Deep Learning for the Classification of ADHD in MRI Data

By Collin Shields
Advised by Dr. Benjamin Nephew

1 INTRODUCTION & BACKGROUND

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder that affects 5.9% of youths (Willcutt, 2012) and 2.5% of adults worldwide (Simon et al., 2018). This disorder is categorized into three subtypes: predominantly inattentive presentation, predominantly hyperactive-impulsive presentation, and combined presentation. The first subtype is associated with a lack of attentiveness, meaning the affected individual will have trouble staying focused and keeping organized. This can cause the individual to suffer academically and struggle to stay employed. The second subtype is associated with a generally impulsiveness, an inability to stay still, and limited patients. This can lead to the individual putting themselves in dangerous situations, as well as negatively impacting their personal relationships. The final subtype is a combination of the first two (CDC, 2023). Those with ADHD generally report a moderate to severe negative impact to their quality of life (Lee et al., 2016). This is in line with many other studies including a study in Denmark that found that those with ADHD were at a much higher risk for suicide (Fitzgerald et al., 2019). Those with ADHD are also three times as likely to develop a nicotine addiction, and 50% more likely to develop a drug or alcohol addiction (Lee et al., 2011). ADHD has also been found to result in lower academic achievement, even when medicated (Fleming et al., 2017). Given how common the disorder is, and how it can impact the lives of affected individuals, proper diagnosis of the disorder is very important.

The diagnosis of ADHD has experienced ongoing development for a very long time, with some of the earliest accounts of the disorder going back to 1775 (Faraone et al., 2021). Since

then, the methods for diagnosing ADHD have become more objective, with modern diagnosis using a standardized self-report or parent-report questionnaire to determine whether the individual has ADHD, as well as the subtype, if applicable (CDC, 2023). This process is described in more detail in the American Psychiatric Association's Diagnostic and Statistical Manual, Fifth edition (DSM-5). The questionnaire and subsequent scoring are called the ADHD Rating Scale or ADHD-RS. This method of diagnosis has been criticized as being subjective due to a lack of biological basis, however it meets the standard criteria for validity of a mental disorder as described by Robins and Guze (Faraone, 2005). Although, this does not mean that misdiagnosis does not occur.

Treatments for ADHD often include prescriptions to drugs that, while very effective for those with ADHD, have a high risk of addiction for those without ADHD. One meta-analysis of 111 studies concluded that nonmedical usage of stimulants is a significant public health problem (Faraone et al., 2020). Misdiagnosis of those with ADHD could exacerbate this issue and expose people to unnecessary risk. There are multiple potential causes of this misdiagnosis. One is that the ADHD tests rely on self-reporting, and as a result it's possible that an individual could lie to get access to the stimulants. Another problem is that a difference in age among tested children can sometimes be misinterpreted as the presence of ADHD. One study found that children that were born in December were much more likely than those born in January to be diagnosed with ADHD. Because ADHD is a neurological condition, rates in diagnosis should not be affected by grade cutoffs. This would imply that diagnosis fails to consider potential developmental immaturity in children being tested. With the existence of misdiagnosis, and potential for substance abuse, there is great merit in finding a concrete method for diagnosis that doesn't require self-reporting or subjective analysis.

One potential avenue for diagnosis is through analysis of magnetic resonance imaging (MRI) scans of the brain. Both structural MRI (sMRI) and functional MRI (fMRI) data has been used before in exploring potential biological markers of the disorder. A basis of diagnosis through this avenue could provide a definitive way to diagnose ADHD and its subtypes and rule out misdiagnosis due to lack of testing or being confused with another disorder. It could also potentially open a path to understanding the biology of ADHD. Unfortunately, the differences found in the scans through manual comparison between those with ADHD and those without

have been small and not clinically usable as indicators of the disorder. Standard analysis is restricted to pre-defined structures and excludes a significant amount of fine-grain data in favor of more identifying chunks of data. Rather than doing standard analysis, the most recent advancements in this field use deep learning to classify the existence of ADHD and its subtypes using MRI data. Deep learning has the advantage of being able to make use of the entirety of the data in an MRI scan.

A notable example of using deep learning for diagnosis is the ADHD-200 consortium competition that was held in 2011. The purpose of the competition was to develop a digital tool that could identify biomarkers of ADHD using functional MRI (fMRI) data. The fMRI scans were accompanied with an identification number, some demographic data, as well as a classification. This classification was one of four groups: Typically Developing Children (TDC), ADHD-Hyperactive/Impulsive, ADHD-Inattentive, or ADHD-Combined. Correctly identifying TDC was referred to as specificity and was rewarded with one point. Correctly identifying the subtype of ADHD was referred to as sensitivity and was also rewarded with one point. Identifying ADHD but incorrectly identifying the subtype was rewarded with half a point. Although it was not explicitly required, machine learning models dominated the competition, and the competition culminated in the John Hopkins University team winning (ADHD-200 Consortium, 2012). Their model had a specificity of 94% but a sensitivity 24%, meaning it could correctly classify 94% of cases as TDC or ADHD, but could only correctly identify 24% of the ADHD subtypes. This model is an insight into how machine learning could be used to identify ADHD diagnosis, however it also makes clear a pitfall of the competition's scoring system. A significant emphasis was placed on specificity over sensitivity. This issue is also mentioned by the John Hopkins team themselves (Eloyan et al., 2012).

