At times the digital Doctor should admit "I don't know"

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

[1]UCSD, Computer Science, San Diego, 92093, United States; [2]Duke, department, city, postcode, country

The meteoric rise of AI and Deep learning raises the possibility that doctors will be replaced computers (1). Geoff Hinton, a famous deep learning researcher said in 2017: “It's just completely obvious that in ten years deep learning is going to do better than Radiologists ... They should stop training radiologists now”.

The predictions of Sebastian Thrun (1, 2), another leader in machine learning, are less disruptive: “... deep learning devices will not replace dermatologists and radiologists. They will augment professionals, offering the expertise and assistance”. In this article we argue for Thrun's prediction and explain why augmentation, rather than replacement, is the approach more likely to prevail.

The question of whether dermatologists will be replaced by computers or be empowered by computers is but a recent incarnation of a debate between AI (Artificial Intelligence) and IA (Intelligence amplification) which has a long history (see inset). To distinguish between AI and IA we use the terms “AI agent” vs. “IA sidekick”. This terminology contrasts agents, which are endowed with agency and can take actions that effect the patient's health, with sidekicks which can provide advice and suggestions, but who are not allowed to take action.

Replacing dermatologists with AI agents can bring cost savings, but is likely to lead to inferior care. One of the reasons is that it is hard for AI to make a human connection with the patient and thereby take into consideration personal, social, financial and mental factors.

On the other hand, IA powered sidekicks IA can help the medical staff detect and diagnose medical problems quickly, efficiently, accurately. This can lead to cost savings, especially for homebound patients suffering from chronic diseases.

Artificial Intelligence and Intelligence Augmentation

The driving question of AI can be summarized as: “are machines capable of behaving in a way that indistinguishable from that of humans, as judged by other humans”. The archetypal test of whether artificial intelligence has been achieved is the Turing Test, in which a human, communicating with another agent through text alone, is unable to tell whether or not the agent is human. A natural consequence of computers being indistinguishable from humans is that they will be replacing humans, causing mass unemployment.

The driving question of IA is whether and how computers can be used to augment humans rather than replace them. Some augmentations are the territory of science fiction. For example, cyborgs whose anatomy is part human, part artificial and can with equal ease solve complex equations or write poetry. Other examples are so mundane are so mundane that they are taken for granted. Examples are the smart phone and google search are ways in which our capabilities are augmented by computers.

The Turing test was published (3) in 1951. A 1956 workshop in Dartmouth college is widely recognized as the beginning of the field of AI. IA appeared on the scene soon thereafter, with Ashby (4) in 1957 Licklider (5) in 1960 and Englbart (6) in 1962.

Arguably, the impact of IA on today's society is much larger than that of AI. Siri, Google search and assisted driving are some of the common apps that augment human ability. On the other hand, the goal of creating a general purpose AI that possesses a human-like capability to reason about new domains seems to be as far as ever. At the same time, AI holds the fascination of many, probably because of it's tantalizing combination of promise and threat.

Central to our approach is a quantification of prediction confidence. Such quantification is needed to avoid premature diagnostic conclusions, and to decide which additional tests or consultations might be needed. Consider a doctor that is asked asked to diagnose a patient with complex or conflicting symptoms. A careful doctor will admit their uncertainty and perform additional tests or ask a specialist. A less careful, overly self confident doctor is likely give an incorrect diagnosis and choose an ineffective or even damaging treatment plan.

An AI agent, trained to be better than the human doctor, might end up behaving like an overly confident doctor. An IA sidekick, aware of it's own limitations, will give advice only when the evidence is strong and otherwise say “I don’t know”.

In the following sections we explore these ideas in more detail. We start with a critique of one of the papers that claims that AI agents can outpeform human diagnosticians.

1. Supervised Learning and the Ground Truth

Deep learning is a special case of supervised learning (see inset), sometimes called input-output learning (7, 8).

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.

