Iranian Journal of Fuzzy Systems

Volume 23, Number 3, (2026), pp. 165-191

Original paper

University of Sistan and Baluchestan

# A multi-level fuzzy explainable prototype network for sex classification across brain atlases

M. Pakravan [1]

[1] *Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran*

mpakravan@modares.ac.ir

**Abstract**

Sex differences in resting-state functional brain organization have been widely reported, yet many deep learning models prioritize predictive performance over interpretability, limiting their ability to reveal underlying neurobiological mechanisms. We propose FXP-Net, a fuzzy explainable prototype-based neural network that combines temporal convolutional modeling with fuzzy prototype learning to enable intrinsically interpretable sex classification from resting-state fMRI. FXP-Net maps windowed ROI time series into a compact latent space in which class-specific prototypes are learned and activated through graded fuzzy membership functions, allowing each prediction to be explained via similarity to a small set of interpretable prototypes rather than opaque feature activations. To enforce interpretability, we introduce a prototype-regularized training objective that encourages intra-class clustering, inter-class separation, prototype diversity, and sparse activation. Across multiple brain atlases and repeated runs, FXP-Net achieves subject-level classification performance competitive with, and in several settings statistically superior to, a strong ConvNet baseline, with peak balanced accuracy exceeding 95%. Importantly, a multi-level interpretability analysis, combining prototype-level and decision-level attributions reveal anatomically consistent patterns across Brainnetome, Glasser, and Gordon parcellations. Quantitative overlap analysis highlights regions within the dorsolateral prefrontal cortex (DLPFC), including lateral area 9/10 and area 46, that demonstrate robust cross-atlas agreement and consistently high importance. These findings are stable across runs and align with prior evidence linking DLPFC organization to sex-related differences in executive control and higher-order cognition. The Python implementation of the proposed FXP-Net architecture is publicly available at: `https://github.com/Mansooreh-Pakravan/FXPNet-Model`.

*Keywords:* Fuzzy prototype learning, explainable deep learning, resting-state fMRI, sex differences, multi-atlas brain analysis.

## 1 Introduction

Sex differences in human brain organization have been repeatedly reported across structural and functional imaging modalities, motivating connectome-level representations for studying sex-related signatures. Early large-scale evidence from diffusion-based structural connectomics demonstrated systematic differences in inter- versus intra-hemispheric connectivity patterns, suggesting that sex effects manifest at the level of distributed network topology rather than isolated regions [47]. With the availability of large public cohorts such as the Human Connectome Project (HCP) and UK Biobank, resting-state fMRI (rs-fMRI) has further enabled fine-grained investigations of sex-related effects in intrinsic functional organization. In particular, consistent differences have been reported across association systems such as the frontoparietal and default-mode networks, which are closely associated with executive control and higher-order cognition [13, 41].

Beyond descriptive neuroimaging analyses, many studies have formulated sex identification as a supervised learning problem using rs-fMRI-derived features. These works have shown that sex can be classified from resting-state connec-

Corresponding Author: M. Pakravan
Received: December 2025; Revised: June 2026; Accepted: June 2026.
https://doi.org/10.22111/ijfs.2026.54363.9632

tivity patterns and that the contributing networks can be characterized, emphasizing that sex information is encoded in distributed functional systems rather than in a single dominant region [28, 47]. Complementarily, sex classification can also be achieved using temporal properties of rs-fMRI signals, such as long-range temporal dependence, indicating that sex-related signatures are present not only in static connectivity but also in the dynamical fluctuations of resting activity [13]. Together, these findings suggest that both spatial topology and temporal dynamics contain complementary information for sex classification.

In parallel, GNNs have become a natural framework for modeling the brain as a structured connectome, enabling topology-aware representation learning and interaction modeling among brain regions. Interpretable GNN architectures have been successfully applied to functional connectivity analysis and disease classification from fMRI data [18, 30], while spatio-temporal attention-based graph models have been proposed to capture dynamic connectome patterns and evolving brain network interactions [25].

To leverage these high-dimensional and distributed signatures, recent work has increasingly adopted deep learning to improve predictive performance and to exploit large-scale neuroimaging datasets. Deep models and advanced machine learning approaches for sex classification have been proposed using functional brain representations derived from large cohorts, demonstrating that neural architectures can learn discriminative patterns from high-dimensional neuroimaging inputs [1, 47]. More broadly, modern machine learning pipelines have been used to relate functional connectivity to demographic and behavioral traits, supporting the view that distributed brain networks provide stable signatures of individual differences [16, 23]. However, despite strong predictive performance, many of these approaches operate as black boxes: their internal mechanisms are difficult to interpret, and the inferred biomarkers can be sensitive to analysis choices such as thresholding, atlas selection, and sample variability [23, 37]. This lack of transparency limits their neuroscientific utility and complicates validation and replication of reported sex effects.

Alongside these modeling advances, XAI has emerged as a critical component in neuroimaging research to ensure that deep learning models produce neuroscientifically meaningful interpretations rather than purely predictive outcomes. Recent work has highlighted the challenges and methodological considerations involved in interpreting machine-learning-derived brain biomarkers, at the same time comprehensive surveys have outlined the development of medical XAI frameworks and their application to brain imaging analysis [44]. More recent reviews have further emphasized the growing role of explainability in improving transparency and reliability in brain disease diagnosis systems [4]. Representative recent studies include explainable connectome transformers for multimodal classification [32], AI-based systematic reviews highlighting methodological trends and pitfalls in neuroimaging pipelines [2], and GNN-driven explainability frameworks designed for identifying functional subnetworks in resting-state fMRI [36, 42]. Collectively, these developments highlight a clear trend toward integrating powerful deep learning architectures with interpretable mechanisms in neuroimaging analysis.

