75227_Auto_Edited.docx

WORD COUNT	6784	TIME SUBMITTED	25-APR-2022 07:21PM
		PAPER ID	85192915

Name of Journal: *Artificial Intelligence in Medical Imaging*
Manuscript NO: 75227
Manuscript Type: MINIREVIEWS

02edEnhancing medical-imaging artificial intelligence through holistic use of time-tested key imaging and clinical parameters: future insights 68

Prakash Nadkarni, Suleman Adam Merchant

Abstract

Much of the published literature in Radiology-related Artificial Intelligence (AI) focuses on single tasks, such as identifying the presence or absence or severity of specific lesions. Progress comparable to that achieved for general-purpose computer vision has been hampered by the unavailability of large and diverse radiology datasets containing different types of lesions with possibly multiple kinds of abnormalities in the same image. Also, since a diagnosis is rarely achieved through an image alone, radiology AI must be able to employ diverse strategies that consider all available evidence, not just imaging information.

Using key imaging and clinical signs will help improve their accuracy and utility tremendously. Employing strategies that consider all available evidence will be a formidable task; we believe that the combination of human and computer intelligence will be superior to either one alone. Further, unless an AI application is explainable, radiologists will not trust it to be either reliable or bias-free; we discuss some approaches aimed at providing better explanations, as well as regulatory concerns regarding explainability ("transparency"). Finally, we look at federated learning, which allows pooling data from multiple locales while maintaining data privacy to create more generalizable and reliable models; and quantum computing, still prototypical but potentially revolutionary in its computing impact.

INTRODUCTION

1. Introduction

As medical knowledge's volume and complexity advances, electronic clinical decision support (CDS) will become increasingly important in healthcare delivery, and increasingly likely to use Artificial Intelligence (AI). Historically, AI approaches have been diverse. However, even senior radiologists, *e.g.*, [1], have inaccurately considered AI, machine learning, and deep learning as synonymous. We therefore summarize these approaches, considering their strengths and weaknesses.

Symbolic approaches: These, the focus of "classical" AI (1950s-1990s), embody the use of high-level abstractions ("symbols") that represent the concepts that humans (often experts) use in solving non-numerical problems. They are most closely related to traditional computer science/software development. In fact, they are mainstream enough that specific terms (instead of "AI") are preferred to describe a given approach. Among the successes:

Business-rule systems (BRS or "Expert Systems") [2]: these allow human experts, working either with software developers or with graphical user interfaces, to embody their knowledge of a particular area to offer domain-specific advice/diagnosis. Robust open-source tools such as Drools [3] are available for building BRS.

Constraint Programming Systems [4]: Constraint satisfaction involves finding a solution to a multivariate problem given a set of constraints on those variables. When the constraints are numeric, techniques such as linear programming [5] (which preceded symbolic AI and is applied in numerous business-operations problems) work better. Some software, such as Frontline Solver(TM) [6] (of which Microsoft Excel's "Solver" add-in is a lightweight version) handles both numerical and symbolic constraints.

Data-driven approaches (Also called "machine learning" or ML): These are used to make predictions, or decisions based on those predictions, by manipulating numbers, or entities transformed into numbers, rather than symbols. They are most useful in

domains where human experts have not formulated problem-solving strategies, but data is available that, if analyzed to discover patterns, can guide such formulation.

Understandably, ML approaches have received a major boost in today's "big data" era. Approaches that employ probabilities, such as Bayesian inferencing [7], have become viable: prior probabilities that could only be guessed at previously (using highly subjective "expert judgment") can now be computed directly from data (e.g., EHRs/public-health registries), with the caveat that these reflect local conditions – *e.g.*, incidence of specific infectious diseases – and will vary with the data source.

All data-driven approaches use iterative mathematical optimization techniques (originally pioneered by Isaac Newton and his contemporaries) to converge onto solutions. In ML parlance, the optimization process is called "training".

ML approaches are subdivided into:

Statistical learning (SL): The use of statistical methods to discover patterns or fit predictive models to data. These techniques originated in the late 19th century (linear regression/correlation), though they have advanced to tackling vast numbers of input variables (also called "features" in ML) and vastly more diverse problems. Human expertise is involved in identifying the features (numeric or categorical) relevant to the problem, and in transforming them to a form suitable for analysis. (For example, a variable comprising of N categories – *e.g.*, gender/race – can be transformed into (N-1) one-or-zero variables using a simple technique called "one-hot encoding"[8].) Almost all SL methods have been developed by researchers with an applied math/statistics background. Individual methods might make specific assumptions about the nature of the variables (e.g., that they have a Gaussian distribution, or that their effects are additive).

Artificial Neural Networks (ANN): (The term "artificial" is typically implied and therefore usually dropped in both the full phrase and the abbreviation.) This family of approaches, which began in the 1950s, also results in the creation of predictive models. It is now prominent enough to deserve its own subsection, below.

