Q&A-informed image-based deep learning for skin disease recognition in the real world

Adria Romero-Lopez, Tory Jarmain, Latif Abid, Jevin Maltais, Albert Jimenez, Jeremy Kawahara, Ghassan Hamarneh, Shuang Ao, Mohamed Akrout, Anirudh Challa, Michael Aristocrat, Katrina Cirone, Stephen Solis, Andrew McCann

Triage

Abstract

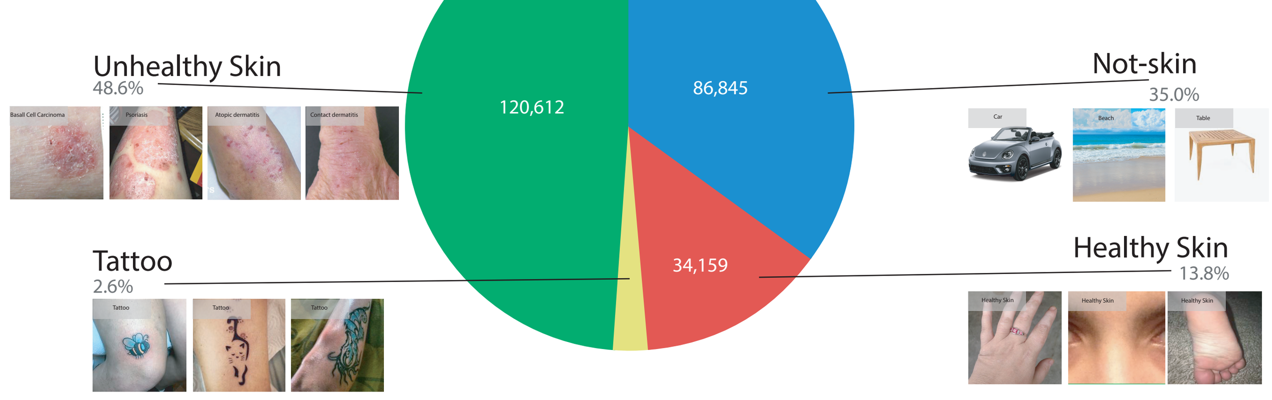
Skin diseases are:
- Among the most common health concerns in the United States
- Indiscriminate of age, gender or race
- Prevalence and cost are comparable to cardiovascular disease, diabetes and obesity
- Malignant skin lesions are of the utmost public and clinical concern
- 1 in 5 Americans will develop skin cancer by the age of 70

The **Triage** approach:
- A smartphone-based self-diagnosis system
- Employs state-of-the-art Machine and Deep Learning:
  - Combines a visual classifier and computational Q&A module
  - Trained on big data (private collection of images and metadata)
- Deployed and already in use at www.triage.com
- Incoming data is continusly reviewed by doctors and fed back into the existing system
- Shown to outperform General Practitioners with visual classification

Data Acquisition

Data Collection

Dataset stats: 200,000+ fully processed



Skin Lesions private collection:
- RGB and square sized images
- Validated images from dermatologists involved in the review process
- Images from dermatology textbooks and dermatology atlases
- Validated data received from dermatologists in Triage's network
- **DermNet New Zealand:** largest contributor to Triage's current dataset

Incoming images review pipeline

Objective
- Expand global network of experienced dermatologists
- Systematically establish expert-consensus labels on incoming images

Process:
1. User uploads an image to Triage system
2. Image is submitted to the Review Pipeline
3. Image is shown to multiple dermatologists for labeling
4. Consensus evaluation:
   - 4.1. If consensus is reached, the image is labaled with the conformed skin condition
   - 4.2. Otherwise, the image is set aside and it will be separately reviewed
5. The reviewed image is removed from the Review Pipeline

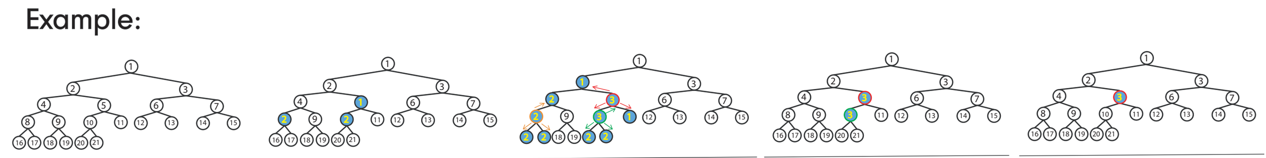
Consensus algorithm

Given:
- Diagnoses $D = \{d_1, d_2, \dots d_N\}$ of submitted image by N reviewers
- Ontology tree T describing clinical hierarchy of skin condition
- Frequency threshold

