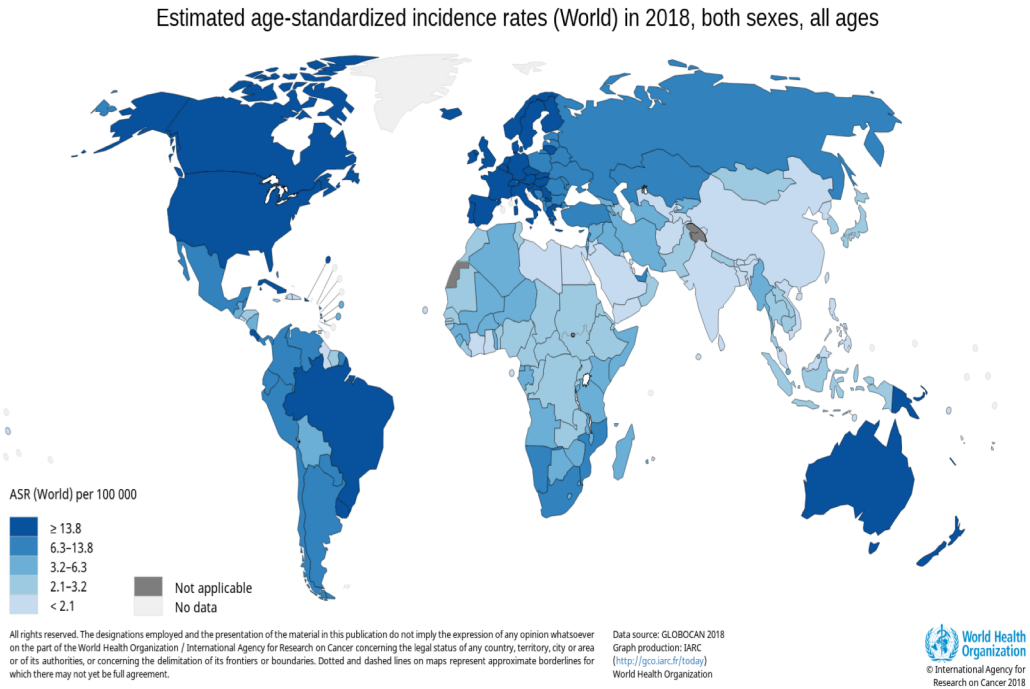
ML Based Pre-screening For Skin Cancer Detection

Pavel Golukhin, Radhika Satapathy, Tyler Ryu, Suyash Dusad

Presented to:
The Grabowski Foundation for MedTech Innovation
December 04, 2019

SKIN CANCER IS THE MOST COMMON CANCER IN THE UNITED STATES AND WORLDWIDE



Facts about Skin Cancer:

- Australia, New Zealand and the USA have the highest incidence rates (ASR) in the world with 181, 171, and 68 ASR respectively
- More than two people die of the disease every hour
- At least one in five Americans will develop skin cancer by the age of 70
- The annual cost of treating skin cancers in the U.S. is estimated at $8.1 billion

SKIN CANCER DETECTION AND TREATMENT DEMAND CONSIDERABLE AMOUNT OF TIME AND MONEY

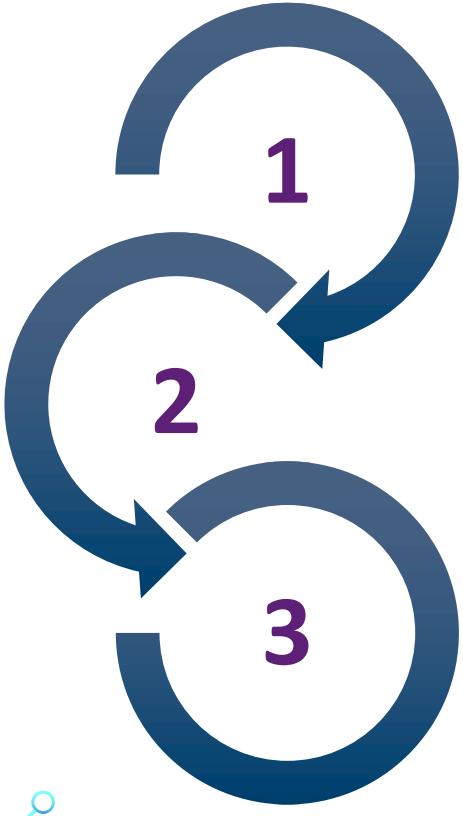
When detected early, the 5-year survival rate for melanoma is 99 percent

The treatment of childhood melanoma is often delayed due to misdiagnosis of pigmented lesions, which occurs up to 40 percent of the time.

