bsi.

...making excellence a habit.™

Medical Device White Paper Series

Recent advancements in AI – implications for medical device technology and certification

Anil Anthony Bharath, Imperial College London

1. Introduction

Clearly, a system whose behaviour is impossible to guarantee seems unsatisfactory from a safety or regulatory perspective. To appreciate why components that incorporate artificial intelligence (AI), and specifically, machine learning, might even be considered, it is helpful to gain a little bit of insight into how modern AI systems are being built, and why they are changing the way in which complex software systems are being engineered. The purpose of this white paper is to provide a very brief overview of where AI is being used in healthcare, and why it might be increasingly seen in medical devices. We also consider what specific additional requirements this might place on regulatory requirements in the near future.

1.1 The nature of the recent advancements in AI

Artificial Intelligence vs Machine Learning: What is the difference?

When we speak of modern AI, we are likely to be thinking of software systems – consisting of combinations of components – that perform in a seemingly intelligent way. Typically, this involves a tight integration of hardware and software that works seamlessly to interpret data, control the actions of devices, or interacts in some way with one or more human users. The components will include those dedicated to user interaction, databases, and software layers that interface with operating systems: at first glance, everything we might expect to find in a modern computer application.

But we must ask the following question: are any of the components based on techniques of machine learning? If so, it is likely that the system's behaviour is impossible to express completely: its behaviour can only be described through its interaction on specific examples of data, or states of environment in which an autonomous agent takes decisions and actions.

We have recently witnessed high-profile events around AI. Among these, we might include a system for diagnosis of pneumonia from chest X-rays (Rajpurkar et al., 2017), a virtual assistant for arthritis sufferers (IBM, 2018) and a system that attains gold-standard performance in early diagnosis of sight-threatening physical changes in the human retina (De Fauw et al., 2018). Much of the impetus for the techniques being applied to healthcare arises through very recent and substantial advances in the state-of-the-art for AI. Moreover, scientific papers in top-ranked scientific journals (e.g. *Nature* and *Science*) have trickled into the wider press, generating significant awareness of progress. Many of these advances in AI rely heavily on machine learning. Compared to human-engineered systems, techniques based on machine learning (ML) develop their own rules of behaviour, by learning from examples or from rewards. Central to many of the techniques being used in modern machine learning is the artificial neural network (ANN): it is built from a relatively small number of types of software units that are designed to mimic the behaviour of biological neurons.

By cascading layers of such units, and providing many parallel channels for the flow of data (e.g. input from sensors, medical image scans, even patient records in the form of text), these layered structures can be taught to make decisions about the presence or absence of a disease (Esteva et al., 2017), keep an eye on radiation dosage during imaging (Tian et al., 2016), or trigger a circuit to regulate a heart rhythm (Figuera et al., 2016). Once a network has been trained to perform its desired task, its behaviour is specified through what are known as *parameters* or weights. Unlike traditional engineering, the final "design" might consist of a network layout accompanied by tens to hundreds of millions of floating point numbers, rather than documented, human readable software.

Why should we opt for an approach that uses machine learning, rather than one that is carefully engineered to specification? There are two primary reasons: first, it is virtually impossible to hand-engineer the rule-set or system design that some types of device software or functionality require. In the medical device or healthcare setting, this arises not from inadequate functional specification, but rather from the wide diversity and complexity found in human anatomy and physiology. If we are to provide the best care for individual patients, we must tailor devices, diagnostics, interventions and therapies on a patient-by-patient basis, and adapt as the patient's state changes. This requires continuous analysis and decision-making, a potentially clear role for automation.

This brings us to the second reason for using machine learning: the performance of complex systems that learn from examples will often better the performance of those that we can hand-engineer. The best example of this is to be found, at present, in case-studies associated with image-based diagnostics (Bello et al., 2019; De Fauw et al., 2018; Rajpurkar et al., 2017).

But, there is a third reason: machine learning is quickly becoming the *de facto* approach to designing complex systems for analysing data from sensors. In short, it is often easier to train a learning system through examples than it is to use rule-based programming. Indeed, even the use of explicit mathematical models within the design of measurement and control systems is arguably being challenged by recent developments in machine learning.

1.2 Ingredients of modern machine learning

Within the past few years, we have witnessed increased use of layered ANNs that are trained with backpropagation (Domingos, 2015). With deep networks (containing many layers), a significant part of the traditional approach to intelligent systems control or diagnostic instrumentation is no longer required: backpropagation replaces the roles of designing and selecting components for filtering sensor data; many parts of a multi-stage design process, developed over decades, are replaced by a network architecture and appropriate training. We are left having to engineer only the simplest of early processing operations, such as amplification and digitisation.

It is often suggested that the availability of large amounts of data, and fast computer hardware, is responsible for the recent advances of machine learning. This is only partly true. There are three additional factors that are very important. Software tools for designing and supporting backpropagation have made significant contributions; these tools are known as "frameworks" for deep machine learning.
The second factor is a culture in machine learning of heavy focus on reproducibility of results, enabled by wide sharing of code and data. This has allowed rapid progress to be made. The final factor, somewhat surprising, has been something of an exodus of programmers from the use of licensed software libraries or commercially produced development platforms to these open-source frameworks. This is partly because there is a very low financial cost of entry to using the tools of modern machine learning to create the components of intelligent systems.