Output:
- Diagnosis consensus (if reached)

Algorithm:
- Using T, expand each $d_i \in D$ to d^*_i , the set of immediate child and parent conditions of d_i
- Consensus: most frequent diagnosis in $D^* = \{d^*_2 \cup d^*_2 \cup \dots \cup d^*_N\}$
- If multiple conditions are equally frequent, choose the node which occurs most frequently in D

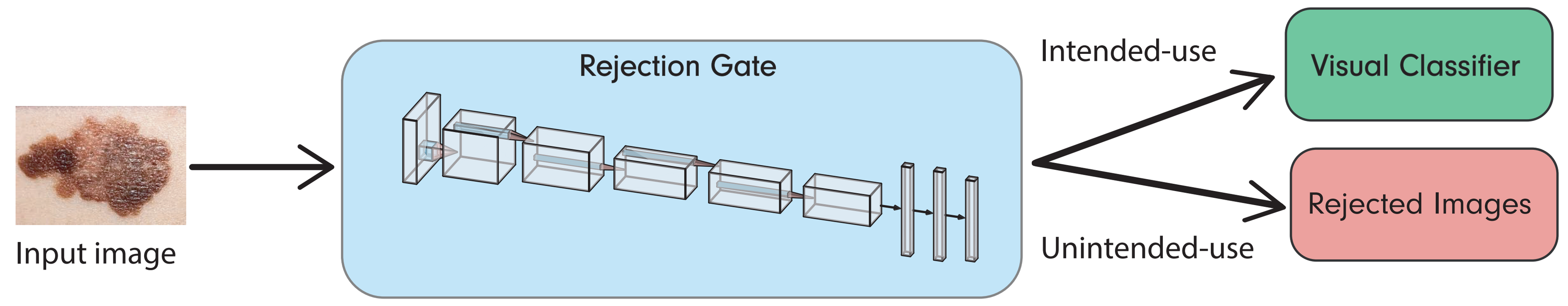
Example:



Ontology tree | 5 experts diagnoses *numbers in yellow indicates # of experts selecting node* | Node expansion *shown as different colored arrows* | Two most frequent nodes with frequency > threshold 2 | Lowest node is selected as the consensus diagnosis

Rejection Gate

- **Gated-logic system** that precedes the final visual classifier
- Classifies images into:
  - **Intended-use** images: Good quality skin lesion images
  - **Unintended-use** images: blurry, poorly-lit, non-skin lesion images and healthy skin images
- Classification approach:
  - **Convolutional Neural Network (CNN)** trained to predict the intended and unintended classes



Benchmarking results

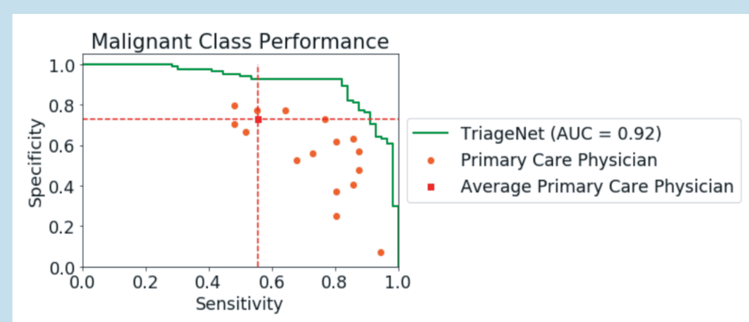
Introduction
- In this study, Triage's AI and a group of general practitioners risk assessment of skin lesions are systematically evaluated to determine average sensitivity and specificity in the detection of melanoma, non-melanoma skin cancer, along with the precancerous actinic keratosis

Materials and Methods
- 140 images of lesions routinely seen in clinical practice from dermnetnz.org's image library were used as test data for the evaluation of general practitioners and Triage's AI's evaluation in terms of specificity and sensitivity in detecting pre-malignant and malignant lesions

Participants
- 15 US board certified general practitioners (family medicine and internal medicine, non pediatrician physicians) were recruited via respondent.io, a platform that sources research participants for qualitative research