There is around 10M people a year to conduct skin cancer tests in the US
It is recommended to conduct professional skin exam with dermatologist annually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination.

An average annual skin cancer treatment cost in the US is $8.1 billion
The annual cost of treating nonmelanoma skin cancer in the U.S. is estimated at $4.8 billion, while the average annual cost of treating melanoma is estimated at $3.3 billion

IN OUR PROJECT WE IDENTIFY THE SKIN CANCER TYPES BASED ON THE PHOTO OF THE DAMAGED SKIN AND ADDITIONAL INFORMATION



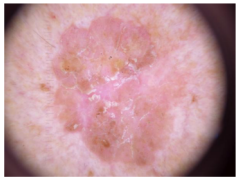
Build a simple, cost efficient tool for pre-screening lesion images

Create the practical instrument for medical practitioners

Make the approach more accessible for patients

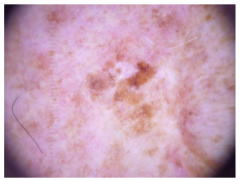
SKIN LESIONS: A CLOSER LOOK...

Actinic keratoses



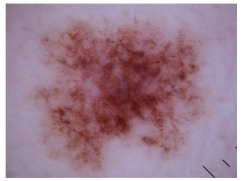
Basal cell carcinoma

Benign keratosis-like lesions

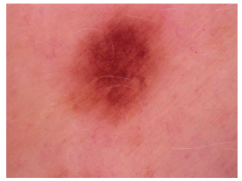


Dermatofibroma

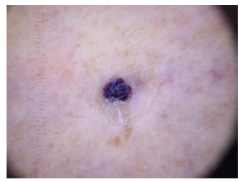
Melanoma



Melanocytic nevi



Vascular lesions



HAM10000 dataset

Primary Data:

10,015 dermatoscopic JPEG images from different populations of the following lesion types:

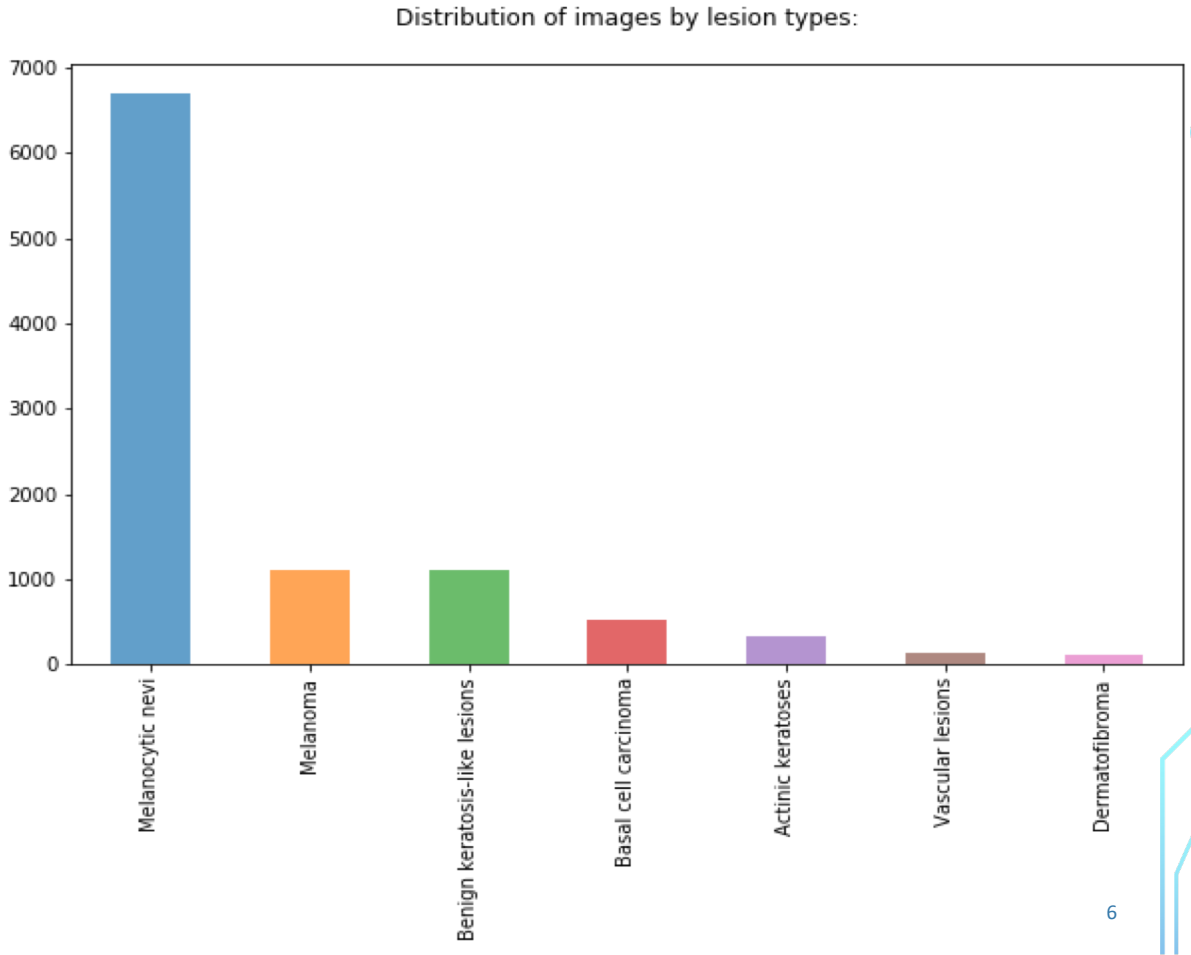
1. Melanocytic nevi
2. Melanoma
3. Benign keratosis-like lesions
4. Basal cell carcinoma
5. Actinic keratoses
6. Vascular lesions
7. Dermatofibroma

Ancillary Data:

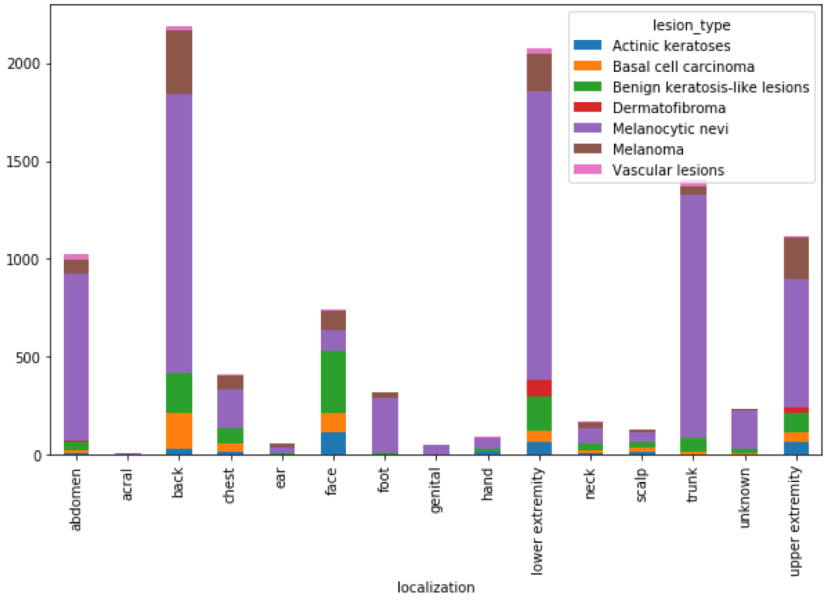
1. Lesion detection method
2. Age of patients
3. Gender of patients
4. Tumor location

