Dermatologist level classification of skin-cancer with deep neural networks

CS 273B

Shreya Shankar, Nitya Mani, Nancy Xu, Noor Siddiqui

Motivation

- 5.4M non-melanoma skin cancer cases each year
- 20% of adults will get skin cancer
- 10K deaths from melanoma
- $8.1B in US healthcare costs
- Stage I has > 98% survival rate
- Stage IV has < 12% survival rate

Background Work

- ImageNet Google Inception v3
- Microscopy techniques
- ML for diagnosis of melanoma + skin cancers
  - k-Nearest Neighbors
  - PCA
  - SVM

Goal

AI assisted dermatology

Early detection is critical for malignant skin cancers

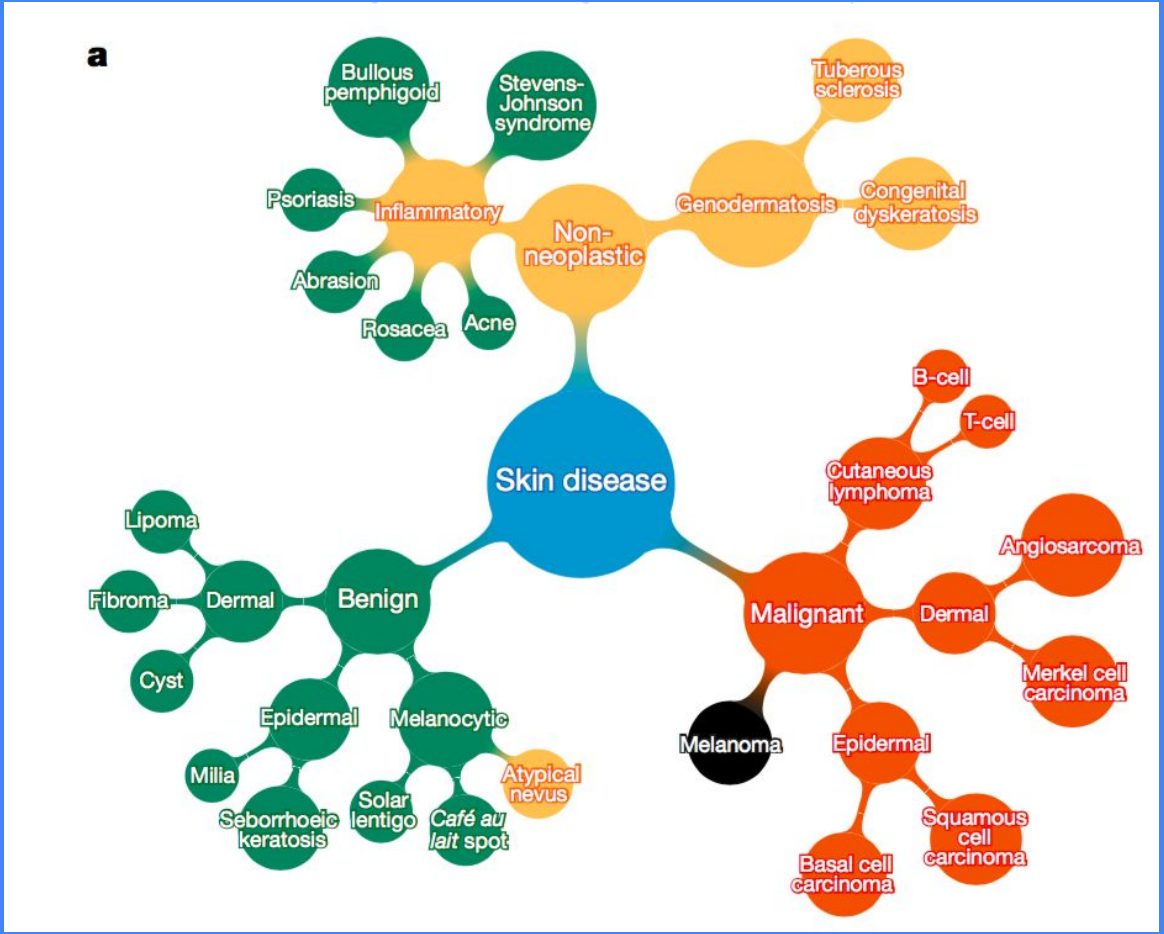
Smartphone enabled diagnostic tool at dermatologist-level accuracy

