What's the point of the project? - A Summary

Issue

The medical field lacks **high quality and quantity training data** to **train AI models**

Why

- **Shortage** of research volunteers
- Lack of medical professionals
- **Expensive** to gather data

Our Solution

- **AI + Crowdsourcing =** Increases the amount of available training data
  - → **Generative AI** helps increase the low number of available images
  - → **Segmentation AI + Public participation in annotating medical images** quickly labels high-quality datasets comparable to an expert

Project Impact

- Gather **large, high-quality** datasets **quickly**
- **Improve the training performance** of other medical AI
- **Adaptable** for various types of medical image

Future Work

- Implement **a robust weighting and verification system** to uphold **labelling quality**
- Explore **alternative segmentation model** to provide base annotation without user prompt
- **Quantitatively** investigate the **efficacy of synthetic images**

Contact Us

A Yang Lab,
Imperial I-X,
Imperial White City Campus,
London W12 0BZ

https://www.yanglab.fyi/

Acknowledgement

We are immensely grateful to our supervisor **Dr. Yang Guang** for his invaluable supervision and guidance and to **Ms. Quan Yuan** for her generous consultation in the visualisation and design of this brochure.

Coupling AI and Citizen Science

Creation of Enhanced Training Dataset for Medical AI Training

BerryinBarry2

Learn More

IMPERIAL

Motivation

Advancements in AI are enhancing medical image analysis, aiding **precise diagnoses** and **more effective treatment plans**.

A major challenge is the **lack for high-quality training datasets**.

AI-based medical image analysis models need **large amounts of labelled data** to achieve **precise and accurate results**.

However, **manually creating** these labelled datasets by **medical experts is costly and time-consuming**.

To overcome this, there's a need for innovative methods to generate **large amounts of high-quality labelled datasets efficiently**.

Aims

Accessibility
Website Based

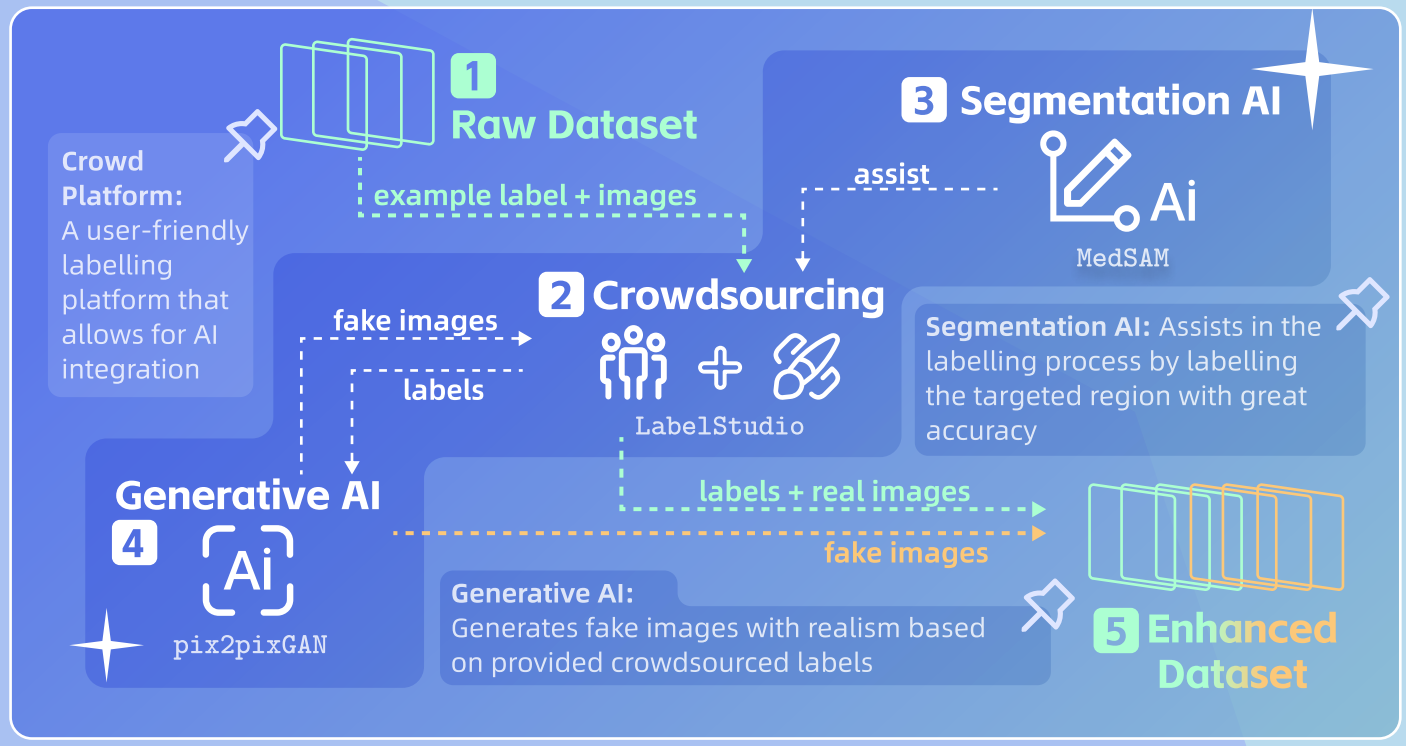
Quantity
Enlarge Existing Dataset

Quality
Accurate Labelling

Efficiency
Low Cost & Less Time

Outcomes

Our **proposed workflow integrates advanced AI and crowd involvement** to enlarge and improve existing unlabelled medical image datasets. This contains **3 main components**:

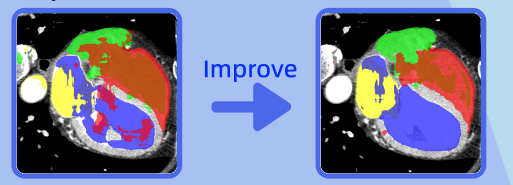


We proposed a **user-friendly online labelling platform**, implemented on LabelStudio, that **gathers volunteers to label medical images** with the **help of Segmentation AI to ensure accuracy**. The labels are then **passed to Generative AI to create fake images**.

By **merging** the fake images and crowd-averaged labels, **a robust and extensive training dataset is created for further medical AI applications**.

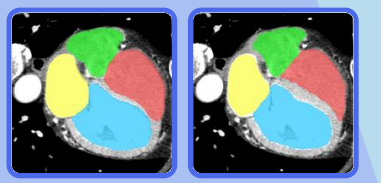
Is this really better?

- **23.2% increase** in test AI model accuracy **with the enhanced dataset**



Original Dataset Enhanced Dataset

- **Comparison** of **crowd-averaged** and **ground-truth labels** show a **76.6% similarity**



Averaged Ground Truth
Heart CT Label