Path Signature Neural Network of Cortical Features for Prediction of Infant Cognitive Scores

Jiale Cheng, Xin Zhang, Hao Ni, Chenyang Li, Xiangmin Xu, Zhengwang Wu, Li Wang, Weili Lin, and Gang Li

Abstract—Studies have shown that there is a tight connection between cognition skills and brain morphology during infancy. Nonetheless, it is still a great challenge to predict individual cognitive scores using their brain morphological features, considering issues like the excessive feature dimension, small sample size and missing data. Due to the limited data, a compact but expressive feature set is desirable as it can reduce the dimension and avoid the potential overfitting issue. Therefore, we pioneer the path signature method to further explore the essential hidden dynamic patterns of longitudinal cortical features. To form a hierarchical and more informative temporal representation, in this work, a novel cortical feature based path signature neural network (CF-PSNet) is proposed with stacked differentiable temporal path signature layers for prediction of individual cognitive scores. By introducing the existence embedding in path generation, we can improve the robustness against the missing data. Benefiting from the global temporal receptive field of CF-PSNet, characteristics consisted in the existing data can be fully leveraged. Further, as there is no need for the whole brain to work for a certain cognitive ability, a top K selection module is used to select the most influential brain regions, decreasing the model size and the risk of overfitting. Extensive experiments are conducted on an in-house longitudinal infant dataset within 9 time points. By comparing with several recent algorithms, we illustrate the state-of-the-art performance of our CF-

PSNet (i.e., root mean square error of 0.027 with the time latency of 518 milliseconds for each sample).

Index Terms—Infant brain, Longitudinal analysis, Cognitive scales, Path signature features

I. INTRODUCTION

THE early development of human brains is attracting more and more attention due to its critical roles in later behavioral and cognitive outcomes [1]–[4]. The cortical structural development during infancy is tightly connected to the acquisition and refinement of information processing abilities as well as visual and language skills [5]. Thus, understanding the relationship between cognitive skills and morphological features of the infant cerebral cortex is of immense importance [6]. Recent works [7]–[11] have concentrated on applying convolutional neural networks and other machine learning based methods to infant neuroimaging data analysis. Nevertheless, only few works [12], [13] related cognitive scores to the infant brain cortical morphology.

Hence, in this work, we aim to propose a machine learning approach to predict cognitive development for each infant by using their longitudinal brain MRI scans. Particularly, with an infant-specific computational pipeline [14], several biologically meaningful cortical measurements can be extracted as input features for each brain region [14]. Meanwhile, five Mullen Scales of Early Learning (MSEL) [15], obtained at 48 months of age, are used to represent cognition skills of each infant. In practice, small-sample-size (SSS) and missing data from different time points during data gathering procedure are often inevitable. Specifically, the SSS problem poses a huge challenge to apply the machine learning based approaches. Therefore, it is desirable for a light-weight and effective feature descriptor to produce a compact and informative representation from the excessive dimension of longitudinal data.

Recent advances [12], [13] are proposed to address these problems by learning the nonlinear mapping function from cortical features to cognitive scores. They have designed the latent space to deal with missing data and high dimensionality issues. Nonetheless, due to the nonlinear dimensionality reduction at early stages, their methods are incapable to analyze the correspondent relationship between brain regions and cognitive scores. Further, their simply weighting features from

X. Zhang and X. Xu are supported in part by the NSFC under grant U1801262; Guangzhou Key Laboratory of Body Data Science under grant 201605030011; Science and Technology Program of Guangzhou under grant 2018-1002-SF-0561; Natural Science Foundation of Guangdong Province under grant 2018A030313295. H. Ni is supported by the EPSRC under the program grant EP/S026347/1 and by the Alan Turing Institute under the EPSRC grant EP/N510129/1. G. Li is supported in part by NIH grants MH117943, MH116225 and MH123202. L. Wang is supported in part by NIH grant MH117943. (Corresponding authors: Xin Zhang and Gang Li)

J. Cheng is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China, and also with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27514 USA (e-mail: jialecheng100@gmail.com).

X. Zhang, C. Li, and X. Xu are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China (e-mail: eexinzhang@scut.edu.cn; lichenyang.scut@foxmail.com; xmxu@scut.edu.cn).

Z. Wu, L. Wang, W. Lin, and G. Li are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: zhengwang.wu@med.unc.edu; li.wang@med.unc.edu; weili.lin@med.unc.edu; gang.li@med.unc.edu).

H. Ni is with the Department of Mathematics, University College London, Gower Street, London WC1E 6BT, UK (e-mail: h.ni@ucl.ac.uk).

Fig. 1. Illustration of the whole framework. Inputs are longitudinal MRI data (grey brains indicating missing data) and output is the predicted cognitive score. The details of our Cortical Feature based Path Signature Neural Network (CF-PSNet) are shown in Fig. 3.

different time points may lead to an insufficient exploration on the dynamic information of longitudinal data.

In this work, the path signature method is applied to describe the dynamic dependencies of longitudinal infant cortical structure. Path signature (PS) comes from rough path theory, a branch of stochastic analysis [16], and can be used as an efficient and principled summary of sequential data (a path) in terms of its effect represents the effects. Recently, several machine learning models incorporated the path signature and all achieved the state-of-the-art performance in various applications [17]–[21]. Inspired by these works, we define paths in longitudinal data along time axis for brain regions since it is explainable and does not require extra parameters to be optimized. Further, path signature features, extracted by a differentiable temporal path signature (TPS) layer, can effectively describe dynamic properties of the cortical growth trajectory. With the help of an attention mechanism to automatically weight brain regions at different time points, encouraging results can be obtained [22].

However, to tackle with the missing data issue and recover the continuously growing patterns in longitudinal analysis, interpolation methods are usually used in pre-processing. The simply interpolated missing data may introduce ambiguity and artifacts on the exploration of dynamic patterns. In previous studies [12], [13], researchers solve this problem by mapping features from all existing visits of the same subject into a uniform latent space for a thorough exploration on the available data. A set of indicators is also used for eliminating the possible influence of missing data on the model performance through back propagation.

Similarly, based on our previous work [22], we stack the light-weight TPS layers to obtain a global receptive field and aggregate features from all existing visits. Compared to [12], [13], their mapping function is a single-layer non-linear function, since it is risky to apply its complicated

substitutes such as RNNs with the SSS problem, while the hierarchical structure of dynamic patterns may be overlooked. In contrast, by stacking TPS layers, we can thoroughly study the multi-scale temporal contextual information at almost no cost. Second, a set of embedding matrices are proposed for indicating the existence of each time point, thus enabling the model to distinguish the existing visits from the missing ones and focus more on the existing scans. Further, considering the limited number of subjects and redundancy of cortical data, we try to select the most influential brain regions in the longitudinal cortical surface data across the entire network. The top K selection module is thus designed to automatically choose few brain regions and give them different weights and further decrease the model size.

