Predicting synesthetic experience from resting-state fMRI: a graph-based deep learning approach

Alex Noah Feldman[1], Aikaterini Melliou[2], Peiwu Qin[2], Yang Li[1], Ercan Engin Kuruoglu[1,*]
[1]iDI, SIGS, Tsinghua University, Shenzhen, China [2]iBHE, SIGS, Tsinghua University, Shenzhen, China

***Abstract*—Synesthesia is a phenomenon that causes sensory crossovers, such as tasting sounds or hearing colors. The presence of an underlying brain signature characterizing it still needs to be investigated. Currently, synesthesia detection relies exclusively on behavioral tasks, limiting both research and potential clinical applications. In this paper, we build various models predicting synesthesia, demonstrating that it can be predicted from Magnetic Resonance Imaging (MRI) and functional MRI (fMRI) recordings (modalities mapping functional and structural properties of the brain), which confirms the presence of distinct neural signatures underlying this condition. We employ Graph Neural Networks (GNNs) as they are the most natural choice for the fMRI data. We also integrate Kolmogorov-Arnold Networks (KANs) to explore their potential for interpreting the results as we aim to facilitate the identification of the underlying neurological biomarkers. Our approach achieves 93% accuracy in synesthesia detection.**

***Index Terms*—Synesthesia, MRI, fMRI, GNN, KAN**

I. Introduction

Synesthesia is a condition in which ordinary stimuli, such as digits or letters, induce concurrent experiences, such as colors. Past research has revealed the diverse neurological implications of synesthesia, from conferring memory advantages through enhanced color processing to sharing atypical sensory sensitivity with autism, while maintaining fundamentally different perceptual processes from seemingly similar experiences such as hallucinations [1, 2, 3]. These findings could lead to new insights into how the brain integrates and processes perceptual information, advancing our overall understanding of human perception and potentially improving treatments for related disorders. Synesthesia appears to result from unusual neurodevelopmental processes, involving genetic influences, distinctive brain connectivity patterns, and differences in cognitive and behavioral patterns. Therefore, investigating synesthesia involves analyzing brain function through both qualitative and quantitative approaches. For this research, we will employ two imaging techniques: traditional Magnetic Resonance Imaging (MRI) to examine brain structure, and functional MRI (fMRI) to study brain activity patterns [4].

MRI captures detailed images of anatomical structures using magnetic fields and radio waves, while fMRI measures blood flow changes in the brain, providing insights into neural activity during resting states or specific tasks and stimuli responses [5]. As a result, the data produced are three to five-dimensional arrays, either stand alone or in time series. In this work, we mostly focus on binary classification, given parcellations' correlations. We will also explore how Kolomogorov-Arnold Netoworks (KANs) can bind with GNNs, combining their respective strengths. Overall, we present an automated approach for detecting synesthesia from neuroimaging data, enabling faster and more objective diagnosis compared to traditional behavioral assessments. This automated classification framework could also significantly advance both clinical practice and research methodology in synesthesia studies, while demonstrating the potential of deep learning architectures for neurological condition detection and analysis.

Previous research has examined both the brain activity associated with synesthetic experiences and anatomical variations in synesthetes' brains. Yet, many of the reported neural differences between synesthetes and non-synesthetes lack solid evidence, primarily because of insufficient sample sizes, errors in statistical analysis, and constraints in research methodology [6]. Recent neuroimaging studies have advanced our understanding of the neural basis of synesthesia. While initial research focused on specific brain regions, evidence from [7] suggests that synesthetic experiences involve distributed networks rather than isolated areas. [8] demonstrated that primary visual cortex activation is not essential for word-induced synesthetic experiences, while [9] showed that non-linguistic auditory stimuli activate the left inferior parietal cortex, a region crucial for multi-modal integration. This collectively supports a network-based perspective of synesthetic processing. The most significant related work was performed by the authors of the dataset that we will use, which analyzed it and was able to classify synesthetes vs. non-synesthetes on biomarkers using simple machine learning algorithms; suggesting a distinct neural signature behind the synesthetic experience [10]. For fMRI image classification in general, the most direct way of processing is by means of graph deep learning models for a node connectivity type of representation. Other approaches, include processing spatial-temporal fMRI volumes. Such tensors can be fed through CNN or transformer based models.

