Exploring Structural Brain Connectivity in Term and Preterm Infants with Explainable AI and Fuzzy Logic

Katherine Birch[1,*], Alberto Durán-López[2], Daniel Bolaños-Martinez[2], Chandresh Pravin[1], Maria Bermudez-Edo[2], Roman Bauer[1] and Suparna De[1,*]

[1]*NICE Research Group, Computer Science Research Centre, School of Computer Science and Electronic Engineering, University of Surrey, GU2 7XH, Guildford, England, United Kingdom*

[2]*Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, 18014, Granada, Spain*

Abstract

Preterm births have been associated with altered neurological development for neonatal infants; this has been implicated in certain neuro-developmental conditions in later life. Advances in brain imaging methods, such as Magnetic Resonance Imaging, have allowed for the analysis of physical connectivity of brain matter in infants shortly after birth. However, commonly used methods of investigating such data rely on a brain network analysis, traditionally based on graph-theoretical approaches, which may fail to capture complex patterns involving both local and global network structures and spatial information. Furthermore, many previous studies of infant brain data rely on a priori selection of specific graph connectivity measures. We propose employing machine learning models such as logistic regression and Graph Neural Networks (GNN) to provide a data-driven approach for classifying preterm and term brain networks at birth. We utilize fuzzy logic, and explainability methods including Shapley Additive Explanations (SHAP) to identify influential regions and connections in decision making. In our analysis, brain regions are represented as spatially embedded nodes, with edges representing strength of structural connections between areas. Using this setup, our model achieves a binary classification accuracy of 88.57%. This performance is further enhanced using a fuzzy boundary between preterm and term classes, achieving an accuracy of 96.19%. This demonstrates that the model can be assisted particularly by adding context to "near-term" born infant cases. These analyses highlight important connections and key nodes, including deep brain structures which are broadly consistent with biological literature.

Keywords

Brain development, Machine learning, Preterm birth, Explainable AI (XAI), Fuzzy logic

1. Introduction

Preterm births disrupt the critical final trimester of infants and carry the potential to affect infant neuro-development, potentially leading to further complications later in life [1, 2, 3, 4]. Recent advances in infant brain imaging methods applied shortly after birth [5], have allowed researchers to map brain matter connectivity into structured network representations suitable for application of computational methods. Traditionally, brain connectivity of infants has been studied using graph theoretical measures, such as the rich-club coefficient and clustering coefficient to expose the structural differences between term and preterm infant brains [5, 6, 7]. Various studies using traditional graph theoretical approaches have suggested that preterm infants exhibit reduced connectivity between hub regions in the brain when compared with term infants, specifically during early stages of birth when these measurements are taken [6, 7, 5, 8].

MAI-XAI@ECAI'25: Multimodal, Affective and Interactive eXplainable AI

*Corresponding author.

https://github.com/Katherine-Birch/Explainable-AI-and-fuzzy-logic-for-preterm-term-structural-brain-connectivity

k.birch@surrey.ac.uk (K. Birch); albduranlopez@ugr.es (A. Durán-López); danibolanos@ugr.es (D. Bolaños-Martinez); c.pravin@surrey.ac.uk (C. Pravin); mbe@ugr.es (M. Bermudez-Edo); r.bauer@surrey.ac.uk (R. Bauer); s.de@surrey.ac.uk (S. De)

http://www.ugr.es/local/mbe (M. Bermudez-Edo); https://www.surrey.ac.uk/people/roman-bauer (R. Bauer); https://www.surrey.ac.uk/people/suparna-de (S. De)

0000-0002-1763-7396 (K. Birch); 0009-0005-8995-869X (A. Durán-López); 0000-0003-0207-2908 (D. Bolaños-Martinez); 0000-0003-1530-0121 (C. Pravin); 0000-0002-2028-4755 (M. Bermudez-Edo); 0000-0002-7268-9359 (R. Bauer); 0000-0001-7439-6077 (S. De)

Many of these previous approaches, while providing valuable insights into certain preterm brain connectivity patterns, have inherent limitations, namely, they require *a priori* selection of specific graph connectivity measures [9]. The pre-selection of particular hypothesis-driven measures [10, 9] raises the potential of overlooking complex or unexpected connectivity structures that may be captured using more data-driven computational methods. Furthermore, these graph theoretic measures are sensitive to the methodological choices, such as thresholding, network normalization, and hub definition [10]. Combined, these methodological variations contribute to a lack of reproducibility across studies and datasets [10].

Beyond the classification of preterm and term infants, in this paper we are primarily interested in identifying the differences in neurological structures that distinguish the two populations. To achieve this, we apply machine learning (ML) explainability techniques that aid in visualizing and understanding the models' decision boundaries [11, 12]. Rather than relying solely on machine learning to extract abstract patterns, we adopt a data-centric approach that iteratively maps model outputs back onto the dataset, and validates neuro-developmental variance between preterm and term infants, as described in the literature [13, 3, 6, 2, 14, 15, 16, 17, 4, 18]. Through our approach that focuses on explainability of the classification task, we hope to enhance the understanding of the neuro-biological differences of term and preterm infants. The findings from this paper have the potential to support future clinical interventions and developmental support strategies. Our contributions are as follows:

- We present an empirical comparison of different ML and GNN architectures for classifying preterm versus term infants from structural brain data. This provides insights into the abilities of various AI models applied to complex, real-world biological data.
- We investigate the impact of incorporating a biologically informed feature engineering strategy. We use atlas-based spatial coordinates (centroids) as additional model inputs.
- We introduce a novel adaptation of fuzzy logic for label smoothing, specifically designed to address continuous development in medical classification. This approach combats the noisy class boundary and considers domain-specific priors, resulting in improved model accuracy.
- We provide insights into model decision-making through the use of SHAP explainability. We do this by aggregating node-level attributions and edge-importance matrices, and projecting them onto the brain atlas to visually interpret the models' decisions. We critically discuss the consistency across different architechtures, and discuss consistency with known neuroscientific literature.
- All code is available on GitHub [1].

2. Related work

Due to the way structural brain data is processed, it often takes the form of a connectivity matrix, with nodes representing brain regions and edges representing connections between them. Graph theoretical measures are commonly used to analyze such brain connectivity data. For example, studies investigate connectivity patterns and population differences by comparing values of graph metrics derived from the structural connectome. In the context of preterm birth, researchers have used measures such as the rich club coefficient, betweenness centrality, and small worldness. These metrics aim to identify important brain regions, developmental patterns, and network efficiency. However, they are often difficult to interpret. When combined, it becomes unclear how much each measure contributes to the observed similarity or difference between networks. For instance, in studies comparing preterm and term infants, findings on the prominence of the rich club are inconsistent. Some report greater [7], while others report lesser [19] rich club organization in preterm infants. These discrepancies may result from differences in normalization, metric definitions, or how multiple measures are integrated.