Results

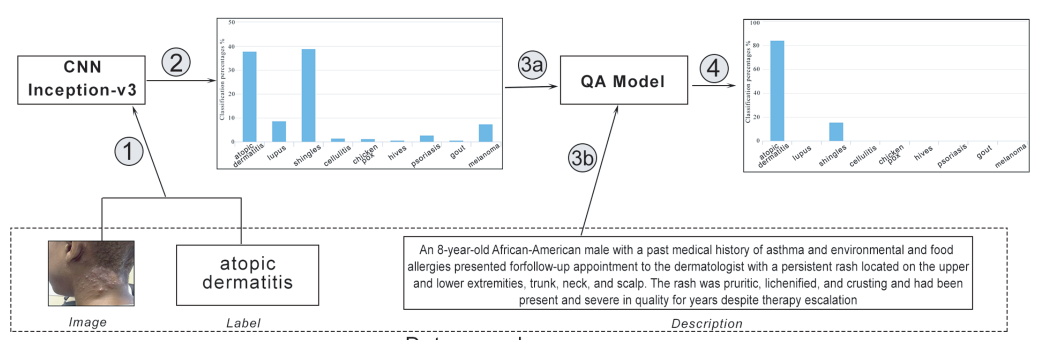
Average weighted results	Accuracy	Precision	Sensitivity	Specificity
General Practicioners	62.0%	67.1%	62.6%	66.1%
Triage	**85.0%**	**85.2%**	**85.0%**	**81.1%**



Conclusions
- In this preliminary experiment, Triage's system demonstrated a superior capability to identifying pre-malignant and malignant conditions from an image when compared to primary care physicians.
- These findings do not represent a head-to-head assessment against primary care physicians as these results are solely based on image recognition. Physicians often leverage patient history, age, sex, and other contextual information to when arriving at a diagnosis.

Q&A model

- Use CNN probabilities as prior to ask questions to patients
- Increase the classification confidence of the CNN
- Imitates the doctor's ability to ask questions needed to arrive at a diagnosis
- Ask the best symptom that maximizes the information gain over symptoms



Data sample

1. **Decision Tree** :
- A search method to determine the best questions
- Increases the CNN top-K accuracy by up to 10%

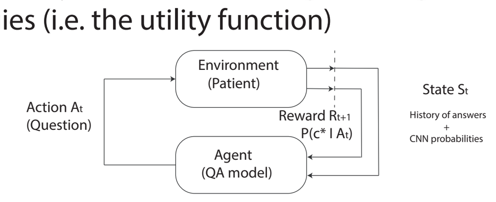
Implementation:
1. Pick the symptom maximizing the information gain
2. Set a cap on maximum number of questions to 10

$$s_j^* = \underset{s_j}{\operatorname{argmax}}\ IG(s_j, C) = \underset{s_j}{\operatorname{argmax}} \left[H(C) - H(C|s_j)\right]$$

3. Choose the conditions with a probability of 95% as a splitting threshold

2. **Reinforcement Learning** :
- It simulates a real-time user interaction
- Increases the CNN top-K accuracy by up to 20%

Implementation:
- Pick the symptom maximizing the long-term reward of splitting strategies (i.e. the utility function)



Visual Classifier

Given:
- A patient's skin lesion image

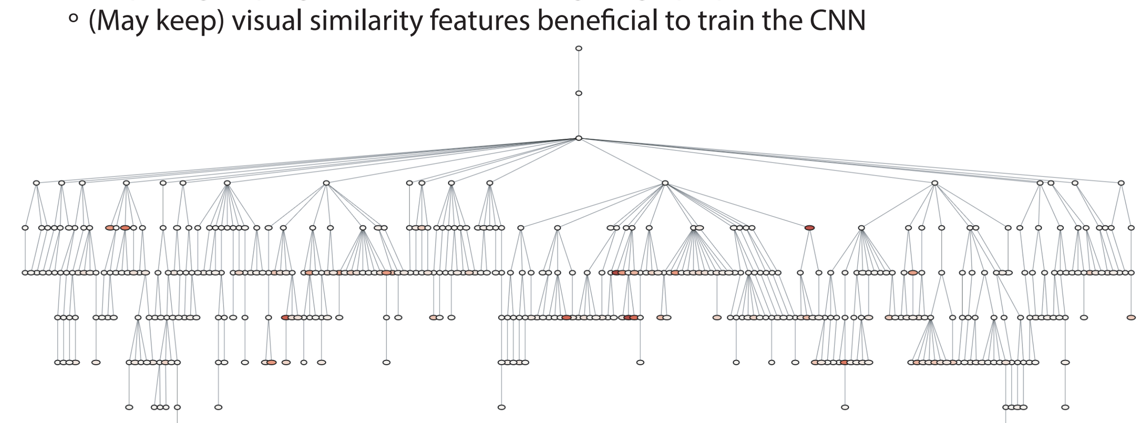
Output:
- Predict the skin disease category

