Why the digital Doctor needs to say ”I don’t know”

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

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

ABSTRACT

The meteoric rise of AI in general and Deep Learning in particular is generating great excitement throughout academia and commerce, and in particular in medicine[?,?]. With some some high-profile claims that AI will soon replace humans in many medical specialties.

In this position paper we present an alternative view. We contrast *Artificial Intelligence* with *Intelligence Augmentation* and argue that the second is more likely to benefit the patient than the first. We provide evidence to this argument and present a vision in which easier decisions are delegated to computers, while the more difficult ones are handled by humans.

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[?]. 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[?,?] 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[?,?]. 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”.

Uncertainty in medical diagnostics

One of the papers that provided evidence that deep neural networks might be able to outperform humans is the work of Esteva et al[?]. 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.

Medical diagnosis is an iterative process elimination. At the start, the doctor has some basic information such as age, medication, and medical history. Based on that information and an interview with the patient, the doctor identifies some diagnostic possibilities. Based on these possibilities, the doctor orders some tests. The tests might allow the doctor to eliminate some possibilities and potentially order more tests. This process is repeated until the possibilities are sufficiently narrow that the doctor can select a treatment plan.

In most cases the correct diagnosis remains unknown. In the better cases, the outcome of the therapy is tracked and recorded. Typically, nothing is recorded.

If ground-truth labels are so hard to collect, is there *any* way to apply machine learning to medicine? In the next section we propose a way in which Machine learning can help medical communities share and evaluate what they know.

1 Background on machine learning

input feature vector, output, training set, test set, ...

Even if we are now able to reliably obtain ground-truth labels, the data quality itself might be another issue. Examples include commonly encountered missing value, noises, resolution, etc. **Yoav :** This should go into a section describing reasons for saying IDK

But the situation could be more delicate nowadays, particularly when it gets more popular to collect as much data as possible, from any possible environments. For example, almost all possible data, including electronic health record, medical imaging, patient monitor, etc, in the hospital environment are considered. In the home-care environment, signals recorded from mobile health devices, daily diary, and others are considered. However, the data validation, calibration and standardization are often not discussed and, rather, implicitly assumed. **Yoav :** It is not clear what is the point of this paragraph. In general, I prefer not to use terms such as ”delicate” and ”subtle” which seem to imply that the authors of the paper are smarter than the reader.

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[?] or alarm[?] system. **Yoav :** This sounds interesting! can you elaborate? Do these systems allow for ”I don’t know”

However, it has been long debate if the recorded biosignals are suitable for this purpose, due to its “blackbox” nature[?,?,?]. **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.

Recently, some delicate artifacts have been reported[?] regarding the pulse transit time that has been shown to reflect blood pressure information[?]. 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?

On quantifying confidence

When a doctor is uncertain of her diagnostics, she will consult other doctors, preferably ones with more experience, seniority or relevant expertise. In a hospital setting an intern would consult a resident who would consult a visiting staff member.

Sometimes, such as in cancer boards, a team of senior doctors is assembled to reach a diagnostics and a treatment plan. Unless the doctors all agree from the start, a discussion ensues. In this discussion doctors present evidence and quantify the confidence of their conclusions.

What does confidence measure in such cases? Consider the statement ”I am confident that the patient has disease X”. One interpretation, which we will call interpretation A is that the probability of disease X given the symptoms is close to one. Note that if one doctor is confident that the disease is X and another is confident that the disease is Y, there is no obvious way a rational discussion can resolve the difference.

We propose a different interpretation of the statement. Under interpretation B, high confidence means that the probability of X is higher than the probability of Y given the symptoms *and given my training and experience*. In other ways, for the my diagnostics to be Y I have to ignore a large fraction of my diagnostics experience.

Conditioning confidence on experience allows for a rational resolution of disagreement. statements such as ”I have more experience than you” or ”Most of my experience is with inner city children” or ”I did my residency treating cases similar to this” are all legitimate argument for why my diagnosis should prevail.

We can partition cases (examples) into *easy* or *hard* based on the distribution of opinions in the group. If there is unanimous agreement we say that the case is very easy.

Medical diagnostics is fraught with uncertainties. At the same time, often a significant fraction of the cases are diagnosed by most doctors identically, easily and confidently.

Expert labeling can be unreliable

The difficulty of collecting ”ground truth” : reliable labels:

- Ground-truth labeling is difficult and subjective. Examples: biopsies and autopsies.
- High inter-rater disagreement.

hautieng : In many clinical scenarios, physicians follow protocols to make clinical decisions. Protocols are usually determined by experts supported by medical evidence.

Yoav : Is there a relationship between cases which can be diagnosed by the protocol and ”easy” cases? A standard machine learning procedure is having labels from medical experts, and those labels come from protocols. **Yoav :** Why just from protocols? Do protocols allow for levels of confidence?

