Early Prediction of Attention Problems and Hyperactivity at 6 years of Age from Brain Morphometry at 1 year of Age and Comparison with Models using Data Imputation
Honors Thesis

Mattheus Victor Guimaraes Bezerra
Department of Computer Science
University of North Carolina at Chapel Hill

October 2020

Abstract

Ongoing longitudinal studies are exploring developmental trajectories of brain morphometry in infants [9]. Recent research has established a link between the development of symptoms related to neurodevelopmental disorders with variations in morphometric measures attained via magnetic resonance imaging (MRI) [5]. In this thesis, I explore predictive models for classification of singletons at high risk (HR) and low risk (LR) for atypical brain development associated with Attention-deficit/Hyperactivity disorder (ADHD). The final models are compared based on standard binary classification schemes. They differ in use of training data, classification groupings, and label for measured behavior. My findings indicate that a predictive signal exists between these morphometric features and behavioral symptoms. When trained with supplementary interpolated data, most model performances improved to make more accurate classifications by expanding the number of training samples. More samples led to better performance, indicating that accurate data imputation can benefit predictive, low-sample studies. However, further research is needed to determine whether a combination of different brain morphometric measurements, additional data, or further model optimization can lead to a sufficiently accurate prediction for clinical use. Supplementary work may also be required via recruitment of infants for additional morphometric measures, parcellation of infant brain structure, recreation of models using different adult-brain parcellation techniques, or creation of a threshold scheme for binary risk classification.

Approved by:
Thesis advisor: Martin A Styner _____
Secondary reader: Marc Niethammer _____

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Objective

Attention-deficit/Hyperactivity disorder (ADHD) is one of the most common neurodevelopmental disorders that is seen in children today. Most recently, the number of children aged 2-17 in the United States diagnosed with ADHD is about 9.4% according to national survey in 2016. Diagnosis via one test is not possible currently, and healthcare providers use American Psychiatric Association's Diagnostic and Statistical Manual, Fifth edition (DSM-5) to diagnose ADHD in standardized fashion [7]. The most common symptoms include inattention, hyperactivity, and impulsivity. Many of these collide with other disorders, and are not exclusive to ADHD. Therefore, additional conditions must be met. The Centers for Disease Control (CDC) divides ADHD into three types based on the strongest characteristics of an individual's symptoms:

- Combined Presentation - Enough symptoms of both criteria inattention and hyperactivity-impulsivity were present for the past 6 months.
- Predominantly Inattentive Presentation - Enough symptoms of inattention, but not hyperactivity-impulsivity, were present for the past 6 months.
- Predominantly Hyperactive-Impulsive Presentation - Enough symptoms of hyperactivity-impulsivity, but not inattention, were present for the past 6 months [6, 7].

Other studies have focused on parent and teacher reporting to help diagnose at ages where children enter school, and have found that certain characteristics can be quantified to determine the risk of a child developing a neurodisorder [17]. Currently, there is strong interest for an early-warning system using biological markers.

1.2 Longitudinal Prediction

During infancy, the human brain undergoes a period of neural development that is crucial for cognitive and behavioral capacities [12]. For most disorders it is generally recognized that early, accurate diagnosis is important for intervention that can assist in the development of afflicted individuals. It has been previously acknowledged that knowledge regarding early brain development from birth to about 4 years of age is limited. Given the growth of the brain in this period, studies such as [19] have emerged to observe structural brain development. Longitudinal magnetic resonance imaging (MRI) characterizes developmental trajectories and is critical in neuroimaging studies of early brain development. Quantification of developmental trajectories allows for analysis and observation of neurodevelopment [9]. This has led to studies correlating neurodevelopmental disorders with morphometric features of the brain attained via MRI [5].

However, missing data points are a common issue in longitudinal studies due for a variety of reasons [22]. Participant attrition and scan failure are

common, and discarding study participants with incomplete data significantly reduces the sample size. Models can often become heavily biased or be trained on data that is not representative of a greater population when faced with too few samples [26]. One solution to deal with the issue of missing data is to interpolate the missing data. This process is known as data imputation, and can be performed in the case of MRIs at either the image or at the measurement level [14]. Data imputation is considered a better way to address holes in data in order to preserve all available samples for research.

In this thesis, I combine these areas of interest by exploring creation of an early-diagnostic system and investigating if there is predictive power in measures of early, atypical brain development associated with ADHD using morphometric features. I compare classification models that predict risk groups for measured Behavior Assessment System for Children Second Edition (BASC-2) behaviors based on their morphometric features and demographics. They differ in use of training data, classification groupings, and label for measured behavior.

2 Methods

2.1 Data

The primary dataset was acquired as part of the University of North Carolina (UNC) Early Brain Development Study which investigates multiple lines of research regarding brain development and neurodevelopmental trajectories [10]. In this study, MRIs were collected from pediatric subjects at ages of 1 and 2 years between 2004 to 2014; similar measures at 4 and 6 years were added later. This study is the first to investigate the relationships between cortical thickness (CT), surface area (SA), and emerging cognitive abilities in a large, healthy sample. Henceforth, this data set shall be referred to as the Gilmore data set.

A similar, supplementary dataset was acquired as part of the Infant Brain Imaging Study (IBIS) study. From this study, MRIs were collected and morphometric features including CT and SA were extracted. These measurements were taken from 6 to 36 months, with most subjects having MRIs at 6, 12, and 24 months [30]. Demographics and a diagnosis for autism spectrum disorder (ASD) were acquired but remain unused in this study.

Previous studies have focused on reporting behavior to quantify developmental characteristics and correlate them to neurodevelopmental disorders such as ADHD. The BASC-2 and Behavior Rating Inventory of Executive Function (BRIEF) use behavioral rating forms to assess a child's emotional and behavioral difficulties, including problems with attention and hyperactivity [17]. The BASC-2 contains demographics for subjects including sex, gestational age (GA), gestation number, birth weight, and ethnicity. Parents report on several indicators such as aggression, hyperactivity (HYP), attention problems (ATP), and more are quantified based on activity in the past 6 months. These indicators are mapped to a mean t-score of 50 and standard deviation of 10. Confidence intervals and percentile ranks also are reported for each scale [33].

2.1.1 Data Processing

Study and Age	Count
Gilmore 1 year old singleton	248
Gilmore 2 year old singleton	188
Gilmore 4 year old singleton	162
Gilmore 6 year old singleton	172
IBIS 1 year old	444
IBIS 2 year old	357

Table 1: Total subjects for each year in each study before any processing was applied. Note that a singleton is a child that is the only one born at one birth. The IBIS data did not guarantee whether or not cases were singletons or twins. See section 2.1.1 for further information.