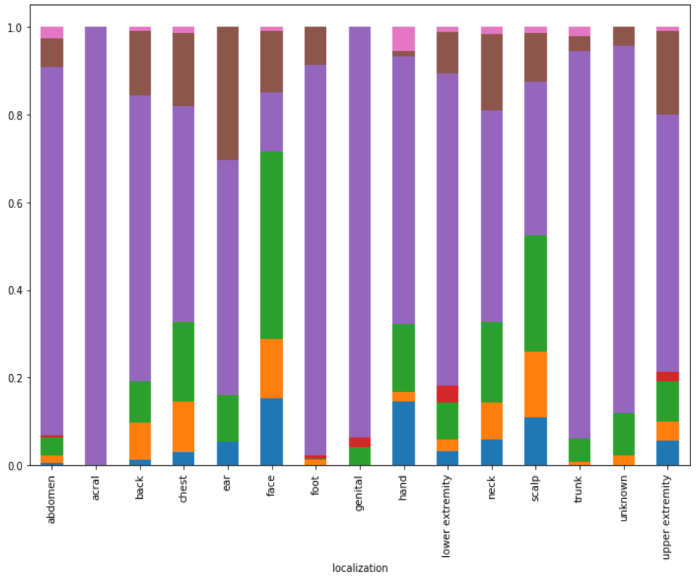
HAM 10000: EXPLORATORY DATA ANALYSIS

- ★ Sample distribution by:
  - Lesion type
  - Localization
  - Age
  - Sex of the patient
  - Validation technology
- ★ Multiple images