1.3 AI development and deployment

Assuming suitable access to data, examples, or an environment for training an autonomous agent (see definition, Table 1), an R&D team will typically build a large number of deep networks, varying numbers of layers, units within the layers, and other factors. Good practice suggests separate validation and test datasets are retained, to tune training and assess performance. Typically, thousands of models might be trained, and it is not uncommon to use cloud services (e.g. virtual machines provided by Amazon or Microsoft) during this process. Reliable training is computationally very intensive, but once all training and validation processes have been done, the trained networks, known as models, can be deployed with more modest resources (Han, Mao & Dally, 2016).

An AI system can then be constructed by using one or more trained networks, encapsulating these with traditional software modules, generally controlling the flow of data around and between the trained networks. Thus, the traditional human engineered software and systems engineering still plays a role, but it is much diminished, representing a fraction of overall system complexity.

2. The use of AI in healthcare and industry

2.1 How is AI being used in medicine and healthcare?

Virtually all active diagnostic devices that use software to interpret sensor data could be expected to benefit from incorporating AI into their controlling software. Devices that are deemed to be "good enough" might, of course, not *need* to be improved. But in principle, devices that can capture and make use of more information about a patient's immediate physiological state, or emerging response to treatment, might be expected to yield better patient outcomes. It is in this setting that increased use of AI can be predicted. For an excellent example, consider the use of glucometers in controlling diabetes, and possibility of optimising insulin delivery (Atlas et al., 2012).

If we were to look at where AI is beginning to have commercial impact, we can look at systems that make intensive use of computation: imaging systems are a good example. Here, patient benefit might primarily lie in the use of assistance in the diagnostic process, and there are several examples. Radiomics is the most high-profile of these (O'Connor et al., 2017), where operators are applied to images to extract quantifiable measures to be used as biomarkers. Biomarkers computed in this way – known as imaging biomarkers – are being suggested to perform patient stratification in order that appropriate therapies can be given (Valdes et al., 2016). Cancer Research UK (CRUK) and the European Organisation for Research and Treatment of Cancer (EORTC) have recently produced a biomarker roadmap (O'Connor et al., 2017) to accelerate clinical translation of imaging biomarkers. One of the potential uses is in patient stratification (Parmar et al., 2015a; Parmar et al., 2015b), but other uses exist in early evaluation of treatment effects, and even determining surgical margins (Taylor et al., 2014); both are potentially critical to patient outcome, and there is increasing reliance on machine learning to implement image-based biomarker detection or measurement (Parmar et al., 2015a; Parmar et al., 2015b).

We can also identify situations where new measurement capabilities are enabled by AI. Free-breathing MRI (FB-MRI) scanning (Tison et al., 2018) is one of these, where there is clear patient benefit. In FB-MRI, optical flow[1] algorithms are applied to correct for chest wall and cavity motion. Existing systems on the market are unlikely to employ algorithms for optical flow based on machine learning. But ANNs, such as FlowNet 2.0 (Ilg et al., 2017), exceed the performance of known hand-engineered algorithms for computing optical flow, and such networks are likely to replace human-designed algorithms in the future.

Looking at the other end of sensor complexity, we can consider rather less sophisticated data capture techniques, and the potential in collecting clinical and laboratory measurements and patient outcomes at a very large scale. Here, it is more likely that techniques of machine learning will also be employed: the quantity of data supports learning, and large-scale data acquisition allows both device level and patient level peculiarities to be "averaged out". Inference of patient risk can be finessed in a way that has only – until recently – been achievable with epidemiological studies, or very large-scale clinical trials. Perhaps the best known deployment of this is the feature of Apple Watch 4 for the detection of atrial fibrillation (AF). A proof of concept of AF detection with sensors on Apple Watches, using data from more than around 9,000 patients, suggests that a deep neural network, specified by just over half a million parameters, was able to infer the presence of AF with moderate-to-high degrees of accuracy (Tison et al., 2018).

[1] Optical flow refers to a specific type of algorithm that estimates motion in sequences of images (i.e. video). It is a "hidden" technology component used in many different ways (e.g. autonomous vehicles and hand-held cameras).

The US National Academies of Sciences, Engineering and Medicine have recommended several steps toward improvements in diagnosis: emphasis on early and correct diagnosis of conditions, reduction in diagnostic errors, and the avoidance of "overdiagnosis" (Balogh, Miller & Ball, 2015). Encoding the diagnostic decision process has traditionally been done by carefully designed diagnostic decision trees, informed by clinical expertise, and experience. Examples of such systems are now widely available, open to the public, in the form of online "symptom checkers" (Semigran et al., 2015; Semigran et al., 2016). These include those that are based solely on patients' observations, and some that even support inclusion of blood test results. Most of these systems are static, and non-probabilistic: they do not learn from outcomes. They are known to have deficiencies, some of which can be corrected by adopting machine learning approaches, natural language processing and maintaining histories on the tendencies of individual patients.

At the time of writing, a rather high-profile example of machine learning in this context is embodied by Babylon Health, a subscription-based health service provider.

Though well known for its online booking services for patient/GP consultations within the United Kingdom's National Health Service (NHS), Babylon operates globally, and is developing AI systems for interacting directly with potential patients using natural-language processing. Although it is difficult to predict how successful individual service-oriented offerings, such as Babylon Health and MomConnect (Barron et al., 2018) will do in the long term, the vision of these systems has certainly influenced the thinking around diagnostic healthcare in the near future, not least because of the focus on operating at a large scale, and the recognition that capturing data around the entire diagnostic process is key to improved diagnostic accuracy.