1.1. Neural Networks (NNs): Deep Learning

NNs are inspired by the microstructural anatomy and functioning of animals' central nervous systems: software that simulates two or more layers of "neuron"-like computational units ("cells"). Each layer's cells send their output to cells in the next – and in approaches called "recurrent NNs", provide "feedback" to earlier layers as well. However, NNs employ mathematical techniques under the hood, notably mathematical "activation functions" for individual cells. The activation function for a neuron typically transforms inputs of large positive or negative numbers into outputs with a smaller range (e.g., zero to one, or ± 1). An activation function may also incorporate a threshold, i.e., the output is zero unless the input exceeds a particular value.

"Deep" NNs, their modern incarnation, have many more layers than older ("shallow") NNs. ("Deep Learning" is ML performed by DNNs). NNs differ from Statistical Learning in two ways.

NNs make few or no assumptions about variables' characteristics: their statistical distributions don't matter, and their inter-relationships may be non-linear (typically, unknown). Consequently, NNs may sometimes yield accurate predictive models where traditional SL fails.

While NNs can use human-expert-supplied features, they don't have to. For image input, DNNs can *discover* features directly from the raw pixels/voxels. The initial layer discovers basic feature such as regional lines, subsequent layers assemble these into shapes, and so on: LeCun *et al*'s classic Nature paper [9] describes this process, which parallels the cat visual cortex's operation, as discovered by Nobelists David Hubel and Torsten Wiesel [10]. After training, the initial layers can be reused for other image-recognition problems, a phenomenon called *Transfer Learning* (TL) [11]: starting training with layers that recognize basic features is faster than starting from scratch.

TL is also widely used in DNN-based Natural Language Processing (NLP) for medical text: BERT [12], a giant DNN trained by a Google team on the entire contents of Wikipedia and Google Books, was used to bootstrap the training of BioBERT, trained on the full text of PubMed and PubMed Central [13]. Choudhary *et al* [14] review medical-imaging applications of *Domain Adaptation*, a special case of TL, where a DNN trained on a set of

labeled images (e.g., relating to a particular medical condition) are reused for images for a different, but related, condition, either as-is or after an accelerated training process.

This gain in power isn't free.

The number of computations involved goes up non-linearly with the number of layers [15], and so much more compute power is required: notably, abundant random-access-memory (RAM) and the use of general-purpose Graphics Processing Units (GPUs) [16], which perform mathematical operations on sequences of numbers in parallel. (In fact, the theoretical advances embodied in diverse modern DNN architectures would be infeasible without powerful hardware.)

DNNs require vastly more data than SL to discover reliable features which human experts may find obvious. Data volume isn't enough: one must also try to eliminate bias by using diverse data. (We address bias in section 3.)

Certain arithmetic-based issues manifest when the number of layers becomes large - production DNNs can have hundreds of layers - and inputs from each layer pass to the next. Underneath the hood, numbers are being multiplied. When a large sequence of numbers that are all either larger or less than 1 get multiplied repeatedly, the product tends to infinity or to zero: for example, 2 multiplied by itself 64 times $\cong 1.88 \times 10^{19}$.

In DNNs, the consequences of repeated multiplication, called the "Exploding Gradient" or "Vanishing Gradient" problems, can thwart the training process. These are both prevented by **Batch Normalization (BN),** which re-adjusts the numerical values of all the outputs of each hidden layer during each iteration of the optimization training, so that the average of the outputs is zero and their standard deviation is one. Apart from speeding learning, BN allows more layers to be added to the DNN, and hence one can tackle harder problems.

Because of their performance characteristics - DNNs have achieved better accuracy than previous methods, on numerous benchmarks, in a variety of domains - most current AI research focuses on DNNs.

Table 1 summarizes the differences between the symbolic, statistical and DNN approaches.

[Table 1 goes here]

1.2. Training in Machine Learning

ML models can be trained in one of two ways:

Supervised Learning: The objective here is to predict a category (presence/absence or severity of a lesion/disease) or a numeric (interval) value. Category prediction is also called "classification". The training data contains the answers: either in the output variable/s for tabular data, or for images, human annotation/Labeling that identifies specific object categories (including their region of interest, if multiple categories coexist within an image.)

Unsupervised Learning: Here, the objective is to discover patterns in the data, thereby achieving dimension reduction (i.e., a more compact, parsimonious representation of the data).

Semi-supervised learning: The drawback of supervised learning is that for unstructured data (narrative text, images) annotation/Labeling is human-intensive, as well as costly if it involves human expertise that must be paid for. Semi-supervised learning uses a combination of (some) labeled and (mostly) unlabeled data, under the assumption that unlabeled data points close to (or in the same cluster as) labeled data points are likely to share the same category/class.