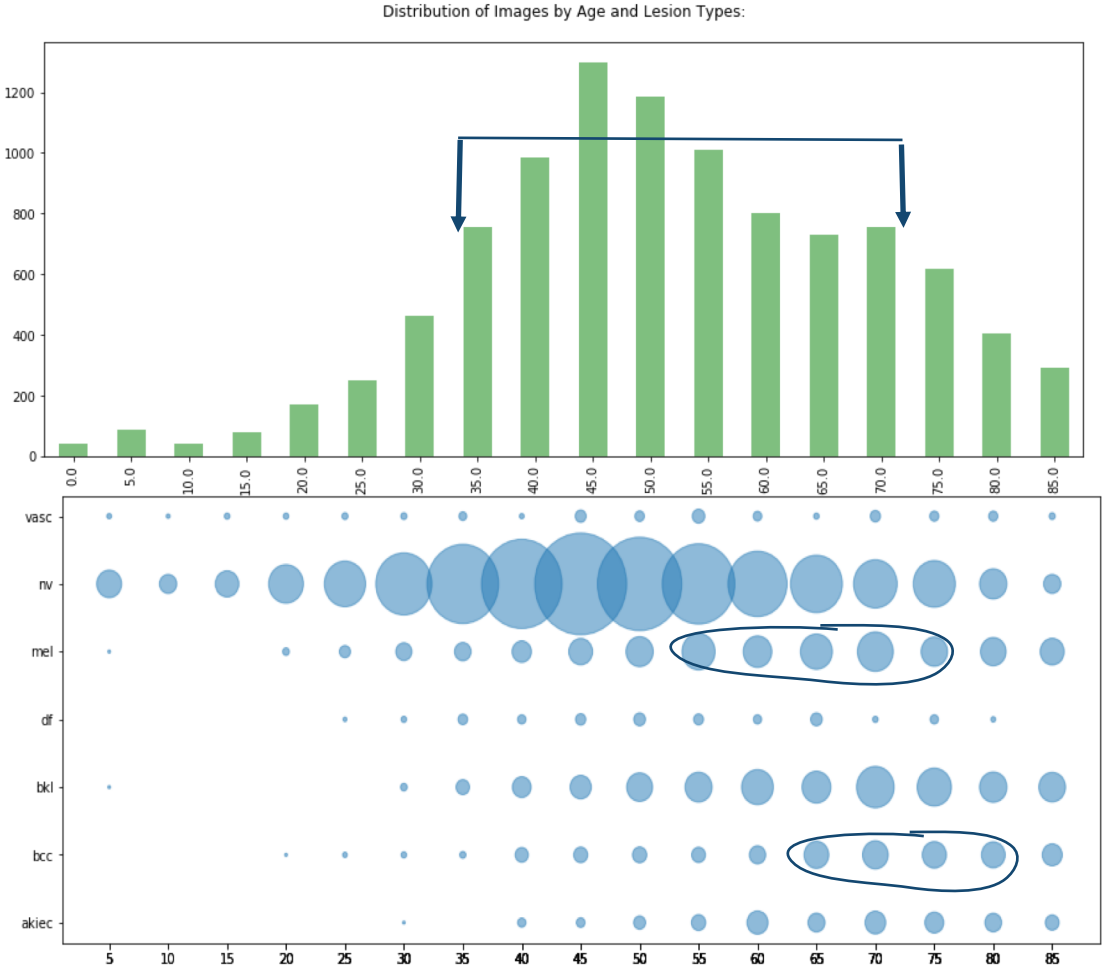


Distribution of Images by Localization...



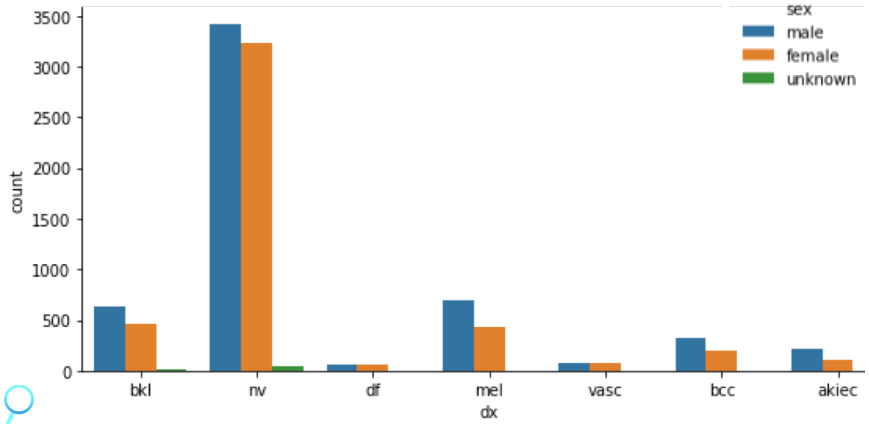


DISTRIBUTION OF IMAGES BY PATIENT AGE CATEGORY...

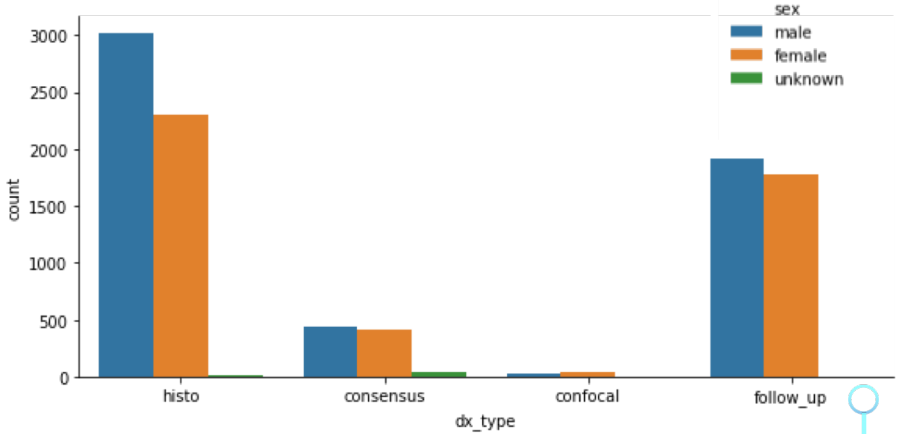


DISTRIBUTION OF IMAGES BY GENDER...

