**ADHD IN CHILDREN CAN BE DETECTED BY MEASURING THEIR PUPIL RESPONSE TO WORKING MEMORY TASKS WITH A MACHINE LEARNING MODEL.**

## Title page

Recognizing ADHD characteristic pupil dilation: a successful machine learning approach using deep convolutional neural networks and Gramian angular fields

## Abstract

Attention Deficit/Hyperactive Disorder (ADHD) is diagnosed in children based on reports from parents, teachers, and close contacts. There is no explicit neurological or biological test within the diagnostic criteria. However, the same neurological circuits that cause ADHD may effect physiological markers in a measurable way. Differences in pupil behaviour have been identified in ADHD1 as well as in other psychiatric conditions such as autism2,3, PTSD4. 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 time series within the data were transformed using the Gramian Angular Summation Field5 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 validity. The methods within this study leverage the sophistication and power of machine learning to classify complex time series and are highly applicable to a range of other conditions.

## Keywords

Pupil dilation ADHD children deception biometrics psychiatry polygraph machine learning

## Introduction

***Diagnosing ADHD***

Attention-deficit/hyperactivity disorder (ADHD) is a chronic psychiatric condition characterised by impaired behavioural inhibition6. The prevalence of ADHD in children is likely to lie between 5% and 10%7. It correlates with many negative academic and social outcomes8 which begin early in childhood and tend to persist into adulthood9. Early intervention and diagnosis mitigates these outcomes and influences disease progression10. Clinical practice guidelines11 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 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

***Limitations of current practice***

Diagnosis of ADHD 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 as gender12,13 or race14. Moreover, environmental effects may influence the severity of ADHD presentation; exercise alleviates hyperactivity15,16 and the time provided for exercise during school varies significantly on both national17,18 and local19 levels. These variables contribute to over-diagnosis or under-diagnosis in certain subpopulations20.

Behavioural reports are the primary criteria for diagnosis. There is no biological test at present which is reliable enough to merit inclusion despite the fact that ADHD is 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 current efforts within limited budget.

***Is 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 under some metrics1,21 in children with ADHD. The  iris dilator muscleis mediated by the locus-coeruleus (LC) which governs norepinephrine (NE) activity in the brain22. Given that the aetiology of ADHD is associated with catecholaminergic transmission defecits, 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 behaviour?***

This study seeks to identify that characteristic pupil response and quantify its utility as a predictor of ADHD. There is indirect evidence for the utility of pupil dilation as a diagnostic indicator of ADHD1. Gaze direction21,23,24 has also previously come to the fore as a useful biometric. 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 network25 program which is highly sensitive to differences between classes comprised of complex or noisy datasets.

## Methods

***Study design***:

This was an analysis of a dataset [published in April 2019](https://www.nature.com/articles/s41597-019-0037-2#article-info): *A pupil size, eye-tracking and neuropsychological dataset from ADHD children during a cognitive task*26*.* That dataset was first analysed by the originator of the experiment, Ossandon T, and the findings published in *Pupil Size Tracks Attentional Performance In Attention-Deficit/ Hyperactivity Disorder*1*.* The primary finding in that study was that pupil size correlated with the subjects’ performance on the tasks and reaction time. Moreover, there were obvious differences in the maximum pupil diameter after probe onset between ADHD and control groups, which reinforces the likelihood for a diagnostic model using that metric.

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 Disorder1

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 make use of that component. Furthermore, by using a neural network, we leave open the possibility that there are unknown differences in pupil behaviour which were embedded in the pupil-time series and are yet to be formally recognized.

***Advantages and disadvantages in neural networks***

The deep convolutional neural network functions as a black box. Although its accuracy can be thoroughly assessed, it is not mathematically correct to draw inferences from the model regarding which features of the pupil-time series specifically it is identifying. 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.

***Participants****:*

50 children age 10-12 were recruited. 28 had been diagnosed with ADHD, and 22 were the non-ADHD control group. The groups were sufficiently similar in age, IQ, and educational level. 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 McCarthy done on healthy individuals. In other studies27 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.