The remainder of this paper is organized as follows. Section II presents the summary and analysis of recent related works and our previous conference version [22]. In Section III, we present the preliminaries of path signature definition and concepts. Section IV describes our dataset and feature extraction details. Section V presents the details of our proposed approach. In Section VI, we elaborate ablation study results and comparisons with state-of-the-art methods. Section VII is devoted to analyze the result by revealing the brain region and cognitive score relationship. Section VIII concludes the paper.

II. RELATED WORK

A. Analysis of Infant Cognitive Scores

In the literature, there has been intensive researches for infant brain development analysis [23], [24]. However, few studies correlate cognition functions with the longitudinal development of the brain. Some researchers used the statistical analysis framework [6], [25]–[27]. For instance, in [6], the authors intended to determine the association between cortical thickness, surface area and measures of cognitive ability through mediation analysis and longitudinal analysis. Factor

analysis has also been used to qualitatively evaluate the impact of different brain features on cognition function and illustrate the high correlation between the anatomical differentiation of white matter fibers and cognitive development [25]. Nevertheless, for the lack of quantitative formulation, they are incapable to predict the cognitive scales for individuals directly.

To enable individual prediction, recent advances of cognitive scores prediction are mainly based on machine learning methods [12], [13] with challenges of the missing data and the trade-off between large dimensional features and the SSS problem. Specifically, in [13], a latent representation is generated for each participant to leverage the complementary information among different time-point. A set of indicators is also introduced to eliminate the loss brought by the incomplete data. Nonetheless, in their formulation, the cortical morphological features extracted from 70 brain regions of each MRI scan are flattened into a 490 dimensional vector, which may cause the loss of structural information. Meanwhile, in [12], researchers proposed a Bag-of-Words [28] based model to decrease the number of features at an early stage with the aim of balancing the overlarge dimensionality of neuroimaging features and limited sample size. They achieved encouraging performance, but the linear transformation they used to gather features of different time points may be incapable to sufficiently explore the dynamic patterns in growth trajectories of brain regions.

B. Path Signature Method

Path signature is an infinite graded sequence of statistics known to characterize data streams which can also be regarded as paths. As a type of noncommutative formal power series, it originated from Chen's study [29] who works on piece-wise smooth path. Lyons first used it to make sense of controlled differential equations driven by very rough paths [16]. The uniqueness of signature was initially proved by [30] for paths of bounded variation and it was extended to paths of finite p -variation for any $p > 1$ in [31]. Recently, the applications of path signature in machine learning is an emerging areas; in particular, the path signature can be used as an effective feature extractor in various tasks [17]–[21]. Specifically, in [32], Diehl applied path signature transformation onto handwritten curves for recognition. Yang pioneered in constructing paths in spatial domain by connecting skeleton nodes in a single frame for skeleton-based human action recognition and obtained great results [20]. However, the choice of constructing the paths from high dimensional data stream to be best accompanied with the the path signature feature is not yet clear and needs further investigation.

C. Comparison with the Previous Work

In our previous work [22], which is the first of building paths on the cerebral cortex structure, we designed the TPS layer to extract the geometrically and analytically meaningful path signature features for describing the growth patterns of brain regions. Together with an attention mask generator, our network achieves a competitive result on an in-house dataset. It provides a new solution on the longitudinal brain analysis and validates its feasibility. However, it is still challenging for

the missing data issue and the trade-off between the excessive feature dimension and limited data. Therefore, in this paper, we extend the previous version by introducing the hierarchical structure, existence embedding and the top K selection module to fully leverage the existing data and decrease model size. More technical details and extended analyses are also added. Our contributions can be concluded as follows,

- We demonstrate the potential of path signature method to have a deeper integration with the machine learning structure and work as a typical sequential descriptor in longitudinal brain analysis.
- We demonstrate the further performance gain enabled by introducing the hierarchical structure, existence embedding and top K selection module to construct the CF-PSNet, compared to the conference version, thanks to its thoroughly exploration and paying more attention on the existing visits and the influential brain regions.
- We perform comprehensive experiments to evaluate the impact of missing data and the number of brain regions used in cognition prediction, and validate the effectiveness our method and its superiority performance compared to the other advanced approaches.

III. PRELIMINARIES OF PATH SIGNATURE

A path X in $\mathbb{R}^d$ is a continuous mapping of finite length from interval $[a, b]$ to a d -dimensional vector space, i.e. $X : [a, b] \rightarrow \mathbb{R}^d$. We use the subscript notation $X_t = \{X_t^1, X_t^2, \dots, X_t^d\}$ to denote the d -dimensional vector for any $t \in [a, b]$, where X_t^i denotes the i^{th} coordinate of X_t and $i \in \{1, 2, \dots, d\}$. Before introducing the signature, let us introduce $S_l(X)_{a,b}$, the l^{th} fold iterated integral of the path X , for the convenience of understanding.

The 1st iterated integral of X^i equals to the increment of X along the i^{th} coordinate, i.e., $X_b^i - X_a^i$, which can be expressed by $\text{Sig}(X)_{a,b}^i = \int_{a < t_1 < b} dX_{t_1}^i$. Meanwhile, the 1st fold iterated integral of X is defined as the collection of $\text{Sig}(X)_{a,b}^i$ for $i \in \{1, 2, \dots, d\}$, i.e.,

$$\begin{aligned} S_1(X)_{a,b} &= \int_{a < t_1 < b} 1 dX_{t_1} \\ &= \int_{a < t_1 < b} S_0(X)_{a,t} dX_{t_1}, \end{aligned} \quad (1)$$

which is a d -dimensional vector. The 0th fold iterated integral $S_0(X)_{a,b}$ is set to 1 conventionally.

Similarly, the 2nd fold iterated integral of X is defined as the integral of $S_1(X)_{a,\cdot}$ against X , which is the collection of all 2nd iterated integrals of X with possible index (i_1, i_2) , i.e., $\left(\text{Sig}(X)_{a,b}^{i_1, i_2}\right)_{i_1, i_2 \in \{1, \dots, d\}}$. The collection of 2nd iterated integrals can be written in the tensor form as follows:

$$\begin{aligned} S_2(X)_{a,b} &= \int_{a < t < b} S_1(X)_{a,t} \otimes dX_t \\ &= \int_{a < t_1 < t_2 < b} dX_{t_1} \otimes dX_{t_2}. \end{aligned} \quad (2)$$

The l -th fold iterated integral of X , $S_l(X)_{a,b}$, is the integral of $S_{l-1}(X)_{a,\cdot}$ against X :

Fig. 2. The illustration of the missing rate distribution for every time point and every subject.

$$\begin{aligned}
 S_l(X)_{a,b} &= \int_{a < t < b} S_{l-1}(X)_{a,t} \otimes dX_t \\
 &= \int_{a < t_1 < \dots < t_l < b} dX_{t_1} \otimes dX_{t_2} \dots \otimes dX_{t_l},
 \end{aligned} \quad (3)$$

whose dimensionality is d^l .

