Should artificial neural networks be used in disease diagnosis and will they replace clinicians in the future?

Abstract

Recent advancements in computer technology have led to considerable progress in the field of artificial neural networks, which are used to perform classification/recognition tasks through machine learning. When combined with the vast amount of clinical data available, neural networks can potentially be a powerful tool for disease diagnosis. Here, the concept of machine learning using neural networks is explained, and its applicability in disease diagnosis is discussed taking into account their inherent advantages and disadvantages. Case studies which evaluate the performance of neural networks versus human clinicians at recognising two common cancers, namely breast and skin cancer, are discussed. Primary research was conducted to understand current opinions and beliefs that need to be overcome and to determine cases where the use of neural networks are likely to be accepted. Although the performance of neural networks running on reasonable hardware have already reached the same level as experienced medical specialists, key technical and implementation challenges, such as the requirement of external validation, still exist for their adoption into healthcare. Furthermore, the majority of patients surveyed want the interaction and emotional reassurance that only a doctor-patient relationship delivers. Therefore, it is unlikely for neural networks to replace clinicians in the near future, but they could be used with highest impact in general practice to quickly identify and classify diseases which are common and where large datasets already exist. Efficiency gains can be expected in situations where general practitioners would suffer from cognitive fatigue (too many patients) or improve the accuracy where the person lacks discrete experience/specialism in recognising the disease.

Introduction

Artificial intelligence (AI) is the field of using algorithm-based applications to simulate a human's mental process and intellectual activity; in essence, AI enables machines to solve problems with knowledge. There are two common types of AI, namely expert systems which are based upon rules, and machine learning which are based upon neural networks. An expert system is a computer system that generates predictions under supervision to emulate human decision making [1]. Expert systems use a knowledge base (data) and an inference engine (a reasoning system based on a set of rules) to arrive at a conclusion. However, they suffer from the knowledge acquisition bottleneck [2], in which its knowledge base and known rules are dependent on humans who contribute these initially [3]. The second type of AI is machine learning [4] and this is where neural networks are prominently used. Machine learning requires a vast amount of data for training, and these systems systematically improve their performance during the process, analogous to a human learning through more *experience*. The overarching goal of machine learning is to be able to outperform humans in making decisions via self-study, without any previous knowledge or rules *a priori*.

Artificial neural networks, also known as connectionist systems, are computing systems inspired by the biological neural networks that constitute animal brains. These neural networks[i] learn to perform a task by considering examples [5], generally without being programmed with task-specific rules. The first ever neural networks have been around since the 1950s, however these neural networks were simple in nature due to the limitations of the computer processing technology at the time, as well as the shortage of data required for training. Due to the increase in computing power, size and richness of available data, the interest in neural networks has been rising in recent years. With the large accumulation of well-documented medical data available today, it is no surprise that neural networks are being considered for various applications in this field. Furthermore, the diagnosis of medical conditions, especially in the case of recognising cancers from examinations, is essentially based on a clinician's experience (that is why the training and qualification of a medical specialist/consultant takes 10-12 years) making the learning ability of neural networks apt for this field.

Fast and accurate disease diagnosis is an integral part of treating any disease as 60-70% of decision making in healthcare is influenced by the results of a diagnosis [6]. It has been shown that diagnosis at early stages of a disease increases the likelihood of a successful recovery; this is especially true for cancer. A study into breast cancer [7] has shown that 90% of women diagnosed at its earliest

[i] For the rest of this work, the word *artificial* is implied whenever neural networks is stated as the biological form of neural networks is not pursued here.

stages survive for at least 5 years after diagnosis, compared to only 15% survival rate of women who were diagnosed during its most advanced stages. With a prompt and accurate diagnosis, clinicians can better assess the risks and benefits of different treatments to deliver individualised health management strategies to treat patients. The overarching goal is to match the right patient with the right treatment at the right time, leading to optimal health and quality of life for the patient.

