

32 Attention deficit hyperactivity disorder (ADHD) is a heterogeneous neurodevelopmental disorder that affects 5%
33 of the pediatric and adult population worldwide. The diagnosis remains essentially clinical, based on history and
34 exam, with no available biomarkers. In this paper, we describe a deep convolutional neural network (DCNN) for
35 ADHD classification derived from the time-frequency decomposition of electroencephalography data (EEG),
36 particularly of event-related potentials (ERP) during the Flanker Task collected from 20 ADHD adult patients and
37 20 healthy controls (HC). The model reaches a classification accuracy of 88%, superior to resting state EEG
38 spectrograms and with the key advantage, compared with other machine learning approaches, of avoiding the need
39 for manual selection of EEG spectral or channel features. Finally, through the use of feature visualization
40 techniques, we show that the main features exciting the DCNN nodes are a decreased power in the alpha band and
41 an increased power in the delta-theta band around 100ms for ADHD patients compared to HC, suggestive of
42 attentional and inhibition deficits, which have been previously suggested as pathophysiological signatures of
43 ADHD. While confirmation with larger clinical samples is necessary, these results highlight the potential of this
44 methodology to develop CNS biomarkers of practical clinical utility.

45

46 **Introduction**

47

48 Attention-deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder characterized by
49 deficits in attention, impulsivity (motor and non-motor) and executive dysfunction. It is associated with high
50 morbidity and disability^{1,2}, and affects up to 5% of adults worldwide³⁻⁵. The diagnosis of ADHD remains
51 essentially clinical, based on history and exam. It can be supported by neuropsychological assessments, but given
52 the heterogeneous cognitive profiles in patients with ADHD, these provide a supportive, not fully diagnostic,
53 function. Significantly, there are many different conditions that present with disordered attention, impulsivity and
54 dysexecutive syndromes, and the range of normal cognitive profiles with variable strengths and weaknesses in
55 these domains is wide, often complicating the differential diagnosis. Hence, a biomarker to reduce the inherent
56 uncertainty of clinical diagnosis would be of great value.

57

58 Electroencephalographic (EEG) signals contain rich information associated with functional dynamics in the brain.
59 The use of EEG in ADHD began more than 75 years ago with Jasper et al.⁶ reporting an increase in the EEG
60 power of low frequencies in fronto-central areas. Since then, human electrophysiological studies using EEG
61 spectral analyses and event-related potentials (ERPs) have established relevant signatures of executive dysfunction
62 in ADHD⁷. In contrast to spontaneous EEG, ERPs reflect changes in the electrical activity of the brain that are
63 time-locked to the occurrence of a specific event, that is, a response to a discrete external stimulus or an internal
64 mental process⁸. ERPs also provide non-invasive neurophysiological measurements with high temporal resolution,
65 allowing to assess dysfunctional brain dynamics, including cognitive processes that may not be apparent at the
66 behavioral level^{9,10}. Indeed, ERPs are commonly used clinically in neurophysiological diagnostic units to support
67 the assessment of neuropsychiatric disorders (e.g., multiple sclerosis¹¹) and sensory disorders (e.g., screening of
68 neonates for hearing impairments¹²).

69

70 Artificial neural networks (ANNs) have recently become a promising application of artificial intelligence (AI) in
71 healthcare¹³. Machine learning, a subtype of AI, and deep learning, a specialized sub-field of machine learning,
72 have been increasingly used in clinical research with promising results. Machine learning can be described as the
73 practice of using algorithms to train a system by using large amounts of data, with the goal of giving it the ability
74 to learn how to perform a specific task, and then make an accurate classification or prediction. Deep learning is a
75 subset of machine learning algorithms that break down the tasks in smaller units (neural networks, NNs) often

76 providing higher levels of accuracy.

77

78 NNs are characterized by their network architecture, defined by the anatomical layout of its connected processing
79 units, the artificial “neurons”, according to a loss or optimization function that specifies the overall goal of the
80 learning process. Connections are “trained”, or taught how to do the desired task, by using a training algorithm
81 that iteratively changes parameters of the NN such that the target function is ultimately optimized based on the
82 inputs the NN receives. There are different types of NNs with different designs and architectures derived from
83 different principles, or conceived for different purposes. The most basic ones are the feed-forward NNs (FNNs), in
84 which activity is propagated unidirectionally layer-by-layer from the input up to the output stage, with no feedback
85 connections within or between layers. We have previously used a specific type of FFNs (feed-forward
86 autoencoders) for the analysis of EEG data with promising outcomes¹⁴. Recurrent Neural Networks (RNN) are
87 another type of NN that, unlike FFNs, are based on architectures with feedback (“recurrent”) connections within or
88 between layers. In related work, we used Echo State Networks (ESNs), a particular type of RNN, to classify
89 Parkinson patients from HC using EEG time-frequency decompositions¹⁵ with successful results. The main
90 limitation of RNNs is, however, their computational cost¹⁶. In addition, one of the main critics to deep NN is their
91 “black-box” nature, i.e., the difficulty in tracing a prediction back to which features are important and
92 understanding how the network reached the final output, which will be later addressed in this study.

93

94 Previous studies have successfully classified ADHD patients from HC using machine learning techniques with
95 accuracies of more than 90%¹⁷⁻²³, but the selection of disease-characterizing features from EEG was done
96 manually after an extensive search in the frequency or time domain. However, EEG signals exhibit non-linear
97 dynamics (chaotic signals that do not behave linearly and cannot be represented as combination of basic sub-
98 signals) and non-stationarity across temporal scales (signals with a mean and variance that do not stay constant
99 over time) that cannot be studied properly using classical machine learning approaches. There is a need for tools
100 capable of capturing the rich spatiotemporal hierarchical structures hidden in these signals. In a previous study²⁴,
101 we trained a machine learning system with pre-defined complexity metrics of time-frequency decompositions of
102 EEG data that showed statistically significant differences between REM Sleep Behavior Disorder (RBD) patients
103 and HC, indicating that such metrics may be useful for classification or scoring. While this approach is useful in
104 several domains, it would be advantageous to use methods where the relevant features are found directly by the
105 algorithms instead of pre-defining them manually.

