Temporal Order-based First-Take-All Hashing for Fast Attention-Deficit-Hyperactive-Disorder Detection

Hao Hu
Department of Computer Science
University of Central Florida
Orlando, Florida
hao_hu@knights.ucf.edu

Joey Velez-Ginorio
Department of Computer Science
University of Central Florida
Orlando, Florida
joeyVelez@knights.ucf.edu

Guo-Jun Qi*
Department of Computer Science
University of Central Florida
Orlando, Florida
guojun.qi@ucf.edu

ABSTRACT

Attention Deficit Hyperactive Disorder (ADHD) is one of the most common childhood disorders and can continue through adolescence and adulthood. Although the root cause of the problem still remains unknown, recent advancements in brain imaging technology reveal there exists differences between neural activities of Typically Developing Children (TDC) and ADHD subjects. Inspired by this, we propose a novel First-Take-All (FTA) hashing framework to investigate the problem of fast ADHD subjects detection through the fMRI time-series of neuron activities. By hashing time courses from regions of interests (ROIs) in the brain into fixed-size hash codes, FTA can compactly encode the temporal order differences between the neural activity patterns that are key to distinguish TDC and ADHD subjects. Such patterns can be directly learned via minimizing the training loss incurred by the generated FTA codes. By conducting similarity search on the resultant FTA codes, data-driven ADHD detection can be achieved in an efficient fashion. The experiments' results on real-world ADHD detection benchmarks demonstrate the FTA can outperform the state-of-the-art baselines using only neural activity time series without any phenotypic information.

Keywords

ADHD detection; First-Take-All; Time series hashing

1. INTRODUCTION

Recent advancements in brain imaging technology, one of the greatest efforts in Neuroscience, aim to uncover features unique to certain neurophysiological phenomena [4, 20]. The intuition is that the neural activity pattern of a healthy human subject ought to appear different from that of a patient suffering from some neural disorder, such as Attention Deficit Hyperactive Disorder (ADHD) and Alzheimer's disease.

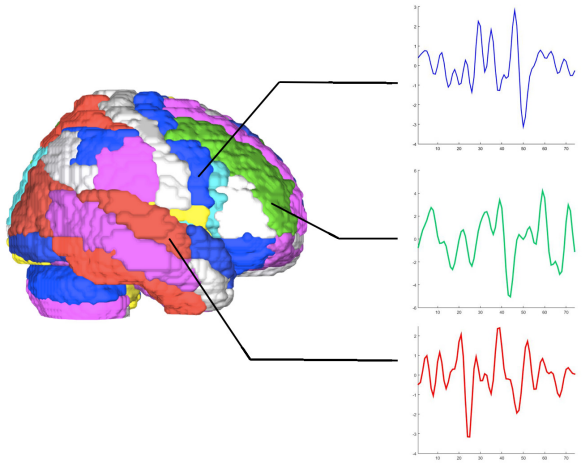


Figure 1: ROIs (AAL[29]) of a human brain and their fMRI time courses. FTA hashes time courses into fixed-size hash codes by encoding temporal order differences between latent patterns among them.

Categorized alongside other neurodevelopmental disorders, ADHD is one of the most common brain diseases. It manifests early and is typically first diagnosed in one's childhood. The predominant effects consist of sustained difficulty in maintaining focus, often offering difficulties assimilating at school, at home, or within the community. Fortunately, despite lacking a cure, brain imaging technology provides an opportunity to timely observe and diagnose ADHD[6, 7].

Several methods for recording sequences of neural activity from the brain exist. Generally, they range from invasive to non-invasive procedures. Whilst the data from invasive techniques such as Electrocorticography (ECoG) or multi-electrode arrays possess higher quality data [24], the opportunity to collect such data seldom arises. By nature of the procedure, it is much more practical for researchers and medical practitioners to opt for non-invasive techniques such as functional Magnetic Resonance Imaging (fMRI). fMRI indirectly measures changes in neural activity by detecting changes in blood flow caused by increased activations of neurons during specific tasks (or resting-state conditions) [27,

*Corresponding author

KDD '16, August 13-17, 2016, San Francisco, CA, USA
 ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939774

9]. In this context, fMRI time-courses [1] offer a feature-rich representation of high level functional organization in the brain (Figure 1).

Considering the representations, we aim to exploit its rich nature for the task of ADHD diagnosis as a pattern classification problem [22]. Utilizing the structure of the resting-state fMRI time-courses, we generate hash codes encoding the temporal structure of the data. These hash codes can then be compared to detect similarities and differences between the fMRI time-courses of healthy patients versus those diagnosed with ADHD. It is known that neural connections exist whether or not regions of the brain are functionally active, hence forth the resting-state fMRI provides a controlled dataset to test for fundamental differences in functional neural networks in the brain.

Tasked with hashing time-series data, our approach centers on fast detection based on retrieval of the similar disorder patterns from a database of brain neural imaging activities. Specifically, we propose the First-Take-All (FTA) hashing method for encoding varied-length fMRI sequences into fixed-size hash codes. The problem of fast matching similar fMRI sequences boils down to a fast search of similar hash codes based on their Hamming distances that can be calculated efficiently.

Specifically, the algorithm first projects an input sequence of varied length onto different subspaces, each representing a sequence of latent patterns. After encoding the temporal order of these patterns to hash the fMRI time-courses, the pattern that appears first among a selection of patterns is used to index the time-course [2]. This scheme can yield a compact encoding of temporal relations between the selected patterns that really matter in distinguishing between healthy individuals and those diagnosed with the ADHD. The optimal pattern projections will be learned to result in the hash codes that minimize the diagnosis errors. In this way, FTA allows for not only high detection rates, but a scalable solution for detecting fMRI sequences. Since the projections are learned as opposed to randomly generated, the solution scales well with large-scale input fMRI sequences using compact hash codes.

Suitably, the objective of this paper is to provide a scalable and efficient [18, 28] solution to the problem of detecting neurodevelopmental disorders (specifically ADHD) via fMRI time-courses. Doing so also reverberates to improved success in brain imaging technologies. The proposed temporal order-based hashing algorithms are much more generic, providing a new framework for fast matching and detection to other forms of time-series data. In this paper, the method has implications on functional neural analysis[14, 17]. Being able to reliably infer causal relationships between brain structures and functions presents an interesting opportunity for further investigations, the range of which include areas of classification outside neurodevelopmental disorders [13]. Ultimately, learning the projections to hash a time-series space efficiently provides important practicality to the design.

[1]In medical imaging terminology, a "time course" refers to an obtained sequence for an imaged area. In this paper, we will use this term interchangeably with "time series" when there is no confusion in the context.

[2]Without loss of generality, we can also designate the second or the third appearing pattern or so on to hash a fMRI sequence. As a convention, we choose the first-appearing pattern in this paper.

