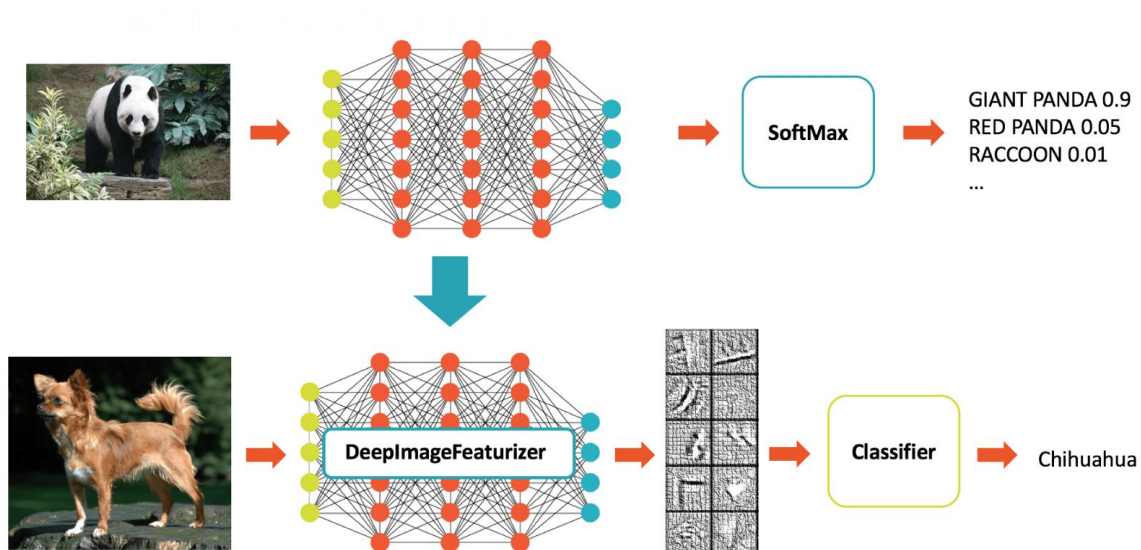


Workflow



Challenges

Reducing dimension? PCA, TL
Transfer Learning (image \rightarrow vector)?
Which models?



Transfer Learning Model Architecture

Model: "model"

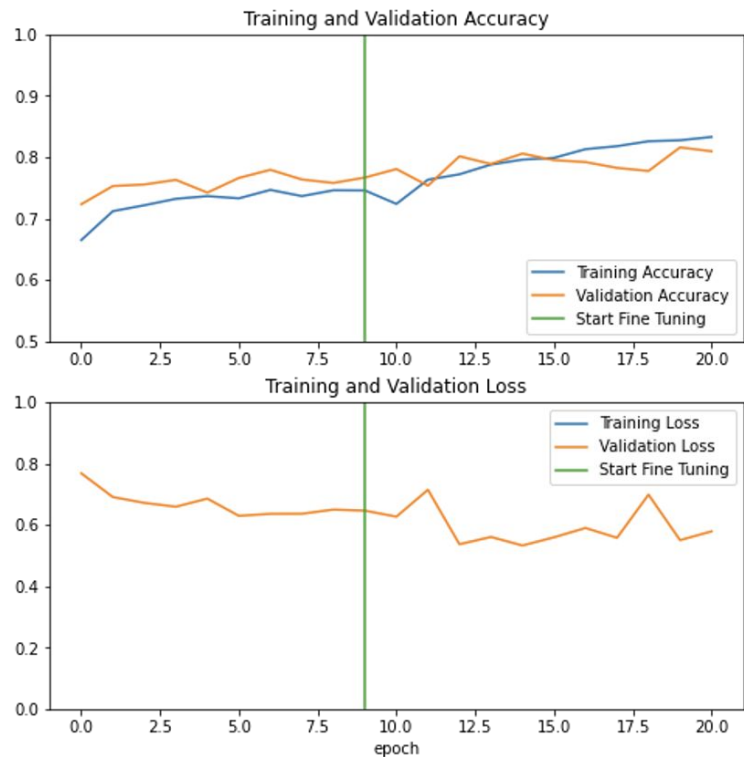
| Layer (type) | Output Shape | Param # |
|------------------------------|-----------------------|----------|
| input_3 (InputLayer) | [(None, 224, 224, 3)] | 0 |
| sequential (Sequential) | (None, 224, 224, 3) | 0 |
| resnet50 (Functional) | (None, 7, 7, 2048) | 23587712 |
| global_average_pooling2d (Gl | (None, 2048) | 0 |
| dropout (Dropout) | (None, 2048) | 0 |
| dense_1 (Dense) | (None, 7) | 14343 |