106

107 With the goal of building a discrimination system that can classify ADHD patients from HCs, here we explore a
108 deep learning approach inspired by recent successes in image classification using Deep Convolutional Neural
109 Networks (DCNNs), a particular type of NN designed to exploit compositional and translationally invariant
110 features in the data that are present in EEG, i.e., features that are recognizable even if their appearance varies in
111 some way¹⁶. These networks were originally developed to deal with image data (2D arrays) from different
112 channels or audio data²⁵, and more recently, EEG data^{26,27}. Similarly, here we train a DCNN with multi-channel
113 two-dimensional time-frequency maps (spectrograms or 2D time-frequency maps), representing EEG spectral
114 dynamics as images with the equivalent image depth provided by multiple EEG channels. These networks treat
115 EEG-channel data as an audio file, and our approach mimics similar uses of deep networks in that domain.
116 Specifically, we use a similar strategy as the one presented by Ruffini et al.²⁸, but instead of using spontaneous
117 EEG spectrograms, we use ERP spectrograms (also called Event-Related Spectral perturbation, ERSP) recorded
118 during a Flanker-Eriksen Task (EFT), a well-established experimental task to assess sustained attention, conflict
119 monitoring and response inhibition. Our assumption is that relevant qualities of ERP data are contained in
120 compositional features embedded in this time-frequency representation. Particularly, we expect that DCNNs may
121 be able to efficiently learn to identify features in the time-frequency domain associated to event-related bursting
122 across frequency bands that may help separate classes, similar to what is known as “bump analysis”²⁹. For
123 comparison purposes, we also trained a RNN based on Long Short-Term Memory (LSTM) networks, which can
124 learn long sequences of data but require higher computational demands, and a Shallow Neural Network (SNN) as
125 a baseline, a more basic type of network with only one layer. We also compared the performance of the ERSP data
126 with a dataset of spontaneous EEG data recorded while the participants were at resting state. Lastly, we propose
127 the utilization of deep learning visualization techniques for the mechanistic interpretation of results, particularly
128 the method popularly known as DeepDream³⁰. This is important to identify pathophysiological features driving the
129 translational and clinical value of the application, and for the optimized further development and acceptance of
130 such techniques in the clinical domain, where black-box approaches have been extensively criticized.

131

132 **Methods**

133 **Participants**

134 A total of 40 participants including 20 healthy adults (10 males, 10 females) and 20 ADHD adult patients (10
135 males, 10 females) participated in the present study (Table 1). The inclusion criteria for ADHD patients consisted

136 of a diagnosis of ADHD made by a board-certified clinician according to the Diagnostic and Statistical Manual of
137 Mental Disorders, Fifth Edition (DSM-5)³¹. Symptom profiles and severity were assessed with the Adult ADHD
138 Self-Report Scale (ASRS-v1.1)³². Patients were either off stimulant medications or, if undergoing treatment with
139 stimulants, were asked to discontinue two days prior to the experiment, under a physician-guided protocol, and
140 allowed to resume afterwards. Psychiatric comorbidities were allowed as long as ADHD was the primary
141 diagnosis. Psychosis, bipolar disorder, substance use disorder and neurological conditions were exclusion criteria.
142 Healthy participants were included if they did not have any psychiatric or neurologic condition and were not
143 taking any psychoactive medications. All participants gave informed and written consent for participation. The
144 study was approved by the Partners HealthCare System's Institutional Review Board and all experiments were
145 performed in accordance with relevant guidelines and regulations at Massachusetts General Hospital.

Table 1. Participant characteristics

	ADHD (n=20)	HC (n=20)	Significance
Demographic	mean (SD)*	mean (SD)*	p value (T test)
Age	43.85 (14.78)	29.90 (10.77)	0.0006
Females	10 (50%)	10 (50%)	0.5
Baseline Scores			
ASRS	62.6 (9.17)	36.47 (11.33)	<0.0001
Current medications – N (%)			
No medication	11 (55%)		
Adderall	2 (15%)		
Vyvanse	2 (10%)		
Concerta	1 (5%)		
Verapamil	1 (5%)		
Aspirin	1 (5%)		
Levothyroxine	1 (5%)		
Modafinil	1 (5%)		

146 Abbreviations. SD: Standard Deviation

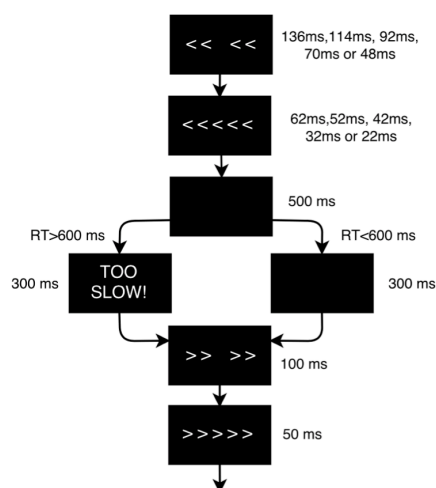
147 (*) All figures are mean (Standard Deviation) unless otherwise specified.

148

149 Experimental Task: Eriksen-Flanker task (EFT)

150 Each patient underwent three identical experimental sessions separated by 1-2 weeks in which they performed the
151 Eriksen-Flanker task (EFT) (Figure 1) while EEG data was recorded. The EFT is a classic behavioral paradigm in
152 which subjects must attend and respond to the direction of a central arrow that is surrounded (“flanked”) by
153 distracting stimuli. The flanking arrows can either have the same (congruent trials) or opposing (incongruent trials)
154 orientation as the central one. Participants are instructed to press the left or right arrow buttons in a keyboard
155 following the direction of the central arrow, ignoring the flankers. In this study there were a total of 140 trials, and
156 each subject had a different, fully random sequence of congruent and incongruent trials, with 2 congruent trials for

157 each incongruent trial, in order to build a tendency towards the prepotent congruent responses and thus increase
158 the difficulty of conflict detection in incongruent trials. Only incongruent trials were used for classification
159 purposes, as they are the ones that most elicit the conflict-related ERP components that characterize the executive
160 function subtasks of selective attention, inhibition and cognitive control³³, primarily impaired in ADHD. The
161 accuracy (percentage of correct/incorrect responses) and the reaction time (RT) were measured for each trial,
162 while also recording EEG data during the task. RT of single trials was introduced into a Generalized Linear Model
163 with Mixed Effects (GLMM) with a Gamma distribution, with Group as a fixed factor (ADHD/HC) and Subject
164 ID as a random intercept. Accuracy was also modeled using a generalized logistic regression model with mixed
165 effects and a binomial distribution.
166



