ADHD in children can be detected by a machine learning model which measures their pupil response to working memory tasks

WOODBURN, Michael1,3; Ossandón, Tomás2; Wainstein, Gabriel2; Rojas-Líbano, Daniel2; KHANNA, Rahul1,3

1Austin Hospital, Victoria, Australia

2Pontifical Catholic University of Chile, Región Metropolitana, Chile

3The University of Melbourne, Victoria, Australia

Contents

[Abstract 1](#_Toc13499666)

[Keywords 1](#_Toc13499667)

[Introduction 2](#_Toc13499668)

[Methods 3](#_Toc13499669)

[Results & Discussion 10](#_Toc13499670)

[Conclusions 14](#_Toc13499671)

[Bibliography 16](#_Toc13499672)

[Acknowledgements 19](#_Toc13499673)

[Tables 19](#_Toc13499674)

# Abstract

Attention Deficit/Hyperactive Disorder (ADHD) is diagnosed in children based on reports from parents, teachers, and close contacts. Although considered a neurodevelopmental disorder, no neurological or biological tests are used for diagnosis. However, it is plausible that causative neurological circuits may affect physiological markers in a measurable way. Differences in pupil behaviour have been identified in ADHD1. The pupil response to a working memory task was recorded from 22 control, and 28 ADHD-diagnosed children. Machine learning methods were used to train a deep convolutional neural network (CNN) to recognize ADHD-characteristic responses. The pupil-time series within the data were transformed using the Gramian Angular Summation Field2 method which enhanced their interpretation by the deep neural network. By assessing the aggregate of many trials of each child, it was possible to establish a diagnostic model for ADHD with modest internal accuracy. The novel machine-learning based diagnostic approach used in this study may be applicable to other disorders.

# Keywords

Pupil dilation ADHD children deception biometrics psychiatry machine learning

# Introduction

## Diagnosing ADHD

Attention-deficit/hyperactivity disorder (ADHD) is a chronic psychiatric condition characterised by impaired behavioural inhibition3. The prevalence of ADHD in children is likely to lie between 5% and 10%4. It correlates with many negative academic and social outcomes5 which begin early in childhood and tend to persist into adulthood6. Early intervention and diagnosis mitigates these outcomes and influences disease progression7. Clinical practice guidelines for doctors8 recommended an algorithmic approach with the following diagnostic steps:

1: The primary care clinician should initiate an evaluation for ADHD for any child 4 through 18 years of age who presents with academic or behavioural problems and symptoms of inattention, hyperactivity, or impulsivity

2. To make a diagnosis of ADHD, the primary care clinician should determine that **Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V-TR) criteria** (emphasis added) have been met

3. In the evaluation of a child for ADHD, the primary care clinician should include assessment for other conditions that might coexist with ADHD

The algorithmic approach usually results in children only being assessed for ADHD after they have developed academic or behavioural problems.

## Limitations of current practice

Diagnosis of ADHD in DSM-V relies upon behavioural reports, often received from parents, teachers, and healthcare workers. This reliance emphasises functional impairment, but also increases vulnerability to reporter bias in regards to traits such as gender9,10 or race11. Moreover, environmental effects may influence the severity of ADHD presentation. For example, exercise alleviates hyperactivity12,13 and the time provided for exercise during school varies significantly on both national14,15 and local16 levels. These variables contribute to over-diagnosis or under-diagnosis in certain subpopulations17.

The centrality of behavioural reports in diagnosis is partly because there is no biological test at present which reliably detects ADHD, despite it being a neurobiological condition. If there were a robust, stable, signature trait of ADHD that could be derived during a single consultation then diagnosis may be made earlier and with greater accuracy. Moreover, if the trait were sufficiently objective, it could reduce the costs of diagnosis and expand the reach of treatment

## The pupil a neurobiological ‘tell’ in ADHD

It is unknown whether such a signature trait exists. Several past studies provide evidence that pupil activity differs in some respects1,18 in children with ADHD. The  iris dilator muscleis mediated by the locus-coeruleus (LC) region which governs norepinephrine (NE) activity in the brain19. Given that the aetiology of ADHD is associated with catecholaminergic transmission deficits, we hypothesize that an underlying difference in this region results in both behavioural and pupillary differences. Therefore it is possible that there is a characteristic pupil response present in children with ADHD that can be detected.

## How useful is the pupil response?

This study seeks to identify that characteristic pupil response and quantify its utility as a predictor of ADHD. Previous research suggests that both pupil dilation1 and gaze parameters18,20,21 have diagnostic utility. If the pupil is useful in diagnosing ADHD, the effect is likely to be subtle. It is also likely that a significant amount of ‘noise’ in the pupil measurements may obscure the underlying characteristic in many cases. Therefore this study makes use of a neural network22 program which is highly sensitive to differences between classes comprised of complex or noisy datasets.

# Methods

Study design

We conducted a secondary analysis of a dataset published by Ossandon et al in Nature Datasets with the original author’s consent.

Ossandan’s primary finding was that pupil size correlated with the subjects’ performance on the tasks and reaction time. Moreover, there were differences in the maximum pupil diameter between ADHD and control groups after probe onset. It is our belief that this difference suggests a diagnostic model is possible.

A close up of a map

Description automatically generated

Figure 1 Maximum pupil diameter measured after probe onset. Each dot represents the mean session value for a subject (i.e., average across trials). Horizontal bars correspond to session averages across subjects. Reproduced with permission from Pupil Size Tracks Attentional Performance In Attention-Deficit/ Hyperactivity Disorder (Ossandon et al)1

This re-analysis sought to build upon the findings in that study. If there are, on average, differences in peak pupil size in the ADHD vs control group, then it was hypothesized that a neural network may be used to generate a predictive algorithm. Furthermore, the use of a neural network would allow hitherto unrecognised complex pupil behaviours to increase diagnostic accuracy.