### The Sternberg-type working memory task

A close up of a piece of paper

Description automatically generated

Figure 2: A single trial of one subject. 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 task26

***Pupil Measurement****:*

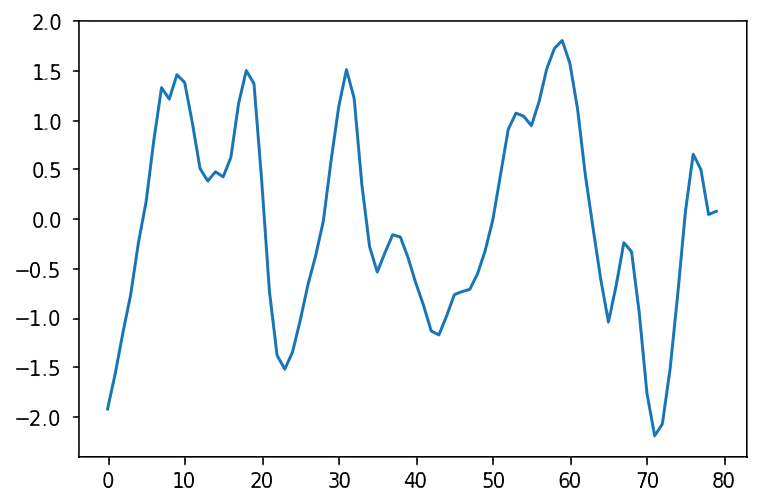
Subject’s heads were fixed by a forehead and chin rest (SR Research LTD) 30cm in front of a computer screen. The screen’s luminosity varied on the order of 1 lux. The narrow range of variation in luminosity controls for luminosity-mediated pupil diameter changes. The Eyelink 1000 (SR Research Ltd, Mississauga, 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.

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

***Pupil-time series processing (original dataset):***

*The original dataset*26 *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.



### A randomly selected pupil-time series

Pupil Size (Z Score)

Time (0.1s)

Figure This is the pupil-time series of a randomly selected trial (subject 10 trial 106)

***Pupil-time series processing (extended):***

*(in addition to those above):*

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. Time series were pared down to reduce the resolution of the data. For each time series, the sampling rate was approximated by a factor of 100. 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.
4. Each time series was used to generate an 80x80 Gramian Angular Summation Field (GASF)5,28. The python library *pyts* was used to compute the values of each cell of the field. The reasons behind this representation of time series merit exploration.

The pupil-time series falls under the broader classification of univariate time series (i.e there is one variable which changes dependent on time). Attempting to classify univariate or even mutlivariate time series as belonging to a certain class is known as Time Series Classification (TSC). Until 2015 the premier algorithm used to do so was Dynamic Time Warping29 (DTW). This algorithm demonstrated good performance where there was strong similarity amongst univariate time series of the same class. However, with the introduction of time series of greater complexity and more abstract relationships within classes the DTW algorithm is no longer considered state of the art30.

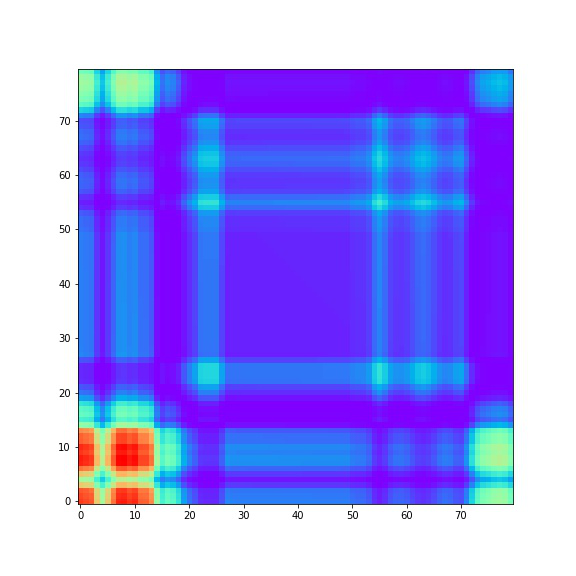
Two competing algorithms have superseded it.

The first is The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification31 (HIVE-COAT). As of 2018 this algorithm achieved the highest accuracy. However, it is also one of the most computationally intensive algorithms and carries with it many impracticalities that make it incompatible with certain datasets.