2.2 How is AI impacting device or system design?

One of the striking trends of the past 2 years – increased demand for employees with skill sets that include machine learning – might be attributed to the "Gartner hype-cycle" (see https://www.gartner.com/en/research/methodologies/gartner-hype-cycle). However, a closer look suggests that there are genuine changes underway in the way systems are being engineered. These changes are seen across a wide range of institutions: in the activity of start-ups, within education and in large corporations. Much of this change is fuelled by the ease with which one can create, train and deploy software modules based on machine learning.

But there is an underlying factor that is worth bearing in mind: as more data becomes available to train a modern machine learning system, the performance of the system – in terms of accuracy – gets better.[2] Thus, continual improvement has become an integral part of the attraction of machine learning in creating components of AI systems that read sensor data, text streams, and then adapt the overall behaviour of the AI system accordingly.

In the experience of this writer, there is increased investment into maintaining, as a core activity, teams who are responsible for data curation and performance monitoring. The dawning realisation is that to improve system performance, one should place strong emphasis on ensuring that data used to train a machine learning system are curated and managed. Let us take an example: for an AI module which is trained to detect an arrhythmia in electrocardiograms. Even after we have a system that has been deployed in a device, we would monitor and curate data that are acquired from many patients, identifying outliers in performance, identify incorrect decisions, and add these into the training set for the next generation of algorithm, perhaps with higher weighting. This becomes almost a process rather than an R&D activity, but it is valuable and important because good data curation and review processes have a very high probability of improving system performance within a predictable timescale: data have become a raw material in a very real sense.

The behaviour of autonomous systems might also be expected to follow similar patterns of development: improvements in technology performance will take several forms, but for systems that, say, perform diagnosis, performance improvements will take three main forms (a) reduction in false positives and negatives in detection subsystems (more accurate diagnosis); (b) reduction in cost of a device; and (c) improvements in power-efficiency/energy consumption, important for embedded controllers or implanted devices. The latter two of these will often require human expertise for the near term; the former becomes a process of data curation and retraining.

2.3 How AI is being used in manufacturing?

Manufacturing at scale requires rapid and accurate monitoring to ensure throughput and reduce downtime. Traditional instrumentation methods still prove themselves valuable in this domain, but increasingly one wishes to monitor for multiple possible signs of manufacturing line/process failure. Systems that can be trained to deal with irrelevant changes in a manufacturing set-up allow greater flexibility within a production-inspection setting; intelligent monitors can be re-trained or fine-tuned when manufacturing processes change.

Looking further into the future, one of the likely outcomes of deep machine learning is likely to be a reduction in the mechanical precision of electro-mechanical devices: instead, control policies, possibly based around deep networks, will be learned in end-to-end form. One consequence of this is that construction costs for robots that operate with high precision are likely to be lower. Further, minor changes in production lines will become easier to accommodate through retraining existing robots or automation. Indeed, rather than being programmed from scratch, a new model of robot might learn to perform well-defined tasks – such as grasping and placing objects during component assembly – by imitating the actions of an existing robot (Arulkumaran et al., 2017), or even of a human.

[2] It is worth providing two minor caveats. First, this assumes that the capacity of the network, including the number of units and layers, has not been reached. If it has, the solution is simply to increase *capacity:* more layers, or more units. The second is that, at some point, there is simply no more available information in the data that would improve patient care; practically, we are seldom at this limit.

3. AI Systems and certification in medical devices

3.1 Adoption of AI in medical devices

There are several drivers for the incorporation of AI into medical devices. Some of these might relate to marketing, and the attraction of AI-equipped systems to first-adopters. In areas where the quantity or complexity of data is high, including systems for intensive care monitoring, biomarker interpretation or visual diagnosis (e.g. MRI, dermatoscopes, ultrasound), assistance to a human interpreter can be provided by systems that can suggest diagnoses, or can retrieve cases which are in some sense, similar. This can support differential diagnosis, particularly for rare forms of some disorder; it is more than simply a retrieval system, as the criteria for matching is likely to be quite difficult to construct. Machine learning, informed by the actions of expert human diagnosticians, is likely to provide scalable and highly effective trained networks to disseminate expert knowledge and capability (Esteva et al., 2017).

The ability of systems based on machine learning to make use of ever increasing amounts of data, or, in this case, patient cases and their subsequent outcomes, is likely to drive improvements in performance – both in diagnosis and treatment planning and monitoring. We note, for example, significant research activity around machine learning for application in intensive care units (ICUs), with the aim of performing patient state or outcome prediction (Johnson et al., 2016). These efforts leverage the intensive and continuous recorded monitoring within such environments; with appropriate approaches to learning from wider electronic health records (Shickel et al., 2018), and long-term outcomes, the degree of improvements in sensitivity and specificity possible by AI is likely to yield diagnostic capabilities that are significantly more cost-effective than human teams for the same level of care.

With the use of reinforcement learning, it is also becoming possible to create autonomous agents – often built around deep neural networks – that are able to perform tasks that are extremely difficult for a human engineering team to code. Systems to recommend patient dosage (Chen, Zeng & Kosorok, 2016; Nemati, Ghassemi & Clifford, 2016) represent one example of the complex tasks – requiring previous case studies, patient interaction and response monitoring – that an autonomous agent can be trained to perform; see, for example, Tseng et al.'s work on dosage adaptation (Tseng et al., 2017).

