When the digital Doctor should admit "I don't know"

Yoav Freund[1] **and Hau-Tieng Wu**[2]

[1]UCSD, department, city, postcode, country; [2]Duke, department, city, postcode, country

Introduction

Digital technology is causing a sea-change in all parts of the medical profession. In particular the meteoric rise of AI in general and deep learning in particular raises the possibility that doctors will be replaced computers (1). The father of deep learning, Geoff Hinton, said in 2017: "It's just completely obvious that that in ten years deep learning is going to do better than Radiologists ... They should stop training radiologists now".

Other deep learning researchers provide a more nuanced perspective. Sebastian Thrun (1, 2) argues that "... deep learning devices will not replace dermatologists and radiologists. They will *augment* professionals, offering the expertise and assistance".

Using computers to augment human intelligence rather replace it, is, at the same time, both heady and boring. On the heady side, consider cyborgs whose anatomy is part human, part artificial and can with equal ease solve complex equations or write poetry. On the mundane side, think of smartphones that are quickly becoming an inseparable part of our person.

The idea of using computers to augment or amplify human intelligence has a very long history. The acronyms AI (Artificial intelligence) and IA (Intelligence Amplification or Intelligence Augmentation) have both become popular in the early 1960's(3, 4). These days, the acronym AI is popular, while the acronym IA is not. However, Sebastian Thrun's statement indicates that the idea of Intelligence augmentation is still on people's mind. We suggest bringing it back.

What would IA look like when applied to medicine? We argue that one important ingredient is to endow AI agents with a degree of humility. Specifically, to allow classifiers, such as DNNs, to say "I don't know".

Labels, ground truth and testing

Yoav : I think this and other technical sections should appear in a separate text box.

the quick brown fox jumps right over the lazy dog. the quick brown fox jumps right over the lazy dog.

Roughly speaking, machine learning (ML) can be divided into *unsupervised* learning and *supervised* learning. In both, the task of the learning algorith is transform a set of *examples* into a *model*. In unsupervised learning the examples are undifferentiated raw measurements. In *supervised* learning, which is our main concern here, each example consists of an *input* and a *label*. In the work of Esteva et al (2) on classification of skin cancer the input is an image of a skin patch and the label is "benign" or "melignant".

Typically, the labels are provided by a human expert. These labels define the *ground truth* and the goal of the learning algorithm is to make predictions that diverge as little as possible from the ground truth. As discussed in the next section, ground truth is usually not available in regular medical practice. In this section we point out a problem with the ground truth use in (2).

Esteva et al (2) set out to show that ML can performs as well as or better than expert dermatologists. This meant that they needed to use for ground-truth a label that is more objective than a dermatologist. To that end they used the diagnosis of a biopsy as ground truth. There is no argument that this is a better ground truth than the opinion of a dermatologist.

The problem with this design is that under normal circumstances, patients get biopsied only if the dermatologist thinks there is a chance of melignancy. Therefor, the set of biopsied examples is biased towards melignancy. It is likely that using a classifier trained in this way on an unfiltered stream of patients will increase the number of patients unnecessarily getting a biopsy.

Uncertainty in medicine

Sources of uncertainty in medical diagnosis.

- **The diagnostic process of elimination**
- **Data Quality, Calibration, resolution** Discuss issue as placement of sensors, lighting when analyzing skin lesions. Sensing back for re-testing.

Hiding Uncertainty

- **Psychological reasons** Both doctor and patient prefer the projection of certitude.
- **Protocols**
- **diagnostic devices** Secrecy of the internal code limits the trustworthiness of the alarms.
- **Alarm Fatigue**

Quantifying Confidence With experience comes confidence. In other words, diagnostic options that contradict the accumulated experience are eliminated.

confidence through aggregation

- A good way to increase diagnostic certainty is to solicit the experience of a diverse group of doctors.
- If there is a clear majority for one diagnostic outcome, then the overall confidence in the diagnostics is high.

- This certainty is very different from the the conditional. probability of the disease given the diagnostic. The first is akin to saying: 95% of the dermatologists would give the same diagnostics. The second defines the probability that, if we had access to ground truth, then 95% of the patients that recieve this diagnostics have the corresponding condition.

Uncertainty in Machine Learning

One can define "confidence" in machine learning. The definition follows a similar logic to the one used for human diagnosticians in the previous section. The yardstick by which we measure confidence of predicting a label is "how much do alternative labels contradict previous experience?". More formally, we ask how much do we need to change the training data so that it supports an alternative label.

Yoav : In box: uncertainty versus accuracy using ROC curves

- Bootstrap samples.
- Samples from different hospitals.
- Easy and hard cases.

Agency and Augmentation

Yoav : Doctors need to adapt. Why would doctors prefer to adapt than to resist? What is the migration path for augmentation in medicine?

- **Computer aided diagnostics** Especially with very large data: ecg for 14 says....

  Pathology.

- **Dissemination of expertise** Computers, trained by experts, can help novices. Serves a function similar to score-cards.

  Teaching young diagnostics

- **Confidence, Trust and adoption of technology**

Summary

1. Siddhartha Mukherjee. A.i. versus m.d.: What happens when diagnosis is automated? *The New Yorker*, April 2017.
2. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
3. W Ross Ashby. An introduction to cybernetics. 1957.
4. Douglas C Engelbart. Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, 1962.