[1] https://github.com/Katherine-Birch/Explainable-AI-and-fuzzy-logic-for-preterm-term-structural-brain-connectivity

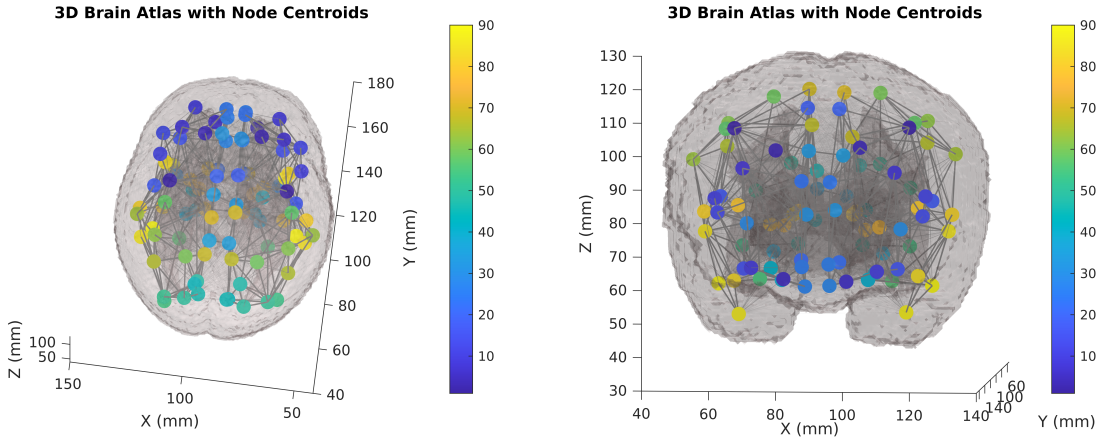

Figure 1: Brain connectivity visualizations from different perspectives. Nodes represent the centroid coordinate of each brain region, and edges denote the connections between them. Here, edges are shown based on their strength (darkness and thickness). Node color scale is showing regions from front-brain (blue) to top-back (green) and lower-back (yellow). These provide coordinate location of nodes relative to each other.

In addition, many of these metrics do not account for the spatial location of brain regions and often require extra graph-based or post hoc analyses to support hypothesis-driven interpretations.

Outside the context of preterm birth, some studies have applied ML models to analyze infant brain data [20, 21]. Support Vector Machine (SVM) was applied to functional brain connectivity data from infants in order to predict between preterm and term [22]. This study had a low sample size, with only 50 infants, and considered functional rather than structural connectivity. One recent study [5] did consider both preterm and term brains; however, it focused on predicting gestational age rather than directly classifying preterm versus term brains. Additionally, the study suffered from a significant lack of data, with only seven infants in each group. The regression model also showed poor performance, with a very low R^2 score, and the authors did not report standard ML metrics, making comparisons difficult.

Recently, researchers have proposed that Graph Neural Network (GNN)-based models may be well suited for tasks involving structural brain data. This approach has been applied in some contexts with varying degrees of success. However, differences between preterm and term brains have not been explored using GNNs. Cui et al. [9] proposed using GNNs and highlighted several ways brain connectivity data can be transformed into graph format. However, the reported results showed low accuracy. Messaritaki et al. [23] also explored this idea, demonstrating various methods for defining structural brain data as a graph. They noted that, despite the apparent suitability for GNNs, brain network nodes often lack meaningful values. They emphasized the importance of considering the strength of connections. Importantly, their work did not address differences between preterm and term infant brains.

Medical literature shows that identifying and predicting differences in infants born before their estimated due date can be challenging. This is due to several factors, including the difficulty of accurately estimating due dates and the common assumption that all pregnancies should ideally last the same amount of time. In reality, gestational age (GA) at birth varies widely, with many births occurring within a window around the estimated due date [16]. Researchers have noted that assessing the potential risks for preterm infants born near the full-term threshold is particularly difficult [18, 14, 17]. In classifying stages of prematurity, fuzzy logic has been proposed in both the biological domain [24, 25, 26] and neuroscience [27, 28]. However, this approach has not yet been applied in the context of infant brain network connectivity.

3. Method

We frame the task of distinguishing preterm from term infant brains as a binary classification problem using structural connectomes derived from Diffusion Tensor Imaging (DTI). Each subject's brain is represented by a connectivity matrix $A_i \in \mathbb{R}^{n \times n}$, where the rows and columns correspond to defined brain regions, and the entries reflect connection strengths between those regions [5]. Since connectomes may differ between preterm and term births, we explore two complementary modeling perspectives: one that operates directly on the matrix representation of A_i, and another that interprets it as a graph structure.

3.1. Matrix-Based Approach

Each structural connectome is encoded as a symmetric adjacency matrix $A_i \in \mathbb{R}^{n \times n}$, where where n is the number of nodes, representing distinct regions in the brain. The entry $A_i[u, v]$ gives the strength of the connection between region u and region v. Since A_i is symmetric and its diagonal entries carry no information (they represent self-connections), we extract a feature vector by listing all entries above the main diagonal, where $u < v$:

$$\mathbf{x}_i = \big(A_i[1,2],\ A_i[1,3],\ \ldots,\ A_i[n-1,n]\big). \tag{1}$$

This vector contains exactly one entry for each unordered pair of regions, capturing all unique connection strengths in the brain network. The resulting feature vector has length $d = \frac{n(n-1)}{2}$.

3.2. Graph-Based Approach

Alternatively, we interpret each structural connectome as a weighted, undirected graph $G_i = \big(V,\ E_i,\ X_v^{(i)},\ X_e^{(i)}\big)$, constructed from its connectivity matrix $A_i \in \mathbb{R}^{n \times n}$:

- V is the fixed set of n nodes, each corresponding to a brain region.
- $E_i \subseteq V \times V$ is the subject-specific edge set, defined by nonzero connections in A_i.
- $X_v^{(i)} \in \mathbb{R}^{n \times f}$ is the feature matrix for the i-th node.
- $X_e^{(i)} \in \mathbb{R}^{|E_i| \times 1}$ is the edge feature matrix. Each edge $(v_j, v_k) \in E_i$ carries a connectivity strength equal to $A_i[j, k]$.

This formulation transforms each structural connectome i into a graph G_i, enabling the use of graph-based learning models to perform binary classification.

3.3. Feature Augmentation and Preprocessing strategies

To improve model performance, we explore both feature augmentation and preprocessing strategies. Table 1 summarizes the techniques applicable to both matrix-based and graph-based formulations. Figure 1 shows brain connectivity visualizations using spatial coordinates from Table 1.

3.3.1. Spatial coordinates

In order to obtain spatial coordinates we calculate node centroids based on each brain region in the infant brain atlas [29, 30] which was used in the initial processing of DTI scans. We first identify unique non-zero integer labels within the atlas NIfTI [2] image. For each brain region (1-90) we identify all voxels belonging to that region, and calculate the centroid of each giving us a mean X,Y,Z node coordinate per region. Applying the affine transformation matrix from the NIfTI header, we are able to obtain the real world spatial coordinates (in mm) from the voxel space centroids. These coordinates are approximate central locations for the corresponding nodes in the already processed connectivity matrices. This

[2]File type: Neuroimaging Informatics Technology Initiative

allows us to take information such as the organisation of nodes relative to one another. The resulting centroids are shown in Figure 1.