3.2 Can systems based on AI be certified for medical use?

The answer to this question needs to consider the types of algorithm that are used in the AI system. Devices that make use of sensor data that is captured from patients might be expected to have input/output mappings that are very well defined. However, as the number of sensor measurements increases, or the history over which a signal is analysed (i.e. time-scale), there is arguably an increased possibility of unexpected device behaviour. In particular, when the definition of a system's performance is based on the examples that it has been exposed to during network training, specific requirements may need to be met in the certification process.

We should make a distinction between "AI" that refers to partially autonomous behaviour based on a well-defined rule set or schema, and AI which incorporates machine learning in some way. For the former of these, one can see medical devices that contain AI as "business as usual". For the latter, there may be extra requirements on aspects of system design that are traditionally considered under software versioning. For example, all other things being equal, it is possible to entirely change a systems' function by altering the weights of the network. Since weight changes are very easy to effect, and the difference between two trained networks can be difficult to detect, there is the possibility for error.

There are two obvious solutions to this conundrum, depending on the complexity of the function performed by the network, and on the complexity of signals or data on which it operates. Let us take the example of "symptoms checkers" (Armstrong, 2018; Elliot et al., 2015). Semigran et al. (2016) proposed the use of patient diagnostic vignettes, examples of possible diagnostic cues that could be provided by patients. This takes the form of lists of patient symptoms for which gold standard diagnoses by experienced human doctors are maintained.[3] One could envisage that such a test could be expanded to many different types of diagnostic data, and indeed the development of techniques that cope with missing or ambiguous data is an area of active research (Campos et al., 2015).

So, one of the likely mechanisms that would be required for certification of the future will be based on agreed sets of sensor signal examples, symptoms, and even images, which represent an open, accepted standard upon which all candidate systems must perform to the same level of agreement with human expert diagnoses. Precisely how such datasets will be established and agreed upon is unclear, but it is likely to be an essential part of the certification process for medical devices which have certain minimum degrees of complexity.

Once a system has been guaranteed to perform to a required standard, information on the software, hardware and operating systems – the entire software stack – should be captured. In addition, since function is largely determined by network weights, these should be uniquely encoded in some way. For example, a hash or some equivalent of a checksum, could be used to produce a unique signature of the weights that define a particular network's behaviour.

Finally, a question naturally arises: should the *data* or *examples* used to train an ML system also be maintained, or subject to scrutiny? Scrutiny of data might sound impractical: the data sizes are large by definition. But tools can be created to find pathological items in datasets, which might lead to bad clinical decisions or device behaviour; and datasets and trained models can be treated as an inseparable pair, and assigned a joint signature.

[3] The complexity of such a test, however, is dramatically increased if symptoms are provided by *patients* who often give conflicting or imprecise descriptions of what they experience. In this case, though the possible outcomes of the test, and the mapping of symptoms to diagnoses might be the same, the added complexity and "noise" in the system needs to be considered. For the case of natural language processing used as input, there is additional complexity around learning for region-specific dialogues, colloquialisms and, the evolving nature of language.

Given the common practice of retraining networks with new data, having a verification of a unique model and data combination seems a sensible idea as part of the certification process. In addition, the data or examples used in training a network should be archived and guaranteed to be retained as part of the system specification.

Online learning remains an open topic in machine learning. In online learning, models are updated as an inference system is live. For the healthcare setting, we should make the distinction between performance across a large population being periodically updated, and adaptive behaviour within an intelligent device. Online learning is more likely to be useful if applied using large amount of data from several individuals, batched together: this is appropriate to the diagnostic setting. On the other hand, we might consider the actions of a specific device with a single patient as adaptive, patient-specific behaviour. The latter is more likely to fall under the category of reinforcement learning; then, we might imagine that certain bounds of behaviour would have to be defined which hold for all patients: adaptation would then be within those agreed bounds. How to establish such bounds remains to be seen.

This raises, of course, questions around the nature of the *examples* themselves, and how the examples used to train a system using machine learning are labelled according to some form of gold standard. Although, in most fields of diagnostic medicine, there are often recognised gold standards for diagnostic tests, treatment decisions by individual clinicians might diverge. But when an AI system is to be deployed for patient triage, how does one establish the "gold standard" for treatment? For palliative or ICUs, how do quality-of-life considerations come into the picture in establishing the "best" outcome? Here, it seems that we are facing a wider challenge, perhaps to medicine as a whole. It might be that the way forward is to have agreement on standardisation and sharing of clinical records, pooled across hospitals and with fine-grained clinical detail. Thus, the gold standards will emerge, and even the rare events, from which both machines and humans can learn, will be captured.

3.3 Recent activity in the regulatory space

A number of pre-market approvals have recently been provided under the US Food and Drug Administration (FDA) "De Novo" pre-market pathway. The algorithm used in the Apple Watch AF detector is one such example. Diabetes management support tools that provide personalised treatment plans (DreaMed) represent another example.

The latter appears to use rather traditional "control-systems" based models of patients as its form of AI, and so – based on the published information – it is unclear to what extent machine learning is used. The astute reader will note that this is one example in which "AI" in a healthcare setting is unlikely to be using (at the time of writing) modern machine learning, where, for example, the model would implicitly be learned from examples of well-controlled patients.