167
168 **Figure 1. Flanker task design scheme.** The flanker arrows were first presented alone for duration of 136ms, 114ms, 92ms,
169 70ms or 48ms depending on the baseline performance of each subject, and were then joined by the target arrow for 62ms,
170 52ms, 42ms, 32ms or 22ms, respectively (the duration of the stimuli was adjusted to the psychometric spot in which each
171 subject reached a performance of 80-85%). Stimulus presentation was followed by a black screen for 500 ms. The time-
172 window for participants' response was 600 ms after target onset. If the participant did not respond within the response window,
173 a screen reading 'TOO SLOW!' was presented for 300 ms. Participants were told that if they saw this screen, they should speed
174 up. If a response was made before the deadline, the 'TOO SLOW!' screen was omitted and the black screen remained on screen
175 for the 300 ms interval. Finally, each trial ended with presentation of the fixation cross for an additional randomly chosen
176 duration (200, 300 or 400 ms) in order to avoid any habituation or expectation by the subject. Thus, trial duration varied
177 between 1070–1400 ms.

178

179 EEG data acquisition and preprocessing

180 EEG was recorded with the Starstim system (Neuroelectronics, Cambridge, MA, USA) from 7 positions covering the
181 primary hubs of the fronto-parietal executive control network (Fp1, Fp2, F3, Fz, F4, P3 and P4) with 3.14cm^2
182 Ag/AgCl electrodes and digitalized with 24-bit resolution at a sampling frequency of 500 samples/second. EEG
183 data was referenced to the right mastoid. Independent component analysis (ICA) was utilized to identify and
184 remove activity associated with blinks, eye movements, and other artifacts. Data was filtered from 1Hz to 20Hz to
185 remove non-neural physiological activity (skin/sweat potentials) and noise from electrical outlets. Trials were
186 epoched within a time frame of 200ms before and 800ms after the stimulus onset. The mean of the pre-stimulus
187 baseline [-200,0]ms was then subtracted from the entire ERP waveform for each epoch to eliminate any voltage
188 offset.

189
190 To create the ERP spectrograms (or ERSP), the Wavelet transform was applied to each single trial as implemented
191 in EEGlab's *newtimef* function, with 1 wavelet cycle at the lowest frequency to 10 cycles at the highest, leading to
192 22 frequency bins logarithmically spaced in the [3, 20]Hz range and 20 linear time bins in the [0, 800]ms range,
193 where 0 represents the onset of the target stimuli in incongruent trials. The input data frames were thus
194 multidimensional arrays of the form [22 Frequency bins] x [20 Time bins] x [7 channels], with 3 minutes of data
195 per subject approximately. For comparison purposes, we also processed with the same parameters a dataset of
196 spontaneous EEG data recorded while the same subjects and ADHD patients were resting with eyes closed (no
197 cognitive task performed).

198

199 Neural network architecture

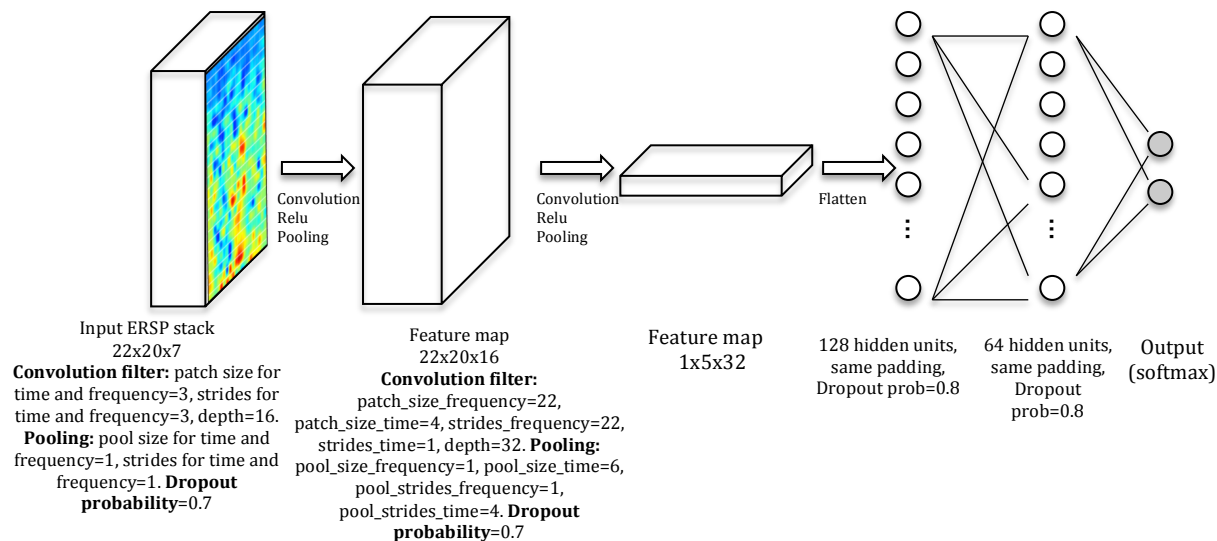
200 The DCNN, implemented in Tensorflow³⁴, is a relatively simple four layer convolutional network, as shown in
201 Figure 2a. In order to avoid overfitting the data (i.e., overtraining the system to the extent that it negatively
202 impacts the performance of the model on new data), we used the so-called "Dropout" method, a regularization
203 technique in which that randomly selected neurons are ignored during training³⁵. The number of iterations in the
204 training process was also limited to the point after which more iterations did not improve training significantly and
205 may lead to overfitting, a method known as "early stopping"³⁶. The patch size of the convolutional filter, the
206 pooling parameters and the number of hidden units indicated in Figure 2, as well as the Stochastic Gradient
207 Descent hyper-parameters (number of steps=600, batch size=32), were determined from our previous work using
208 EEG spectrograms²⁸, but no fine-tuning or optimization of parameters was applied.

209

210 We compared the DCNN's performance with a Shallow Neural Network (SNN) (Figure 2c), a more basic machine
 211 neural network with a hidden layer, and with a RNN consisting of stacked LSTM^{16,37}, a type of RNN capable of
 212 using information about events in the past (memory) to inform predictions in the future (Figure 2b).

