Deep Learning for Mental Illness Detection Using Brain SPECT Imaging

Felisa J. Vázquez-Abad[1], Silvano Bernabel[1], Daniel Dufresne[2*], Rishi Sood[3], Thomas Ward[3†], and Daniel Amen[4†]

[1] Department of Computer Science
Hunter College CUNY, New York, NY, USA
[2] School of Mathematics and Statistics
University of Melbourne, Australia
[3] Amen Clinics, New York, NY, USA
[4] Amen Clinics, Costa Mesa, CA, USA

Abstract. We apply deep learning to the detection of mental illness with meaningful results, using SPECT (Single Photon Emission Computed Tomography) images of the brain. The data consists in scans from patients with attention deficit hyperactivity disorder (ADHD), major depressive disorder (MDD) and obsessive compulsive disorder (OCD), plus scans of healthy brains. We focus here on the application of a deep convolutional neural network (CNN). The main challenge in using CNN models for medical diagnosis is often the number of samples not being sufficiently large to ensure high accuracy. We propose a soft classifier for using the machine. Instead of a binary output "yes/no" for each condition, we add an intermediate outcome, which says that the machine yields a weak result. The "Red Zone" corresponds to a positive result (condition is present) and the "Green Zone" corresponds to a negative, each with a preassigned statistical confidence level; the "Amber Zone" is an ambiguous outcome, where the scan is assigned a likelihood of having the condition. This information is then passed to the doctors for further analysis of patients.

Keywords: Deep learning, SPECT scans, artificial intelligence, mental health

1 Introduction

One tool used in psychiatry today is SPECT (Single Photon Emission Computed Tomography) brain images. Our research consists in using SPECT images for visual recognition; we apply artificial intelligence to determine if there is indeed important information embedded in these images that is critical for achieving accurate diagnosis. Our algorithms use **only** the SPECT images to arrive at a

[*] Corresponding author: daniel@ozdaniel.com

[†] Daniel Amen and Thomas Ward are co-last authors.

patient's diagnosis, while conventionally psychiatrists gather a clinical history in order to achieve a diagnosis.

SPECT imaging has been used to evaluate a variety of conditions since the 1970s. In this paper, we find that machine learning analysis of SPECT images on its own yields accuracies between 82% and 99% when diagnosing ADHD.

We believe our algorithms would provide a first indication to primary care doctors that would enable them to refer patients to psychiatrists earlier in the development of illnesses such as ADHD. We aim at building algorithms that will be easy to compute and can be attached to the SPECT scanners, providing the results as well as the images immediately. With this in mind, it is our goal to create algorithms that will help early detection, yet accept input of small dimensions and compute relatively fast.

In deep learning, it is usually recommended to use at least $50,000$ samples to achieve good accuracy. For example, in [2] brain scans are used to train a deep network that classifies tumor cells. Those authors use 2.5 and 3.4 million samples (each corresponds to a pixel in a brain image). In our project, each whole brain SPECT image is just one sample, because the goal is to detect mental illness of the whole brain, rather than classifying smaller regions as tumors as in [2]. The Amen Clinics has accumulated brain SPECT images over the past 28 years. The largest data set available corresponds to patients with ADHD, and it has images of 1,583 subjects. There are very few images of healthy brains (93). The first challenge is thus the small sample sizes available to train the machine. Furthermore, verifying accuracy with the test scans suffers from the imbalance between the number of healthy samples compared to the rest. For example, if the proportion of healthy samples in the test set is only 10%, then a learning algorithm that always predicts illness will achieve 90% accuracy (it only gets the 10% healthy diagnostics wrong).

The second challenge that we face is classification. Mathematically, a classification is well defined only when its classes are disjoint and cover the whole sample space (i.e. it forms a partition of the sample space). However, medical conditions are not mutually exclusive. That is, a patient with ADHD may also suffer from MDD, a suicidal attempter may also suffer from OCD, and so on. The average number of conditions per patient for the Amen Clinic is actually 4.2. The goal is to eventually include over 20 conditions. In theory, multiple conditions pose no problem, one simply considers all possible combinations of the medical conditions, which necessarily form a partition. However, the latter is not feasible in this study, because many classes would have extremely few samples.

