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".

Artificial Intelligence and Intelligence Augmentation

Using computers to augment human intelligence rather replace it is both tantalizing and mundane. 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.

What would IA look like when applied to medicine? that is the question we aim to answer here. We argue that an important ingredient of the answer is to introduce to AI agents a level of humility. Specifically, to design classifiers, such as DNNs, to say "I don't know".

Labels, ground truth and testing

Supervised Learning and ground truth

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 the focus of this article, each example consists of an *input* and a *label.* 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.

Skin cancer diagnosis using Deep Neural Networks

One of the papers that provided evidence that deep neural networks might be able to outperform humans is the work of Esteva et al (2). They trained a Deep neural network to classify images of skin into three categories: benign, malignant and non-cancerous. The network was then tested, along with twenty five dermatologists on images which were labeled by a pathologist analysis of the biopsy. The neural network outperformed the human dermatologist. This is, without a doubt, an impressive finding. However, it is based on a retrospective analysis, in other words, an analysis of historical data. To predict the performance of the DNN when used in a dermatology practice we need to how a dermatologist, or any other diagnosticians, arrives at their final diagnostics.

In their famous work, Esteva et al. set out to show that a classifier trained by machine learning can performs as well as or better than expert dermatologists. In this application of supervised learning each example consists of an input image of a skin patch and an output label that is "benign" or "melignant"

As they wanted to compare the system to human dermatologists they needed a better ground truth than that provided by the dermatologists. To that end they used the diagnosis of a biopsy as ground truth. It is arguable that this label is more accurate than the one given by the dermatologist, even though it depends on the human judgement of the pathologist.

However, even if we assume that pathologists labels are more reliable than dermatology labels, the requirement that each example corresponds to a biopsy introduces a significant bias. 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

Patient Monitoring
Take the patient monitor widely used in the intensive care unit (ICU), operation room (OR), or emergency room (ER) as an example. It is now common to analyze biosignals recorded from the patient monitor to train an intelligent (5) or alarm (6) system. **Yoav :** Say a few words about what is an itelligent alarm system. Does it adapt to the patient. Is this related to alarm Fatigue? However, it has been long debate if the recorded biosignals are suitable for this purpose, due to its "blackbox" nature (7–9). **Yoav :** Is the problem with the biosignals or with the system? Regarding black-boxes, as long as we allow the box to output I don't know when the prediction is unstable. Also, per-example explanations can be derived from the features that contribute the most from the score, assuming the learner is biased towards sparse classifiers classifier is a sparse one, such as boosting or Lasso.

Blood pressure monitoring
Recently, some delicate artifacts have been reported (10) regarding the pulse transit time that has been shown to reflect blood pressure information (11). This problem should be viewed as a more complicated version of the "heterogenous noise" issue commonly considered in the statistical literature. **Yoav :** the way this is written, it sounds off topic. Is there an aspect of automation or ML that will connect it to the article?

Inter-rater agreement
in practice physicians need to make a decision when sitting on the "gray area" that is not covered by the protocol. Over this gray area, different physicians may make different decisions based on their experience or the information they have at hand. In some cases, physicians can achieve a reliable decision making, probably with sufficient clinical information (12) or if only the major information is needed (13).

Yoav : Do we need to explain that sometimes inter-rater agreement is high?

Identifying sleep stages
The American Academy of Sleep Medicine (AASM) publishes criteria for manual sleep stage annotation and sleep apnea detection. This annotation is based on manual analysis of biosignals(14, 15). However, it is well known that the inter-rater agreement rate of sleep stage annotation among experienced experts is only about 76% over normal subjects and about 71% over subjects with sleep apnea (16).

Low inter-rater agreement
This low inter-rater agreement can be found in many clinical problems (17–19). **Yoav :** Please elaborate.

The issue underlying the inter-rater agreement is subtle. [physiological knowledge, phase transition, available information, treatment target, economic consideration]

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.

Uncertainty vs. accurracy
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.
5. Alistair E W Johnson, Mohammad M. Ghassemi, Shamim Nemati, Katherine E. Niehaus, David Clifton, and Gari D. Clifford. Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, 104(2):444–466, 2016.
6. William Fleischman, Bethany Ciliberto, Nicole Rozanski, Vivek Parwani, and Steven L Bernstein. Emergency department monitor alarms rarely change clinical management: An observational study. *The American journal of emergency medicine*, 2019.
7. Jeffrey M. Feldman. Can clinical monitors be used as scientific instruments? *Anesthesia and Analgesia*, 103(5):1071–1072, 2006.
8. Kirk H. Shelley and Steven J. Barker. Disclosures, what is necessary and sufficient? *Anesthesia and Analgesia*, 122(2):307–308, 2016.
9. Maxime Cannesson and Steven L. Shafer. All boxes are black. *Anesthesia and Analgesia*, 122(2):309–317, 2016.
10. Yu-Ting Lin, Yu-Lun Lo, Chen-Yun Lin, Martin G Frasch, and Hau-Tieng Wu. Unexpected sawtooth artifact in beat-to-beat pulse transit time measured from patient monitor data. *PloS one*, 14(9), 2019.
11. Heiko Gesche, Detlef Grosskurth, Gert Küchler, and Andreas Patzak. Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method. *European journal of applied physiology*, 112(1):309–315, 2012.
12. Sangeeta Mehta, John Granton, Stephen E Lapinsky, Gary Newton, Kristofer Bandayrel, Anjuli Little, Chuin Siau, Deborah J Cook, Dieter Ayers, Joel Singer, et al. Agreement in electrocardiogram interpretation in patients with septic shock. *Critical care medicine*, 39 (9):2080–2086, 2011.
13. Monika Atiya, Tobias Kurth, Klaus Berger, Julie E Buring, and Carlos S Kase. Interobserver agreement in the classification of stroke in the women's health study. *Stroke*, 34(2): 565–567, 2003.
14. C. Iber, S. Ancoli-Isreal, A. Chesson Jr., , and S. Quan. *The AASM Manual for Scoring of Sleep and Associated Events-Rules: Terminology and Technical Specification*. American Academy of Sleep Medicine, 2007.
15. R. B. Berry, D. G. Budhiraja, and et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. *J Clin Sleep Med*, 8(5):597–619, 2012.
16. Robert G Norman, Ivan Pal, Chip Stewart, Joyce A Walsleben, and David M Rapoport. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, 23(7):901–908, 2000.
17. Maria Brosnan, Andre La Gerche, Saurabh Kumar, Wilson Lo, Jonathan Kalman, and David Prior. Modest agreement in ecg interpretation limits the application of ecg screening in young athletes. *Heart Rhythm*, 12(1):130–136, 2015.
18. Mika Venhola, Mikko Reunanen, Seppo Taskinen, Tuija Lahdes-Vasama, and Matti Uhari. Interobserver and intra-observer agreement in interpreting urodynamic measurements in children. *The Journal of urology*, 169(6):2344–2346, 2003.
19. Diana Carolina Moncada, Zulma Vanessa Rueda, Antonio Macías, Tatiana Suárez, Héctor Ortega, and Lázaro Agustín Vélez. Reading and interpretation of chest x-ray in adults with community-acquired pneumonia. *The Brazilian Journal of Infectious Diseases*, 15(6):540–546, 2011.