Study and Age	BASC-2	Count
Gilmore 1, 2 year old singleton	Mixed Availability	129
Gilmore 1, 2, 6 year old singleton	Available	85
Gilmore 1, 6 year old singleton	Available	42
Gilmore 2, 6 year old singleton	Available	34
IBIS 1, 2 year old	Unavailable	290

Table 2: Total subjects available declined for every group after implying restrictions on longitudinal data for specific time points. The test set for this study is the union of rows 2 and 3. Therefore I want the Gilmore singletons that were able to complete a 1 year MRI and 6 year MRI. Criteria for inclusion can be found in Section 2.1.1.

In order to create my early-diagnostic system based upon morphometric features, two systems of information were required: the morphometric features at age 1, and a later ATP t-score from the same subject. I processed data by removing any subjects that had not completed the BASC-2 or did not have morphometric features available at 1 and 2 years. Several subjects also did not complete their morphometric measurements at 4 and 6 years old. Individuals that were missing only their 1 year morphometric measurements but still had their 2 year morphometric measurements were kept for models employing data imputation, as listed in 2.2.4. Table 1 shows the pre-processing counts of each group available in the data sets. These individuals participated in the Gilmore study and had their MRIs at the designated ages. These samples are not assumed to have any assessment via the BASC-2 nor BRIEF.

Due to several incomplete cases, the number of usable subjects was greatly reduced. Table 2 shows the counts available after implementing the criteria described above. As seen, the number of subjects with necessary data dramatically reduces the sample size available. Row 1 shows the 129 singletons that

started the study and remained for at the second year's scans. This does not give any information about their 6-year scans. When we begin to account for the remaining subjects that completed their 6-year assessment and scan, only 85 remained.

Row 3 accounts for subjects that had morphometric features at 1 year of age, completed their BASC-2, but missed their 2-year scan. Row 4 includes subjects that have available features at 2 years of age, but not at 1 year of age. These participants also have their BASC-2 available, but may have entered the study late.

137 subjects remained with usable morphometric data that had their 1 year MRI and BASC-2/BRIEF assessments. This is the union of rows 2 and 3. The 4 year old morphometric measurements were not considered due to a significantly lower sample size and evidence from [9, 34] confirming that the first 2 years of life are the most critical period of postnatal brain development. The IBIS data did not have BASC-2 assessments and therefore could not be used in the testing set, yet was used for the training the imputation models.

2.1.2 Data Classification

Typicality	Count	Behavior Assessment System for Children (BASC-2)
Typical	115	Attention Problems(ATP)
Atypical	22	Attention Problems(ATP)
Typical	123	Hyperactivity(HYP)
Atypical	14	Hyperactivity(HYP)

Table 3: Total subject counts for the different risk groups. Note that t-score varies between measured assessment, so we separate classification labels by typicality and BASC-2 assessment. We total 137 usable subjects. The description of the risk group definition is given in Section 2.1.2.

Subjects were given classification labels based on their BASC-2 ATP and HYP t-scores separately. high risk (HR) was assigned a label of 1 for binary classification while low risk (LR) were assigned a label of 0. Those with t-scores ≥ 65 were deemed to be developing atypically which correlates to greater risk for ADHD. This is considered to be 1.5 standard deviation(s) (SD) higher than the mean. This is consistent with classification from [17] that averaged a t-score of 71.19 and SD of 9.07 for individuals diagnosed with ADHD. Subjects with a t-score ≤ 65 were considered to be typical which correlates to LR as consistent with the ratings where non-ADHD subjects averaged a t-score 52.08 and SD of 8.13. Table 3 shows the available classes for each typicality and BASC-2 score.

Subjects were also binned into a simplified version of the binary classification. The middle t-scores in this instance were removed. The critieria for these non-contiguous groups is slightly different. Atypicality (assigned a label of 1) associated with HR was assigned to individuals with BASC-2 t-scores ≥ 65. This is the same as for contiguous group setting above. However, typical

Typicality	Count	Behavior Assessment System for Children (BASC-2)
Typical	107	Attention Problems(ATP)
Atypical	22	Attention Problems(ATP)
Typical	111	Hyperactivity(HYP)
Atypical	14	Hyperactivity(HYP)

Table 4: Total subject counts for a simplified binary classification. Again, t-score varies between measured assessment, so we separate classification labels by typicality and BASC-2 assessment. We total a separate amount of subjects in this instance. ATP has 129 usable subjects and HYP has 125 usable subjects. The thresholds for t-scores are given in Section 2.1.2.

subjects associated with LR were assigned a label of 0 if their t-scores were $\leq$60. In doing so, we remove the subjects with scores between 60 and 65 and thus fall outside of one standard deviation but may not be at HR for ATP and HYP. Subjects were given classification labels based upon their BASC-2 ATP and HYP scores separately.

2.1.3 Data Parcellation

The adult human brain can be subdivided into multiple regions of interest (ROI). Divisions are called parcellations, but these can be divided further into sub-parcellations. The Destrieux parcellation featured in [4] is based on the cortical surface using standard internationally accepted nomenclature and criteria. Its labels are consistent with anatomical rules as well as automated computational parcellation.

Both the IBIS and Gilmore data sets readily featured Destrieux parcellations for the morphometric features. The Destrieux method subparcellates the sulcal and gyral parts of the cortex into smaller entities based on classical anatomical descriptions. Because it has so many smaller entities, it can give a precise definition of a cortical surface as compared to other parcellations [4]. Other parcellations were either unavailable between both datasets, or did not include the morphometric features needed.

Through the Destrieux parcellations, 148 ROI (74 from either side of the brain) have become available[1] for use as input features. These are listed and thoroughly described in [4]. Each ROI had its according CT and SA. This parcellation is also featured in Figure 1.