The data for supervised learning consists of a large collection (input,output) pairs. For medical diagnosis, the inputs is medical information for the patient (Heart rate, blood tests, X-ray images etc.) and the output is the diagnosis. This output is considered the “ground-truth” and is assumed to represent the undisputed truth.

Here lies the the first difficulty with applying supervised learning to medical diagnosis. In most real-world scenarios the diagnosis does is not an objectively measurable fact, rather, it represents the conclusion drawn by a fallible human diagnostician. We will return to this issue in the next section.

The other important assumption made in supervised learning is that the generated classifier is tested using the same distribution of examples as that of the training set.

We now consider a study in deep neural networks which claims to show that DNNs can perform diagnostics as well as, or better, than human diagnosticians. In a highly cited paper in the journal Science (2) provides evidence supporting the claim that computers can diagnose skin cancer as well or better than board certified dermatologists.

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 performed comparably to, and sometimes better than the human dermatologist. To provide ground truth, the patients were biopsied and the piopsies were diagnosed by pathologists.

A fundamental problem with the experiment is in the way the data was collected. The data used in the experiment was retrospective, i.e. it was collected from the records of past patients for which both a skin image and a biopsy were available. Normally, patients get biopsied only if the dermatologist thinks there is a significant chance of **malignancy**. As a result, a retrospective study that is based on patients for whom a biopsy was taken is likely to over-represent malignant patients and therefor be biased. If an image-based classifier is trained on the biased data, its performance on unbiased test data is likely to be worse. Specifically, when the classifier is applied to skin images of undiagnosed patients it is likely to over-diagnose them as malignant. The practical implication would be that more patients than necessary will be biopsied.

As we elaborate on in the next section, in medical diagnostics the ground truth is usually not available, all that we have to go on are the opinions of human diagnosticians.

DRAFT

2. Uncertainty in medicine

For the most part, it is hard to associate ground truth with medical diagnostics. This is evident studies of inter-rater agreement (see inset). In studies of this kind multiple doctors produce diagnostics based identical medical information without communicating with each other.

Inter-rater agreement
A common method for measuring the level of agreement is an inter-rater agreement studies. In such studies several doctors are provided with the same patient file and are asked to give a diagnosis.
A common measure of of the agreement between two raters is the Cohen's kappa coefficient, usually denoted by κ. **Yoav :** How is the kappa coefficient generalized to studies where there are more than two raters ? Kappa is computed from two quantities: $0 \leq a \leq 1$ is the fraction of patient files on which the two raters agree, and $0 \leq c \leq 1$ is the fraction of agreements that would occur by chance. The definition of kappa is $\kappa = \frac{a-c}{1-c}$.
If $\kappa = 1$ The raters always agree, if $\kappa = 0$ the rate of agreement corresponds to chance, and if $\kappa < 0$ then the rate of agreement is lower than chance, i.e. the two raters tend to have different opinion. An interpretation of κ recommended by Cohen (9) is: $\kappa \leq 0$: no agreement, $0 < \kappa \leq 0.20$:none to slight agreement, $0.2 < \kappa \leq 0.40$: fair agreement, $0.4 < \kappa \leq 0.60$ as moderate agreement, $0.6 < \kappa \leq 0.80$: substantial agreement, and $0.8 < \kappa \leq 1.00$: perfect agreement.
Yoav : can we give a list of kappa values for some common diagnostics here?

In addition, diagnosis is not an input-output mapping. Rather, it is an iterative process which reduces uncertainty over time. To illustrate this, consider the diagnostics of a patient that is treated in an out-patient clinique.. When a patient arrives at a clinique for the first time, all diagnostics are possible. After a physical exam and an interview with a doctor, , many possibilities are eliminated. In simple cases, this is enough for the doctor to confidently choose a treatment. In more complex cases, the doctor might ask for multiple tests and visits, refer the patient to a specialist, consult colleagues, journals and books etc. To choose a treatment plan, the set of possible diagnostics has to be reduced however, it does not have to be reduced to a single diagnostics, as multiple diagnostics might share a treatment plan.