The contributions of this paper are:

1. We propose a novel FTA hashing algorithm to hash time series with varied length into fixed-size hash codes by encoding the temporal order of the latent patterns inside the time series.

2. In order to acquire the optimal projections, we formulate it as a learning problem whose training loss can be minimized in an efficient fashion.

3. We perform extensive experiment studies on benchmarks of ADHD detection and demonstrate the superior performance of the proposed FTA hashing with several evaluation metrics.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. The ADHD detection paradigm including FTA hashing algorithm is introduced and discussed in Section 3. Section 4 includes the learning algorithm for searching optimal projections. Experiments and performance studies are presented in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

In this section, we review several related topics pertinent to the task of ADHD detection. Among these, the overall theme fosters support for classification analysis of fMRI time-courses.

It is known that multivariate data mining and machine learning methods have been used to approach the classification of fMRI data[13, 7]. Similar to the approach in the paper, multivariate methods presume from a core tenet of neuroscience that neural data encodes itself across larger functional regions in the brain. Intriguingly, the motivation behind utilizing multiple ROIs in test and training exists within other works as well; as ROI selection can be viewed as a form of feature selection [13, 25]. Similar works [7] also suggest the widespread adoption of multivariate techniques including Support Vector Machines (SVMs) [5] and Linear Discriminant Analysis (LDA) [3] in spite of the traditional uni-variate alternatives. Fueling this shift was a dissatisfaction with how univariate models rely exclusively on the information contained in time-courses contained in individual voxels[7].

In tune with our method's focus on preserving temporal structure is the Dynamic Time Warping (DTW) to find similar patterns between two time series [2]. The idea is to create a sequence alignment algorithm that preserves and efficiently discovers knowledge from potentially large data archives. The approach used in this paper (FTA) maintains similar motivations, insofar as it aims to solve the task of presenting a model for efficient and scalable knowledge discovery in the domain of neural data.

The related works presented offer several intriguing points of support to our investigation. Foremost, the premise has been set for the intuitions behind fMRI analysis for ADHD detection[7]. For example, [10] extracts features from fMRI time courses to improve the ADHD detection rate. Meanwhile, the method proposed by [12] combines both unsupervised and supervised algorithms and achieves best performance in ADHD-200 Global Competition. This reliably shows that the problem of knowledge discovery in brain imaging can be improved through utilization of Machine

Learning models. In addition, other efforts suggest that a detection scheme for analysis of fMRI time-courses can be expanded to other neural datasets and disorders[10, 7]. This offers an expanded utility to the model presented in this paper, in which the FTA can be used on other problems within Neuroscience and a host of other topic areas, all whilst preserving efficiency. Lastly, DTW offers an element of distinction between former methods in time-series analysis versus those of which focus specifically on preserving temporal order during encoding of time-series data [2]. In doing so, FTA provides clear advantages over methods which eschew these temporal features [16].

3. FIRST-TAKE-ALL: ADHD DETECTION BY TIME-SERIES HASHING

In this paper, we introduce a novel hashing-based paradigm to automatically identify ADHD subjects. The structure of our method can be summarized as following: First, brain atlases containing a number of Regions of Interest (ROIs) will be constructed and corresponding time courses of each ROI will be extracted based on the ADHD subjects' resting state fMRI data of the brain. Then, we propose a new temporal order-preserving hashing algorithm called First-Take-All to hash time courses into binary sequences. With those sequences, we compare the distance (similarity) between them to determine whether a patient is an ADHD subject.

3.1 Brain Atlas Construction and Time Course Extraction

In brain neuroanatomy, many approaches, such as automated anatomical labeling (AAL) [29], Eickhoff-Zilles (EZ) [11], Talairach and Tournoux (TT) [21] and Harvard-Oxford (HO) [15], have been proposed to construct brain atlases by using structural anatomic or functional information. After that, voxels in the regions that have structural or functional similarities will be grouped into Regions of Interest (ROIs) and the time courses of these ROIs can be extracted from the voxels of the subjects' resting state fMRI data. Since the brain atlas construction and time-course extraction are not the focus of this paper, we will employ the existing brain atlases pre-constructed by the neuroanatomy community. The details will be presented in Section 5.

With time courses of ROIs extracted, every subject can be mapped to a unique vector (multivariate) sequence which describes the brain activities of that subject. For example, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$ can be a subject's time-courses of ROIs, where each $\mathbf{x}_t \in \mathbb{R}^D, t = 1, \cdots, T$ represents the brain activities of that subject in D different ROIs. For our convenience, we will refer to Time Courses of ROIs as **TC**s in the rest of this paper for short.

3.2 Temporal Order-Preseving Hashing

Now we propose a hashing algorithm for time-series data which can map a TC into a fixed-size hash codes regardless of its original length. It can be roughly divided into 2 parts. First, a TC will be projected into several subspaces to produce a set of projection sequences. Then, we generate hash codes for entire TC by conducting an operation called First-Take-All (**FTA**) on those projection sequences.

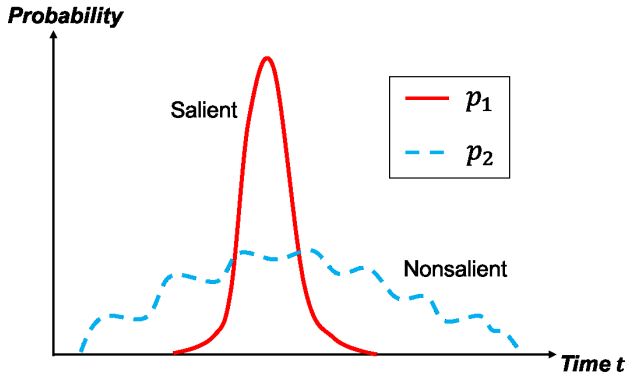


Figure 2: Comparison between projections generated by salient and nonsalient patterns. The red solid line represents the projection of a salient pattern with a small variance, while the blue dotted line represents the projection of a non-salient pattern with a large variance over the time axis.

Sequence Projection

Consider a TC $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$ of length T described in Section 3.1. The first step is to project $\mathbf{X}$ into several subspaces defined by an optimized projection matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K] \in \mathbb{R}^{D \times K}$. Each $\mathbf{w}_k \in \mathbb{R}^D$ is a projection vector taken from $\mathbf{W}$ that generates a sequence $\mathbf{s}_k = \mathbf{w}_k^\intercal \mathbf{X}$ for $k \in \{1, \cdots, K\}$. The way to find the optimal $\mathbf{W}$ will be given in section 4.

Intuitively, each sequence $\mathbf{s}_k$ represents the score over the occurrence of a latent pattern[3] $\mathbf{w_k}$, and any TC is composed of a sequence of temporally-ordered patterns. The **orders** of certain **unknown** neural activation patterns often matter in ADHD detection. Thus, we seek to find these relevant patterns as well as compactly represent their orders in a hash code space, where the similarity between TCs can be directly computed by their Hamming distance.