In general, the signature of a path is a graded infinite series containing all the l fold iterated integrals. In practice, for dimensionality reduction, we may truncate the signature up to a finite degree l , which is denoted as $Sig_l(X)_{a,b}$:

$$Sig_l(X)_{a,b} = (S_0(X)_{a,b}, S_1(X)_{a,b}, \dots, S_l(X)_{a,b}). \quad (4)$$

The dimension of $Sig_l(X)_{a,b}$ defined in Equation (4) is:

$$n_{PS} = \frac{d^{l+1} - 1}{d - 1}. \quad (5)$$

In practice, the discrete time series are more common, which can be embedded into the path space by the linear interpolation. The corresponding signature of the embedded path can be computed explicitly by Chen's identity [29] as follows:

$$Sig(X)_{a,b}^{i_1, i_2, \dots, i_l} = \frac{1}{l!} \prod_{j=1}^l (X_b^{i_j} - X_a^{i_j}), \quad (6)$$

which is a non-linear feature set of the time series data. Specifically, in this application, we interpret X as the growth trajectories of brain regions extracted from the longitudinal MRI scans. It is noteworthy that the 1st fold iterated integral of growth trajectories are the variations of biological measurements, while the linear combination of the 2nd iterated integrals, $\frac{1}{2}(Sig(X)_{a,b}^{i_1, i_2} - Sig(X)_{a,b}^{i_2, i_1})$, is equal to the area enclosed by the curve (X^{i_1}, X^{i_2}) and a chord connecting the starting and the ending points of the path, which can be used to evaluate the curvature of the growth trajectories.

The signature of a path is an effective feature set of the streamed data containing many algebraic and analytic properties. It remains invariant under time re-parameterization of path X [33]. It means that the signatures capture the information on the path trajectory while removing infinite-dimensional noise caused by speed variation. Intuitively, the signature of a path plays a role as the non-commutative polynomial on the path space. Further, in practice, the signature feature set can be used to handle time series of variable length, and variation caused by the time parameterization. Therefore, it can be used as a

fixed dimensional and principled descriptor to summarize the path information over intervals effectively and offer dimension benefits especially for long time series. For interested readers, [34] and [35] are suggested to find more details.

IV. DATASET AND FEATURE EXTRACTION

In this study, we used T1w and T2w MR images of 23 infants collected at 9 different longitudinal time points (i.e., their 0, 3, 6, 9, 12, 18, 24, 36 and 48 months after birth) using a 3T Siemens scanner (TIM TRIO) with a 32-channel head coil. Specifically, we acquired the T1w images with the following parameters: 144 sagittal slices, repetition time (TR) = 1900 ms, echo time (TE) = 4.38 ms, flip angle = 7° , acquisition matrix = 256×192 , and voxel size = $1 \times 1 \times 1 \text{ mm}^3$, while the T2w images were acquired with the following parameters: 64 axial slices, TR = 7380 ms, TE = 119 ms, flip angle = 150° , acquisition matrix = 256×128 , and voxel size = $1.25 \times 1.25 \times 1.95 \text{ mm}^3$. For more information, please refer to [36].

These collected MR images were processed by an infant MRI computational pipeline¹ [14], [37] to extract 7 morphological measurements of the cerebral cortex. The key steps include intensity inhomogeneity correction [38], skull stripping and cerebellum removal [39], tissue segmentation [40], separation of left/right hemisphere, topology correction [41], and inner and outer surfaces reconstruction [42]. Then, the 7 different morphological cortical measurements are calculated at each vertex of the reconstructed cortical surfaces, including cortical thickness (THI), mean curvature (CUR), local gyrification index (LGI), vertex area (ARE), vertex volume (VOL), sulcal depth in string distance (SDS) and sulcal depth in Euclidean distance (SDE) [43], [44], which are widely used to quantify early brain development [45]. Afterwards, we can parcellate the cerebral cortex into 70 anatomically meaningful regions of interest (ROIs) with the help of a 4D infant cortical surface atlas [46] for dimensionality reduction. A 70×7 feature map is then extracted for each available scan corresponding to 70 different brain regions and their 7-dimensional feature representations. Finally, we can get a cohort of dynamic feature maps as shown in Fig. 1 by concatenating these feature maps along the time axis. Considering that we have missing visits as illustrated in Fig. 2, an interpolation method is applied to enable the following process works properly. Specifically, by averaging the existing data among all subjects at each visit, we can obtain a 9-point cortex growth trajectory template $\mathbf{t} = \{t_{0M}, t_{3M}, \dots, t_{48M}\}$. For an missing visit t_i^n of subject n , we fill it with its two nearest existing visit t_j^n, t_k^n as follows,

$$t_i^n = t_j^n + (\mathbf{t}_i - \mathbf{t}_j) * \frac{t_j^n - t_k^n}{t_j^n - t_k^n}. \quad (7)$$

Further, the five Mullen cognitive scores for each participant were estimated at their 4 years old to quantify the development of their cognitive functions, i.e., Fine Motor Scale (FMS), Receptive Language Scale (RLS), Expressive Language Scale (ELS), Visual Receptive Scale (VRS) and Early Learning Composite (ELC), which are firmly correlated to the brain morphological measurements mentioned above [15].

¹<http://www.ibeat.cloud/>

Fig. 3. The illustration of CF-PSNet. The network structures of the three streams in multi-stream neural network, which are surrounded by different colors, are identical. The 1×1 output is the final cognitive score we predict. Best viewed in color.

V. CORTICAL FEATURE BASED PSNET

In this section, we will elaborate a novel CF-PSNet consisting of stacked TPS layers to form a hierarchical temporal representation for each participant to predict their cognitive scores accurately with the extracted cortical measurements as shown in Fig. 1. Before exploring the dynamic patterns in the growth trajectories of brain regions with PS, we conduct feature redundancy removal and brain regions selection by a single-layer non-linear mapping and a top K selection module considering the SSS problem. It is noteworthy that an existing embedding matrix is applied within the first TPS layer we used to indicate missing visits. Assuming that the most informative brain regions for each task changes along time, the group fully connected layers and attention mask generator are also applied to treat each time point and each brain region differently, which is following our previous work [22]. The multi-scale temporal representation is processed with a multi-stream network and finally fused by a fully connected layer whose output are cognitive scores predicted.

A. Top K Selection Module

First, we briefly review the top K selection module, as illustrated in Fig. 3, which is proposed to select the top K influential brain regions for each task based on features over all 9 time points. As shown in Fig. 4, features of the same brain region are flattened into a 36-dimensional feature vector and then fed into a fully connected layer to produce a scalar as its influence coefficient under a certain cognitive task. In terms of these influence coefficients, top K influential ROIs are selected to be fed into following layers. Further, we multiply influence coefficients with the corresponding feature vectors of different ROIs as an initial weight. By selecting the top K influential brain regions, the excessive dimension of input features is substantially decreased. Then, parameters needed in the next following layers are accordingly reduced.