Medical diagnosis techniques follow two general components, namely data gathering and processing the data [8]. The data gathering process ranges from asking simple questions about a patient's medical history to conducting a physical examination with various diagnostic tests. The second step involves processing the results and findings to conclude and communicate the diagnosis. This essay focusses on assessing and comparing neural networks to other methods during the second phase, that is, the performance of neural networks in examining the gathered information/tests and arriving at a diagnosis for the patient. The effectiveness of neural networks in the diagnosis procedure will be discussed through the evaluation of studies and articles, which show promising performance of neural networks for use cases in various clinical research environments. Neural networks are also benchmarked against human clinicians and primary research was conducted to obtain the opinion of patients in the perceived benefits and hurdles for adoption of this technology in medical diagnosis.

Discussion

Neural network operating concepts

Neural networks at their core consist of a set of generic algorithms which mimic the neurone interactions in the human brain that have roles in recognising patterns. Their main function is to cluster and classify data according to similarities among inputs and it is this ability that can be utilised to diagnose medical conditions according to the group of symptoms presented. To produce a result, the neural network must be trained or taught beforehand by providing a very large dataset with both the input problem and the output answers. This is similar to a test containing just the input (in this case the symptoms and presentations of the disease) and what the desired output (the medical diagnosis) should be, without providing *how* the outputs are derived. The neural network then repeatedly does this test *making small changes to itself* until it achieves the desired output from the given input with high enough accuracy. The two most common ways neural networks adapt themselves to produce the desired output are:

- *Backpropagation*, which involves the neural network tuning itself every time it gets an answer wrong such that it does not repeat the same mistake again, or,

- *Genetic evolution*, which is the application of multiple neural networks that do the test and those that underperform are removed from the group, and only those that remain are replicated and randomly altered to mimic the process of evolution.

Deep learning is the term used to describe neural networks which learn from large, very diverse, highly unstructured (eg. images, text or sound or any combination of these) and yet seemingly inter-connected data sets (as they lead to a common outcome), to predict and classify information.

Understanding the inherent advantages and disadvantages of neural networks

It is important to recognise the technical advantages and disadvantages of neural networks to ensure that they are appropriately applied to a specific application.

The requirement of neural networks to learn through observation of datasets limits their application to problems with pre-existing datasets or where datasets can be obtained (with assumption in a practical digital format or can be digitised for machine learning). When presented with small datasets, neural networks are unable to generalise a pattern between the individual cases and suffers overfitting [9], that is, the neural network directly correlates the given answer with the individual case. The analogy here is memorising an answer to a given question rather than knowing how to arrive at the answer through generalisation. A separate issue, known as *imbalanced* data sets [10], also arises with rare conditions. For example, even if the best neural network today could diagnose a *rare* (<1%) disease accurately 99% of the time, the number of false positives (at 1%) exceeds the number of positive diagnosis due to the rare nature (which is <1% occurrence) of the disease. The above problems can be overcome by increasing amount of data for learning, however for rare medical diseases in which such data can only be collected on the very rare occasions when these are actually diagnosed, neural networks will not be able to perform satisfactorily.

Neural networks rely on raw computer processing power to achieve its performance, and the hardware requirements scale directly with the complexity of the neural network and data input size. With the availability of reasonably priced multi-core/multi-threaded processors and parallel computing, this is no longer an issue compared to earlier times. We are also entering an age in which powerful infrastructure exists in the cloud where online “deep learning” capabilities (in the Google Cloud and Amazon Web Service) are already being offered cost effectively, thereby eliminating the need to own powerful hardware at the point of use or *in-situ*. It is recognised that the use of cloud computing in this context introduces other issues with respect to patient privacy which are outside the scope of this work.

The robustness of neural networks depends also on the quality (in terms of consistent noise content) of the data presented. For example, a study on machine learning reacting to adversarial attacks [11] shows how through the deliberate addition of static noise onto an image, an image recognition neural network can completely change its output whereas a human would still correctly identify the image as if it never changed. This problem can very easily be fixed through simply adding such cases to the training data, however with so many variations, it is difficult to completely eradicate the problem. In the medical field, the quality of the data will often be the result of the examination process and depend not only on the signal-to-noise ratio of the instrument used but also the skill of the human examiner (for example, in freehand ultrasound imaging). This issue highlights the vulnerability of neural networks to varying quality of datasets, whether these are introduced inadvertently via limitations in the examination process chain or deliberately through hacks or attacks.