Our contributions are:

– **Validation of the SPECT technology for diagnosis.** In this paper, we train a binary deep convolutional network for each of the three conditions ADHD, MDD and OCD. For deep learning we used compressed images from the scans exclusively. We did not use all of the 3D information from the SPECT scans, only the 2D surface images, to reduce the dimension of the input. Surprisingly, this seems to be enough for the deep CNN to distinguish the condition being tested from healthy. The results show high accuracy in prediction, see next section.

- **Social and economic impacts.** Our results show that SPECT scans can help early detection of mental illness, which can then improve treatment. There are great social and economic benefits in having tools that allow earlier treatment of mental illness. In particular, our method will allow family doctors and school counselors to refer potential patients to appropriate treatment early on.
- **Amber zone.** In this paper, one binary classifier is trained to distinguish SPECT scans that show each particular condition from the healthy. We express the result of the algorithm as a raw score. Usually deep learning networks apply a final activation function to that raw score and then a threshold comparison; we do not perform this step. Instead, we keep the set of raw scores found when training the algorithm as a frequency distribution; when implementing the algorithm this frequency distribution is used to express the outcome as an estimate of the probability that the scan shows the condition. In our view, this is a more useful way to proceed than just providing a yes/no output. In practice, a patient's scan would be fed into several algorithms, each detecting a particular condition; the estimated probabilities obtained would give the doctor information about each condition and an approximate ranking of the conditions from which the patient most likely suffers. For each algorithm, if the probability is above a certain level (say, 80%) then the result is labeled a clear positive (the "Red" zone), if it is lower than another level (say 20%) it is a clear negative (the "Green" zone), and if the probability is anywhere in between then the result is ambiguous (the "Amber" Zone). This effectively builds "soft" classifiers to provide better information to the medical practitioner, rather than provide only a binary output. In this particular study this is a way to avoid using a multi-class algorithm, which as we have explained above is not possible here due to the small number of samples.

Related work. An important difference between this and other research efforts for image processing is that we do not propose a "hard" classifier. The closest work that we have found to our CNN model is reported in [2]. Those authors use deep CNN models to classify each pixel of the brain scans in five classes: non-tumor, necrosis, edema, non-enhancing tumor or enhancing tumor (five classes). They propose an architecture with two scales (called the TwoPath model), and another one with a cascading model. Another one is Byra *et al.* [1], who studied automating and improving liver steatosis assessment. Their study compares their CNN model with two well-known methods of assessing liver steatosis in ultrasound images: the hepatorenal sonographic index (HI) and the Gray-level co-occurrence matrix (GLCM)

This paper extends a project that was submitted at the 2017 CUNY IBM-Watson[¶] competition, and which was awarded Third Prize (the team leader was Thomas Ward, Felisa Vázquez-Abad was the academic mentor).

[¶] Watson is a system developed by International Business Machines Corporation

2 Main results: CNN models for single conditions

Each model's architecture presented below was chosen through cross-validation. Then the best performing design in terms of accuracy was selected. During the testing phase, since we had limited negative samples, we couldn't employ cross-validation, instead we used hold-out sample sets for the ADHD, MDD, and OCD models. The negative (Healthy) samples are constant throughout all hold-out sets, and the positive samples correspond to the condition of related model. These hold-out sample sets were selected randomly once, and there was no human inspection of the test samples that would introduce bias into the models. Therefore the models never see the test samples, and provide an accurate measure of their validity.