To model the temporal order of patterns, first we need to locate the moment they appear. Here we use softmax to compute the probability that a pattern k appears at time t:

$$p_{k,t} = \frac{\exp(\mathbf{w}_k^\intercal \mathbf{x}_t)}{\sum_{t'=1}^{T} \exp(\mathbf{w}_k^\intercal \mathbf{x}_{t'})} \tag{1}$$

Let $\mathbf{u} = [1/T, 2/T, \cdots, t/T, \cdots, 1]^\intercal$ be the normalized timescale, each entry of which denotes a relative time moment on the range of $[0, 1]$ in an input sequence of length T. Then, the expected moment m_k that the pattern k appears can be calculated as

$$m_k = \mathbb{E}_{t \sim p_{k,t}} \left[\frac{t}{T} \right] = \sum_{t=1}^{T} \frac{t \cdot p_{k,t}}{T} = \mathbf{u}^\intercal \mathbf{p}_k \tag{2}$$

where $\mathbf{p}_k = [p_{k,1}, \cdots, p_{k,T}]^\intercal$ is a vector containing the probability of pattern k appearing at each moment.

Note that a pattern related with ADHD detection usually corresponds to a salient pattern of neural activation in brain ROIs. Thus, we expect that it should have a sharp appearance in the projection sequence such as the p_1 shown in

[3]These patterns are latent because they are unlabeled.

figure 2. Accordingly, we propose to minimize the **variance of pattern occurrence**

$$v_k = \mathrm{Var}_{t\sim p_{k,t}}\left[\frac{t}{T}\right] = \sum_{t=1}^{T}\frac{(t-m_k)^2\cdot p_{k,t}}{T^2} \quad (3)$$

together with the other criteria to learn the projection matrix $\mathbf{W}$ in Section 4. This regularization term is more likely to generate a salient pattern that really matters in detecting ADHD.

First-Take-All Temporal-Order Comparison

Now putting the expected appearing moments of K patterns into $\mathbf{m} = [m_1, \cdots, m_K]$, we wish to develop an ordinal hashing algorithm directly encoding their temporal order for a TC. Specifically, we perform a First-Take-All (FTA) comparison to rank the patterns by their temporal order – the pattern whose expected appearing moment comes first wins the FTA comparison, and its index is used to hash the entire TC.

For example, when we have two projected sequences (i.e., $K = 2$), FTA simply encodes the pairwise order between two corresponding patterns. When $K > 2$, FTA makes a higher-order comparison to decide which pattern appears first. Note that the output FTA hash code is not binary; instead it is a K-ary code. For this reason, we call K the FTA base.

For a pairwise FTA comparison involving only two patterns, knowing the first coming pattern completely encodes the temporal order between these two patterns. This can be generalized to high-order comparison if more than two patterns are involved for choosing the first-coming pattern. Such a high-order FTA comparison could generate more compact code to distinguish between different types of TCs. For example, suppose there are three types of TCs have different orders of patterns 1 – 2 – 3 – 4, 1 – 3 – 2 – 4 and 1 – 4 – 3 – 2, respectively. Then an effective FTA comparison only needs to make a order-3 comparison between the last three patterns 2, 3 and 4, which will output the FTA code 2, 3 and 4 to distinguish these three types of TCs. In this case, there is no need to make comprehensive comparisons between all possible pairs of patterns.

However, the patterns whose orders matter in classifying different types of TCs are unknown a-priori. A suitable group of patterns must be learned so that the same type of TCs will have similar pattern orders. We will discuss the detail about the learning of these patterns in Section 4.

Mathematically, the index of the first-appearing pattern can be expressed as

$$\mathbf{h} = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^\intercal \mathbf{m} = \boldsymbol{\theta}^\intercal[\mathbf{u}^\intercal\mathbf{p}_1|\cdots|\mathbf{u}^\intercal\mathbf{p}_K] = \mathbf{u}^\intercal\mathbf{P}\boldsymbol{\theta} \quad (4)$$

where $\boldsymbol{\theta} \in \{0,1\}^K, \mathbf{1}^\intercal\boldsymbol{\theta} = 1$ and $\mathbf{h}$ is an 1-of-K indicator of the FTA winner – its unique nonzero entry is indexed by the first-appearing pattern in the input TC, and $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_K]$.

An algorithmic description for the entire FTA hashing procedure is shown in Algorithm 1. Multiple FTA codes can be generated with a set of projection matrices. The code length L in the algorithm represents the number of hash codes generated for a TC.

Figure 3 illustrates the FTA comparison when K is set to 3. Here p_1, p_2 and p_3 in Figure 3 are the probability of each pattern appearing over the time axis t. From them, we can get their expected pattern appearing moments m_1, m_2 and m_3, which are approximately shown in the figure, where we have $m_2 < m_3 < m_1$. By the FTA comparison, we choose the index of m_2 as the hash code for the TC, which is 2.

Algorithm 1 First-Take-All Hashing

1: **Input:** TC $\mathbf{X}$, code length L, a set of projection matrices $\{\mathbf{W}_i\}_{i=1}^L$
2: **Initialize:** $b \leftarrow$ empty sequence
3: **for** $i = 1$ to L **do**
4: $\quad \mathbf{S} = \mathbf{W}_i^\intercal\mathbf{X}$
5: $\quad$ for each row $\mathbf{s}_k$ of $\mathbf{S}$, calculate m_k through Eq.(1) and Eq. (2)
6: $\quad k^* \leftarrow \arg\min_{1\le j\le K} m_k$.
7: $\quad b \leftarrow bk^*$ (concatenation)
8: **end for**
9: **return** b

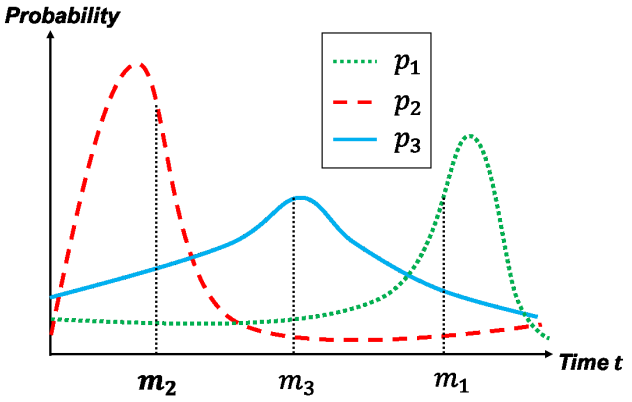


Figure 3: Illustration for First-Take-All Temporal-Order comparison when K is set to 3

Before the end of this section, let us analyze the computational complexity of hashing a sequence by FTA. First, it costs $O(TDK)$ to apply the projection matrix to an input TC $\mathbf{X}$. Then finding the expected moments of K projected sequences costs $O(TK)$. It also costs $O(K)$ to find the first-appearing pattern which wins FTA comparison out of K candidates. Hence, FTA totally costs $O(TDK)$ to hash an input TC up to a constant factor.