Recent advances in neuroimaging analysis have increasingly incorporated transformer-based architectures, graph neural networks (GNNs), and explainable artificial intelligence (XAI) frameworks better to capture the complex spatiotemporal structure of brain activity and to improve interpretability of predictive models. Transformer-based models have shown strong potential for modeling fMRI data and connectomic representations due to their ability to capture long-range dependencies and global contextual relationships across distributed brain regions. Recent studies have explored transformer architectures for learning neural representations from intracranial recordings and fMRI data, including self-supervised transformer models for neural signal representation learning [46], transformer-based prediction of task-related activity from resting-state dynamics [26], hierarchical vision-transformer architectures adapted for 4D fMRI data [24], and transformer-based embedding frameworks for human brain functional representations [50].

Despite this progress, current approaches to rs-fMRI-based sex classification still exhibit several important limitations. First, most models prioritize predictive accuracy and provide explanations at a single level, typically post-hoc, decision-level attribution without explicitly modeling stable internal concepts that correspond to population-level sex-related patterns. Second, the robustness of inferred biomarkers with respect to common analysis choices, such as atlas selection and thresholding, is rarely assessed systematically, even though such choices can substantially affect which regions and networks appear important [6, 43]. Third, while uncertainty and graded evidence are intrinsic to noisy and heterogeneous brain signals, many pipelines rely on single-value connectivity estimates and hard thresholding, which may obscure subtle yet consistent effects and reduce reproducibility across folds, datasets, and experimental runs. Addressing these limitations requires frameworks that (i) provide multi-level interpretability by separating internal concept-level representations from sample-specific decision-level attributions, and (ii) explicitly incorporate uncertainty and cross-atlas validation to identify robust, biologically plausible biomarkers.

A complementary direction for enhancing robustness and interpretability is the integration of fuzzy reasoning into learning pipelines. Fuzzy-set theory provides a principled mechanism for representing graded evidence and uncertainty [48], and has long been used to improve transparency in pattern recognition and time-series classification through interpretable rules and soft membership assignments. Interpretable fuzzy structures have been exploited in diverse

settings such as fuzzy soft-set-based modelling and decision making [9, 10] and fuzzy data envelopment analysis with transparent efficiency scores under uncertainty [39]. These studies leverage fuzzy information granules and rule-based constructions to maintain human-readable semantics while handling noisy, imprecise data. In neuroimaging applications, fuzzy formulations are particularly appealing because brain signals are inherently variable, noisy, and heterogeneous across subjects and sessions, and relevant evidence may be distributed and partially overlapping. In our own recent work, FuzzyCAL and related models [34, 35], we followed the same philosophy in a neuroimaging context by integrating fuzzy logic with causal attention GNNs to obtain graded, concept-level explanations for disorder classification from rs-fMRI. This growing body of research supports the broader view that combining fuzzy reasoning with data-driven models is an effective strategy for enhancing robustness and interpretability in complex connectomic pattern-recognition tasks.

Recent advances underscore the synergy between principled learning and application-specific constraints. For instance, structured models have improved medical signal classification [11], while optimized centroid updates have enhanced the stability of prototype-based representations [5]. Additionally, recent work on edge-computing pattern recognition highlights the need for efficient learning pipelines [33]. Inspired by these developments, we propose FXP-Net for sex classification from rs-fMRI signals. Our model integrates end-to-end deep learning with an interpretable fuzzy prototype-based framework for robust evidence formation.

Motivated by these challenges and ideas, we propose a multi-atlas, multi-level interpretability framework for sex classification from rs-fMRI. Our approach introduces FXP-Net, a fuzzy, explainable prototype network that explicitly integrates three widely used cortical parcellations, Brainnetome, Gordon, and Glasser, each reflecting a different anatomical or functional perspective on brain organization. Rather than treating atlas variability as a nuisance, we leverage cross-atlas agreement as a robustness criterion, identifying regions that consistently emerge as important across parcellation schemes. Moreover, we distinguish between two complementary levels of explanation. At the *prototype level*, we analyze how brain regions contribute to stable, class-specific latent concepts learned by the model, capturing population-level patterns that define male and female prototypes. At the *decision level*, we quantify how regions directly influence the final classification outcome for individual samples. By jointly analyzing these two levels and grounding them in a fuzzy representation of graded evidence, we disentangle stable representational evidence from sample-specific decision drivers and explicitly incorporate uncertainty into the interpretability process. This design aims to provide a more transparent, reliable, and neuroscientifically grounded understanding of sex differences in functional brain organization.

While prior studies have made important contributions to sex classification from resting-state fMRI, they mainly differ from our work in either modeling strategy, interpretability, or atlas analysis. For example, Dhamala et al. [13] and Weis et al. [47] investigated sex-related connectivity patterns using conventional connectivity-based analyses, while Leming and Suckling [28] and Ryali et al. [41] employed deep learning models to improve classification performance and identify relevant brain organization patterns. Ritchie et al. [40] reported large-scale sex differences in brain structure, but their study was not based on fuzzy, prototype-based explainability for rs-fMRI classification.

**Novelty of the Proposed Framework** The proposed framework introduces three methodological novelties. First, it uses a fuzzy prototype-based model that represents brain patterns through soft membership to latent prototypes, rather than hard assignment or purely post-hoc feature attribution. Second, it provides a two-level interpretability analysis by separating prototype-level explanations from decision-level explanations, allowing both conceptual and prediction-specific interpretation. Third, it evaluates ROI importance across three atlases and introduces cross-atlas agreement as a robustness criterion, enabling the identification of regions whose relevance is stable across different parcellation schemes. To the best of our knowledge, this combination of fuzzy modeling, multi-level interpretability, and multi-atlas robustness analysis has not been previously explored in sex classification studies based on rs-fMRI.