The Ethics Committee of the School of Medicine of the Pontificia Universidad Católica de Chile (Protocol number 11082) approved the study methods and found that they were consistent with the Declaration of Helsinski.

## Advantages and disadvantages in neural networks

The deep convolutional neural network (CNN) functions as a black box. Although its accuracy can be thoroughly assessed, it is mathematically fraught to draw inferences from the model regarding which features of the pupil-time series differ between classes. Thus it was chosen in order to maximize diagnostic potential with a necessary loss in explanatory power. The neural network’s utility is in its real world application.

Dynamic Time Warping23 (DTW) was an alternative to the CNN. This algorithm has in the past demonstrated good performance where there was strong similarity amongst univariate time series of the same class. However, given that our time series is of great complexity and more abstract relationships may exist within the classes the DTW algorithm was not considered state of the art24.

The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification25 (HIVE-COAT) was another alternative. As of 2018 this algorithm achieved the highest accuracy. However, it is also one of the most computationally intensive algorithms and was therefore prohibitively resource intensive to use within this study.

We chose to use resnet22. It is a deep CNN which achieved victory in the 2017 Kaggle ImageNet competition. It had become apparent that although Resnet was developed to recognize the presence of real objects in photographic images, it was chosen due to its demonstrable effectiveness24,26 in classifying univariate series encoded as images. In order to make use of that serendipitous benefit, the time series in our dataset were encoded to GASFs, which can be represented as colour images.

Participants

50 children were recruited. 28 had been diagnosed with ADHD, and 22 were the non-ADHD control group. There were no significant differences in age, IQ, and educational level between the two groups. Parents gave informed consent to allow their children to participate, and children’s assent was formalized with a signed document. Children were advised to cease medication 24 hours prior to the day of the experiment.

Task

Subjects were tested on their performance in a Sternberg-type delayed visuo-spatial working memory task. The children were instructed to memorize 1 or 2-dot arrays in which dots appeared in any of the 4x4 grid squares. After the memorization arrays a ‘distractor’ image was presented in 75% of cases, which contained content of either array type, emotional content, or neutral content. In 25% of cases there was no distractor. They were then shown a probe array and asked to press a button to indicate whether they believed a dot had been present in the same specific grid square within the previous three arrays.

This task is designed to specifically challenge the working memory of the participants and is adapted from a similar study by Dolcos and McCarthy27 on healthy individuals. In other studies28 similar working memory tasks have revealed performance differences between ADHD and control groups. This may be due to the high demand in the task for sustained attention.

Each subject was required to perform 160 trials.

A close up of a piece of paper

Description automatically generated

Figure 2: A single trial of one subject performing a Sternberg-type working memory task. After seeing the probe, subjects were challenged to indicate whether the position of the dot had been presented in one of the previous arrays. Reproduced with permission from A pupil size, eye-tracking and neuropsychological dataset from ADHD children during a cognitive task29

## Pupil Measurement

Subject’s heads were fixed by a forehead and chin rest (SR Research LTD) 30cm in front of a computer screen. The screen’s luminosity varied on the order of 1 lux. The narrow range of variation in luminosity controls for light-mediated pupil diameter changes. The Eyelink 1000 (SR Research Ltd, Mississauga, Ontario, Canada) eye-tracking device was used to record pupil diameter with a sampling frequency of 1000Hz. In optimal conditions, within the 8s task, 8000 measurements of pupil diameter were taken.

Pupil-time series processing (original dataset)

*The original dataset*29 *was published in Nature Datasets. These steps were taken during the creation of that dataset.*

1. Periods of blink in which no pupil data was taken were interpolated with the Matlab® *spline* function.
2. Pupil-time series were baseline adjusted.
3. Pupil-time series were smoothed with a bandpass Butterworth filter between 0.025Hz and 4Hz
4. Outliers were defined as measurements for which the derivative function lay more than 3 standard deviations outside the norm. They were discarded under the assumption that they were artefacts.
5. Pupil-time measurements were then normalized using a Z-score.
6. Pupil-time measurement were stored as time series with labels indicating the subject, group, and trial in a .mat file.

Pupil Size (Z Score)

Time (0.1s)

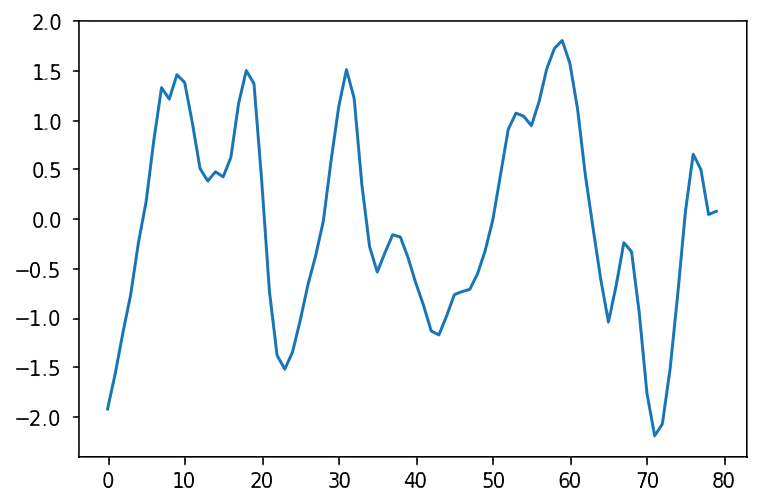


Figure 3 This is the pupil-time series of a randomly selected trial (subject 10 trial 106). The length of the trial is 8s and all arrays and probes were displayed at fixed times during these recordings (see figure 2)