4. LEARNING OPTIMAL PROJECTIONS

A learning algorithm to find out the optimal projections $\mathbf{W}$ will be presented in this section, including the formulation of the optimizing problem and its efficient solution.

4.1 Training Loss

Given a TC $\mathbf{X}$ and the expected appearing moments of K patterns $\mathbf{m} = [m_1, \cdots, m_K]$, which is acquired from a fixed $\mathbf{W}$. Then we can apply the following softmin to calculate the probability that the kth pattern will appear first:

$$h_k \triangleq P(\text{pattern } k \text{ comes first}|\mathbf{X}) = \frac{\exp(-m_k)}{\sum_{k'=1}^{K}\exp(-m_{k'})} \quad (5)$$

The smaller the m_k, the more likely the pattern k will appear first.

Now consider a pair of TCs $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$, along with their label s_i and s_j. We hope that through our learning algorithm, the resultant FTA hash codes can reflect the label similarity between two TCs. In other word, when $s_i = s_j$, there is a greater chance that the same pattern will appear first in both $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$; otherwise, different patterns will appear first if $s_i \neq s_j$ if $s_i \neq s_j$.

Mathematically, it is easy to see that $h_k^{(i)} h_k^{(j)}$ is the probability that the kth pattern will appear first in both $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$. Suppose h^{ij} represents the probability that the same pattern will appear first in both TCs. It can be computed as by summing up over all patterns

$$h^{ij} = \sum_{k=1}^{K} h_k^{(i)} h_k^{(j)} \tag{6}$$

Our goal is to maximize h^{ij} when $s_i = s_j$ and to minimize it when $s_i \neq s_j$. This results in the following objective function

$$O^{ij} = (1 - h^{ij})^{s_{ij}} (h^{ij})^{(1-s_{ij})} \tag{7}$$

Here, $s_{ij} = 1$ *i.f.f.* $s_i = s_j$; and $s_{ij} = 0$ otherwise. We wish to minimize it for all TC pairs. Suppose the training set is $\mathcal{T} = \{\mathbf{X}^{(i)}, s_i\}_{i=1}^{N}$ with N TCs. Then the total logarithmic training loss over $\mathcal{T}$ becomes

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{N} \left[s_{ij} \log\left(1 - h^{ij}\right) + (1 - s_{ij}) \log\left(h^{ij}\right) \right] \tag{8}$$

Minimizing it can minimize the pairwise diagnosis errors on the training set incurred by the resultant FTA hash codes.

4.2 Projection Orthogonality

In addition to the minimization of training loss, it is worth mentioning that the redundancy between the learned patterns also affects the FTA hashing performance. With a set of redundant patterns learned, their projection sequences would be highly correlated or even identical to one another. This could reduce the degree of the temporal order being distinguished between different patterns. In this case, a smaller perturbation or local warping would change the temporal orders significantly, thereby degenerating the FTA's performance in presence of noises.

To improve the resiliency of FTA against perturbations or noises on TCs, we wish to reduce the redundancy between patterns by minimizing the following normalized inner products between projection vectors

$$\Omega = \sum_{k \neq k'=1}^{K} \left(\frac{\mathbf{w}_k^\intercal \mathbf{w}_{k'}}{\|\mathbf{w}_k\| \|\mathbf{w}_{k'}\|} \right)^2 \tag{9}$$

Clearly, minimizing it can make the learned projection vectors as orthogonal to each other as possible, thereby minimizing the redundancy between the corresponding patterns [4].

4.3 Putting Together

In addition to the training loss (8) and the projection orthogonality (9), we also consider to minimize the variance of pattern occurrences as shown in Eq. (3). This can be expressed as the following total variance over the training set:

$$\mathcal{V} = \sum_{i,k=1}^{N,K} v_k^{(i)}$$

where $v_k^{(i)}$ is the occurrence variance of pattern k in the ith TC of training set. As aforementioned, minimizing the variance of pattern occurrences can generate salient patterns for ADHD diagnosis.

Then putting them together, we can define the following minimization problem to learn the projection matrix $\mathbf{W}$

$$\begin{aligned} \min_{\mathbf{W}} \mathcal{F} \triangleq \mathcal{L} \quad & \cdots \text{ training loss} \\ + \gamma\, \Omega \quad & \cdots \text{ Projection Orthogonality} \\ + \eta\, \mathcal{V} \quad & \cdots \text{ Variance of pattern occurences} \end{aligned} \tag{10}$$

where two positive coefficients γ and η are two hyper parameters that control the contributions of the projection orthogonality and the minimization of pattern occurrence variance.

Algorithm 2 Learning Optimal Projections

1: **Input:** training TC set $\chi = \{\mathbf{X}^{(i)}, s_i\}_{i=1}^{N}$, K, learning rate α
2: **Initialize:** Randomly initialize $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$, $\mathbf{w}_k \in \mathbb{R}^{D \times 1}, k = 1, \cdots, K$
3: **repeat**
4: Select training pair $\mathbf{X}^{(i)}$, $\mathbf{X}^{(j)}$.
5: $P^{(i)} = [p_{k,t}^{(i)}]_{K \times T}$, $P^{(j)} = [p_{k,t}^{(j)}]_{K \times T}$, calculate $p_{k,t}^{(i)}$ and $p_{k,t}^{(j)}$ based on Eq.(1)
6: for each $\mathbf{w}_k, k = 1, \cdots, K$, compute $\frac{\partial \mathcal{F}}{\partial \mathbf{w}_k}$ with $p_{k,t}^{(i)}$ and $p_{k,t}^{(j)}$ based on Eq.(12), (13), (14), (15), (16), (17), (18), (19).
7: $\nabla_{\mathbf{W}} \mathcal{F} \leftarrow [\frac{\partial \mathcal{F}}{\partial \mathbf{w}_1}, \cdots, \frac{\partial \mathcal{F}}{\partial \mathbf{w}_K}]$
8: $\mathbf{W} \leftarrow \mathbf{W} - \alpha \nabla_{\mathbf{W}} \mathcal{F}$
9: **until** Convergence.

4.4 Optimization

We adopt the stochastic gradient descent method to minimize $\mathcal{F}$ as to find the optimal projection $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K] \in \mathbb{R}^{D \times K}$. For each training iteration, we randomly pick up a pair of TCs and their labels and calculate the gradient $\nabla_{\mathbf{W}} \mathcal{F}$. Since $\mathcal{F}$ is differentiable w.r.t. $\mathbf{W}$, it allow us to calculate $\nabla_{\mathbf{W}} \mathcal{F}$, leading to an efficient learning procedure. All equations needed to calculate $\nabla_{\mathbf{W}} \mathcal{F}$ can be found in Appendix A and the entire optimization procedure is described in Algorithm 2.

