3.0 Challenges

With such an impressive showing of studies wherein machine learning algorithms were leveraged within healthcare for the classification, inspection, and detection of several serious radiological, dermatological, haemotological and many other such diseases, it might come across as rather surprising to learn that real world deployments of machine learning algorithms within clinical practice are rare [1]. There are many reasons for this, and it is these inherent challenges that plague the widespread adoption and deployment of machine learning algorithms into the field of healthcare as a whole. Within this section several such key challenges will be highlighted and examined through both an ethical and legalistic lens, as well as pertaining to the disconnect between the design of machine learning algorithms and the understanding of those algorithms by medical practitioners.

3.1 Adaptation to a Clinical Setting

As outlined by Kelly et al. in [1], many challenges arise pertaining to the translation of machine learning models into clinical practice. While many studies have been performed that showcase seemingly impressive results within a clinical setting, in reality these studies have been retrospective in nature, that is they have used historically labelled data to train and test algorithms. In reality to best understand how these AI systems perform within a true clinical setting, *prospective* studies must be performed. Prospective studies represent the dynamic nature of a clinical setting, wherein patients are monitored as their medical characteristics or circumstances change as a function of time. In reality, the performance of these AI systems is likely to be worse in prospective studies as compared to retrospective studies, as the real-world data collected from a prospective study will vary substantially from the retrospective data used to train the algorithm [1].

Another problem that presents itself relates to the nature of machine learning literature. As a common practice within the machine learning community, many studies that are performed are published first to pre-print servers, and not to peer-reviewed journals [1]. Due to the stringent nature of the medical community, peer-reviewed evidence will be imperative for the trust and adoption of AI systems within the broader context of the healthcare industry. The common, so called "gold-standard" of medical research is the use of randomized, controlled trials [1, 2], which is not common practice within the machine learning community. As such, high quality reporting of machine learning studies is required for the adoption of these systems into the healthcare industry. For clinicians and medical practitioners to trust and adopt these models into their daily use, there must exist clear and concise reporting on every aspect of these models, such that their usefullness within a clinical setting can be quantified [1].

Clinicians and medical practitioners are foremost qualified in medicine, and are not well versed in the knowledge and terminology of machine learning models; as such, many of the metrics that are often utilized within the machine learning community have no tangible meaning to them. As outlined by Kelly et al., "accuracy does not necessarily represent clinical efficacy" [1]. Area under the curve of a receiver operating characteristic curve is a widely used metric within the machine learning community, however it not necessarily the best metric to represent clinicial applicability as it is not easily understandable.

Similarly, other such metrics such as specificity and sensitivity recorded at a given model operating point, might have no meaning to a doctor who is not well versed in these commonly used terminologies. Arguably the most important factor is that these measure do not reflect what is most important to patients, namely whether the use of the model results in a beneficial change in the care of the patient [1]. Therefore for ML algorithms to be widely adopted within the healthcare industry, these metrics must become more conducive of the clinical understanding of efficacy. Medical students and practicing clinicians should be properly trained in introductory machine learning courses to provide them with the knowledgebase to understand, digest, and adopt these models into their daily workflow.

There are numerous other challenges related to the adoption of machine learning models within a clinical setting. There exists difficulty in comparing models due to the variable methodologies used on datasets with differing distributions. The "dataset-shift" problem also exists, where the introduction of AI models will generate a shift in medical practices, which will cause a change in the dataset used to train these models: effectively chasing a moving target. Current AI models are also far from generalisability, as there exist many unknown variables such as differences in equipment, laboratory equipment, electronic health record systems (EHRs), as well as differences in their codebases. Finally, there exist many logistical difficulties in implementing AI systems, as most healthcare data is not currently stored in a form that is easily digestable by these systems. Data is stored in numerous different mediums, such as medical imaging archival systems, pathology systems, EHRs, insurance databases, and other locations. The amalgamation of all of this data into a single, unifed dataset is a very difficult undertaking [1].

When these AI systems make false decisions in automated diagnoses, the results can sometimes be quite harmful to those affected. Due to the nature of EHRs, the data that these machine learning algorithms are trained on is not always the highest quality, or contains numerous inaccuracies. These so called "data errors" can lead to potentially disastrous results, and these "decision errors" have tangible, real effects that affect real patients [3]. Similarly, Habehh & Gohel highlight that the probabilty of error in prediction as well as the impact of this error are downsides of using these models within a clinical setting. Due to how these models rely on probablistic distributions, there is skepticism as to the validity of the predictions made by these models, and the implications of ML-based algorithms failing to operate as intended have severe and potentially fatal repercussions. Finally, they highlight how the rise of ML within the healthcare industry could fuel a decline in the personal relationship that many patients have between the healthcare practitioner, resulting in potential rejection from the public [4]. For these aforementioned reasons, these ML models are far from widespread adaptation into clinical settings, from the perspective of healthcare practitioners and clinicians.

3.2 The Black Box Problem

Another challenge with the use of machine learning algorithms within the healthcare industry is that of the *black box problem*. While this could fall under the section titled Clinical Adaptation, or perhaps the section titled Legal Issues, the author(s) felt it pertinent that this problem recieve its own dedicated section. While the black box problem does factor into why these models are not widely adopted within the healthcare industry from the viewpoint of a clinician, and it also has heavy legal implications, the problem is so inherently intertwined into the inner workings of machine learning that it was given its own section.