Statistical learning techniques can be either supervised or unsupervised. Examples of supervised techniques are: Multivariate linear regression/general linear models, which predict interval values; logistic regression and support vector machines, which predict categories; K-nearest neighbor and Classification and Regression Trees (CART), which predict either. Unsupervised SL methods include clustering algorithms, principal components/factor analysis and Latent Dirichlet Allocation.

DNNs, which need very large amounts of data, have motivated the development of semi-supervised methods. They are intrinsically suited for classification. For interval-value prediction with image data, they typically perform or assist in segmentation (which can work with/without supervision), after which numeric volumes can be computed from the demarcated voxels.

1.2.1. Preprocessing

Before training, the data is typically pre-processed with one or more steps. Pre-processing makes the training (and hence predictions) more reliable. The strategies used depend on the kind of data (numeric *vs* image). Some strategies are general, while others are problem specific (we occasionally refer to the latter). Among these steps are:

Detecting suspected erroneous values including unrealistic outliers (e.g., non-physiological clinical-parameter values). The adage "Garbage In, Garbage Out" applies to all facets of computing.

Replacing missing/erroneous values ("imputing"): An entire subfield of applied statistics is devoted to this problem. Strategies include picking the average value across all data points, average value for the individual patient, interpolated values (for time-series data), *etc.* In general, SL algorithms, many of which mandate either imputing all missing values or dropping the data point/s in question, are more vulnerable to missing values than DL.

Standardizing: adjusting numeric values so that disparate variables are represented on the same scale. For variables with a Gaussian ("Normal") distribution, each value is subtracted from the variable's mean and the result divided by the variable's standard deviation, with the sign preserved. For non-Gaussian variables, the value is subtracted from the median and divided by the inter-quartile range. (Batch normalization, discussed earlier, was inspired by standardizing.)

For images, editing out artefacts extraneous to the content to be analyzed - *e.g.*, superimposed text labels or rulers to indicate object size. We come back to this issue later.

1.2.2. Sources of Error: Overfitting and Hidden stratification

A strength of DNNs, stated earlier, is their ability to discover features from raw data. Sometimes, this can also be a weakness: *overfitting* occurs when any ML model is led astray by incidental but irrelevant features in the input. Apart from working unreliably with a new dataset, an overfitted model often making mistakes that humans never would. A DNN for diagnosing skin malignancies used a ruler/scale's presence to infer

cancerous lesions, whose dimensions are usually recorded diligently [17]. Similarly, textual labels on plain musculoskeletal radiographs were confused with internal-fixation implants, lowering accuracy, [18].

Several strategies minimize the risk of overfitting, in addition to making reporting of results more honest:

Cross-validation: The training data is partitioned into a certain number, N (e.g., 10), of approximately equal slices. The training is conducted N times, each time sequentially withholding 1 slice (i.e., only the remaining N-1 slices are used), and the results are averaged.

Withholding of test data from training: A portion of the data is completely withheld from the training process. After the ML model is fully trained with the training data, it is evaluated with the test data, and results are (or should be) reported against the test data only.

Regularization: This is a general term for computational techniques that reduce the likelihood of overfitting during the operation of the training algorithm's optimization phase. The most well-known and general approach is to *penalize model complexity*: the fewer the number of variables that remain in the final trained model, the less the complexity. Originally applied to linear and logistic regression [19], where Lasso and Ridge Regression respectively include penalties that are linear and quadratic in the final number of variables, it is also used for DL.

A regularization approach specific to DLs is **Dropout:** disabling a certain fraction of neurons in hidden layers of a multilayer network during each cycle of training. Li *et al* [20] provide theoretical reasons why dropout can interfere with batch normalization, discussed above, resulting in performance degradation. They recommend that dropout be employed only after the last hidden layer where BN is used, and that the proportion of disabled neurons not exceed 50% (and should usually be much smaller).

A related problem, ***Hidden Stratification*** [21] occurs when a category contains sub-categories ("strata") unrecognized during problem analysis: here, performance on some strata may be poor. Thus, Rueckel *et al* [22]cite an example of severe pneumothorax being

recognized accurately only in those images where a chest tube (inserted to provide an outlet for trapped air) is present [23]. While mild pneumothorax is treated conservatively without a tube, misdiagnosing a yet-to-be-treated, severe pneumothorax has serious consequences.

Nakkiran *et al* [24] had earlier observed the phenomenon of *"double descent."* For some problems, when a DNN classifier is trained on increasingly larger datasets, performance intially gets worse. Later, when the training dataset has become much larger, performance gets better. This could be explained by hidden stratification. The somewhat-larger dataset is heterogenous in unconsidered ways, but the instances of minority sub-categories are too few to learn from, so they only serve to degrade performance. With much larger datasets, these instances become numerous enough to yield a signal that the DNN can use to discriminate more accurately.