In order to apply a supervised learning method, such as DNN, to the diagnostic problem, we need to define a ground-truth label for each patient. But that is easier said than done. As the final output of the diagnostic process is a treatement plan, we would like to know what is the best treatment plan. Unfortunately, we can only use a single treatment plan to treat the patient, so the most that we might be able to infer is whether the chosen treatment was effective. Even if the patient improved, the cause might have been unrelated to the treatment. It might be due to a change in diet or reduction in stress. Moreover, in most cases, there are few or none followup visits and as a result there is no data as to whether the patient has a lasting improvement in health.

For these reasons, we propose a different formulation of the problem. Rather than predicting the "correct" diagnosis we predict the distribution of diagnosis across the doctors. This distribution represents the confidence of the prediction. In other words, if we predict the doctors to be unanimous, we assign to the prediction a high level of confidence. If, on the other hand, we predict that there will be as many diagnoses as there are doctors, then our prediction is useless. It is akin to saying "I don't know".

Yoav : worked up to here

It is certainly expected that physicians can achieve a reliable decision making, probably with sufficient clinical information (10) or if only the major information is needed (11). However, in many cases, the quality of decision making might be jeopardized due to various reasons, among which the uncertainty in medicine is non-negligible.

DRAFT

Uncertainty due to signal quality

Medical devices use a variety of bio-sensors that record, display and distribute different biometric signals, ranging from vital signs such as heart rate, oxygen saturation, temperature and blood pressure, to high-frequency waveforms such as ECG, EEG, respiratory signal and arterial blood pressure. High frequency signals suffer from artifacts and other signal quality problems, some of which depend on the patient. In some cases, these problems can be handled by the human diagnostician. In other cases such as artifact removal of EEG is still an active research problem (12).

Another common source of signal quality issue is how the sensor is placed. While there have been several standards, ranging from the well-known ECG systems (13) and international 10–20 EEG systems (14) to recently smart clothing system for telemedicine (15), it is not always possible to achieve a precise sensor placement for biomedical signal collection due to various reasons. This uncertainty might be tolerable for some clinical applications; for example, an imprecise ECG sensor placement might not impact the identification of some types of arrhythmia from the ECG signal, like atrial fibrillation. However, this uncertainty might cause troubles in other applications; for example, an imprecise placement of the deep brain stimulation lead inside subthalamic nucleus might downgrade the Parkinson disease treatment outcome.

DRAFT

Alarm fatigue

Patient monitors are bedside medical devices that monitor patients that are at risk but currently stable, freeing the medical staff to attend to the patients whose status is critical. Unfortunately, Patient monitors suffer from signal quality issues and tend to generate false alarms at a high rate. Over time, this can result in the staff not responding to alarms, potentially resulting in great damage to the patient. This phenomenon, called alarm fatigue (or alarm overload) is a major problem in hospital care (16, 17).

DRAFT

Protocol limitation

~~**Yoav :** For readers that are not MD, we should explain what are protocols, how they are generated, and whether all or some of their~~

According to NCI dictionaries, protocol means a detailed plan of a scientific or medical experiment, treatment, or procedure. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/protocol.

In clinics, it is a document that guides decision making, including criteria regarding diagnosis, management, and treatment. It exists in different areas of healthcare with different formats. In a loose sense, it could be understood as an algorithm solving a given mathematics problem. However, unlike the relationship between an algorithm and a mathematics problem, a medical protocol might not cover every situation and provide all possible solutions, and have several limitations.