However, due to the complexity of human body and its intricate interaction with the environment, a protocol might not cover everything, **Yoav :** Do you mean all patients? and 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[?] or if only the major information is needed[?]. **Yoav :** is this with or without a protocol? What is meant here by ”reliable”, is it that most of the doctors agree? Or is it the accuracy as meaured against ”ground truth”? Do you have an estimate of how often ground truth is available?

But in other cases, a low inter-rater agreement rate might happen. Take the sleep study as an example. There has been criteria for sleep stage annotation and sleep apnea event annotation by reading continuous biosignals continuously updated by the American Academy of Sleep Medicine (AASM)[?,?]. 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[?]. This low inter-rater agreement can be found in many clinical problems[?,?,?]. **Yoav :** It would be very interesting to know whether the disagreement is distributed uniformly over all of the cases: i.e. 70/30 split over all cases, or are there cases that have a 90/10 split and can be though of as ”easy”, while other cases have a 55/45 split and are therefor ”hard”

A direct consequence of this low inter-rater agreement is a questionable trained “artificial intelligence”. It is possible that we magically obtain a dataset that contains information that is sufficient for the decision making, while the information is too subtle so that it is not considered in the protocol, and we also magically obtain labels from a magical master that can see though all the information and provide the correct decision. However, by doing a simple math, we shall not count on such a magic and should come back to the protocol itself.

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

Yoav : Can you describe a particular interesting / illuminating / convincing case?

Data size vs. data diversity

The prevalent methodology for estimating the test error of a deep neural network (or any other learning algorithm) is to collect a large dataset of labeled examples, split this data, at random, into a training set and a test set, train the DNN on the training set and test the result on the test set. The reported error is the error on the test set. Comparisons of the accuracy of the DNN to human accuracy are usually based on computing the test error in that way.

This random train/test methodology (RTTM) is valid under the assumption that the training set and the test set are both drawn from the same stationary distribution. However, the assumption rarely holds in practice. In practice, data collected in different hospitals has different distributions. Differences arise from different patient populations, different protocols, differences in the digitization instruments and many other causes. In order to estimate the true test error, the training set and the test set should be collected from *different* hospitals.

Replication Issues with some published results

Not all claims regarding deep learning in medicine can be trusted.

Skin cancer detection[?]

Pneumonia detection paper solely from X-ray (Stanford)

2 When the best answer is ”I don’t know”

Papers in DNN research often claim that the generated neural network performs better than humans. In this section we present some of the critiques of this claim and propose the remedy of outputting ”I don’t know” on the harder examples.

In the following subsections we discuss the important role of abstention in medical practice and the ways in which abstention can be formalized and used in machine learning.

Medical Augmentation

The importance of saying “I don’t know” in medical practice. (medical augmentation)

Cancer Boards

Second opinions

Reducing alarm fatigue

Bed-side alarm system. (adaptive systems for reducing alarm fatigue?)

Using Ensembles

Using ensembles of classifiers to quantify uncertainty.

3 Agency, trust and adaptation

A quote from Robert Rechter's book "The digital doctor"[?]:

> Harvard psychiatrist and leadership guru Ronald Heifetz has described two types of problems: technical and adaptive. Technical problems can be solved with new tools, new practices, and conventional leadership. Baking a cake is a technical problem: follow the recipe and the results are likely to be fine. Heifetz contrasts technical problems with adaptive ones: problems that require people themselves to change. In adaptive problems, he explains, the people are both the problem and the solution. Leadership, he once said, requires mobilizing and engaging people around a problem "rather than trying to anesthetize them so you can go off and solve it on your own."

Rechter continues to say that the digitization of medicine "the Mother of All Adaptive Problems". In other words, for AI to be widely adapted, doctors and nurses ("medic" in the following) need to positively engage in its adaptation. Declaring that AI will soon replace medics, positions AI in an adversarial stance towards medics and is likely to make them more resistant to the adoption of AI technology[?].

Moreover, as argued above, claims that AI can perform diagnosis more accurately than most medical professionals are overblown. On the other hand, if we allow the AI system to *abstain* from prediction on the hard cases, high accuracy on the easier cases. Using AI to classify the easy cases can reduce the work load on the doctor or nurse, and free more time to deal with the hard cases.

This approach is often called IA, which stands for "Intelligence Amplification" or "Intelligence Augmentation". In this approach the role of the computer is to assist, rather than replace the human. The medic remains the agent responsible for the treatment of the patient, the medic delegates some of the work to the IA agent, but sets a threshold on the confidence level such that when the confidence level of the agent is low, it re-engages the medic.

The patient-medic relationship is strengthened, because the medic can devote more time to the more difficult cases.

4 Summary