The black box problem relates to the interpretability of a machine learning model, and ones own ability to understand and explain the reasoning as to why and how the output of the model was produced. At its core, this problem stems from how these models are formulated. Large swathes of input data are fed in to train the model to detect patterns that are sometimes anomalous in nature that would be otherwise difficult for a human practitioner to detect. These models then take unforeseen test data and either classify or make a decision based on the type of algorithm. But intrisinic to the functioning of a machine learning algorithm is the manner in which this learning takes place. While we can observe the input and output of the model, we have no way of understanding what is happening within the model itself: this is what is known as the hidden layers.

These hidden layers learn new representations of the input, and each hidden layer is a re-representation of the input, as information is extracted at higher and higher levels of abstraction. As human operators we have no idea what occurs within the hidden layer. For a deep learning model with millions or even billions of parameters, one can imagine the complexity of the interactions between these parameters and the abstraction that occurs. Therefore the effectiveness of these AI models in the healthcare industry is limited entirely by their own inability to explain their decision making in a way that is understandable [1, 5]. As one can imagine, this is entirely problematic within the medical industry. A patient is not going to agree to a treatment without first understanding the decision making process that went into determining which treatment was best for them.

As outlined by Kelly et al., the best performing algorithms are unfortunately the least explainable (deep learning), whereas models with poorer performance are often more explainable (linear regression, decision trees) [1]. If an algorithm ends up making an incorrect decision, or incorrectly recommends something to a healthcare practitioner and they act upon that decision, there is no legal basis for explaining or defending the system as we simply do not know what lead to that decision. This makes it difficult for scientists to understand how data correlates to their predictions, and jeopardizes the faith that the public has in the medical system [2]. Therefore the fact that these algorithms cannot be supervised by medical professionals represents a glaring challenge with their implementation and widespread adoption [6].

3.3 Discrimination and Bias

Another challenge that plagues the implementation of ML into the field of healthcare is that of discrimination and bias. As demonstrated by algorithms implemented into non-medical fields, problematic decisions have been made by algorithms that reflect biases that are inherently woven into the data used to train them [1, 2, 7]. These biases stem from systemic problems with the collection of data, such as under-representation of minorities due to racial biases in dataset development, which in turn will lead to subpar prediction results [2]. Similarly, the implementation of AI systems affect groups that are disadvantaged by factors such as race, gender, and socioeconomic background. Such examples of this within the healthcare industry are mortality prediction algorithms with accuracy that varies by ethnicity [8] and dermatological algorithms that classify both benign and malignant moles with accuracy that could rival a board-certified dermatologist, but with severe underperformance on images of these types of moles when tested on people of colour [9, 10].

The unfairness of an algorithm can be classified into three constituents, such as (1) model bias, where the model elects to best represent the majority but not the under-represented, marginalized groups; (2) model variance, due to inadequate data stemming from minority communities; and (3) outcome noise, such as the effect of unobserved variables that interact with the models predictions [8]. Algorithms may become biased if there have been no medical studies done in certain populations, and as such the data is lacking [7]. The result of this is when the algorithm functions as intended, many people benefit. However when the decision making process falters, those that are negatively affected will be from marginalized communities that are under-represented within the very dataset used to train the algorithm, which compounds societal inequities [11].

Price and Cohen outlined bias in action by considering the allocation of scarce medical resources amongst multiple patients. They propose a scenario wherein a particular minority group responds worse off to the proposed medical intervention then other groups, and as such failure to collect information on this minority group would potentially lead to the algorithm giving the minority group more priority than if the data had been included. If the minority group actually responds better to the intervention, the opposite effect could result [12]. It is therefore important that strategies to minimize the discrimination of marginalized communities and bias stemming from disparities in datasets between races be implemented throughout the development of machine learning algorithms for use in the healthcare sector.

A bias that was surprising to the author(s) was one that did not stem from the data itself, but rather the intent of those designing the algorithms, as highlighted by Char and Shah in [7]. They highlight how the medical industry is entering into a new age of privatized development, wherein TALK ABOUT THE CHAR ARTICLE

3.4 Data Privacy and Ethics

3.5 Legal Issues

References

[1] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, p. 195, Oct. 2019.

[2] B. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, and S. Abdullah, "Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector," *Biomedical Materials & Devices*, vol. 1, pp. 731–738, Sept. 2023.

[3] O. Ali, W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi, "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities," *Journal of Innovation & Knowledge*, vol. 8, p. 100333, Jan. 2023.

[4] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *Current Genomics*, vol. 22, pp. 291–300, Dec. 2021.

[5] J. Guan, "Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance," *Chinese Medical Sciences Journal*, vol. 34, pp. 76–83, June 2019.

[6] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Medical Ethics*, vol. 22, p. 122, Sept. 2021.

[7] D. S. Char, N. H. Shah, and D. Magnus, "Implementing Machine Learning in Health Care — Addressing Ethical Challenges," *New England Journal of Medicine*, vol. 378, pp. 981–983, Mar. 2018.

[8] I. Chen, F. D. Johansson, and D. Sontag, "Why Is My Classifier Discriminatory?," in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Feb. 2017.

[10] H. A. Haenssle, C. Fink, A. Rosenberger, and L. Uhlmann, "Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al.," *Annals of Oncology*, vol. 30, pp. 854–857, May 2019.

[11] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *Journal of Global Health*, vol. 8, p. 020303, Dec. 2018.

[12] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature Medicine*, vol. 25, pp. 37–43, Jan. 2019.