The American Academy of Sleep Medicine (AASM) publishes criteria for manual sleep stage and sleep apnea annotation from the gold standard sleep study instrument, the polysomnogram (PSG). This annotation is based on manual analysis of biosignals recorded from the PSG (18, 19). The AASM is a protocol that has been extensively applied, with rigorous scientific support, and updated regularly according to latest evidences. A detail sleep profile is critical for sleep quality enhancement, or even medical condition improvement. However, it is well known that even with the well established protocol, the inter-rater agreement rate of sleep stage annotation among experienced experts, in terms of percentage of epoch-by-epoch agreement, is only about 76% over normal subjects and about 71% over subjects with sleep apnea, while the Cohen's kappa is 65% over normal subjects and about 59% over subjects with sleep apnea (20). Among many reasons, the one that is directly related to the intelligent system development is how the criteria are "described" in the protocol. For example, it is described in the protocol that if the delta wave occupies more than 20% of a given 30-second epoch of the electroencephalogram during sleep, that 30-second epoch is defined to be the N3 stage. 20% of a given 30-second epoch is 6 seconds. What about if the delta wave occupies 5.99-, or 6.01-seconds? What about if the delta wave sustains for 10 seconds, but it is divided into two consecutive 30-second epochs? When sitting on the "gray area" that is inherited from the protocol, sleep experts need to make a decision based on their experience or the information they have at hand, and this leads to medical uncertainties, and hence the inter-rater, or even intra-rater disagreement.

Another protocol limitation is the "extrapolation error"; that is, when we apply the developed protocol to the population different from the population that we collect the evidence for the protocol (21). Such extrapolation error usually comes from the variability among subjects. If such variability is big, it limits the development of a more quantitative protocol (22), and different protocols might be needed for different situations.

Medical uncertainty as manifest low inter-rater agreement consequence, can be found in many clinical problems (see blocks on ...) Besides the above-mentioned reasons, there are more. For example, in some situations, when the needed information is missing, it is challenging to make a differential diagnosis (31). Despite the variety of reasons, the key message here is that medical uncertainty is a non-negligible fact in medicine.

A direct consequence of the low inter-rater agreement rate is that the trained intelligent system might be questionable. It is clear that such intelligent system is questionable and might raise concerns. Recently, various regulations in this regard have been proposed (32, 33).

Now, suppose we are able to eliminate all challenges from data calibration and validation issues, and we can provide as much information as possible to train the intelligence system. Even under this assumption, it is clear that the system still suffers from the protocol limitation or knowledge gap issues. Can such system be useful in clinics? To answer this question, we should not forget that physicians also follow the same protocol and have knowledge gaps. Depending on the clinical problems, and the experience of physicians under consideration, the agreement rate varies. Usually, intern doctors know the least, while a senior attending knows the most. It is natural that we trust a senior expert more, but it does not mean that we do not trust a junior intern doctor.

knowledge gap

~~**Yoav :** is this an example of uncertainty because of knowledge gap? Wouldn't it be be~~
Urodynamic studies provide the best bladder and sphincter functional data for urologists to decide how to treat patients at risk for renal damage (23). While it has been extensively studied and applied in clinics, the main issue that plagues this field of urodynamics is the lack of precise definition of a detrusor contraction or overactive contraction (23). The lack of a well defined definition is due to the lack of quantitative study from the pathophysiological perspective, so the definition is still based on "expert opinion". For example, usually an overactive contraction represents itself as a "bump" in the detrusor pressure signal. However, what is the breath and height of a bump should we call it an overactive contraction? How to distinguish a true overactive contraction from an artifact? While there have been several reference information, like abdominal pressure, that could help us identify artifacts, but it can only explain a small portion of them. Unsurprisingly, this fundamental issue has led to a significant inter-rater disagreement (22, 24).

Another example is the ongoing pandemic, COVID-19 (25), while this article is prepared. Back in Jan 2020, when it was first reported in China, nobody had a clue how it will generate damage to human body, not to mention how to treat a patient. All medical practices, ranging from diagnosis to treatment to vaccine were all made based on experience. For example, the quinine and remdesivir were considered potential for hospitalized patients. A lot of data could be collected, but there is a huge gap between the data and what's going on. Knowledge was quickly accumulated in the past few months. For example, we know more about hydroxychloroquine, remdesivir, and many other candidate drugs; see, for example (26**?**). However, due to the lack of knowledge, even if there exist some protocols, for example (27–30) and several others, there are still many white and unknown details in the overall medical practice.