2.2 Models

I considered 8 models when predicting risk for atypicality in ATP and HYP. Models were divided based on their use of interpolated data, BASC-2 assessment, and continuity of subject classification labels. A model was created for

[1]Parcellation is a form of dimensionality reduction. This helps to prevent overfitting in models with low numbers of samples

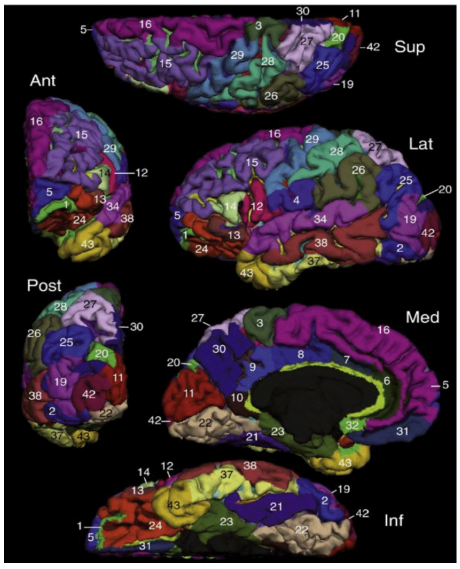

Figure 1: This image is a pial view of the manual parcellation of one hemisphere in Destrieux's training set. Numerical indices refer to the anatomical regions, and we see views for the superior (Sup), anterior (Ant), lateral (Lat), posterior (Post), medial (Med), and inferior (Inf) regions. For a greater understanding or to continue discussion, see the original article [4].

each combination of these attributes, and all models are binary classifiers. Every model was constructed as a Python3 program using the `scikit-learn` [32] and `tensorflow` [25] packages via the `keras API` [2].

Non-linearities exist in male and female brain development. Males in particular tend to exhibit symptoms of and be diagnosed with ADHD at a far higher rate than females [7]. Neural networks and deep-learning models in are strongly suited to learning multiple disjoint, non-linear associations between features (See Section 2.2.1). They are becoming increasingly used when studying problems in neurodevelopment. For example [13, 20, 29] are all examples of use of neural networks for prediction of neurodevelopmental disorders, and in particular [13] has correlated morphometric brain SA to prediction of ASD as early as at 2 years old. Next, I will give an introduction to neural networks, deep-learning, and autoencoders.

2.2.1 Overview of Neural Networks and Deep Learning

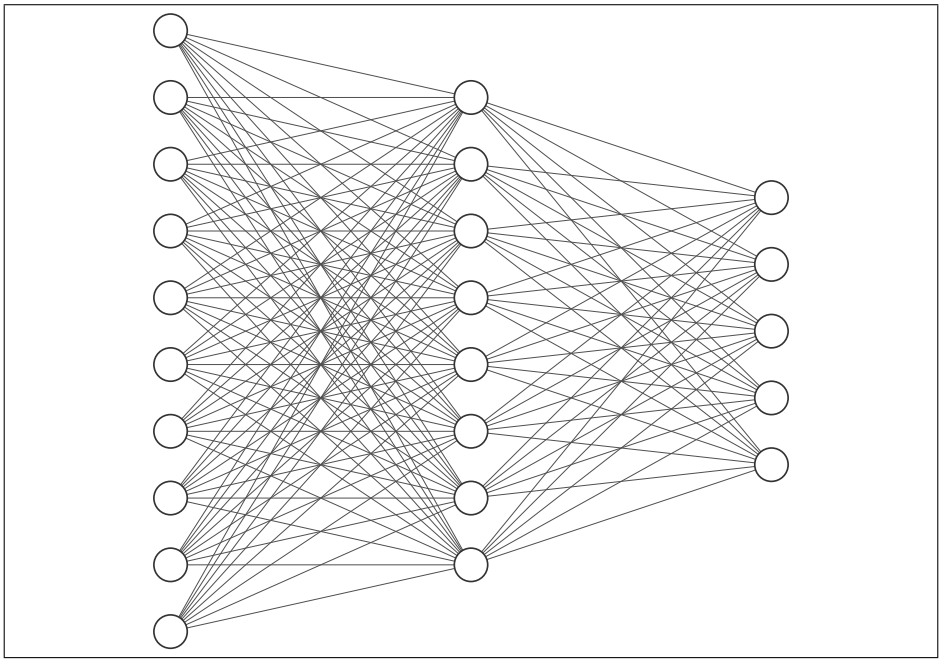

Figure 2: This figure was generated from [23]. This sample neural network is considered shallow because there is only one hidden layer. It has an input layer of 10 nodes, a hidden layer of 8 nodes, and output layer of 5 nodes. For further information see Section 2.2.1.

Neural networks are composed of layers of nodes that can be passed to linear and non-linear functions [15]. Hidden layers appear in between the input and output layers of nodes. Neural networks are considered "deep" when there are

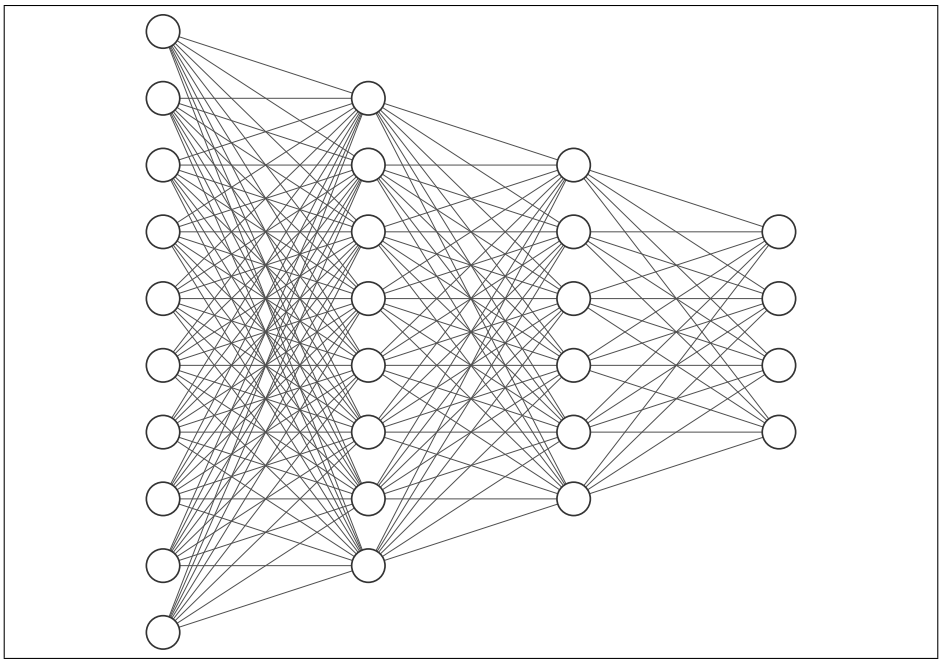

Figure 3: This figure was generated from [23]. This sample neural network is considered deep because there are multiple hidden layers. It has an input layer of 10 nodes, a hidden layer of 8 nodes, another hidden layer of 6 nodes, and output layer of 4 nodes. For further information see Section 2.2.1.

stacks of these layers, and "shallow" when there is only one hidden layer. Deep learning uses networks that stack these hidden layers [8]. We focus on deep learning since these are the networks generated and reviewed in Section 3.

