Connectome Transformer with Anatomically Inspired Attention for Parkinson's Diagnosis

Diego Machado-Reyes
machad@rpi.com
Biomedical Engineering, Rensselaer Polytechnic Institute
Troy, New York, USA

Mansu Kim
mansu.kim@catholic.ac.kr
Department of Artificial Intelligence, Catholic University of Korea
Bucheon, Republic of Korea

Hanqing Chao
chaoh@rpi.edu
Biomedical Engineering, Rensselaer Polytechnic Institute
Troy, New York, USA

Li Shen
li.shen@pennmedicine.upenn.edu
Dept. of Biostatistics, Epidemiology & Informatics, Univ. of Pennsylvania
Philadelphia, Pennsylvania, USA

Pingkun Yan*
yanp2@rpi.edu
Biomedical Engineering, Rensselaer Polytechnic Institute
Troy, New York, USA

ABSTRACT

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease in the United States. The structural or functional connectivity between regions of interest (ROIs) in the brain and their changes captured in brain connectomes could be potential biomarkers for PD. To effectively model the complex non-linear characteristic connectomic patterns related to PD and exploit the long-range feature interactions between ROIs, we propose a connectome transformer model for PD patient classification and biomarker identification. The proposed connectome transformer learns the key connectomic patterns by leveraging the global scope of the attention mechanism guided by an additional skip-connection from the input connectome and the local level focus of the CNN techniques. Our proposed model significantly outperformed the benchmarking models in the classification task and was able to visualize key feature interactions between ROIs in the brain.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; *Health informatics*; *Imaging*.

KEYWORDS

Brain connectivity, Parkinson's disease, Deep learning, Transformer

*Correspondence to Li Shen and Pingkun Yan. Data used in the work were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (ppmi-info.org). This work was supported in part by 1) NIH [T32 GM067545] training fellowship to DMR, 2) NIH [U01 AG068057, R01 AG066833] and NSF [IIS 1837964] to LS, 3) NSF [OAC 2046708] to PY.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
BCB '22, August 7–10, 2022, Northbrook, IL, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9386-7/22/08…$15.00
https://doi.org/10.1145/3535508.3545544

ACM Reference Format:
Diego Machado-Reyes, Mansu Kim, Hanqing Chao, Li Shen, and Pingkun Yan. 2022. Connectome Transformer with Anatomically Inspired Attention for Parkinson's Diagnosis. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3535508.3545544

1 INTRODUCTION

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease in the United States, affecting 2-3% of the population ≥ 65 years of age [14]. Brain imaging has increasingly been used for PD and related disorders diagnosis [2], from which brain connectomes can be extracted and used to study brain disorders [19]. The alterations in the connectivity patterns could be indicative of an underlying disease such as PD [6]. However, brain connectivity analysis can be very challenging due to its complexity, large size and sparsity. Traditional computational approaches such as machine learning (ML), multi-layer perceptrons (MLP) and convolutional neural networks (CNN) struggle when dealing with highly sparse data with many input features. Therefore, in this work, we propose a novel transformer model for PD diagnosis and biomarker identification from structural connectome data.

The proposed model integrates a connectome-encoding convolutional layer and a joint attention mechanism incorporating anatomical guidance, in order to capture critical PD-related brain connectivity patterns and create highly representative embeddings from brain connectivity data. Combining both techniques allows the proposed model to capture local and global patterns in the data. Previous approaches using CNNs have achieved good results [9]. However, CNNs have a limited receptive field and struggle to capture meaningful patterns in long-range connections. Similarly, traditional MLPs become prohibitively computationally expensive when dealing with very large inputs due to their fully connected architecture. On the other hand, transformer models have the potential to learn complex structures from brain connectivity data due to their ability of capturing long-range interactions among the input features. Nevertheless, transformer models struggle to capture local