DR

Managing uncertainty in medicine. Medical diagnosis is often uncertain or inconclusive. On the other hand, a doctor responsible for a patient's health has to make decisions in spite of this uncertainty. If the uncertainty presents a sufficiently small risk, the doctor can choose a treatment. Otherwise the doctor might consult other doctors, a medical journal or a book.

Inter-Rater agreement

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.

To better understand the process and the possible place of AI in it, we turn to the Kahaneman's (34) "Thinking Fast Thinking Slow" and to Vordermark book on medical decision making (35).

Medical diagnosis can be divided into two main types: recognition and elimination. Recognition is a fast mental process that is partially unconscious where the one correct diagnosis presents itself in the doctors mind. Sometimes the doctor is not able to explain their recognition in words, which hinders discussion and documentation. As recognition typically points to a single diagnosis, there is a danger that the recognized diagnosis will hide other possible diagnoses. Elimination, on the other hand, is a slow deliberate process which starts with all possible diagnoses and gradually eliminates unlikely ones based on patient history, examination and test results. As Elimination is deliberative, it is easier to discuss and document it.

In both recognition and elimination, past experience plays an important role. This experience is based on medical practice as well as knowledge learned from lectures or books.

IA can aid the doctor both in Recognition and in Elimination. On the Recognition side, an IA can sift through massive data and point the diagnostician to suspicious areas.

On the Elimination side, an IA system could help carefully and systematically eliminate diagnoses. This can help the doctor stay aware of possibilities that are not obvious, for differential diagnosis.

How to quantify IDK? We should discuss how to quantify the confidence, or certainty, a physician has when making a decision. Clearly, experience leads to confidence. With more experience aggregated, diagnostic options that contradict the accumulated experience are eliminated, and hence more problems that need to be handled by the elimination process can be handled by the recognition process. However, facing our complicated human body, it is almost not possible for any single physician to aggregate all necessary experience to be confident about anything, so IDK is still an option. A practical and simple way to increase diagnostic certainty is to solicit the experience of a diverse group of doctors via discussion. If there is a clear majority for one diagnostic outcome, then the overall confidence in that diagnostics is high. While this voting procedure might guarantee the optimal outcome, it eliminates the uncertainty during the whole procedure. With this certain procedure, even if the outcome is negative, it can be traced back and accumulate evidence and experience.

Certainty and conditional probability

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 receive this diagnostics have the corresponding condition.

DRAFT

3. 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.

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

4. Human decisions and Intelligence augmentation

Computer has been an integral part of medical practice for decades. From electronic medical records (EMR) to medical instrumentation to billing, hospitals and cliniques cannot function without computers. By some measures computers can already make better diagnosis than human doctors. The question is not whether computer diagnostics will become part of medical practice, the question is how.

Some claim that human doctors and nurses are heading to extinction, following the fate of manufacturing jobs and bank cashiers. Our prediction is that computers will change the nature of medical work, but that it will increase, rather than decrease, the number of healthcare workers, especially in the care of chronic disease and aging, and exploring the nature of our complicated human body.

We believe computers can perform accurate diagnosis for cases where different doctors are likely to agree. In other cases that are in the diagnostic gray area, the computer will output “I don’t know” and transfer the responsibility to the doctor. In most cases, the doctor cannot say “I don’t know” because she is responsible for the patients health. On the other hand, resolving the diagnostic question is not her only choice. She can consult another doctor or the literature, ask for additional tests, or decide on a treatment based on available information. Deciding between these options requires much more than diagnostic information. It involves understanding the patient’s emotional, mental and financial state, the patient’s support system, the strengths and weaknesses of the hospital in which this is taking place etc. Such exploration and results will be fed back to the system to reduce the gray area, which is similar to training an intern doctor in the hospital.

Over time, computers will be able to take into consideration more and more of this complex information. However, for the foreseeable future, it is unlikely that computers will be given the responsibility to make medical decisions. Computers will take on much of the diagnostics and alarm tasks, improving the accuracy and timeliness of the doctors actions. Computers will output IDK in gray areas and will leave the decision making to the human doctor. Giving the computer the authority to make decisions currently done by human doctors will not only deprive the patient the human attention of the doctor, but also put patients in risk.