Several recent FDA "De Novo" approvals are for image-based tools of AI. But the FDA has signalled a different approach that slightly shifts the centre of mass for certification – particularly for AI-related technologies – from a full focus on the "device" or its "performance", to a focus on the company, and its processes. The FDAs Digital Health Software Precertification Pilot Program programme was launched in 2019; in its current form, it leverages the idea of software as a medical device (SaMD), but pays particular attention to the culture of a company sponsoring the SaMD entity, including organisational excellence. This is a surprising move, but appears designed to support companies offering AI for patient management tools, diagnostic or analytical systems which arguably have no physical embodiment in the form of a device, but which could determine the course of treatment, and therefore carry life-or-death implications. A related move from the FDA is the "Breakthrough Device" designation, another pre-market process. Outcomes of public workshops and guidance on these schemes are available from the FDA website.

4. Summary, further reading and references

4.1 Summary

Because of the potential for medical device performance to be systematically improved through capturing data on patient outcomes, and for the ability of machine learning to integrate many sources of patient information, we can expect to see devices that incorporate machine learning to increasingly appear on the healthcare market.

Community effort will be required to agree on ground-truth or labels assigned to data that are used in training AI systems; this may be a role for learned societies. This will be a continuous process, since data and acquisition technologies themselves are constantly evolving. We can expect that curation and versioning of datasets will require dedicated resource that is industry wide.

From a regulatory perspective, it is likely that a key piece of information required to analyse the underlying reasons for adverse events from a device incorporating machine learning will be the data used in its training. With the possible exception of autonomous vehicles, there is perhaps no other domain for which forensic scrutiny of the quality of training data will be quite so critical as for the emerging generation of intelligent healthcare systems.

4.2 Further sources and reading

4.2.1 Technical material

We have evolved; when the author was a PhD student, long hours would be spent photocopying articles to be later perused with a ruler or highlighter (the kind that has real, fluorescent ink – this might date me a bit). My students get their education from four key sources:

- On-line video lecture courses, such as those offered for free by *Coursera*. The *Coursera* co-founder, Professor Andrew Ng of Stanford, has developed several courses on machine learning. Many are accessible to those with an engineering background, even if a tad rusty. I would also suggest his evolving book for those seeking to lead teams for application-specific machine learning (free, as it evolves as a writing project, chapter by chapter). I personally like the style of Geoffrey Hinton's lectures.

- *arXiv* preprints – these are usually full versions of quite technical papers that, not to put too fine a point on it, bypass the peer review process. Some papers on arXiv do make it into print, some don't. In essence, the "peer review" comes from the citation counts of students, academics and researchers as assessed by Google Scholar, and a large contributor to citations is the presence of associated code with each paper.

- Code repositories, provided by students or researchers. This has had a huge effect on the reproducibility of results, and therefore on the advance of machine learning, and particularly deep learning.

- Blogs. Somewhat bewilderingly, technical blogs are widely used by researchers. When someone finds a problem with a piece of code, or learning problem, and a "workaround", they often write a blog about it. These get indexed by search engines. There is no guarantee of correctness, of course, but because of the way that search rankings are produced, the blogs that are trusted and correct tend to rise to the top search results.

4.2.2 Good review articles

Much of the "introductory" material on machine learning is aimed at a technical audience. However, I have found that one of the best pieces of writing is from a computer scientist who strongly encourages wider involvement from those outside the field in shaping ideas for the future: Pedro Domingos' "The Master Algorithm" (Domingos, 2015) provides a summary of the main camps of AI, explaining how machine learning fits into this. Although the book points to a lack of any form of *general* AI, it does explain technical material in a very accessible way.

With regard to healthcare based around the use of natural language processing, Strickland presents an excellent critique (Strickland, 2019), identifying failings of some highly publicised attempts at AI for healthcare, yet acknowledges the potential for approaches that make use of machine learning in creating well-defined components of complete systems. See also a well-curated set of questions relevant to radiology (https://appliedradiology.com/articles/the-real-questions-to-ask-an-ai-platform-vendor).

Finally, for those working in the field of AI for healthcare, who wish to support machine learning as a key technology for building system components, one of the best known proponents and educators in the ML field is writing a book entitled "Machine Learning Yearning" (Andrew Ng). It is aimed at people who are tasked with leading projects that will make use of machine learning (https://www.mlyearning.org/); one can even sign up to receive chapters as they are written!

References

Armstrong, S. (2018) The apps attempting to transfer NHS 111 online. *BMJ: British Medical Journal (Online)*, 360, k156.

Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. (2017) Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine,* 34 (6), 26–38.

Atlas, E., Nimri, R. et al. (2012) Monitoring device for management of insulin delivery. U.S. Patent Application No. 13/203,273.

Balogh, E. P., Miller, B. T. & Ball, J. R. (eds) (2015) *Improving Diagnosis in Health Care.* Washington, DC: The National Academies Press. ISBN 978-0-309-37769-0, DOI:10.17226/21794.

The Caldicott Committee (1999) *Report on the Review of Patient-Identifiable Information.*

Barron, P., Peter, J. et al. (2018) Mobile health messaging service and helpdesk for South African mothers (MomConnect): History, successes and challenges. *BMJ Global Health,* 3, e000559.

Bello, G. A., Dawes, T. J. W. et al. (2019) Deep-learning cardiac motion analysis for human survival prediction. *Nature Machine Intelligence,* 1 (2), 95–104.

Campos, S., Pizarro, L., Valle, C., Gray, K. R., Rueckert, D. & Allende, H. (2015) 'Evaluating imputation techniques for missing data in ADNI: A patient classification study', in A. Pardo and J. Kittler (eds) *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2015. Lecture Notes in Computer Science,* vol. 9423. Cham, Springer..