The second algorithm to demonstrate a high level of accuracy is Resnet25, a deep convolutional neural network which achieved victory in the 2017 Kaggle ImageNet competition. It has become apparent that although Resnet was developed to recognize the presence of real objects in photographic images, it is also useful at recognizing characteristic traits in data which has been encoded to an image. In order to make use of that serendipitous benefit, the time series in our dataset were encoded to GASF.

1. A figure was created with the *matplotlib.pyplot* library to represent the GASF pictorially. The ‘rainbow’ colormap was used as it is one of many colormaps which are reasonably interpretable by humans and have been used to train convolutional neural networks (CNN) to a high level of accuracy32. The resultant directory contained 5873 GASFs each corresponding to a single timeseries measured during one task done by a subject.

### The corresponding GASF for the above pupil-time series



Time (0.1s)

Time (0.1s)

Figure The colour of the cells is proportional to the value of the GASF at that coordinate



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

***Labelling of GASFs***

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 DataBunch***: **Organizing Data for Machine Learning.**

Subjects were divided into training and validation groups. Out of 50 subjects, 6 were randomly assigned to the validation set. The randomization process was designed such that 3 were selected from the ADHD group and 3 were selected from the Control group. The selected subjects were 23, 13, 24, 31, 42, 38. Therefore, the training set consisted of the remaining 46 subjects. On a programming level, the sets comprised of the labelled trials of the set’s respective subjects. In this permutation of subjects, 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, as this size is preferred by the CNN. 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.

***Choosing a learner:***

The resnet50 convolutional neural network was chosen due to its demonstrable effectiveness30,32 in classifying univariate series encoded as GASFs. It 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.*

The data of subjects in the validation set were specifically analysed in order to assess the fraction of each of their 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 was done to explore the accuracy on subject trials that the model had neither seen in the training set, nor received feedback on for its predictions towards the validation set. In summary, 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.

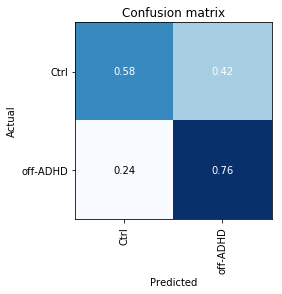
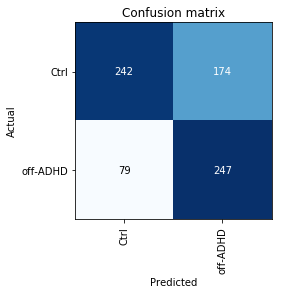
One subject’s trials were selected and removed from the databunch. This was the test subject. Remaining trials were randomly distributed to either the training set or the validation set. The CNN was trained and validated from scratch on these trials. It was then tested on the trials of the test subject, 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 databunch (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 databunch, each model is sufficiently similar in that they have all trained or validated 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 trained or validated 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:

## Results and discussion

Figure These two confusion matrices represent the same results. However, the left has been normalised to better represent accuracy. "Ctrl" denotes trials belonging to the healthy control group. "off-ADHD" denotes trials belonging to the ADHD group.



### Classification accuracy per trial

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.

If ones takes the classification of a trials as ADHD as the positive result, 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 pupil behaviour in 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.

### ROC Curve for trial classification

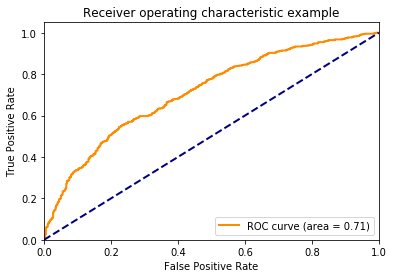


Figure This ROC curve is derived from the 742 trials GASFs in the validation set and the model's predicted probability that they belong to the ADHD class.

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 efficacy in identifying ADHD-characteristic trials. Given that the values are not derived from any type of parameterized statistic model, it is difficult to judge the statistical significance and p-value of this result. It satisfies the basic requirement that ROC curves must be derived from more than 100 observations.33

## 3. Diagnostic accuracy on validation group

Figure The fraction of trials (considering only the trials which were available as GASFs) predicted to belong to the ADHD class. 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, the majority of trials were correctly classified.

The trials of subjects 42 and 31 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). When accounting for error of one standard deviation (under Bernoulli assumptions, using the plugin method) we are unable to reject the null for those subjects, where the null hypothesis states that the model was unable to classify trials for that subject above the level of pure chance.