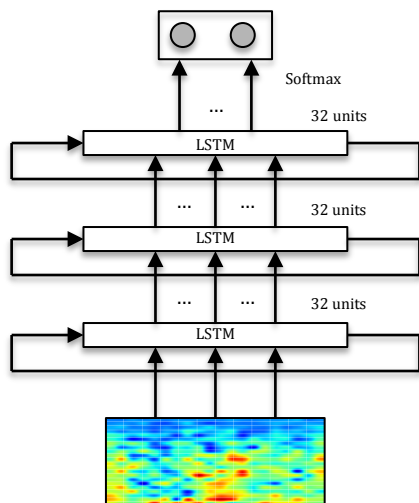
213

214 a) DCNN

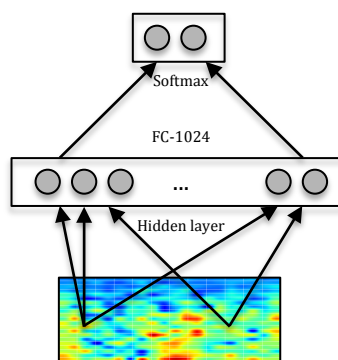


215
 216
 217
 218

b) RNN



c) SNN



219
 220

221 **Figure 2. Network architectures. a) DCNN** model displaying input, convolution with pooling layers, and hidden-unit layers.
 222 The first two layers perform the convolution, the Rectified Linear Units (ReLU) function and the pooling processes for feature
 223 extraction. The last two layers with 128 and 64 hidden nodes perform the class classification in HC or ADHD. For each trial (or
 224 frame), the classifier outputs the probability of the frame belonging to each class (using the softmax function¹⁶) and, after
 225 averaging over frames per subject, we obtained the probability of the subject belonging to each class. Classification was
 226 performed by choosing the class with maximal probability. **b) RNN** consisting of three stacked layers of LSTM cells, where

227 each cell uses as input the outputs of the previous one. Each cell used 32 hidden units, and dropout was used to regularize it. e)
228 SNN architecture used for comparison with one layer of 1024 units.

229

230 Performance assessment

231 The performance metrics assessed for each architecture were accuracy (probability of good a classification) and
232 area under the curve (AUC)³⁸. Classification performance was validated using the leave-pair out cross-validation
233 (LPO), a method for model selection and performance assessment of deep learning algorithms that consists of
234 training the network N=20x20=400 times (all possible combinations of pairing 1 HC with 1 ADHD), holding one
235 sample from each group out from the training set at a time, and measuring the performance using the held out pair
236 as a test set³⁸.

237

238 To account for the significant differences in age between the ADHD and HC groups, we applied the Inverse
239 Probability Weighting (IPW) method³⁹, which assigns different weights to the subjects in the training process
240 according to their propensity score⁴⁰. The IPW method lead to the same performance without adjustment in all
241 architectures, thus ruling out the effect of age as a confounding factor.

242

243 Feature visualization

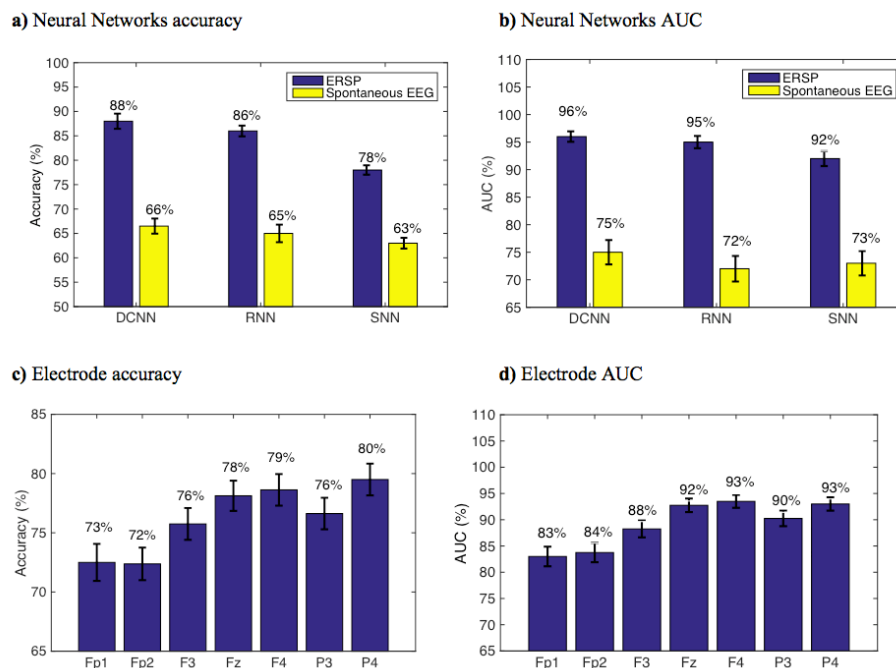
244 Once the network was trained, it was used to find out what type of inputs optimally excite the output nodes using a
245 method popularly known as “DeepDream”, which refers to the generation of synthetic images that produce desired
246 activations in a trained deep network by exaggerating small features within them³⁰. The algorithm maximizes a
247 particular class score using gradient descent, starting from a null or random noise image. In particular, we
248 computed the DeepDream spectrograms averaged over N=400 experiments by maximizing the output logits after
249 30 iterations in steps of 1, initializing with different random images (seeds).