Despite the above drawbacks, neural networks have many advantages over discrete algorithms (or rules) and humans. One such advantage is their ability to be used for problems in which humans are unable to describe the logic to reach a solution. For example, the recognition of objects in images or videos is something people can do easily, however, individuals are unable to describe how they are able to identify such objects; they just *know*. In this aspect, neural networks can provide a solution through the building of knowledge/inferences from given data sets. Another crucial advantage to neural networks is their ability to work quickly without cognitive fatigue. Cognitive fatigue is used to describe when a person feels mentally exhausted and as a result, the person is affected negatively in terms of productivity and reliability at performing cognitive tasks such as absorbing information or solving problems [12]. As neural networks do not suffer from cognitive fatigue and metal exhaustion, they are able to perform tasks consistently and quickly no matter how much work is given to them. This allows for around the clock usage and thus makes them highly effective at automating complex repetitive tasks such as common classification problems (for example, looking for and deciphering the number plate of a speeding car in a video). Note that one of the most promising applications for neural networks in medical diagnosis is the classification of cancers from images.

The Black Box problem

Neural networks are *black box* in nature [13] which create specific issues in complying with current best practices when used with medical diagnosis. In computing and engineering, the term black box is used to describe a device, system or object that can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings [10]. The black box problem applies to neural networks as studying the structure of the network does not provide any insight as to *how* the network produces its outputs. This is particularly important in medical diagnosis, because when it comes to

making diagnosis with best practice, there is heavy emphasis for *transparency* and *auditability* in the decision making process and the need to justify why/how a predictive model should be used in the clinical environment [14]. Currently, there exists only a few methods for seeing inside the black box to obtain the inherent predictive model, such as saliency maps [15]. These methods are crude as they are limited to highlighting the location of salient features but fail to define the pathological (root-cause) characteristics themselves, which would allow for a definitive predictive model to be reverse-engineered. Furthermore, this becomes increasingly difficult with deep learning as there are often too many individual data points driving the network's predictions to identify specific covariates and this level of explanation/transparency of how the network arrives at its output is just not compatible with its black box nature.

Due to the inability to interrogate a deep learning neural network, caution should be applied when making assumptions on the network's ability to generalise and be applicable in a wide range of situations. For example, a neural network could incorrectly form associations with confounded non-pathological properties. In other words, the neural network does not deduce the desirable cause-and-effect relationship between its given variables, but instead works solely on an observed association or correlation between them (correlation does not imply causation).

The black box nature of neural networks also makes manual alteration or changes difficult as it is almost impossible to understand and therefore change what a network is doing to reach its output. It is hard to make minor algorithmic tweaks to improve the accuracy or reliability of a well performing network; the only way is to train the network with even more data. This makes it even more difficult to understand how the network arrives at its prediction which reduces the trust that we can have of neural networks.

As explained above, the black box problem is a major factor in determining if neural networks could replace clinicians in the future. Since it would be near impossible to determine how the network made a diagnosis, if a diagnosis were incorrectly made, it would be tricky to *debug* how such an error had occurred. For human clinicians, it is relatively easy to go back to think through which decisions caused the diagnosis to be invalid. This is significant in improving the rate of true positives and true negatives (also known as sensitivity and specificity respectively [16]), thus reducing false positives and false negatives. Such incorrect diagnoses have dire consequences in healthcare as they can cause unnecessary psychological stress to patients, unnecessary use of medical equipment and recourses, or in the worst case, not detecting an underlying condition until it is too late for the patient.