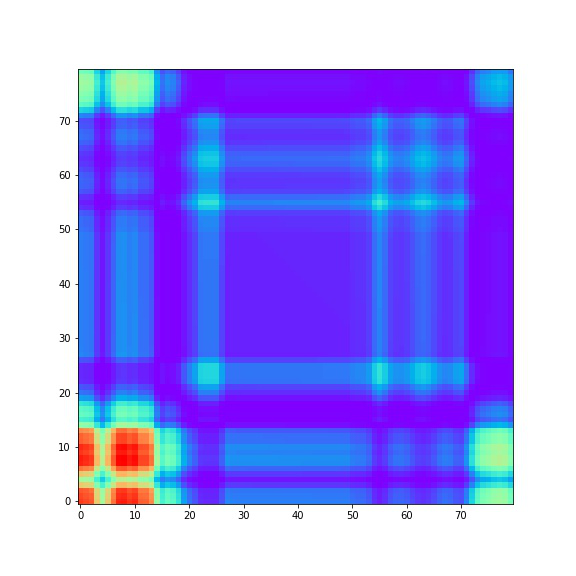
## Pupil-time series processing (extended)

*(Further processing was performed for this publication in addition to those above):*

### Cleaning pupil-time series

1. Pupil-time series were screened for missing data. Any pupil-time series in which there was more than 1s continuous of missing data was discarded. For the remaining time series, linear interpolation was used to fill in missing pupil data.
2. Pupil-time series were scaled. Pupil Z scores were observed to fall unfailingly between values -3 and 3. Therefore, these were chosen as the minima and maxima, and all series were scaled proportionately to those values to ensure all adjusted pupil-time series lay between -1 and 1. i.e a value of -3 became -1, and 3 became +1. Any values lying outside of those ranges were clipped to values of -1 or 1.
3. Data compression by a factor of 100 was applied to reduce computational demands. Therefore, for the 8000 measurements taken in 8s, 80 were preserved at equal intervals. Given the smoothness of the time series, this loss of resolution was tolerable and is unlikely to result in a loss of useful information.

### Encoding to a GASF Matrix

1. Each time series was used to generate an 80x80 Gramian Angular Summation Field (GASF)2,30. The python library *pyts* was used to compute the values of each cell of the field. A GASF is a type of matrix that represents a 1-dimensional array (for instance, a univariate time series) as a 2-dimensional grid. This type of data encording was chosen in order to allow us to use the resnet50 deep CNN.
2. An image was created with the *matplotlib.pyplot* library to represent the GASF pictorially. The ‘rainbow’ colourmap was chosen as it is reasonably interpretable by humans and has previously been used to train convolutional neural networks (CNN) to a high level of accuracy26.

Time (0.1s)

Time (0.1s)

Figure 4 The corresponding GASF for the above pupil-time series. The colour of the cells is proportional to the value of the GASF at that coordinate. Trial is subject 10 trial 106.



Figure 5 The rainbow colourmap in the python library matplotlib. Values are low (left, blue) to high (right, red)

## Labelling of GASFs

The resultant directory contained 5873 GASFs, each corresponding to a single timeseries measured during one task done by a subject (table 1).

All GASFs belonging to the ADHD group were labelled ‘off-ADHD’ to signify that they reflected an ADHD subject’s trial without medication. All that were belonging to the control group were labelled ‘Ctrl’.

Not all subjects were equally represented in this set due to inequal distribution of missing data. The majority of timeseries of subjects 14, 35, and 45 specifically were culled in the data processing stage. However, given that all subjects performed 160 trials, retention was sufficient for this experiment. The mean number of GASFs represented per subject was 117 of the potential 160.

## *Creation of a workable Dataset*: Organizing Data for Machine Learning.

Subjects were divided into training and validation groups. Out of 50 subjects, 6 were randomly assigned to the validation group. The randomization process was designed such that 3 were selected from the ADHD group and 3 were selected from the Control group. The selected validation group contained subjects 23, 13, 24, 31, 42, 38. Therefore, the training group consisted of the remaining 46 subjects. The “sets” contained the trials (as GASFs) of all members in the group. The training set was of size 5131 and the validation set was of size 742.

GASFs were rescaled (down) to size 224x224 in order to increase speed of processing. Given the timeseries was 80 units in length, an image width of 224 pixels theoretically preserved all cell colours (and therefore values) represented in the GASF. No other transforms were applied, which preserved proportionality within the images.

## Generating the model:

The model trainer was run through the fast.ai library on a K80 GPU cloud server provided by Salamander.ai. ‘Tuning’ of the CNN by adjusting the learning rate achieved a small addition in accuracy. The outcome measured was the accuracy of classification of trials to either the ADHD or Control groups. This trained CNN is considered the prime model for this study. Class predictions for individual trials are shown in the results: *Classification accuracy per trial.*

## Deriving a useful prediction for each subject

The data of subjects in the validation set were specifically analysed in order to assess the fraction of each of the subject’s trials which were classified as ADHD. See results: *Fraction of trials classified ADHD for each subject in the validation set.*

## Testing on a subject in the blind

To test the model more rigorously, an additional division was made so that there were now three sets: training, validation, and test. This allows us to use the *Cross-vaidation* technique *(*also known asOut-of-sample testing) which tests whether the model can be generalized to independent data. In brief, the model is not allowed to interact with the test set in any way before it was used to assess the model’s accuracy. This simulates a real-life application of the model.