2. The Need for a Holistic, System based Approach

   Most recent research in radiology AI has focused on DNNs: the following is just a brief list of DL applications. (This list is not intended to be comprehensive.)

   Binary (Yes/no) classification: Elbow fractures [25] , rib fractures [26], orthopedic implants [27], pneumothorax [28], pulmonary embolism [29], lung cancer [30], pulmonary tuberculosis (where several commercial applications exist) [31].

   Multi-category classification (grading/staging): anterior cruciate ligament injuries [32], hip fracture [33].

   Segmentation with quantitation: Pulmonary edema [34],epicardial fat [35, 36]; gliomas [37, 38]; liver metastases [39, 40]; spleen [41], and brain infarcts [42].

   While impressive, much more is needed to apply AI to realistic problems, especially when intended for deployment in teleradiology scenarios where onsite skill/experience is often lacking. We summarize the issues here before discussing each issue in detail.

   The focus on DNN applications that perform only a single task, while proliferating the number of publications in the literature, does little to advance the likelihood of practical deployment.

Depending on the problem, humans use multiple problem-solving strategies. Similarly, realistic solutions must combine multiple AI approaches, in addition to old-fashioned software engineering (such as intuitive and robust user interfaces).

Good radiologists are also good clinicians. AI must be able to use all available evidence, including collective wisdom gained over decades of experience.

Both humans and AI can be biased; this susceptibility must be recognized. Among the numerous ways to reduce bias, one must consider explainability – the ability to clearly describe the workings of a particular application to a subject-matter expert unfamiliar with AI technology.

2.1. The Limitations of Uni-tasking

As Krupinski notes [1], most DNNs in radiology uni-task. Thus, a DNN specialized for rib-fracture recognition will, even if outperforming radiologists, ignore concurrent tuberculosis, pneumothorax, or Flail Chest, unless trained for the same. For that matter, DNN tuberculosis (TB) diagnosis considering only consolidation/cavitation/mediastinal lymph nodes may miss TB in children. In one series of pediatric patients with pleural effusions, 22% had TB; in 41% of these, effusion was the only radiologic TB sign [43]. We have noticed that these effusions may be lamellar and track upwards, akin to pleural thickening, without being overtly visible, unlike the usual pleural effusions. In fact, in our experience, a lamellar effusion in a child is a good pointer towards the presence of a Primary Complex of TB.

No clinical radiologist uni-tasks: "Savant Syndrome" describes humans with exceptional skill in one area who are mentally challenged otherwise. Overspecialized DNNs suffer, in effect, from perceptual blindness. This phenomenon can be induced experimentally in normal humans by overwhelming their cognitive abilities: in a famous experiment, where subjects had to watch a basketball-game video and count the number of passes one team made, half the subjects failed to notice an intermingling gorilla-suited actor in the center of several scenes [44].

Based on general-purpose vision (GPV) studies, features learned in one specialized uni-tasking recognition problem (e.g., cats) transfer poorly to a related problem (e.g.,

recognizing horses). GPV has advanced because of the public availability of datasets, most notably ImageNet [45], which contain a vast number of object categories, often with multiple categories per image. The images are annotated by crowdsourcing: each object is indicated with a bounding box. Any DL approach expecting to perform well in a challenge to identify these objects cannot be over-specialized. (Unfortunately, DNNs trained on ImageNet perform very poorly with radiology images: transfer learning is not guaranteed to work.)

We believe that focusing short-term on research publications addressing relatively simple problems (with much research being PhD-thesis-driven) retards overall progress. Historically, symbolic AI's notorious addiction to this approach, accompanied by hype that greatly outpaced actual achievement, led to several "AI Winters" [46, 47], steep funding drops following disillusionment. McDermott (a symbolic AI researcher) raised such concerns in a famous 1976 paper, "Artificial Intelligence Meets Natural Stupidity"[48].

2.1.1. Moving toward multi-tasking

There is no reason (besides the costs of compensating radiologists for their time) why radiographic modality-specific ImageNet equivalents cannot be created. Collections of images for trauma patients where multiple lesions are likely to be present may be a good starting point. One could also reuse the vast amount of existing annotated images for uni-tasking-DL research: federated DL (see section 5.1) may help to test new, broader, lesion-recognition algorithms.

While DNNs excel at the important subtask of pattern recognition, they alone would not suffice to move radiology AI into the clinic, as now discussed.

2.2. The Right Strategy for the Right Subtask

Decades of research in cognitive psychology, especially observations of human expertise, have shown that humans use different strategies to different problems. In his classic, "Conceptual Blockbusting" [49], James L.Adams identifies strategies as varied as: general-purpose critical thinking; knowledge of science and mathematics (including calculus); visualization; and applying ethical constraints.