2.1 Direct (or "vanilla") CNN model (no transfer learning)

We built a standard CNN model for binary image classification, described in Figure 1, using the Python package Keras, with Tensorflow as backend, and running on a GeForce GTX 1080 Ti GPU. The training set had 1,266 images of ADHD patients and 74 healthy controls, while the test set had 314 and 19, respectively. Training achieved 100% accuracy, with 21,759 parameters (execution time 8'44 minutes).

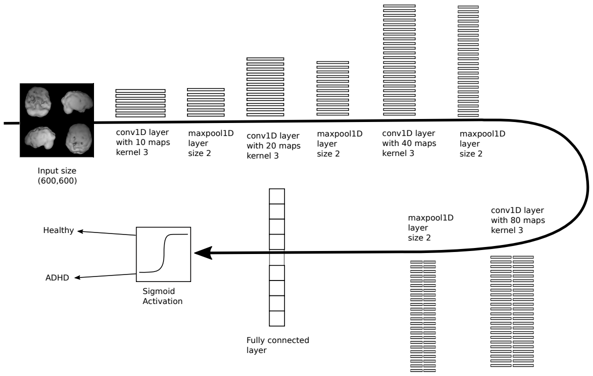

Fig. 1. Direct CNN model

The accuracy for the test sample is an important measure, as it provides information on how the algorithm performs when using it for classification: it provides the fraction of correct classifications. The available data is divided into a training set and a test set. The parameter values θ that are optimal for the training data are then fixed for the model (they are "the machine"). With these fixed parameters, each of the samples in the test set is used as input x to the algorithm, and a final raw score is obtained. This is a real number $v(x, \theta)$, positive or negative. The last step applies an activation function, in this case the sigmoid function:

$$\sigma(v(x,\theta)) \;=\; \frac{1}{1+e^{-v(x,\theta)}}.$$

This is a number between 0 and 1, commonly called the "score". By default, a test image is classified as negative (resp. positive) if the score is less than (resp. larger than or equal to) 0.5.

To better understand the performance of the algorithm, given that we have significant imbalance between negatives (healthy) and positive (ill) images, we provide accuracies by class, rather than overall accuracy.

Table 1 provides accuracies by class, rather than overall accuracy. For instance, the row for the ADHD algorithm shows that of all the true ADHD test images, 99.4% were correctly classified as positives and of all the healthy test images, 78.9% were correctly classified as negative. Errors of this magnitude are to be expected when there are so few negative samples. Each model has 21,759 parameters.

Model Name	Train Time	Train Size	Test Size	Train Acc. Pos.	Train Acc. Neg.	Test Acc. Pos.	Test Acc. Neg.
ADHD	8'44s	1,340	336	100%	100%	99.4%	78.9%
MDD	3'48s	601	151	100%	100%	98.5%	84.2%
OCD	4'20s	699	176	100%	100%	97.5%	100.0%

Table 1. CNN models training times, sample sizes and accuracies.

Figure 2 displays the results obtained on the training and testing sets as histograms. It is clear that the CNN separates the positives and negatives adequately.

These results validate the SPECT imaging technology for aid in diagnosis. Our algorithms use only the images, yet as shown on Figure 2 the raw scores obtained show a clear separation between patients with the condition tested against healthy ones.

2.2 Cross-validation with few samples

Here we demonstrate how having very few samples (in our case for the healthy set) makes using validation of a model misleading. We show in each of our models (one for each condition we study) that validation is not straightforward. Instead this is one of the reasons why we propose the "amber zone" for soft classification, as detailed in the following section.