Total params: 23,602,055

Trainable params: 14,343

Non-trainable params: 23,587,712

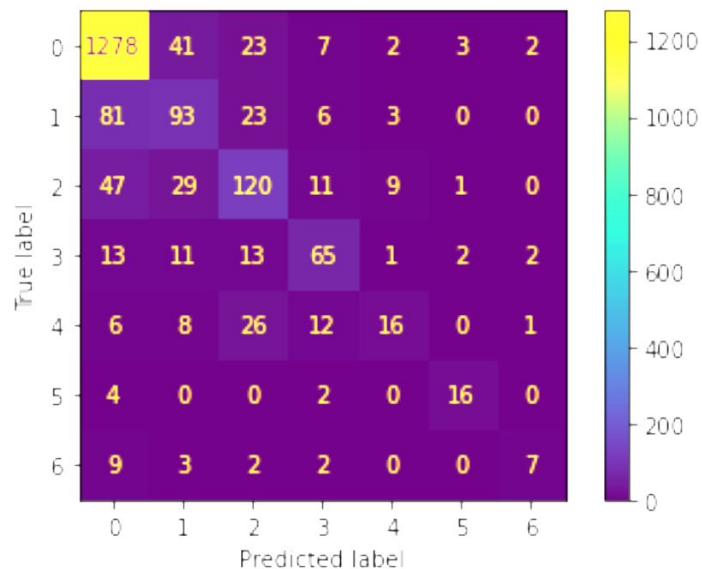
20 epochs



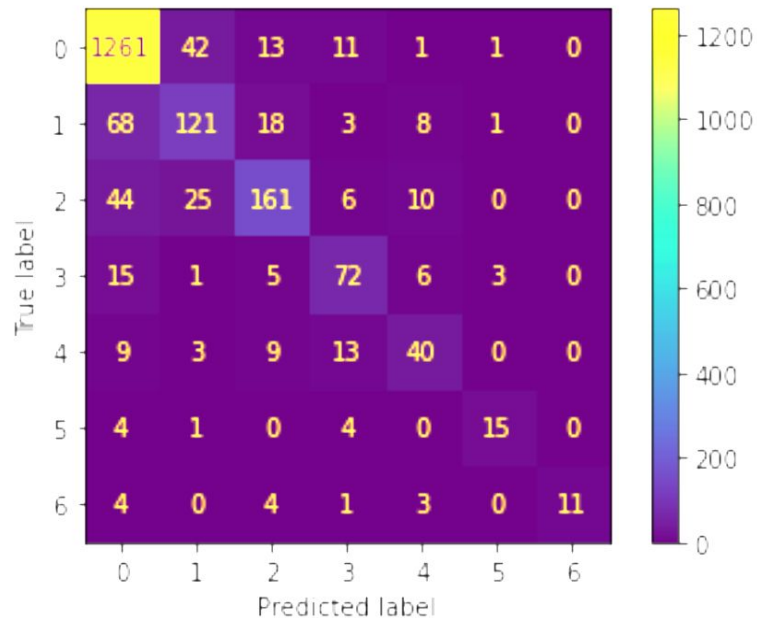
50 epochs



Model Performance

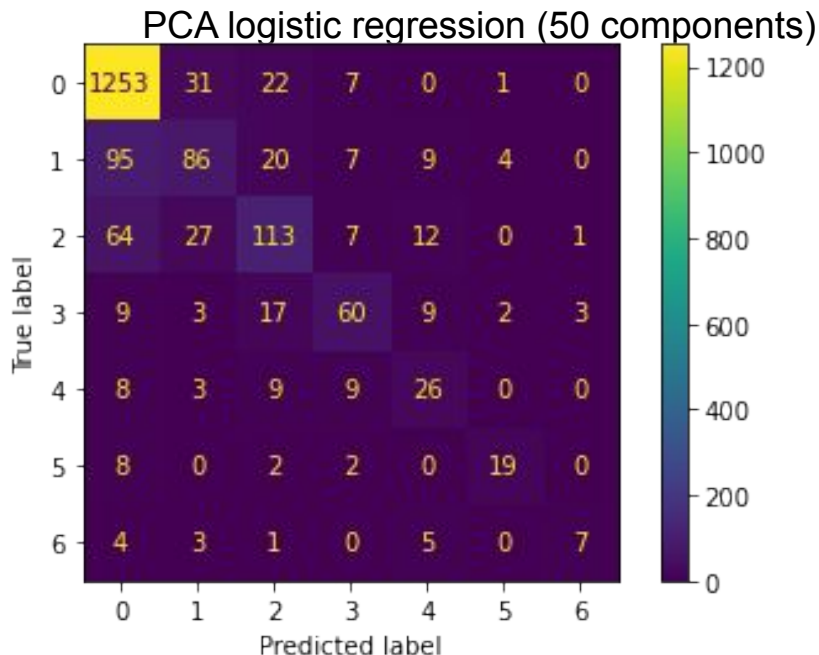