Our paper is structured as follows: Section II shows the dataset structure and how the preprocessing is carried out, Section III elaborates on the methods tested and Section IV provides the respective results. Finally, Section V presents the future steps to be taken and Section VI concludes this manuscript.

*Corresponding author - kuruoglu@sz.tsinghua.edu.cn

II. Methods

The data used in our project consist of graphs, which in turn have nodes accompanied by features, explained in detail in the relevant section. We, thus, explore graph related algorithms.

We allocated 97 samples for the training, 15 for validating and 15 for testing. After single-epoch training, we evaluated the performance of the models on the training and testing sets with a final validation evaluation. See our GitHub repository [11] for implementation details.

Potentially, the data could be represented in other formats and methodologies beyond graph-based approaches could be considered as well. If we treat the adjacency matrix as an image, we could use image processing methods, such as transformer or CNN-based techniques. However, we should be careful when considering the kernel size and attention mechanisms due to the position-dependent nature of patterns within these matrices. Sequential data transformation could also facilitate NLP-based processing. Still, graph-based processing remains the most natural approach for this type of data.

A. Baseline : Vanilla GraphConv

Our baseline model consists of 3 GraphConv layers as defined by *torch_geometric* [12], and a final linear layer as can be seen in Figure 1. This GraphConv layer is an implementation of the graph neural network operator, as defined in [13]. The input to this model is composed of the node features and the node edges. Local and global neighborhoods are extracted, and the model iteratively computes a coloring based on these neighborhoods.

B. Brain Graph Neural Networks

We will use BrainGNN, a specialized GNN designed to process brain connectivity data by incorporating region-of-interest (ROI) awareness and specific pooling mechanisms [14]. Originally, classification on the graph is achieved by first embedding node features into a low-dimensional space, then coarsening or pooling nodes and summarizing them. The summarized vector is then fed into a multi-layer perceptron (MLP) as can be seen in Figure 2 (left).

1) Layers: The BrainGNN architecture is composed of blocks of ROI-aware Graph Convolutional Layer (Ra-GConv) layers, Node Pooling (R-pool) layers and a readout layer.

a) ROI-aware Graph Convolutional layer: Ra-GConv extends vanilla GraphConv by incorporating ROI-specific weight learning, allowing different embedding weights for different regions rather than using the same weights for all nodes. As each input is parcelled with the same atlas and the ROIs are ordered identically for every subject, we can use pseudo-position of one-hot encoding to maintain the geometric distribution of each ROI. In this way, the Ra-GConv layer embeds node features, edge features and pseudo-positions.

b) Node Pooling layer: As we need to reduce the dimensionality, we keep the most impactful ROIs and remove noisy nodes.

c) Readout layer: It summarizes the node feature vectors into a single vector, which is fed into the classifier. This is the layer that will be later replaced by the KAN layers.

2) Loss Functions: BrainGNN also provides the following loss function components : cross-entropy loss, unit loss, group-level consistency loss, TopK pooling loss.

C. Kolmogorov–Arnold Networks

1) A theoretical introduction: Kolmogorov-Arnold Networks (KANs) are a new approach to deep learning, described in a recently published paper [15]. KANs can be used either in their vanilla flavor, as an alternative to MLPs, or by integrating them in the above-mentioned GNNs. In contrast to MLPs, which have fixed activation functions on nodes and learnable weights on the edges, KANs have learnable activation functions on edges and no linear weights at all; every weight parameter is replaced by a univariate function parameterized as a spline. KANs are grounded in the Kolmogorov-Arnold Representation Theorem, which states that every multivariate continuous function on a bounded domain can be written as a finite composition of continuous functions of a single variable and addition operations. This theoretical foundation guarantees that KANs can approximate the complex functions needed for brain state classification while maintaining interpretability through their univariate components. Owing to the fact that KANs are just combinations of MLPs and splines, they are able to learn the compositional structure and the univariate functions. Furthermore, the splines can be made arbitrarily accurate by means of grid extension, without suffering the curse of dimensionality.

2) Implementation and integration into BrainGNN: In our implementation, we modify the original architecture by keeping the Ra-GConv layers and the R-pool layers. We, then, remove the readout layer and replace it with 2 KAN layers (Alg. 1).