Deep networks can approximate the class of compositional functions as well as shallow networks but with an exponentially lower number of training parameters and sample complexity. A deep network also does not need to have the same architecture as the function it is trying to approximate, implying there is a larger range of training set sizes for which deep networks can learn [27]. Deep neural networks are capable of learning non-linear associations and even multiple disjoint associations. They can improve prediction and have several other problem benefits [8].

The multi-layer perceptron (MLP) is the origin of binary classification via neural networks. It is a logistic regressor where the input layer is transformed using a learned non-linear transformation. Use of hidden layers is sufficient to make a MLP a universal approximator [3]. These hidden layers output to a final node(s) for which a predicted classification can be made. Figure 4 shows an example of a binary classification neural network.

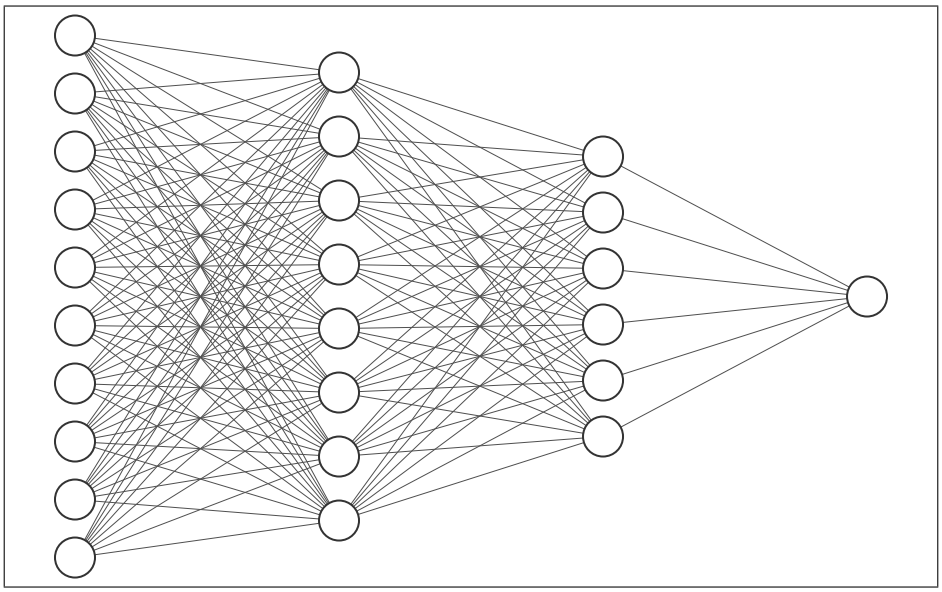

Figure 4: This figure was generated from [23]. This sample binary classification network reduces dimensionality until the final layer which consists of one node. A sigmoid activation function is used to place final classifications in a specified range which can be thresholded for predictions. It is considered deep because there are multiple hidden layers. It has an input layer of 10 nodes, a hidden layer of 8 nodes, another hidden layer of 6 nodes, and output layer of 1 nodes. For further information see Section 2.2.1.

2.2.2 Overview of Autoencoders

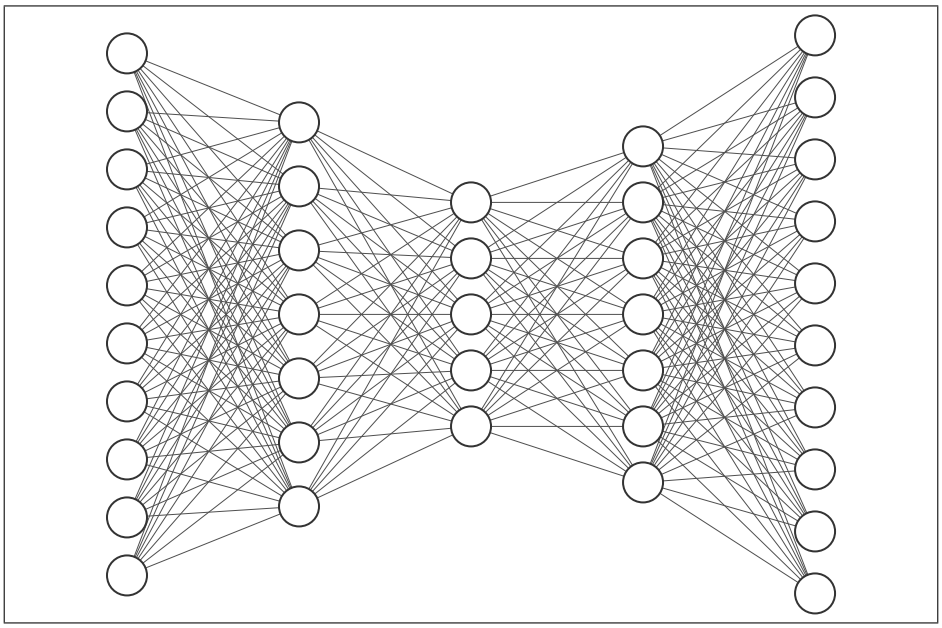

Figure 5: This autoencoder consists of multiple layers for dimensionality reduction. It finds an encoded representation of its input, and then reconstructs information based on the encoded representation. This figure reduces dimensionality and then reconstructs information in the following order: $10 \longrightarrow 8 \longrightarrow 5 \longrightarrow 8 \longrightarrow 10$. For more information, see Section 2.2.2.

Autoencoders are a type of neural network (see Section 2.2.1) used to learn data encodings in an unsupervised manner [21]. Autoencoders learn an encoding for a set of data, usually by reducing dimensionality and training the network to ignore signal noise. After reduction, the data is then reconstructed from its encoded representation.

Autoencoders can have special cases that are not necessarily used to reconstruct data. Several variations of the autoencoder architecture exist for a variety of purposes. In my case, I created an autoencoder that was used instead as a generative model. Instead of just learning an encoded representation of the the data, the encoded representation was then used to reconstruct and predict a past or future data point. Studies such as [31] describe the use of autoencoders to learn continuous feature representations from medical records to make predictions.

2.2.3 Input Data

All models used 1 year old LR and HR subjects as training inputs. Every model had an input layer size of 298 features. 296 of these features came from the SA

and CT derived through the Destrieux parcellation (See Section 2.1.3). Lastly the sex and GA demographics from the BASC-2 were also included in these features.