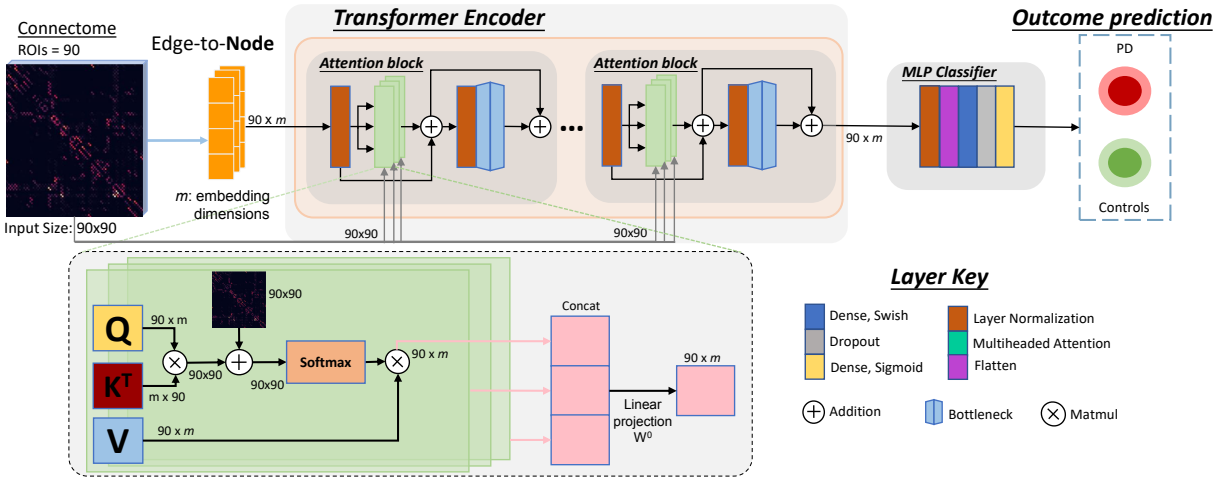

Figure 1: Proposed connectome transformer encoder model and classifier.

patterns due to the lack of induction bias. More recent visual transformer methods have incorporated techniques from CNN models, such as convolutional filters and pooling [12].

A key challenge in the field is to quantify the learned relationships between ROIs to understand the underlying disease mechanisms; moreover it is relevant to capture the local level structures from the information-rich connectome to maximize the learned patterns. In order to alleviate these challenges, our proposed model implements the following technical contributions. (1) Our method introduces transformers into brain network analysis to efficiently learn and explicitly identify the complex relationships between regions of interest, which also provides interpretability of the DL model. (2) Our method integrates brain connectivity-specific CNN techniques into the transformer model to leverage the local focus capabilities of CNNs and enhance the global feature learning process of transformers. (3) A joint attention mechanism is proposed to directly include the essential features from the original anatomical input into the computed attention to fully utilize the brain connectivity.

Our proposed method was used to classify PD patients from healthy controls and the learned biomarkers connections were corroborated in the literature. Our method was validated using a landmark PD biobank: the Parkison's Progression Markers Initiative (PPMI). Our model was able to achieve promising imaging biomarker identification and prediction accuracy in the PD classification task, outperforming traditional ML methods and CNNs. More importantly, the interpretability of our model enables identifying a set of clinically relevant imaging biomarkers and interaction patterns that provide valuable insights into the underlying biological mechanisms of PD. It may lead the community to form new hypotheses for subsequent molecular and clinical investigations.

2 CONNECTOME TRANSFORMER

The proposed connectome transformer encoder and classifier consist of three main stages as shown in Fig. 1. First, each input connectome is encoded into an initial representation through an edge-to-node layer [9]. Next, in order to learn PD-relevant connectivity patterns, the connectome encoded representation is used as input to the transformer encoder. In this module, we use skip connections from the input connectome to the multi-head attention (MHA) mechanisms (shown in green) which directly incorporate connectivity data to the attention mechanisms. Finally, an MLP is used to classify the learned embedded representations from each input connectome as PD or healthy control (HC).

2.1 Connectivity Tokenization

Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ represent the input of our transformer, where n is the number of ROIs. The edge-to-node layer aims to obtain highly informative tokens (*i.e.* feature/ROI representations) from the input connectomes. The edge-to-node layer was implemented following the published code [10]. This implementation can be simply regarded as a column-wise 1D convolution with kernel size $1 \times n$ over the input connectome with m filters to obtain n tokens of embedding dimension m. The tokenization can be formulated as: $\boldsymbol{t_i} = W_t \boldsymbol{x}_{i,\cdot} + \mathbf{b}$ where $W_t \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

2.2 Transformer Encoder

The transformer encoder is constructed by several repeated attention blocks (see Fig. 1). Each of these blocks contains a normalization layer followed by a multi-head joint attention layer and a skip connection from the normalization layer to the output of the attention block. Then, a second normalization layer is applied and the output is passed to a feed-forward block (denoted as the bottleneck icon).