DISTRIBUTION BY GENDER AND LESION TYPE



DISTRIBUTION BY GENDER AND VALIDATION TECHNOLOGY



INSIGHT FROM EDA

1. Imbalanced data
2. Data contains several strata of information
3. No anomalies in data values
4. Few lesions with multiple images

CHALLENGES OF MEDICAL IMAGE CLASSIFICATION

INSUFFICIENT LABELED SAMPLES

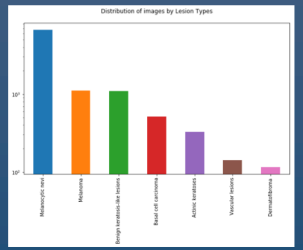
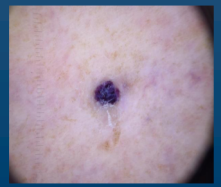


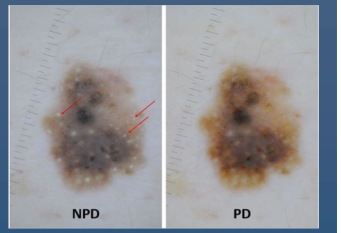
IMAGE SIZE

28 x 28 x 3
784 features

450 x 600 x 3
810,000 features

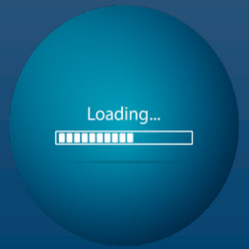


COLOR VARIATION



ARTIFACTS

TIME & COMPUTE CAPABILITIES



Color & Variation Image Source: https://dermoscopedia.org/Differences_between_polarized_and_non_polarized_dermoscopy

MODELS SELECTED

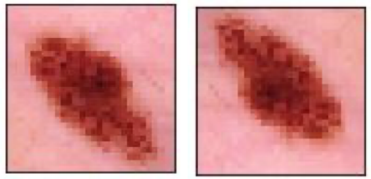
Model characteristics	K-nearest neighbors	Naive Bayes	Logistic Regression	SVM	Neural Net
Explainability	✓	✓	✓		
Model complexity				✓	✓
Speed	✓	✓	✓		
Ease of Training	✓	✓	✓	✓	
Accuracy in high dimension space				✓	✓
Memory efficient				✓	
Likelihood of Overfitting				✓	

MODEL RESULTS - Base Models

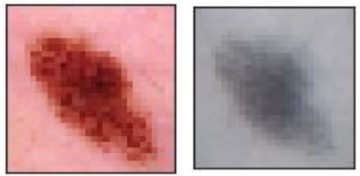
Model results		K-nearest neighbors	Naive Bayes	Logistic Regression	SVM	Neural Net
Overall Accuracy		68%	48%	70%	67%	93%
Melanoma (Cancerous class)	Accuracy	22%	26%	18%	0%	14%
	Recall	16%	34%	13%	0%	8%
Basal cell carcinoma (Cancerous class)	Accuracy	34%	21%	40%	0%	24%
	Recall	31%	37%	35%	0%	19%

MODEL IMPROVEMENTS...