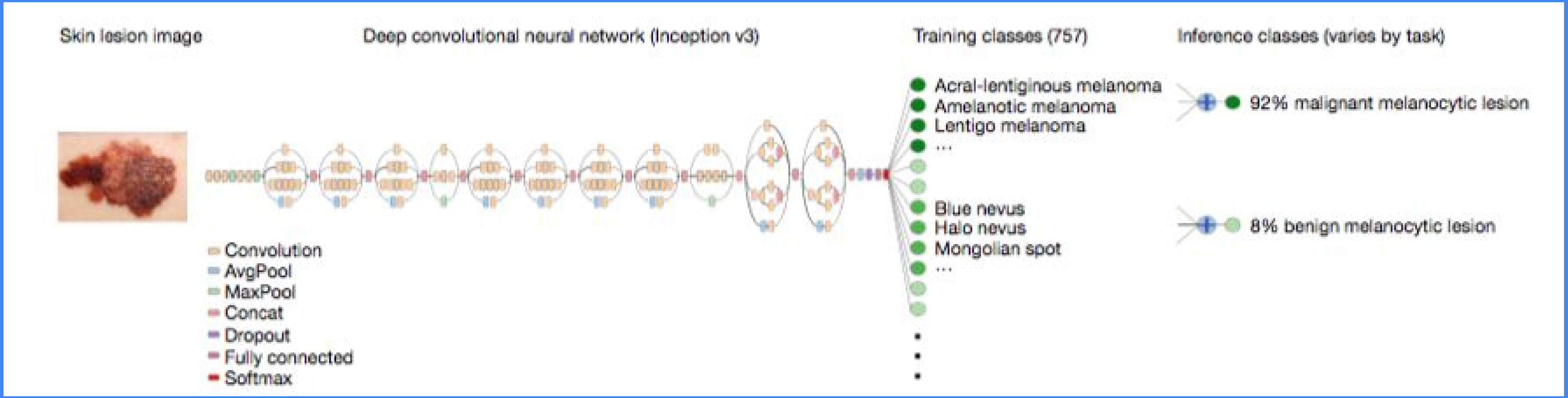
Dataset

- > 130K images of skin
- From: 18 different clinician-curated, open-access online repositories, as well as from clinical data from Stanford University Medical Center
- ISIC Dermoscopic Archive, Edinburgh Dermofit Library, Stanford Hospital, Google Images
- Labelled by hand and verified with biopsy
- 757 disease classes, 2,032 distinct diseases