Table 1
Feature augmentations and preprocessing strategies.

Feature	Description
Spatial Coordinates	Adds (x, y, z) atlas-based centroids as node features, providing spatial information.
Node Degree	Adds the number of edges per node: $d_v = \sum_u \mathbb{I}[A[v,u] > 0]$, encoding local connectivity.
Thresholding	Removes weak edges below a threshold ϵ: $A[v,u] = 0$ if $A[v,u] < \epsilon$.
Undersampling	Balances class distribution by randomly removing samples from the majority class. See Figure 6 in Appendix A for a histogram demonstrating distribution of GA in the dataset.

3.4. Fuzzy logic

In the standard formulation, the gestational age (GA) label is defined as y_i = preterm if $\mathrm{GA}_i < \tau$ and y_i = term if $\mathrm{GA_i} \geq \tau$, with the cutoff set at $\tau = 37$ weeks [31]. However, GA is recorded in whole weeks, and biological maturation is continuous, so infants at 36^{+6days} and 37^{+0days} weeks often exhibit nearly identical connectomes [16]. Recent studies have shown that functional brain development evolves gradually around this gestational threshold, without a sharp boundary [15]. As a result, models trained on these hard labels tend to struggle for subjects with $\mathrm{GA_i}$ near τ, where small dating errors introduce label noise and the decision boundary becomes arbitrary. To address this, we define a soft target:

$$y_i^{\text{soft}} = \sigma\left(\frac{\mathrm{GA}_i - \tau}{T}\right) = \frac{1}{1 + \exp\left(-\frac{\mathrm{GA}_i - \tau}{T}\right)}, \tag{2}$$

where T is a temperature parameter that controls the smoothness of the transition, as shown in Figure 2.

Training with these soft targets smooths the decision boundary around 37 weeks, which may reduce the impact of GA-recording errors. This approach encourages the model to learn graded changes in connectivity rather than an abrupt jump, and it aims to improve robustness in the late preterm window (35–37 weeks), where clinical and connectomic differences lie on a continuum. A comparison of the fuzzy and traditional boundaries is shown in Figure 3.

3.5. Explainability

While high classification performance demonstrates that structural connectomes carry discriminative information, understanding why a model makes its decisions is important for trust and biological insight. In brain networks, explainability reveals which connections or regions drive the prediction of preterm versus term status, guiding neuroscientific hypotheses and potential biomarkers.

We employ SHapley Additive exPlanations (SHAP) [11] to decompose the prediction $f(\mathbf{x})$ as:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j, \tag{3}$$

where ϕ_0 is the baseline output and ϕ_j is the SHAP value of feature j.

After computing the SHAP matrix $\Phi \in \mathbb{R}^{N \times d}$ for N subjects and d features (edges and global covariates), we identify the most important edges and nodes as follows:

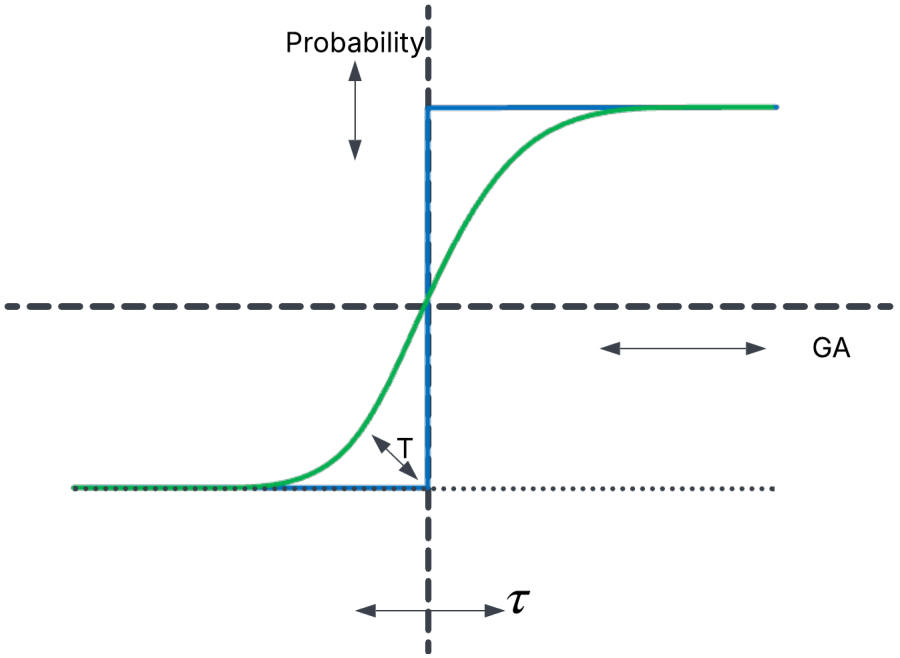

Figure 2: Sigmoidal mapping from gestational age to soft target y_i^{soft}. This illustrates the variables transition point (τ) and steepness (T) and how they interact with GA. From this interaction we determine probability, and this serves as the fuzzy label.

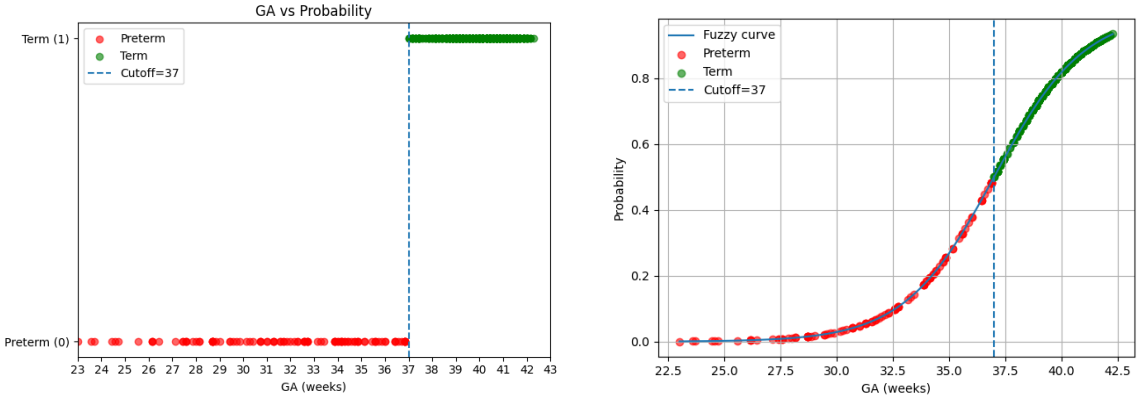

Figure 3: The left-hand image shows traditional binary labels, where all subjects are classified as 0 (preterm) or 1 (term) based on the 37-week threshold. The right-hand image shows the same subjects after applying fuzzy logic, with a smoothed transition around the threshold.