## 4. Diagnostic accuracy on test subject

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. This is consistent with the machine learning principle that models will tend to perform less accurately on the test set than the validation set; when those subjects were removed from the validation and training sets and placed in the test set, predictions of their trials decreased in accuracy.

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 off the trials in aggregate. This indicates that in reality, it is judicious to thoroughly assess a subject across many trials and to look at the overall proportion of trials classified as ADHD.

## 4. Diagnostic accuracy on test group

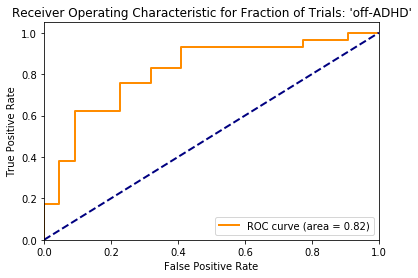


Figure The corresponding ROC curve for the graph Fraction of trials classified as ADHD for each subject in the test set

## Conclusions

***Internal Validity***

There **are** characteristic differences between pupil behaviour in children with ADHD and healthy controls. These differences are sufficient 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’, diagnostic accuracy increases. Accuracy is likely to improve with the number of trials a subject performs. We were fortunate to have a very high number of trials per subject in this study. Taking the proportion of trials which were classified as off-ADHD was a pragmatic choice but fundamentally arbitrary. Future analysis of the data may uncover another method to process the predictions of the 160 trials of a single subject which will yield greater diagnostic accuracy. For instance, there may be one single pathognomonic pupil behaviour that reveals itself in one out of the one-hundred-and-sixty trials which can indicate subject ADHD with great accuracy. A learner/model which comfortably assesses all 160 trials at once and returns a single prediction per subject could feasibly uncover and make use of such a pathognomonic trait. The engineering of such a learner 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 code used to process the pupil data for this analysis has been made available at \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

***Improving the Model:***  
  
Further data collection, incorporating results from other similar experiments, is likely to improve the external validity of this model. Data gathered from a conscientiously standardized set of experiments will inevitably vary in certain aspects; systematic error can be introduced via a number of minor experimental differences 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 be trained to recognize different experimental setups and make correct predictions regardless.

***By 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 implausible that the present advantage which it offers should extend much further into the future. More than likely resnet50 will be superseded by a targeted time series classification model.

***By changing the data to image encoding***

Whilst GASF is the standard practice encoding technique for univariate time series, there is increasing evidence that a combination of two different encoding techniques yields superior accuracy after training34. This may be due to the capacity of encoders to emphasize to the CNN specific attributes of the underlying data. 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 *generalizable* to 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 dataset26 in raw coordinate form but was not included in this analysis. There is some evidence that ADHD subjects are less able to suppress saccades21,35,36. 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 dataset.

A more fruitful measurement to pursue may be EEG. The utility of EEG in diagnosing ADHD is increasingly obvious, with a 2014 study *Use of EEG to Diagnose ADHD*37concluding that development in ECG analytics may progress EEG towards clinical deployment. Since EEGs are are multivariate timeseries, a multidimensional approach to resnet50 is required. Such an approach has been successful30 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 combinations of hardware and software4 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: autism2,3, PTSD4, and Parkinsons38. Many possibilities are likely to exist beyond these.

Outside of medicine, there are applications in fields like deception detection. Meta-analysis39 has firmly established that the pupil can act as a cue to deception40,41. It is trivial to train neural network to recognize the signature pupil behaviour of a deceptive subject or deceptive answer. 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 another42. 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.

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. doi:10.1038/s41598-017-08246-w

2. 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. doi:10.1002/aur.1745

3. 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. doi:10.1007/s10803-012-1741-3

4. Cascardi M, Armstrong D, Chung L, Paré D. Pupil Response to Threat in Trauma-Exposed Individuals With or Without PTSD. *J Trauma Stress*. 2015. doi:10.1002/jts.22022

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

6. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychol Bull*. 1997. doi:10.1037/0033-2909.121.1.65

7. Scahill L, Schwab-Stone M. Epidemiology of Adhd in School-Age Children. *Child Adolesc Psychiatr Clin N Am*. 2018. doi:10.1016/s1056-4993(18)30106-8