For each model, the total sample was divided into a training sample (80%) and a test sample (20%). In what follows, we describe the results for ten different random splits of the available data. The average accuracy for positive and

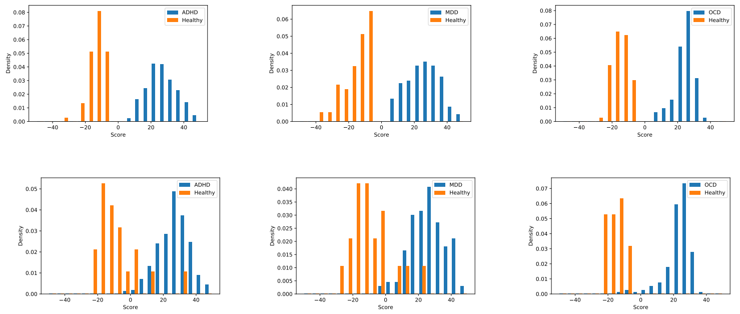

Fig. 2. Histograms of raw scores for ADHD, MDD, and OCD models. Upper row are training scores, lower row shows testing scores.

negative samples for the ADHD, MDD, and OCD models are (99.1%, 77.9%), (98.3%, 77.9%), and (99.1%, 92.1%) respectively.

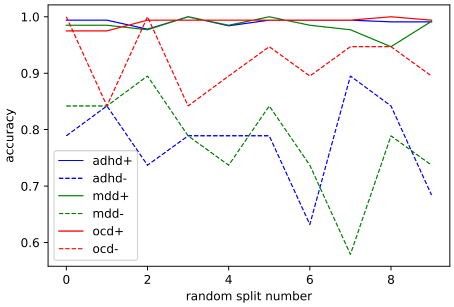

Fig. 3. Accuracy for ten randomly selected splits for models ADHD, MDD, and OCD.

Importantly, Figure 3 shows the results per split. Clearly the mean accuracy does not provide enough information: the variance of the accuracy for negative samples (the images for the healthy) is too big for the ADHD and the MDD models and it lowers the statistical significance of the averages.

According to our medical team, there is a "diamond shape" pattern in OCD SPECT images. This pattern could explain why the OCD model has higher accuracy and smaller variance for the healthy images, than either the ADHD or MDD model. We speculate that a reason for the low accuracy values for healthy images when tested for ADHD or MDD could be that some SPECT images of healthy subjects appear to have patterns that resemble mild affliction of ADHD or MDD. Interestingly, it appears that the patterns in the healthy images are quite different from those of OCD.

2.3 The amber zone

"Training" a CNN means fitting the parameters, using the samples in the training set (the latter is distinct from the test set). The fitting is performed using the standard stochastic gradient or other optimization method. Denote by X the input vector (an image), and let $v(X, \theta)$ denote the result from the CNN prior to the final activation, that is, $v(X, \theta) \in \mathbb{R}$ (the raw score). Neural network classification techniques are inspired by logistic regression (a well-known statistical technique). Logistic regression **assumes** that the probability of an image being positive is the function $\sigma(v(X, \theta)) = 1/(1 + e^{-v(X,\theta)})$. Machine learning uses a sample of X's to fit the parameters needed to compute $v(X, \theta)$. Denote $\hat{\theta}$ the resulting parameter estimate.

Once training has been performed, the resulting algorithm (= machine) can be applied to produce a score for each new image that it receives. This is the algorithm that practitioners will be using when they classify new images, for which there is no diagnostic yet.

While the statistical model for the optimization assumes that the probability that an image x being positive (ADHD patient) is $\sigma(v(x, \hat{\theta}))$, this assumption may not agree with reality. The model for logistic regression is mathematically and computationally convenient in order to carry out the tuning of the algorithm's parameters in an efficient way. However, for a random individual i with SPECT scan $X(i)$, we believe it is very likely that Prob(i has ADHD) $\neq \sigma(v(X(i), \hat{\theta}))$.

To illustrate this, we show in Figure 4 a typical cumulative distribution of the raw scores obtained, against the sigmoid function. The step function is the empirical distribution. For example, a raw score of -8 will give an estimate of 0.001 for the probability of having ADHD when using the sigmoid, while the true fraction of individuals with a score of -8 or less that did have ADHD is almost 20%. Calling $\sigma(v(X(i), \hat{\theta}))$ a "probability" may lead to misinformation. We believe that the analysis of the conditional histograms can provide better information than the scores alone.