Case study: Breast cancer diagnosis

A study in 2019 performed by Shen *et al* [17] describes the use of a machine learning neural network trained with a dataset containing 2478 mammography images which was obtained from CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) [18, 19]. Here, the issues discussed above with the application of neural networks manifests themselves clearly. Due to computational limits (one NVIDIA 8 GB Quadro M4000 graphics card in a realistic use-case scenario), the mammograms fed into the neural network had to be downsized from 4000x3000 (12 Megapixels) to merely 1152x896 pixels (1 Megpixel) with potential loss of information. The study also highlighted that their original neural network model they used was much more likely to overfit and took much longer to train and ended up combining two different neural network models for the study. Although the study stated that the proportion of cancer cases in each data set was consistent, it is unclear if this meant the proportions matched the proportions of cancer cases that occurred in reality or if it was to keep each dataset consistent with one another. For the purposes of this case study, it is assumed there were no imbalanced data sets since the typical occurrence of breast cancer is high at 1 in 8 (12%) [20], and is the most prevalent cancer today.

The study by Debono *et al* [21] in 2014 involved giving 10 human radiographers 500 images from BreastScreen's NSW Sydney West database in a blind/independent test to detect breast cancer. The radiographers all came from Westmead Breast Cancer Institute and had a range of radiographic (median = 28 years) experience, which included mammographic (median = 13 years) and breast screening (median = 8 years) activities. The study accounted for cognitive fatigue of the human radiographers by providing the images in small batches of 30-55 mammograms. There was also no time limit for diagnosis, thus reducing the impact of stress or pressure on the results. To compare these results with that of the neural network, we assume that the occurrence of breast cancer and image quality (eg. in terms of perceived contrast, noise) in the Australian database is similar to the Curated Breast Imaging database used for the neural network. We also assume Australian radiographers are similar in performance to other radiographers worldwide to extend this comparison in general terms.

	Definition	Neural network [18]	Human radio-graphers range [19]	Human radio-graphers average [19]
Sensitivity	% of true positives found	86.1%	76% to 92% (16% variation)	82%
Specificity	% of true negatives found	80.1%	74.8% to 96.2% (21.4% variation)	89%

Table 1: Neural network vs human radiographer performance in detecting breat cancer from mammography

Table 1 immediately shows that the results from human radiographers have significant variation (up to >20%) across individual radiographers as they vary in experience and expertise. To achieve a more consistent result with humans, more than one radiographer will have to analyse each mammogram making the diagnosis process more time consuming or less efficient. The insight here is that neural networks are more useful for reproducible and fast diagnosis of many mammography images due to their faster computation and consistent operation. It is also noted, however, that radiographers achieved a higher average in specificity than the neural network; this suggests that radiographers were better at true negatives which in turn reduces the number of false positives. Reducing false positives is important as they can lead to extensive further investigation or unnecessary treatments, which are costly and timely, often causing unnecessary distress for the patient [22]. This suggests that, for this neural network, a human radiographer usefully complements the neural network by checking it to reduce false positives, whilst keeping the consistency and high sensitivity (true positives) of the network.

Needless to say, the *best* human radiographer in 2014 outperforms the 2019 neural network by a margin that cannot be ignored (better by 5.9% for sensitivity and 16.1% for specificity). However, it must be noted that this high level of performance for humans is only achieved through years of training and specialist on-the-job experience (median 28 years radiography, of which 13 years in mammography and 8 years in breast screening).

Case study: Skin diseases

A 2009 study by Kabari *et al* [23] explores the use of a neural network for diagnosing skin diseases. This network was designed with a range of inputs representing symptoms for various skin diseases and had a range of outputs that represented the corresponding skin diseases. The training

data for the network was collected from Olivet Clinic, Port Harcourt and from the National Skin Centre for Dermatology in Nigeria. After training, the neural network was then tested on 20 random test cases network and achieved a 90% success rate. Recent work in 2020 by Zhang *et al* [24] exhibits 97% sensitivity for the detection of skin cancer with a neural network. The same generalisations in the previous case study is used here to enable comparison with human clinicians. Tran *et al* [25] found that human dermatologists typically had sensitivities of 77-96%, with general practitioners much worse only achieving sensitivities of 24-70%. Again, the neural networks of today are in the same ballpark of performance as the specialist clinician (ie. dermatologist) but the worrying statistic is that general practitioners can be far behind, and so general practitioners need to adopt a variety of strategies to ensure early detection of skin diseases/cancer [26].