Logistic Regression,
overall accuracy: 84%
Melanoma recall: .44



Gradient Boosted Trees
Overall accuracy: 80%
Melanoma Recall .58

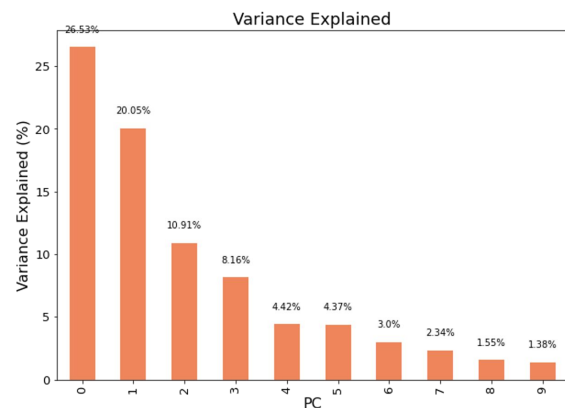
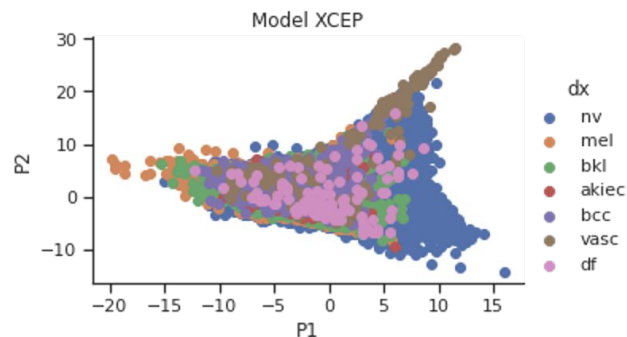
Reducing the feature set