The psychologists Daniel Kahneman and Amos Tversky, founders of "behavioral economics" (Kahneman got a Nobel– Tversky was deceased by then) postulate two modes of thinking. These are "System 1" – "lower level", rapid, intuitive, and reflex ("short-cut")– and "System 2" – "higher level", slow, deliberate, considering multiple sources of information, and requiring concentration. (We return to this work later.) As noted by Lawton [50],DNNs embody System 1 thinking, while statistical and symbolic approaches embody System 2. Both must be used together.

What applies to humans also applies to electronic systems. Symbolic, statistical and NN approaches have been combined in several ways:

In new domains where little practical human experience has accumulated, statistical learning has led to discovery of patterns that can then be encoded as rules or in decision trees, which originated symbolic AI.

While symbolic AI can identify differential diagnosis for a given clinical presentation, statistical AI, using data from local sources or from the literature, can compute probabilities to rank these diagnoses, as well as sensitivity/positive predictive value of individual findings (including test results) to suggest the way forward.

Symbolic approaches are easier for human experts to understand (because they parallel deliberative human problem-solving approaches), and so are often used to "explain" patterns discovered by DNNs. (We discuss explainability in Section 4.)

In radiology AI, Rudie *et al* combine DNN with symbolic/statistical AI (Bayesian networks) for differential diagnosis of brain lesions. Doing this on a large scale across multiple radiology domains has the potential to improve clinical decision making.

2.3. Using All Available Evidence

In sufficiently diverse patient populations, attribution of diagnoses to detected radiographic lesions requires evidence from history, physical exam, non-radiology investigations, plus knowledge of prevalence. Our recommendation to combine all such information to make better decisions is not unique: Kwon *et al* [51] also suggest a Radiology AI that approach that combines multiple evidence sources (imaging plus

clinical variables) for COVID-19 prognostication, while Jamshidi *et al* [52] also recommend a combined approach for COVID-19 diagnosis and treatment.

We provide examples below.

An upper-lobe cavity on a chest X-ray could suggest neoplastic processes, mycobacterial infection, intracellular fungal infection (histoplasma, coccidiosis), *etc.* Serological confirmation plus newer technologies (e.g., GenXPert for tuberculosis [53]) assist diagnosis.

The failure to elicit a proper history can be expensive and traumatizing. One of us (S.A.M.) encountered a young girl who had been repeatedly evaluated under general anesthesia for possible ectopic ureter localization, because of failure to make one simple observation on the plain radiograph. A subsequent Multidetector CT exam concluded erroneously that the incontinence was due to a vesicovaginal fistula, which is extremely rare in children, more so if acquired. This erroneous diagnosis could have been avoided by a simple observation (a slight gap in the pubic symphysis) and one simple question: when did symptoms start? (From birth.) This suggested the correct diagnosis: female epispadias, which a pediatric surgeon confirmed.

Recognizing Midline shift (MLS), plus trans-tentorial and other herniations, allows better triaging for intracranial bleeds or head trauma [54, 55]). Xiao *et al* [56] describe an algorithm to MLS of the brain on CT, with a sensitivity of 94 % and specificity of 100 %, comparable to radiologists.

In head injury, ear-nose-throat bleeds / pneumocephalus suggest basilar skull fractures [57], which are non-displaced and difficult to detect unless looked for diligently.

Pneumothorax diagnosis by DNNs [58], while useful, could increase accuracy for Tension Pneumothorax by additionally looking for simple radiological signs like - inversion of the diaphragm, tracheal shift/shift of mediastinal structures to the opposite side (Figure 1).

AI for rib-fracture recognition [59] can be complemented by the clinical finding of "Flail Chest", which seriously impairs respiratory physiology [60] and may occur when three or more ribs are broken in at least two places.

2.4. Combining AI with Other Technologies

A major thrust of medical AI is in making other technologies, both existing and novel, much "smarter", reducing error by assisting manual tasks and decision-making performed by the radiologist or operator.

Applications in Interventional Radiology: The Royal Free Hospital in London employs an AI-backed keyhole procedure for stenting, coupled with Optical coherence tomography (OCT). While OCT allows viewing the inside of a blood vessel, the AI software automatically measures vessel diameter to enhance decision-making by the interventionist. [61]. Similar roles are possible in interventions such as robotic intussusception–where visualization of the ileocecal junction and reflux into terminal ileum could be taken as end points of the procedure.

AI-assisted 3-D Printing of biological tissue such as heart valves, blood vessel grafts and possibly complete organs is discussed in [62].

3. Biases in Radiology

Artificial Intelligence needs real Intelligence to guide it. Truly intelligent humans are distinguished from the merely smart by intellectual humility and flexibility: as noted in Robson's "The Intellect Trap" [63], they constantly consider the possibility of being wrong, and abandon long-held beliefs when these are invalidated by new evidence. Tetlock's work on human expertise also emphasizes flexibility's importance; both in adapting to reality, as well as in problem-solving strategies. As discussed in section 2.2, AI approaches must be flexible too.