The main contributions of this work are summarized as follows:

- We propose an interpretability framework that integrates decision-level and prototype-level explanations for deep learning–based sex classification from rs-fMRI.
- We conduct a multi-atlas analysis (Brainnetome, Gordon, and Glasser) and introduce cross-atlas overlap as a criterion for identifying robust and anatomically consistent regions.
- Using this framework, we identify stable regions across explanation levels and atlases, highlighting dorsolateral prefrontal cortex (DLPFC) territories, particularly lateral area 9/10 and area 46 as interpretable markers of sex-related functional differences.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing pipeline, model architecture, and the proposed interpretability framework. Section 3 presents quantitative classification results and

detailed interpretability analyses across atlases and explanation levels. In Section 4, we discuss the neurobiological implications of the identified regions and relate our findings to existing literature. Finally, Section 5 concludes the paper and Section 6 outlines limitations and directions for future research.

# 2 Materials and methods

## 2.1 Materials

This study is based on the publicly available *Human Connectome Project (HCP) 1200 Subjects Release* [45], which provides high-quality multimodal MRI data from healthy young adults aged 22–35 years. All data were acquired as part of the WU–Minn HCP consortium under protocols approved by the Washington University Institutional Review Board, with informed consent obtained from all participants.

Each subject completed four resting-state fMRI (rs-fMRI) runs, each comprising 1200 time frames (approximately 14.5 minutes per run), resulting in a total of 4800 volumes per subject. Two runs were acquired in one session and two in a separate session on a different day. During scanning, participants were instructed to remain still with eyes open and fixate on a crosshair. Phase-encoding directions alternated between left-to-right (LR) and right-to-left (RL) across runs to mitigate susceptibility-related distortions.

All rs-fMRI data were acquired on a 3 T Siemens Connectome Skyra scanner using a multiband gradient-echo EPI sequence (TR = 720 ms, TE = 33.1 ms, flip angle = 52°). The data were processed using the HCP minimal preprocessing pipelines [19], including motion correction, distortion correction, spatial normalization, and projection to the grayordinate space (approximately 91k cortical and subcortical grayordinates).

To further assess the potential confounding effect of head motion, we quantified the mean Framewise Displacement (FD; [38]) for all participants using the publicly available motion parameters provided with the HCP dataset. Motion regressors were extracted, and FD was computed according to standard procedures. A two-sample statistical comparison of mean FD between female and male participants showed no significant difference $p > 0.05$. This finding is consistent with previous large-scale analyses of the HCP S1200 cohort, which likewise reported no significant sex-related differences in mean FD [47]. Taken together, these results indicate that the observed differences in fuzzy connectivity patterns are unlikely to be driven by residual head motion artifacts.

After excluding subjects with missing or ambiguous sex labels, the final sample consisted of 550 females and 656 males ($N = 1206$), yielding a moderately balanced sex distribution. Age distributions for female and male participants showed substantial overlap within the targeted young adult range. No significant age difference was observed between groups, reducing the likelihood that age acts as a confounding factor in the subsequent sex-difference analyses.

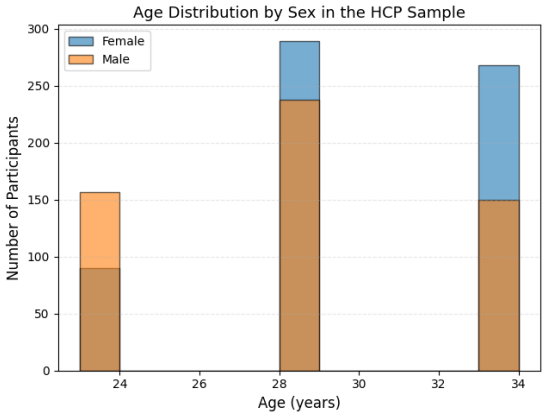


Figure 1: Age distribution of female and male participants in the analyzed HCP cohort. Both groups span a similar age range (22–35 years) with substantial overlap, indicating minimal age-related confounding in sex-based comparisons.

Resting-state fMRI BOLD time series were analyzed after atlas-based parcellation using three widely adopted brain atlases (Brainnetome, Gordon, and Glasser). Table 1 summarizes the main characteristics of the brain atlases used in this study and highlights their conceptual differences in terms of parcellation strategy. Using all three atlases enables us to evaluate the robustness and interpretability of sex-related findings across distinct spatial and functional representations of the brain.

Table 1: Brain atlases used in this study and their conceptual differences.

| Atlas | #ROIs | Coverage | Parcellation principle |
|---|---|---|---|
| **Brainnetome** [15] | 246 | Cortex + subcortex | Connectivity-driven atlas derived from connectional architecture (structural and functional connectivity profiles), providing fine-grained brain regions. |
| **Glasser (HCP-MMP1.0)** [20] | 360 | Cortex only | Multi-modal parcellation combining myelin maps, cortical thickness, task activations, and connectivity information from the HCP dataset. |
| **Gordon** [21] | 333 | Cortex (with subcortical extensions) | Resting-state correlation-based parcellation designed to delineate large-scale functional systems and networks. |

For each subject and each acquisition run, the data were represented as a matrix $\mathbf{Y} \in \mathbb{R}^{N_{\mathrm{ROI}} \times T}$, where $T$ denotes the number of time points and $N_{\mathrm{ROI}}$ is the number of regions of interest (ROIs) defined by the selected atlas. All preprocessing steps were applied at the ROI time-series level and were identical across runs, folds, and atlases to ensure methodological consistency.