The above paragraph depicts the training algorithm for a projection matrix resulting in one FTA hash code. The above learning algorithm can be used as a subroutine in a standard ensemble method like AdaBoost. This will yield multiple FTA codes for a TC. The similarity between TCs can be computed with the Hamming distance between their concatenated FTA codes. Then, ADHD can be fast detected by retrieving the similar TCs from a labeled database.

5. EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed method by conducting experiments on ADHD 200

[4]The projection orthogonality can also be imposed as a hard constraint in the optimization problem. However, by experiments, we found that the performance is more stable by posing it as a soft term that penalizes the projection redundancy in the objective function.

dataset, a dataset developed for ADHD detection. First we give a brief introduction on ADHD dataset. Then we discuss the experiment setting. We compare the proposed method with several supervised and unsupervised baselines with different evaluation metrics. Finally, we study the impact of the hyper-parameters K and L on the performance.

5.1 Datasets and Background

We evaluate the proposed FTA approach on ADHD-200 dataset. ADHD-200 was initially prepared by ADHD-200 Consortium[23] for the ADHD-200 Global Competition, a competition that aimed to improve the understanding of the neural basis of ADHD through the implementation of the scientific discovery. It contains 776 records of the resting-state fMRI and anatomical data across 8 independent imaging sites, 491 of which come from typically developing individuals and 285 from children and adolescents diagnosed with ADHD (ages: 7-21 years old). Accompanying phenotypic information includes: diagnostic status, dimensional ADHD symptom measures, age, sex, intelligence quotient (IQ) and lifetime medication status. Preliminary quality control assessments (usable vs. questionable) based upon visual time-series inspection are included for all resting state fMRI scans. An additional 197 individuals from six imaging sites were released without the diagnosis labels during the competition for testing purposes and their labels were released separately afterwards. More information on the dataset can be found at http://fcon_1000.projects.nitrc.org/indi/adhd200/.

In order to bring the ADHD-200 Global Competition to a wider audience, The Neuro Bureau[5] made preprocessed versions of the competition data freely available to the general public to help those whose specialities lay outside of resting-state fMRI analysis to bypass technical obstacles. There are several preprocessed datasets available which were preprocessed by different pipelines. In order to fairly compare the proposed method with the baselines, we choose the dataset preprocessed by Athena pipeline[1] which is also used by the baseline methods. More information about the Athena pipeline can be found at Neuro Bureau's website[6].

5.2 Experimental Setting and Baselines

For the sake of fair comparison, we follow the experiment setting similar with the baselines. For the proposed FTA hashing, we determine the values of hyper parameters K, L, γ and η by conducting 5-fold cross validation on the training set. As mentioned in section 3.1, the TCs we used for evaluation were extracted from the pre-constructed brain atlas which was built with automated anatomical labeling (AAL)[29]. The way to extract TCs is averaging the time courses within each ROI voxel[6]. Note that the AAL atlas was constructed using anatomic and cyto-architectonic information and did not incorporate functional information. Thus the resultant TCs do not contain any prior phenotypic information which may impact the evaluation. Based on the cross validation, FTA base K and code length L are set to 2 and 200 respectively.

We compare the proposed method with following algorithms:

- Dynamic Time Warping (DTW)[26]: A well known time series alignment technique that computes optimal distance between two time series of different lengths while preserving their temporal order.
- Derivative Dynamic Time Warping (DDTW)[19]: This method uses derivatives of the original time series to improve alignment by DTW.
- Canonical Time Warping (CTW)[31]: This method combines canonical correlation analysis (CCA) with DTW.
- Method from Johns Hopkins University (JHU)[12]: The winner of ADHD-200 Global Competition which achieved the state-of-the-art performance.
- Attributed Graph Distance Measure (AGDM)[10]: This method proposed a graph based feature called Attributed Graph Distance Measure which can be used to classify ADHD subjects.

Note that DTW, DDTW and CTW are unsupervised methods while JHU and AGDM are supervised. Following the settings of these baselines, we evaluate the proposed method with four statistical metrics: Prediction Accuracy, Specificity (= True Negative Rate), Sensitivity (= True Positive Rate) and J-Statistic (= Specificity + Sensitivity - 1). Among them the Prediction Accuracy is the primary metric for scoring. Note that the detection rate used in [10] is identical to prediction accuracy.

5.3 Comparison with Unsupervised Baselines

We begin our evaluation by demonstrating comparison results with the unsupervised baselines. As mentioned in Section 5.1, the training set consists of 8 subsets collected from different sites while the test set consists of 6 subsets. We implement FTA in Matlab and use the Matlab implementation of DTW, DDTW and CTW provided by [30] to conduct the experiments. The training and testing are performed on the training and testing subsets across all the eight imaging sites. The hardware configuration for the experiment is Intel i7-4790 CPU at 3.6GHz and 8GB RAM. Table 1 shows all four evaluation metrics and the computing time of three unsupervised baselines and the FTA.

Metric	DTW	DDTW	CTW	FTA
Accuracy	0.4678	0.4386	0.4678	**0.6140**
Specificity	0.7127	0.6915	0.8085	**0.9149**
Sensitivity	**0.2987**	**0.2987**	0.1818	0.2727
J-Statistics	0.0115	-0.0098	-0.0097	**0.1876**
Time(s)	149.5502	267.8659	47996.5866	**20.771**

Table 1: Performance comparison between the unsupervised baselines and the FTA.

As shown, the FTA outperforms all three baselines by nearly 15% to 18% on the prediction accuracy, a significant improvement to these well known time-series alignment methods. This demonstrates the FTA can better encode the temporal patterns which are important for predicting ADHD subjects than the unsupervised baselines. All four methods have a high specificity but a relatively low sensitivity. The FTA can reach the best specificity of 0.9149 but with a sensitivity of 0.2727. This is reasonable since there are much more TDC subjects than ADHD children diagnosed in the

[5]http://www.neurobureau.org/

[6]http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline

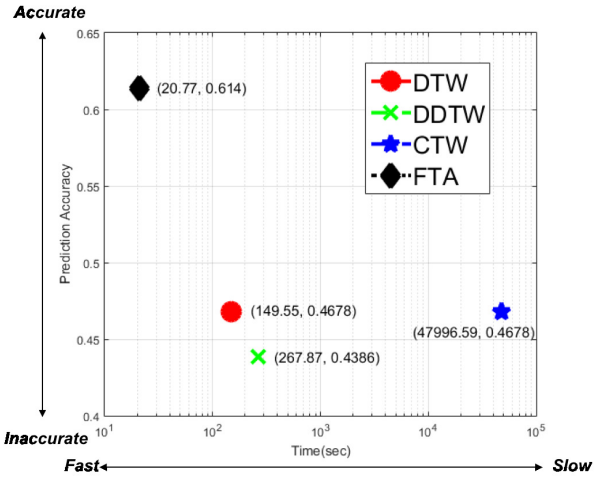


Figure 4: Prediction Accuracy versus Execution Time. Comparison between FTA and unsupervised baselines

training set (491 TDC and 285 ADHD children) thus the ability of a classifier to detect ADHD children is dampened – the prior distribution of TDC is biased by a large number of TDC samples in a general population. Such an unbalance between TDC and ADHD subjects can be relieved by imposing a larger penalty on missing ADHD subjects. However, we do not perform such a re-balance for the sake of a fair comparison with the other baselines.