In contrast, the previous method (Deriving a useful prediction for each subject) allowed the model to receive feedback on its predictions towards our subjects of interest *as it learned.*

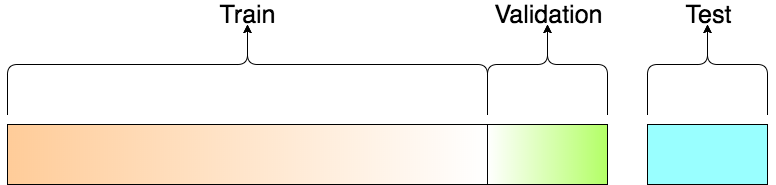
**

Figure 7 The separation of sets. The model can only interact with the training and validation sets during learning. Reproduced from Tarang Shah31.

One subject’s trials were selected and removed from the dataset. This was the test subject. Remaining trials were randomly distributed to either the training set or the validation set. The CNN recreated from scratch with these two these sets. It was then tested on the trials of the test subject, and the fraction of trials which were predicted to be ADHD-characteristic was returned. This process was repeated for all subjects, so that there were always 49 subjects within the dataset (comprised of the training and validation sets) and one without (the test set). See results: *Fraction of trials classified as ADHD for each subject in the test set.*

Although a new model is created each time a new subject is selected to be removed from the dataset, each model is sufficiently similar in that they have all learned on the same 48 subjects (and differ on 1 subject). Theoretically, these models will all be slightly less accurate on average than the model which learned on all 50 subjects, referred to earlier in this paper as the prime model. Therefore, in creating a ROC for these models, we make the approximate assumption that all models are the same. See results: *Diagnostic accuracy on test group*. This is done with the understanding that the AUC will be underestimated, which biases the null hypothesis. Moreover, it allows us to allocate one subject at a time to the ‘test set’ which simulates testing the model on a child who has not been involved in the development of the model. In an optimal scenario, the number of subjects could be increased to create a large ‘test set’ that could be used to create an ROC without creating 50 almost-identical models.

Code for all the above methods is available at <https://github.com/woodytwoshoes/ADHD_pupil>

# Results & Discussion

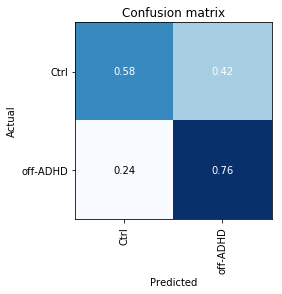
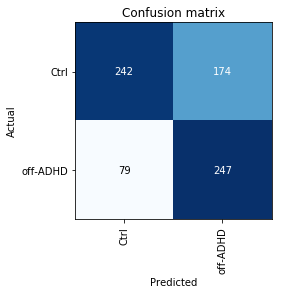


Figure 8 Classification accuracy per trial. These two confusion matrices represent the same results. However, the left has been normalised. "Ctrl" denotes the healthy control group. "off-ADHD" the ADHD group.

## Predictions for individual trials

This confusion matrix shows the model’s predictions for which class trial pupil-time series belonged to: ADHD (off medication) or Control. All predictions are made on the trials in the validation set.

This model achieved a sensitivity of 0.76 and specificity of 0.58. Overall accuracy is 0.67, the average of these two statistics.

The model demonstrated better accuracy in classifying off-ADHD trials than Control trials. It performed well above chance, confirming that there is an underlying difference in ADHD subjects which affects the pupil during this task. However, the significant error rate suggests a high amount of intrapersonal and interpersonal variation in this difference, which is consistent with the results graphed in Ossandon et al1 2019.

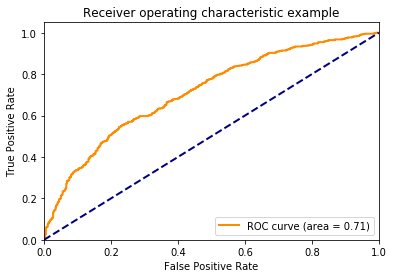


Figure 9 ROC Curve for trial classification. This ROC curve is derived from the 742 trials GASFs in the validation set and the model's returned probability that they belong to the ADHD class.

The CNN analyses a trial pupil-time series and outputs a single prediction and associated probability values for each possible class (off-ADHD and Control)

When we take the trial probability value for the class ‘off-ADHD’ as the statistic of interest, we can construct the above ROC curve. It demonstrates a modest accuracy in identifying ADHD-characteristic trials. It satisfies the basic requirement that ROC curves must be derived from more than 100 observations.32 Given that the values are not derived from a classic parameterized statistical model, it is difficult to judge the statistical significance (via p-value) of this result.

## Predictions for trials as grouped by subject/child

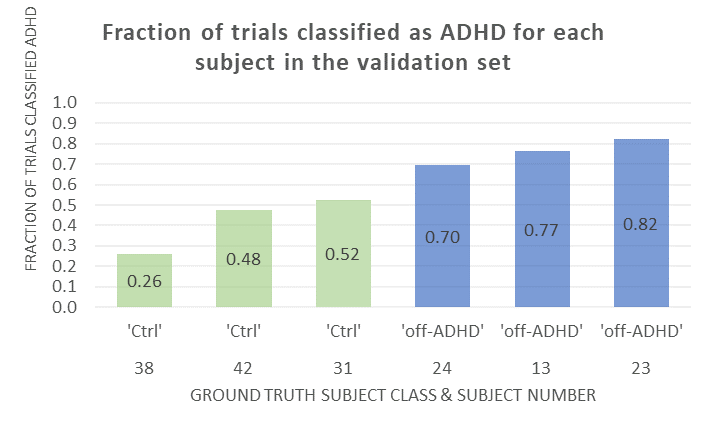


Figure 10 Diagnostic accuracy on validation group. The fractions of trials (considering only the trials which were available as GASFs) predicted to belong to the ADHD class are shown. If the trial was not predicted to belong to the ADHD class, it was necessarily predicted to belong to the control class.