Chen, G., Zeng, D. & Kosorok, M. R. (2016) Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association,* 111 (516), 1509–1521.

De Fauw, J., Ledsam, J. R. et al. (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine,* 24 (9), 1342–1350.

Domingos, P. (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* London, UK, Allen Lane Publishers (Penguin Press). ISBN 978-0-465-06570-7.

Elliot, A. J., Kara, E. O. et al. (2015) Internet-based remote health self-checker symptom data as an adjuvant to a national syndromic surveillance system. *Epidemiology & Infection,* 143 (16), 3416–3422.

Esteva, A., Kuprel, B. et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542 (7639), 115–118.

Figuera, C., Irusta, U. et al. (2016) Machine learning techniques for the detection of shockable rhythms in automated external defibrillators. *PLoS One,* 11 (7), e0159654.

Han, S., Mao, H. & Dally, W. J. (2016) Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations.* See, also, preprint on arXiv:1510.00149.

IBM (2018) Arthritis Research UK launching a cognitive virtual assistant to provide personalised support. (Press Release). Available from: https://www.ibm.com/case-studies/arthritis-research-uk. [accessed 21 October 2018]. See also https://www.versusarthritis.org/get-help/arthritis-virtual-assistant/

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A. & Brox, T. (2017) Flownet 2.0: Evolution of optical flow estimation with deep networks. *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2.

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A. & Clifford, G. A. (2016) Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104 (2), 444–466.

Nemati, S., Ghassemi, M. M. & Clifford, G. D. (2016) Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. *Conference Proceedings: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, 2978–2981.

O'Connor, J. P., Aboagye, E. O. et al. (2017) Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology,* 14 (3), 169–186.

Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. (2015a) Machine learning methods for quantitative radiomic biomarkers. *Scientific Reports*, 5, 13087.

Parmar, C., Grossmann, P., Rietveld, D., Rietbergen, M. M., Lambin, P. & Aerts, H. J. (2015b) Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in Oncology*, 5, 272.

Rajpurkar, P., Irvin, J. et al. (2017) CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

Semigran, H. L., Linder, J. A., Gidengil, C. & Mehrotra, A. (2015) Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ: British Medical Journal,* 351, h3480.

Semigran, H. L., Levine, D. M., Nundy, S. & Mehrotra, A. (2016) Comparison of physician and computer diagnostic accuracy. *JAMA Internal Medicine*, 176 (12), 1860–1861.

Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. (2018) Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics,* 22 (5), 1589–1604.

Strickland, E. (2019) IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum,* 56 (4), 24–31.

Taylor, F. G., Quirke, P. et al. (2014) Preoperative magnetic resonance imaging assessment of circumferential resection margin predicts disease-free survival and local recurrence: 5-year follow-up results of the MERCURY study. *Journal of Clinical Oncology,* 32 (1), 34–43.

Tian, X., Segars, W. P., Dixon, R. L. & Samei, E. (2016) Convolution-based estimation of organ dose in tube current modulated CT. *Physics in Medicine & Biology,* 61 (10), 3935–3954.

Tison, G. H., Sanchez, J. M. et al. (2018) Passive detection of atrial fibrillation using a commercially available smartwatch. *Journal of the American Medical Association (JAMA) Cardiology,* 3 (5), 409–416.

Tseng, H. H., Luo, Y., Cui, S., Chien, J. T., Ten Haken, R. K. & Naqa I. E. (2017) Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics,* 44 (12), 6690–6705.

Valdes, G., Luna, J. M., Eaton, E., Simone II, C. B., Ungar, L. H. & Solberg, T. D. (2016) MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Scientific Reports,* 6, 37854.

Table 1 Jargon Buster

Term	Lay explanation
Artificial intelligence (AI)	A collection of software components whose collective function mimics that of biological cognition. Examples include everyday technologies such as Apple's Siri, Amazon's Alexa, IBMs Watson, or Google's AlphaGo/AlphaZero.
Autonomous agent	A software algorithm or set of algorithms that is designed to decide on and take actions within some sort of environment. An example would be a "chatbot", whose environment might be restricted to interacting via text messages with a human patient.
Artificial neural network (ANN)	A type of processing architecture based on a very simplified computational model of a biological neuron. In its simplest form, it consists of a weighted summation of input floating-point values, followed by a non-linear "activation" function, yielding a floating-point output. The single perceptron is one such simple model, and multi-layer perceptrons are the predecessors to todays' modern deep networks. Convolutional neural networks (CNNs) are a variant in which weights are shared to cover space or time.
Backpropagation	A remarkably successful algorithm which may be seen as an approach to numerical optimisation in the presence of uncertainty. Goes hand-in-hand with implementations of artificial neurons such that derivatives of weights with respect to outputs of the neuron can be easily calculated, and cascaded up the network to estimate derivatives of weights with respect to the error produced by an untrained or partially trained network. These derivatives, given specific data, are sometimes referred to as "gradients".
Bayesian network	A graph-based model of dependencies and causalities between states (e.g. of a patient), causal relationships, and observations, or measurements. The graph is probabilistic, and updates can be passed through the network to update probabilities using Bayesian inference. Very suitable for situations where some observations are missing, there is conflict, and epidemiological "background" might be changing, for example, under epidemic conditions.
CNN	A form of ANN that is very well suited for sequentially ordered data, images, and audio signals. Relies on the operation of convolution to repeat weight patterns over space. Often combined with some fully connected layers.
Deep learning framework	Encompasses what used to be called "libraries" and "environments". There are some new concepts, in that deep networks are specified by architecture descriptions, and there are therefore tools for describing the architecture of a network, for training, data loading, checkpointing during training, estimating and reporting gradients, and so on. Each type of artificial neuron in a framework will be designed to have analytically computable derivatives, and each unit also has a data flow path not only from input to output, but also implicitly from output back to input: the latter may be seen as a "channel" for the backpropagation signal during training.
Hyper-parameters	The parameters that are used in training a neural network. This might include the initialisation of the weights before training, the rate of learning, or other parameters of the training algorithm (e.g. momentum for back-propagation). Hyper-parameters may also be used to refer to the architecture of the network, such as the number of units in a particular layer, or the numbers of layers.
Machine Learning	A sub-branch of AI in which the rules by which a decision or action are taken are learned through examples, a training process. There is generally minimal specification of the rules of input/output mapping at the time the system is in use (sometimes referred to as "inference time", or rather incorrectly as "test time"). Instead, human engineering effort is confined to an overall architecture (for the case of deep networks), or in selecting the best objective function or loss function(s) to be used in training. There are several subclasses of machine learning algorithms: supervised, semi-supervised are but two examples. There is a further important subclass of algorithms known as reinforcement learning; adversarial network training and curiosity-based learning are other examples of emerging techniques.