Some of the digitization of the medicine has come between patients and doctors. A common impression from the learning perspective is that physicians need to record more activities and hence reduce the amount of time on interacting with patients. However, we believe that a properly designed IA that knows IDK can move medicine in the opposite direction, letting the computer make the common noncontroversial diagnostics and giving the patient more time to interact with the patient.

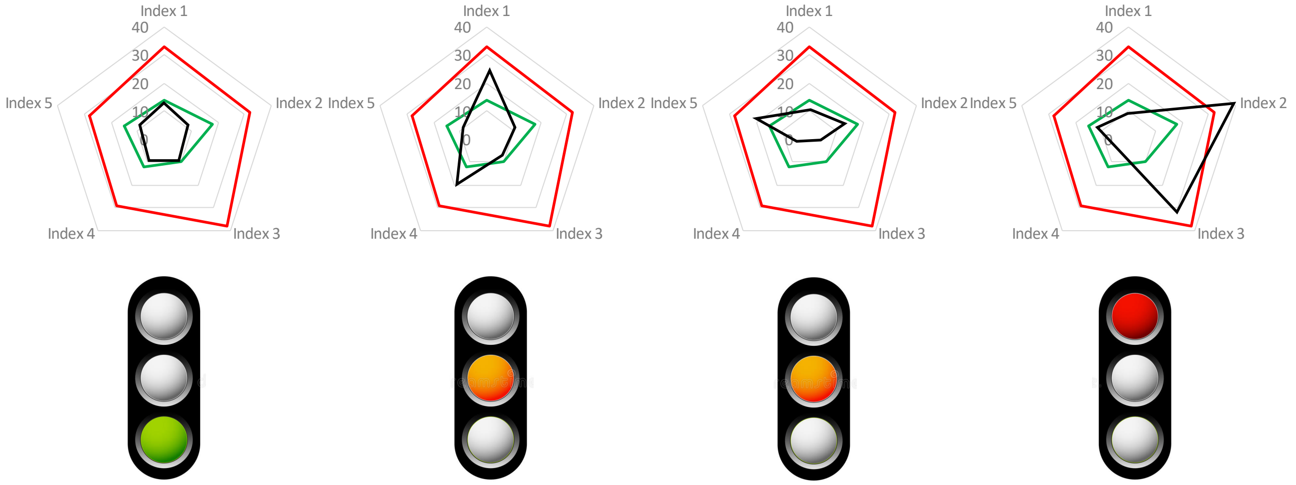

Fig. 1. An illustration of how an alarm system works as a radar system. The red and green pentagons indicate the danger and safe region of five indices. A set of indices inside the green pentagon is safe (shown as a green light). If any index is outside the red pentagon, the patient is surely in danger (shown as a red light). However, if any index is outside the green pentagon but inside the red pentagon, the patient is in a “gray zone”, or marginal situation, the system might not be able to decide the situation, and will report IDK (shown as a yellow light).

For IA technology to be widely adopted, the nurses and doctors that use them should experience an improvement in their practice with the IA system. One example of such system is that the display of the diagnostics computer uses a three color code to identify the pre-defined status. In this system, green indicates a confident negative diagnostic, red corresponds to a confident positive diagnosis, and yellow corresponds to IDK, meaning that the computer cannot confirm or reject the diagnostic outcome. See Figure 1 for an illustration of such a system with a radar display. The thresholds that define these three ranges depend on our knowledge, and the data uncertainty and protocol issues should be taken into account.

With the IA system with IDK, healthcare providers could focus their time on patients overall situation, communication for life plan, or other interactions, and intervene the medical diagnostics when the IA system says IDK.

We finish this section with a few application areas which seem ready for applications of IA.