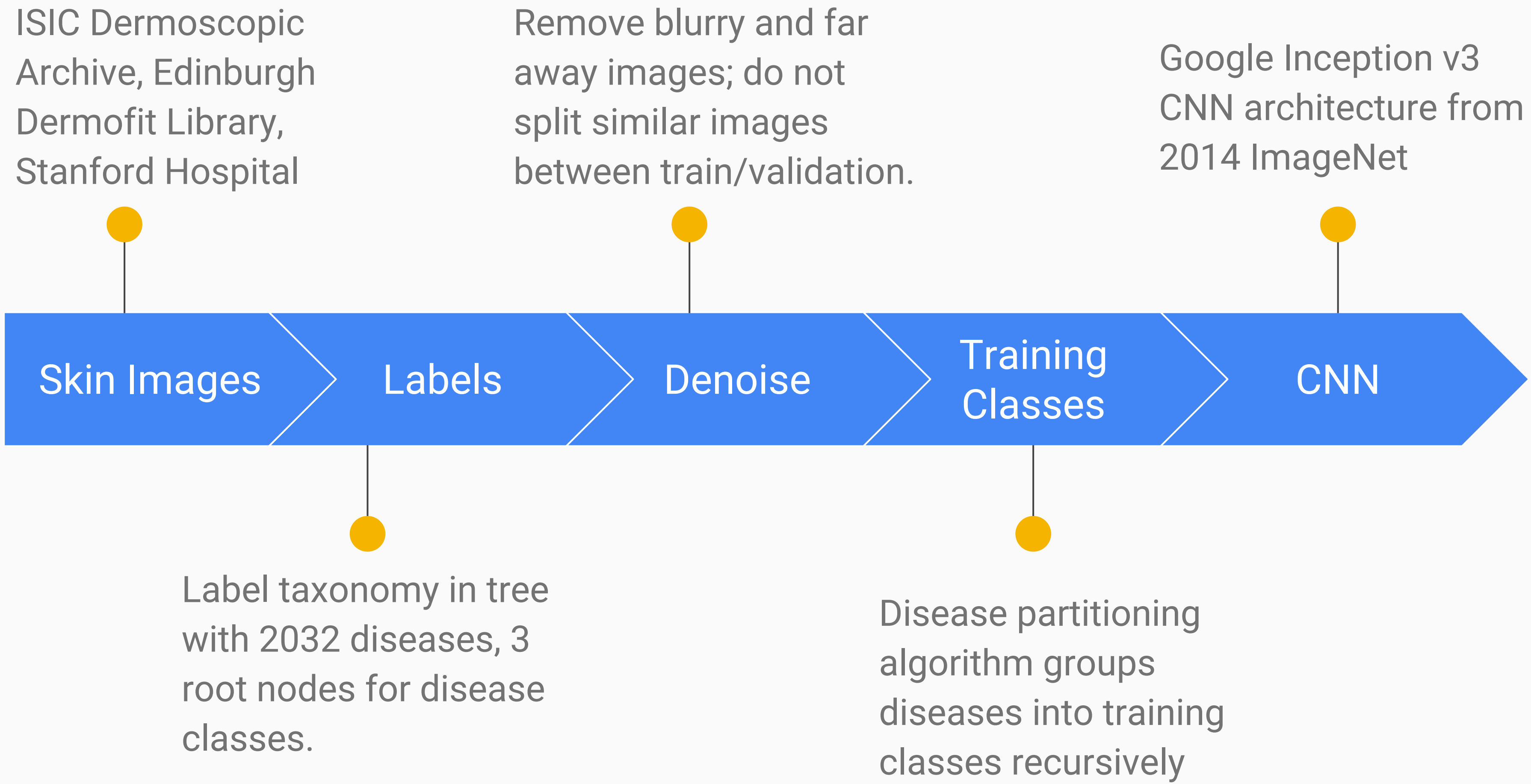
Methodology

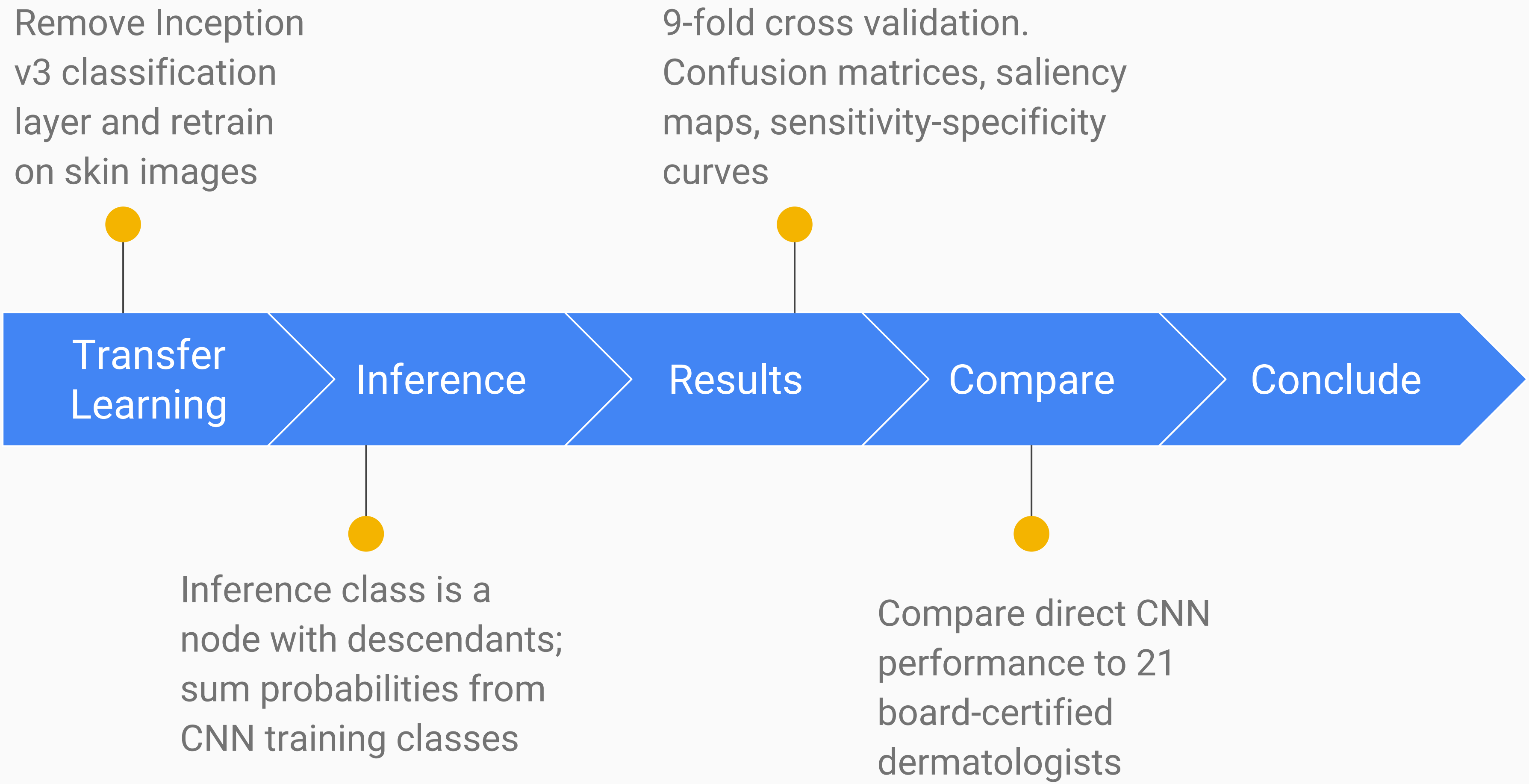
Label Taxonomy




Skin lesion image
Deep convolutional neural network (Inception v3)
Training classes (757)
Inference classes (varies by task)
Acral-lentiginous melanoma
Amelanotic melanoma
Lentigo melanoma
...
Blue nevus
Halo nevus
Mongolian spot
...
92% malignant melanocytic lesion
8% benign melanocytic lesion
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax

CNN


ISIC Dermoscopic Archive, Edinburgh Dermofit Library, Stanford Hospital
Remove blurry and far away images; do not split similar images between train/validation.
Google Inception v3 CNN architecture from 2014 ImageNet
Skin Images
Labels
Denoise
Training Classes
CNN
Label taxonomy in tree with 2032 diseases, 3 root nodes for disease classes.
Disease partitioning algorithm groups diseases into training classes recursively


Remove Inception v3 classification layer and retrain on skin images
9-fold cross validation. Confusion matrices, saliency maps, sensitivity-specificity curves
Transfer Learning
Inference
Results
Compare
Conclude
Inference class is a node with descendants; sum probabilities from CNN training classes
Compare direct CNN performance to 21 board-certified dermatologists

Validation

Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.5%
CNN - PA	**72.0%**

Classifier	Nine-way accuracy
Dermatologist 1	53.3%
Dermatologist 2	55.0%
CNN	48.9%
CNN - PA	**55.3%**