Contributors

BSI is grateful for the help of the following people in the development of the white paper series.

Author

Anil Anthony Bharath (FIET) is Professor of Biologically-Inspired Computation and Inference at Imperial College London. He holds a first degree from UCL (Electronic & Electrical Engineering) and a PhD from Imperial. He has conducted research on computer vision and pattern recognition and has worked on medical informatics. He co-founded Cortexica Vision Systems, a company that applies biologically-inspired algorithms and deep machine learning for "visual search": the ability to query image databases based on the visual appearance of items. His current research interest is around architectures of convolutional neural networks; recent publications have reviewed generative adversarial networks and deep reinforcement learning. Professor Bharath was commissioned via Imperial Consultants to provide his independent opinion for this report.

Peer reviewers

John Wilkinson OBE has been Director of Devices at the Medicines and Healthcare products Regulatory Agency since February 2012. Prior to this, he was Chief Executive of Eucomed, the European medical technology industry association. He is a member of the Competent Authorities Medical Devices (CAMD) Executive, which is seeking to enhance collaboration between European member states and the European Commission in developing and managing the EU medical devices regulatory system.
His earlier experience included the role of Director General of the Association of British Healthcare Industries and a number of roles in the medical devices industry, both in the UK and the United States, with Becton Dickinson and the BOC Group. These were followed by a period as Chief Executive of an early stage medical imaging company.
John holds a first degree in Zoology from the University of Aberdeen and an MBA from the University of Warwick. He was awarded an OBE for services to the medical devices industry in the 2010 New Year's honours list.

Kevin Holochwost has been a research and development manager and quality assurance director for prominent medical device manufacturers and has experience in all parts of new product realisation. He has baccalaureate and master degrees in physics and applied mathematics from Stony Brook and Cornell University. He has worked with BSI in the active devices team for several years, where he is one of BSI's many experts on embedded or standalone software and radiation devices.

Joshua Schulman, PhD is Vice President for Clinical, Regulatory and Quality Affairs at MaxQ-AI, a developer of machine learning and AI software devices which assist clinicians in improving patient outcomes and optimising healthcare resources. He was previously Regulatory Affairs Manager at Omrix Biopharmaceuticals, a Johnson and Johnson company which manufactures biological hemostats and sealants. At Omrix, he was responsible for regulatory activities in the United States and EU for biologics and devices. He also served as an Associate at Strategic Analysis, Inc. and at Booz Allen Hamilton, Inc. where he developed R&D strategies and managed research efforts for industry and government and focused on human research policy. In parallel, he trained in the Regulatory Affairs program at Johns Hopkins University. Dr. Schulman completed undergraduate studies at Columbia University, and a PhD and post-doctoral fellowship in Physiology and Neuroscience at New York University. At NYU, he focused on identification of biomarkers for human central nervous system diseases treatments using magnetoencephalography.

Jane Edwards, Head of Communications, Global Product Management, BSI
Jane holds a BSc in Chemistry and an MBA from Durham University. She has over 13 years' experience in the medical device industry, having previously worked for Coloplast in their ostomy and continence business. Jane's experience includes working within the pharmaceutical, chemical and telecoms industries for Glaxo Wellcome, ICI and Ericsson, allowing her to bring depth of knowledge from across many industries and technologies. Her current role in BSI allows her to work with technical reviewers across all disciplines ensuring that all BSI communications are accurate and relevant. She is a member of the European Medical Writers Association.

Paul Sim, Medical Devices Knowledge Manager, BSI Standards
Paul has worked in the healthcare industry for over 35 years, joining BSI in 2010 to lead the organisation in Saudi Arabia where it had been designated as a Conformity Assessment Body. Later, he managed BSI's Unannounced Audits programme. Since October 2015, he has been working with both the Notified Body and Standards organisations looking at how best to use the knowledge, competencies and expertise in both. Previously he held senior RA/QA leadership positions at Spacelabs Healthcare, Teleflex Medical, Smiths Medical and Ohmeda (formerly BOC Group healthcare business). Paul is a member of the Association of British Healthcare Industries (ABHI) Technical Policy Group and Convenor of the ABHI ISO TC 210 Mirror Group. He is Convenor of the BSI Committee that monitors all of the work undertaken by ISO TC 210, and Convenor of the BSI Subcommittee dealing with quality systems. As UK Delegation Leader to ISO TC 210, he is also actively involved in the work of national, European and international standards' committees.