The complete forward pass can be expressed as:

$h = \Phi_1(x_{GNN})\ y = \Phi_2(h)$

where x_{GNN} is the output from BrainGNN layers.

Algorithm 1 Proposed Model Architecture

```
model: Network(
  (n1): Sequential(Linear, ReLU(), Linear)
  (conv1): MyNNConv
  (pool1): TopKPooling
  (n2): Sequential(Linear, ReLU(), Linear)
  (conv2): MyNNConv
  (pool2): TopKPooling
  (fc1): Linear
  (bn1): BatchNorm1d
  (fc2): Linear
  (bn2): BatchNorm1d
  (lin_in_1): Linear
  (kan_in_1): NaiveFourierKANLayer()
  (lin_in_2): Linear
  (kan_in_2): NaiveFourierKANLayer()
  (lin_out): Linear
)
```

III. Dataset and Features

A high-quality and dimension data set was published recently [4]. It is a neuroimaging database consisting of 102

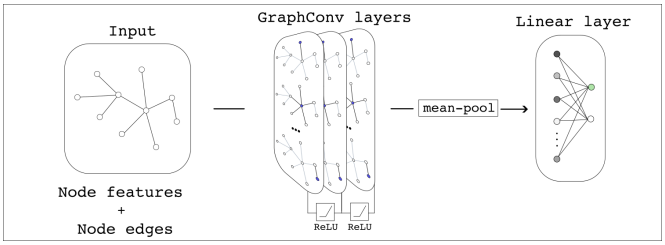

Fig. 1: Simple vanilla GraphConv model

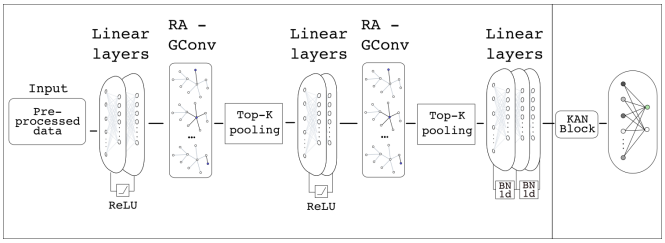

Fig. 2: BrainGNN model (left) with KAN addition (right) [14]

Model	Train acc.	Train Loss	Train F1 Score	Test acc.	Test Loss	Test F1 Score	Params
Baseline VanillaGNN	0.7900	0.2852	0.8827	0.8519	0.2852	0.9200	62786
Unscented BrainGNN Without KAN	**0.9280**	**0.0010**	**0.9560**	0.9330	**0.0100**	**0.9660**	141922
Fragrant BrainGNN With KAN	0.7938	0.0038	0.8837	**0.9333**	0.0216	0.9655	29796770

TABLE I: Performance comparison across model variants. The table shows training and testing metrics (accuracy and loss), F1 scores for balanced evaluation given class imbalance, and model complexity in terms of trainable parameters. VanillaGNN serves as the baseline, while other variants test the contribution of different architectural components.

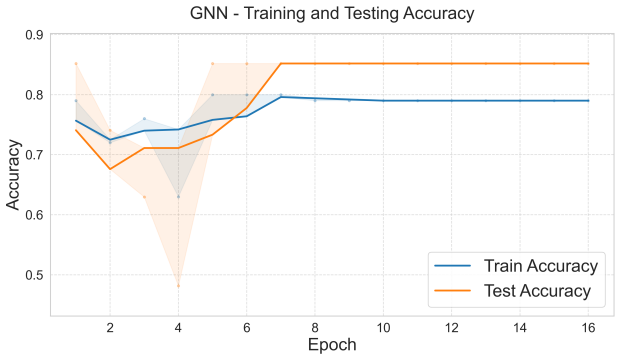

Fig. 3: Training and testing accuracy of the baseline vanilla model

synesthetic brains and 25 control participants, using state-of-the-art 3T MRI protocols from the Human Connectome Project (HCP). It consists of structural (T1- and T2-weighted) images together with approximately 24 minutes of resting state data per participant. In addition, a 'deep phenotype' is provided which includes detailed information about each participant's synesthesia - including the specific type(s) they have (10 types were considered in the study) associated with clinical and cognitive measures [4]. This dataset also provides the HCP-derived brain parcellations, and more specifically, subject-specific parcellations according to the HCP-MMP1, which has 180 labels per hemisphere, subject-specific node time series, and subject-specific connectomes.

