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 those responsible for designing ML algorithms for private sector healthcare corporations might be inclined towards malpractice to bolster their algorithm performance. There may exist temptation towards teaching machine learning systems to guide users towards clinical actions that would not necessarily lead to better care, but would lead to better quality metrics and hospital ratings. These systems might be designed in a manner that when they are being scrutinized by health regulators, they skew the data provided for public evaluation. The authors highlight Volkswagen's algorithm for passing emissions test by reducing their nitrogen oxide emissions when tested. Therefore, one must carefully consider the delivate balance between turning a profit and bolstering ratings, and providing improved healthcare for patients. Those responsible for desigining ML for use in healthcare are most likely not healthcare professionals themselves, and therefore need to seriously consider this tension.

3.4 Data Privacy and Ethics

Arguably the largest concern with the use of machine learning in the healthcare field is that of data privacy and ethics. As we all know, ML algorithms require copious amounts of data to sufficiently train them to a deployable state. In the context of healthcare, there are often issues with acquiring such a large repository of data. This stems from the fact that patient records are often regarded as confidential, and as such many healthcare organizations are not yet ready to offer up this personable information [2]. The deployment of ML systems into the healthcare field has vast implications towards the privacy and confidentiality of data. As early as 1982 some researchers within the medical field were describing the once Hippocratic cornerstone of medicine, confidentiality, as a decrepit and dated concept [13], an opinion that might be more relevant now than ever.

Char and Shah highlight how the implementation of ML as a decision making support tool into the healthcare field will make confidentiality an increasingly difficult concept to maintain, as patients whose data are not recorded cannot benefit from machine learning analyses [7], and the training of ML models will require as much data as possible to ensure clinical accuracy. There also lies the issue in conglomerating health records from all over into one central location, as health records are important and vulnerable data that might be targeted by hackers. There must exist sufficient security measures to prevent data breaches and protect this vulnerable, confidential information [2, 3]. To protect against these kinds of breaches, under the Health Insurance Portability and Accountability Act (HIPAA) patient information is modified by removing a set of 18 specified identifiers, such as names and emails [12]. Murdoch states in [6] however that several emerging computational strategies can be used to re-identify individuals in health data breaches, even if the information has been anonymized and scrubbed of identifiers, with success rates as high as 86% [6, 14–17].

But what if a patient does not want their data to be used to train a ML model? Patient consent therefore becomes a glaring issue, with problems related to informed consent [5, 18], consent to data collection [3, 6, 18], privacy and consent [6], and the extent of consent [12, 19] being topics of debate. Informed consent in the context of data management refers to a persons right to know where and how their data is being used, as well as what kinds of data is being used. In terms of data collection, a patient should have consent as to whether or not their information can be used in the training of an ML model. In most recent years, Google has gotten into trouble with their predatory data collection tactics regarding healthcare data being harvested from patients unknowingly.

In 2017 Google's DeepMind partnered with the British NHS Foundation to use machine learning as an assistive tool in the management of acute kidney injury. What users of this app did not know, however, was that their data was inappropriately being annexed from the United Kingdom to the United states, and this private data was collected without the consent or agency of the user [6, 18]. As Murdoch states in [6], corporations may not always be encouraged to maintain privacy protection if they can monetize the data or otherwise gain from it, provided that the legal penalties are not high enough to offset this behaviour. In a 2018 survey of four thousand American adults, it was found that only 11% were willing to share their health data with tech companies, compared to 72% with physicians. Similarly, only 31% were either "somewhat confident" or "confident" in the abilities of tech companies to safeguard their data [6].

Furthmore, there are concerns about the extent of data collection, and where the line is drawn. Racine et al. highlight that the amount and intrusiveness of health data collection may grow over time as more and more healthcare facilities implement these technologies, highlighting how the development of AI will be used to justify extensive digitization of health status, healthcare practices, treatment use, and even the data collected from wearable devices [19]. It is therefore important to consider consent as a dynamic concept, where a patient can track the use of their data and reaffirm their consent if they still wish to. Giving a patient the ability to monitor the evolution of their data as a function of time will allow the patient to have more control over their data, and help to build a trusting relationship between the patient and the healthcare practitioner or ML engineer responsible for the algorithm using their data. Finally, another issue pertaining to consent is whether or not patients should have a role in deciding how their data should be used [12], specifically should they be able to decide what gets used from their dataset, and if so, how should it be used?

4.0 Case Study: Bias in ML

In a paper titled *Why Is My Classifier Discriminatory?*, Chen et al. examined the concept of fairness in the context of data, and sought to reduce discriminatory classification by not simply constraining the models used, but rather through augmenting the data collection methodologies used [8]. Their paper dealt with automatic decision support systems, like those discussed within this survey. They state that it is often hoped that there will be reduced bias and improved accuracy by training models on observational data, but this is not always the case. Other factors such as data quality and the choice of model however may encode unintentional discrimination, which has a systemically disparate impact [8].

The authors evaluted fairness with respect to protected groups of individuals defined by attributes such as gender and ethnicity, and this paper sought to address the impact of data collection and processign on discrimination and fairness, as this concept was less discussed within the literature compared to constraining the model itself. They describe discrimination in terms of differences in prediciton across all protected groups. In terms of automatic decision support systems, it can often be difficult to balance the tradeoff between predictive accuracy and fairness, as the predictions influence high-states decisions when used in areas such as the healthcare industry [8].

A model was defined as fair if its errors were distributed similarly accross all protected groups, as measured by the cost function γ. Discrimination was defined as referring to specific kinds of differences in the predictive power of models when applied to these different protected groups. Without diving too deep into the mathematical formulation of their predictive methodology, they based their predictions on a set of covariates:

$$X \in \chi \subseteq \mathbb{R}^{\mathrm{k}}$$

and a *protected attribute*:

$$A \in \mathcal{A}$$

Where X represents the medical history of a patient in critical care, A is the self-reported ethnicity, and Y is mortality. Predictions learned from a training set *d* are denoted as:

$$\hat{Y}_d := h(X, A)$$

Fairness was quantified using a popular cost-based definition known as the equalized odds criterion, which states that a binary classifier $\hat{Y}$ is fair if its false negative rates (FNR) and false positive rates (FPR) are equal across all groups. FPR and FNR were defined for a member of a protected group $a \in \mathcal{A}$ as:

$$\text{FPR}_a(\hat{Y}) := \mathbb{E}_X[\hat{Y} \mid Y = 0, A = a], \quad \text{FNR}_a(\hat{Y}) := \mathbb{E}_X[1 - \hat{Y} \mid Y = 1, A = a]$$

The cost function γ was defined as:

$$\gamma_a \in \{\text{FPR}_a, \text{FNR}_a, \text{ZO}_a\}$$

Where ZO_a is the *zero-one loss*, which is a measure of how many mistakes a hypothesis function h makes on the training set, and is given by:

$$\text{ZO}_a(Y) := \mathbb{E}_X[\mathbf{1}(Y \neq \hat{Y}) \mid A = a]$$

They utilized the MIMIC-III dataset, which is a collection of all clinical notes from 25,879 adult patients from the Beth Israel Deaconess Center, and used this dataset to predict hospital mortality of patients in critical care. Fairness was studied with respect to five self-reported ethnic groups, namely Asian (2.2%), Black (8.8%), Hispanic (3.4%), White (70.8%), and Other (14.8%). From the dataset, notes collected in the first 48 hours of an intensive care unit (ICU) stay were used, and patients that stayed for longer than 48 hours were considered. Features were identified by utilizing the tf-idf statistics of the 10,000 most used words from the notes, which assigned a weight to each word based on its frequency relative to the entire document. Their model was trained using 50% of the datam with hyper-parameters selected on 25%, and tested on 25%. The model that they used was a logistic regression model with L1-regularization, and they found that it achieved an AUC of 0.81.

Their findings were that Other and Hispanic groups had the highest and lowest generalized zero-one loss, respectively. The largest ethnic group (White) did not have the best accuracy, whereas the smaller ethnic groups tended towards the extremes. Despite having similar base-rates, Black and Hispanic error rates differed drastically. By increasing the training data size by subsampling and splitting the data, holding out at least 20% for testing, they found that differences in loss decreased with additional training data. Following this, they used clustering to identify clusters where the difference in prediction errors between protected groups was large. This revealed subpopulations for which the differences in zero-one loss were high, which allowed for a more targeted method to reducing discimination. Their work highlights that manipulation of data can be sufficient to improve fairness while maintaining predictive accuracy [8]

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.

[13] M. Siegler, "Confidentiality in Medicine — A Decrepit Concept," *New England Journal of Medicine*, vol. 307, pp. 1518–1521, Dec. 1982.

[14] E. Check Hayden, "Privacy loophole found in genetic databases," *Nature*, Jan. 2013.

[15] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science (New York, N.Y.)*, vol. 339, pp. 321–324, Jan. 2013.

[16] Y. Erlich, T. Shor, I. Pe'er, and S. Carmi, "Identity inference of genomic data using long-range familial searches," *Science*, vol. 362, pp. 690–694, Nov. 2018.

[17] S. Ji, Q. Gu, H. Weng, Q. Liu, Q. He, R. Beyah, and T. Wang, "De-Health: All Your Online Health Information Are Belong to Us," June 2019.

[18] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," in *Artificial Intelligence in Healthcare*, pp. 295–336, Elsevier, 2020.

[19] E. Racine, W. Boehlen, and M. Sample, "Healthcare uses of artificial intelligence: Challenges and opportunities for growth," *Healthcare Management Forum*, vol. 32, pp. 272–275, Sept. 2019.