1. **Edge importance matrix.** We compute the mean absolute SHAP value for each edge feature:

$$M_{uv} = \frac{1}{N} \sum_{i=1}^{N} \left| \phi_{j(i,u,v)}^{(i)} \right|, \tag{4}$$

where $j(i, u, v)$ maps the pair of regions (u, v) to the corresponding index in $\mathbf{x}$. We then symmetrize the matrix by setting $M_{uv} = M_{vu}$.

2. **Node-level aggregation.** For each node v, we define:

$$I_v = \frac{1}{n-1} \sum_{u \neq v} M_{uv}, \tag{5}$$

which captures the average importance of all edges incident on v.

For SHAP importance threshold δ, edges with $M_{uv} > \delta$ and nodes with top I_v values identify the most influential connections and regions driving the model's decisions. This explainability pipeline highlights the features that most strongly inform the classifier [12].

4. Experiments

4.1. Dataset

We use data from the Developing Human Connectome Project 2nd release [32] which was processed in [5], and made available on their GitHub [3]. It comprises structural brain data from 524 infants, shortly after birth, acquired via DTI. Connectivity is represented as symmetric adjacency matrices between 90 cortical and subcortical regions. For each subject, we also have the following metadata: gestational age at birth (GA), postmenstrual age at scan (PMA), sex, session ID, and subject ID. We derive the preterm/term label using GA≤37 weeks as preterm and GA>37 weeks as term. Table 2 summarizes all variables, including their array shape and type.

Table 2
Description of dataset variables.

Variable	Shape	Description
SCmu	(90, 90, 524)	Connectivity matrices for 524 subjects
ga	(524, 1)	Gestational age at birth (weeks)
mu	(1, 524)	Mean edge weight
pma	(524, 1)	Postmenstrual age at scan (weeks)
ses	(1, 524)	Session identifiers (string)
sex	(524, 1)	Sex (0 = female, 1 = male)
sub	(1, 524)	Subject identifiers (string)

4.2. Design

We apply the proposed methodology to the dataset described above. We use four baseline ML models: Logistic Regression (LR), SVM, Multi-Layer Perceptron (MLP), and Random Forest (RF) [33]. We also evaluate three GNN architectures: Graph Convolutional Network (GCN), Graph Attention Network (GAT) and Graph Isomorphism Network (GIN) [34, 35, 36]. Each execution uses 5-fold cross-validation (CV), with hyperparameters selected as shown in Table 1. We use stratified sampling and use a 20% / 80% training/test split. Additionally, we integrate fuzzy logic into the best-performing models from both the ML and GNN approaches. We use accuracy, precision, recall and F1-score as the evaluation metrics for all experiments. In addition, we calculate weighted averages, in order to account for unbalanced classes [37]:

$$\text{Metric}_W = \frac{(\text{Metric}_0 \times \text{Class size}_0) + (\text{Metric}_1 \times \text{Class size}_1)}{\text{Class size}_0 + \text{Class size}_1}. \tag{6}$$

[3] https://github.com/CoDe-Neuro/Predicting-age-and-clinical-risk-from-the-neonatal-connectome

Table 3
Impact of Spatial Information on ML Model Performance. We show LR, SVM, MLP, RF. We also show GNN variants: GCN, GAT, and GIN. Each model is assessed with and without the inclusion of spatial node coordinates. All values are shown from 0 to 1. Best results are in bold. Weighted averages consider unbalanced classes [21 preterm; 84 term].

Model	Class	Precision	Recall	F1-score	Overall (weighted average)			
					Accuracy	Precision	Recall	F1-score
LR	Preterm	0.70	0.76	0.73	0.89	0.89	0.89	0.89
	Term	0.94	0.92	0.93				
LR + Spatial	Preterm	0.85	0.81	**0.83**	**0.93**	**0.93**	**0.93**	**0.93**
	Term	0.95	0.96	**0.96**				
SVM	Preterm	0.61	0.90	0.73	0.87	0.90	0.87	0.88
	Term	**0.97**	0.86	0.91				
SVM + Spatial	Preterm	0.20	**1.00**	0.33	0.20	0.04	0.20	0.07
	Term	0.00	0.00	0.00				
MLP	Preterm	0.60	0.14	0.23	0.81	0.78	0.81	0.76
	Term	0.82	0.98	0.89				
MLP + Spatial	Preterm	0.00	0.00	0.00	0.80	0.64	0.80	0.71
	Term	0.80	**1.00**	0.89				
RF	Preterm	**1.00**	0.19	0.32	0.84	0.87	0.84	0.79
	Term	0.83	**1.00**	0.91				
RF + Spatial	Preterm	**1.00**	0.10	0.17	0.82	0.85	0.82	0.75
	Term	0.82	**1.00**	0.90				
GCN	Preterm	**0.91**	0.48	0.63	0.89	0.89	0.89	0.87
	Term	0.88	0.99	0.93				
GCN + Spatial	Preterm	0.00	0.00	0.00	0.80	0.64	0.80	0.71
	Term	0.80	**1.00**	0.89				
GAT	Preterm	0.79	**0.71**	**0.75**	**0.90**	**0.90**	**0.90**	**0.90**
	Term	**0.93**	0.95	**0.94**				
GAT + Spatial	Preterm	0.85	0.52	0.65	0.89	0.88	0.89	0.87
	Term	0.89	**0.98**	0.93				
GIN	Preterm	0.00	0.00	0.00	0.80	0.64	0.80	0.71
	Term	0.80	**1.00**	0.89				
GIN + Spatial	Preterm	0.00	0.00	0.00	0.80	0.64	0.80	0.71
	Term	0.80	**1.00**	0.89				

Table 4
Performance metrics with fuzzy logic. We select the two best performing models from Table above: LR with Spatial and GAT. Values are shown from 0 to 1.

Model	Class	Precision	Recall	F1-score	Overall (weighted average)			
					Accuracy	Precision	Recall	F1-score
LR + Spatial + fuzzy	Preterm	**0.95**	**0.86**	**0.90**	**0.96**	**0.96**	**0.96**	**0.96**
	Term	**0.97**	**0.99**	**0.98**				
GAT + fuzzy	Preterm	0.92	0.52	0.67	0.90	0.90	0.90	0.88
	Term	0.89	**0.99**	0.94				

4.3. Results

In Table 3 we present the ML and GNN results alongside those obtained using the spatial-coordinate (described in Table 1). We focus on those atlas coordinates from Table 1 because this strategy was the only one to improve model performance. Based on these results, we select the best-performing algorithms from each approach (LR with spatial coordinates and GAT) and apply fuzzy-logic labeling, as shown in Table 4.