In order to model the brain as a graph, the nodes are defined as brain regions of interest (ROIs) and the edges are defined as the functional connectivity between those ROIs, computed as the pairwise correlations of the fRMI time series. However, brain graphs are not translation invariant, and thus different embeddings must be used over different nodes. To create such a graph, first the brain is parcelled into N ROIs based on its T1 structural MRI. ROIs are graph nodes V = $v_1, ..., v_N$ which are preordered. As brain ROIs can be aligned by brain parcellation atlases based on their locations in the structure space, the brain graphs can be described as ordered aligned graphs. An undirected weighted graph is defined as $G = (V, E)$, where E is the edge set, i.e., a collection of (v_i, v_j) linking vertices from v_i to v_j. In this setting, G has an associated node feature set that can be represented as matrix $H = [h_1, ..., h_N]$, where h_i is the feature vector associated with node v_i. For every edge connecting two nodes, $(v_i, v_j)\epsilon E$, its strength is given by $e_{ij}\epsilon R$ and $e_{ij} > 0$. We also set $e_{ij} = 0$ for $(v_i, v_j) \notin E$ and therefore the adjacency matrix $E = [e_{ij}]\epsilon R^{NxN}$is well defined [14].

Given the HCP-derived brain parcellations, we follow the matlab code provided along with the dataset to extract full and partial correlations between the ROIs [4]. For each of the 360 nodes, the full correlations are used as node features, while the partial correlations are used to build the graph edges, as suggested in [14]. The edge weights are normalized per subject, and only the most positive (10%) and most negative (10%) partial correlations are kept, while the rest are discarded, in order to create a sparse matrix.

As a result, the final dataset used consists of 127 graphs. Each graph has 360 nodes. Each node is associated with a feature (a vector of 360 Pearson correlations), a pseudo-position (one-hot encoded ROI index) and edges to other nodes given by thresholded partial correlations.

IV. Results

Our experiments focused on binary classification of synesthetic versus non-synesthetic brains using resting-state fMRI

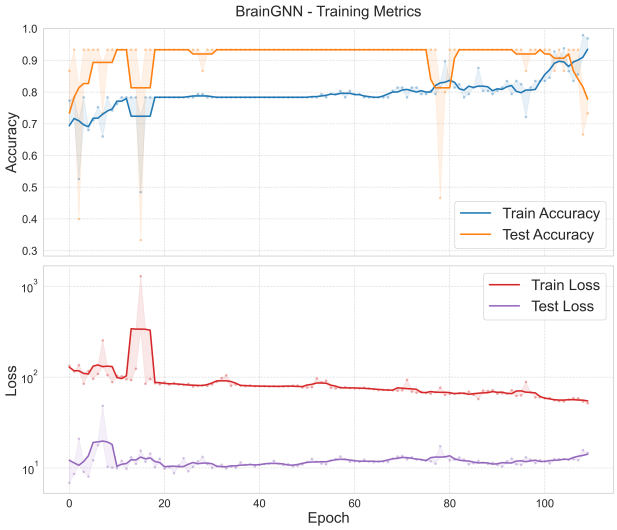

Fig. 4: Training and testing accuracy of the unscented BraingGNN model

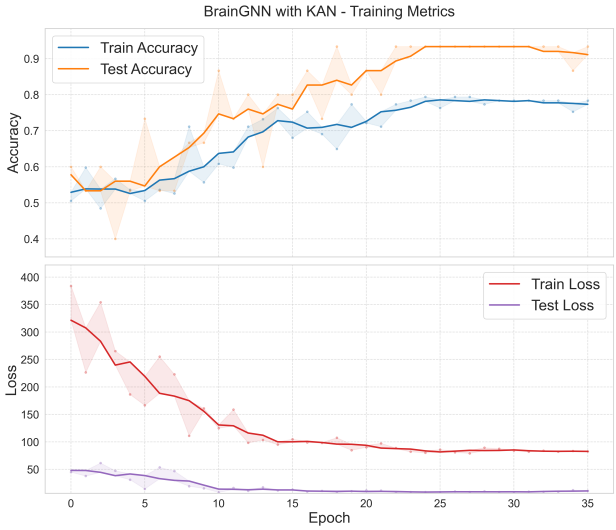

Fig. 5: Training and testing accuracy of the fragrant BrainGNN model with KAN layers