To reduce slow scanner drifts and very low-frequency trends, each ROI time series was first linearly detrended along the temporal dimension. We further removed ultra-slow fluctuations by regressing out a low-order temporal nuisance model consisting of an intercept, a linear trend, and a small set of sinusoidal harmonic components. This trend-plus-harmonics regression effectively suppresses long-period oscillations and baseline drifts that are unlikely to reflect neural activity, while preserving the temporal structure relevant for functional analysis. Temporal filtering was applied using a zero-phase band-pass filter to retain frequencies in the conventional resting-state range (0.008–0.09 Hz). Zero-phase filtering was employed to avoid phase distortions and to preserve the temporal alignment of BOLD fluctuations across ROIs. Then, each ROI time series was standardized by z-scoring across time, yielding zero-mean and unit-variance signals. This normalization step ensures that all ROIs contribute on a comparable scale and prevents regions with intrinsically higher variance from disproportionately influencing the learning process.

For model training and evaluation, the preprocessed time series were segmented into overlapping sliding windows of fixed length. These windows were treated as individual samples during optimization, while subject identity was retained to allow robust aggregation of predictions and interpretability measures at the subject level during evaluation.

## 2.2 Methods

Figure 2 depicts the full methodological pipeline used in this study.

The analysis begins with rs-fMRI time series, which undergo windowing and standard preprocessing to generate temporally segmented input samples. These samples are mapped into a latent representation that serves as the common feature space for subsequent modeling.

From this shared latent space, two complementary modeling paths are explored. First, a baseline classifier is trained to quantify reference performance independent of interpretability constraints. Second, the proposed FXP-Net incorporates fuzzy prototype learning, enabling interpretable decision formation through prototype-regularized training. The fuzzy prototypes aim to capture characteristic latent-space patterns associated with each class.

Model outputs from both paths are aggregated at the subject level, after which cross-validation and evaluation procedures are performed. This schematic overview provides a structured representation of all major components of the methodology; the following subsections describe each stage in detail.

### 2.2.1 Baseline model: Temporal ConvNet

As a baseline, we employ a one-dimensional temporal convolutional network (ConvNet) directly applied to atlas-parcellated rs-fMRI time series. For each temporal window, the model receives an input tensor

$$\mathbf{X} \in \mathbb{R}^{B \times N_{\mathrm{ROI}} \times T},$$

where $B$ is the batch size, $N_{\mathrm{ROI}}$ is the number of cortical regions, and $T$ is the window length. Each ROI is treated as a separate input channel, allowing the network to learn multiregional temporal patterns jointly.

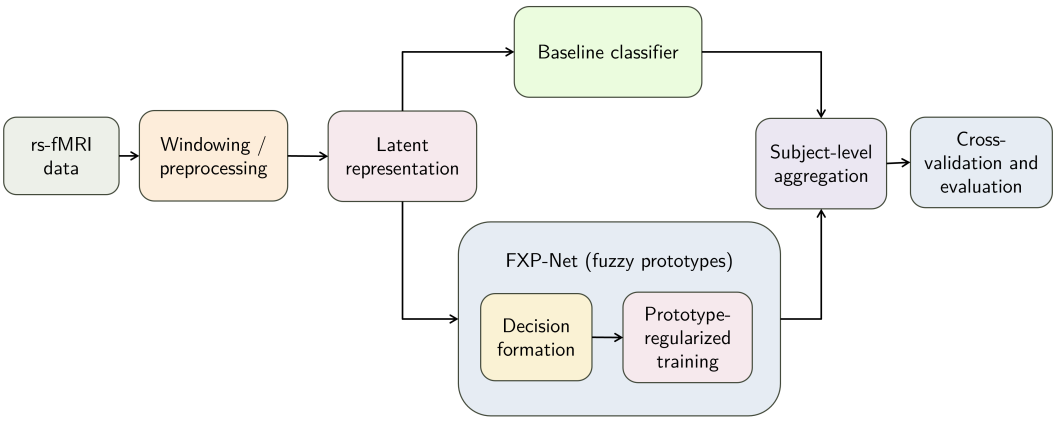


Figure 2: Overview of the methodological pipeline. rs-fMRI time series are windowed and preprocessed to generate standardized input segments, which are encoded into a latent representation. From this shared representation, two modeling paths are considered: a baseline classifier and the proposed FXP-Net with fuzzy prototypes, which performs prototype-regularized training and interpretable decision formation. Outputs from both models are aggregated at the subject level before cross-validation and evaluation.

The network begins with a temporal convolutional block that extracts mid-level temporal features. A first convolution layer, Conv1d($N_{\text{ROI}} \to 256$, $k=7$) with temporal padding, is followed by batch normalization, ReLU activation, and max pooling (stride 2). A second convolution, Conv1d($256 \to 256$, $k=7$), refines these features, after which an additional max pooling stage (kernel size 4, stride 2) further compresses the temporal resolution.

A second convolutional block expands the representational capacity through a temporal convolution Conv1d($256 \to 512$, $k=5$) with ReLU activation, followed by max pooling (kernel size 4, stride 2). This yields a compact 512-channel temporal representation.

To produce a window-level embedding, global average pooling is applied across time, generating a fixed 512-dimensional feature vector. Dropout ($p = 0.5$) is then used to regularize training, and a fully connected layer maps the pooled features to two logits corresponding to female and male classes.

In summary, the baseline ConvNet provides a strong discriminative model capable of learning hierarchical temporal dynamics in multivariate BOLD signals, yet it lacks mechanisms for interpretability—motivating the introduction of fuzzy prototype-based reasoning in the proposed FXP-Net.

### 2.2.2 Proposed model: FXP-Net (Fuzzy Explainable Prototype Network)

To achieve *intrinsic interpretability* without sacrificing classification performance, we propose FXP-Net, a fuzzy prototype-based neural architecture for resting-state fMRI analysis. FXP-Net is explicitly designed so that each prediction can be explained in terms of similarity to a small number of learned prototypes in a low-dimensional latent space.

Figure 3 illustrates the FXP-Net architecture at a conceptual level. The model learns a compact latent embedding where a small set of class-specific fuzzy prototypes provides interpretable evidence for the male vs. female decision. A residual linear head is included to maintain discriminative robustness.