4.4. Explainability

We use SHAP Equations 4 and 5 to compute attribution scores, deriving edge-importance matrices and aggregating node-level importances to interpret each model prediction. Figure 4 contrasts feature importance in the LR and GAT models through both heatmaps and Atlas-based network plots. The top row displays SHAP edge-importance heatmaps, highlighting which connections most drive each model's predictions. The bottom row renders nodes at their Atlas coordinates, sized and colored by SHAP score to reveal the most (red) and least (green) influential regions (node indices correspond to Table 5). Figure 5 presents the SHAP summary plot indicating influence of individual connections on the LR model outcome. Negative SHAP values indicate the model is pushed towards class 0 (preterm) while positive values classify towards class 1 (term). Red indicates high feature values while blue indicate low. Finally, Table 5 presents the top 10 and bottom 10 nodes ranked by SHAP importance for both the LR and GAT models. Regions shown in bold indicate agreement between the two models. LR and GAT concur on 8 of the 10 least important nodes, whereas they align on 4 of the 10 most important regions.

Table 5
Top 10 and bottom 10 nodes by SHAP importance for both LR and GAT models, with bold indicating regions of agreement (8/10 least important; 4/10 most important. SHAP values calculated through the combined metric. These regions are displayed in Atlas coordinate space in Figure 4 panels (c) and (d).

LR			GAT		
Top 10					
Node Index	**Node Name**	**SHAP Value**	**Node Index**	**Node Name**	**SHAP Value**
78	**Thalamus R**	**0.22**	90	Temporal Inf R	0.39
31	Cingulum Ant L	0.21	61	Parietal Inf L	0.38
32	Cingulum Ant R	0.20	**7**	**Frontal Mid L**	**0.37**
34	Cingulum Mid R	0.18	**30**	**Insula R**	**0.35**
36	Cingulum Post R	0.18	82	Temporal Sup R	0.33
30	**Insula R**	**0.18**	**78**	**Thalamus R**	**0.33**
68	Precuneus R	0.18	29	Insula L	0.33
8	Frontal Mid R	0.16	62	Parietal Inf R	0.33
74	**Putamen R**	**0.16**	**74**	**Putamen R**	**0.33**
7	**Frontal Mid L**	**0.16**	71	Caudate L	0.31
Bottom 10					
54	**Occipital Inf R**	**0.06**	50	Occipital Sup R	0.09
21	Olfactory L	0.06	87	Hippocampus L	0.08
11	**Frontal Inf Oper L**	**0.06**	**75**	**Palladium L**	**0.08**
70	**Paracentral Lob R**	**0.06**	**37**	**Frontal Inf Oper L**	**0.08**
75	**Palladium L**	**0.05**	**49**	**Occipital Sup L**	**0.07**
69	Paracentral Lob L	0.05	**54**	**Occipital Inf R**	**0.07**
80	**Heschl R**	**0.05**	45	Cuneus L	0.07
49	**Occipital Sup L**	**0.04**	**70**	**Paracentral Lob R**	**0.06**
37	**Hippocampus L**	**0.04**	**80**	**Heschl R**	**0.03**
79	**Heschl L**	**0.01**	**79**	**Heschl L**	**0.03**

5. Discussion

From Table 3 and Table 4 we see that LR is the best performing algorithm overall. Most models do well on average, but SVM, MLP and RF have trouble with the preterm class. Among the GNNs, GCN achieves moderate performance across both classes. GAT outperforms GCN, likely because its attention mechanism highlights the most informative graph connections. Adding spatial coordinates improves LR performance, as the atlas location data provides useful information, but it does not benefit the GNN models because it introduces additional complexity causing overfitting in some variants. For example, we note that GIN overfits, suggesting that more complex GNN models do not necessarily outperform simpler models like LR in this context. However, future work could broaden the scope through further experimentation with alternative architectures. Moreover, incorporating fuzzy logic into