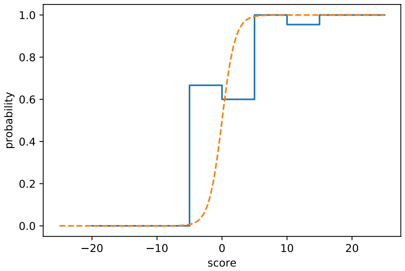

Fig. 4. Empirical probability of having ADHD versus the sigmoid.

Although the model's parameter vector $\hat{\theta}$ provides the best fit for a sigmoid function, the true probability is not necessarily a sigmoid. This is the case in general for all parametric statistical estimation. Furthermore, a binary classifier must ultimately convert the raw scores into binary values, which is done using a threshold. Under the sigmoid model, the "default" threshold for the raw score $v(X, \hat{\theta})$ is chosen as 0.0. This implies that, if the model were true, the classifier is weighing equally the type I and type II errors. It is apparent from Figure 2 that we may choose other thresholds to achieve a different balance between errors. In particular, when using these techniques for detection in order to help the medical practitioners for diagnosis, we think it is better to provide a different criterion. That is, instead of a *hard* classifier, we propose a *soft* classifier. Specifically, we suggest using boundaries $\underline{v}$, $\bar{v}$ such that for patient i,

$$\begin{aligned}
&\mathrm{Prob}(i \text{ has } ADHD \,|\, v(X(i), \hat{\theta}) \leq \underline{v}) \leq \alpha_1 \\
&\mathrm{Prob}(i \text{ has } ADHD \,|\, v(X(i), \hat{\theta}) \geq \bar{v}) \geq 1 - \alpha_2.
\end{aligned}$$

(Here α_1, α_2 are chosen levels of significance, for instance 0.05.) These two thresholds will be provided to the medical practitioner as the result of the analysis of the brain SPECT images: if the raw score $v(X, \hat{\theta}) < \underline{v}$ the algorithm will suggest "healthy", if $v(X, \hat{\theta}) > \bar{v}$ then it will suggest "ADHD". Within the "ambiguity zone" between $\underline{v}$ and $\bar{v}$, that we call the Amber Zone, the algorithm will provide an estimate of the true probability, that is: $p(v) = \mathrm{Prob}(i \text{ has } ADHD \,|\, v(X(i), \hat{\theta}) = v)$.

In practice one would probably want to choose the empirical score distribution for the combined training and test sets, in order to obtain a better estimate of the probabilities. Smoothing techniques could also be used to turn the step function into a continuous one.

3 Conclusion and Future Work

We have shown that deep learning can help diagnose ADHD, MDD and OCD based on SPECT scans. Our results are effective in testing each of the conditions against healthy samples, especially in light of the relatively small sample sizes which were available. Typically, medical doctors arrive at diagnoses using several clinical data, while our algorithms used only the scan images of the brain. This validates the SPECT technology, as it shows that it must detect physiological characteristics of those conditions. Using our algorithms, physicians' diagnoses would potentially be more accurate, faster and more economical. Eventually this approach could lead to better predictions of possible mental illness conditions, which would in turn improve patients' outcomes.

Each of our algorithms is trained to detect a specific mental condition. In machine learning, it is customary to choose a threshold (say, 0.0) and to classify raw scores into positive/negative according to raw scores. We do use this procedure to train our algorithms but when implementing them, we express the raw scores from training and testing as a frequency distribution (distinct for each condition tested). Then, we find where the raw score of the new scan falls within that empirical distribution. If the new raw score is smaller than a specific negative value we label the result a "clear negative" (Green Zone); if the score is larger than a specific positive value the score is labeled a "clear positive" (Red Zone); scores that fall in between are labeled "ambiguous" (Amber Zone). The practitioner is given the probability achieved by the score on the frequency distribution, meaning the higher that probability, the more likely a positive diagnostic becomes. The Green and Red Zones levels are determined by looking at Type I and II errors. We believe this is more useful in practice than a simple yes/no result, especially when a practitioner is faced with the outcomes of multiple algorithms, each detecting different conditions.