Xiu *et al* [27] performed a systematic review of studies published between 2012 and 2019 using meta-analysis to evaluate the diagnostic accuracy of deep learning algorithms versus healthcare professionals in classifying diseases using medical imaging. Like in the above case studies, it was concluded that the performance of neural networks is already on par with the performance of healthcare professionals. The paper also brought to light similar challenges found in this essay, that is the generalisations needed due to differing datasets across the studies in order make head-to-head comparisons. The authors also noted the lack of published parameters of the neural networks preventing others from replicating results, and this has led to the subject of creating documentation guidelines for prediction models in clinical use [28]. The analysis also found biases with respect to internal versus external validation of the performance; internally validated studies stated overly optimistic accuracies for *both* neural networks and healthcare professionals. In the highly regulated clinical environment, external validation of the neural network will always be needed before qualifying neural networks for widespread use.

A common trend in the literature is that neural networks are only applied to common cancers (eg. breast and skin are used in the case studies here) and barely any predictive studies exist on new or rare diseases. This shows the inherent limitation of neural networks where without enough data or a high enough rate of occurrence, problems arise from imbalanced data [20].

A further study by Shen *et al* [29] confirms that neural networks today have performance that is on par with healthcare experts, especially for image recognition where object identification is the main outcome of medical diagnosis. Neural networks trained for classification of physical attributes was found to be the most consistent and reduces the cognitive burden on human experts, increasing the efficiency of healthcare delivery. The paper also made a strong point that neural networks cannot exist without human engagement as the final diagnosis needs to have real world implications.

Survey on peoples' opinions on the use of AI for disease diagnosis.

The importance of human engagement is highlighted from the survey[ii] conducted here. A total of 68 respondents, aged 18 to 65, showed that only 24% of people would trust an AI's (neural network implied) diagnosis over a general practitioner's as shown in Figure 1. The result did not vary significantly depending on respondents' generation/age (<25 years, or >25 years) or whether the respondent had a medical background. This is a particularly surprising result, given that the younger, tech-savvy and computer-reliant generation of respondents did not trust AI over a general practitioner. The irony here is that the literature review showed that the general practitioner performed much worse compared to neural networks, which were on par with specialists.