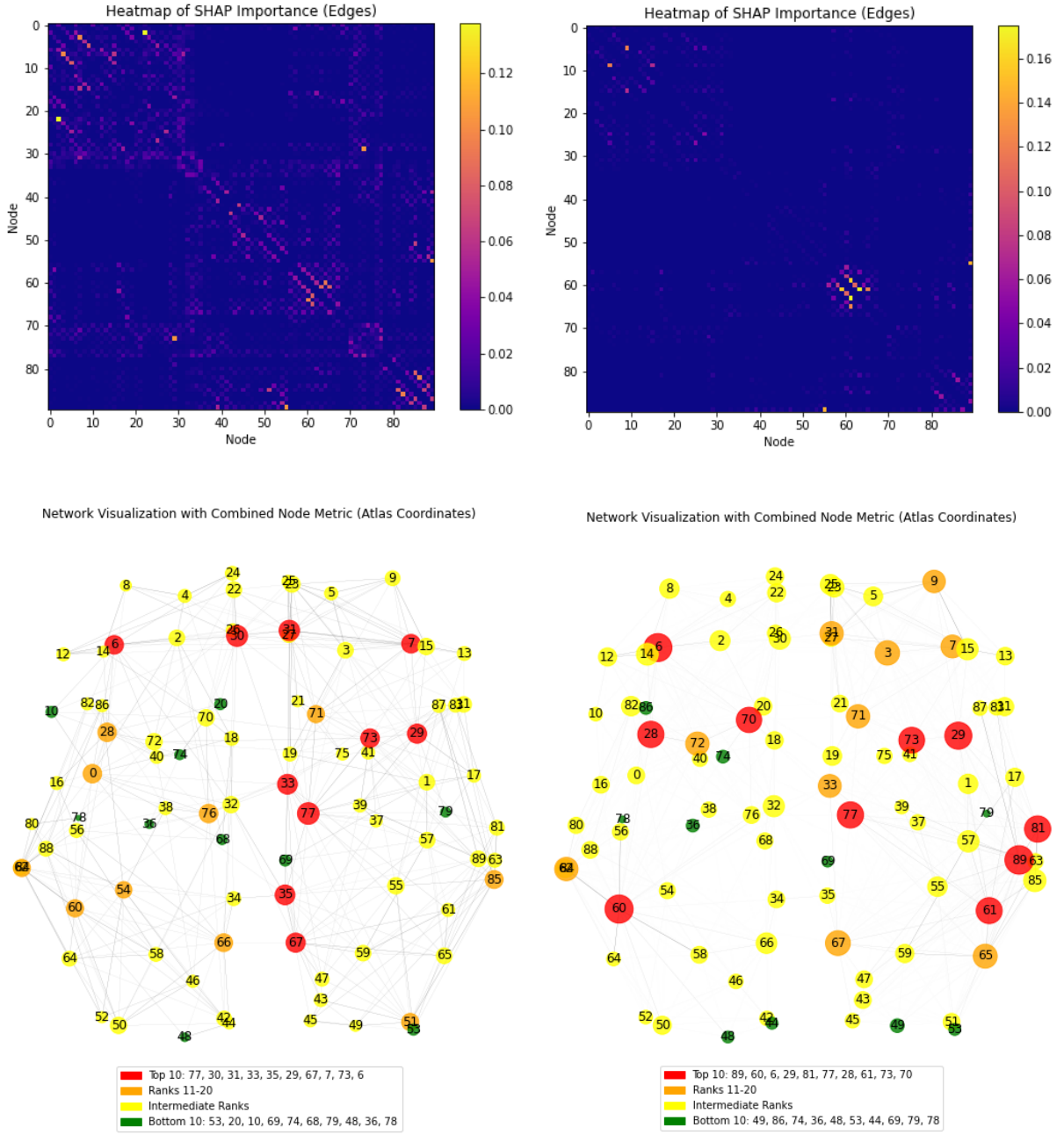

Figure 4: The top panels show SHAP edge-importance heatmaps for LR (left) and GAT (right). The bottom panels (LR - left; GAT - right) depict nodes in Atlas coordinates space colored red (top 10), green (bottom 10), and yellow (others). The regions in lower panels (left) and (right) correspond to the region names in Table 5 (NB// Node numbers here from 0-89).

LR further enhances its performance by allowing smoother transitions around the decision threshold and better handling uncertainty in feature values. One prior study applied SVM to functional brain data and reported an accuracy of 84% [22]. In our experiments, SVM achieved a comparable accuracy of approximately 86% but was nonetheless outperformed by other methods. Although we focus on structural rather than functional development, we observed that SVM exhibited a pronounced bias toward the majority class, which is an important limitation given the relative scarcity of data in this

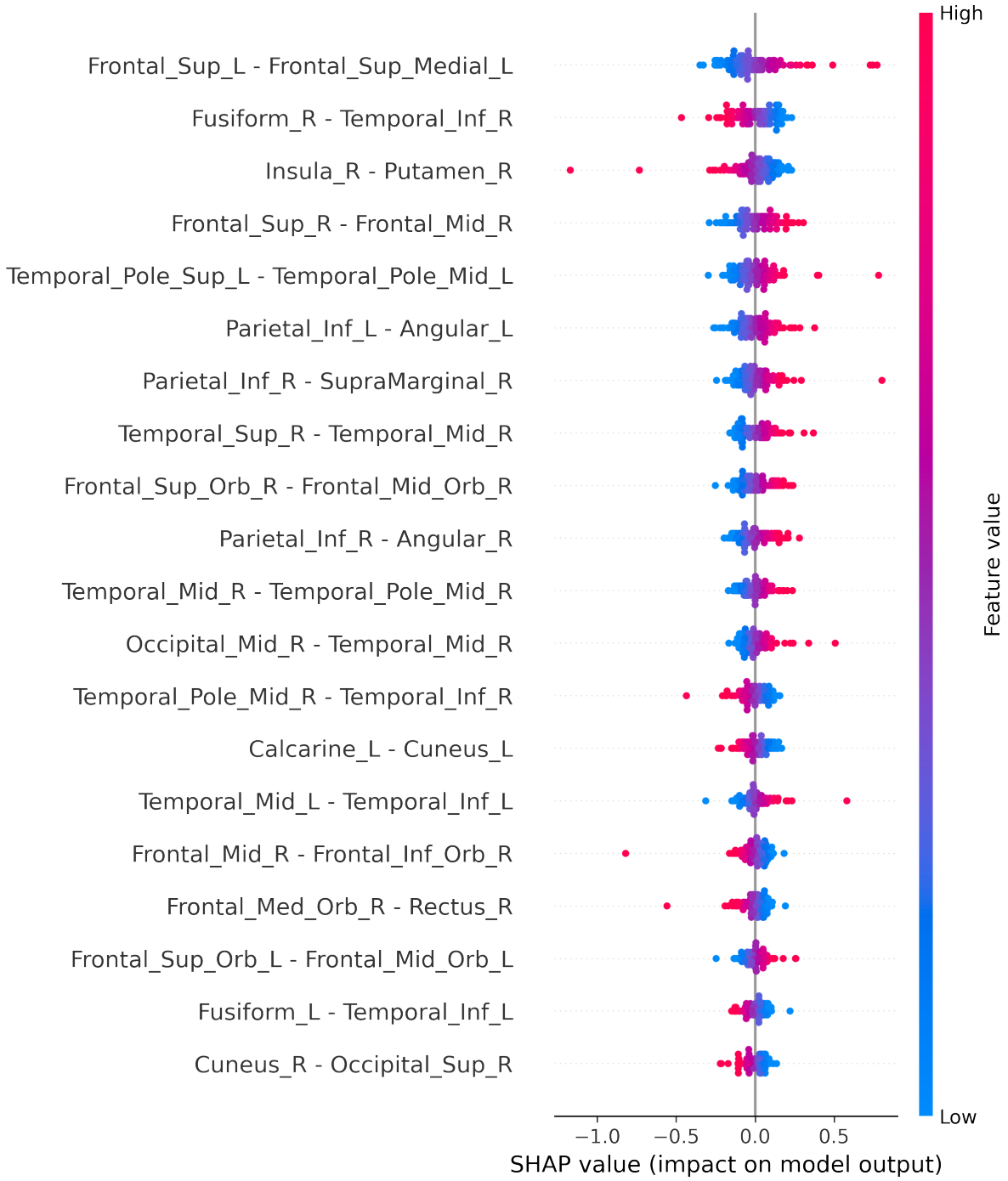

Figure 5: SHAP summary plot visualizing each edge connection's impact on the best model output (LR). Negative SHAP values indicate the model is pushed towards class 0 (preterm) while positive values classify towards class 1 (term). Red indicates high feature values while blue indicate low.

domain.

In order to further understand these results, we consider the main regions and connections that contribute to the classification. We note that while SHAP is not enough to confirm any causal relationships, there are a number of interesting findings which are supported by biological literature. From SHAP analysis of LR and GAT models, the regions indicated are predominantly from the deep brain structures like putamen and thalamus, confirming previous findings [13, 38]. Additionally, it demonstrates that the models consistently attend to specific regions, suggesting notable differences between preterm and term brain connectivity. Many of the connections which were highlighted as important, were

between nodes which were also identified as important. Table 5 demonstrates that there are regions consistently attended to regardless of LR or GAT model, but also indicates that the regions that are most consistent across models are the ones least influential which provides valuable insights into the effect of preterm birth on different brain regions. Moreover, regions least important in distinguishing between the classes include regions which are generally understood to be well developed early before birth, for example Heschl's gyrus, occipital lobe areas, and the temporal pole, corroborating Gilmore et al. [13]. These are associated with vision, hearing and other senses [38]. While regions most influential in the classification are understood to be present early in development, connections between these regions undergo significant development closer to the time of birth [39]. These regions have also been suggested as contributors in conditions such as Autism Spectrum Disorder (ASD) [1]. The thalamus is one key region highlighted in our results. Neuroscientific studies have suggested that connective differences in the thalamus are linked to epilepsy [40], and it is well established that the risk for epilepsy is increased by preterm birth [3, 2]. Another region which was particularly indicated in the LR model was cingulum, which has previously been implicated in developmental conditions following preterm birth [4]. Moreover, Figure 5 illustrates that the connections between certain regions are influential in the classification, and demonstrates that higher feature values on certain edges can indicate either preterm or term depending on which regions they connect. For example this suggests that for higher feature values of the connection between Frontal_Sup_L and Frontal_Sup_Medial_L, the model predicts cases of term infants. For the same edge with lower feature values, the model is more likely to classify as preterm. Many of the regions indicated in Figure 5 are also indicated in Figure 4, suggesting that not only are the nodes important, but suggesting that important nodes also connect to other important nodes [8, 7].