Another team was able to use exclusively demographic data to identify diagnosis and produced a higher accuracy than all other teams. Although they were disqualified due to this strategy being out-of-line for the general spirit of the competition, this indicated that the data provided by the consortium may have been biased. Other models may unintentionally have leveraged this bias despite the goal of identifying biomarkers for ADHD. Gender was a significant indicator of diagnosis, as the data set had disproportionately more male ADHD cases than female ADHD cases. This discrepancy in diagnosis between genders coincides with

findings from other studies as well, as is described by meta-analysis from Willcutt (2012) and Simon et al. (2018). IQ also had a role in this, as the IQ of the those diagnosed with ADHD was 7-10 points lower on average to those who were not (ADHD-200 Consortium, 2012), which is consistent with findings from Frazier et al. (2004). However, these indicators are just one part of a large and complex system and cannot be reliably used for diagnosis.

Our goal is to use deep learning to identify ADHD and its subtypes. We propose that instead of using one network to do all the above, it may be more efficient to use two separate networks that perform different functions. Specifically, one that can identify between typically developing individuals and individuals with ADHD, and one that can identify subtypes of ADHD. By separating out the classification groups like this, we hope to achieve higher accuracy overall. We will not be using demographic data in the hopes that the model produced will only use the biological data in its classifications. This should aid in preventing some biases from affecting our results.

The deep learning model we will be using is the convolutional neural network used in the more recent Alzheimer's disease study (Liu et al., 2022). This model was used to identify Alzheimer's in the hopes that it would be possible to predict the development of Alzheimer's early on. Our plan is to use the pretrained network for our own training, with the hopes that some of the patterns in the data might be transferable. There are major differences between the data used for the Alzheimer's study and our own data. These differences include a large gap in the age ranges between the two studies, differences between using fMRI and sMRI data, and an overall difference in classification goals. These differences might affect our training and effectiveness of our model, but it could also provide new insight into how well pre-trained CNNs can adapt to a change in environment.

2 METHODS

2.1 DATA

We will be using the publicly available data set from ADHD-200 consortium for this study. The ADHD-200 consortium data is organized into a set of MRI anatomical and resting-state functional scans, as well as top-level tsv files that contains information about each of the

subjects, including ADHD presence and type, as well as other collected information such as gender, handedness, and IQ. For the purposes of this study, we decided to use anatomical scans, rather than the functional scans. The reasoning for this is described in more detail in section 2.3.

This data is split between many different institutions, and so for the purposes of this project the scans will be pooled together in one folder and the tsv files will be combined into one file, ALL.tsv, using an automated python script, fileMover.py. ALL.tsv, was then split into two parts using TSVSplitter.py: ALL_val.tsv and ALL_train.tsv. These files contained a random selection of 10% and 90% of ALL.tsv respectively. This was done to isolate a set of training data and validation data. We also made another file that included only positive ADHD diagnosis with the goal of training two separate networks with specialized jobs. One would use binary classification between controls and ADHD subtypes, and the other would use trinary classification of the ADHD subtypes themselves.

The tsv files were also cleaned using TSVFixer.py, which would remove entries from the files whose anatomical scans didn't meet the minimum size requirements of 96x96x96. Finally, there was also a problem with duplicate entries, which were removed from the tsv file manually. This entry duplication is assumed to be an issue with fileMover.py, and if there are no duplicate entries in the tsv files, the CNN will ignore any duplicate scans.

2.2 Deep Learning Model

A 3D CNN, composed of convolutional layers, instance normalization (Ulyanov et al., 2017), ReLUs and max-pooling layers, was designed by Liu's team to perform classification of Alzheimer's disease and mild cognitive impairment and normal cognition cases. In a preliminary work they showed that the proposed architecture is superior to state-of-the-art CNNs for image classification (Beekly et al. 2004). The proposed architecture contains several design choices that are different from the standard convolutional neural networks for classification of natural images: (1) instance normalization, an alternative to batch normalization (Ioffe and Szegedy, 2015), which is suitable for small batch sizes and is empirically observed to achieve better performance; (2) small kernel and stride in the initial layer for preventing losing information in small regions; (3) wider network architecture with more filters and less layers for the diversity of

the features and ease of training. These techniques all independently contribute to boosting performance.