Fig. 4. The illustration of the structure of the top K selection module. In practice, we fixed K at 10 after comparison. Figure legends please refer to Fig. 3.

B. Temporal Path Signature Layer

Stacked temporal path signature layers are then proposed to extract multi-scale dynamic information and generate discriminative representation for each participant. In each TPS layer, for the first step, K paths corresponding to K most influential ROIs are defined along the time axis as shown in Fig. 5 and then, like what CNN does, split by overlapping sliding windows with the size of W and a sliding stride of 1 as illustrated in Fig. 6. Consequently, for each path, $T^i = (T^{i-1} + 1 - W)$ sub-paths are generated to further explore local temporal properties, each covers a temporal receptive field of a $\bar{T}^i = 9 - i \times (W + 1)$ time points period. In this work, we denote i as the index of TPS layers, while $T^0 = 9$ for input. Then, for every sub-path, Equations (4) and (6) together with Chen's identity are employed to compute its corresponding path signature features and obtain an output with the dimension of $n_{PS} = (4^{l+1} - 1)/(4 - 1)$ according to Equation (5) with d being set as 4. The outputs of different sliding windows are gathered to construct a new path, shown

Fig. 5. The illustration of the way we defined paths. Each red cube is denoted as a 4-dimensional feature vector for a brain region at a certain time point.

in Fig. 6 for details. Afterwards, an 1×1 convolutional layer is again introduced to conduct a feature transformation as well as dimensionality reduction from n_{PS} to 4 for SSS problem. Notably, this is the only set of parameters needed in the TPS layer. Given the short time series we have in this study, we only apply two TPS layers in CF-PSNet and generate a global temporal representation which covers all existing visits. We hope that it can be tested with longer sequence in future works.

Additionally, as the existence of missing visits, it would be better to highlight the position of the interpolated visits in paths explicitly. To this end, inspired by positional encoding in [47], we design an existence embedding in path generation to lift the path into a higher dimension. The existence embedding is a 1-dimension sequence which has the same length with the path generated to indicate the more reliable existing visits, so that it can be concatenated with the path. There are many choices of existence embedding. In this work, we use the cumulative sequence:

$$\begin{aligned} Path &= \{1, \circ, 3, 4, \circ, \circ, 7, \circ, 9\} \\ Binary &= \{0, 1, 0, 0, 1, 1, 0, 1, 0\} \\ Embedding &= \{0, 1, 1, 1, 2, 3, 3, 4, 4\} \end{aligned}$$

In the example above, *Path* denotes for the path we generated, while $\circ$ denotes for the missing visits. We use *Binary* to suggest the existence of each time point and compute the existence embedding as the cumulative sum of the *Binary* sequence.

C. Multi-stream Neural Network with Group Fully Connected Layers

With stacked TPS layers, hierarchical temporal information is extracted. A multi-stream neural network is then included in our CF-PSNet to process the raw features and multi-scale temporal PS features separately, believing that each stream transforms a set of features aggregated to a certain level. Considering that MR scans are collected sparsely along time, the influence of each brain region may vary in different time points. Therefore, we introduce group fully connected layers in all streams in Fig. 3, regard features sharing the same receptive

Fig. 6. The illustration of path signature features extraction. In practice, the sliding window size and truncated level are fixed at $W = 5$ and $k = 2$ respectively.

Fig. 7. The illustration of the structure of the attention mask generator. The 9×1 and 9×10 output are separately intermediate outputs and attention masks.

field in time domain as a group and process them separately by applying a group-specific fully connected layer. For instance, the numbers of groups are 9, 5, 1 respectively for Stream I, II and III when the sliding window size W is set to 5. The output of each group-specific fully connected layer is a set of features which encodes the cortical structure at corresponding developmental periods. At the end of the multi-stream neural network, the informative vectors produced by all streams are concatenated and fused with a fully connected layer and output the predicted cognitive score y (depicted in purple in Fig. 3).

D. Attention Mask Generator

For emphasizing the most influential brain region in each developmental stage, an attention mask generator is introduced as a parallel branch of the multi-stream neural network in Fig. 7. In this branch, a set of intermediate cognitive score $y_i, i \in \{1, 2, \dots, 9\}$ is obtained for each time point of the input data by applying the group fully connected layers sequentially with the group number of 9. The importance of each brain region at a certain time point is implicitly learned by the weights of the corresponding group-specific layer. Thus, we sum over parameters of group fully connected layers, which green dash lines surrounded in Fig. 7 along input channels and generate a $9 \times K$ attention mask corresponding to 9 time points and K ROIs. For features in Stream II and III whose receptive fields

TABLE I

PERFORMANCE COMPARISON OF CF-PSNET WITH DIFFERENT SEQUENCE ENCODERS. THE LAST COLUMN SHOWS THE TOTAL TIME COST IN MILLISECONDS (MS) FOR PROCESSING EACH SAMPLE

Sequence encoders of CF – PSNet	Metrics	VRS	FMS	RLS	ELS	ELC	Average	Time
Bi-LSTM [48]	RMSE	0.049	0.047	0.032	0.047	0.058	0.046	888ms
	R ²	0.878	0.912	0.957	0.795	0.941	0.878	
Transformer [47]	RMSE	0.038	0.047	0.050	0.049	0.040	0.045	625ms
	R ²	0.867	0.834	0.873	0.811	0.892	0.856	
GRU [49]	RMSE	0.041	0.043	0.042	0.038	0.047	0.042	739ms
	R ²	0.912	0.868	0.909	0.852	0.879	0.884	
TPS layer	RMSE	0.008	0.019	0.036	0.026	0.026	0.023	518ms
	R ²	0.996	0.971	0.955	0.896	0.974	0.958	

TABLE II

THE IMPACT OF DIFFERENT W AND l OF THE TPS LAYER ON THE MODEL SIZE

$W \backslash l$	1	2	3	4	5
2		18.6k	19.2k	21.2k	29.4k
3		16.3k	16.9k	18.9k	27.1k
4		14.0k	14.5k	16.6k	24.8k
5		11.7k	12.2k	14.3k	22.5k

TABLE III

THE IMPACT OF DIFFERENT W AND l OF THE TPS LAYER ON THE PREDICTION PERFORMANCE

$W \backslash l$	1	2	3	4	5
2		0.039	0.041	0.042	0.045
3		0.043	0.042	0.036	0.032
4		0.032	0.030	0.039	0.041
5		0.023	0.029	0.038	0.042

are more than a single time point, averaging are conducted over the corresponding receptive field of attention mask. Then, element-wise multiplications are conducted between groups of features and attention masks to weight brain regions differently for different time stages and time scales. Notably, in this work, the intermediate output $\hat{y} = (y_1, y_2, \dots, y_9)$ is used to assist generating attention masks, but not the final output.