All numerical features excluding sex were rescaled using minmax normalization found in the `scikit-learn MinMaxScaler` class [32]. To reduce bias, subjects were purposely transformed to the fit of the IBIS study's MinMax scaling. The final scaled input data is summarized in Figure 6.

Cortical Thickness (CT)
Surface Area (SA)
Gestational Age (GA)
Sex (Mapped to a 0/1 index)

Figure 6: Above are the input features into each classification network. Notably CT, SA, GA were scaled relative to a separate population IBIS to avoid any bias that could be gathered from training on a population scaled to its own distribution. Gestational age GA and sex were acquired from a separate demographics sheet of BASC-2 and BRIEF participants. See Section 2.2.3 for details.

2.2.4 Individual Model Configuration

Each network had an input layer, two hidden layers, a dropout layer, and an output layer.

The output layer consisted of a single node with a sigmoid activation function. Sigmoidal nonlinearities are often included in the nodes of the output layer so that the network produces outputs in a fixed, finite range. Scaling can also occur without affecting the generality of the network [3]. In my case, I produced a number in the interval [0,1]. This output was rounded based on a threshold of 0.5 to produce a single classification of typical(0) or atypical(1). This threshold is summarized below where p represents an output from the sigmoidal node.

$$\begin{cases} p \geq 0.5 & \text{HR for atypical brain development associated with symptoms for ADHD} \\ p < 0.5 & \text{LR for atypical brain development associated with symptoms for ADHD} \end{cases}$$

In Table 2 there is a clear imbalance in the number of HR and LR subjects. To deal with the class imbalance I used Synthetic Minority Over-sampling Technique (SMOTE) which creates synthetic samples based on the k-nearest neighbors, and was configured to only create synthetic samples for the minority (HR) class. [1]. SMOTE generated samples were used only in training and not testing.

Models were compiled using Adam for optimization. Adam is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments [18]. I used the binary cross-entropy loss-function

available via tensorflow to measure the difference between network predictions and actual classification risk groups [25].

Model hyper-parameters were set using a grid-search with custom iterator to include indicies for samples generated from SMOTE. The gridsearch was set with the separate population of singleton 2 year data, which was used to try and predict risk groups associated with ATP. `GridsearchCV` is available through [32]. A separate gridsearch was created for models based on their use of non-contiguous data. Therefore, the binary classifiers for contiguous ATP and HYP data all used the same optimized configuration. Similarly, the simplified binary classifiers all used the same optimized configuration. No further attempts to optimize the network architecture were made.

All gridsearches for optimized hyperparameters were done using the `StratifiedKFold` class which cross-validated between 8 folds. These grid-searches also occurred on the available Gilmore samples that contained their 2 and 6 year data points seen in rows 1 and 3 of Table 2. to avoid any overfitting. This solves a different problem by attempting to predict risk groups based on 2 year old CT, SA, GA, and sex data.

2.2.5 Autoencoder Configuration

For generation of the interpolated samples, an autoencoder (see Section 2.2.2) was created. Demographics are already available from the BASC-2 so GA and sex were not interpolated. Only morphometric features such as CT and SA from the Destrieux parcellation were interpolated. One in each autoencoder separately.

Multiple configurations were considered for the autoencoder. Firstly, I used principle component analysis (PCA) to estimate the size of the encoded representation. Note that PCA is the process of computing the principal components and using them to perform a change of basis on the data. PCA is commonly used for dimensionality reduction by obtaining lower-dimensional data via projecting each data point to first few principal components. This preserves the data's variation as much as possible [16, 24]. An autoencoder configuration was create using two hidden layers, with the calculated PCA representing approximately 73% of variance for CT and 62% of variance for SA.

Using the set of recommended encoded layer sizes, I created multiple autoencoders. The first was a model that trained on scaled Gilmore and IBIS data (scaled relative to the IBIS data set) that took 1 year data as input and output 1-year data. A similar model was constructed for the 2-year data. This model was intended to find the best encoded representation of the data. A separate decoder was then constructed for model to predict the 1-year data from the 2-year data and vice versa. Both had similar measures of mean absolute loss, which is derived from the average distance to the original data point from the interpolated one. A combined decoder was also created that input encoded 1-year and 2-year old data, to output 2-year and 1-year data respectively. This had a similar level of loss.

Another model that was created was a traditional autoencoder that input 1-

year data and output 2-year data with no pretraining or separate architectures. This model performed within the same range as those described above. It was reasoned that the similar results despite having separate architectures were an indicator of sufficient data to represent the differences in 1 and 2-year data. For the final interpolated models, the architecture directly predicting 1-year data from their 2-year samples was used.

3 Results

The configuration for all models can be read in Section 2.2.4. Every model was evaluated via an 8-fold cross validation scheme through use of scikit-learn's `StratifiedKFold` class [32]. Each fold tested on 2-3 atypical subjects and 8-10 typical subjects, while training on the remaining set of subjects and generated data via SMOTE. The cross validation scheme allows for every subject to be evaluated exactly once in whichever fold they did not appear in the training data. The stratification of each fold refers to balancing of classes, so that each fold maintains similar amounts of typicals and atypicals in testing.

Results are presented using standardized binary classification measures[2]. More formally: accuracy (ACC), sensitivity (SEN)[3], specificity (SPC)[4], positive predictive value (PPV)[5], and negative predictive value (NPV). Their formulas are given in the following figure. Each is defined in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). In this case TP refers to the case that a subject is correctly predicted as atypical in whichever category is being measured. Similarly, TN refers to the case that subject is correctly predicted as typical in whichever category is being measured. In each case we evaluate models primarily on PPV followed by higher relative binary classification measures. PPV is favored due to its relevance in potential clinical diagnosis [29].

ACC Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

SEN Sensitivity $= \frac{TP}{TP+FN}$

SPC Specificity $= \frac{TN}{TN+FP}$

PPV Positive predictive value $= \frac{TP}{TP+FP}$

NPV Negative predictive value $= \frac{TN}{TN+FN}$

Figure 7: Binary classification measures. See Section 3 for details.

[2]The Wikipedia article https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers provides more information

[3]Also known as recall or true positive rate

[4]Also known as selectivity or true negative rate

[5]Also known as precision

Models without interpolation first appear as a measurement of how many studies would perform with near minimal samples as a result of participant attrition and unusable morphometric data. This listing is followed by a section where the previous models reappear now training with extra interpolated data from real subjects. These imputed missing samples allow us to include usable BASC-2 data from subjects that may have entered the study late or have an unavailable MRI at 1 year old. These models give researchers a better idea of how models may perform when provided with a greater number of samples and members of the minority (atypical) class.