Next we perform efficiency evaluation for all four methods. Since all three baselines adopt Dynamic Programming (DP) to perform similarity search, it can be expected that they will have a much slower speed than the proposed FTA hashing which performs similarity search by the Hamming distance. Figure 4 shows the Prediction Accuracy versus the Execution Time for all the methods. Compared with the baselines, FTA achieves more than 30% higher prediction accuracy on the entire test set and at least 7 times faster than DP based baselines. That suggests the proposed FTA hashing can be used for fast detection of ADHD subjects.

5.4 Comparison with Supervised Baselines

Now we compare the proposed method with two supervised baselines: JHU[12] and AGDM[10]. The approach proposed by JHU is a weighted combination of several algorithms including CUR decompositions, random forest, gradient boosting and support vector machine et al. It adopts both fMRI data and the accompanying phenotypic information like IQ to predict the ADHD subjects. On the contrary, another baseline AGDM only requires fMRI data to make the prediction. According to [10], features called AGDM were extracted from fMRI data to encode the brain network structure first. Then they were used to train a SVM classifier which made the final prediction. Since two baseline methods have different experiment settings, we follow their individual settings to make a fair comparison.

FTA vs JHU

Similar to Section 5.3, we train and test the FTA on all the training subsets across the eight sites. More specifically, we randomly sample 3,000 pairs from all the 8 training subsets and use them to find the optimal projections. We also report the evaluation metrics following the definitions from the ADHD-200 Global Competition [23]. The comparison results between the JHU and the FTA are summarized in Table 2.

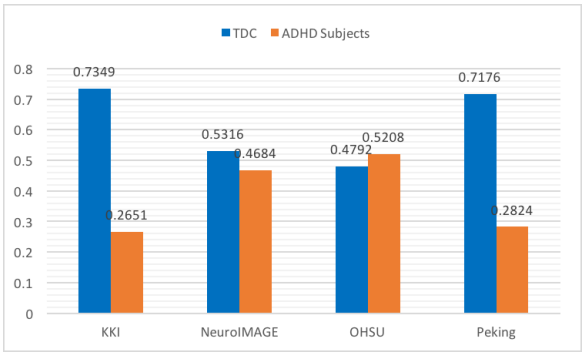


Figure 5: Percentage of TDC/ADHD Subjects in each training subset.

Metric	JHU	FTA
Prediction Accuracy	0.6102	**0.6140**
Specificity	**0.94**	0.9149
Sensitivity	0.21	**0.2727**
J-Statistics	0.15	**0.1876**

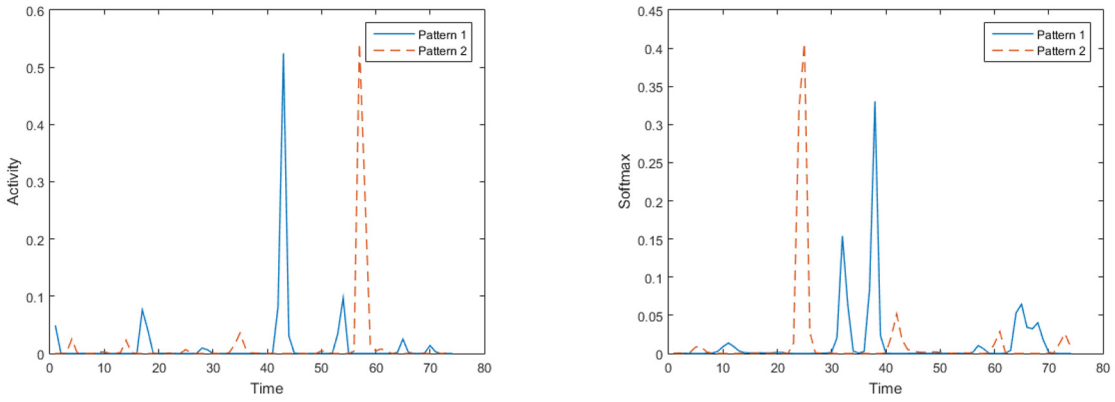
Table 2: Performance evaluation of JHU and FTA

From the table, we can see the proposed FTA outperforms the JHU on three metrics – prediction accuracy, sensitivity and J-Statistics. This is an impressive result, considering FTA only uses fMRI time courses to achieve such performance, whereas the JHU method involves both fMRI time courses and phenotypic information. Especially, compared with the JHU result, FTA improves the sensitivity by 29.85% while still keeping a competitive specificity (over 0.9). Such results demonstrate that encoding the temporal orders of the different neural activity patterns is definitely a helpful clue to identify the ADHD subjects.

We further justify this claim by showing some real examples of temporal orders of learned patterns. Figure 6 shows the sequences of two patterns projected from the TCs corresponding to the TDC and ADHD subjects. Figure 6a comes from a TDC subject while Figure 6b is obtained from a ADHD subject. It is clear that these two patterns exhibit different temporal orders between the TDC and ADHD subjects. Since the FTA can encode the temporal order of patterns, the resultant hash codes can characterize the neural activity difference between different types of TCs.

FTA vs AGDM

Unlike the JHU method, the AGDM conducted their experiments on the individual subsets. There are totally eight subsets collected by different imaging sites and AGDM chose four of them to evaluate their approach. The chosen subsets are collected by 1) Kennedy Krieger Institute (KKI), 2) NeuroIMAGE, 3) Oregon Health & Science University (OHSU) and 4) Peking University (Peking) respectively. To evaluate the proposed method, we follow the same experiment setting as the AGDM, and train and test on each subset separately.



a. Typically Developing Children

b. ADHD Patient

Figure 6: Comparison of the temporal order of two patterns between the TDC and ADHD subjects. It illustrates how the order of these two patterns matters in detecting ADHD.