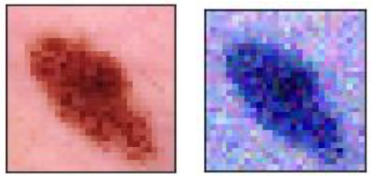
Image augmentation



Flipping

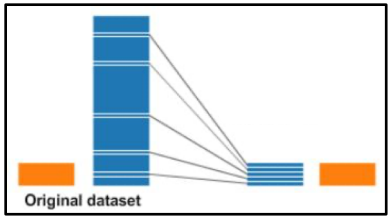


Blurring

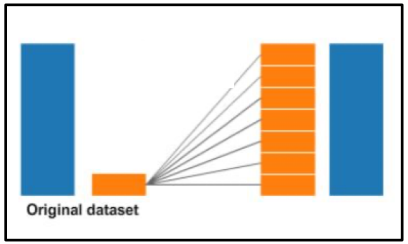


Random noise

Data Balancing

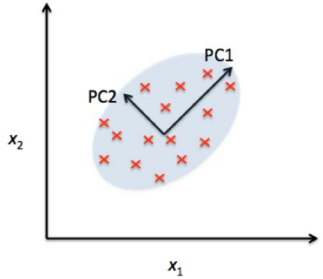


Under sampling



Over sampling

Principal component & Hyperparameter tuning



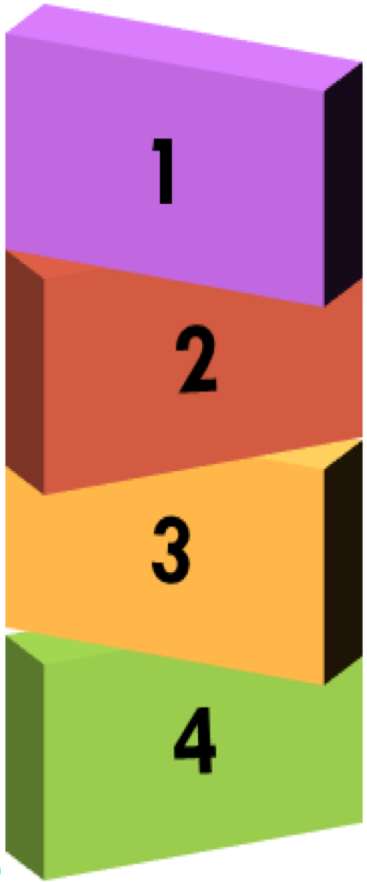
PCA

Hyperparameter tuning

MODEL RESULTS - Post Preprocessing

Model results		K-nearest neighbors		Naive Bayes		Logistic Regression		SVM		Neural Net	
		Base data	Proces-sed	Base data	Proces-sed	Base data	Proces-sed	Base data	Proces-sed	Base data	Proces-sed
Overall Accuracy		68%	87%	48%	67%	70%	68%	67%	76%	93%	81%
Melanoma (Cancerous class)	Accuracy	22%	75%	26%	22%	18%	18%	0%	38%	14%	25%
	Recall	16%	68%	34%	16%	13%	11%	0%	26%	8%	82%
Basal cell carcinoma (Cancerous class)	Accuracy	34%	78%	21%	22%	40%	30%	0%	51%	24%	25%
	Recall	31%	71%	37%	15%	35%	23%	0%	48%	19%	56%

NEXT STEPS



Gaps in our model and how we plan to address them

- Run at 99% explained variance
- Fully execute models
- Try an ensemble
- Cross-validate final model

Improvement techniques yet to explore

- Alternate ways to balance data
- Some kind of weighing scheme to better identify malignant classes
- Binary Vs multi-class problem

Ancillary data we have not used

- Localization
- Age
- Sex/Gender

Identify an acceptable accuracy for a viable product

- Survey with doctors

REFERENCES

- *Machine Learning Methods for Histopathological Image Analysis*, Daisuke Komura ⁎, Shumpei Ishikawa, Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, Computational and Structural Biotechnology Journal 16 (2018) 34–42
- *A Review of the Quantification and Classification of Pigmented Skin Lesions: From Dedicated to Hand-Held Devices*, Mercedes Filho, Zhen Ma and João Manuel R. S. Tavares* Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
- *Automated System for Prediction of Skin Disease using Image Processing and Machine Learning*,International Journal of Computer Applications (0975 – 8887) Volume 180 – No.19, February 2018
- *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research 16 (2002) 321–357
- *SMOTE for high-dimensional class-imbalanced data*, Blagus and Lusa BMC Bioinformatics 2013, 14:106
- *Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset*, International Journal of Computer Applications (0975 – 8887) Volume 77– No.3, September 2013
- *Data augmentation in dermatology image recognition using machine learning*, 1st Lt. Pushkar Aggarwal, University of Cincinnati, Cincinnati, Ohio, Accepted: 28 April 2019
- *The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions*, Philipp Tschandl1*, Cliff Rosendahl2, Harald Kittler1, October 26, 2019

THANK YOU. QUESTIONS?