We did not apply multiple classification because we did not have enough samples. When dealing with multiple and possibly overlapping conditions, it is necessary to combine the results from the different algorithms that would be used in parallel. Using our first model presented in Section 2 for various conditions (here we used ADHD, MDD, or OCD), each brain scan image is assigned a raw score for each of the mental illnesses (in this case, three raw scores). We performed several experiments (not shown) mixing three conditions (healthy, ADHD and OCD), all with similar results. If all samples from various mental illnesses are mixed in the histogram, the resulting plot shows multimodal distributions, and it would be very difficult to draw meaningful conclusions. Instead, the histograms corresponding to each individual condition are plotted using different colors. It can be seen that each of the algorithms (say, one trained to detect ADHD) produces histograms that have a relatively smooth shape for each of the conditions of the subsamples (here, there would be four subsamples: healthy, ADHD only, OCD only, and patients with both ADHD and OCD). Typically the healthy inputs show empirical distributions that are skewed to the left and are well separated from the non-healthy raw scores. Those of other conditions present distinct shapes and they overlap. The overlaps are significant and further analysis of the raw scores (such as clustering techniques) may yet provide insight into the mental condition of patients. To illustrate this idea, Figure 5 shows a scatter plot with the ADHD model scores on the x-axis and the OCD model scores on the y-axis. From the scatter plot in Figure 5 we may already draw some ideas for future research. Looking only at the isolated conditions (all but the green dots), it is apparent that the space is divided into three clusters, corresponding to red (healthy), blue (ADHD) and orange (OCD) raw scores for the SPECT images. We performed a dozen experimental designs combining various conditions and methods for training, and all of them showed more or less the same pattern. We plan to further analyze high dimensional clustering when adding more conditions to the study. It may well turn out that clustering results from the binary machines will yield interesting new knowledge for psychiatrists.

Co-morbidity is common in mental illnesses, and it is not unusual to see mixed results for some patients, so the Amber Zone may yield insight about the

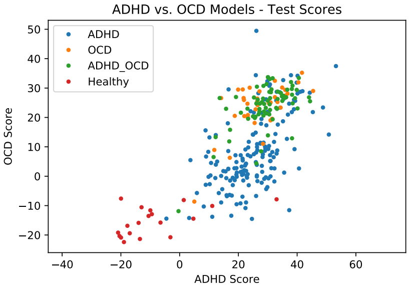

Fig. 5. Correlation between ADHD and OCD models scores.

patients condition. We plan to build a stochastic model to generate simulated functional images of the brain, based on its structure, to use in the pre-training phase, in order to improve the results, given extra conditions and small sample sizes.

The complexities behind diagnosis of mental illness are profound even for medical specialists. No two cases will ever be the same. This paper shows that artificial intelligence can help doctors ask meaningfully better questions in order to provide a more individualized plan for the patient. These tools will enhance the ability of medical doctors to provide an accurate diagnosis but are not meant to replace them with blind classification machines.

This study focuses on SPECT images because of the expertise of our team in this technology and the availability of data. However, our machine learning methodology can also be applied to other type of medical imaging. Future research may involve using different inputs for comparison.

Acknowledgement

This work was partially supported by the CUNY Institute for Computer Simulation, Stochastic Modeling and Optimization.

References

1. Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michaelowski, A., Paluszkiewicz, R., Ziarkiewicz-Wroblewska, B.: Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images.(report). International Journal of Computer Assisted Radiology and Surgery **13**(12) (2018)
2. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical Image Analysis **35**, 18 – 31 (2017). DOI https://doi.org/10.1016/j.media.2016.05.004
3. Pelham, W., Foster, E., Robb, J.: The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. Journal of pediatric psychology **32**(6), 711–727 (2007)