This graphs shows the fraction of trials which were predicted to belong to the ‘off-ADHD’ class for each of the subjects in the validation set. Note that there was a clear linear boundary between the ADHD (off medication) class and the control class. This was partly the result of a randomization that was favourable in this instance. For all subjects except subject 31 (fraction ADHD= 0.52), the majority of trials were correctly classified. The trials of subjects 31 and 42 (fraction ADHD= 0.48) were classified with accuracy close to 50%, reflecting the model’s poorer ability to classify control trials correctly (a weakness identified in results part 1).

## Predictions for trials as grouped by subject/child when using the train-valid-test set method:

Figure 11 Diagnostic accuracy on test subjects. All trials for each subject which were available as GASFs were considered. Note the trend that for subjects with ADHD (blue,dark), a higher fraction of their trials were classified as ‘off-ADHD’. Values are given in table 1.

This graph shows predictions derived from the training-validation-test method.

The above graph shows the proportion of trials classified as ‘off-ADHD’ for each subject by the model created while that subject was held apart in the test set.

Note that the same subjects 23, 13, 24, 31, 42, 38 appear in this graph as appeared in the graph: *Fraction of trials classified ADHD for each subject in the validation set*. However, the proportion of their trials which have been classified as ‘off-ADHD’ has shifted towards 0.5. The model is less accurate when making predictions on trials it had not previously seen.

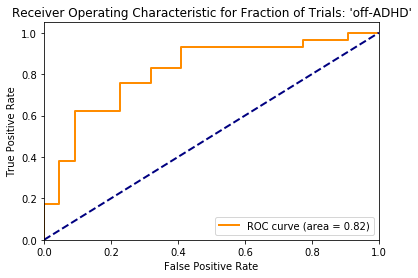


Figure 12 Diagnostic accuracy on test group. The corresponding ROC curve for the graph: Fraction of trials classified as ADHD for each subject in the test set

When we take the fraction of a subject’s trials which have been classified ‘off-ADHD’ as our diagnostic statistic, we achieve an even greater area under the curve (0.82) in the ROC than we did for individual trials (AUC = 0.71). This improvement in sensitivity and specificity is the result of considering a large number of trials for each subject, and making a judgement based on the trials in aggregate. This indicates that in reality, it is judicious to thoroughly assess a subject across many repeated trials and to look at the overall picture of classifications. The ROC should be interpreted with its limitations in mind; it does not technically satisfies the basic requirement that ROC curves must be derived from more than 100 observations,32 as only 50 subjects were included.

## Experimenting with a different test statistic

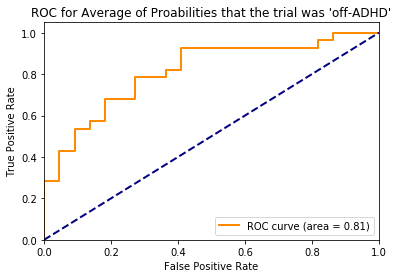


Figure 13 A ROC was generated with the mean of the probability values of all a subject's trials

For each trial that the model sees, it returns both a prediction (‘off-ADHD’ or ‘Control’) and the corresponding probabilities assigned to each class. By taking the fraction of trials classified ‘off-ADHD’, as the test statistic for a subject, the certainty of the model for each trial is not taken into account (only whether the model was greater or less than 50% certain that the trial belonged to the ADHD class). As an additional experiment, the probability values (which give the probability that the trial belongs to the ADHD class) for each of a single subject’s trials were collected, a mean was found, and that mean assigned to the subject. From that mean probability and the ground truth subject class, an ROC was graphed.

The new ROC had no greater AUC than the initial one. Therefore, it was apparent that the new statistical method was not superior to the first, despite the fact that it incorporated certainty values returned by the model.

# Conclusions

## Internal Validity

There arecharacteristic differences between pupil behaviour in children with ADHD and healthy controls. These differences are sufficient to train a mode to differentiate ADHD vs Control trials with somereliability in children within the same study. By taking the fraction of trials of a single subject which were classified ‘off-ADHD’, the accuracy per child increases.

Taking the proportion of trials which were classified as off-ADHD was a pragmatic choice of statistic, but unlikely to be ideal. A model which is able to assess all 160 trials at once and returns a single prediction per subject could feasibly uncover and make use of a superior statistic. In an extreme case there may be a pathognomonic pupil trait which appears in only one trial which is sufficient to diagnose ADHD in the subject. A better model may recognize it. The engineering of such a model is an exercise for a future study.

## External Validity

These differences in the pupil may be generalizable and may be of diagnostic value more broadly. It is not clear whether the pupil data is predictive *in addition* to the DSM criteria. The most significant barrier is experimental setup; it is unclear how the model created to predict ADHD under the present experimental conditions will fare under novel conditions. A standard set of experimental conditions should be established. This set of conditions should particularly emphasize consistent luminosity, timing of tasks, and pre-test preparation of the child. Furthermore, the methods used to process pupil data should be easily replicable and available online. To this end, the data pre-processing code has been made available at <https://github.com/woodytwoshoes/ADHD_pupil>

The best practical use for our model is to distinguish ADHD from other conditions which may mimic it. Because our model receives inputs from the neurophysiology of a subject, it may aid in differentiating children with true ADHD from those with hearing problems, dyslexia, sleep disorders, autism, and mood disorders33. Given that the prescription of stimulants may exacerbate34–36 many of these conditions, a model which biases the correct diagnosis is especially valuable.