E. Loss Function

All modules mentioned above are learned simultaneously with the loss function defined as:

$$Loss(y, \hat{y}, Y) = \lambda \cdot \sum_{t=1}^9 |Y - y_t| + |Y - y|, \quad (8)$$

which consists of two set of $L1$ loss to constrain the predicted cognitive scores y and the intermediate output $\hat{y}$ separately. Their ground truth are cognitive scales estimated at 48 months after birth which are denoted as Y . By constraining the intermediate output $\hat{y}$ with Y , we assuming that the cognitive function development at 48 months can be predicted by a single visit before. λ is a hyper-parameter introduced to balance the two $L1$ loss.

VI. EXPERIMENTS

In this section, we first conduct a series of experiments to illustrate how hyper-parameters of the path signature feature extraction influence the model size and experimental results. Then, with the help of an ablation study, the effectiveness of our network modules (e.g., the TPS layer, top K selection module and the existence embedding) is proved. Further, by comparing CF-PSNet with several recent algorithms, we show that our method achieves state-of-the-art performance.

A. Model Configurations

As described in Section IV, we conduct experiments on an in-house longitudinal dataset. Following the experimental settings in [13], we perform leave-one-out validation and use the Root Mean Squared Error (RMSE) between the predict values and ground truth of five cognitive scores as the evaluation metric. Learning rate is tuned in $\{10^{-3}, 10^{-4}, 10^{-5}\}$ with Adam as our optimizer. ReLU is employed as the non-linear activation for hidden neurons. After comparison, we fix $\lambda = 0.1$. In all our experiments, the number of epochs is at a maximum of 400.

Notably, we normalize five cognitive scores with their maximum and minimum values separately within a $[0, 1]$ range to have a unified comparison setting. The mean $\pm$ standard deviation for 5 tasks are 0.6615 ± 0.2467 , 0.5842 ± 0.2571 , 0.5565 ± 0.2913 , 0.6910 ± 0.2014 , 0.6087 ± 0.2628 respectively. An open-source package, *signatory*², is used to provide efficient path signature computation with backpropagation.

B. Comparison on the TPS Layer

In this work, for the first time, we introduce path signature method to extract rich dynamic information hidden in longitudinal data of structural MRI. In this subsection, we will evaluate its capability by selecting different hyper-parameters and comparing with other sequential models. First, we run our method with different sliding window sizes W and truncated levels l . It is noteworthy that two TPS layers in Fig. 3 are sharing the same choices of W and l in practice. Nevertheless, our method achieves great performance. To illustrate the impact of these two hyper-parameters on model

²Code available at github.com/patrick-kidger/signatory

TABLE IV
PERFORMANCE COMPARISON OF CF-PS NET WITH DIFFERENT POOLING MODULES

Pooling modules of CF – PSNet	VRS	FMS	RLS	ELS	ELC	Average	Parameter Size
Our Method w/o pooling	0.065	0.081	0.115	0.079	0.109	0.090	136.7k
DiffPool [50]	0.085	0.062	0.085	0.022	0.069	0.064	11.9k
Top K selection	0.008	0.019	0.036	0.026	0.026	0.023	11.6k

TABLE V
THE ABLATION STUDY OF CF-PSNET (IN TERMS OF RMSE)

Attention module	Stream II	Stream III	Top K selection	VRS	FMS	RLS	ELS	ELC	Average
✓	✓	✓	×	0.065	0.081	0.115	0.079	0.109	0.090
✓	×	×	✓	0.026	0.041	0.072	0.040	0.036	0.043
✓	✓	×	✓	0.047	0.014	0.044	0.022	0.039	0.033
×	✓	✓	✓	0.018	0.036	0.045	0.053	0.050	0.040
✓	✓	✓	✓	0.008	0.019	0.036	0.026	0.026	0.023

Fig. 8. Impact of the number of brain regions selected (K) for the top K selection module on the model performance.

Fig. 9. Impact of the number of visits in input starting from different ages. Experiments are conducted with the model containing a single TPS layer.

Fig. 10. Impact of the missing rate on the model performance. The error ranges and average values of 3 runs are illustrated.

size, some examples are shown in Table. II. It can be seen that, the number of parameters decreases with a smaller truncated level and a larger sliding window size. Then, we compare the performance of CF-PSNet with different hyper-parameter choices in terms of averaging RMSE of all five prediction tasks. In Table III, the best performance of each row improves when we increase the sliding window size, since a larger W may lead to a larger receptive field, more existing visits and lighter fully group connected layers. In most cases, the PS of a higher level typically characterizes more trivial details and does not leads to further improvements. Based on the discussion above, we selected $W = 5$ and $l = 2$ in following experiments.

To validate the effectiveness and efficiency of our TPS layer, we select some well-known sequence encoders (i.e., Bi-LSTM [48], GRU [49] and Transformer [47]) as our substitutes. In practice, we replaced each TPS layer in Fig. 3 with single layer of other substitutes to have a fair comparison. In Table I, we found out that our path-signature-based method shows superior performances in various metrics without introducing extra parameters under the SSS problem.

C. Comparison on the Top K Selection Module

On the top of CF-PSNet, a pooling layer is proposed to select the most influential brain regions for each cognitive functions and eliminate redundancy. DiffPool [50] are selected

to be compared with the top K selection module as both of them do not require the adjacent information which is lacked in structural MRI data. In Table IV, we find that the top K selection module and DiffPool both effectively decrease the number of parameters. Nonetheless, the top K selection module outperforms DiffPool by a large margin. The reason might be that the former one preserves the cortical structural information by retaining the features of top K influential ROIs, whereas the latter generates K clusters using features from all brain regions, some of which may be useless or containing noises. Further, we test the top K selection module by varying the selection of K . In Fig. 8, the average RMSE for all five cognitive scores is illustrated. It can be observed that with more brain regions selected, the model needs more parameters to be optimized, while the performance is deteriorating. Of note, the top K selection module only contains 36 parameters to optimize, while the model size decreased for more than 90% after applying this module.

D. Model Analysis

In Fig. 9 and Fig. 10, several experiments are conducted to investigate the impact of feeding different number of visits and adjusting the missing rate on the model performance. Specifically, in Fig. 9, we use the CF-PSNet with single TPS layer as the model configuration, as it requires at least 9 time points to apply the stacked TPS layers. Notably, model

Fig. 11. The illustration of the error range and distribution of CF-PSNet in five runs. The horizontal and vertical axes are actual Mullen scores and predicted scores respectively.

TABLE VI

PERFORMANCE COMPARISON BETWEEN CF-PSNET AND BASELINES (IN TERMS OF RMSE). WE USED † TO DENOTE THE EXPERIMENT CONFIGURATION ADOPTED IN [12].