Algorithm:
- Training class selection using an Ontology Partitioning Algorithm
- Data augmentation (rotations, flips, zooms)
- Class weighting to tackle the unbalance within the dataset
- Ensemble of multiple diverse CNN architectures

Inference:
Our diagnosis system averages the predictions from different model architectures to reach the final image predictions

Dermatologic Ontology

- Based on the **DermO ontology**
  - Graph containing the relationships of a broad coverage of dermatologic diseases
  - Phenotypic relationships of skin disorders: sets of observable characteristics
- **Triage Ontology**
  - Keep dermatologic information among condition nodes
  - Propose groupings of nodes accounting for graph pairwise relations
  - (May keep) visual similarity features beneficial to train the CNN

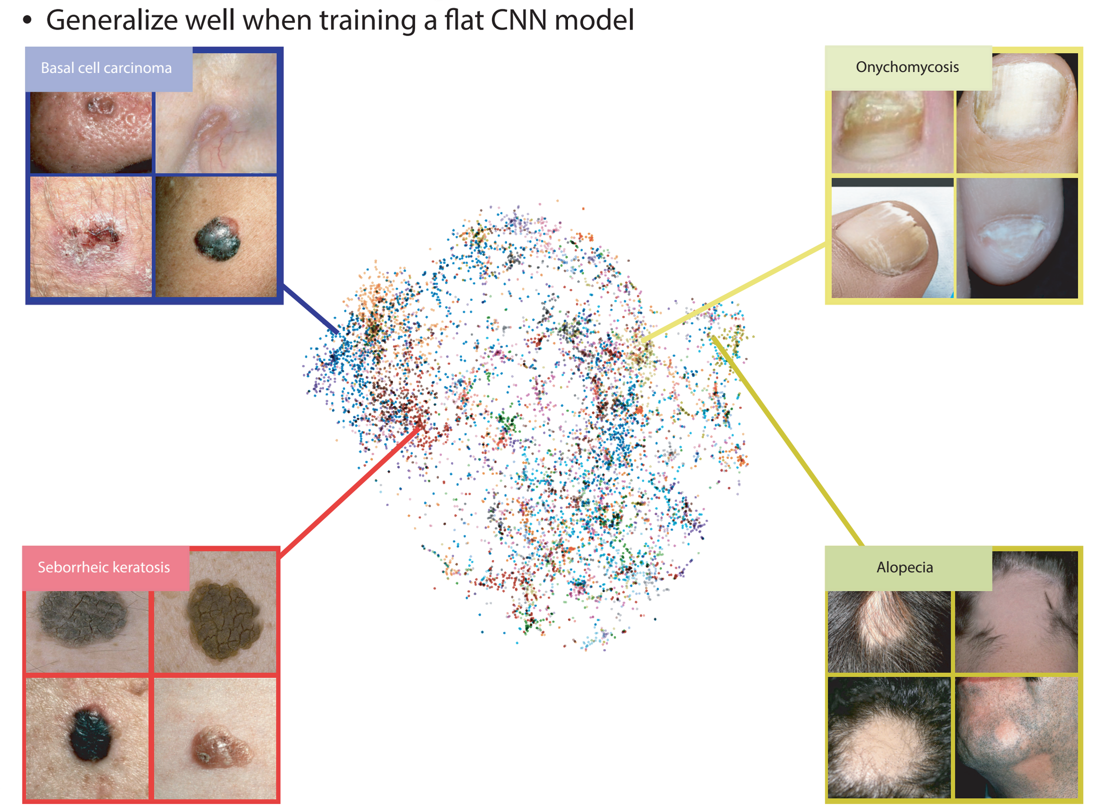


Triage Ontology visualization

Ontology Partitioning algorithm

Motivation
- Class imbalance and insufficient number of images per certain classes

Approach
We propose a grouping of conditions that will act as training classes. These groupings should:
- Share the same clinical meaning
- Generalize well when training a flat CNN model



t-SNE visualization of the 117-way classifier's last dense layer

Challenges and Future Work

Data Understanding
- How to generate efficient training classes:
  - Class imbalance and minimum number of images
  - High intra-class variance

Image metadata
- Encode metadata into visual classifier to be used as privileged information

Domain Adaptation
- Incoming images are visually different from our training & test set (smartphone, lighting, background, blur)

Online Learning
- Re-train models as we receive and process new batches of images

Try out Triage now! Sign up at www.triage.com

NVIDIA | DermNet New Zealand All about the skin | Ontario | MaRS | LEO Innovation Lab