- **Computer aided diagnostics for large-scale data**

  Medical imaging devices such at digital X-ray, CT, EMR and scanning microscope generate many gigabytes of data for each patient. Radiologists and pathologists spend their days analyzing these images to diagnose the patient. The large size and high resolution of the images on the one hand, and the time limitation on the analyst on the other imply that the analyst has to quickly narrow down the suspicious region, increase the chance of missing dangerous abnormalities.

  IA can help the pathologist by suggesting locations in the high resolution image that might contain cancer nodules ().

  directing her attention to the parts of the image that are

- **Adaptive Patient monitors**

  Alarm fatigue is a well known issue medical providers encounter when working with patient monitors. It is frequently named as a threat to patient safety (36, 37), and a lot of research has been carried out toward this problem (16, 38–40). By further accumulating knowledge, reducing data uncertainty, and improving protocol, it is expected that the gray zone a well developed IA system has is small, and the alarm fatigue issue is alleviated since it only makes an alarm when it runs into IDK. There are many other aspects such an IA system equipped with IDK could help. Since the system knows IDK, it knows what is affirmative. When a medical decision made by a physician falls in the affirmative area, the IA system could help doubly confirm if the decision has any risk not considered by the physician. Such alarm, when sufficiently accurate, could help improve patient risk and healthcare quality. Eventually, this IA system could be evolved into a second opinion provider to healthcare providers.

- **Dissemination of expertise**

  Computers, trained by experts, can help novices. A well-trained IA system equipped with IDK can provide confirmed answers to inexperienced physicians, and serve a function similar to score-cards. Moreover, it can be applied to areas with scarce health resource. The system can provide local healthcare providers knowledge they do not know, and be connected back to physicians with richer medical knowledge when it runs into IDK. On the high level, eventually, we can view an IA system with IDK as a medical specialist full of knowledge and do not make mistake when it knows the answer. When it encounters IDK, it will not hide it. The feedback from experienced physicians, or newly developed knowledge, could be input to decrease the gray areas, and reduce the chance of encountering IDK. Such system in the beginning behaves like an intern doctor, and teaching it is like teaching young diagnostics. Due to the brain capacity and physical limitation, it is impossible for a single physician to know everything in every field, and it is possible that even a very experienced physician could make a mistake. Such a well trained IA system can eventually serve as a reliable second opinion provider to experienced physicians.

DRAFT

5. 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. AM Turing. Can digital computers think? typescript with annotations of a talk broadcast on bbc third programme, 15 may. Technical report, AMT/B/5. In [39]. URL: http://www. turingarchive.org/browse. php/B/5, 1951.
4. W Ross Ashby. An introduction to cybernetics. 1957.
5. Joseph CR Licklider. Man-computer symbiosis. IRE transactions on human factors in electronics, (1):4–11, 1960.
6. Douglas C Engelbart. Augmenting human intellect: A conceptual framework. Menlo Park, CA, 1962.