The following subsections describe each component of the FXP-Net architecture in detail and explain how these stages are combined to form the complete learning and decision-making pipeline.

**(1) Temporal feature extraction.** Given a sliding-window input

$$\mathbf{X} \in \mathbb{R}^{B \times N_{\text{ROI}} \times T}, \tag{1}$$

where $B$ is the batch size, $N_{\text{ROI}}$ the number of atlas-defined brain regions, and $T$ the window length, FXP-Net first applies a lightweight temporal ConvNet backbone. This backbone consists of a sequence of two one-dimensional convolutions along the temporal axis with intermediate pooling operations, followed by global average pooling over time. The output is a fixed-length feature vector

$$\mathbf{h} = f_\theta(\mathbf{X}) \in \mathbb{R}^{B \times d_f}, \tag{2}$$

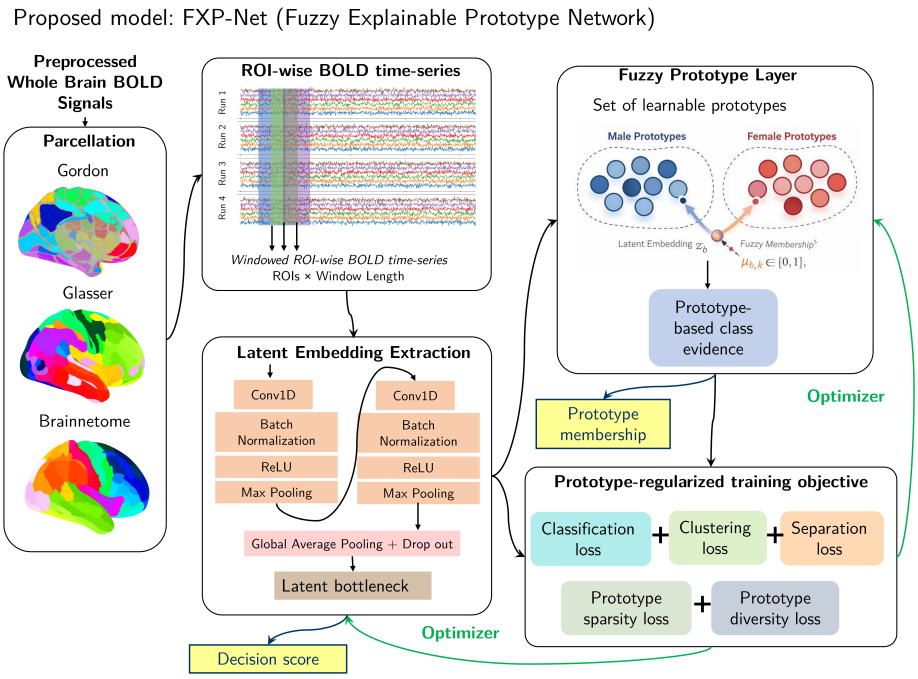


Figure 3: High-level architecture of the proposed FXP-Net. A parcellated fMRI ROI time-series window is encoded by a lightweight temporal CNN, compressed by a bottleneck into a latent embedding, and matched to class-specific fuzzy prototypes to obtain membership-based evidence. A residual linear head provides an auxiliary discriminative pathway, and the final logits combine both prototype and linear evidence. This structure supports prototype-level (membership-based) and decision-level (logit-based) explanations.

with feature dimension $d_f = 256$. The backbone architecture closely mirrors the baseline ConvNet, ensuring comparable temporal modeling capacity.

**(2) Latent bottleneck.** To obtain a compact and structured representation, the backbone features are projected into a low-dimensional latent space via a bottleneck mapping

$$\mathbf{z} = g_\phi(\mathbf{h}) \in \mathbb{R}^{B\times d}, \qquad d = 32, \tag{3}$$

where $g_\phi(\cdot)$ is implemented as a linear layer followed by a ReLU nonlinearity. This bottleneck limits representational freedom and encourages the network to encode each window using a small number of latent factors, which is essential for stable prototype learning and interpretability.

**(3) Fuzzy prototype layer.** FXP-Net learns a set of $K$ prototypes $\{\mathbf{p}_k\}_{k=1}^K$ in the latent space, with $\mathbf{p}_k \in \mathbb{R}^d$. For each latent embedding $\mathbf{z}_b$, the model computes a *fuzzy membership score* to each prototype:

$$\mu_{b,k} = \exp\left(-\frac{\|\mathbf{z}_b - \mathbf{p}_k\|_2^2}{2\sigma_k^2}\right), \quad \mu_{b,k} \in (0,1], \tag{4}$$

where $\sigma_k$ is a learnable prototype-specific scale parameter.

In the above formulation, $\sigma_k$ controls the spread of the membership function associated with the $k$-th prototype. Conceptually, this parameter determines how quickly the membership value decreases as the latent representation moves away from the prototype center in the embedding space. A smaller $\sigma_k$ produces a sharper membership function, resulting in more localized and selective prototype assignments, whereas a larger $\sigma_k$ yields smoother and more diffuse memberships. Consequently, $\sigma_k$ regulates the degree of fuzziness in the prototype–sample association and influences how broadly latent representations can contribute to multiple prototypes, thereby promoting smoother decision boundaries and improved robustness to variability in the latent space.

These memberships provide a graded notion of similarity, allowing each input window to be partially associated with multiple prototypes. The vector of memberships $\boldsymbol{\mu}_b = (\mu_{b,1}, \ldots, \mu_{b,K})$ is then passed to the subsequent classification

layer, so that class decisions are formed directly from fuzzy degrees of belonging to the learned prototypes rather than from hard, crisp assignments.

Importantly, both the prototype centers $\mathbf{p}_k$, and their spreads $\sigma_k$ are learned jointly with the rest of the network parameters via backpropagation: they are initialized randomly and updated by gradient-based optimization (Adam) using the classification loss. Thus, the parameters of the fuzzy logic component are data-driven and optimized end-to-end, without manual specification of membership functions.

