Lessons from Deploying AI in Healthcare

James Zou
Stanford University
March 5, 2022
Example 1: deploying cardiology AI

Article | Published: 25 March 2020

**Video-based AI for beat-to-beat assessment of cardiac function**

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley & James Y. Zou

*Nature* **580**, 252–256(2020) | Cite this article

Computer vision assesses cardiac ultrasound

Algorithm output

EchoNet assessed chamber area

Algorithm mimics clinical workflow

EchoNet-Dynamic

Input

Model

Output

Idea: use temporal segmentation to focus attention of model.

Ouyang et al. Nature 2020
Achieves expert performance in new hospital

Predicting heart failure

Examples

Example 2: AI to improve telemedicine

COVID-19 \rightarrow 50x \text{ increase in digital visits}
Many patient photos for telemedicine are poor quality

• Manual review of photos prior to the physician encounter consumed >2000 hours in 2021 at Stanford

• Poor quality photos disrupt clinic workflow

• Improving teledermatology = improving access to care
Truelmage = Online check deposit for dermatology
TrueImage Workflow

Please move to brighter lighting

Move to brighter lighting

Photo submitted to your clinician

Vodrahalli et al, PSB 2021
TrueImage Algorithm

Skin & lesion Segmentation

Blur Features

Lighting Features

Zoom/Crop Features

Linear / SVM / RF Classifier

‘Good’?

‘Blur’?

‘Lighting’?

‘Zoom/Crop’?

Deep learning + ensembling

Vodrahalli et al, PSB 2021
Prospective study at Stanford

TruelImage filters 80% of poor quality photos; takes <1 minute per patient
What does improvement look like?

The following issues were detected with your image: blur, lighting

GOOD image, thank you! no need to retake.
Deploying TrueImage at Stanford clinics

Launch Gradio interface on HIPAA compliant servers

Your private AWS/GCP machine creates tunnel and public link

Your authorized users can now access the model

The following issues were detected with your image: blur, lighting
Example 3: AI to design clinical trials

Evaluating eligibility criteria of oncology trials using real-world data and AI

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou

Nature 592, 629–633 (2021) | Cite this article
Google launches AI health tool for skin conditions 5/18/21

Breakthrough development will assist users in self-diagnosing issues ranging from acne to melanoma

A woman checks birthmarks on her back. Derm Assist will be free to all internet users, whether they are Google users or not © Albina Gavrilovic/Getty
AI dermatology apps

Original reported: **0.93 AUC**

Stanford patients: **0.60 AUC**

ModelDerm

Roxana Daneshjou
Why did the Derm AI performance crater?

- 0.93 AUC
  - 0.16 visual annot. (prev) vs. biopsy (Stanford)
    - label mistakes
  - 0.12 common diseases (prev) vs. all (Stanford)
    - data shift
  - 0.05 worse on darker skin
    - sensitivity: 0.41 (white) and 0.12 (dark)
  - few dark skin samples in orig. train/test

Roxana Daneshjou
Data used to test 130 FDA-approved AI

93/130 did not report multi-site evaluation
Only 4 prospective studies

Wu et al. *Nature Medicine* 2021
Large variability in cross site performance

<table>
<thead>
<tr>
<th>Site</th>
<th>Stanford (N=19K)</th>
<th>Boston (N=23K)</th>
<th>NIH (N=11K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>0.90 ± 0.01</td>
<td>0.87 ± 0.01</td>
<td>0.85 ± 0.02</td>
</tr>
<tr>
<td>Baylor</td>
<td>0.83 ± 0.01</td>
<td><strong>0.89 ± 0.01</strong></td>
<td>0.84 ± 0.02</td>
</tr>
<tr>
<td>NIH</td>
<td>0.78 ± 0.01</td>
<td>0.76 ± 0.02</td>
<td><strong>0.88 ± 0.02</strong></td>
</tr>
</tbody>
</table>

Wu et al. *Nature Medicine* 2021
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.
2. Understand why AI makes systematic mistakes.
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

1. Data used to train dermatology AI

- AI paper
- private data
- public data

Daneshjou et al. JAMA Derm 2021
1. Transparent dataset and code

Largest public dataset of medical videos.

Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

2. Why did the model make this mistake?

Correct label → Zebra

Conceptual explanation of mistakes
Conceptual explanation of mistakes

Mistakes made by the model

**Label: Allergic Contact Dermatitis**
Pred: Stasis Edema (19%)

**Label: Fixed Eruptions**
Pred: Erythema Nodosum (35%)

**Label: Mucinosis**
Pred: Aplasia Cutis (9%)

**Label: Sarcoidosis**
Pred: Nevus Sebaceous of Jadassohn (36%)

Output of our AI mistake explainer

Abubakar Abid, Mert Yuksekgonul
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.

2. Understand why AI makes systematic mistakes.

3. AI often optimizes the wrong objective

 optimization over a **fixed, single-site** validation dataset

Abubakar Abid
Optimize for human usage instead!

Data Acquisition and Labeling → Split Data into Train/Validation → Model Building → Model Validation → Real-World Usage

optimize over real-world user data

Abubakar Abid
Human-in-the-loop evaluation of ML impact

Would you biopsy the lesion?

Participant’s initial response (response 1)

“AI” advice given

Participant’s final response (response 2)

Kailas Vodrahalli
Worse AI can be better for humans

Human accuracy improvement

![Accuracy Delta](Image)

Dataset 1 Dataset 2

Uncalibrated = overconfident models

Human confidence in correct answer

![Confidence in Correct Label Delta](Image)

Dataset 1 Dataset 2

Kailas Vodrahalli
Using gradio

```python
import gradio
app = gradio.Interface(classify_skin_image, inputs="image", outputs="label")
app.launch(share=True)
```

url: [www.gradio.app/test12543](http://www.gradio.app/test12543) can be shared

Abid et al *Nature Machine Intelligence* 2020
Using gradio

import gradio
app = gradio.Interface(classify_skin_image, inputs="image", outputs="label")
app.launch(share=True)

url: www.gradio.app/test12543 can be shared

Abid et al Nature Machine Intelligence 2020
gradio used for Stanford’s 1st real-time AI trial
Lessons for deploying trustworthy medical AI

1. Understand what data is used to develop the AI.
2. Understand why AI makes systematic mistakes.
Resources and thanks
Papers and codes available www.james-zou.com

Disparity in dermatology AI
Daneshjou et al. JAMA Dermatology 2021

Data transparency for biomedical AI
Wu et al Nature Medicine 2021

Video-based AI for cardiac assessment.
Ouyang et al. Nature 2020

Explaining model mistakes
Abid, Yuksekgonul, Zou. In review 2022

Gradio for human-in-the-loop AI
Abid et al. Nature Machine Intelligence 2020

Thanks to: NIH, NSF CAREER, Sloan, Chan-Zuckerberg