AI Ethics Journal

Vol 1, Issue 1
Fall 2020

Medical/Healthcare Applications

Ethical Issues Arising due to Bias in Training A.I. Algorithms in Healthcare and Data Sharing as a Potential Solution

Bilwaj Gaonkar Ph.D[1], *Kirstin Cook B.A.*[1], *Luke Macyszyn M.D.*[1]

[1] *Department of Neurosurgery, University of California, Los Angeles*

DOI: https://doi.org/10.47289/AIEJ20200925

Abstract

Machine learning algorithms have been shown to be capable of diagnosing cancer, Alzheimer's disease and even selecting treatment options. However, the majority of machine learning systems implemented in the healthcare setting tend to be based on the supervised machine learning paradigm. These systems tend to rely on previously collected data annotated by medical personnel from specific populations. This leads to 'learnt' machine learning models that lack generalizability. In other words, the machine's predictions are not as accurate for certain populations and can disagree with recommendations of medical experts who did not annotate the data used to train these models. With each human-decided aspect of building supervised machine learning models, human bias is introduced into the machine's decision-making. This human bias is the source of numerous ethical concerns. In this article, we describe and discuss three challenges to generalizability which affect real world deployment of machine learning systems in clinical practice. First, there is bias which occurs due to the characteristics of the population from which data was collected. Second, the bias which occurs due to the prejudice of the expert annotator involved. And third, the bias by the timing of when A.I. processes start training themselves. We also discuss the future implications of these biases. More importantly, we describe how responsible data sharing can help mitigate the effects of these biases – and allow for the development of novel algorithms which may be able to train in an unbiased manner. We discuss environmental and regulatory hurdles which hinder the sharing of data in medicine – and discuss possible updates to current regulations that may enable ethical data sharing for machine learning. With these updates in mind, we also discuss emerging algorithmic frameworks being used to create medical machine learning systems, which can eventually learn to be free from population- and expert-induced bias. These models can then truly be deployed to clinics worldwide, making medicine both cheaper and more accessible for the world at large.

History

Received **30 March 2020**

Accepted **16 September 2020**

Published **23 September 2020**

Keywords

Bias in AI, Ethics, Data sharing, Machine Learning, Algorithmic Bias, Generalizability, AI Training Models, Healthcare, Healthcare Innovation

Contact

Bilwaj Gaonkar:

Department of Neurosurgery, University of California Los Angeles, UCLA Stein Eye Institute, Edie & Lew Wasserman Building, 300 Stein Plaza Driveway, Los Angeles, CA 90095, United States

Email: bgaonkar@mednet.ucla.edu

Acknowledgements

The authors thank Dr. Subramanian Iyer for interesting discussions pertaining to implications of AI Ethics in fields other than medicine to AI in medicine.

Disclosure of Funding

This research was sponsored by NIH R21 EB026665. Luke Macyszyn is a board member and investor of Theseus AI. Dr. Gaonkar is a passive investor in Theseus AI. Theseus AI is a start-up spun out of UCLA intending to commercialize work done at the Macyszyn laboratory.

AI Ethics Journal

Artificial Intelligence, a term increasingly congruent with machine learning – has a wide variety of applications in medicine (Rigby 2019). Machine learning may be defined as the set of algorithms which can 'learn' from data to perform a diagnosis or render a treatment decision - as opposed to being explicitly programmed to do so. This has led to the development of systems which can diagnose Alzheimer's disease (Gaonkar & Davitzikos, 2013; Gaonkar, Davatzikos et al., 2015), which can delineate tumors on images (Bakas et al., 2016), compute natural history efficiently (Gaonkar et al., 2019; Attiah et al., 2019) predict recurrence of cancer (Maczyszyn et al, 2015) and personalize medical treatment (Davenport & Kalakota, 2019). Most of these applications in healthcare stem from the supervised learning paradigm – wherein the machine learns from a suitably large set of data annotated by human experts (Sotiras et al., 2016).

For example, one of the most common applications of artificial intelligence in healthcare involves the application of supervised machine learning as a tool to predict which treatment protocols will succeed based on patient attributes as captured using patient data (Davenport & Kalakota, 2019). Operationalizing such an application requires that we first find a suitable approach to capture information pertaining to the patient in a mathematical form – this process is often called feature extraction (Sotiras et al., 2016; Guyon, 2006). With recent developments in deep neural networks (LeCun, Bengio et al., 2015), some of the feature extraction processes have been automated in narrow domain-specific settings like computer vision (Soh, 2016) or natural language processing (Boag, Wacome et al., 2015). However, manual feature extraction is valuable when translating real-world scenarios into machine-interpretable ones – such as the process of recording patient notesthrough an encounter or when interpretability of the extracted features is desired, as in cases of medical treatment decision support and selection. Features extracted by this process are used to train machine learning algorithms to carry out narrowly defined tasks.

Despite the enormous investments made and progress achieved, machine learning algorithms have failed to achieve translatability from the laboratory to the clinic. A factor in the delayed deployment of these algorithms is the work that still needs to be done to create a maximally generalizable machine. One challenge to generalizability is algorithmic bias. That said, machines are not inherently unethical. It is when humans impose their own views and opinions, whether consciously or not, upon a machine that the resultant model can incorporate these human decisions in the tasks asked of them. Therefore, ethical considerations of artificial intelligence, mainly supervised machine learning, stem from the human decision making within the workflow of creating a model. There are several places in which human input and design can transfer human bias to the model. One of these places where many different types of bias may be taught to a model is in and surrounding the training data collection process. Three types of bias to address related to training data and the resulting machine generalizability are sample bias, annotator bias, and temporal bias. For each of these types of bias, there are both technical and ethical aspects that must be remedied to avoid creating models that amplify these biases. We describe these biases in detail in the next section.

Biases in Training Machine Learning Models

Sample Bias

While machine learning (ML) has been touted as a more objective method of aiding diagnosis and treatment in the medical arena, the objectivity of ML models - especially the

AI Ethics Journal

supervised learning models - is limited by the objectivity of the training data. With the rapid increase in the popularity of ML and efforts to deploy ML models for clinical use, it is necessary to critically evaluate the data used for training these models to ensure that these algorithms are not learning biases from the training datasets. An algorithm becomes unethical to use when implemented in situations for which it was not appropriately trained but is expected to output reliable predictions. In general, a model can be considered biased if used in a setting that was not reflected by the data on which it was trained. Predictions made by this ML model are unreliable because the model has not been trained on similar enough data previously, and thus has not 'learned' accurate predictions for these cases. It is possible that this same model could be ethically implemented in another population that has a similar distribution to the population of the training data. When the cause of the model's bias is due to differences in training and test dataset population distributions, this is termed *sample bias*. Whether a model is used in an ethical way can be defined by the extent that the distribution of the chosen training data mirrors that of the target population.

When training models for use in a clinical setting, oftentimes the source of the data are the electronic health records (EHR). This source of data can be problematic if some of the data are not available or are metadata since machine learning models will learn minute patterns that may arise in these data inconsistencies and anomalies (Gianfrancesco, Tamang et al., 2018). If there are patterns within the training data set from which the machine learns, the resulting model could recognize these incorrect patterns within real-world data and output erroneous predictions (Cabitza, Rasoini et al., 2017). It is possible that the model could also fail to identify other patterns in the real-world data if the training data's distribution differs enough from that of the population for which it is used.

Because inconsistencies and inaccuracies are a reality when using EHR to gather data, depending solely on one indication - like a particular diagnostic code or test result - for even well-defined health outcomes (e.g. hypertension) does not always lead to accurate classifications (Wong, Horwitz, et al., 2018). Two proposed solutions to allow incomplete EHR data to train more predictive models are to include more features in the training data and to use multiple data types to identify a single target (Wong, Horwitz et al., 2018). Including more features is particularly useful when the relationship between the target and the variables is intricate. An example of the utility of this approach is evident when the same treatment is utilized for multiple ailments (Wong, Horwitz et al., 2018; Brown, Haines et al., 2015). Researchers often include information like medication use, procedures, demographics, and sometimes genomic factors, in addition to the standard diagnostic marker features in algorithms (Wong, Horwitz et al., 2018; Wei, Texeira et al., 2016; Shivade et al., 2014).

The second proposed tactic to make incomplete EHR data more useful - using multiple data types to teach an algorithm to predict one target - creates the opportunity for classification of this target variable from multiple types of inputs. Having multiple data types contribute to classification of the target variable can act as a way to lessen the impact that one of these features being omitted or misrepresented in the EHR might have on the model's performance. Some of these data types, like unstructured data, are difficult to extract in a form useful for training algorithms. Machine learning is particularly valuable in this data pre-processing step. Unstructured data like images and free text were estimated to make up 80% of the patient data in EHR systems, and these formats are not easily queried in an automated way (Wong, Horwitz et al., 2018; Murdoch & Detsky, 2013).

AI Ethics Journal

This unstructured data requires manual human work to find and arrange it in a way that can be used in the algorithm training process (Ford, Carroll et al., 2016; Araujo et al., 2017). This can be time-consuming and labor-intensive, which can mean that preparing adequate amounts of this data for training may be cost-prohibitive. Since this means that the amount of training data would be smaller than if the data collection could be automated, the chances of this smaller amount of data having a distribution that differs from that of the general population is greater. This has a greater probability of resulting in a more biased and less generalizable ML model. The more data that is used to train the model, the greater the model's predictive accuracy, and the more ethical the resulting ML model (Halevy, Norvig et al., 2009).

Another threat to generalizability due to sample bias is that vulnerable populations, such as those from lower socioeconomic statuses, are more likely to have incomplete EHR data in any given health care system. This means that models trained using this EHR data are not representative of and cannot be ethically used for patients from these vulnerable populations (Gianfrancesco, Tamang et al., 2018; Arpey, Gaglioti et al., 2017). This would render the models trained using a single EHR system's data non-generalizable, with the possibility of inaccurate and/or biased predictions when tested with data from another population (e.g. a different facility's EHR system). A proposed remedy for this - and to create more generalizable algorithms at a single institution - was to train with more external data that represent more diverse populations (Gianfrancesco, Tamang et al., 2018). There are also people who do not seek healthcare, so no data exists for models to be trained to be inclusive and representative of the populations that these individuals belong to. A model's performance evaluated from an ethical standpoint necessitates clear definitions of the model's intended use, a commitment to its deployment in only this specified way and setting, and appropriate training with representative data. If the data used to train the model is too differently distributed than the population for which the model will be used within, then the algorithm's performance is jeopardized and untrustworthy. In an A.I. enabled world, data shared by one individual might benefit another, and the lack of such sharing may impede another individual from getting the best possible care. This leads to a central ethical question of whether data sharing should be mandated to ensure best possible care for all, or if A.I.-enabled healthcare should be available only to those communities for which adequate data are available to train the required models.

Annotator Bias

When two different physicians see the same patient with the same symptoms, they may treat the patient with slightly different drugs. A machine learning algorithm trained on the first physician may also 'learn' to treat the patient like the first physician. Conversely, if the algorithm was trained by the second physician it will treat accordingly. Thus, the algorithmic models suffer from annotator bias. It is important to distinguish this from sample bias, which occurs due to bias in picking samples for training. One of the motivations for integrating machine learning algorithms into healthcare workflows is objectivity in diagnoses and treatment decisions. However, because annotator bias creates objective algorithms, using supervised machine learning becomes a challenge. Since algorithmic outputs of supervised learning models are completely dependent on what the model has learned from the training data, human-annotated training data can create a biased model. Annotator bias is often expressed due to Automation bias, the human inclination to indiscriminately accept outputs of the algorithm, which can

AI Ethics Journal

lead to incorrect decision making (Bond et al., 2018; Goddard, Roudsari et al., 2012, Parasuraman, Molloy et al., 1993). Any errors made by the original data annotators of machine learning algorithms could be propagated if human oversight over algorithmic decisions is not maintained.

Temporal Bias

A third type of bias which is infrequently discussed in machine learning literature is temporal bias. Algorithms, especially modern deep learning algorithms, are effectively immortal learning machines. A human physician may train for ten years – practice for 40 years and pass on some of what they have learned to the next generation of physicians. Eventually, the human physician will die – and nuanced information about their practice will die with them. A machine learning algorithm, on the other hand, is effectively immortal. A machine algorithm with access to enormous processing power and memory capacity can continue to 'learn' over multiple generations of humans. Thus, a healthcare system belonging to a community which deploys a machine learning algorithm earlier will be able to:

- Train A.I. technologies for a longer period of time.
- Train A.I. technologies that are able to utilize multi-generational information to inform diagnosis and treatment intervention.
- Collect data for a longer period of time.

This in turn means that such societies will confer a 'healthcare' advantage to their constituents. Consider the case of two hypothetical societies: Atlantis - which has machine learning models trained to use geolocation, movement frequency, and facial image data to recognize individuals with symptoms associated with a viral respiratory infection - and Baltia, which lacks such technology or tracking. In this case, Atlantis is able able to effectively track individuals carrying the virus and ask such individuals to self-isolate, thus controlling the spread. Baltia, on the other hand, has to start building the necessary A.I. infrastructure while being saddled with an epidemic. Atlantis is likely to emerge from the epidemic in less time and with less damage than Baltia. This difference is not because of a difference in technological capability between the two societies, but rather because Baltia implemented the same A.I. much later than Atlantis. This is an example of temporal bias.

Temporal bias may also be considered in the context of an algorithm which continues to run for multiple generations of humans. Recall that an algorithm can continue to collect data and train on such data over multiple generations of humans, unlike a human physician who will eventually die along with their knowledge of treating a particular community. Since an algorithm is only limited by computational memory and processing power, a community, which starts collecting data and training algorithms a few generations prior to another, has a tremendous advantage. This type of temporal bias is illustrated in the example of A.I. algorithms using genetic information for making medical decisions. This advantage will most likely be applicable to future generations of a community, for whom data pertaining to their ancestry has already been digested by the algorithm. Going back to our hypothetical example, suppose that Atlantis collected genetic data and trained algorithms for multiple generations and Baltia did not. The trained algorithm will be able to make more accurate decisions for an individual from Atlantis, based on the particular genetic history of the individual. Since historical genetic data is simply not available for individuals in Baltia, they will be disadvantaged - even if at some future point Atlantis shares their algorithm and their data.

AI Ethics Journal

Effects of Bias and Data Sharing as a Potential Solution

Sample Bias

We will continue utilizing our hypothetical societies, Atlantis and Baltia. The individuals in Atlantis share medical data with A.I. developers who train algorithms that aid in or deliver healthcare to members of Atlantis. Now, if these developments improve healthcare delivery in Atlantis, agents from Baltia may attempt to import and implement the A.I. models used in Atlantis. This approach may not yield the expected advantages in Baltia due to sample bias stemming from differences between Atlantis' training data and Baltia's algorithmic input data. Furthermore, if the population characteristics of Baltia diverge significantly from Atlantis, such importation may even lead to adverse effects for the healthcare system in Baltia. In an alternative scenario, if medical data were shared and used by both societies - assuming no violation of privacy - the initial models trained may be able to ethically serve both societies since both would be represented in the training data. Depending on the algorithms used, sharing data could be mutually beneficial for Atlantis ***and*** Baltia.

Annotation Bias

A.I.-based medicine will most likely be deployed on a wide scale in several large healthcare systems in the future. Such A.I. may be initially trained using data generated by a small set of annotators. The accuracy of such annotations will ultimately drive the accuracy of the A.I. solution being implemented, and using a small set of annotators would inevitably lead to annotation bias in the trained models. A potential solution would be crowdsourcing annotation. Crowdsourcing, which evolved as a method of obtaining data quickly and cheaply, in theory can have the additional benefit of involving a larger and more diverse set of annotators. This diversity can contribute to the *wisdom of crowds* effect of aggregating imperfect judgments to create a collective intelligence (Surowiecki, 2004). Beyond utilizing *wisdom of crowds* as a passive byproduct of crowdsourcing, it has been proposed that computational methods be implemented in ways that intentionally incorporate the requirements for *wisdom of crowds* (diversity, independence, decentralization, and aggregation) in order to develop more representative, fair, and ethical training data generation (Lease, 2011; Mason, Vaughan et al., 2014). Ideally, crowdsourcing with data quality curation could blunt the effect of annotator bias in training clinically useful machine learning models. Note that crowdsourcing by definition requires widespread data sharing – thus making data sharing central to the future development of medical A.I. decision making systems.

Temporal bias

A.I. algorithms are sensitive to initialization, the amount of data on which training is performed, and the time used to train on that data. In general, the more time used for training, the better trained the algorithm - even if the same data was used for training (Goodfellow, Bengio et al., 2016). Furthermore, the longer that a community collects data, the larger the volume of data that is collected and annotated. Returning to the hypothetical Atlantis and Baltia, if Atlantis started training the algorithm and collecting data prior to Baltia, there would be an advantage to Atlantis. This type of bias may be partially resolved if Atlantis shares the numerical parameters of a trained algorithm and data it has collected over time with Baltia. However, the effects of this type of bias cannot be completely mitigated by data sharing. For example (as stated before), a machine learning algorithm which uses multi-generational genetic data to deliver personalized treatments in Atlantis may not be able to be as effective in Baltia – which has only just started collecting genetic data.

AI Ethics Journal

Current Hurdles to Data Sharing

HIPAA

The Health Insurance Portability and Accountability Act (**HIPAA**) of 1996 is an American legislation that pertains to the protection of sensitive patient information. HIPAA levies stiff penalties on clinical organizations for non-compliance. This has permeated a culture of fear with medical providers as well as organizations – who silo their data in response. These 'siloes' make it difficult for any machine learning algorithm, even within a given healthcare system, to access patient data. It also makes data sharing across institutions a relatively difficult and long drawn effort. In the long run, this will ultimately lead to a relative disadvantage compared to societies where such data sharing is encouraged. Even in its current form, HIPAA allows for sharing data under associate agreements, and for direct sharing of anonymized data. If organizations and communities are able to enact frameworks for responsible data sharing that are deemed HIPAA-compliant, cross-institutional data sharing can become a norm - and the ultimate result will be improved healthcare for all individuals. Another potential approach is to update the legislation itself to explicitly allow for sharing data for purposes related to A.I. implementation in healthcare.

Technical Challenges

Even if all the data were accessible, it would be difficult to seamlessly share all this data and train an algorithm on it. Medical data is heterogeneous, and each patient generates everything from physician notes to X-Rays to MRIs, often across multiple healthcare systems. This data is enormous in size, making training machine learning algorithms on the entirety of this data, a technical challenge. So far, the increase in the amount of data captured from patients has outpaced both improvements in hardware and in software. Thus, technical challenges to the seamless mobility of data will only increase in the near future. This may necessitate the evolution of machine learning frameworks that are portable, a possibility that is already being explored by the machine learning community in work on federated learning (Yang, Liu et al., 2019; McMahan, Moore et al., 2017).

Conclusion

We conclude that sample bias, annotation bias and temporal bias pose important ethical considerations that must be addressed before the world dives headlong into implementing A.I.-enabled medicine. A potential approach to addressing some of these challenges is to implement frameworks for sharing data across institutions nationally, and possibly even internationally. However, data sharing must be implemented responsibly by ensuring that privacy is not violated, and that technical challenges do not stymie the derivative rewards. In this section we detail what such frameworks could look like and how one could go about implementing data sharing without completely sacrificing privacy.

Data Anonymization and Sharing

Anonymization of data prior to sharing is the simplest and most commonly used technique for data sharing in the current environment. For certain types of data, anonymization is possible. For example, laboratory reports can be anonymized by removing the name, address and medical record numbers of the associated patient. The remaining data cannot be traced back to the patient and can be safely shared across institutions. This is possible only because such reports are highly structured at any given healthcare facility and patient identifiers are stored in very specific places in these data. Removing these identifiers anonymizes the data. Similar types of structured data are found in other relatively narrow domains of healthcare such as radiology and pathology. In radiology, the DICOM (Larobina &

AI Ethics Journal

Murino, 2014) standard stores private health information (PHI) as a part of the image meta-data which can be easily removed for the radiological scans. Anonymization procedures could be established for these on a per institution basis – and such procedures could form the basis of anonymization software that enable data sharing in a seamless fashion across institutions, cities and nations without privacy compromises.

While anonymizing and sharing in these narrow domains is relatively easy, it remains extremely difficult to automate the anonymization of the physician note, which is the most important and richest source of healthcare data. Machine learning has been applied to attempt anonymization of physician notes with less than impressive results. Unfortunately, sharing the physician note across institutions through anonymization will require human readers to go through each note and anonymize it. Alternately, physicians will need to write notes in a structured manner to enable easy anonymization. Both of these possibilities are difficult to implement in existing healthcare institutions where physicians are over-stretched, and budgets operate on razor thin margins. Nevertheless, creating a structured template for medical notes where PHI is restricted to one subsection would rapidly and easily facilitate research and data sharing, without placing the burden on the physician. In short, data anonymization can facilitate data sharing in narrow subdomains of healthcare where private health information is stored in a structured fashion. In other subdomains, it is necessary to define national and international standards for anonymization of data and create tools to adhere to these standards. Compliance to well-defined standards using tools created specifically for such compliance can make the sharing of data easier and ensure that such sharing does not violate privacy of individual patients.

Cloud computing and AI.

Is privacy breached if personal data is never read by another human, but only read by an algorithm? This question is relevant to the data sharing debate in healthcare and beyond. Healthcare institutions in particular have slowly started to move from local data archival to cloud-based data archival using established cloud computing vendors such as Amazon Web Services (AWS), Microsoft Azure and Google Cloud. As increasing numbers of healthcare providers make the switch, these vendors are likely to build up enormous repositories of private health information. Most of these vendors have substantial expertise in AI and machine learning as well and could design algorithms which train on data from multiple institutions. However, doing so without breaching individual level privacy will likely remain a challenge. Privacy is generally understood to be the condition of being free from being observed or disturbed by other humans. Does this extend to algorithms? This question should ultimately be addressed by philosophers and ethicists since it concerns the definition of privacy itself.

Emerging Technological solutions.

Federated learning- (Yang, Liu et al, 2019), differential privacy- (Abadi et al., 2016) and homomorphic encryption-based machine learning (Aono, Hayashi et al., 2017) are three emerging technical paradigms that provide a solution to cross-institutional data sharing. Each of these paradigms aims to share data without adversely affecting user privacy. Federated learning is a machine learning approach where machine learning models are trained on local batches, but gradient updates are shared centrally in a common model. Differential privacy espouses a class of approaches which ensures that the model is robust to a change in one sample of the training data. Lastly, homomorphic encryption-based machine learning allows for learning directly from encrypted data. Each of these paradigms is an active area

AI Ethics Journal

of work in the machine learning community: each of these approaches also comes at a cost. There is added complexity in training algorithms, and also the possibility that a simpler algorithm trained with data sharing outperforms algorithms designed within the aforementioned paradigms. In short, we conclude that establishing a clear legal framework to enable data sharing in an anonymous fashion, rethinking privacy in an algorithmic context, and developing machine learning algorithms that respect privacy while training by sharing digested data instead of raw data are all potential solutions to implement data sharing for machine learning in the modern healthcare setting.

References

[1] Ashraf, A. *et al.* (2015). "Breast DCE-MRI kinetic heterogeneity tumor markers: Preliminary associations with neoadjuvant chemotherapy response," *Translational Oncology, 8*(3), 154-162

[2] Gaonkar, B. and Davatzikos, C. (2013, Sep.). "Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification," *Neuroimage*, *78*, 270–283.

[3] Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94.

[4] Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference.*

[5] Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair.

[6] Zadrozny, B. (2004, July). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning* (p. 114).

[7] Kallus, N., & Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. *35th International Conference on Machine Learning, ICML 2018.*

[8] Gostin, L. O., & Nass, S. (2009). Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. *JAMA - Journal of the American Medical Association*, *301*(13), 1373-1375.

[9] Rigby, M. J. (2019). Ethical dimensions of using artificial intelligence in health care. *AMA Journal of Ethics*, *21*(2), 121-124.

[10] Gaonkar, B., Shinohara, R. T., Davatzikos, C., & Alzheimers Disease Neuroimaging Initiative. (2015). Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Medical Image Analysis*, *24*(1), 190-204.

[11] Gaonkar, B., & Davatzikos, C. (2013). Analytic estimation of statistical significance maps for support vector machine based multivariate image analysis and classification. *Neuroimage*, *78*, 270-283.

[12] Bakas, S. *et al.* (2015, October). GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In *BrainLes 2015* (pp. 144-155). Springer, Cham.

[13] Gaonkar, B. *et al.*, Quantitative Analysis of Spinal Canal Areas in the Lumbar Spine: An Imaging Informatics and Machine Learning Study. *American Journal of Neuroradiology*, *40*(9), 1586-1591.

[14] Gaonkar, B., Beckett, J., Villaroman, M. H. S. D., Ahn, B. S. C., and Edwards, B. S. M. (2019). Quantitative analysis of neural foramina in the lumbar spine: an imaging informatics and machine learning study. *Radiology: Artificial Intelligence*, *1*(2), 180037.

[15] Attiah, M. *et al.*, (2019). Natural history of the aging spine: a cross-sectional analysis of spinopelvic parameters in the asymptomatic population. *Journal of Neurosurgery: Spine*, *32*(1), 63-68.

[16] Macyszyn, L. *et al.*, (2015). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology*, *18*(3), 417-425.

AI Ethics Journal

[17] Sotiras, A. *et al.*, (2016). Machine learning as a means toward precision diagnostics and prognostics. In *Machine learning and medical imaging* (pp. 299-334). Academic Press.

[18] Guyon, I. *Feature Extraction Foundations and Applications*. 2006.

[19] LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436-444.

[20] Soh, M. "Learning CNN-LSTM Architectures for Image Caption Generation," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2016.

[21] Boag, W., Wacome, K., Naumann, T., and Rumshisky, A. (2015). CliNER: a lightweight tool for clinical named entity recognition. *AMIA Joint Summits on Clinical Research Informatics (poster)*.

[22] Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, *178*(11), 1544-1547.

[23] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA - Journal of the American Medical Association*, *318*(6), 517-518.

[24] Wong, J., Horwitz, M. M., Zhou, L., & Toh, S. (2018). Using machine learning to identify health outcomes from electronic health record data. *Current Epidemiology Reports*, *5*(4), 331-342.

[25] Lanes, S., Brown, J. S., Haynes, K., Pollack, M. F., & Walker, A. M. (2015). Identifying health outcomes in healthcare databases. *Pharmacoepidemiology and Drug Safety*, *24*(10), 1009-1016.

[26] Wei, W. Q., Teixeira, P. L., Mo, H., Cronin, R. M., Warner, J. L., & Denny, J. C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, *23*(1), 20-27.

[27] Shivade, C. *et al.* (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, *21*(2), 221-230.

[28] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA - Journal of the American Medical Association*, *309*(13), 1351-1352.

[29] Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1007-1015.

[30] Araujo, T. *et al.*, (2017). Classification of breast cancer histology images using convolutional neural networks. *PloS one*, *12*(6), e0177544.

[31] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.

[32] Arpey, N. C., Gaglioti, A. H., & Rosenbaum, M. E. (2017). How socioeconomic status affects patient perceptions of health care: a qualitative study. *Journal of Primary Care & Community Health*, *8*(3), 169-175.

[33] Bond, R. R. *et al.* (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology*, *51*(6), S6-S11.

[34] Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121-127.

[35] Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1-23.

[36] Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business.* Doubleday.

[37] Lease, M. (2011). On quality control and machine learning in crowdsourcing. *Human Computation*, *11*(11).

[38] Mason, W., Vaughan, J. W., & Wallach, H. (2014). Computational social science and social computing. *Machine Learning*, *95*, 257–260.

[39] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* Cambridge: MIT Press.

AI Ethics Journal

[40] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-19.

[41] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017, April). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics in PMLR, 54,* 1273-1282.

[42] Larobina, M., & Murino, L. (2014). Medical image file formats. *Journal of Digital Imaging*, *27*(2), 200-206.

[43] Abadi, M. *et al.* (2016, October). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308-318).

[44] Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2017). Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, *13*(5), 1333-1345.