Sites	Prediction Accuracy		Specificity		Sensitivity		J-Statistic	
	AGDM	FTA	AGDM	FTA	AGDM	FTA	AGDM	FTA
KKI	0.5455	**0.8182**	0.625	**1**	**0.3333**	**0.3333**	-0.0417	**0.3333**
NeuroImage	0.48	**0.8**	0.6429	**0.8571**	0.2727	**0.7273**	-0.0844	**0.5844**
OHSU	0.8235	**0.8676**	0.8929	**0.9643**	0.5	**0.6667**	0.3929	**0.6310**
Peking	0.5882	**0.6176**	0.9259	**1**	0.2083	**0.25**	0.1342	**0.25**
Average	0.6281	**0.7759**	0.8312	**0.9554**	0.2727	**0.4943**	0.1003	**0.4497**

Table 3: Performance evaluation on individual sets

Table 3 shows the comparison results between the FTA and the AGDM on individual sets. Apparently, the FTA significantly outperforms the AGDM on every evaluation metrics. Specifically, the average prediction accuracy and the average sensitivity of FTA outperform the AGDM by around 15% and 20% respectively. This means more than 75% of subjects can be correctly classified, and nearly half of ADHD patients can be successfully detected by the FTA. Moreover, the specificities of FTA on all four subsets are obviously boosted compared with the AGDM. Especially on KKI and Peking, the specificity achieves 1 – all TDC subjects are correctly identified.

Similar results can be found about the sensitivity. On the NeuroIMAGE and the OHSU subsets, the sensitivities have reached over 0.7 and 0.6, much closer to the corresponding specificity. Consider our discussion in Section 5.3 – the low sensitivity is caused by high ratio of TDC in the training set. We calculated the percentage of TDC and ADHD subjects on each subset, which are illustrated in Figure 5. It shows that the ratio of TDC/ADHD subjects on NeuroIMAGE and OHSU is close to 1, while it is much higher on KKI and Peking. This implies that the low sensitivity is caused by a high ratio of TDC/ADHD.

5.5 Parameter Sensitivity Analysis

Finally, we evaluate the FTA's performance by varying its hyper-parameters. In particular, we study the impact of the code length L and FTA base K on the performance. Besides the AAL, we also conduct experiments with TCs extracted from 3 other pre-built brain atlas: Talairach and Tournoux (TT)[21], Harvard-Oxford (HO)[15] and CC200[8].

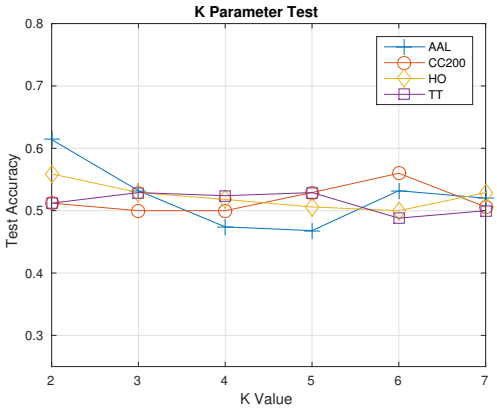


Figure 7: Test accuracy across different ROIs, using different K values

Prediction Accuracy vs K

By varying the FTA base K, it can be shown that it affects the test accuracy of the FTA model. For a fixed code length of 200, a variety of K values were used: from 2 through 7. The results show differences in impact of K for different ROI atlases. For instance, with the CC200, an increased K tends to improve the accuracy, whereas an increased K in the AAL results in a subtle decrease in the test accuracy

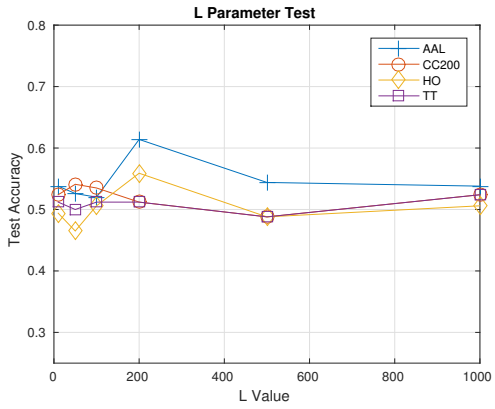


Figure 8: Test accuracy across different ROIs, using different L values

followed by an increase after $K = 6$. Overall, all four ROI atlases, begin to converge to a median of performance when K approaches 7. Throughout the tests, the results indicate only small dependencies contingent on the values of K.

Considering K is the number of patterns involved in generating a hash code, assessing its role offers a way to tune the amount of discriminant information available at any given comparison. From Figure 7, it seems that for slight increases in K (e.g., from K = 4 to K = 6 for CC200) the amount of information increases, offering an increased test accuracy. However, there appears to be a point at which further increasing the value of K no longer improves the test accuracy (beyond $K = 6$). Ultimately, this leads to a decline in performance due to the redundancies in the comparison of a too large number of patterns.

Prediction Accuracy vs L

Now let us assess the code length parameter L. Fixed at $K = 2$, a variety of code lengths were tested, ranging from $L = 10$ to $L = 1000$. In Figure 8, it can be seen that L acts differently across different ROIs, especially between the range of 10 to 200 code length. Incidentally, it is shown that this range experiences the most abrupt shifts in performance, whilst the range 200-1000 shows either constant or steady trends. In the case of AAL and HO, they are particularly sensitive to this parameter. Within the $L = 10$ to $L = 200$ range, the performance reaches its peak and then drops subsequently. Extending to the set of other ROIs, Figure 8 shows that this trend (to a lesser degree) is exhibited on CC200 and TT as well.

Resulting from this experiment, it can be shown that varying code lengths has impact on performance, with tuning required particularly in the range of compact codes from $L = 10$ and $L = 200$. Yet similar to K, simply increasing the value indefinitely does not result in a better test accuracy because of higher redundancies that might exist in longer codes.

6. CONCLUSION

In this paper, we propose a novel FTA algorithm to hash time series of varied length into fixed-size codes. We use the resultant hash codes to efficiently detect the ADHD subjects by fast retrieving the similar subjects. To maximize its performance, we design an effective algorithm to learn the optimal projections $\mathbf{W}$. The results of extensive experimental evaluations show the proposed FTA outperforms both unsupervised and supervised methods on the ADHD-200 dataset for identifying ADHD subjects, beating the winning algorithm in the ADHD-200 Global Competition.

7. ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

8. REFERENCES

[1] P. Bellec, C. Chu, F. Chouinard-Decorte, D. S. Margulies, and C. R. Craddock. The neuro bureau adhd-200 preprocessed repository. *bioRxiv*, page 037044, 2016.

[2] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] G. Buzsaki. *Rhythms of the Brain*. Oxford University Press, 2006.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[6] S. Cortese, C. Kelly, C. Chabernaud, E. Proal, A. Di Martino, M. P. Milham, and F. X. Castellanos. Toward systems neuroscience of adhd: a meta-analysis of 55 fmri studies. *American Journal of Psychiatry*, 2012.