3.1 Attention Problems

Training Data	Model	ACC	SEN	SPC	PPV	NPV
Non-Interpolated	Contiguous	83.2	40.9	90.4	47.4	89.7
	Simplified Binary	78.3	22.7	89.7	31.3	85.0
Interpolated	**Contiguous**	**86.1**	**50.0**	**93.0**	**58.0**	**90.7**
	Simplified Binary	**81.4**	**31.8**	**91.6**	**43.8**	**86.7**

Table 5: The models training without interpolated data appear above, while those training with interpolated data appear below. Measures are given in percentages. See Figure 7 to understand how binary classification measures were derived. Here, the best performing models between interpolated and non-interpolated data-using counterparts are bolded. Binary classification measurements are calculated based on the conglomerate of fold testing data.

Two models exist for predicting attention problems without interpolated data: One model with a simplified set of subjects with t-scores $\geq$65 or $\leq$60 (simplified binary model), and a contiguous model that divides subjects based on t-scores $\geq$65 (contiguous binary model). Model configurations can be reviewed in 2.2.4 and input data can be reviewed in Section 2.2.3.

As shown in Table 5, the contiguous binary model achieved the following results: ACC: 83.2, SEN: 40.9, SPC 90.4, PPV 47.4, and NPV 89.7. The simplified binary model performed slightly worse in all measures with the following results: ACC: 78.3, SEN: 22.7, SPC 89.7, PPV 31.3, and NPV 85.0. Note these measures are given in percentages.

Table 5 also gives the results for the models trained with interpolated data. The contiguous binary model achieved the following results: ACC: 86.1, SEN: 50.0, SPC 93.0, PPV 58.0, and NPV 90.7. The simplified binary model performed slightly worse with measures decreasing with the following results: ACC: 81.4, SEN: 31.8, SPC 91.6, PPV 43.8, and NPV 86.7. A notable significant increase is the rise in PPV as compared to the model without interpolation.

Training Data	Model	ACC	SEN	SPC	PPV	NPV
Non-Interpolated	Contiguous	87.6	14.3	95.9	28.6	90.8
	Simplified Binary	86.4	14.3	95.5	28.6	89.8
Interpolated	**Contiguous**	**89.1**	**14.3**	**97.6**	**40.0**	**90.9**
	Simplified Binary	**88.0**	**14.3**	**97.3**	**40.0**	**90.0**

Table 6: The models training without interpolated data appear above, while those training with interpolated data appear below. Measures are given in percentages. See Figure 7 to understand how binary classification measures were derived. Here, the best performing models between interpolated and non-interpolated data-using counterparts are bolded. Binary classification measurements are calculated based on the conglomerate of fold testing data.

3.2 Hyperactivity

As shown in Table 6, the contiguous binary model achieved the following results: ACC: 87.6, SEN: 14.3, SPC 95.9, PPV 28.6, and NPV 90.8. The simplified binary model performed slightly worse in almost all measures with the following results: ACC: 86.4, SEN: 14.3, SPC 95.5, PPV 28.6, and NPV 89.8. The primary difference is in about 5% for PPV in favor of the contiguous model.

Table 6 again shows the models trained with interpolated data. The contiguous binary model achieved the following results: ACC: 89.1, SEN: 14.3, SPC 97.6, PPV 40.0, and NPV 90.9. The simplified binary model achieved the following results: ACC: 88.0, SEN: 14.3, SPC 97.3, PPV 40.0, and NPV 90.0. In this case the simplified binary model performed better than the contiguous model.

4 Discussion

Models using combinations of binning techniques and interpolated data were evaluated to determine the ability of morphometric features in predicting symptoms related to ADHD. Separate models were created for prediction of ATP and HYP. My results indicate that models using simplified binary data for separation of low and high risk subjects tend to perform worse than models with all subjects included, but all models evidence some possible signal leaving room for future studies. More generally, these results echo the hypothesis that symptoms related to neurodevelopmental disorders can be predicted via morphometric features; other studies hold similar views such as in Mostapha and Hazlett et al. where ASD is predicted from morphometric features. My results also suggest that greater model optimization and investigation must occur before predictive models can achieve clinical strength.

The low values of PPV indicate that these models are not suited for clinical diagnosis, but can portray a signal of prediction for ATP and HYP. Overall the models performed relatively poorly in classification of HR individuals which

may be due to a variety of reasons. My study was limited by available data as a combination of morphometric features and BASC-2 assessments were required. Models may have performed better given more non-imputated samples, and performance may have improved if more HR subjects were available. This may be reinforced as the performance of all models trained on interpolation featured comparable to improved PPV indicating that increased samples may have resulted in a greater number of HR predictions. AS per Table 2 atypical class size was roughly 1/5 to 1/9 the size of the typical class, which may not be statistically significant for classification. This issue is consistent with ideas presented in Section 1.2 and [22, 26]. Additional data allows for models to learn more complex associations with less risk of overfitting and better generalization. Binary classification measures related to the LR class increased slightly when introduced to more samples.

Models involving the prediction of HYP performed considerably worse than models predicting ATP which may be attributed to the gridsearch settings being optimized for prediction of ATP from 2-year singleton measurements. A separate model using optimized hyperparameters for HYP may perform better, and could be performed on a separate population of singletons or a different data set e.g. 2-year singleton data for minimal bias but optimal performance. The model using interpolated data in the contiguous classifier for ATP also did worse than its counterpart that did not use interpolated data. This may be attributed to added noise within the data from subjects for which the model does not necessarily know the label of. If not used smartly, interpolated data can hinder the model through confusing samples, rather than improve performance. Adding subjects in the intermediate t-score range forces the model to make a decision on classification, for which cleaner data is better.

My models did contain some subtle biases that should be noted within this section. Gridsearches were conducted on a fixed seed, of which testing models were also configured. While minor, this is still considered a bias towards the arbitrarily selected seed.

Furthermore, it should be noted that the simplified binary classifiers cannot be accepted as predictive models. They do not include intermediary data, and therefore do not provide any decision making for actual prediction of ATP or HYP. Simplified models evaluate signal and extrema from the training process and seeing if it helps. However, we want to validate on all subjects and see what happens to them in a testing phase for actual prediction.