data processed into graph representations. The primary task was to accurately identify synesthetic individuals from controls based on functional connectivity patterns observed during rest, with no external stimuli presented. As mentioned above, the baseline model makes use of the node features and the node edges only. No pseudo-positions, nor edge weights are fed to the model. Given the simplicity of the model, it does learn fast, and after just a few epochs, there is no more progress. The performance both on the training and testing set remains equal to always predicting synesthesia (the ratio of 102 synesthetes and 25 controls gives a 0.816 success rate if the model always predicts synesthesia).

A. *Unscented (BrainGNN without KAN layers)*

BrainGNN, as originally published, can learn better, as it is provided with more information. Edge features and pseudo-positions seem to help the model overcome the imbalance in the data. However, the model is prone to overfitting.

B. *Fragrant (BrainGNN with KAN layers)*

By integrating the KAN layers in the model, we can actually see no further improvement. However, there was a consistent good generalization over the testing dataset, while the performance on the training dataset was average. This could be attributed to the constraints that come with the loss function, and severely penalize the KAN too.

V. Discussion and Future Directions

The reliable classification of synesthetic brains from resting-state fMRI data reveals important insights about the neurological basis of synesthesia and provides opportunities for advancing detection methodologies. It, notably, allows us to study how specific patterns of brain connectivity give rise to distinct subjective experiences. The high accuracy achieved confirms that distinctive patterns in functional brain connectivity result from synesthesia, persisting even without a triggering stimulus. This supports that synesthesia reflects fundamental differences in neural architecture rather than momentary cross-activation patterns. By identifying the neural correlates of these unusual perceptual experiences, we may gain critical insights into one of neuroscience's most profound questions: how neural activity transforms into conscious perception. Synesthesia provides a natural window into how brain connections create unique perceptual experiences, helping us understand the link between neural activity and how we experience the world.

To this end, in the future, we plan to curate the imbalance in the dataset by means of data collection and data augmentation. While our study used fMRI data, incorporating electroencephalogram (EEG) recordings could be beneficial for clinical applications. Despite a lower spatial resolution, EEG's accessibility and ease of use could make synesthesia diagnosis more widely available. Moreover, using time-series averages may not fully utilize our rich temporal data. We could improve performance by extracting overlapping temporal patches, adding controlled noise to simulate brain activity variations, and using variable-length sliding windows to capture different temporal dependencies.

The integration of KAN layers, though not improving raw performance metrics, demonstrated promising generalization capabilities - suggesting that it may be valuable for clinical applications. More sophisticated versions of KAN will also be considered. Moreover, as it has been observed that the brain network's underlying graphical model is non-Gaussian, one could consider incorporating Cauchy graphical model or Cauchy-GNN [16] in future works. Such approaches could be combined with adaptive methods like Graph Signal Adaptive Message Passing (GSAMP) [17] to handle the non-Gaussian temporal dynamics through spatiotemporal localized updates that process time-varying graph signals under both Gaussian and impulsive noise conditions.

VI. Conclusion

Our study demonstrated that synesthetic experience can be reliably detected from resting-state fMRI, achieving 93.3% accuracy in classification. BrainGNN was able to capture the subtle differences in functional connectivity characterizing synesthetic brains, while KAN integration showed promise for clinical applications. Overall, the high-accuracy achieved confirms that synesthesia leaves detectable neural "fingerprints" even during rest, indicating fundamental differences in neural organization. Our graph-based deep learning approach provides insights into how brain activity becomes conscious experience and allows us to measure objectively what once could only be described subjectively.

Data and Code Availability

All the code can be found on GitHub [11], with instructions on downloading the required data from the original sources.