Category	Methods	VRS	FMS	RLS	ELS	ELC	Average
Multi-task	TrMTL [51]	0.279	0.276	0.192	0.217	0.136	0.220
	MtJFS [52]	0.276	0.273	0.189	0.214	0.134	0.217
	NN [53]	0.219	0.259	0.165	0.196	0.182	0.204
	RMTL [54]	0.146	0.200	0.178	0.188	0.137	0.170
	LPMvRL [13]	0.162	0.189	0.139	0.165	0.138	0.158
	CF – PSNet(<i>multi – task</i>)	0.117	0.120	0.148	0.096	0.079	0.117
Multi-task†	MTMLR [12]	0.170	0.180	0.200	0.170	0.170	0.178
	CF – PSNet(<i>multi – task</i>)	0.116	0.114	0.161	0.122	0.069	0.117
Single-task	BrainPSNet [22]	0.046	0.075	0.095	0.063	0.057	0.067
	CF – PSNet	0.008	0.019	0.036	0.026	0.026	0.023

achieves the best performance when we feed in time points from 0M to 18M after birth. One possible explanation is the excessive missing rates in the 24M and 36M as illustrated in Fig. 2. Further, we try to adjust missing rate by randomly drop existing visits. The initial missing rate is around 0.306, while we add 2% missing visits each time. A sharp decrease on the model performance can be observed as the missing rate grows, while the performance with existence embedding is more steady suggesting its effectiveness. Three different runs are conducted for the reliability.

Ablation study. Further, an ablation study was conducted to explore the ability of several network modules in CF-PSNet. Best parameter setting is used based on foregoing experiments. According to Table V, all modules of CF-PSNet contribute to the final result. The top K selection module we used successfully remove the redundant brain regions. We introduce multi-scale temporal information by TPS layers in Stream II and III which are effective in predicting cognitive scores.

E. Comparison with the State-of-the-art Methods

Several recent algorithms are selected as baselines to validate the performance of our method, including: 1) NN (nearest neighbour) [53]; 2) MtJFS (Multi-Task Learning with Joint Feature Selection) [52]; 3) TrMTL (Trace-Norm Regularized Multi-Task Learning) [51]; 4) RMTL (Robust Multi-Task Feature Learning) [54], 5) MTMLR (Multi-Task Multi-Linear Regression) [12] and 6) LPMvRL (Latent Partial Multi-view Representation Learning) [13]. It is noteworthy that in these methods, five cognitive scores are learned simultaneously with the belief that these scores are essentially inter-related and can benefit each other for the prediction tasks. Hence, we proposed *CF – PSNet(multi – task)* to predict these five scores at once and have a fair comparison. From Table VI, we find that *CF – PSNet(multi – task)* outperforms the other algorithms in three tasks and Average under the same settings which validates the performance of our method. Additionally, since MTMLR [12] used different morphological features in its experiments, we adjusted the data configuration accordingly, which is denoted as *Multi – task†* in Table VI. The huge margin between *CF – PSNet(multi – task)* and

Fig. 12. The illustration of the importance distribution among different brain regions w.r.t. five cognitive scores respectively. The color bars are used to evaluate the contribution of each brain region. The sum of the importance scores of all brain regions for each task is 1.

Fig. 13. The illustration of influence of anatomical features w.r.t. five Mullen scores.

CF - PSNet may indicate that the important features for these five tasks are not the same and some task-specific brain regions exist. We conducted the paired t-test between the result using model in our conference version [22] and *CF-PSNet*. The p-values are small (< 0.05) in most tasks, i.e., 0.029 for VRS, 0.001 for FMS, 0.674 for RLS, 0.001 for ELS, 0.001 for ELC proving the improvement is significant. The scatter plots of 10 different runs for predicting the cognitive scores are depicted in Fig. 11 demonstrating that the scores are predicted fairly good.

In general, we conclude our experimental results as follows. 1) Features of multiple temporal resolution may contribute to the predicting accuracy of cognitive scores. TPS layers can effectively and efficiently extract the dynamic information hidden in cortical morphological growth trajectories. 2) Every cognition function may be associated with few brain regions.

Most of ROIs are redundant. 3) Further, there are some general and task-specific brain regions for these five tasks.

VII. RESULT ANALYSIS

In this subsection, we aim to investigate the relationship between cerebral cortical structure and cognition functions in early postnatal stages using testing data and well trained models. Motivated by [55], we first try to analyze the most important morphological features for each cognitive function by computing the gradient of features of input data w.r.t. the final loss with backpropagation. Within this procedure, morphological features with higher gradients are regarded to be paid more attention by the neural network. As a result, in Fig. 13, we can find that curvature collects more attention of our model for all cognitive functions. The abbreviations of features and labels have been introduced in Section IV.

Additionally, as we introduce a top K selection module in *CF-PSNet* to select brain regions for each cognitive function, we used top 10 influential ROIs ranked by the TopKPooling layer of each infant to make a vote and obtained a set of influence scores for 70 ROIs for each tasks. Then, the Pearson correlation coefficients are calculated between influence scores distributions of two hemispheres, each with 35 ROIs. Consequently, the Pearson coefficients for VRS, FMS, RLS, ELS and ELC are 0.2677, 0.1431, 0.2761, 0.3461, and 0.1825 respectively, which may indicate that there is a difference between the influence of these two hemispheres on the cognitive abilities [56]. In Fig. 12, the distribution of influence scores among different ROIs are depicted³. If we compared the most important brain regions among five cognitive abilities, it can be observed that the influence scores

³The indices of ROIs: <https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation>

distribution of RLS, ELS and FMS are similar, while they can be distinguished with the distribution of VRS and ELC easily, which validates the above assumption that the most important brain regions of different cognitive abilities may not be the same [6].

Notably, for the convenience of illustration, in Fig. 12, we summed up the influence scores for corresponding ROIs in the left and right hemispheres. It can be noticed that, the top 5 important brain regions in Fig.12 for five cognitive functions are distributed in 10 of 35 ROIs, shown in Table VII for details. The FMS, which measures a child's ability to manipulate small objects with one hand (unilaterally) and two hands (bilaterally) [57], is closely relevant with the superior temporal sulcus, fusiform gyrus, frontal cortex, parahippocampal gyrus and pericalcarine cortex in our results. Previous researches have observed a significant activation in the fusiform and parahippocampal gyri of toddlers in a finger-tapping fine motor task [58], which is consistent with our work. Besides, individuals with movement disorder showed higher gyrification in the superior temporal sulcus and lateral frontal cortex [59], which indirectly confirmed their relationship. It also can be found that the banks of the superior temporal sulcus, pericalcarine, and lateral orbital frontal cortex have a tight correlation with the ELS function, which is constructed to measure speaking ability and language formation. In recent advances [60], [61], a similar conclusion has been drawn by analyzing the fMRI connectivity with the expressive language tasks in young children and adolescents. Further, a significant correlation has been revealed between the receptive language related functions (consistent functions with RLS) and the left fusiform and bilateral posterior cingulate with age increasing [62]. Moreover, researchers also found that the congruous words would elicit extensive response in the lateral orbital frontal cortex measured by RLS [63]. These observations may indicate the intrinsic correlation between the RLS with the fusiform, posterior cingulate, and lateral frontal cortex. Additionally, in previous studies [64], [65], the entorhinal cortex volume has been verified as a strong indicator of the VRS and RLS functions impairment in Alzheimer's disease, which may implicate the same findings as shown in Table VII. To sum up, the most influential brain regions for each cognition function we selected are generally consistent with previous works, which makes this study more convincing.