250

251 **Results**

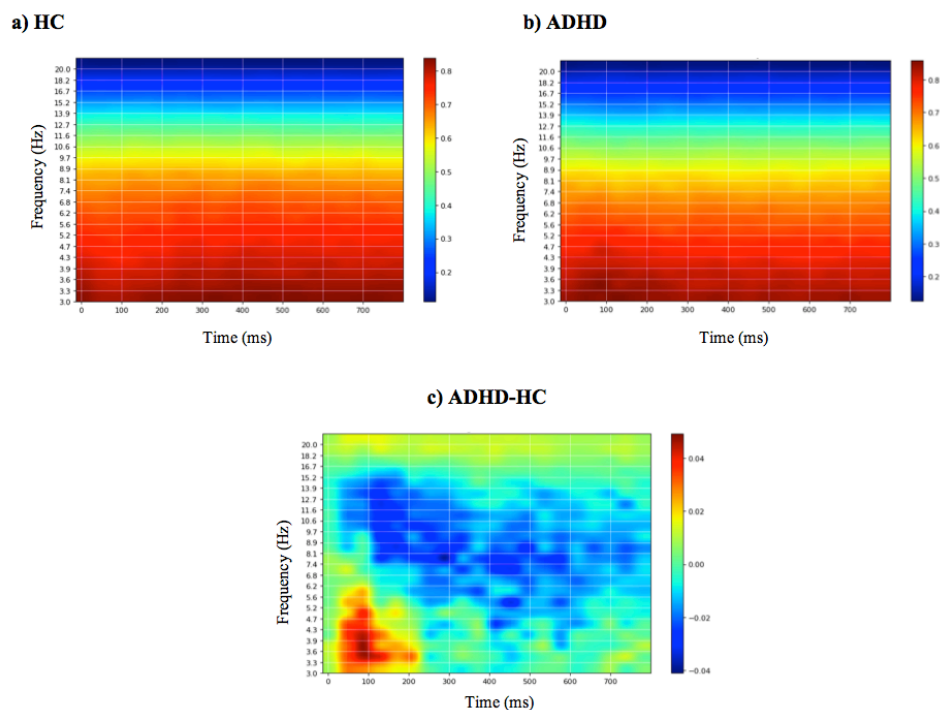
252 The results from classification using different methods and datasets are detailed in Figure 3a and 3b, showing that
253 the DCNN trained with ERSPs reached an accuracy of 88% (AUC=96%), very similar to the RNN performance
254 (Accuracy=86%, AUC=95%) and outperforming the SNN (Accuracy=78%, AUC=92%). In comparison with
255 spontaneous EEG spectrograms, ERSPs provided better performance for all architectures. To assess the
256 performance of each individual channel, we also trained the DCNN with ERSP data from single channels and

257 found that frontal (F3, Fz and F4) and parietal electrodes (P3, P4) provide the best performance compared to
258 frontopolar (Fp1, Fp2) electrodes (Figure 3c and 3d).
259



260
261 **Figure 3. Performance assessment.** Neural networks accuracy (a) and AUC (b) with ERSP and spontaneous EEG data.
262 Electrode accuracy (c) and AUC (d) in a DCNN trained with ERSP data from single channels. Error bars indicate mean square
263 error.

264
265 The mean DeepDream ERSP averaged over channels can be seen in Figure 4 (see Table S1 for individual
266 channels). The difference between groups reveals that the main feature that optimally excites the network nodes is
267 an increased power for the ADHD group in the delta-theta band (3-7 Hz) around 100 ms and a decreased power in
268 the alpha band (7-12 Hz) along the entire time course, with a residual decrease in theta and beta. Note that the
269 patterns shown in the DeepDream ERSP are very similar to the patterns of the ERSP computed from the real data
270 (Table S1), thus showing that the network is actually learning real neurophysiologically identifiable differences
271 between groups.



272

273 **Figure 4. Mean DeepDream ERSP.** Mean DeepDream ERSP averaged over channels generated after 400 experiments for
274 healthy controls (a), ADHD patients (b) and their difference (c). Color bar units = dB.

275

276 Behaviorally, the mean reaction time was significantly slower for ADHD compared to HC ($RT_{ADHD}=368ms$,
277 $RT_{HC}=321ms$, $\beta=46ms$, $CI=[38,53]ms$, $p<0.001$), which can be expected with this type of population with
278 attention deficits, but the mean percentage of responses for each group was not significantly different
279 ($Accuracy_{ADHD}=62%$, $Accuracy_{HC}=65%$, $\beta=0.12$, $CI=[0.02, 0.26]$, $p=0.10$).

280

281 **Discussion**

282 In this study we present a viable deep learning model for effective discrimination of patients with ADHD,
283 providing a new tool for the analysis of EEG dynamics in ADHD and supporting the potential of deep learning
284 strategies for biomarker development in neuropsychiatry. We deem this approach to be particularly interesting for
285 various reasons. First, it largely mitigates the need for EEG feature selection (spectral bands, time ranges, specific
286 ERP components and channels). Second, results with ERSPs represent an improvement over spontaneous EEG
287 spectrograms (e.g. subject Accuracy with DCNN was 88% for ERSP vs. 66% for spontaneous resting-state data).
288 Third, the performances of the proposed DCNN and RNN systems are very similar and they outperform the SNN
289 used for comparison. Finally, through the use of feature visualization, we identify neurophysiologically
290 interpretable features that can be extracted from the model, providing further validation and evidence that the

291 network performance is not driven by noise or artifact signals in the data and providing a mechanistic model with
292 added value to understand pathophysiology.

293

294 The higher accuracy provided by DCNN and RNN compared to SNN proves that the complex deep approaches
295 with more layers and units provide better performance than more shallow networks. The similar performance of
296 DCNN compared to RNN shows, however, that the higher computational demands of RNN do not provide better
297 performance than the DCNN approaches, thus proving DCNN as a more efficient method than RNN.

298

299 The fact that ERSP data provide better performance than spontaneous EEG data with all architectures also shows
300 that event-related data from a highly yield task that elicits the primary executive functions impaired in ADHD is a
301 better predictor than spontaneous EEG data recorded while the participants are at resting state.

302

303 Finally, through the use of feature visualization we show that the main spectral features picked up by the DCNN
304 nodes are a decrease in alpha activity over the entire time course and an increased delta-theta activity around 100
305 ms for ADHD patients compared to HC. There is evidence that an increased alpha activity (or alpha Event-Related
306 Synchronization, ERS) in conflict and inhibitory tasks is related with an improved inhibition of the prepotent
307 response, reflecting a top-down inhibitory control process⁴¹. Therefore, we interpret the decrease in alpha power in
308 ADHD as a deficit in cognitive control. On the other hand, the increased delta-theta activity is localized to 100ms
309 and is probably related to the increase in N100 amplitude in the time domain (Table S1). N100 is a visual sensory
310 evoked potential that is thought to index sensory analysis of simple stimulus features and whose amplitude is
311 influenced by selective attention⁴². The increased delta-theta power in that latency suggests that ADHD patients
312 manifest specific alterations in the process of early selection of visual task stimuli⁴³. Given that there were no
313 significant differences in the percentage of correct responses between ADHD and HC, we interpret this as a
314 compensation strategy to offset inhibitory deficits by shifting more attention to the task^{44,45}.