[7] D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

[8] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

[9] J. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. Stam, S. M. Smith, and C. Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

[10] S. Dey, A. R. Rao, and M. Shah. Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. *Frontiers in neural circuits*, 8, 2014.

[11] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, and K. Zilles. A new spm toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25(4):1325–1335, 2005.

[12] A. Eloyan, J. Muschelli, M. B. Nebel, H. Liu, F. Han, T. Zhao, A. Barber, S. Joel, J. J. Pekar, S. Mostofsky, et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. 2012.

[13] J. A. Etzel, V. Gazzola, and C. Keysers. An introduction to anatomical roi-based fmri classification analysis. *Brain research*, 1282:114–125, 2009.

[14] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9673–9678, 2005.

[15] J. A. Frazier, S. Chiu, J. L. Breeze, N. Makris, N. Lange, D. N. Kennedy, M. R. Herbert, E. K. Bent, V. K. Koneru, M. E. Dieterich, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 2005.

[16] T.-c. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[17] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

[18] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 285–289. ACM, 2000.

[19] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. In *Sdm*, volume 1, pages 5–7. SIAM, 2001.

[20] S. Klöppel, A. Abdulkadir, C. R. Jack, N. Koutsouleris, J. Mourão-Miranda, and P. Vemuri. Diagnostic neuroimaging across diseases. *Neuroimage*, 61(2):457–463, 2012.

[21] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox. Automated talairach atlas labels for functional brain mapping. *Human brain mapping*, 10(3):120–131, 2000.

[22] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[23] M. P. Milham, D. Fair, M. Mennes, S. H. Mostofsky, et al. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012.

[24] G. A. Ojemann, J. Ojemann, and N. F. Ramsey. Relation between functional magnetic resonance imaging (fmri) and single neuron, local field potential (lfp) and electrocorticography (ecog) activity in human cortex. *Front Hum Neurosci*, 7:34, 2013.

[25] R. A. Poldrack. Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, 2(1):67–70, 2007.

[26] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993.

[27] R. P. N. Rao. *Brain-Computer Interfacing: An Introduction*. Cambridge University Press, New York, NY, USA, 2013.

[28] C. A. Ratanamahatana and E. Keogh. Making time-series classification more accurate using learned constraints. SIAM, 2004.

[29] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[30] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[31] F. Zhou and F. Torre. Canonical time warping for alignment of human behavior. In *Advances in neural information processing systems*, pages 2286–2294, 2009.

APPENDIX

A. EQUATIONS FOR LEARNING OPTIMAL PROJECTIONS

This section contains the equations required to derive the gradient of $\mathcal{F}$ w.r.t. $\mathbf{W}$ in learning the optimal projections.

Here, we use $\mathcal{L}^{ij}$ to denote the logarithmic training loss for a pair of TCs $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$, i.e.,

$$\mathcal{L}^{ij} = s_{ij} \log\left(1 - h^{ij}\right) + (1 - s_{ij}) \log\left(h^{ij}\right) \tag{11}$$

The following equations can be calculated by applying the chain rule of the derivatives on $\mathcal{F}$ of Eq. (10).

$$\frac{\partial \mathcal{F}}{\partial \mathbf{w}_k} = \sum_{i,j=1}^{N} \frac{\partial \mathcal{L}^{ij}}{\partial \mathbf{w}_k} + \gamma \frac{\partial \Omega}{\partial \mathbf{w}_k} + \eta \frac{\partial \mathcal{V}}{\partial \mathbf{w}_k} \tag{12}$$

$$\frac{\partial \Omega}{\partial \mathbf{w}_k} = \sum_{k \neq k'=1}^{K} \frac{2\mathbf{w}_k^\intercal \mathbf{w}_{k'}}{\|\mathbf{w}_k\|^4 \|\mathbf{w}_{k'}\|^2} \left[\left(\mathbf{w}_k^\intercal \mathbf{w}_k\right) \mathbf{w}_{k'} - \left(\mathbf{w}_k^\intercal \mathbf{w}_{k'}\right) \mathbf{w}_k \right] \tag{13}$$

$$\begin{aligned} \frac{\partial \mathcal{V}}{\partial \mathbf{w}_k} = \sum_{i,k=1}^{N,K} \sum_{t=1}^{T} \frac{1}{T^2} \Bigg[& 2\left(t - m_k\right) \left(-\frac{\partial m_k^{(i)}}{\partial \mathbf{w}_k}\right) p_{k,t}^{(i)} \\ & + (t - m_k)^2 \left(p_{k,t}^{(i)} \mathbf{x}_t^{(i)} - p_{k,t}^{(i)} \left(\sum_{t'=1}^{T} p_{k,t'}^{(i)} \mathbf{x}_{t'}^{(i)} \right) \right) \Bigg] \end{aligned} \tag{14}$$

$$\frac{\partial \mathcal{L}^{ij}}{\partial \mathbf{w}_k} = \begin{cases} -\frac{1}{1-h^{ij}} \frac{\partial h^{ij}}{\partial \mathbf{w}_k} & s_i = s_j \\ \frac{1}{h^{ij}} \frac{\partial h^{ij}}{\partial \mathbf{w}_k} & otherwise \end{cases} \tag{15}$$

$$\frac{\partial h^{ij}}{\partial \mathbf{w}_k} = \sum_{k'=1}^{K} h_{k'}^{(i)} \frac{\partial h_{k'}^{(j)}}{\partial \mathbf{w}_k} + \sum_{k'=1}^{K} h_{k'}^{(j)} \frac{\partial h_{k'}^{(i)}}{\partial \mathbf{w}_k} \tag{16}$$

$$\frac{\partial h_k^{(i)}}{\partial \mathbf{w}_k} = h_k^{(i)} \left(-\frac{\partial m_k^{(i)}}{\partial \mathbf{w}_k}\right) - \left(h_k^{(i)}\right)^2 \left(-\frac{\partial m_k^{(i)}}{\partial \mathbf{w}_k}\right) \tag{17}$$

$$\frac{\partial h_l^{(i)}}{\partial \mathbf{w}_k} = -h_l^{(i)} h_k^{(i)} \left(-\frac{\partial m_k^{(i)}}{\partial \mathbf{w}_k}\right), \text{ when } l \neq k \tag{18}$$

$$\frac{\partial m_k^{(i)}}{\partial \mathbf{w}_k} = \sum_{t=1}^{T} \frac{t}{T} \left(p_{k,t}^{(i)} \mathbf{x}_t^{(i)} - p_{k,t}^{(i)} \left(\sum_{t'=1}^{T} p_{k,t'}^{(i)} \mathbf{x}_{t'}^{(i)} \right) \right) \tag{19}$$