8. 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. doi:10.1136/adc.2004.059006

9. Kooij JJS, Bijlenga D, Salerno L, et al. Updated European Consensus Statement on diagnosis and treatment of adult ADHD. *Eur Psychiatry*. 2019. doi:10.1016/j.eurpsy.2018.11.001

10. McGoey KE, Eckert TL, DuPaul GJ. Early intervention for preschool-age children with ADHD: A literature review. *J Emot Behav Disord*. 2002. doi:10.1177/106342660201000103

11. 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. doi:10.1016/j.yneu.2012.05.077

12. Arms E, Bickett J, Graf V. Gender bias and imbalance: Girls in US special education programmes. In: *Gender and Education*. ; 2008. doi:10.1080/09540250802190180

13. Hoff Sommers C. *The War Against Boys: How Misguided Feminism Is Harming Our Young Men*.; 2001. doi:10.1176/appi.ps.52.9.1266

14. Stevens J, Harman JS, Kelleher KJ. Race/Ethnicity and Insurance Status as Factors Associated with ADHD Treatment Patterns. *J Child Adolesc Psychopharmacol*. 2005. doi:10.1089/cap.2005.15.88

15. Medina JA, Netto TLB, Muszkat M, et al. Exercise impact on sustained attention of ADHD children, methylphenidate effects. *ADHD Atten Deficit Hyperact Disord*. 2010. doi:10.1007/s12402-009-0018-y

16. Wigal SB, Emmerson N, Gehricke JG, Galassetti P. Exercise: Applications to Childhood ADHD. *J Atten Disord*. 2013. doi:10.1177/1087054712454192

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

18. VicGov. School Life. https://www.study.vic.gov.au/en/living-in-victoria/school-life/Pages/default.aspx.

19. Nakajima T, Sasaki T, Nakagome K, 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. doi:10.1016/S0024-3205(00)00603-2

20. Hamed AM, Kauer AJ, Stevens HE. Why the diagnosis of attention deficit hyperactivity disorder matters. *Front Psychiatry*. 2015. doi:10.3389/fpsyt.2015.00168

21. Fried M, Tsitsiashvili E, Bonneh YS, et al. ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Res*. 2014. doi:10.1016/j.visres.2014.05.004

22. 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. doi:10.1016/j.neuron.2014.09.033

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

24. 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.

25. 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. doi:10.1109/CVPR.2016.90

26. 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. doi:10.1038/s41597-019-0037-2

27. 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. doi:10.1080/13803390701754720

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

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

30. 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. doi:10.1007/s10618-016-0483-9

31. 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. doi:10.1109/ICDM.2016.74

32. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. *Data Min Knowl Discov*. 2019. doi:10.1007/s10618-019-00619-1

33. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978. doi:10.1016/S0001-2998(78)80014-2

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

35. 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. doi:10.1152/jn.00192.2003

36. 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. doi:10.1089/cap.2016.0162

37. Lenartowicz A, Loo SK. Use of EEG to Diagnose ADHD. *Curr Psychiatry Rep*. 2014. doi:10.1007/s11920-014-0498-0

38. Giza E, Fotiou D, Bostantjopoulou S, Katsarou Z, Karlovasitou A. Pupil light reflex in Parkinson’s disease: Evaluation with pupillometry. *Int J Neurosci*. 2011. doi:10.3109/00207454.2010.526730

39. DePaulo BM, Malone BE, Lindsay JJ, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. *Psychol Bull*. 2003. doi:10.1037/0033-2909.129.1.74

40. DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper H. Cues to deception. - ProQuest. *Psychol Bull*. 2003. doi:10.1037//0033-2909.129.1.74

41. Vrij A. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection. In: *Interpersonal Communication*. ; 2014. doi:10.1515/9783110276794.321

42. G S. To Explain or to Predict? http://projecteuclid.org/euclid.ss/1294167961. Accessed June 30, 2019.

## List of abbreviations used (if any)

## Acknowledgements

## References in either the Harvard or Vancouver style

## Illustrations and figures (if any)

## Tables (if any)

## 5000 words ± 10% (ie between 4500-5500 words); this excludes references, content pages, figures and tables

## Further details in “Instructions for Students”