315

316 Note that the DeepDream spectrograms generated for Fp1 and Fp2 are substantially different and provide lower
317 performance than the other positions (F3, F4, Fz, P3, P4), which may be explained by the lower signal quality of
318 frontopolar positions due to blinks, muscle artefacts and sweat. The lower performance of F3 and P3 electrodes
319 compared to F4 and P4 may also be related to the lower power scale in their DeepDream spectrogram, respectively

320 (Table S1). This may suggest the existence of inter-hemispheric differences in the features driving the
321 discrimination between ADHD and HC.

322

323 Similar studies have explored the application of deep learning to EEG signals. For example, DCNNs have been
324 used for epilepsy prediction and monitoring⁴⁶, mental workload classification⁴⁷ and motor imagery classification^{48,49}.

325 Deep neural networks have also shown convincing results in classifying psychiatric disorders such as dementia⁵¹
326 and ADHD⁵²⁻⁵⁷, mostly with MRI data. To our knowledge, this is the first study using a deep learning approach

327 with EEG event-related spectral data to discriminate adult ADHD patients from HC with no prior selection of
328 EEG features and the combination with feature visualization techniques to provide further mechanistic evidence of

329 the underlying pathophysiology driving the classification. This is particularly important, as it not only allows to
330 develop clinical tools but also to delineate pathological signatures and disease mechanisms.

331

332 One of the limitations of this study is the relatively small size of the dataset, with the consequent limitation on the
333 network due to susceptibility to overfitting. Although “early stopping” and regularization should mitigate this

334 issue, further improvements could be achieved with bigger datasets. Another limitation is the age difference
335 between the two groups. While the mean ages are well after the period of brain maturation when myelination and

336 ADHD symptoms are still changing, and well before a geriatric threshold when other type of biological changes
337 (including normal aging) may affect cognition, we addressed this possible confounder using Inverse Probability

338 Weighting. The age difference was an artifact caused by the fact that the two cohorts were recruited prospectively
339 for independent studies (though at the same time and with the same exact protocol and hardware) and then

340 analyzed together retrospectively to address the proposed questions, hence the lack of appropriately age-matched
341 controls. Future prospective validation studies should use larger cohorts and randomize age-matched controls.

342

343 It is also worth mentioning that, although the current work considerably eliminates the need for manual extraction
344 of features, it is still focused on classification during high yield incongruent trials of a specific task. While this

345 requires a priori knowledge constrains, if validated with higher definition EEG and bigger datasets, it may be a
346 helpful diagnostic and biomarker development strategy (i.e. choosing high yield events of a high yield task) with

347 practical future procedural advantages (i.e. it would be easy to implement it in clinical settings with currently
348 existing tools, such as tasks for neuropsychological assessments and standard EEG for electrophysiological

349 diagnosis).

350

351 Our findings may also have several implications from the clinical perspective by bringing new information to
352 inform the clinician's decisions. Although the networks in this study have been trained with a small dataset of 40
353 subjects, if validated with bigger datasets this approach could be used to support the diagnosis of ADHD on a
354 single-patient basis. The fact that the current networks have been trained with low-resolution EEG datasets (7
355 channels) of short duration (3 minutes) would make it easy to implement them not only in an EEG clinical unit,
356 but possibly by an outpatient clinician, eliminating the need to get longer or higher quality data with sophisticated
357 and clinically unpractical EEG systems. However, even if these deep learning systems are properly validated in the
358 future, clinicians should view their output as statistical predictions, not as a ground truth, and they should judge
359 whether the prediction applies to that specific patient and decide if additional data or expertise is needed to inform
360 that decision.

361
362 Future work should include the exploration of this approach with larger datasets as well as a more systematic study
363 of network architecture and regularization schemes. This includes the use of deeper architectures, improved data
364 augmentation methods, alternative data segmentation and normalization schemes. With regards to data
365 preprocessing, we should consider improved spectral estimation using more advanced techniques such as state-
366 space estimation and multitapering⁵⁸, and the use of cortical or scalp-mapped EEG data prior creation of
367 spectrograms.

368
369 Finally, we note that we make no attempt to fully-optimize our architecture in this study. In particular, no fine-
370 tuning of hyper-parameters has been carried out using a validation set approach, a task we reserve for future work
371 with larger datasets. Our aim was to validate the idea that deep learning approaches can provide value for the
372 analysis of time-frequency representations of EEG, and particularly ERSP data, for the effective discrimination of
373 ADHD.

374

375 **Data availability**

376 The datasets generated and analyzed during the current study are available from the senior author (JAC) on
377 reasonable request.

378

379 **Acknowledgements**

380 This research was partly supported by NIH grants (RO1 MH112737, R21 DA042271, R21 AG056958 and R21
381 MH113018) and the Louis V. Gerstner III Research Scholar Award to JAC.

382

383 **Author contributions**

384 LD contributed with the processing of the data, the implementation of the deep learning systems, and the writing
385 of the article. GR contributed with the conception and design of the deep learning systems and the revision of the
386 manuscript. JC contributed with the conception and design of the study, the supervision of data acquisition and the
387 findings, and the critical revision of the manuscript. All authors gave the approval to the final version of the
388 manuscript to be published.

389

390 **Competing Interests**

391 LDV is an employee at Neuroelectrics and a PhD student in the Camprodon Lab. GR is a co-founder of
392 Neuroelectrics, a company that manufactures the EEG device used in this study. JAC is a member of the scientific
393 advisory board for Apex Neuroscience Inc.

394

395 **References**

396

397 1 Shamay-Tsoory, S. G. & Aharon-Peretz, J. Dissociable prefrontal networks for cognitive and affective
398 theory of mind: a lesion study. *Neuropsychologia* **45**, 3054-3067,
399 doi:10.1016/j.neuropsychologia.2007.05.021 (2007).

400 2 Biederman, J. *et al.* Functional impairments in adults with self-reports of diagnosed ADHD: A controlled
401 study of 1001 adults in the community. *The Journal of clinical psychiatry* **67**, 524-540 (2006).

402 3 Fayyad, J. *et al.* Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder.
403 *The British journal of psychiatry : the journal of mental science* **190**, 402-409,
404 doi:10.1192/bjp.bp.106.034389 (2007).

405 4 Kessler, R. C. *et al.* The prevalence and correlates of adult ADHD in the United States: results from the
406 National Comorbidity Survey Replication. *The American journal of psychiatry* **163**, 716-723,
407 doi:10.1176/ajp.2006.163.4.716 (2006).