As is standard in deep learning for image classification (Goodfellow et al., 2016), they performed data augmentation via Gaussian blurring with mean zero and standard deviation randomly chosen between 0 and 1.5, and via random cropping (using patches of size $96 \times 96 \times 96$).

The model was trained using stochastic gradient descent with momentum 0.9 (as implemented in the torch.optim package) to minimize a cross-entropy loss function. They used a batch size of 4 due to computational limitations. They used a learning rate of 0.01 with a total of 60 epochs of training which were chosen by grid search based on validation set performance. During training, the model with the lowest validation loss was selected.

2.3 Study Structure

This study was conducted as a child to an overhead study run by Dr. Benjamin Nephew at WPI. The main goal of the study was to replicate the findings in the Liu et al. (2022) study, and multiple subgroups were created based on student interests to apply the modified deep learning pipelines to different topics chosen by those groups. Each group was tasked with collecting data and applying the modified pipeline to it. This is what is being referred to when other groups are mentioned.

2.4 Model Usage

We used the pretrained model described above with the hopes of transferring its learning over to our model. We did this in the hopes that the model would retain what it had learned from training on anatomical scans and repurpose its existing knowledge for new classifications. Because the model was trained on anatomical data, we decided to use anatomical scans for our own training to maximize the amount of transferrable data patterns. Our plan was to train each of our proposed neural networks for about 100 epochs with a learning rate of 0.01 and a batch size of 16. We would then observe the results and tweak values to maximize the model's effectiveness.

We faced some immediate problems while configuring the model to do binary classification of ADHD vs control. The first was that all the tsv files provided different categories with different information, which meant that only some data was reliably present in all files. These categories were participant_id, gender, age, and participant_category, where the first and last variables are the ones necessary for running the pipeline. Another problem we encountered was with the naming schemes of the files. The pipeline required all scan files to have the same naming structure. There were only a handful of variations, so the pipeline was modified to just search for all potential file names for any specific scan file.

As was mentioned previously, we also encountered an issue where some of the scan files were not large enough for the pipeline, which resulted in the scan being skipped and eventually the program terminating due to different resulting batch sizes. The first change made to try to fix this problem was to have the program that moved files and merged the tsv files exclude scans that weren't marked as "pass" in quality control, although this didn't seem to fix the issue. Due to inconsistencies in the labels for the quality control columns, as well as complications in how they were organized, we finally decided to manually remove any scans that were not large enough. This happened to only include one scan.

The final problem that we encountered was that the validation accuracy of the network would not improve at all throughout training, no matter how many epochs it trained for. The validation accuracy would always come out to about 50% every epoch, and batch accuracy would remain about the same throughout. Another group with the same problem identified this behavior as the model always predicting all positive or all negative while training. A variety of approaches were attempted to fix this issue, such as modifying the learning rate to be higher and lower than 0.001, changing the batch size, as well as some general fixes to different parts of the code. In the end, none of these changes had an effect, which was also reported by the other groups working on their own projects.

2.5 ETHICS

All the data used in this study is publicly available and de-identified.

3 RESULTS

Due to time constraints, we were unable to get the deep learning pipeline to produce results. With more time, we could have tried more approaches to fixing the program, applied the fMRI scans instead of the anatomical scans, or even tried using the random forest machine learning model instead of the CNN. As for why the networks seemed to be unable to learn, there are a couple of potential reasons behind this.

The first explanation behind the lack of learning is that anatomical data cannot be used to identify classify ADHD vs controls in deep learning. The John Hopkins team in the ADHD-200 consortium, for example, only used fMRI data. It's possible that brain structure is not useful enough for identifying ADHD. Extending from this, transferring the learning from the Alzheimer's study could also have caused pipeline to fail. The data may have been too different, resulting in the training process struggling. This could explain why the network would classify all scans as ADHD or all scans as TPC while training, though it wouldn't explain the lack of improvement.

The next explanation is that the data used in this study was not adequate, or that it wasn't filtered or handled correctly. This is our first experience working with deep learning, and as a result it's possible the data was not separated or cleaned properly resulting in noise that inhibiting the learning process.

The final explanation is that modifications made to the original CNN program in creation of the binary and multiclass pipelines resulted in erroneous behavior. This is reinforced by the fact that multiple other groups seemed to have the same issue as us. Their models would only produce 50% validation accuracy and didn't seem to learn at all. This could also explain the behavior of classifying scans as either all positive or all negative.

Given the fact that other groups experienced the same issue, we conclude that the most likely explanation is that the pretrained model could not translate what it had learned to other functions such as ADHD classification. This issue was potentially compounded by inefficiencies in using anatomical MRI scans over fMRI scans for identifying ADHD.