Each head j in a MHA block first generates query $Q^j = \{\boldsymbol{q}_i^j\}_{i=1}^n$, key $K^j = \{\boldsymbol{k}_i^j\}_{i=1}^n$, and value $V^j = \{\boldsymbol{v}_i^j\}_{i=1}^n$ vectors for each of the ROIs, which can be calculated by $\boldsymbol{q}_i^j = W_Q^j \boldsymbol{t}_i$, $\boldsymbol{k}_i^j = W_K^j \boldsymbol{t}_i$, and $\boldsymbol{v}_i^j = W_V^j \boldsymbol{f}_i$ respectively, where $W_Q^j, W_K^j \in \mathbb{R}^{m \times d_k}$, and $W_V^j \in \mathbb{R}^{m \times d_v}$ are learnable parameters. Moreover, each head j receives the input connectome X through a skip connection. Then, for each query, an output is calculated as a weighted sum of all the value vectors, where the weights are computed as the similarity between the query and each key plus the input connectome. Such an operation enables

the MHA block to aggregate information across all ROIs according to the query. The output of the j-th attention head is computed as: $\text{head}^j(\boldsymbol{q}_i^j, K^j, V^j, X) = \text{softmax}\left(\frac{\boldsymbol{q}_i^{jT} K^j}{\sqrt{d_k}} + X\right) V^T$.

The input connectomes are added to the raw attention scores obtained from the query and key multiplication at the MHA layer as illustrated in Fig. 1 and described in the equation above. This anatomically-guided joint attention is applied at each of the attention blocks of the transformer encoder. The intuition behind the skip connections from the input connectome is to keep the relevant features that could be lost at deeper stages of the network, considering that the connectomes already contain key feature interaction information similar to the raw attention. Thus, an anatomically guided joint attention can be obtained. This process is analogous to the U-Net [17], where the low-level image features can be directly passed to later part of the decoder to keep the information.

The outputs of all attention heads are then concatenated and projected to get the final output of the MHA block, $\text{MHA}(\boldsymbol{q}_i^j, K^j, V^j) = \text{concat}(head^1, ..., head^h)W_O$, where $W_O \in \mathbb{R}^{hd_v \times m}$ is a learnable projection matrix. Since each head has respective parameters, the MHA block is able to jointly consider different types of correlations in the input feature [18]. The feed-forward block is an inverse bottleneck structure composed of two dense layers with one Swish [16] activation function in between layers. The full process of the l-th layer in our transformer encoder is formulated as: $F_l' = \text{MHA}(\text{LN}(F_{l-1})) + F_{l-1}, F_l = \text{FF}(\text{LN}(F_l')) + F_l'$, where $\text{LN}(\cdot)$ is the normalization layer [1]. The output of the transformer encoder is flattened into a vector and fed to an MLP classifier.

3 RESULTS AND DISCUSSION

3.1 Dataset

Our study used the diffusion tensor imaging (DTI) data and the corresponding PD diagnosis from 153 participants from PPMI. The dataset presented a considerable imbalance with 112 PD patients and the remaining 41 as HC. The mean age for the PD patients was 61.05±9.28 and for the HC was 61.15±10.44. We processed the data using the tractography algorithm implemented in FSL probtrackX to extract fiber information and construct structural connectivity matrices (connectomes). Moreover, in order to remove the confounding variables from the population substructure, the mean connectivity matrix of the combined training and validation sets was calculated and subtracted from the training, validation, and testing set. Data augmentation using the local synthetic instances (LSI) method [3] was performed on the training set.

3.2 Classification results

Evaluation Strategy We used the area under the receiver operating characteristic curve (AUC) for performance evaluation. We ran 10×10-fold stratified cross-validation to ensure accurate results for the small and imbalanced dataset used to train the network (70% train, 10% validation 20% test). Random states for the data splitting were set for fair comparison across models by evaluating on the same data splits. Mean AUCs on the test partition were calculated for each one of the 10-fold experimental units and an overall mean of the mean AUCs is reported together with standard deviations.

Table 1: Comparison over 10 rounds of 10-fold cross-validation. Mean AUCs with significant difference (α=0.05 and 0.005) are denoted with '*' and '', respectively.**

Model	Mean AUC ± SD	Median AUC
PCA + RF	0.520 ± 0.035 **	0.509
PCA + SVM	0.527 ± 0.043 **	0.516
BrainNetCNN [9]	0.540 ± 0.021 **	0.538
Our model	**0.631 ± 0.039**	**0.644**
- Joint Attention	0.588 ± 0.029 *	0.604
- E2N layer	0.563 ± 0.028 **	0.568
- E2N, - Joint attention	0.570 ± 0.052 **	0.558
- LSI	0.530 ± 0.032 **	0.532

Paired samples Wilcoxon test was used for significance testing for comparison between proposed and baseline models ($\alpha = 0.05$).