## Improving the Model:

The model can be improved by the following means.

1. Gathering more data  
   Further data collection, incorporating results from other similar experiments, is likely to improve the external validity of this model. Data gathered from a standardized set of experiments will inevitably vary in certain aspects; systematic error can be introduced via a number of minor experimental differences and due to pupil sensitivity. However, these variations can be overcome with a large enough dataset which may empower the neural network to identify the true underlying characteristic of the ADHD pupil behind the noise and error. This may be the fundamental advantage of a machine learning based approach over a classic heuristic-based approach (for instance, measuring the maximum peak pupil dilation after probe presentation). In a classic approach, a change in experimental conditions might cause enough distortion in the data that the criteria for diagnosis be met erroneously or missed. If the prediction is being made by a neural network, it can learn to make correct predictions despite different experimental conditions.

### Changing the learner

As neural networks are improved, additional accuracy may be achieved by simply substituting the resnet50 architecture for a superior model. In ideal conditions, a model designed specifically for GASF representation of time series would be adopted for this analysis. As things stand, the resnet50 was developed to perform well on more conventional image classification. Its utility in this study is a serendipitous quirk of its design. It is likely that resnet50 will be superseded by a tailored time series classification model.

### Changing the encoding technique

Whilst GASF is a good encoding technique for univariate time series, there is increasing evidence that a combination of two different encoding techniques yields superior accuracy after training37. This may be due to the capacity of different encoders to emphasize different specific attributes of the underlying data to the CNN. Other encoders with strong potential are Markov Transition Fields and Recurrence Plots. Given that resent50 can be modified to encompass multidimensionality, it is also theoretically possible to feed all 160 trials per subject into the learner as a single item. The potential accuracy of a model trained on 160-dimensional items is unknown.

## Implications for other ADHD Biometrics

These methods are generalizableto other traits within ADHD children: gaze direction, motor agitation, body temperature, facial expression, etc. Of the former, gaze direction is the logical next biometric to incorporate within this model. Gaze position is contained within the original dataset29 in raw coordinate form but was not included in this analysis. There is some evidence that ADHD subjects are less able to suppress saccades18,38,39. However, these findings were elicited in experiments designed to challenge subjects to suppress the saccades, and there is no guarantee that the same effect would be present in this experiment’s dataset.

Another potentially fruitful measurement to pursue may be EEG. The utility of EEG in diagnosing ADHD is increasingly seen as feasible, with a 2014 study *Use of EEG to Diagnose ADHD*40concluding that the limiting factor is ECG analytics but progress in this area may legitimise EEG as a diagnostic tool. The necessary standards which must be met to create a diagnostic model are laid out in Boutros 201441. Since EEGs are multivariate timeseries, a multidimensional approach to a CNN would be required (if the CNN method is chosen). Such an approach has been successful24 in other studies, though not yet in EEG.

## Application to further biometrics:

In this study, the classes which we are trying to distinguish between are ADHD and non-ADHD. We are only looking at one measure: pupil diameter. Consider the techniques used more broadly. Multiple biometrics can be input simultaneously to train a scaled-up version of the same neural network, yielding a potentially maximally accurate classification. These methods may be used to engineer a more powerful combination of hardware and software with a high degree of accuracy in discriminating between classes. Ultimately with a large enough dataset it may be possible to move beyond simple classification and towards a scalar measure such as severity of illness or prognosis. For example, a correlating measured physiology with a score on the Wechsler Intelligence Scale for Children.

Consider a separate study in which the same data was gathered from other classes of individual instead (not ADHD). Possible options are those which have been already implicated in pupil differences: autism42,43, PTSD44, and Parkinsons45.

Outside of medicine, there are applications in fields like deception detection. Meta-analysis46 has firmly established that the pupil can act as a cue to deception47,48. It is trivial to train a neural network to recognize the signature pupil behaviour of a deceptive subject or deceptive answer by using the methods created in this study. More biometrics could potentially be gathered to further increase accuracy.

## The future of machine learning in medicine:

The rise of machine learning is having a vast trans-industry impact. Techniques were once only available to those with a high level of on-site computer power and PhD-level programming skills. Now, however, they are accessible to any healthcare professional with the motivation to learn.

Traditional inference statistics focus on the validation of hypotheses. Machine learning, however, seeks to use one set of observations to predict another49. Given that medicine is fundamentally a practical science, prediction is especially useful in our field. In fact, as machine learning is used more and more in medical science, a greater emphasis may be placed on whether results are *useful* in practice, as opposed to explanatory. This study seeks to justify its aims by invoking the possibility of a real world diagnostic model for ADHD that relies on neurophysiology – a model that might arise out of methods laid out in this paper.