408 5 Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J. & Rohde, L. A. The worldwide prevalence of
409 ADHD: a systematic review and metaregression analysis. *The American journal of psychiatry* **164**, 942-
410 948, doi:10.1176/ajp.2007.164.6.942 (2007).

411 6 Jasper, H. H., Solomon, P. & Bradley, C. Electroencephalographic analyses of behavior problem
412 children. *American Journal of Psychiatry* **95**, 641-658, doi:10.1176/ajp.95.3.641 (1938).

- 413 7 Lenartowicz, A. & Loo, S. K. Use of EEG to Diagnose ADHD. *Current psychiatry reports* **16**, 498,
414 doi:10.1007/s11920-014-0498-0 (2014).
- 415 8 Fabiani, M., Gratton, G. & Federmeier, K. D. in *Handbook of psychophysiology, 3rd ed.* 85-119
416 (Cambridge University Press, 2007).
- 417 9 Sanei, S. & Chambers, J. A. *EEG Signal Processing.* (John Wiley & Sons Ltd, 2013).
- 418 10 Woodman, G. F. A brief introduction to the use of event-related potentials in studies of perception and
419 attention. *Attention, perception & psychophysics* **72**, 2031-2046, doi:10.3758/app.72.8.2031 (2010).
- 420 11 Pokryszko-Dragan, A. *et al.* Event-related potentials and cognitive performance in multiple sclerosis
421 patients with fatigue. *Neurological sciences : official journal of the Italian Neurological Society and of*
422 *the Italian Society of Clinical Neurophysiology* **37**, 1545-1556, doi:10.1007/s10072-016-2622-x (2016).
- 423 12 Paulraj, M. P., Subramaniam, K., Yaccob, S. B., Adom, A. H. & Hema, C. R. Auditory evoked potential
424 response and hearing loss: a review. *The open biomedical engineering journal* **9**, 17-24,
425 doi:10.2174/1874120701509010017 (2015).
- 426 13 Durstewitz, D., Koppe, G. & Meyer-Lindenberg, A. Deep neural networks in psychiatry. *Molecular*
427 *Psychiatry*, doi:10.1038/s41380-019-0365-9 (2019).
- 428 14 Kroupi, E. *et al.* in *HBP Student Conference - Transdisciplinary Research Linking Neuroscience, Brain*
429 *Medicine and Computer Science* (Viena, Austria, 2017).
- 430 15 Ruffini, G., Ibañez, D., Castellano, M., Dunne, S. & Soria-Frisch, A. EEG-driven RNN Classification for
431 Prognosis of Neurodegeneration in At-Risk Patients. *Artificial Neural Networks and Machine Learning –*
432 *ICANN 2016*, 306-313 (2016).
- 433 16 Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning.* (MIT Press, 2016).
- 434 17 Mueller, A., Candrian, G., Kropotov, J. D., Ponomarev, V. A. & Baschera, G. M. Classification of
435 ADHD patients on the basis of independent ERP components using a machine learning system. *Nonlinear*
436 *Biomedical Physics* **4**, S1, doi:10.1186/1753-4631-4-s1-s1 (2010).
- 437 18 Tenev, A. *et al.* Machine learning approach for classification of ADHD adults. *International journal of*
438 *psychophysiology : official journal of the International Organization of Psychophysiology* **93**, 162-166,
439 doi:10.1016/j.ijpsycho.2013.01.008 (2014).
- 440 19 Jahanshahloo, H. R., Shamsi, M., Ghasemi, E. & Kouhi, A. Automated and ERP-Based Diagnosis of
441 Attention-Deficit Hyperactivity Disorder in Children. *Journal of Medical Signals and Sensors* **7**, 26-32
442 (2017).

- 443 20 Nazhvani, A. D., Boostani, R., Afrasiabi, S. & Sadatnezhad, K. Classification of ADHD and BMD
444 patients using visual evoked potential. *Clinical neurology and neurosurgery* **115**, 2329-2335,
445 doi:10.1016/j.clineuro.2013.08.009 (2013).
- 446 21 Sadatnezhad, K., Boostani, R. & Ghanizadeh, A. Classification of BMD and ADHD patients using their
447 EEG signals. *Expert Systems with Applications* **38**, 1956-1963,
448 doi:<https://doi.org/10.1016/j.eswa.2010.07.128> (2011).
- 449 22 Ahmadlou, M. & Adeli, H. Wavelet-synchronization methodology: a new approach for EEG-based
450 diagnosis of ADHD. *Clinical EEG and neuroscience* **41**, 1-10, doi:10.1177/155005941004100103
451 (2010).
- 452 23 Abibullaev, B. & An, J. Decision support algorithm for diagnosis of ADHD using
453 electroencephalograms. *Journal of medical systems* **36**, 2675-2688, doi:10.1007/s10916-011-9742-x
454 (2012).
- 455 24 Ruffini, G. *et al.* Algorithmic complexity of EEG for prognosis of neurodegeneration in idiopathic rapid
456 eye movement behavior disorder (RBD). *bioRxiv* (2018).
- 457 25 Oord, A. v. d., Dieleman, S. & Schrauwen, B. in *Proceedings of the 26th International Conference on*
458 *Neural Information Processing Systems - Volume 2* 2643-2651 (Curran Associates Inc., Lake Tahoe,
459 Nevada, 2013).
- 460 26 Tsinalis, O., Matthews, P., Guo, Y. & Zafeiriou, S. *Automatic Sleep Stage Scoring with Single-*
461 *Channel EEG Using Convolutional Neural Networks.* (2016).
- 462 27 Vilamala, A., Madsen, K. H. & Hansen, L. K. Deep Convolutional Neural Networks for Interpretable
463 Analysis of EEG Sleep Stage Scoring. *2017 International workshop on Machine Learning for signal*
464 *processing* (2017).
- 465 28 Ruffini, G. *et al.* Deep Learning With EEG Spectrograms in Rapid Eye Movement Behavior Disorder.
466 *Frontiers in neurology* **10**, 806 (2019).
- 467 29 Dauwels, J., Vialatte, F., Musha, T. & Cichocki, A. A comparative study of synchrony measures for the
468 early diagnosis of Alzheimer's disease based on EEG. *NeuroImage* **49**, 668-693,
469 doi:<https://doi.org/10.1016/j.neuroimage.2009.06.056> (2010).
- 470 30 Alexander, M., Christopher, O. & Mike, T. Inceptionism: Going Deeper into Neural Networks. *Google*
471 *Research Blog* (2015).