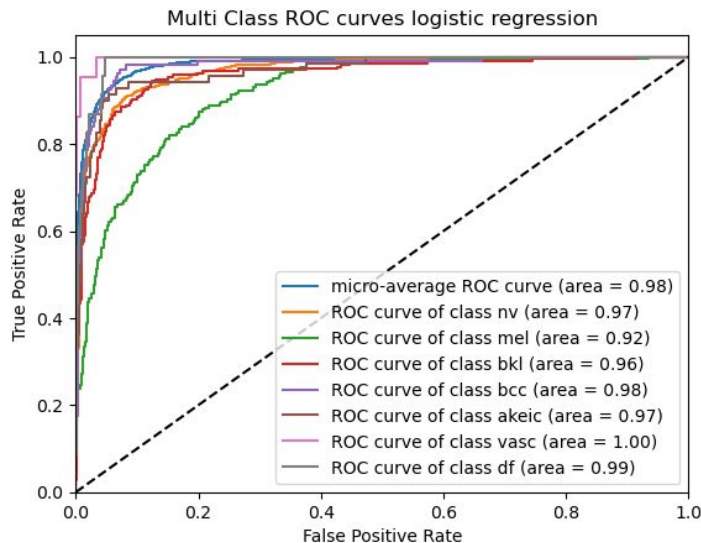
overall accuracy: 79%

Melanoma recall: .42

Balanced multiclass accuracy: 58.7%



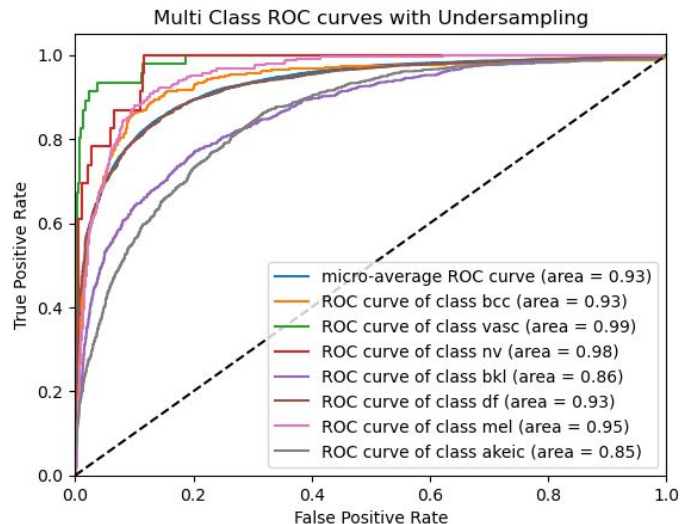
Model Performance



Overall Accuracy: 84.7%

Balanced Multiclass Accuracy: 69%

Melanoma Recall : .44



Overall Accuracy: 69.9%

Balanced Multiclass Accuracy 68.1%

Melanoma Recall : .58

Conclusions:

Comparable to these fine fellows:



| | | | | | | |
|----|--|--|--|-----|-------|---|
| 30 | BioImaging-KHU BioImaging-KHU | Deep Learning with Adapted InceptionResNetV2 | | No | 0.705 | ▼ |
| 31 | Dominiks AI team Dominiks AI team | Above dermatologist-level classification of malignant melanomas with deep neural | | Yes | 0.703 | ▼ |
| 32 | Mammoth Mammoth | Old fashion | | No | 0.703 | ▼ |
| 33 | Redha Ali, Russell C. Hardie, Manawaduge Supun De Silva, and Temesguen Messay Ke Redha Ali, Russell C. Hardie, Manawaduge Supun De Silva, and Temesguen Messay Ke | Combining Deep and Handcrafted Image Features for Skin Cancer Classification | | No | 0.701 | ▼ |
| 34 | CNR-ISASL-Lecce CNR-ISASL-Lecce | Deep Convolutional Neural Network with Stochastic Gradient Descent Optimization | | No | 0.699 | ▼ |
| 35 | Hosei University, Iyatomi lab Hosei University, Iyatomi lab | SEResNet101 w/ mean_teacher + SEResNet152 w/o mean_teacher | | No | 0.695 | ▼ |
| 36 | Manu Goyal Manu Goyal | DeepLabV3+ with Priority strategy based on benign/malignant and number of images | | No | 0.695 | ▼ |
| 37 | PA_Tech PA_Tech | deep convolutional neural network with transfer learning | | No | 0.692 | ▼ |
| 38 | Opsins Opsins | Transfer learning based CNN | | No | 0.691 | ▼ |
| 39 | UNIST_BMIPL UNIST_BMIPL | Multiscale Lesion Segmentation and Application to Skin Cancer Classification | | No | 0.687 | ▼ |
| 40 | Nitwit AI Nitwit AI | Inception V3 | | No | 0.685 | ▼ |

Follow Up

- Cluster Bottlenecks

- Multi - GPU clusters
- Single consumer GPU vs 6 CPU workers offers ~10X speedup

- Changes in technology:

- From spark_tensorflow_distributor to Horovod Runner in spark 3.1

- Spark Level data augmentation:

- implement a useable upsample function

Questions?

