IT UNIVERSITY OF COPENHAGEN

Annotation Task Design in Medical Imaging

Thor Wessel Lindberg (17858, mawl@itu.dk)

April 24, 2024

Computational Diagnostics

Medical Image Analysis

Machine Learning

Expert Diagnostics

Crowdsourcing

Mechanical Turk

Research Question

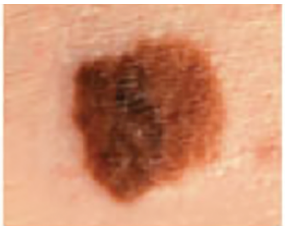
How design decisions are perceived by annotators

and whether design impacts annotation quality

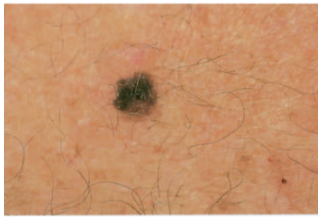
A is for Asymmetry

B is for Border

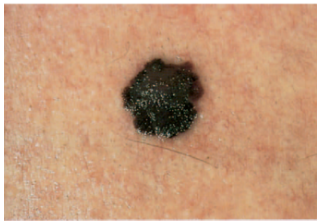
C is for Color



D is for Diameter or Dark



E is for Evolving (Before)



E is for Evolving (After)

Image Dataset

ISIC 2017 Challenge

ENHANCE project

Implementation

Swift language

Distribution

Annotation Tasks

1. *Fail or pass*
2. *Similarity*
3. *Likert scale*

Comparison

Accuracy scores

ISIC ground truth

subset	mturk	annotators	difference	task
1-20	0.28	0.70	0.42	1
21-40	0.48	0.65	0.17	2
41-60	0.48	0.63	0.15	3

$2 \rightarrow 1$

Annotation Quality

1. Intermediary classifications (1) are perceived as benign classifications (0).

2. Annotators are given an intermediary option between benign or malignant.

subset	mturk	annotators	difference	task
1-20	0.72	0.85	0.13	1
21-40	0.58	0.60	0.02	2
41-60	0.67	0.74	0.07	3

$1 \rightarrow 0$ and $2 \rightarrow 1$

Platform

Distribution

Scalability

Design

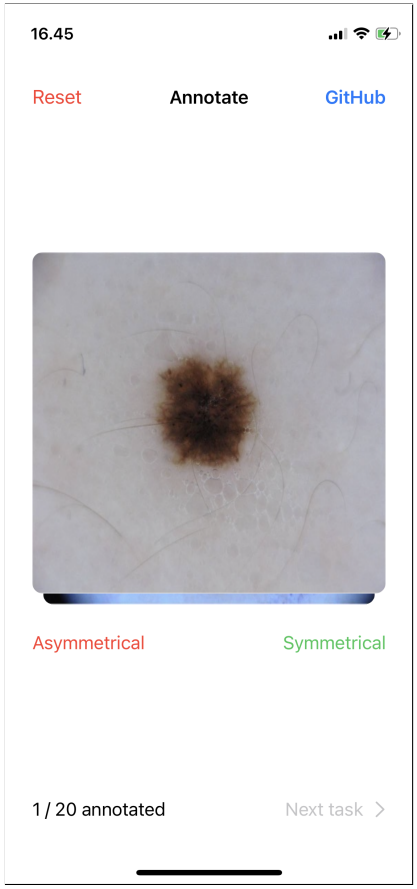
Sampling Braindr

Control vs. Payoff

Facilitating Annotation

Education

Tasks



Annotator Perception

Impact of perception on data quality

"How should skin lesion annotation be presented?"

Research Process

Survey assumptions

"Why is task design not considered or documented?"

Minimum	*Maximum*
slow	fast
stepwise	fluent
instant	delayed
uniform	**diverging**
constant	inconstant
mediated	direct
spatial separation	**spatial proximity**
approximate	precise
gentle	powerful
incidental	targeted
apparent	covered