**(4) Prototype-based class evidence.** An equal number of prototypes is allocated to each class. With $P = 8$ prototypes per class, the total number of prototypes is $K = 16$. Prototype memberships are aggregated within each class to form prototype-based logits.

To allow flexible contribution of individual prototypes, we introduce learnable nonnegative weights:

$$w_k = \mathrm{softplus}(\alpha_k) \geq 0, \tag{5}$$

where the softplus function is defined as

$$\mathrm{softplus}(x) = \log(1 + e^x). \tag{6}$$

This smooth transformation ensures strictly nonnegative prototype weights while preserving differentiability for stable gradient-based optimization.

The prototype-based evidence for class $C$ is computed as

$$\ell_{b,c}^{\mathrm{proto}} = \sum_{k \in \mathcal{P}_c} w_k \, \mu_{b,k}, \tag{7}$$

where $\mathcal{P}_c$ denotes the set of prototypes assigned to class $c \in \{0, 1\}$.

**(5) Residual linear head.** To preserve predictive accuracy and stabilize training, FXP-Net includes a residual linear classifier operating directly on the latent representation:

$$\boldsymbol{\ell}_b^{\mathrm{lin}} = \mathbf{W}\mathbf{z}_b + \mathbf{b} \in \mathbb{R}^2, \tag{8}$$

where $\mathbf{z}_b \in \mathbb{R}^d$ denotes the latent embedding of the $b$-th input window, $\mathbf{W} \in \mathbb{R}^{2\times d}$ is a learnable weight matrix, and $\mathbf{b} \in \mathbb{R}^2$ is a bias vector. The resulting vector $\boldsymbol{\ell}_b^{\mathrm{lin}}$ represents the class logits produced by the linear branch for the two classes (male and female).

**(6) Final decision rule.** The final class logits are obtained by combining prototype-based and linear evidence:

$$\boldsymbol{\ell}_b = \lambda \, \boldsymbol{\ell}_b^{\mathrm{proto}} + (1 - \lambda) \, \boldsymbol{\ell}_b^{\mathrm{lin}}, \tag{9}$$

where $\lambda \in [0, 1]$ controls the relative contribution of interpretable prototype evidence..

The index $b$ denotes the $b$-th window in the mini-batch, such that all logits and prototype memberships are defined at the window level.

**(7) Prototype-regularized training objective.** FXP-Net is trained using a composite objective that jointly optimizes classification accuracy and prototype interpretability. Consider a mini-batch of $B$ input windows $\{(\mathbf{x}_b, y_b)\}_{b=1}^{B}$, where $\mathbf{x}_b$ denotes the $b$-th window and $y_b \in \{0, 1\}$ is its ground-truth class label. For each window, the model outputs (i) final class logits $\boldsymbol{\ell}_b \in \mathbb{R}^2$, (ii) a latent embedding $\mathbf{z}_b \in \mathbb{R}^d$, and (iii) fuzzy prototype membership scores $\mu_{b,k} \in (0, 1]$ for $k = 1, \ldots, K$ prototypes. Each prototype $\mathbf{p}_k \in \mathbb{R}^d$ is assigned to exactly one class via a fixed prototype–class mapping.

The overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + w_{\mathrm{clust}}\mathcal{L}_{\mathrm{clust}} + w_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}} + w_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} - w_{\mathrm{sparse}}\mathcal{L}_{\mathrm{sparse}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the standard cross-entropy loss computed from the final logits $\boldsymbol{\ell}_b$, and the remaining terms act as prototype-specific regularizers. The scalar coefficients $w_{\mathrm{clust}}, w_{\mathrm{sep}}, w_{\mathrm{div}}, w_{\mathrm{sparse}} \geq 0$ control the relative strength of each regularization component.

**Classification loss.** The classification term encourages correct predictions at the window level. Let $B$ denote the mini-batch size and $\ell_{b,c}$ the logit corresponding to class $c$ for the $b$-th sample. The classification loss is defined as

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B}\sum_{b=1}^{B} \log \frac{\exp(\ell_{b,y_b})}{\sum_{c\in\{0,1\}} \exp(\ell_{b,c})}. \tag{11}$$

**Prototype regularization terms.** Let $\mathcal{P}_{y_b}$ denote the set of prototypes assigned to class $y_b$, and $\mathcal{P}_{\neg y_b}$ the set of prototypes assigned to the opposite class. The prototype regularization terms are defined as follows:

**(i) Clustering loss.** The clustering loss encourages each latent embedding $\mathbf{z}_b$ to be close to at least one prototype of its own class:

$$\mathcal{L}_{\mathrm{clust}} = \frac{1}{B}\sum_{b=1}^{B} \min_{k\in\mathcal{P}_{y_b}} \|\mathbf{z}_b - \mathbf{p}_k\|_2^2 . \tag{12}$$

**(ii) Separation loss.** To increase inter-class margins, the separation loss penalizes proximity to prototypes of the opposite class:

$$\mathcal{L}_{\mathrm{sep}} = \frac{1}{B}\sum_{b=1}^{B} \exp\!\left(-\min_{k\in\mathcal{P}_{\neg y_b}} \|\mathbf{z}_b - \mathbf{p}_k\|_2^2\right). \tag{13}$$

**(iii) Prototype diversity loss.** To avoid redundant or collapsed prototypes, diversity is enforced by penalizing high cosine similarity between distinct prototype vectors:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{K(K-1)} \sum_{k\neq k'} \left(\frac{\mathbf{p}_k^\top \mathbf{p}_{k'}}{\|\mathbf{p}_k\|_2 \, \|\mathbf{p}_{k'}\|_2}\right)^2 . \tag{14}$$

**(iv) Prototype sparsity loss.** Finally, sparsity is encouraged in the normalized membership distribution

$$\tilde{\mu}_{b,k} = \frac{\mu_{b,k}}{\sum_{k'=1}^{K} \mu_{b,k'}}, \tag{15}$$

via the squared $\ell_2$ norm

$$\mathcal{L}_{\mathrm{sparse}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} \tilde{\mu}_{b,k}^2, \tag{16}$$

which promotes peaked (low-entropy) prototype activations such that each window is explained by only a small number of prototypes.