Tversky and Kahneman emphasize that, because of its reflex nature, System 1 thinking is prone to bias. Also, because System 2 requires sustained mental effort (which can cause fatigue), System 1 often contaminates System 2 thought, leading to errors or bias. Busby *et al* [64] cite this work in their excellent article on bias in radiology. An early paper by Egglin and Feinstein considers context bias in radiology [65], where certain aspects of patients' initial presentation to their clinicians led radiologists to give less weight to alternative diagnoses.

Electronic applications can be biased just as humans are. The sources of bias are several.

Symbolic approaches may reflect the biases of their human creators.

Machine-learning approaches that rely on humans to specify relevant features/input variables may be biased if the features chosen are inappropriate, or if relevant features are omitted.

If features are discovered entirely by DL, the data itself may be biased or non-representative. An early version of Facebook's artificial-vision system misidentified bare-chested black males as "primates" [66]because of too few samples in the training data.

4. Explainability of AI

Explainability is the ability to describe the internal workings of a particular AI model (which may apply one or more techniques to a practical problem) to a human expert who intimately knows the problem's-domain but not AI technology. Molnar's book on Interpretable ML [67] is an excellent reference. From this perspective, ML techniques are classified into *"white-box"* (explainable in terms resembling ordinary language), and *"black-box"* models, which cannot be readily explained, because they rely on complex mathematical functions/concepts.

4.1. What determines "Black-Box" *vs* "White-Box"?

Explainability is determined by the following factors:

The choice of technique. In general,

Symbolic AI (and techniques that display output as symbols, such as decision trees) are most understandable/explainable.

Statistical techniques are less explainable. Tversky and Kahneman found in their studies of cognitive errors that people find statistical concepts – such as the phenomenon of regression to the mean due to random processes– more difficult to understand than symbols. In the real-life example of the "Monty Hall problem" [68], at least 1,000 PhDs, including the great mathematician Paul Erdos, had difficulty believing the correct answer, which is an application of Bayesian reasoning that causes a revision of posterior probabilities when new evidence arrives. Therefore, the explainer must often educate the human expert in statistics before addressing the specifics of the application.

In DNNs, the "explanation" is actually a large set of numbers, corresponding to the weights of the inputs of each "neuron" to the neurons to which it connects, along with descriptions of the mathematical transformation/s involved. This is so far removed from everyday experience as to be practically incomprehensible (though there is active research in converting this information into explanatory visuals).

The classification of a particular technique as "black-box" or "white-box" is somewhat arbitrary, depending on the beholder, and on the domain expert's background knowledge. For example, Loyola-Gonzales [69] classifies Support Vector Machines (SVMs) as "black-box". However, SVMs, developed by applied statistician Vladimir Vapnik's group at Bell Labs [70], are mathematically very closely related to regression [71], but try to optimize a different mathematical function (maximized separation between instances of different classes *vs* minimized sum-of-least-squares deviations between observed and predicted values). Multivariate regression (linear, logistic, *etc.*) is taught in enough practically oriented college-level statistics courses for non-statisticians (e.g., business majors, life scientists, medical researchers) to be widely understood.

The complexity of individual problems:

Any model with hundreds of input variables (such as the regression models used by macro-economists) will be intrinsically hard to comprehend.

Business-Rule systems are naturally expressed in ordinary language, and so are in principle, highly explainable. However, R1, devised by McDermott [72] to configure Digital Equipment Equipment's VAX minicomputers based on a customer's needs, eventually used 2,500 rules. Proving that a BRS is internally consistent - that is, no rule contradicts any other rule in the system- is known to be combinatorically hard. "Understanding" the principles of a large BRS does not make it any easier to debug if its output is incorrect.

Whether human-understandable input needs to be modified into an unfamiliar form to make it amenable to computation. This is the case with SVMs when employed for optical character recognition: the image of each letter is converted to a set of numeric

features. In the extreme case, radiographic images are transformed by DNNs from individual pixels into hundreds of features that are "discovered" from the raw data, with each subsequent layer in the DNN representing composite features of increasing complexity.

4.2. The Consequences of Non-Explainability

The concerns about explainability are closely tied to two risks:

Bias: if you cannot explain the application (to a human expert, or to a jury if the application's use is challenged legally), how can you show that it is not biased? "Because the computer says so" is unpersuasive.