VIII. CONCLUSION

In this paper, we propose a novel model, so-called CF-PSNet, using longitudinal cortical features of infants to predict their cognitive scores. A path-signature-based layer, i.e. TPS layer, is introduced to explore the anatomical and geometric properties hidden in the cortical developmental trajectories. Based on hierarchical temporal features, a multi-stream model is constructed to combine information of various time points and time scales and generate an informative representation for each participant. By applying the existence embedding, the ambiguity and artifacts caused by interpolation is certainly removed. Further, prior knowledge is used to help reduce the hypothesis space for SSS problem. Considering

TABLE VII

THE TOP FIVE INFLUENTIAL ROIS FOR EACH COGNITION FUNCTION IN FIG. 12. AS THREE BRAIN REGIONS SHARE EQUAL INFLUENCE COEFFICIENT, THERE ARE SEVEN ROIS LISTED FOR RLS.

ROIs	VRS	FMS	RLS	ELS	ELC
Banks of the superior temporal sulcus		✓	✓	✓	
Entorhinal cortex	✓		✓		
Fusiform gyrus	✓	✓	✓	✓	✓
Inferior parietal cortex	✓				
Lateral orbital frontal cortex	✓	✓	✓	✓	✓
Lingual gyrus			✓		
Parahippocampal gyrus	✓	✓		✓	✓
Pars triangularis					✓
Pericalcarine cortex		✓	✓	✓	✓
Posterior-cingulate cortex			✓		

different ROIs' influence on cognition abilities along time, we use a top K selection module and an attention mask generator to respectively select ROIs on global and weight them differently at corresponding developmental stages. The experimental result shows the state-of-the-art performance of our method. Besides it illustrates the impact of missing rate and the trade-off between model size and the dimensionality of features, which validates the great potential of the integration of machine learning and path signature method in longitudinal brain analysis. Additionally, in this study, the relationship between cortical morphology and cognition functions is also presented.

REFERENCES

- [1] H. C. Hazlett *et al.*, "Early brain development in infants at high risk for autism spectrum disorder," *Nature*, vol. 542, no. 7641, pp. 348–351, 2017.
- [2] I. Rekik, G. Li, P.-T. Yap, G. Chen, W. Lin, and D. Shen, "Joint prediction of longitudinal development of cortical surfaces and white matter fibers from neonatal mri," *NeuroImage*, vol. 152, pp. 411–424, 2017.
- [3] S. C. Jha *et al.*, "Environmental influences on infant cortical thickness and surface area," *Cerebral Cortex*, vol. 29, no. 3, pp. 1139–1149, 2019.
- [4] F. Wang *et al.*, "Developmental topography of cortical thickness during infancy," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 855–15 860, 2019.
- [5] J. Kagan and N. Herschkowitz, *A young mind in a growing brain*. Psychology Press, 2006.
- [6] J. B. Girault *et al.*, "Cortical structure and cognition in infants and toddlers," *Cerebral cortex*, vol. 30, no. 2, pp. 786–800, 2020.
- [7] D. C. Dean III *et al.*, "Estimating the age of healthy infants from quantitative myelin water fraction maps," *Human brain mapping*, vol. 36, no. 4, pp. 1233–1244, 2015.
- [8] L. Wang *et al.*, "Links: Learning-based multi-source integration framework for segmentation of infant brain images," *NeuroImage*, vol. 108, pp. 160–172, 2015.
- [9] Y. Meng, G. Li, Y. Gao, W. Lin, and D. Shen, "Learning-based subject-specific estimation of dynamic maps of cortical morphology at missing time points in longitudinal infant studies," *Human brain mapping*, vol. 37, no. 11, pp. 4129–4147, 2016.
- [10] M. D. Shen *et al.*, "Increased extra-axial cerebrospinal fluid in high-risk infants who later develop autism," *Biological psychiatry*, vol. 82, no. 3, pp. 186–193, 2017.
- [11] H. Yang, F. Wang, and P. Jiang, "Accurate anatomical landmark detection based on importance sampling for infant brain MR images," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 5, pp. 1078–1086, 2017.
- [12] E. Adeli, Y. Meng, G. Li, W. Lin, and D. Shen, "Multi-task prediction of infant cognitive scores from longitudinal incomplete neuroimaging data," *NeuroImage*, vol. 185, pp. 783–792, 2019.