Baselines The proposed transformer encoder model was compared against well-established ML and recent DL models. The former includes random forest (RF) and support vector machine (SVM) as classifiers, where principal component analysis (PCA) was first applied for feature dimensionality reduction. RF and SVM were implemented using Sci-Kit Learn [15] and a basic systematic hyperparamter tuning. On the other hand, BrainNetCNN [9] was used for the DL category, following the public Keras-TensorFlow implementation [5]. Our proposed model was trained using focal loss [11] and AdamW [13] optimizer. In our work, focal loss was employed to handle the imbalance in positive and negative cases. The focal loss has a great capability of dealing with imbalanced datasets and focusing on hard negative samples for more robust feature learning. The model architecture was implemented using 2 attention blocks with 12 heads each, bottleneck embedding dimensions of 512, learning rate of 0.00001, and weight decay of 0.0001. The model was trained with batch size of 8, and 100 epochs. The model was implemented using Keras Tensorflow v.2.4.

Results Table 1 shows the proposed model performed significantly better than the competing methods, achieving a mean AUC of 0.633 and median AUC of 0.644, almost a 10% increase over the highest performing baseline model. BrainNetCNN achieved the next best performance with a mean AUC of 0.540, followed by SVM and RF. The classification results show the proposed model can effectively learn informative representations and capture PD-related patterns from the connectomes. This is expected as traditional ML models such as RF and SVM struggle at capturing complex non-linear interaction patterns. Furthermore, BrainNetCNN did not perform as well as our proposed model probably due to its limitation to a local level focus and the lack of capabilities for capturing long-range interactions between the ROIs. Our model is able to capture the complex patterns and feature interactions at local and global levels due to its transformer backbone and convolutional tokenization.

3.3 Ablation studies

The ablated components were the edge-to-node (E2N) tokenizer, the skip connections from input connectomes to the MHA modules (denoted as joint attention), and the data augmentation technique LSI. As it can be seen in the bottom section of Table 1, while the

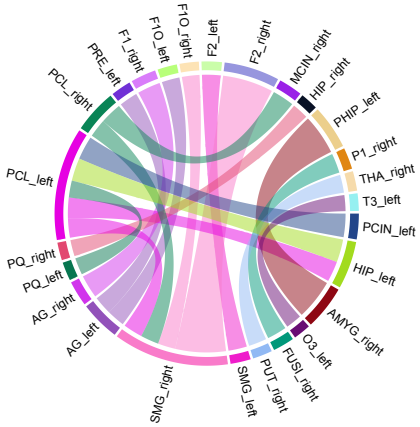

Figure 2: Chord plots of attention scores for the top fold showing the learned ROI interactions.

absence of the joint attention decreases the performance of the model, the model performs its lower when the E2N layer is removed. This is expected as the joint attention should aid the model to maintain key features from the input connectome in the deeper stages of the network; however, it is not essential to the feature relationship learning at the MHA mechanism. On the other hand, the E2N layer, plays a vital role in the network, as it provides the original feature encodings. While transformers are highly proficient at capturing the global interactions in the feature set, they struggle at the local level. Moreover, the lack of both E2N and joint attention layers performs better than the absence of the E2N layer alone. The lack of E2N layer would lead to using the original connectome as input, rendering the skip connections less useful. Therefore, the absence of both strategies could lead to a simplified model that performs better than the one without the E2N layer but that includes the connectome skip connection. Finally, as transformer models are known to be data-hungry due to their very large number of parameters, the sample size is key to their performance. The absence of the data augmentation technique, LSI, significantly hurdles the proposed model performance.