Given that neural networks exhibit the ability to learn complex, non-linear features, there are a variety of factors to be explored to improve network performance. Learning appropriate hyperparameters could be accomplished via more robust, fine-tuned gridsearches. Additional sampling approaches featuring some way to involve the minority class may also help. Adjusting model complexity and adjusting imputation measures such as SMOTE and the autoencoders may also provide more opportunities with minority class samples and balancing class weights. Exploring different data options and acquiring a higher number of valid samples can also improve network performance. Additional pretraining with a separate population may allow the model to adapt for singleton 1-year features.

I substituted the most appropriate parcellation I had in order to use the morphometric features available, but further studies may look to expand on my results by computing other parcellations, e.g. Gordon parcellation, and evaluating the performance of models that make use of this new data [11].

The use of CT may also call for review. Previous studies have shown links between SA and neurodevelopmental disorders (primarily ASD), but CT has only been shown to be predictive of symptom severity particularly in ASD [28]. CT was chosen due to previously observed differences in brain morphometry between patients With ASD and healthy individuals [35]. It may be the case that models only using SA and related features may perform at a higher level. There are also other features available e.g. intra-cranial cavity volume (ICV), surface complexity index (SCI), or derived extra-axial cerebrospinal fluid (EA-CSF).

Heterogeneity of the brain at 1 year of age may also affect results, as individual developmental differences may mask biomarkers and create noise within the data. Models may err in choosing boundaries for atypicality in features. Noise from processing steps such as interpolation of data via autoencoders may also contribute to a difficulty for identifying atypicality in the current models.

To end, I find evidence of a signal in models using morphometric features for classification of risk associate with symptoms related to ADHD. There is evidence that these types of models can correctly predict development of symptomps in HR which may lead to further diagnosis via BASC-2 or BRIEF examinations for ADHD or other related neurodevelopmental disorders. These "rough" models are not robust enough to be considered in clinical evaluation, but suggest further exploration can lead to classification networks capable of early prediction of attention problems and hyperactivity at 1 year of age from brain morphometry.

5 Conclusion

In this study I introduced the lack of predictive models associated with symptoms of ADHD. I explored why longitudinal studies are necessary for furnishing data, but often faced with issues of attrition and reduced sample sizes. Extrapolating ideas regarding longitudinal studies and quantification of developmental trajectories from [9], I combine the correlation of neurodevelopmental disorders with morphometric features [5] to create models that attempt to predict symptoms related to diagnosis of ADHD. When faced with lowered data counts for reasons explored in [22] I use data imputation techniques discussed in [14] via creation of autoencoders to generate singleton data. During my study, I gained a knowledge of machine learning, deep learning, and generative models, and reviewed a parcellation technique for dimensionality reduction of morphometric features. I compare my developed models based on their use of training data and classification groupings, but did not achieve performance worthy of clinical consideration. The reasons for this shortcoming amount to ideas involving network optimization and structure, lack of available training data, lack of HR

subjects, potential noise in model training, and biases introduced in the process. I also considered other factors that could be expanded upon in the future.

This study concludes on the findings that the possibility of clinical-strength classifier may exist for symptoms related to diagnosis of ADHD. I have contributed to this area research by showing a predictive signal among morphometric features, ATP, and HYP may exist. Further research can expand or explore many possibilities based on current research in the field and this work.

5.1 Further Work

In Section 2.1.3 I discuss the Destrieux parcellaion of the adult human brain. Further work is required to generate appropriate infant brain parcellations and evaluate their performance classification models for symptoms of neurodevelopmental disorders. It is also the case that this study could be reperformed with other parcellations like that developed in [11]. In Section 4 I note several possible options for model optimization due to configuration choices explained throughout Section 2 but more thoroughly examined in Section 2.2.4. A threshold within binary classifiers (mentioned in Section 2.2.4 may also be explored for symptoms related to ADHD via methods similarly explored in [29] for models predicting ASD.

Ultimately, further research is needed to determine whether a combination of brain morphometric measurements, introduction of new data, or additional classifier optimization will lead to accurate prediction of symptoms for clinical use.

References

[1] N. V. Chawla et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. DOI: 10.1613/jair.953. URL: https://doi.org/10.1613$%5C%$2Fjair.953.

[2] François Chollet. *keras*. 2015. URL: https://github.com/fchollet/keras.

[3] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (Dec. 1989), pp. 303–314. DOI: 10.1007/bf02551274. URL: https://doi.org/10.1007$%5C%$2Fbf02551274.

[4] Christophe Destrieux et al. "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature". In: *NeuroImage* 53.1 (Oct. 2010), pp. 1–15. DOI: 10.1016/j.neuroimage.2010.06.010. URL: https://doi.org/10.1016%5C%2Fj.neuroimage.2010.06.010.

[5] "Development of cortical shape in the human brain from 6 to 24 months of age via a novel measure of shape complexity". In: *NeuroImage* 135.15 (June 2016), pp. 163–176. DOI: 10.1016/j.neuroimage.2016.04.053. URL: https://doi.org/10.1016/j.neuroimage.2016.04.053.

[6] *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, 2013. DOI: doi.org/10.1176/appi.books.9780890425596. URL: doi.org/10.1176/appi.books.9780890425596.

[7] Centers for Disease Control and Prevention. *Attention-Deficit / Hyperactivity Disorder (ADHD)*. 2020. URL: https://www.cdc.gov/ncbddd/adhd/index.html (visited on 09/24/2020).

[8] M.W Gardner and S.R Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". In: *Atmospheric Environment* 32.14-15 (Aug. 1998), pp. 2627–2636. DOI: 10.1016/s1352-2310(97)00447-0. URL: https://doi.org/10.1016$%5C%$2Fs1352-2310$%5C%$2897$%5C%$2900447-0.

[9] John H. Gilmore et al. "Longitudinal Development of Cortical and Subcortical Gray Matter from Birth to 2 Years". In: *Cerebral Cortex* 22.11 (Nov. 2011), pp. 2478–2485. ISSN: 1047-3211. DOI: 10.1093/cercor/bhr327. eprint: https://academic.oup.com/cercor/article-pdf/22/11/2478/1044202/bhr327.pdf. URL: https://doi.org/10.1093/cercor/bhr327.

[10] Jessica B. Girault et al. "Cortical Structure and Cognition in Infants and Toddlers". In: *Cerebral Cortex* 30.2 (July 2019), pp. 786–800. DOI: 10.1093/cercor/bhz126. URL: https://doi.org/10.1093/cercor/bhz126.

[11] Evan M. Gordon et al. "Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations". In: *Cerebral Cortex* 26.1 (Oct. 2014), pp. 288–303. DOI: 10.1093/cercor/bhu239. URL: https://doi.org/10.1093$%5C%$2Fcercor$%5C%$2Fbhu239.