Including fuzzy logic in the model is important, as not only does it improve the accuracy and precision of the classification, but it allows for individual differences. Individuals close to the term cut off of 37 are difficult to classify, which may suggest that some are more similar to the term class than others of the same GA. This is important for further studies to take into account, and may also explain why it is difficult to predict pediatric outcomes for this particular group of infants [14, 17].

Future work could consider alternative explainable models, and compare the insights with those we presented from SHAP, as well as comparisons to the neuroscientific knowledge. Moreover, our study is limited by data scarcity, so future studies should aim to incorporate any further data which becomes available. While we addressed the issue of class inbalance through undersampling, further work could explore alternative methods for counteracting this.

6. Conclusion

In this paper, we compare matrix-based and graph-based classifiers for distinguishing preterm from term infant brain connectomes and show that adding spatial atlas coordinates improves model performance. We apply a fuzzy-logic boundary around 37 weeks' GA to the best LR and GAT models to smooth the decision threshold, raising accuracy from 88.57% to 93.33% with spatial coordinates and to 96.19% when adding fuzzy logic, which lets the model learn gradual maturational changes. Among graph neural networks, GAT outperforms GCN but still falls short of LR, demonstrating that a simple linear model can separate both classes without added complexity.

Using SHAP edge and node importance scores, we identify a consistent set of deep-brain regions (thalamus, putamen, cingulum) and their connections as the primary classification drivers, aligning with known neuro-developmental findings on preterm risk. This interpretable mapping suggests potential biomarkers for early developmental screening. Future work can expand the cohort and treat gestational age as a continuous variable to refine sensitivity in the late-preterm window and support more personalized assessments.

Acknowledgments

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) DTP Studentship 2753824 for the University of Surrey.
Data were provided by the developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial.

Declaration on Generative AI

During the preparation of this work, the authors used GenAI for small text revisions and clarity. The authors reviewed and edited any content and take full responsibility for the publication's content.

References

[1] M. Schuetze, M. T. M. Park, I. Y. Cho, F. P. MacMaster, M. M. Chakravarty, S. L. Bray, Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder, Neuropsychopharmacology 41 (2016) 2627–2637.

[2] G. Ball, J. P. Boardman, D. Rueckert, P. Aljabar, T. Arichi, N. Merchant, I. S. Gousias, A. D. Edwards, S. J. Counsell, The effect of preterm birth on thalamic and cortical development, Cerebral cortex 22 (2012) 1016–1024.

[3] M. A. Ream, L. Lehwald, Neurologic consequences of preterm birth, Current neurology and neuroscience reports 18 (2018) 1–10.

[4] H. J. Lee, H. Kwon, J. I. Kim, J. Y. Lee, J. Y. Lee, S. Bang, J.-M. Lee, The cingulum in very preterm infants relates to language and social-emotional impairment at 2 years of term-equivalent age, NeuroImage: Clinical 29 (2021) 102528.

[5] Y. Taoudi-Benchekroun, D. Christiaens, I. Grigorescu, O. Gale-Grant, A. Schuh, M. Pietsch, A. Chew, N. Harper, S. Falconer, T. Poppe, et al., Predicting age and clinical risk from the neonatal connectome, NeuroImage 257 (2022) 119319.

[6] D. Batalle, E. J. Hughes, H. Zhang, J.-D. Tournier, N. Tusor, P. Aljabar, L. Wali, D. C. Alexander, J. V. Hajnal, C. Nosarti, et al., Early development of structural networks and the impact of prematurity on brain connectivity, Neuroimage 149 (2017) 379–392.

[7] V. R. Karolis, S. Froudist-Walsh, P. J. Brittain, J. Kroll, G. Ball, A. D. Edwards, F. Dell'Acqua, S. C. Williams, R. M. Murray, C. Nosarti, Reinforcement of the brain's rich-club architecture following early neurodevelopmental disruption caused by very preterm birth, Cerebral Cortex 26 (2016) 1322–1335.

[8] R. Bauer, M. Kaiser, Nonlinear growth: an origin of hub organization in complex networks, Royal Society open science 4 (2017) 160691.

[9] H. Cui, W. Dai, Y. Zhu, X. Kan, A. A. C. Gu, J. Lukemire, L. Zhan, L. He, Y. Guo, C. Yang, Braingb: a benchmark for brain network analysis with graph neural networks, IEEE transactions on medical imaging 42 (2022) 493–506.

[10] B. C. Van Wijk, C. J. Stam, A. Daffertshofer, Comparing brain networks of different size and connectivity density using graph theory, PloS one 5 (2010) e13701.

[11] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[12] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[13] J. H. Gilmore, R. C. Knickmeyer, W. Gao, Imaging structural and functional brain development in early childhood, Nature Reviews Neuroscience 19 (2018) 123–137.