Failure: DNNs that process images often make unexplained, bizarre mistakes – misidentifications or failure to identify, as noted by Heaven D[73]. Explanations for such mistakes' origins are not obvious in "post-mortems" even to DNN experts. One approach to forestalling such errors is to deliberately attempt to fool image-classification DNNs by generating "fakes" using another "adversary" DNN to make tweaks (minor or not-so-minor) to authentic images, which are then supplied as training input to the classification-DNN [74].However, while adversarial networks have reduced misidentifications, they do not offer cast-iron guarantees that a mistake will never be made. As in the cliché, absence of evidence (of defects) is not evidence of absence.

Failure can have consequences ranging from the merely frustrating to the near-apocalyptic. A famous example of the latter was the Soviets' satellite-based Early-Missile-Warning System, which, in 1983, flagged 5 missiles from US sites heading toward the USSR [75]. A retaliatory nuclear strike, which would have started World War 3, was averted by Lt. Col. Stanislav Petrov, who reasoned that this was a false alarm – an intentional US attack would need many more missiles – and disobeyed standing orders (to relay the warning up the command-chain) by deciding to wait for confirming evidence, which never arrived.

4.3. Approaches Toward Making "Black-Box" AI More Explainable

In general, such approaches are specific to the problem being addressed, as Molnar makes clear.

One can show the impact of the values of individual input variables/features on the output variable (e.g., categorization, risk score) using a technique called Deep Taylor Decomposition (DTD) [76], based on the Taylor series taught in intermediate-level Calculus. Lauritsen *et al* [77] use DTD as part of an explanation module for predicting four categories of acute critical illness in inpatients based on EHR data. DTD works when the number of input variables is modest (this paper used 33 clinical parameters), and the variables correspond to concepts in the domain. It would not be useful for very numerous, transformed, or automatically discovered variables.

Sometimes, a detailed technical explanation may not be necessary: one can simply test with enough test cases where the system's output matched that of human experts. For images, delineating areas of interest with highlight boxes can draw the user's attention. (This is a standard technique employed by object-recognition systems on benchmark datasets such as ImageNet.) This technique has the drawback that in case of erroneous diagnosis, merely drawing the user's attention to regions of interest may not suffice.

Also, "absence of evidence is not evidence of absence". For a "black-box" system with a critical bug that manifests under uncommon circumstances, you will discover the problem only when it happens. In a complex-system (non-AI) context, Jon Bentley, in his classic work "Programming Pearls" [78] cites a colleague who implemented what he thought was a performance optimization in a FORTRAN compiler. Two years later, the compiler crashed during use. The colleague traced the crash to his "optimization", which had never been invoked in the interim and crashed the very first time it was activated in production.

Loyola-Gonzales [69] suggests combining a white-box and black-box approach (the order depending on the problem) in a pipeline, so that the output of the first is processed into a more human-understandable approach by the second.

4.4. Regulatory Concerns

Certain software applications for tasks previously requiring specialized human skills have already received FDA approval and are in wide use. For example, smartphone-deployable electrocardiogram (EKG)-interpretation programs report standard EKG parameters as well as a few abnormal signals such as Ventricular Premature Beats. Given the increasing deployment of Software as a Medical Device (SaMD), and the possibility of catastrophic medical error when operated (semi-) autonomously, national regulatory bodies are naturally concerned about standardizing the processes of development and testing of SaMD to prevent such errors.

The FDA has specified an action plan, including guidelines for best ML practices, version control when the algorithm is changed, and protection of patient data [79]. The European Commission's proposal for regulation is much wider, encompassing uses of AI across all of society [80]: Human Rights Watch has criticized this proposal [81] on the grounds that it currently does not offer sufficient protection for the social safety net when such software functions autonomously to make decisions concerning, for example, eligibility of individuals for benefits.

5. Future Directions

5.1. Federated Machine Learning

ML in general, and DL specifically, need lots of data to achieve desired accuracy. Volume alone does not suffice: the data must also be sufficiently diverse (i.e., coming from multiple locales) to minimize bias. The obvious solution, physical pooling of data. faces the following barriers:

Data privacy - which is less of an issue with digital radiography, where DICOM metadata containing identifiable information can be removed.

Mistrust – a formidable hurdle when academic or commercial consortia bring rivals together.

The technique of *Federated Learning* (FL), originally pioneered by Google as an application of their well-known MapReduce algorithm [82] allows iteratively training an ML model across geographically separated hardware: the ML algorithm is distributed,

while data remains local, thereby ensuring data privacy. It can be employed for both statistical and deep learning.

Typically, a central server coordinates computations across multiple distributed clients. At start-up, the server sends the clients initialization information. The clients commence computation. When each client is done, it sends its results back to the server, which collates all clients' results. For the next iteration, the server sends updates to each client, which then computes again. The process continues until the ML training completes convergence.