- [13] C. Zhang, E. Adeli, Z. Wu, G. Li, W. Lin, and D. Shen, "Infant brain development prediction with latent partial multi-view representation learning," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 909–918, 2018.
- [14] G. Li, L. Wang, F. Shi, J. H. Gilmore, W. Lin, and D. Shen, "Construction of 4d high-definition cortical surface atlases of infants: Methods and applications," *Medical image analysis*, vol. 25, no. 1, pp. 22–36, 2015.
- [15] E. Braaten, *The SAGE encyclopedia of intellectual and developmental disorders*. SAGE Publications, 2018.
- [16] T. J. Lyons, "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.
- [17] T. Lyons, H. Ni, and H. Oberhauser, "A feature set for streams and an application to high-frequency financial tick data," in *Proceedings of the 2014 International Conference on Big Data Science and Computing*, 2014, pp. 1–8.
- [18] S. Lai, L. Jin, and W. Yang, "Toward high-performance online hccr: A cnn approach with dropdistortion, path signature and spatial stochastic max-pooling," *Pattern Recognition Letters*, vol. 89, pp. 60–66, 2017.
- [19] M. Liu, L. Jin, and Z. Xie, "Ps-lstm: Capturing essential sequential online information with path signature and lstm for writer identification," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, 2017, pp. 664–669.
- [20] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, and J. Chang, "Leveraging the path signature for skeleton-based human action recognition," *arXiv preprint arXiv:1707.03993*, vol. 1, 2017.
- [21] C. Li, X. Zhang, L. Liao, L. Jin, and W. Yang, "Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8585–8593.
- [22] X. Zhang *et al.*, "Infant cognitive scores prediction with multi-stream attention-based temporal path signature features," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 134–144.
- [23] M.-s. Choe *et al.*, "Regional infant brain development: an MRI-based morphometric analysis in 3 to 13 month olds," *Cerebral Cortex*, vol. 23, no. 9, pp. 2100–2117, 2013.
- [24] D. Holland *et al.*, "Structural growth trajectories and rates of change in the first 3 months of infant brain development," *JAMA neurology*, vol. 71, no. 10, pp. 1266–1274, 2014.
- [25] S. J. Lee *et al.*, "Common and heritable components of white matter microstructure predict cognitive function at 1 and 2 y," *Proceedings of the National Academy of Sciences*, vol. 114, no. 1, pp. 148–153, 2017.
- [26] G. Ball *et al.*, "Thalamocortical connectivity predicts cognition in children born preterm," *Cerebral cortex*, vol. 25, no. 11, pp. 4310–4318, 2015.
- [27] A. Carlson, K. Xia, M. Azcarate-Peril, B. Goldman, and R. C. Knickmeyer, "Infant gut microbiome associated with cognitive development," *Biological Psychiatry*, vol. 83, pp. 148–159, 2018.
- [28] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 591–606, 2008.
- [29] K.-T. Chen, "Integration of paths—a faithful representation of paths by noncommutative formal power series," *Transactions of the American Mathematical Society*, vol. 89, no. 2, pp. 395–407, 1958.
- [30] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Annals of Mathematics*, pp. 109–167, 2010.
- [31] H. Boedihardjo, X. Geng, T. Lyons, and D. Yang, "The signature of a rough path: uniqueness," *Advances in Mathematics*, vol. 293, pp. 720–737, 2016.
- [32] J. Diehl, "Rotation invariants of two dimensional curves based on iterated integrals," *arXiv preprint arXiv:1305.6883*, 2013.
- [33] T. J. Lyons, M. Caruana, and T. Lévy, *Differential equations driven by rough paths*. Springer, 2007.
- [34] D. Levin, T. Lyons, and H. Ni, "Learning from the past, predicting the statistics for the future, learning an evolving system," *arXiv preprint arXiv:1309.0260*, 2013.
- [35] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," *arXiv preprint arXiv:1603.03788*, 2016.
- [36] M. Woodburn *et al.*, "The maturation and cognitive relevance of structural brain network organization from early infancy to childhood," *NeuroImage*, p. 118232, 2021.
- [37] G. Li, L. Wang, F. Shi, W. Lin, and D. Shen, "Simultaneous and consistent labeling of longitudinal dynamic developing cortical surfaces in infants," *Medical image analysis*, vol. 18, no. 8, pp. 1274–1289, 2014.
- [38] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in mri data," *IEEE transactions on medical imaging*, vol. 17, no. 1, pp. 87–97, 1998.
- [39] Q. Zhang, L. Wang, X. Zong, W. Lin, G. Li, and D. Shen, "Frnet: Flattened residual network for infant mri skull stripping," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 999–1002.
- [40] L. Wang *et al.*, "Volume-based analysis of 6-month-old infant brain mri for autism biomarker identification and early diagnosis," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 411–419.
- [41] L. Sun *et al.*, "Topological correction of infant white matter surfaces using anatomically constrained convolutional neural network," *NeuroImage*, vol. 198, pp. 114–124, 2019.
- [42] G. Li *et al.*, "Consistent reconstruction of cortical surfaces from longitudinal brain mr images," *Neuroimage*, vol. 59, no. 4, pp. 3805–3820, 2012.
- [43] —, "Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age," *Journal of Neuroscience*, vol. 34, no. 12, pp. 4228–4238, 2014.
- [44] A. E. Lyall *et al.*, "Dynamic development of regional cortical thickness and surface area in early childhood," *Cerebral cortex*, vol. 25, no. 8, pp. 2204–2212, 2015.
- [45] G. Li *et al.*, "Computational neuroanatomy of baby brains: A review," *NeuroImage*, vol. 185, pp. 906–925, 2019.
- [46] Z. Wu, L. Wang, W. Lin, J. H. Gilmore, G. Li, and D. Shen, "Construction of 4D infant cortical surface atlases with sharp folding patterns via spherical patch-based group-wise sparse representation," *Human brain mapping*, vol. 40, no. 13, pp. 3860–3880, 2019.
- [47] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [49] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [50] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *arXiv preprint arXiv:1806.08804*, 2018.
- [51] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 457–464.
- [52] A. Evgeniou and M. Pontil, "Multi-task feature learning," vol. 19, 2007, p. 41.
- [53] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [54] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 42–50.
- [55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.
- [56] J. Cernáček, "Ontogenic aspects of functional asymmetry of the hemispheres," *Ceskoslovenska neurologie a neurochirurgie*, vol. 53, no. 3, pp. 151–155, 1990.
- [57] R. Dumont, C. L. Cruse, V. Alfonso, and C. Levine, "Book review: Mullen scales of early learning: Aps edition," *Journal of Psychoeducational Assessment*, vol. 18, no. 4, pp. 381–389, 2000.
- [58] E. Redle *et al.*, "Functional mri evidence for fine motor praxis dysfunction in children with persistent speech disorders," *Brain Research*, vol. 1597, pp. 47–56, 2015.
- [59] A. Treble, J. Juranek, K. K. Stuebing, M. Dennis, and J. M. Fletcher, "Functional significance of atypical cortical organization in spina bifida myelomeningocele: relations of cortical thickness and gyrification with iq and fine motor dexterity," *Cerebral Cortex*, vol. 23, no. 10, pp. 2357–2369, 2013.
- [60] K. Blackmon *et al.*, "Focal cortical anomalies and language impairment in 16p11.2 deletion and duplication syndrome," *Cerebral Cortex*, vol. 28, no. 7, pp. 2422–2430, 2018.
- [61] V. Youssofzadeh, J. Vannest, and D. S. Kadis, "fmri connectivity of expressive language in young children and adolescents," *Human brain mapping*, vol. 39, no. 9, pp. 3586–3596, 2018.

- [62] J. Zhuang, M. A. Johnson, D. J. Madden, D. M. Burke, and M. T. Diaz, "Age-related differences in resolving semantic and phonological competition during receptive language tasks," *Neuropsychologia*, vol. 93, pp. 189–199, 2016.
- [63] M. Friedrich and A. D. Friederici, "Early n400 development and later language acquisition," *Psychophysiology*, vol. 43, no. 1, pp. 1–12, 2006.
- [64] L.-K. Yeung *et al.*, "Object-in-place memory predicted by anterolateral entorhinal cortex and parahippocampal cortex volume in older adults," *Journal of Cognitive Neuroscience*, vol. 31, no. 5, pp. 711–729, 2019.
- [65] M. Di Paola *et al.*, "Episodic memory impairment in patients with alzheimer's disease is correlated with entorhinal cortex atrophy," *Journal of neurology*, vol. 254, no. 6, pp. 774–781, 2007.