4 REFERENCES

Faraone, S. V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M. A., Newcorn, J. H., Gignac, M., Al Saud, N. M., Manor, I., Rohde, L. A., Yang, L., Cortese, S., Almagor, D., Stein, M. A., Albatti, T. H., Aljoudi, H. F., Alqahtani, M. M. J., Asherson, P., … Wang, Y. (2021). The World Federation of ADHD International Consensus Statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, *128*, 789–818. https://doi.org/10.1016/j.neubiorev.2021.01.022

CDC. (2023, September 27). *Symptoms and diagnosis of ADHD*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/adhd/diagnosis.html

ADHD-200 Consortium. (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in Clinical Neuroscience. *Frontiers in Systems Neuroscience*, *6*. https://doi.org/10.3389/fnsys.2012.00062

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A. D., Joel, S., Pekar, J. J., Mostofsky, S. H., & Caffo, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, *6*. https://doi.org/10.3389/fnsys.2012.00061

Liu, S., Masurkar, A. V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., Fernandez-Granda, C., & Razavian, N. (2022). Generalizable deep learning model for early alzheimer's disease detection from structural mris. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-20674-x

Willcutt, E. G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: A Meta-Analytic Review. *Neurotherapeutics*, *9*(3), 490–499. https://doi.org/10.1007/s13311-012-0135-8

Simon, V., Czobor, P., Bálint, S., Mészáros, Á., & Bitter, I. (2018). Prevalence and correlates of adult attention-deficit hyperactivity disorder: Meta-analysis. *British Journal of Psychiatry*, *194*(3), 204–211. https://doi.org/10.1192/bjp.bp.107.048827

Faraone, S. V. (2005). The Scientific Foundation for understanding attention-deficit/hyperactivity disorder as a valid psychiatric disorder. *European Child & Adolescent Psychiatry*, *14*(1), 1–10. https://doi.org/10.1007/s00787-005-0429-z

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). *Instance Normalization: The Missing Ingredient for Fast Stylization*. https://doi.org/10.48550/arXiv.1607.08022

Beekly, D. L., Ramos, E. M., van Belle, G., Deitrich, W., Clark, A. D., Jacka, M. E., & Kukull, W. A. (2004). The national Alzheimer's coordinating center (NACC) database: an Alzheimer disease database. *Alzheimer Disease & Associated Disorders*, *18*(4), 270-277.

Ioffe, S., Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research 37: 448-456. https://proceedings.mlr.press/v37/ioffe15.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. https://scholar.google.com/scholar_lookup?&title=Deep%20learning&publication_year=2016&author=Goodfellow%2CI&author=Bengio%2CY&author=Courville%2CA&author=Bengio%2CY

Ford-Jones, P. C. (2015). Misdiagnosis of attention deficit hyperactivity disorder: 'normal behaviour' and relative maturity. *Paediatrics & Child Health*, *20*(4), 200–202. https://doi.org/10.1093/pch/20.4.200

Lee, Y., Yang, H.-J., Chen, V. C., Lee, W.-T., Teng, M.-J., Lin, C.-H., & Gossop, M. (2016). Meta-analysis of quality of life in children and adolescents with ADHD: By both parent proxy-report and Child self-report using pedsqlTM. *Research in Developmental Disabilities*, *51–52*, 160–172. https://doi.org/10.1016/j.ridd.2015.11.009

Lee, S. S., Humphreys, K. L., Flory, K., Liu, R., & Glass, K. (2011). Prospective association of childhood attention-deficit/hyperactivity disorder (ADHD) and substance use and abuse/dependence: A meta-analytic review. *Clinical Psychology Review*, *31*(3), 328–341. https://doi.org/10.1016/j.cpr.2011.01.006

Fitzgerald, C., Dalsgaard, S., Nordentoft, M., & Erlangsen, A. (2019). Suicidal behaviour among persons with attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, *215*(4), 615–620. https://doi.org/10.1192/bjp.2019.128

Fleming, M., Fitton, C. A., Steiner, M. F., McLay, J. S., Clark, D., King, A., Mackay, D. F., & Pell, J. P. (2017). Educational and health outcomes of children treated for attention-deficit/hyperactivity disorder. *JAMA Pediatrics*, *171*(7). https://doi.org/10.1001/jamapediatrics.2017.0691

Faraone, S. V., Rostain, A. L., Montano, C. B., Mason, O., Antshel, K. M., & Newcorn, J. H. (2020). Systematic review: Nonmedical use of prescription stimulants: Risk factors, outcomes, and risk reduction strategies. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59*(1), 100–112. https://doi.org/10.1016/j.jaac.2019.06.012

Frazier, T. W., Demaree, H. A., & Youngstrom, E. A. (2004). Meta-analysis of intellectual and neuropsychological test performance in attention-deficit/hyperactivity disorder. *Neuropsychology*, *18*(3), 543–555. https://doi.org/10.1037/0894-4105.18.3.543