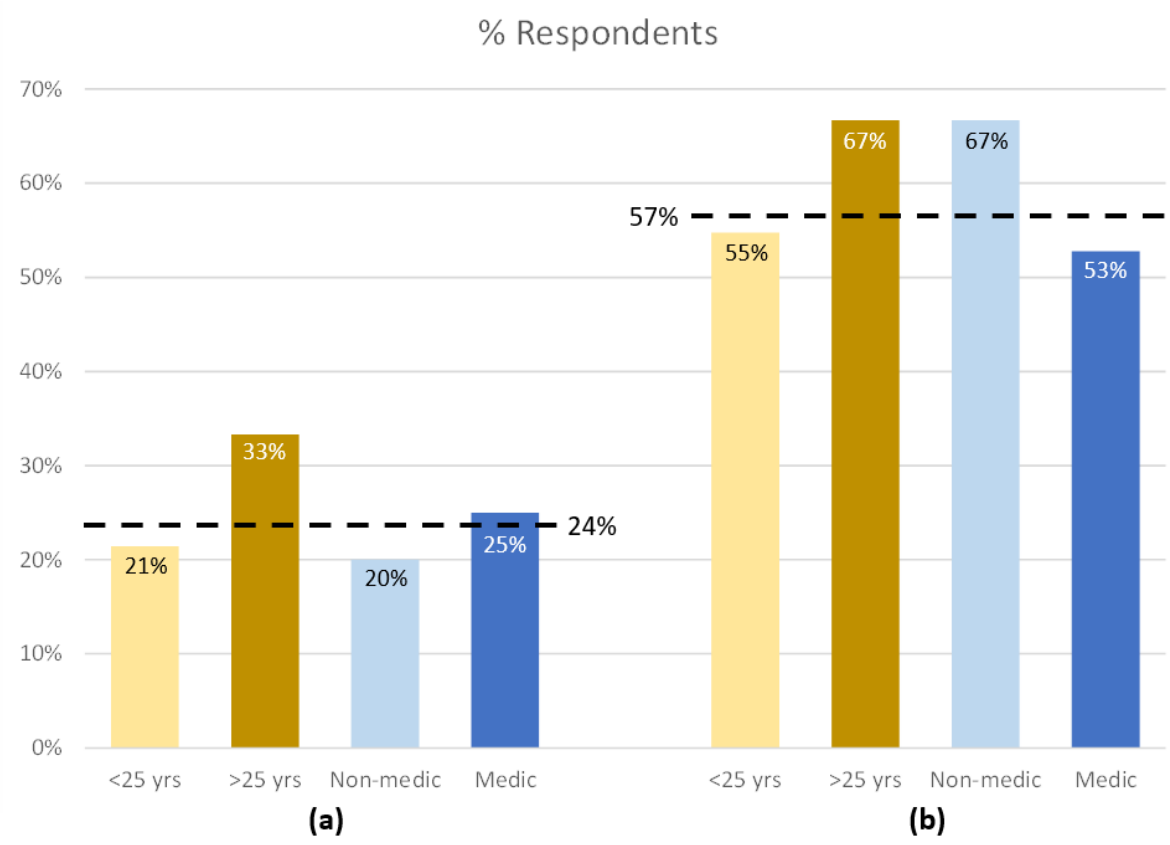

Figure 1: Results of primary research survey; dotted line shows combined average. Responses to (a) Would you trust AI over a general practitioner? (b) Do you think AI would outperform clinicians at diagnosing cancer, skin diseases, diabetes?

[ii] This survey was performed on the social media account of a medical student, and so the respondents likely had above average competence/understanding in the field of medical diagnosis. The sampling method is opportunistic, that is, people with an opinion probably responded whereas others who were impartial ignored it. The demographic of the participants was mainly under 25s, with a small number of older relatives participating.

From the reasons given, it was interesting to read that respondents would in fact be *stressed* if neural networks replaced doctors and so in reality cannot be used in isolation to relieve the workload of general practitioners in community healthcare. Participants highly valued the opinion from a general practitioner as it would provide re-assurance to them personally, some explicitly stating that they have an unconscious distrust for any diagnosis from a machine. This is important as it shows how neural networks would not be able to offer the empathy and reassurance that is provided by a human doctor and how it cannot replace the trust in doctor-patient relationships. Thorsen *et al* [30] found that in general practice, consultations are done to fulfil the expectations of the patient in terms of re-assurance, communication and human interaction. General practitioners have a major role in the mental, social and emotional well-beings of patients which as yet cannot be fulfilled with an AI solution.

On the other hand, when asked specifically if AI (neural networks implied again) would outperform clinicians at classifying specific diseases compared with a clinician, more than half (57%) of respondents would responded positively, again with little variation across age/generation and medic/non-medic. This suggests that in a narrow-spectrum diagnosis, where the disease has clear progression and symptoms, people believe a machine could provide accurate results. In particular, for this scenario, one participant wrote: “With its non-subjective diagnosis based on data, I'm confident of its use. Health professionals may be influenced by limitations of resources or be affected by emotional situations or their personality. Of course, our choice of doctors are based on our wisdom and trust of persons or the team that does the diagnosis…Time is also important, with AI we can access diagnosis quickly, with tests we always have a long wait which causes anxiety.”

Overall, respondents were hopeful that neural networks could be used in combination with clinicians with positive impact, such as reducing diagnosis times and providing objective/independent confirmation; these result in enabling early intervention and causing less anxiety for the patient.