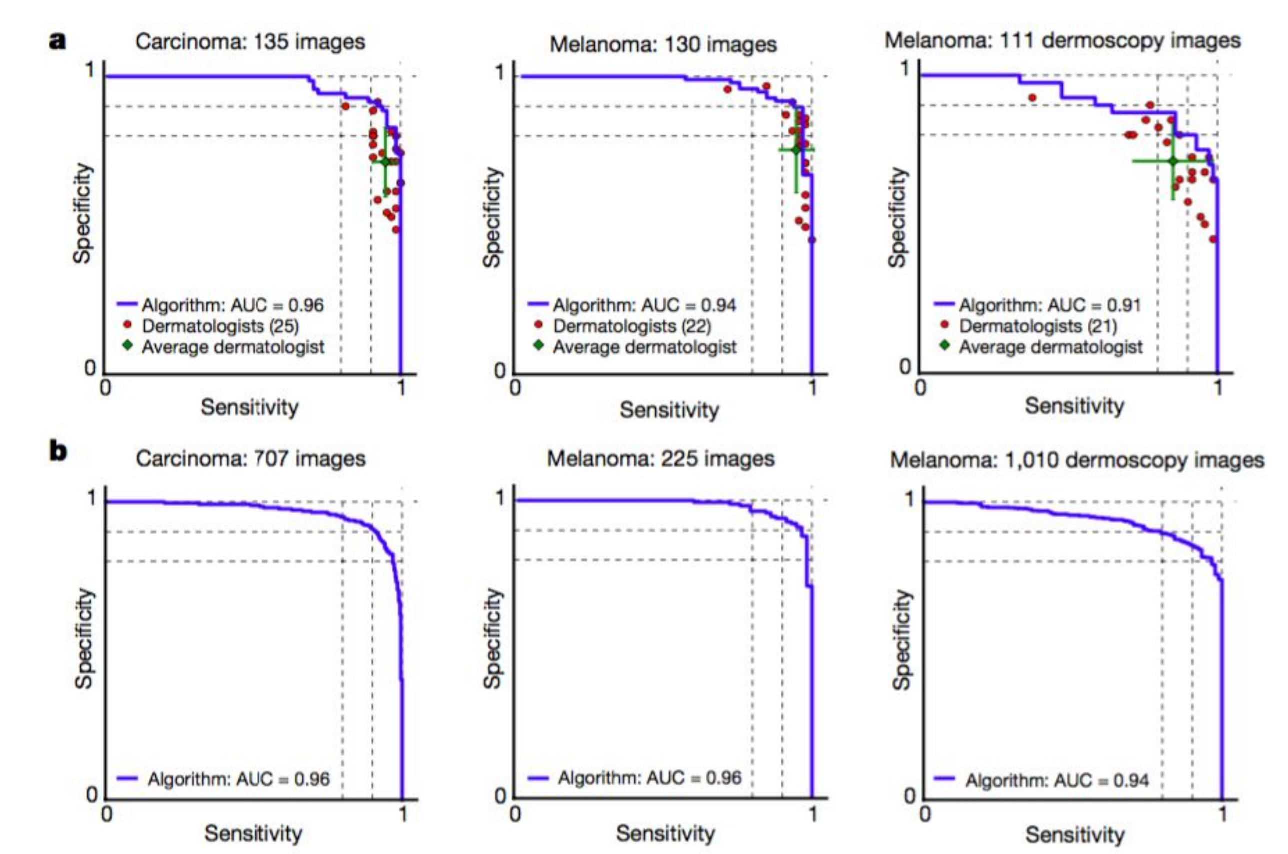
Disease classes:
three-way classification