7. Andrew Ng. What artificial intelligence can and can't do right now. Harvard Business Review, 9, 2016.
8. Eric Topol. Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK, 2019.
9. Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3):276–282, 2012.
10. 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.
11. 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.
12. Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. Methods for artifact detection and removal from scalp eeg: A review. Neurophysiologie Clinique/Clinical Neurophysiology, 46(4-5):287–305, 2016.
13. Ary L Goldberger, Zachary D Goldberger, and Alexei Shvilkin. Clinical electrocardiography: a simplified approach e-book. Elsevier Health Sciences, 2017.
14. George H. Klem, H. Lüders, Herbert H. Jasper, and Christian Elger. The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. Electroencephalography and clinical neurophysiology. Supplement, 52:3–6, 1999.
15. Krisjanis Nesenbergs. Architecture of smart clothing for standardized wearable sensor systems. IEEE Instrumentation & Measurement Magazine, 19(5):36–64, 2016.
16. Maria Cvach. Monitor alarm fatigue: an integrative review. Biomedical instrumentation & technology, 46(4):268–277, 2012.
17. EXECUTIVE BRIEF. Top 10 health technology hazards for 2020. 2020. URL https://www.ecri.org/landing-top-10-patient-safety-concerns-2020. Top 10 patient safety concerns of 2020 from ECRI Institute.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. Paul Abrams. Describing bladder storage function: overactive bladder syndrome and detrusor overactivity. Urology, 62(5):28–37, 2003.
24. Anne G Dudley, Mark C Adams, John W Brock III, Douglass B Clayton, David B Joseph, Chester J Koh, Paul A Merguerian, John C Pope IV, Jonathan C Routh, John C Thomas, et al. Interrater reliability in interpretation of neuropathic pediatric urodynamic tracings: an expanded multicenter study. The Journal of urology, 199(5):1337–1343, 2018.
25. Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). International Journal of Surgery, 2020.
26. Jason D Goldman, David CB Lye, David S Hui, Kristen M Marks, Raffaele Bruno, Rocio Montejano, Christoph D Spinner, Massimo Galli, Mi-Young Ahn, Ronald G Nahass, et al. Remdesivir for 5 or 10 days in patients with severe covid-19. New England Journal of Medicine, 2020.
27. World Health Organization et al. Population-based age-stratified seroepidemiological investigation protocol for coronavirus 2019 (covid-19) infection, 26 may 2020. Technical report, World Health Organization, 2020.
28. World Health Organization et al. Protocol for assessment of potential risk factors for coronavirus disease 2019 (covid-19) among health workers in a health care setting, 23 march 2020. Technical report, World Health Organization, 2020.
29. Kento Nakajima, Hideaki Kato, Tsuneo Yamashiro, Toshiharu Izumi, Ichiro Takeuchi, Hideaki Nakajima, and Daisuke Utsunomiya. Covid-19 pneumonia: infection control protocol inside computed tomography suites. Japanese Journal of Radiology, pages 1–3, 2020.
30. Mohamed E Awad, Jacob CL Rumley, Jose A Vazquez, and John G Devine. Perioperative considerations in urgent surgical care of suspected and confirmed covid-19 orthopaedic patients: Operating room protocols and recommendations in the current covid-19 pandemic. JAAOS-Journal of the American Academy of Orthopaedic Surgeons, 28(11):451–463, 2020.
31. 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.
32. W Price and II Nicholson. Black-box medicine. Harv. JL & Tech., 28:419, 2014.
33. Roger Allan Ford, W Price, and II Nicholson. Privacy and accountability in black-box medicine. Mich. Telecomm. & Tech. L. Rev., 23:1, 2016.
34. Daniel Kahneman. Thinking, fast and slow. Macmillan, 2011.
35. Jonathan S Vordermark II. An Introduction to Medical Decision-Making: Practical Insights and Approaches. Springer Nature, 2019.
36. Sue Sendelbach and Marjorie Funk. Alarm fatigue: a patient safety concern. AACN advanced critical care, 24(4):378–386, 2013.
37. Keith J Ruskin and Dirk Hueske-Kraus. Alarm fatigue: impacts on patient safety. Current Opinion in Anesthesiology, 28(6):685–690, 2015.
38. Christine Weirich Paine, Veena V Goel, Elizabeth Ely, Christopher D Stave, Shannon Stemler, Miriam Zander, and Christopher P Bonafide. Systematic review of physiologic monitor alarm characteristics and pragmatic interventions to reduce alarm frequency. Journal of hospital medicine, 11(2):136–144, 2016.
39. Yong Bai, Duc Do, Quan Ding, Jorge Arroyo Palacios, Yalda Shahriari, Michele M Pelter, Noel Boyle, Richard Fidler, and Xiao Hu. Is the sequence of superalarm triggers more predictive than sequence of the currently utilized patient monitor alarms? IEEE Transactions on Biomedical Engineering, 64(5):1023–1032, 2016. a new alarm system called superalarm to avoid alarm fatigue.
40. Xiao Hu. An algorithm strategy for precise patient monitoring in a connected healthcare enterprise. NPJ digital medicine, 2(1):1–5, 2019.

DRAFT