[12] Chiaki Hasegawa et al. "Developmental Trajectory of Infant Brain Signal Variability: A Longitudinal Pilot Study". In: *Frontiers in Neuroscience* 12 (2018), p. 566. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00566. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00566.

[13] Heather Cody Hazlett et al. "Early brain development in infants at high risk for autism spectrum disorder". In: *Nature* 542.7641 (Feb. 2017), pp. 348–351. DOI: 10.1038/nature21369. URL: https://doi.org/10.1038%5C%2Fnature21369.

[14] Yoonmi Hong et al. "Longitudinal Prediction of Infant Diffusion MRI Data via Graph Convolutional Adversarial Networks". In: *IEEE Transactions on Medical Imaging* PP (Apr. 2019), pp. 1–1. DOI: 10.1109/TMI.2019.2911203.

[15] J J Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. ISSN: 0027-8424. DOI: 10.1073/pnas.79.8.2554. eprint: https://www.pnas.org/content/79/8/2554.full.pdf. URL: https://www.pnas.org/content/79/8/2554.

[16] Daniel Hsu, Sham M. Kakade, and Tong Zhang. *A Spectral Algorithm for Learning Hidden Markov Models*. 2012. arXiv: 0811.4413 [cs.LG].

[17] Kelly Pizzitola Jarratt, Cynthia A. Riccio, and Becky M. Siekierski. "Assessment of Attention Deficit Hyperactivity Disorder (ADHD) Using the BASC and BRIEF". In: *Applied Neuropsychology* 12.2 (2005), pp. 83–93. DOI: 10.1207/s15324826an1202$_$4. URL: https://doi.org/10.1207/s15324826an1202_4.

[18] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].

[19] Rebecca C. Knickmeyer et al. "A Structural MRI Study of Human Brain Development from Birth to 2 Years". In: *JNeurosci* 28.47 (Nov. 2008), pp. 12176–12182. DOI: 10.1523/JNEUROSCI.3479-08.2008. URL: https://doi.org/10.1523/JNEUROSCI.3479-08.2008.

[20] Ben D. Knoble. "Early Prediction of Autism Spectrum Disorder at 12 months of Age from Brain Morphometry". Apr. 2020. DOI: 10.17615/f2mz-mz17. URL: https://doi.org/10.17615/f2mz-mz17.

[21] Mark A. Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE Journal* 37.2 (1991), pp. 233–243. DOI: 10.1002/aic.690370209. eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690370209. URL: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370209.

[22] NM Laird. "Missing data in longitudinal studies". In: *Statistics in medicine* 7.1-2 (1988), pp. 305–315. ISSN: 0277-6715. DOI: 10.1002/sim.4780070131. URL: https://doi.org/10.1002/sim.4780070131.

[23] Alexander LeNail. "NN-SVG: Publication-Ready Neural Network Architecture Schematics". In: *Journal of Open Source Software* 4.33 (Jan. 2019), p. 747. DOI: 10.21105/joss.00747. URL: https://doi.org/10.21105%5C%2Fjoss.00747.

[24] Panos P. Markopoulos et al. "Efficient L1-Norm Principal-Component Analysis via Bit Flipping". In: *IEEE Transactions on Signal Processing* 65.16 (Aug. 2017), pp. 4252–4264. ISSN: 1941-0476. DOI: 10.1109/tsp.2017.2708023. URL: http://dx.doi.org/10.1109/TSP.2017.2708023.

[25] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[26] Tyler H Matta, John C Flournoy, and Michelle L Byrne. "Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies". In: *Developmental cognitive neuroscience* 33 (Oct. 2018), pp. 83–98. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2017.10.001. URL: https://europepmc.org/articles/PMC6969275.

[27] H. Mhaskar, Q. Liao, and T. Poggio. "When and Why Are Deep Networks Better Than Shallow Ones?" In: *AAAI*. 2017.

[28] Elaheh Moradi et al. "Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data". English. In: *NeuroImage* 144.A (2017). EXT="Tohka, Jussi", pp. 128–141. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2016.09.049.

[29] Mahmoud Mostapha. "Learning from Complex Neuroimaging Datasets". PhD thesis. University of North Carolina at Chapel Hill, 2020.

[30] IBIS Network. *IBIS*. 2019. URL: https://www.ibis-network.org/index.html (visited on 09/15/2020).

[31] Milad Zafar Nezhad et al. "A Predictive Approach Using Deep Feature Learning for Electronic Medical Records: A Comparative Study". In: *ArXiv* abs/1801.02961 (2018).

[32] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[33] Cecil R. Reynolds and Randy W. Kamphaus. *Teacher Rating Scales Report Behavior Assessment System for Children, Second Edition*. Dec. 2008.

[34] Fiona Richardson. "A young mind in a growing brain. Jerome Kagan, & Elinore Chapman Herschkowitz. Lawrence Erlbaum, Mahwah, NJ, 2005. ISBN 080585309X". In: *Infant and Child Development* 15.5 (2006), pp. 555–557. DOI: 10.1002/icd.487. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/icd.487. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/icd.487.

[35] Daan van Rooij et al. “Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group”. In: *American Journal of Psychiatry* 175.4 (2018). PMID: 29145754, pp. 359–369. DOI: 10.1176/appi.ajp.2017.17010100. eprint: https://doi.org/10.1176/appi.ajp.2017.17010100. URL: https://doi.org/10.1176/appi.ajp.2017.17010100.

Abbreviations

ACC	accuracy
ADHD	Attention-deficit/Hyperactivity disorder
ASD	autism spectrum disorder
ATP	attention problems
BASC-2	Behavior Assessment System for Children Second Edition
BRIEF	Behavior Rating Inventory of Executive Function
CDC	Centers for Disease Control
CT	cortical thickness
DSM-5	American Psychiatric Association's Diagnostic and Statistical Manual, Fifth edition
EA-CSF	extra-axial cerebrospinal fluid
FN	false negatives
FP	false positives
GA	gestational age
HR	high risk
HYP	hyperactivity
IBIS	Infant Brain Imaging Study
ICV	intra-cranial cavity volume
LR	low risk
MLP	multi-layer perceptron
MRI	magnetic resonance imaging
NPV	negative predictive value
PCA	principle component analysis
PPV	positive predictive value
ROI	regions of interest
SA	surface area
SCI	surface complexity index

SD	standard deviation(s)
SEN	sensitivity
SMOTE	Synthetic Minority Over-sampling Technique
SPC	specificity
TN	true negatives
TP	true positives
UNC	University of North Carolina