0. Benign single lesions
1. Malignant single lesions
2. Non-neoplastic lesions

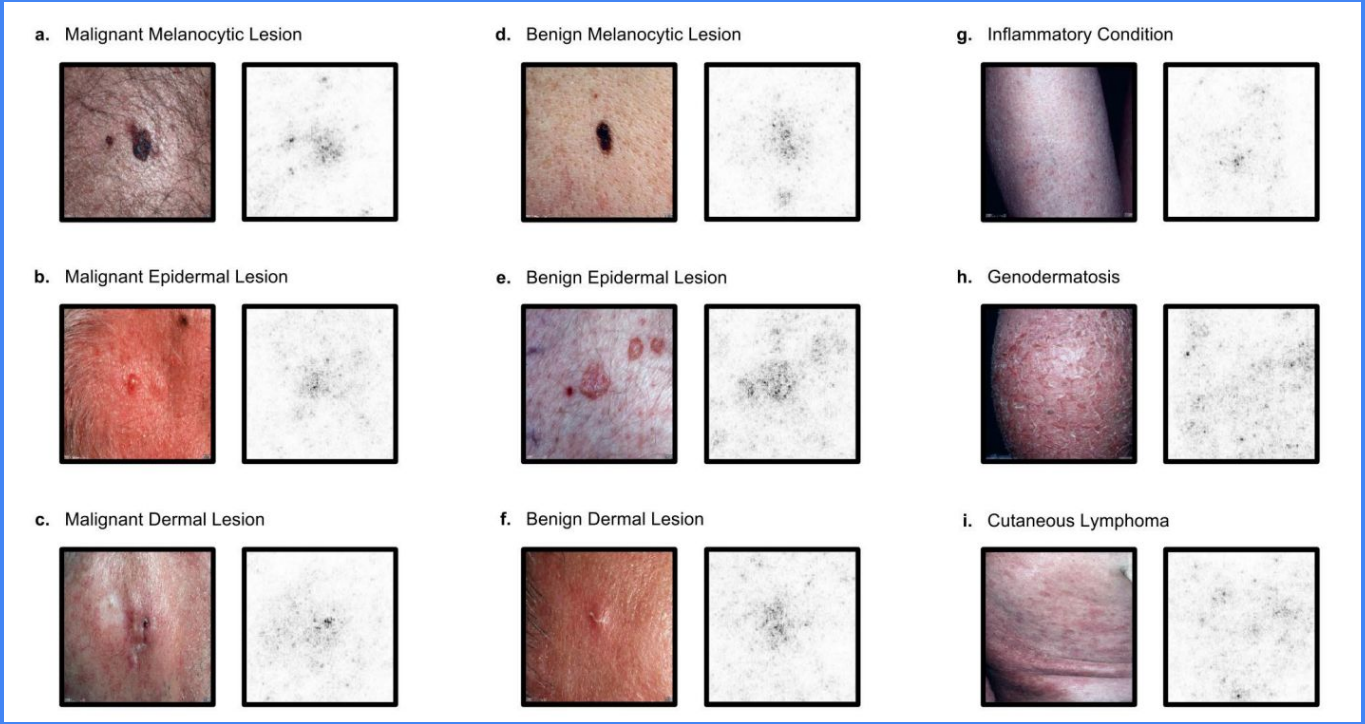
Disease classes:
nine-way classification

0. Cutaneous lymphoma and lymphoid infiltrates
1. Benign dermal tumors, cysts, sinuses
2. Malignant dermal tumor
3. Benign epidermal tumors, hamartomas, milia, and growths
4. Malignant and premalignant epidermal tumors
5. Genodermatoses and supernumerary growths
6. Inflammatory conditions
7. Benign melanocytic lesions
8. Malignant Melanoma

Main Results



Main Results



Strengths

91% AUC minimum for CNN sensitivity-specificity curve

Significant advance compared to previous methods that use PCA, k-nearest-neighbors, or other regression based models

Weaknesses

Binary classification does not translate to real world

Dermatologists use other modalities to diagnose - results do not imply that CNN outperforms dermatologist diagnosis

Issues

Test set includes “gold standard biopsy proven images”, but the training set does not

Paper hopes to be able to aid in the diagnosis skin cancer in stage I; however no results support this specific claim

Extensions: Deploy solution to the real world

Classifier should be more general

Should be capable of being deployed in the field (robust to lower quality images, capable of rejecting images that are too blurry)

Ideally do the processing locally (on the mobile device) as opposed to sending it to the server

Extensions: New Analysis

Analyze the performance of the classifier on other skin diseases in the taxonomy commonly misdiagnosed by dermatologists

Analyze the sensitivity/specificity of the classifier on identifying images from different stages (I, II, III, IV) of skin cancer as compared to dermatologists

Extensions: New Algorithms

Additional data augmentations (the authors only use random reflections/rotations)

Benchmark the performance on different architectures (ResNet V2, etc)

Questions?