# Bibliography

1. Wainstein G, Rojas-Líbano D, Crossley NA, Carrasco X, Aboitiz F, Ossandón T. Pupil Size Tracks Attentional Performance in Attention-Deficit/Hyperactivity Disorder. Sci Rep. 2017;

2. Wang Z, Oates T. Encoding Time Series as Images for Visual Inspection and Classification Using Tiled Convolutional Neural Networks. AAAI Work. 2015;

3. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. Psychol Bull. 1997;

4. Scahill L, Schwab-Stone M. Epidemiology of Adhd in School-Age Children. Child Adolesc Psychiatr Clin N Am. 2018;

5. Harpin VA. The effect of ADHD on the life of an individual, their family, and community from preschool to adult life. Arch Dis Child. 2005;

6. Kooij JJS, Bijlenga D, Salerno L, Jaeschke R, Bitter I, Balázs J, et al. Updated European Consensus Statement on diagnosis and treatment of adult ADHD. Eur Psychiatry. 2019;

7. McGoey KE, Eckert TL, DuPaul GJ. Early intervention for preschool-age children with ADHD: A literature review. J Emot Behav Disord. 2002;

8. Kevill J. ADHD: Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents. Yearb Neurol Neurosurg. 2012;

9. Arms E, Bickett J, Graf V. Gender bias and imbalance: Girls in US special education programmes. In: Gender and Education. 2008.

10. Hoff Sommers C. The War Against Boys: How Misguided Feminism Is Harming Our Young Men. Psychiatric Services. 2001.

11. Stevens J, Harman JS, Kelleher KJ. Race/Ethnicity and Insurance Status as Factors Associated with ADHD Treatment Patterns. J Child Adolesc Psychopharmacol. 2005;

12. Medina JA, Netto TLB, Muszkat M, Medina AC, Botter D, Orbetelli R, et al. Exercise impact on sustained attention of ADHD children, methylphenidate effects. ADHD Atten Deficit Hyperact Disord. 2010;

13. Wigal SB, Emmerson N, Gehricke JG, Galassetti P. Exercise: Applications to Childhood ADHD. J Atten Disord. 2013;

14. Manufacturers IPEA. 2018 Survey on Recess - Voice of Play.

15. VicGov. School Life [Internet]. Available from: https://www.study.vic.gov.au/en/living-in-victoria/school-life/Pages/default.aspx

16. Nakajima T, Sasaki T, Nakagome K, Takazawa S, Ikebuchi E, Ito Y, et al. Comparison of the effects of zolpidem and zopiclone on nocturnal sleep and sleep latency in the morning: A cross-over study in healthy young volunteers. Life Sci. 2000;

17. Hamed AM, Kauer AJ, Stevens HE. Why the diagnosis of attention deficit hyperactivity disorder matters. Frontiers in Psychiatry. 2015.

18. Fried M, Tsitsiashvili E, Bonneh YS, Sterkin A, Wygnanski-Jaffe T, Epstein T, et al. ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. Vision Res. 2014;

19. Reimer J, Froudarakis E, Cadwell CR, Yatsenko D, Denfield GH, Tolias AS. Pupil Fluctuations Track Fast Switching of Cortical States during Quiet Wakefulness. Neuron. 2014;

20. Nechushtai R, Elit M, Systems SM. United States Patent US 6,652,458 B2. 2001;1(12).

21. Pishyareh E, Tehrani-Doost M, Mahmoodi-Gharaie J, Khorrami A, Rahmdar SR. A comparative study of sustained attentional bias on emotional processing in adhd children to pictures with eye-tracking. Iran J Child Neurol. 2015;

22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016.

23. Berndt D, Clifford J. Using Dynamic Time Warping to Find Patterns in Time Series. In: KDD-94 Workshop on Knowledge Discovery in Databases. 1994.

24. Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov. 2017;

25. Lines J, Taylor S, Bagnall A. HIVE-COTE: The hierarchical vote collective of transformation-based ensembles for time series classification. In: Proceedings - IEEE International Conference on Data Mining, ICDM. 2017.

26. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov. 2019;

27. Dolcos F. Brain Systems Mediating Cognitive Interference by Emotional Distraction. J Neurosci. 2006;

28. Brocki KC, Randall KD, Bohlin G, Kerns KA. Working memory in school-aged children with attention-deficit/hyperactivity disorder combined type: Are deficits modality specific and are they independent of impaired inhibitory control? J Clin Exp Neuropsychol. 2008;

29. Rojas-Líbano D, Wainstein G, Carrasco X, Aboitiz F, Crossley N, Ossandón T. A pupil size, eye-tracking and neuropsychological dataset from ADHD children during a cognitive task. Sci data. 2019;

30. Wang Z, Oates T. Imaging time-series to improve classification and imputation. In: IJCAI International Joint Conference on Artificial Intelligence. 2015.

31. Shah. T. Train, Validation and Test Sets [Internet]. 2017 [cited 2019 Jul 8]. Available from: https://tarangshah.com/blog/2017-12-03/train-validation-and-test-sets/

32. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;

33. Felt BT, Biermann B, Christner JG, Kochhar P, Van Harrison R. Diagnosis and management of ADHD in children. Am Fam Physician. 2014;

34. Biederman J, Klein RG, Pine DS, Klein DF. Resolved: Mania is mistaken for ADHD in prepubertal children. J Am Acad Child Adolesc Psychiatry. 1998;

35. Sarampote CS, Efron LA, Robb AS, Pearl PL, Stein MA. Can Stimulant Rebound Mimic Pediatric Bipolar Disorder? J Child Adolesc Psychopharmacol. 2002;

36. Lecendreux M, Konofal E, Bouvard M, Falissard B, Mouren-Siméoni MC. Sleep and alertness in children with ADHD. J Child Psychol Psychiatry Allied Discip. 2000;

37. Oguiza I. Large benchmark Time Series Classification studies: summary [Internet]. Fast.ai Forums. 2018. Available from: https://forums.fast.ai/t/time-series-sequential-data-study-group/29686/66?u=michaelwoodburn