Acknowledgments

We thank the support from the National Natural Science Foundation of China (32350410397); Shenzhen Medical Research Funds (D2301002); Science, Technology, Innovation Commission of Shenzhen Municipality (JCYJ20240813112016022, JCYJ20220530143014032, JCYJ20230807113017035, JCYJ20220530143002005, KCXFZ20211020163813019); Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan (JC2022009); Tsinghua University SIGS Start-up fund (QD2022024C); Shenzhen Ubiquitous Data Enabling Key Lab (ZDSYS20220527171406015); and Bureau of Planning, Land and Resources of Shenzhen Municipality (2022) 207.

References

[1] J. Ward et al. "Atypical sensory sensitivity as a shared feature between synaesthesia and autism". In: *Sci Rep* 7 (Mar. 2017), p. 41155. DOI: 10.1038/srep41155.

[2] K. Lunke and B. Meier. "A persistent memory advantage is specific to grapheme-colour synaesthesia". In: *Sci Rep* 10 (Feb. 2020), p. 3484. DOI: 10.1038/s41598-020-60388-6.

[3] M. Del Rio et al. "The mechanisms underlying conditioning of phantom percepts differ between those with hallucinations and synesthesia". In: *Sci Rep* 14 (Mar. 2024), p. 5607. DOI: 10.1038/s41598-024-53663-3.

[4] C. Racey et al. "An Open Science MRI Database of over 100 Synaesthetic Brains and Accompanying Deep Phenotypic Information". In: *Scientific Data* 10 (Nov. 2023), p. 766. DOI: 10.1038/s41597-023-02664-4.

[5] N. K. Logothetis. "What we can do and what we cannot do with fMRI". In: *Nature* 453.7197 (June 2008), pp. 869–878. ISSN: 1476-4687. DOI: 10.1038/nature06976.

[6] J. M. Hupé and M. Dojat. "A critical review of the neuroimaging literature on synesthesia". In: *Frontiers in Human Neuroscience* 9 (Mar. 2015). DOI: 10.3389/fnhum.2015.00103.

[7] R. Rouw, H. S. Scholte, and O. Colizoli. "Brain areas involved in synaesthesia: A review". In: *Journal of Neuropsychology* 5 (Sept. 2011), pp. 214–242. DOI: 10.1111/j.1748-6653.2011.02006.x.

[8] J. A. Nunn et al. "Functional magnetic resonance imaging of synesthesia: activation of V4/V8 by spoken words". In: *Nature Neuroscience* 5 (2002).

[9] J. Neufeld et al. "The neural correlates of coloured music: A functional MRI investigation of auditory–visual synaesthesia". In: *Neuropsychologia* 50 (Jan. 2012), pp. 85–89. DOI: 10.1016/j.neuropsychologia.2011.11.001.

[10] J. Ward et al. "Synesthesia is linked to large and extensive differences in brain structure and function as determined by whole-brain biomarkers derived from the HCP (Human Connectome Project) cortical parcellation approach". In: *Cerebral Cortex* 34.11 (2024), bhae446.

[11] Alex Feldman and Aikaterini Melliou. *Synesthesia Oracle*. Version 1.0.0. 2025. URL: https://github.com/alexTrod/Synesthesia-Oracle.

[12] M. Fey and J. E. Lenssen. "Fast Graph Representation Learning with PyTorch Geometric". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.

[13] C. Morris et al. "Weisfeiler and leman go neural: Higher-order graph neural networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 4602–4609.

[14] X. Li et al. "BrainGNN: Interpretable Brain Graph Neural Network for fMRI Analysis". In: *Medical Image Analysis* 74 (Dec. 2021), p. 102233. DOI: 10.1016/j.media.2021.102233.

[15] Z. Liu et al. *KAN: Kolmogorov-Arnold Networks*. Accepted at ICLR 2025. 2025. URL: https://openreview.net/forum?id=Ozo7qJ5vZi.

[16] T. Muvunza, Y. Li, and E.E. Kuruoglu. "Cauchy Graphical Models". In: *Proceedings of The 12th International Conference on Probabilistic Graphical Models*. Ed. by Johan Kwisthout and Silja Renooij. Vol. 246. Proceedings of Machine Learning Research. PMLR, Nov. 2024, pp. 528–542. URL: https://proceedings.mlr.press/v246/muvunza24a.html.

[17] Yi Yan, Changran Peng, and Ercan E Kuruoglu. "Graph signal adaptive message passing". In: *IEEE Signal Processing Letters* (2025).