Professor Harold Thimbleby, See Change Fellow in Digital Health, based at Swansea University, Wales. He is Expert Advisor on IT to the Royal College of Physicians, a member of the World Health Organization's Patient Safety Network. Harold has been an expert witness in NHS criminal cases. He has honorary fellowships of the Royal College of Physicians, of the Edinburgh Royal College of Physicians, and of the Royal Society of Arts. He is a popular speaker and has given invited talks in over 30 countries around the world.

Published white papers

- *The Proposed EU Regulations for Medical and In Vitro Diagnostic Devices: An Overview of the Likely Outcomes and Consequences for the Market,* Gert Bos and Erik Vollebregt

- *Generating Clinical Evaluation Reports – A Guide to Effectively Analysing Medical Device Safety and Performance*, Hassan Achakri, Peter Fennema and Itoro Udofia

- *Effective Post-market Surveillance – Understanding and Conducting Vigilance and Post-market Clinical Follow-up,* Ibim Tariah and Rebecca Pine

- *What You Need to Know About the FDA's UDI System Final Rule,* Jay Crowley and Amy Fowler

- *Engaging Stakeholders in the Home Medical Device Market: Delivering Personalized and Integrated Care,* Kristin Bayer, Laura Mitchell, Sharmila Gardner and Rebecca Pine

- *Negotiating the Innovation and Regulatory Conundrum,* Mike Schmidt and Jon Sherman

- *The Growing Role of Human Factors and Usability Engineering for Medical Devices: What's required in the New Regulatory Landscape?* Bob North

- *ISO 13485: The Proposed Changes and What They Mean for You,* Bill Enos and Mark Swanson

- *The Differences and Similarities between ISO 9001 and ISO 13485,* Mark Swanson

- *How to Prepare for and Implement the Upcoming MDR: Dos and Don'ts,* Gert Bos and Erik Vollebregt

- *How to Prepare for and Implement the Upcoming IVDR: Dos and Don'ts,* Gert Bos and Erik Vollebregt

- *Planning for Implementation of the European Union Medical Devices Regulations – Are you prepared?* Eamonn Hoxey

- *Cybersecurity of Medical Devices,* Richard Piggin

- *The European Medical Devices Regulations – what are the requirements for vigilance reporting and post-market surveillance?* Eamonn Hoxey

- *General Safety and Performance Requirements (Annex 1) in the New Medical Device Regulation – Comparison with the Essential Requirements of the Medical Device Directive and Active Implantable Device Directive,* Laurel Macomber and Alexandra Schroeder.

- *Do you know the requirements and your responsibilities for medical device vigilance reporting? – A detailed review on the requirements of MDSAP participating countries in comparison with the European Medical Device Regulation 2017/745,* Cait Gatt and Suzanne Halliday

- *Technical Documentation and Medical Device Regulation - A Guide for Manufacturers to Ensure Technical Documentation Complies with EU Medical Device Regulation 2017/745,* Dr Julianne Bobela, Project Associate; Dr Benjamin Frisch, Senior Associate; Kim Rochat, Senior Partner; and Michael Maier, Senior Partner; all at Medidee Services SA

- *Nanotechnology – what does the future look like for the medical devices industry?,* Professor Peter J Dobson, the Queen's College, Oxford, with Dr Matthew O'Donnell, BSI

- *Developing and Maintaining a Quality Management System for IVDs,* Melissa Finocchio, Project Portfolio Leader, bioMérieux

- *Digital maturity in an age of digital excitement; Digital maturity goes beyond excitement to quality,* Professor Harold Thimbleby

Forthcoming white papers

- *Recognising and reducing digital risk in healthcare: We need to up our game to make digital innovation safe and effective*, Prof Harold Thimbleby
- *Classification issues explained for the IVD market,* Mika Reinikainen
- *Impact and Potential for 3D Printing and Bioprinting in the Medical Devices Industry*, Kenny Dalgarno
- *The convergence of Pharma and Medical Devices*, Barbara Nasto
- *Risk management for medical devices and the new ISO 14971,* Jos van Vroonhoven (working title)

About BSI Group

BSI (British Standards Institution) is the business standards company that equips businesses with the necessary solutions to turn standards of best practice into habits of excellence. Formed in 1901, BSI was the world's first National Standards Body and a founding member of the International Organization for Standardization (ISO). Over a century later it continues to facilitate business improvement across the globe by helping its clients drive performance, manage risk and grow sustainably through the adoption of international management systems standards, many of which BSI originated. Renowned for its marks of excellence including the consumer recognized BSI Kitemark™, BSI's influence spans multiple sectors including aerospace, construction, energy, engineering, finance, healthcare, IT and retail. With over 70,000 clients in 150 countries, BSI is an organization whose standards inspire excellence across the globe.
BSI is keen to hear your views on this paper, or for further information please contact us here: julia.helmsley@bsigroup.com

This paper was published by BSI Standards Ltd

For more information please visit:
http://www.bsigroup.com/en-GB/our-services/medical-device-services/BSI-Medical-Devices-Whitepapers/

BSI National Standards Body
389, Chiswick High Road
London W4 4AL
United Kingdom

T: +44 (0) 845 086 9001
E: cservices@bsigroup.com
bsigroup.com