**(8) Optimization and learning.** Training FXP-Net corresponds to solving a well-defined optimization problem. Let $\Theta$ denotes the full set of trainable parameters of the model (feature extractor weights, projection parameters, prototype vectors, prototype widths, and classifier weights). The objective of learning is to *minimize* the total loss function defined in Eq. (11):

$$\min_{\Theta} \; \mathcal{L}(\Theta). \tag{17}$$

This formulation follows the standard view in machine learning, where learning is equivalent to solving a continuous optimization problem over a high-dimensional parameter space. The gradient of the loss with respect to the parameters is computed by backpropagation, and the parameters are updated in the direction that reduces the loss.

We use the Adam optimization algorithm to solve this minimization problem. Adam is a first-order gradient-based optimizer that adaptively adjusts the learning rate for each parameter by keeping track of (i) the exponentially weighted average of past gradients (the "first moment") and (ii) the exponentially weighted average of squared gradients (the "second moment"). These adaptive estimates allow Adam to stabilize the optimization trajectory and accelerate convergence, especially in non-convex problems typical of deep neural networks.

An update step in Adam can be written generically as

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \, \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{18}$$

where $\hat{m}_t$ and $\hat{v}_t$ denote the bias-corrected moment estimates. Through repeated application of these updates, the model parameters gradually move toward values that minimize the loss. Thus, learning in FXP-Net is directly achieved by solving this optimization problem via Adam.

To ensure stable optimization and a balanced contribution of the different loss components, the key hyperparameters of the model (including the regularization weight $\lambda$ and the prototype-related coefficients) were selected through empirical tuning based on preliminary experiments on the training set. For each hyperparameter, we evaluated a range of candidate values and chose the configuration that provided the best validation performance while maintaining consistent and stable convergence. We observed that the model's performance remained relatively robust within a reasonable neighborhood of the selected values, indicating that the algorithm does not exhibit excessive sensitivity to small variations in these hyperparameters.

Overall, this composite objective enforces a structured latent space in which samples are attracted to class-consistent prototypes, repelled from opposite-class, prototypes remain mutually distinct, and a sparse and interpretable set of prototype activations supports each prediction.

Because each prediction is mediated through a small set of activated prototypes with high membership scores, FXP-Net is interpretable by construction. Prototype-level analysis reveals *which latent patterns* the model has learned, while decision-level analysis reveals *which brain regions* most strongly influence the final classification. This design provides a transparent bridge from temporal BOLD dynamics to latent prototypes and ultimately to atlas-defined brain regions.

**Algorithmic summary of FXP-Net.** To provide an operational view of the proposed approach, Algorithm 1 presents the full training procedure of FXP-Net and its relation to the equations introduced above.

**Algorithm 1** Training procedure of FXP-Net

**Require:** Mini-batch $\{\mathbf{x}_b, y_b\}_{b=1}^{B}$, model parameters $\theta, \phi, \{\mathbf{p}_k, \sigma_k, w_k\}_{k=1}^{K}$, learning rate $\eta$
**Ensure:** Updated parameters after one training step

**for each** window $\mathbf{x}_b$:
  Extract temporal features $\mathbf{h}_b = f_\theta(\mathbf{x}_b)$
  Compute latent embedding $\mathbf{z}_b = g_\phi(\mathbf{h}_b)$
**for each prototype** $k = 1, \ldots, K$:
  Compute fuzzy membership $\mu_{b,k} = \exp(-\|\mathbf{z}_b - \mathbf{p}_k\|_2^2/(2\sigma_k^2))$
Compute prototype-based logits $\ell_{b,c}^{\text{proto}} = \sum_{k \in \mathcal{P}_c} w_k \, \mu_{b,k}$
Compute residual linear logits $\boldsymbol{\ell}_b^{\text{lin}} = \mathbf{W}\mathbf{z}_b + \mathbf{b}$
Combine prototype and linear evidence $\boldsymbol{\ell}_b = \lambda \, \boldsymbol{\ell}_b^{\text{proto}} + (1 - \lambda)\boldsymbol{\ell}_b^{\text{lin}}$
Compute total loss $\mathcal{L} = \mathcal{L}_{\text{CE}} + w_{\text{clust}}\mathcal{L}_{\text{clust}} + w_{\text{sep}}\mathcal{L}_{\text{sep}} + w_{\text{div}}\mathcal{L}_{\text{div}} - w_{\text{sparse}}\mathcal{L}_{\text{sparse}}$
Compute gradients $\nabla\theta, \nabla\phi, \nabla\mathbf{p}_k, \nabla\sigma_k, \nabla w_k$
Update parameters using Adam:

$$\theta \leftarrow \text{Adam}(\theta, \nabla\theta), \quad \phi \leftarrow \text{Adam}(\phi, \nabla\phi),$$

$$\mathbf{p}_k \leftarrow \text{Adam}(\mathbf{p}_k, \nabla\mathbf{p}_k), \; \sigma_k \leftarrow \text{Adam}(\sigma_k, \nabla\sigma_k), \; w_k \leftarrow \text{Adam}(w_k, \nabla w_k).$$

## 2.3 Training and evaluation protocol

For each atlas (Brainnetome, Gordon, and Glasser), parcel-wise resting-state BOLD time series were extracted for all subjects with complete data across four independent runs. To ensure temporal consistency across subjects, all time series were truncated to the minimum common length within each run.