Conclusion

In general, neural networks today perform on par as human specialists with tens of years of experience in classification problems such as detecting cancers from images, achieving sensitivities and specificities well beyond 80% (eg. for breast cancer) and even reaching 97% (eg. for skin cancer). This performance is available today, running neural networks on typical workstation-class hardware, albeit some data management methods such as image size reduction is necessary to keep the computational requirements reasonable.

Despite the large amount of evidence in the literature showing the promise of neural networks for medical diagnosis, the actual implementation of neural networks has not reached commonplace. The challenge in introducing neural networks into the highly regulated field of clinical/medical diagnostics is due to the black box nature of the neural network, which inherently lacks transparency. It is very difficult to validate the network's internal reasoning (no simple set of rules) at arriving at an outcome or make adjustments to how the neural network operates, except by providing more training data. Note that external validation and transparency are pre-requisites for any diagnostic tool to be used with patients in a clinical environment. The applicability of neural networks is also limited to diseases where large data sets exist and the disease has a high rate of occurrence (which need to be higher than the false positives of the neural network); therefore, neural networks cannot be used to detect rare or new diseases.

A survey was conducted where it was shown that the majority of respondents would not *trust* a neural network's diagnosis over that of a general practitioner, despite evidence in the literature that neural networks exceed the performance of a general practitioner in correctly diagnosing conditions from images (to the extent that it is on par with experienced human specialists). The reason is that the patient's expectation from a general practitioner is not only for a diagnosis, but also for the emotional re-assurances that a human interaction can only bring. However, in a clinical/specialist setting, the majority of respondents preferred the neural networks due to its objectiveness, independence and speed.

In conclusion, although neural networks can perform disease diagnosis for classification of common cancers and diseases at the level of experienced human specialists, they cannot yet replace clinicians. The expected emotional support and doctor-patient relationship that is very much present during the patient journey is still a vital component of healthcare. Just like drug development, neural networks will need to go through intense regulatory scrutiny and this process is made difficult by their black box nature. The most effective use case for neural networks today is therefore to address human cognitive fatigue and improve clinical efficiency by complementing medical professionals (especially for general practitioners where there is highest gain in terms of overall accuracy and mental workload) in potential identification of common diseases, although the decisive diagnosis would still need to be delivered by the human clinician.

References

1. P. Jackson, "*Introduction to Expert Systems*", 3rd edition, Addison-Wesley (1999).
2. C. Wagner, "*Breaking the Knowledge Acquisition Bottleneck Through Conversational Knowledge Management*", Information Resources Management **19**(1), 70 (2006).
3. L.R. Medsker, "*Hybrid Intelligent Systems*", 1st edition, Springer (1995).
4. M. Kubat, "*An Introduction to Machine Learning*", 2nd edition, Springer (2017).
5. S.I. Gallant, "*Neural network learning and expert systems*", 1st edition, The MIT Press (1993).
6. C. Goodman, E. Faulkner, C. Gould, A. Smith, C. Aguiar, C. Nelson, A. Grover, A. Berlin, R. Phillips and A. Horan, "*The Value of Diagnostics Innovation, Adoption and Diffusion into Health Care*", The Lewin Group. Inc, (2005).
7. "*Why is early diagnosis important?*" [internet article dated 26 June 2018]. Cancer Research UK [cited 20 Feb 2020]. From: https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important
8. J. P. Langlois, "Making a diagnosis", 2nd edn. New York City: Springer, Boston, MA (1997)
9. R. Caruana, S. Lawrence and L. Giles, "*Overfitting in Neural Nets: Backpropagation, Conjugate Gradient and Early Stopping*", Association for Computing Machinery NIPS'00 Proceedings of the 13th International Conference on Neural Information Processing **13**, 402 (2000).
10. M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker and G.D. Tourassi, "*Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance*", Neural Netw. **21**(2-3), 427 (2009).
11. P. Panda, I. Chakraborty and K. Roy, "*Discretization based Solutions for Secure Machine Learning against Adversarial Attacks*", IEEE Access **7**, 70157 (2019).
12. M. Tanaka, A. Ishii and Y. Watanabe, "*Effects of Mental Fatigue on Brain Activity and Cognitive Performance: A Magnetoencephalography Study*", Anatomy & Physiology: Current Research **S4**, DOI: 10.4172/2161-0940.S4-002 (2015).
13. Y. Bathaee, "*The Artificial Intelligence Black Box and the Failure of Intent and Causation*", Havard Journal of Law and Technology **31**(2), 890 (2018).
14. E.W. Steyeberg and Y Vergouwe, "*Towards better clinical prediction models: seven steps for development and an ABCD for validation*", European Heart Journal **35**(29), 1925 (2014).
15. K. Simonyan , A. Vedaldi and A. Zisserman, "*Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*", presented at Workshop at International Conference on Learning Representations (2014), arXiv:1312.6034 [cs.CV].
16. A.G. Lalkhen and A. McCluskey, "*Clinical tests: sensitivity and specificity*", Continuing Education in Anaesthesia Critical Care & Pain **8**(6), 221-223 (2008).
17. L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride and W. Sieh, "*Deep Learning to Improve Breast Cancer Detection on Screening Mammography*", Scientific Reports **9,** 12495 (2019).