- 472 31 American-Psychiatric-Association. *The Diagnostic and Statistical Manual of Mental Disorders*. 5th
473 Edition edn, (American Psychiatric Publishing, 2013).
- 474 32 Kessler, R. C. *et al.* The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short
475 screening scale for use in the general population. *Psychol Med* **35**, 245-256 (2005).
- 476 33 Kopp, B., Rist, F. & Mattler, U. N200 in the Flanker task as a neurobehavioral tool for investigating
477 executive control. *Psychophysiology* **33**, 282-294 (1996).
- 478 34 Abadi, M. *et al.* in *Proceedings of the 12th USENIX conference on Operating Systems Design and*
479 *Implementation* 265-283 (USENIX Association, Savannah, GA, USA, 2016).
- 480 35 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to
481 Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929-1958
482 (2014).
- 483 36 Prechelt, L. in *Neural Networks: Tricks of the Trade* (eds Genevieve B. Orr & Klaus-Robert Müller)
484 55-69 (Springer Berlin Heidelberg, 1998).
- 485 37 Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735-1780,
486 doi:10.1162/neco.1997.9.8.1735 (1997).
- 487 38 Antti, A., Tapio, P., Willem, W., Bernard De, B. & Tapio, S. 3-13 (PMLR, 2009).
- 488 39 Linn, K. A., Gaonkar, B., Doshi, J., Davatzikos, C. & Shinohara, R. T. Addressing Confounding in
489 Predictive Models with an Application to Neuroimaging. *The international journal of biostatistics* **12**, 31-
490 44, doi:10.1515/ijb-2015-0030 (2016).
- 491 40 Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in
492 Observational Studies. *Multivariate behavioral research* **46**, 399-424,
493 doi:10.1080/00273171.2011.568786 (2011).
- 494 41 Klimesch, W., Sauseng, P. & Hanslmayr, S. EEG alpha oscillations: The inhibition–timing hypothesis.
495 *Brain Research Reviews* **53**, 63-88, doi:<https://doi.org/10.1016/j.brainresrev.2006.06.003>
496 (2007).
- 497 42 Rugg, M. D., Milner, A. D., Lines, C. R. & Phalp, R. Modulation of visual event-related potentials by
498 spatial and non-spatial visual selective attention. *Neuropsychologia* **25**, 85-96 (1987).
- 499 43 Yordanova, J., Heinrich, H., Kolev, V. & Rothenberger, A. Increased event-related theta activity as a
500 psychophysiological marker of comorbidity in children with tics and attention-deficit/hyperactivity

- 501 disorders. *NeuroImage* **32**, 940-955,
502 doi:<https://doi.org/10.1016/j.neuroimage.2006.03.056> (2006).
- 503 44 Broyd, S. J. *et al.* The effect of methylphenidate on response inhibition and the event-related potential of
504 children with attention deficit/hyperactivity disorder. *International journal of psychophysiology : official*
505 *journal of the International Organization of Psychophysiology* **58**, 47-58,
506 doi:10.1016/j.ijpsycho.2005.03.008 (2005).
- 507 45 Prox, V., Dietrich, D. E., Zhang, Y., Emrich, H. M. & Ohlmeier, M. D. Attentional processing in adults
508 with ADHD as reflected by event-related potentials. *Neuroscience Letters* **419**, 236-241 (2007).
- 509 46 Liang, J., Lu, R., Zhang, C. & Wang, F. in *IEEE International Conference on Healthcare Informatics*
510 *(ICHI)* 184-191 (2016).
- 511 47 Ma, T. *et al.* The extraction of motion-onset VEP BCI features based on deep learning and compressed
512 sensing. *J Neurosci Methods* **275**, 80-92, doi:10.1016/j.jneumeth.2016.11.002 (2017).
- 513 48 Bashivan, P., Rish, I., Yeasin, M. & Codella, N. *Learning Representations from EEG with Deep*
514 *Recurrent-Convolutional Neural Networks*. (2015).
- 515 49 Tabar, Y. R. & Halici, U. A novel deep learning approach for classification of EEG motor imagery
516 signals. *J Neural Eng* **14**, 016003, doi:10.1088/1741-2560/14/1/016003 (2017).
- 517 50 An, X., Kuang, D., Guo, X., Zhao, Y. & He, L. in *Intelligent Computing in Bioinformatics*. (eds De-
518 Shuang Huang, Kyungsook Han, & Michael Gromiha) 203-210 (Springer International Publishing).
- 519 51 Vieira, S., Pinaya, W. H. & Mechelli, A. Using deep learning to investigate the neuroimaging correlates
520 of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev* **74**, 58-75,
521 doi:10.1016/j.neubiorev.2017.01.002 (2017).
- 522 52 Deshpande, G., Wang, P., Rangaprakash, D. & Wilamowski, B. Fully Connected Cascade Artificial
523 Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional
524 Magnetic Resonance Imaging Data. *IEEE transactions on cybernetics* **45**, 2668-2679,
525 doi:10.1109/tcyb.2014.2379621 (2015).
- 526 53 Han, X., Zhong, Y., He, L., Yu, P. S. & Zhang, L. in *Brain Informatics and Health*. (eds Yike Guo *et al.*)
527 156-166 (Springer International Publishing).
- 528 54 Hao, A. J., He, B. L. & Yin, C. H. in *2015 IET International Conference on Biomedical Image and Signal*
529 *Processing (ICBISP 2015)*. 1-6.
- 530 55 Kuang, D. & He, L. in *2014 International Conference on Cloud Computing and Big Data*. 27-32.

- 531 56 Kuang, D., Guo, X., An, X., Zhao, Y. & He, L. in *Intelligent Computing in Bioinformatics*. (eds De-
532 Shuang Huang, Kyungsook Han, & Michael Gromiha) 225-232 (Springer International Publishing).
- 533 57 Zou, L., Zheng, J., Miao, C., Mckeown, M. J. & Wang, Z. J. 3D CNN Based Automatic Diagnosis of
534 Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access* **5**, 23626-
535 23636, doi:10.1109/ACCESS.2017.2762703 (2017).
- 536 58 Kim, S.-E., Behr, M. K., Ba, D. & Brown, E. N. State-space multitaper time-frequency analysis.
537 *Proceedings of the National Academy of Sciences* **115**, E5, doi:10.1073/pnas.1702877115 (2018).
538