38. Munoz DP, Armstrong IT, Hampton KA, Moore KD. Altered Control of Visual Fixation and Saccadic Eye Movements in Attention-Deficit Hyperactivity Disorder. J Neurophysiol. 2006;

39. Bucci MP, Stordeur C, Septier M, Acquaviva E, Peyre H, Delorme R. Oculomotor Abnormalities in Children with Attention-Deficit/Hyperactivity Disorder Are Improved by Methylphenidate. J Child Adolesc Psychopharmacol. 2017;

40. Lenartowicz A, Loo SK. Use of EEG to Diagnose ADHD. Current Psychiatry Reports. 2014.

41. Boutros N, Fraenkel L, Feingold A. A Four-Step Approach for Developing Diagnostic Tests in Psychiatry: EEG in ADHD as a Test Case. J Neuropsychiatry Clin Neurosci. 2014;

42. Dinalankara DMR, Miles JH, Nicole Takahashi T, Yao G. Atypical pupillary light reflex in 2–6-year-old children with autism spectrum disorders. Autism Res. 2017;

43. Daluwatte C, Miles JH, Christ SE, Beversdorf DQ, Takahashi TN, Yao G. Atypical pupillary light reflex and heart rate variability in children with autism spectrum disorder. J Autism Dev Disord. 2013;

44. Cascardi M, Armstrong D, Chung L, Paré D. Pupil Response to Threat in Trauma-Exposed Individuals With or Without PTSD. J Trauma Stress. 2015;

45. Giza E, Fotiou D, Bostantjopoulou S, Katsarou Z, Karlovasitou A. Pupil light reflex in Parkinson’s disease: Evaluation with pupillometry. Int J Neurosci. 2011;

46. DePaulo BM, Malone BE, Lindsay JJ, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. Psychological Bulletin. 2003.

47. DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper H. Cues to deception. - ProQuest. Psychol Bull. 2003;

48. Vrij A. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection. In: Interpersonal Communication. 2014.

49. G S. To Explain or to Predict? [Internet]. [cited 2019 Jun 30]. Available from: http://projecteuclid.org/euclid.ss/1294167961

## List of abbreviations used:

ADHD: Attention Deficit Hyperactivity Disorder  
CNN: Convolutional Neural Network  
GASF: Gramian Angular Summation Field

# Acknowledgements

We would like to acknowledge the Pontificia Universidad Católica de Chile for facilitating the experiment and the original authors for co-authoring this paper. Furthermore, our gratitude goes to Francisco Ingham and Jeremy Howard for developing the software and tutorials for fast.ai.

# Tables

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | Class | Ratio: ADHD:All Predictions | GASF count |
| 1 | off-ADHD | 0.71 | 129 |
| 2 | off-ADHD | 0.65 | 110 |
| 3 | off-ADHD | 0.80 | 102 |
| 4 | off-ADHD | 0.62 | 87 |
| 5 | off-ADHD | 0.30 | 144 |
| 6 | off-ADHD | 0.76 | 89 |
| 7 | off-ADHD | 0.49 | 118 |
| 8 | off-ADHD | 0.53 | 103 |
| 9 | off-ADHD | 0.50 | 127 |
| 10 | off-ADHD | 0.52 | 99 |
| 11 | off-ADHD | 0.51 | 105 |
| 12 | off-ADHD | 0.82 | 131 |
| 13 | off-ADHD | 0.81 | 81 |
| 14 | off-ADHD | 0.68 | 25 |
| 15 | off-ADHD | 0.48 | 141 |
| 16 | off-ADHD | 0.65 | 136 |
| 17 | off-ADHD | 0.61 | 123 |
| 18 | off-ADHD | 0.54 | 128 |
| 19 | off-ADHD | 0.38 | 131 |
| 20 | off-ADHD | 0.62 | 143 |
| 21 | off-ADHD | 0.81 | 101 |
| 22 | off-ADHD | 0.58 | 81 |
| 23 | off-ADHD | 0.74 | 117 |
| 24 | off-ADHD | 0.63 | 128 |
| 25 | off-ADHD | 0.55 | 130 |
| 26 | off-ADHD | 0.54 | 141 |
| 27 | off-ADHD | 0.73 | 49 |
| 28 | off-ADHD | 0.81 | 134 |
| 29 | Ctrl | 0.56 | 124 |
| 30 | Ctrl | 0.56 | 137 |
| 31 | Ctrl | 0.53 | 134 |
| 32 | Ctrl | 0.48 | 139 |
| 33 | Ctrl | 0.51 | 144 |
| 34 | Ctrl | 0.51 | 140 |
| 35 | Ctrl | 0.80 | 5 |
| 36 | Ctrl | 0.57 | 138 |
| 37 | Ctrl | 0.45 | 138 |
| 38 | Ctrl | 0.31 | 139 |
| 39 | Ctrl | 0.42 | 146 |
| 40 | Ctrl | 0.43 | 129 |
| 41 | Ctrl | 0.43 | 143 |
| 42 | Ctrl | 0.45 | 143 |
| 43 | Ctrl | 0.29 | 130 |
| 44 | Ctrl | 0.43 | 132 |
| 45 | Ctrl | 0.00 | 1 |
| 46 | Ctrl | 0.35 | 140 |
| 47 | Ctrl | 0.66 | 125 |
| 48 | Ctrl | 0.37 | 143 |
| 49 | Ctrl | 0.52 | 132 |
| 50 | Ctrl | 0.45 | 138 |