FL's drawbacks are Internet-based communication overhead, which limits training speed, and greater difficulty of analysis of any detected residual bias. Ng *et al* [83] provide a detailed technology overview. Sheller *et al* [84]use FL to replicate prior analysis of a 10-institution brain-tumor-image-dataset derived from The Cancer Genome Atlas(TCGA). Sarma *et al* [85] describe 3-institution FL-based training on whole-prostate segmentation from MRIs, while Navia-Vasquez *et al* [86] describe an approach for Federated Logistic Regression.

In balance, FL's finessing of data privacy issues enables addressing of problems at scales not previously possible, with the greater data volume and diversity ensuring better accuracy and generalizability.

5.2. Quantum Computing

See our previous work, Merchant *et al* [87], for an exploration of this rapidly progressing and revolutionary field. Here, we only provide a basic introduction and address some issues not covered in that paper.

Quantum mechanics describes the rules governing the properties and behavior of matter at the molecular and subatomic levels. Established technologies such as digital photography and nuclear radiography (based on the photoelectric effect), the integrated circuit (based on semi-conduction of electricity by certain materials), and the laser (based on coherent emission of photons) are all applications of quantum mechanics.

Quantum computing (QC) uses the phenomenon of *quantum superposition*, in which matter at the atomic/subatomic level can exist (briefly) in two different states

simultaneously, as the basis for computing hardware design. Unlike the bit in an ordinary computer, which can be either 1 or 0, the quantum bit ("qubit") can be both 1 and 0 simultaneously, so that an array of N qubits could represent 2^N states simultaneously.

QC can, in theory, help solve certain computational problems (called NP-hard problems, where NP = "non-deterministic polynomial" [88]). The time taken to solve an NP-hard problem by brute force (i.e., trying out every possible solution, which is the only way to solve such a problem exactly) increases exponentially as the problem size grows linearly. For example, cracking the widely used Advanced Encryption Standard-256 (with 256 bits) would take all the world's (non-quantum) computers working together, longer than the age of the Universe. In 1994, Peter Shor's theoretical work [89] showed that a "quantum computer" with enough qubits could solve a particular NP-hard problem (factoring the product of 2 Large prime numbers, used in AES-256) in polynomial time, making cryptographic attacks feasible.

The physical challenge is to maintain the qubits stable for a sufficiently long time to accomplish some computation (thus far, such stability has been achieved at temperatures close to absolute zero). In addition, for a computer based on qubits, prototypical work suggests that replacing the conducting elements (the interconnecting wires in an integrated circuit) with light-conducting elements (so-called optical computing [90]) may be the way forward [91].

There are also theoretical considerations as to the kinds of problems for which QC will offer benefits. Thus, Aaronson [92] points out that we don't yet know if the class of problems involved in the optimization (training) phase of DNNs will benefit: while we can hope that they do, the simulations must still be performed to show that this will be the case. Similar concerns are echoed by Sarma [93], who expresses uncertainty about the timeline for QC to become commercially feasible.

Despite the risks of hype and disillusion., it may be worth remembering Arthur C. Clarke's dictum about the future: "If an elderly but distinguished scientist says that something is possible, he is almost certainly right; but if he says that it is impossible, he

is very probably wrong." [94]. If quantum computing becomes commercially viable, almost every aspect of computing (and therefore, every technology that depends on computing) will benefit vastly. The Quantum Internet, Intelligent Edge devices, Edge Computing, Quantum Artificial Intelligence, Quantum Artificial Intelligence Algorithms and their applications in Augmented Reality/Virtual Reality and a more immersive Metaverse experience (for teaching/simulations, actual interactions *etc.*); are some of the exciting future developments/enhancements based on Quantum Computing that we have discussed in our previous paper [87].

CONCLUSION

Combining the wisdom (of both knowledge and meta-knowledge – i.e., problem-solving strategies) gained over the years, with the tremendous versatility of AI algorithms will maximize the utility of AI applications in medical imaging for everyday clinical care. However, scaling up the use of multiple algorithmic strategies and sources of evidence is challenging. Because of its sheer diversity and volume, radiologists' experiential knowledge is very hard to encode in a form that allows instant retrieval. This difficulty applies even to its subset, "artificial general intelligence" (AGI), also known as "common sense". Common sense, apart from being not so common across humans, turns out to be surprisingly hard to implement, because of the sheer breadth of information that must be encoded into computable form.

We see two ways forward: the first long-term and less feasible, the second possible today.

Allocating massive effort and resources to create medical/radiology AGI.

Using software technology (including AI) to extend the human mind, much as access to Web search engines has vastly democratized access to considerable specialized knowledge.

In the latter approach, AI technology can be ubiquitous, integrated, and often functioning behind the scenes for tedious, monotonous and time-consuming tasks (as suggested by Krupinski [1], but still leaving humans in control of critical decisions.

75227_Auto_Edited.docx

ORIGINALITY REPORT

0%

SIMILARITY INDEX

PRIMARY SOURCES

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS