Ethical Concerns of Using Machine Learning to Diagnose Alzheimer's Disease

Maya Gonzalez
mgonzalez3@oxy.edu
Occidental College

Abstract

There are various ethical concerns to consider when incorporating a Machine Learning (ML) model into the decision making process of a real clinical scenario. Using ML to diagnose Alzheimer's disease (AD) from image data is one problem space that has the potential to prevent unnecessary invasive procedures or to diagnose a patient before symptoms appear in hopes to slow down the progression. The concerns that are arguably most important are the ones concerning the safety of the patient. There are many factors and decisions being made in the overall process even before a patient is introduced into the equation. The general process includes data collection, research and model building, and then implementation into a clinical setting.

1 Ethical Considerations

1.1 Data Set Collection

To start off at the beginning of the process, the data sets available are too small for training purposes. Current research mainly trains on one of several popular data sets, which contain only several hundreds of participants. It is generally accepted that a model is more generalizable and has better performance if it is trained on more data. For example, ImageNet is an image data set used for training and evaluation, and contains over 14 million annotated images. Although ImageNet is used to advance general image recognition whereas the problem space of AD classification is more specialized, the discrepancy in data size is notable. It is also important to consider the specific data each model is trained on. If there is a model that is determined to be the 'best' at the AD classification problem, it is likely highly dependent on the data trained on. Would the results still be the same if the model was tested on a different data set? Is there a chance that all participants in one data set have something in common due to a sampling bias? The model may perform well on a certain population group if that was the only data provided for training.

It is acknowledged that most medical datasets are trained on a subset or specific group of the general population. The lack of racial and ethnic diversity in the participants that make up a dataset is a problem that extends into the space of brain scan databases. For example, the ADNI dataset is one of the largest AD datasets available and widely used in research. However, the participant data is biased, as demonstrated by the participant makeup. According to the reported ADNI participant demographics [4], of the 822 total subjects sampled from, 764 identify as White. Only 39 of the total participants identified as black or African American and 14 identified as Asian. Of the 822 participants, 21 identified as Hispanic or Latino. Males accounted for 478 of the participants while females made up 344 of the total participants. The lack of participant sampling which accurately reflects the general population reduces the ability to generalize the models trained on these data sets.

The work of [2] Li et al. analyzes the potential causes and consequences of using neuroimaging data sets that lack racial and ethnic diversity. The researchers used two datasets, the Human Connectome Project and the Adolescent Brain Cognitive Development, that deliberately included individuals of different races and ethnicities, yet still included a majority of white participants. Predictive models were trained on one group that was composed of majority white American (WA) patients and another composed solely of African American (AA) patients. The model trained only on AAs improved in prediction accuracy for AA patients, yet was still lower than the accuracy level of the model trained on a dataset mainly composed of WA patients. The authors argue that researchers should work to build data sets that are more representative of the general population, even if it requires extra work to study from diverse cohorts. However, definite conclusions can not be drawn from their work therefore the researchers suggest that further research is needed in this area.

Although the discrepancy in gender is not as drastic as the racial discrepancy, it is still important to consider if this discrepancy can lead to any biased results. There are established anatomical differences in the brains of male and females. Liu et al. [3] used two structural-neuroimaging datasets to analyze over 2000 brain scans. The reserachers identified differences in gray matter volume in the cortex for men and women. These differences may contribute to differences in cognition and information processing, but that is out of scope for this paper. These structural differences

appear on MRI scans, which is why considering participant composition based on gender is important when analyzing the dataset used for training.

One can imagine the consequences if s a data set only contained data from one racial group or one gender. It is important to consider if there are any other factors that would introduce a bias into the data that is being used to train on. It is necessary to rule out all possibilities of introducing a bias into the model before any ML model can be used in the decision making process of a real case. If not, these biases will be encoded into the final decision and can have life threatening consequences.

Gebru et. al [1] proposed a standard method of documenting datasets used for Machine Learning. The goal of utilizing this documentation method is to increase transparency of the researchers and within the community. Partaking in this documentation process in the creation of neuroimaging datasets may help to mitigate some of the concerns raised about data bias. More work is need to be done to ensure that the datasets used for training are unbiased and well-researched.

1.2 A Lack of Collaboration

There are various researchers attempting to tackle the problem of using ML to diagnose AD from imagery data. In the literature, there are issues of reproducibility of other researcher's code and models. It is difficult to try and reproduce a published model since all aspects of the model may not be mentioned in detail. It is necessary to evaluate the model and be able to reach the same results, which is difficult and even unlikely if there is a lack of collaboration between researchers. There is a lack of collaboration between everyone involved in the process: data collectors, ML researchers, and medical practitioners. There should be regulatory bodies and guidelines in place for the entire project from start to finish. An understanding of each step in the process would help ensure that all the project stages are working cohesively and that there are no blindspots. A regulatory body can be responsible for making decisions to some of the lingering questions such as how the algorithm is decided upon. Is this a choice that the medical team can decide upon or is there a universal model that is decided on and used everywhere?

1.3 Integration into a Clinical Setting

Regardless of the model's performance, there is a broader question of how the model would be implemented in a medical setting. More specifically, at what stage would the model's decision be incorporated? How much influence should the decision of a model have over that of the medical professionals? Most medical professionals and doctors involved in the decision making process may not have experience or familiarity with ML. They may see this tool as a black box, in that they have no idea how the decision is being reached and therefore may not trust the ML model. Yang et. al [5] investigated how Decision Support Tools (DSTs) are integrated into clinical settings by specifically looking at the decision of implanting a artificial heart in patients. Based on design studies in multiple hospitals, there was no voiced resistance to the inclusion of DSTs within a team's workflow. However, the authors related general issues of integrating DSTs into multidisciplinary decision meetings as a way of complementing clinical decisions. They argued that there is a need to design DSTs as both functional and integrated into clinician's workflows so that the decision making can be supported by DSTs rather than being hindered by it. In the broader context of using ML in clinical settings, regardless if clinicians don't trust the tool at all or if they place too much weight in the model's decision, the medical practitioners may be unnecessarily swayed by the model's decision and therefore incorporate bias in the ultimate decision.

The mathematical underpinnings of ML and the deployed model are difficult to understand for both the medical practitioners and the patients. The patients involved should be aware that a ML model is being used to assist in the decision making process. What if the model made an incorrect decision which led to an unnecessary operation and risked the patient's life? Would the doctors lose faith in the tool or would the patient have grounds to press legal charges? Who is held responsible if an incorrect decision from the ML model results in an unnecessary life lost? These are all questions that must be considered and answered before a ML model can be introduced into a hospital with real patients.

Once deployed into a clinical setting, another issue emerges of who has access to this technology. MRI scans are an expensive procedure in the US healthcare system. There may also be additional costs tied to using the predictive model. Only those individuals with the available money would be able to take advantage of this resource. This lack of accessibility further deepens the barrier of access to medical innovations. If a tool is only able to help a subset of the general population who have the available resources, then a question arises of how useful this will actually be.

2 Conclusion

Due to the lack of regulatory oversight in place, it is not yet ethically feasible to use a ML model to assist in the decision making process of Alzheimer's disease. More work is needed to ensure the entire process is sound and can be used safely in a medical setting with real patients. The outlined questions and more must be carefully considered by a di-

verse group of qualified individuals. It is not worth risking a patient's life to make wishful strides in Machine Learning.

References

[1] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[2] Jingwei Li et al. "Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity". In: *Science advances* 8.11 (2022), eabj1812.

[3] Siyuan Liu et al. "Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans". In: *Proceedings of the National Academy of Sciences* 117.31 (2020), pp. 18788–18798.

[4] *Menu of demographics tables - adni*. URL: https://adni.loni.usc.edu/wp-content/uploads/2012/08/ADNI_Enroll_Demographics.pdf.

[5] Qian Yang, Aaron Steinfeld, and John Zimmerman. "Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–11.