3.4 Support by known mechanisms

The visualization of the learned attention scores by our model allowed for key insight into the learned relationships between features. The top performing model (i.e. highest AUC in the test set) from the 10×10-fold cross-validation framework is visualized in Fig. 2. The joint attention scores were first averaged across the samples in the test set and then the maximum values from the resulting mean attention matrices were selected for a final model joint attention matrix. Chord plot was drawn using the circlize R library [7]. Fig. 2 shows several key feature connections captured by the model signaling possible biological relationships between ROIs towards PD development. The proposed model detected strong connections between the angular gyrus (AG - left and right) and the superior frontal gyrus (F1O - left and right). Previous clinical literature has associated these regions with altered glucose metabolism for PD patients and general cognitive decline [4]. Moreover, previous findings indicate that the amygdala (AMYG) and parahippocampal gyrus (PHIP) may play a key role in PD patients' cognitive-emotional deficits [20]. Finally, the proposed model found multiple strong connections involving the supramarginal gyrus (SMG) and ROIs in the frontal lobe (*i.e.*, PCL, F1O, F2), which have been shown to be important biomarkers for PD pathophysiology [8].

4 CONCLUSION

The ever-growing number of people affected by PD coupled with the limited understanding of the pathogenesis and biomarkers for PD detection raises the importance of developing models that allow for PD detection while providing insight into the key biomarkers behind PD development. In this work, a transformer encoder coupled with CNN techniques is proposed to classify PD patients from healthy controls, while simultaneously providing insight into the learned relationships between the ROIs analyzed. While few studies have been done for PD patient classification and biomarker identification using diffusion MRI-derived connectomes; our results show the predictive power of the structural connectivity data and shed light on future directions integrating connectomes with other data modalities for robust PD detection and biomarker identification.

One limitation of this work is the small sample size that limits the capability of the models to learn multiple generalizable features and consequently the limited classification capabilities of the models. Nevertheless, the proposed model was able to provide key insights into the underlying mechanisms for PD development and putative biomarkers for further clinical analysis. For future work, we will integrate different data modalities to achieve an enhanced view of the PD landscape.

REFERENCES

[1] Jimmy Lei Ba et al. 2016. Layer Normalization. *arXiv:1607.06450 [cs, stat]* (July 2016). arXiv: 1607.06450.
[2] Natasha S. R. Bidesi et al. 2021. The role of neuroimaging in Parkinson's disease. *Journal of Neurochemistry* 159, 4 (2021), 660–689.
[3] Colin J. Brown et al. 2015. Prediction of Motor Function in Very Preterm Infants. In *MICCAI 2015 (Lecture Notes in Computer Science)*. Cham, 69–76.
[4] Qiu-Yue Dong et al. 2021. Glucose metabolism a potential biomarker for subjective cognitive decline. *Alzheimer's Research & Therapy* 13, 1 (April 2021), 74.
[5] Amine Echraibi. 2017. BrainCNN. https://github.com/AmineEchraibi/BrainCNN.
[6] Alex Fornito et al. 2015. The connectomics of brain disorders. *Nature Reviews Neuroscience* 16, 3 (March 2015), 159–172.
[7] Zuguang Gu et al. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 19 (Oct. 2014), 2811–2812.
[8] Carl D. Hacker et al. 2012. Resting state functional connectivity of the striatum in Parkinson's disease. *Brain: A Journal of Neurology* 135, Pt 12 (2012), 3699–3711.
[9] Jeremy Kawahara et al. 2017. BrainNetCNN: Convolutional neural networks for brain networks. *NeuroImage* 146 (Feb. 2017), 1038–1049.
[10] J Kawahara et al. 2019. ann4brains. github.com/jeremykawahara/ann4brains.
[11] Tsung-Yi Lin et al. 2018. Focal Loss for Dense Object Detection. *arXiv:1708.02002 [cs]* (Feb. 2018). arXiv: 1708.02002.
[12] Yang Liu et al. 2021. A Survey of Visual Transformers. *arXiv:2111.06091 [cs]* (Nov. 2021). arXiv: 2111.06091.
[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]* (Jan. 2019). arXiv: 1711.05101.
[14] C. Marras et al. 2018. Prevalence of Parkinson's disease across North America. *npj Parkinson's Disease* 4, 1 (July 2018), 1–7.
[15] F. Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[16] Prajit Ramachandran et al. 2017. Searching for Activation Functions. *arXiv:1710.05941 [cs]* (Oct. 2017). arXiv: 1710.05941.
[17] Olaf Ronneberger et al. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015 (Lecture Notes in Comp Sci)*. 234–241.
[18] Ashish Vaswani et al. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.
[19] Fang-Cheng Yeh et al. 2021. Tractography Methods and Findings in Brain Tumors and Traumatic Brain Injury. *NeuroImage* 245 (Dec. 2021), 118651.
[20] N. Yoshimura et al. 2005. The amygdala of PD patients silent in response to fearful facial expressions. *Neuroscience* 131, 2 (2005), 523–534.