[14] W. A. Engle, A recommendation for the definition of "late preterm"(near-term) and the birth weight–gestational age classification system, in: Seminars in perinatology, volume 30, Elsevier, 2006, pp. 2–7.
[15] X. Hou, P. Zhang, L. Mo, C. Peng, D. Zhang, Sensitivity to vocal emotions emerges in newborns at 37 weeks gestational age, Elife 13 (2024) RP95393.
[16] A. M. Jukic, D. D. Baird, C. R. Weinberg, D. R. McConnaughey, A. J. Wilcox, Length of human pregnancy and contributors to its natural variation, Human reproduction 28 (2013) 2848–2855.
[17] S. Karnati, S. Kollikonda, J. Abu-Shaweesh, Late preterm infants–changing trends and continuing challenges, International Journal of Pediatrics and Adolescent Medicine 7 (2020) 38–46.
[18] J.-M. Moutquin, Classification and heterogeneity of preterm birth, BJOG: An International Journal of Obstetrics & Gynaecology 110 (2003) 30–33.
[19] J. S. de Almeida, D.-E. Meskaldji, S. Loukas, L. Lordier, L. Gui, F. Lazeyras, P. S. Hüppi, Preterm birth leads to impaired rich-club organization and fronto-paralimbic/limbic structural connectivity in newborns, NeuroImage 225 (2021) 117440.
[20] D. Scheinost, A. Pollatou, A. J. Dufford, R. Jiang, M. C. Farruggia, M. Rosenblatt, H. Peterson, R. X. Rodriguez, J. Dadashkarimi, Q. Liang, et al., Machine learning and prediction in fetal, infant, and toddler neuroimaging: a review and primer, Biological psychiatry 93 (2023) 893–904.
[21] B. Surányi, L. Kovács, L. Szilágyi, Segmentation of brain tissues from infant mri records using machine learning techniques, in: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), IEEE, 2021, pp. 000455–000460.
[22] C. D. Smyser, N. U. Dosenbach, T. A. Smyser, A. Z. Snyder, C. E. Rogers, T. E. Inder, B. L. Schlaggar, J. J. Neil, Prediction of brain maturity in infants using machine-learning algorithms, NeuroImage 136 (2016) 1–9.
[23] E. Messaritaki, S. I. Dimitriadis, D. K. Jones, Optimization of graph construction can significantly increase the power of structural brain network studies, NeuroImage 199 (2019) 495–511.
[24] E. Vlamou, B. Papadopoulos, Fuzzy logic systems and medical applications, AIMS neuroscience 6 (2019) 266.
[25] J. L. Pappas, Biological taxonomic problem solving using fuzzy decision-making analytical tools, Fuzzy sets and systems 157 (2006) 1687–1703.
[26] J. Cao, T. Zhou, S. Zhi, S. Lam, G. Ren, Y. Zhang, Y. Wang, Y. Dong, J. Cai, Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—a comprehensive review, Information Sciences 662 (2024) 120212.
[27] S. S. Godil, M. S. Shamim, S. A. Enam, U. Qidwai, Fuzzy logic: A "simple" solution for complexities in neurosciences?, Surgical neurology international 2 (2011) 24.
[28] K. Munir, A. de Ramón-Fernández, S. Iqbal, N. Javaid, Neuroscience patient identification using big data and fuzzy logic–an alzheimer's disease case study, Expert Systems with Applications 136 (2019) 410–425.
[29] F. Shi, P.-T. Yap, G. Wu, H. Jia, J. H. Gilmore, W. Lin, D. Shen, Infant brain atlases from neonates to 1-and 2-year-olds, PloS one 6 (2011) e18746.
[30] A. Schuh, A. Makropoulos, E. C. Robinson, L. Cordero-Grande, E. Hughes, J. Hutter, A. N. Price, M. Murgasova, R. P. A. Teixeira, N. Tusor, et al., Unbiased construction of a temporally consistent morphological atlas of neonatal brain development, BioRxiv (2018) 251512.
[31] J.-A. Quinn, F. M. Munoz, B. Gonik, L. Frau, C. Cutland, T. Mallett-Moore, A. Kissou, F. Wittke, M. Das, T. Nunes, et al., Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data, Vaccine 34 (2016) 6047–6056.
[32] E. J. Hughes, T. Winchman, F. Padormo, R. Teixeira, J. Wurie, M. Sharma, M. Fox, J. Hutter, L. Cordero-Grande, A. N. Price, et al., A dedicated neonatal brain imaging system, Magnetic resonance in medicine 78 (2017) 794–804.
[33] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, A. Hussain, A hybrid mental health prediction model using support vector machine, multilayer perceptron, and random forest algorithms, Healthcare Analytics 3 (2023) 100185.
[34] S. Zhang, H. Tong, J. Xu, R. Maciejewski, Graph convolutional networks: a comprehensive review,

Computational Social Networks 6 (2019) 1–23.
[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).
[36] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, arXiv preprint arXiv:1810.00826 (2018).
[37] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (2009) 427–437.
[38] P. Rea, Essential clinical anatomy of the nervous system, Academic Press, 2015.
[39] A. Taymourtash, E. Schwartz, K.-H. Nenning, D. Sobotka, R. Licandro, S. Glatter, M. C. Diogo, P. Golland, E. Grant, D. Prayer, et al., Fetal development of functional thalamocortical and cortico–cortical connectivity, Cerebral Cortex 33 (2023) 5613–5624.
[40] L. V. Marcuse, M. Langan, P. R. Hof, F. Panov, I. Saez, J. Jimenez-Shahed, M. Figee, H. Mayberg, J. Y. Yoo, S. Ghatan, et al., The thalamus: Structure, function, and neurotherapeutics, Neurotherapeutics (2025) e00550.

A. Histogram of Gestational Age

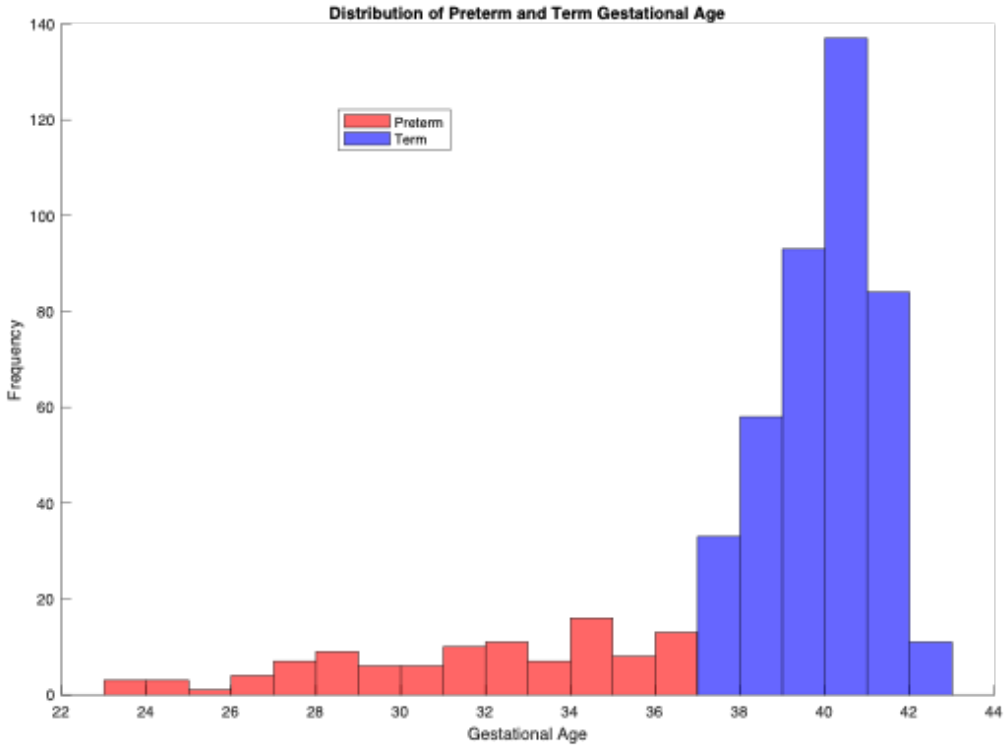

Figure 6: Gestational age histogram (including cases between 35–37 weeks). Demonstrates class imbalance.