18. R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Goroboy and D.L. Rubin, "*A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research*", Scientific Data **4**, 170177 (2017).
19. M. Heath, K. Bowyer, D. Kopans, R. Moore, W.P. Kegelmeyer and M.J. Yaffe, "*The Digital Database for Screening Mammography*", Medical Physics Publishing Proceedings of the Fifth International Workshop on Digital Mammography, pp 212-218 (2001).
20. "*How common is Breast Cancer?"* [internet article dated 8 Jan 2020]. American Cancer Society [cited 31 May 2020]. From: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html
21. J.C. Debono, A.E. Poulos, N. Houssami, R.M. Turner and J. Boyages, "*Evaluation of radiographers' mammography screen-reading accuracy in Australia*", J. Med. Radiat. Sci. **62**(1), 15 (2015).
22. J.G. Elmore, M.B. Barton, V.M. Moceri, S. Polk, P.J. Arena, S.W. Fletcher, "*Ten-Year risk of False Positive Screening Mammograms and Clinical Breast Examinations*", N. Engl. J. Med. **338**(16), 1089 (1998).
23. L.G. Kabari and F.S. Bakpo, "*Diagnosing Skin Disease Using an Artificial Neural Network*", IEEE Proc 2nd International Conference on Adaptive Science & Technology (ICAST), pp 187-191 (2009).
24. N. Zhang, Y.X. Cai, Y.Y. Wang, Y.T. Tian, X.L. Wang and B. Badami, *"Skin cancer diagnosis based on optimized convolutional neural network"*, Artif Intell Med **102**:101756 (2020).
25. H. Tran, K. Chen, A.C. Lim, J. Jabbour and S. Shumack, "*Assessing Diagnostic skill in Dermatology: A Comparison Between General Practitioners and Dermatologists*", Australas J. Dermatol. **46**(4), 230 (2005).
26. M-L. Rubusam, M. Esch, E. Baum and S. Bosner, "*Diagnosing skin disease in primary care: a qualitative study of GPs' approaches*", Family Practice **32**(5), 591 (2015).
27. X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bryunseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E.J. Topol, L.M. Bachmann, P.A. Keane and A.K. Denniston, "*A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis*", Lancet Digital Health **1**(6):e271 (2019).
28. K.G.M. Moons, J.A.J. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D.G. Altman, J.B. Reitsma and G.S. Collins, "*Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist*", PLoS Med **11**(10), e1001744 (2014)
29. J. Shen, C.J.P Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S.Y. Wong, P-H. Fang and W-K Ming, "*Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review*", JMIR Med Inform **7**(3), e10010 (2019).
30. H. Thorsen, K. Witt, H. Hollnagel and K. Malterud, "*The purpose of the general practice consultation from the patient's perspective – theoretical aspects*", Family Practice **18**(6), 638 (2001).