To capture fine-grained temporal dynamics, a sliding-window strategy was employed. Each subject's time series was segmented into overlapping windows of length 256 time points with a step size of 256 time points, resulting in 50% overlap between consecutive windows.

This duration was chosen based on the established principle that window lengths must be at least equal to $1/f_{min}$ (where $f_{min}$ is the minimum frequency of the filtered BOLD signal, typically $\sim 0.01$ Hz) to avoid spurious fluctuations in dynamic connectivity [29]. A length of 256 TRs provides an optimal trade-off between achieving sufficient statistical stability for connectivity estimation and maintaining adequate temporal resolution to detect transient functional states [22, 49]. Additionally, sequence lengths of 256 (a power of 2) are computationally advantageous for hierarchical representation learning in modern deep learning architectures [17]. Each window was treated as an independent sample with dimensions $N_{\text{ROI}} \times 256$.

Model evaluation was conducted using stratified 5-fold cross-validation at the *subject level*. Subjects were partitioned into training and validation folds while preserving the class distribution (male/female) in each fold. All windows belonging to a given subject were assigned exclusively to either the training or validation set, thereby preventing information leakage. For each atlas, the cross-validation split was fixed and reused across all runs to ensure consistency, and all reported results were obtained exclusively from validation subjects.

For each combination of atlas, run, and cross-validation fold, an independent FXP-Net model was trained. The network received windowed inputs of size $N_{\mathrm{ROI}} \times 256$ and produced binary predictions corresponding to biological sex. Training was performed using the Adam optimizer with a learning rate of $1 \times 10^{-4}$ and a batch size of 32. A maximum of 20 training epochs was used, with early stopping based on validation loss (patience of 5 epochs) to prevent overfitting and unnecessary computation. To ensure reproducibility, all experiments were conducted using a fixed random seed (2025) and deterministic computation settings.

To mitigate potential class imbalance arising from unequal numbers of training windows per class, a weighted cross-entropy loss was employed. Class weights were computed exclusively from the training data of each fold using an inverse-frequency strategy, thereby penalizing misclassification of the minority class more strongly and encouraging balanced decision boundaries. The number of training samples per run corresponded to the total number of sliding windows generated from the training subjects in each fold.

In addition to the weighted classification loss, FXP-Net incorporates a prototype-based fuzzy regularization mechanism designed to promote interpretability and structured latent representations. The latent embedding dimension was set to 32, and each class was represented by 8 learnable prototypes, resulting in a total of 16 prototypes. Prototype memberships were modeled using a differentiable fuzzy assignment mechanism controlled by a temperature parameter of 1.0 and an initialization scale of 1.0.

The overall training objective combines a weighted cross-entropy classification loss with a prototype-based regularization term that encourages structured and interpretable latent representations. In addition, standard regularization techniques are applied to improve generalization, including weight decay and dropout within the network.

At the subject level, window-level predictions belonging to the same subject are aggregated by averaging the predicted posterior probabilities across all windows (mean voting), followed by thresholding to obtain a binary subject-level prediction. Evaluation metrics, including accuracy, balanced accuracy, F1-score, precision, and recall, are computed at the subject level in order to mitigate variability introduced by individual temporal windows.

This aggregation strategy reduces the influence of temporal fluctuations and window-level noise in resting-state fMRI signals, yielding subject-level predictions that more reliably reflect consistent patterns across the entire recording. The same aggregation procedure is applied to both the baseline models and FXP-Net to ensure a fair comparison.

The entire training and evaluation pipeline is repeated independently for each run and across all cross-validation folds, resulting in multiple evaluations per atlas. Final results are reported as mean ± standard deviation across folds to assess both performance and stability.

Table 2 summarizes the full set of implementation details and hyperparameter settings used throughout the experiments, including optimization parameters, prototype configuration, regularization terms, data segmentation settings, and subject-level aggregation rules. This tabulated summary provides a concise reference that facilitates transparency, comparison across atlases, and reproducibility of the FXP-Net pipeline.

## 2.4 Interpretability at prototype-level vs. decision-level

To provide a rigorous yet transparent explanation of FXP-Net predictions, we report interpretability at two complementary levels: *prototype-level* (concept-wise evidence) and *decision-level* (output-driven, end-to-end evidence). As illustrated in Fig. 4, both branches use the same input (subject-level ROI time-series windows) and the same trained classifier. However, they differ in the *attribution target* used during backpropagation. In both cases, window-level attributions are first temporally pooled to obtain one importance value per ROI. These scores are then aggregated across validation subjects and across folds and runs to produce stable atlas-specific ROI importance maps.

**Prototype-level (concept-level attribution).** The FXP-Net architecture maintains a set of class-specific prototypes $\{p_k\}_{k=1}^{K}$ in a latent space and, for each input window $x$, produces: (i) prototype membership (activation) scores $\mu_k(x) \in (0, 1]$, and (ii) class logits. Prototype-level interpretability addresses the question: *which ROIs most strongly drive the activation of a given prototype (concept)?* Concretely, for each prototype $k$, we compute a Grad×Input attribution with respect to the scalar target $\mu_k(x)$:

$$A^{(k)}_{\mathrm{proto}}(x) \;=\; \left| \frac{\partial \mu_k(x)}{\partial x} \odot x \right|. \tag{19}$$

Table 2: Implementation details and training hyperparameters of FXP-Net.

| **Component** | **Configuration** |
|---|---|
| Input window size | $N_{\text{ROI}} \times 256$ time points |
| Sliding window step | 128 time points (50% overlap) |
| Cross-validation | Stratified 5-fold (subject-level split) |
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-4}$ |
| Batch size | 32 |
| Maximum epochs | 20 |
| Early stopping | Patience = 5 (based on validation loss) |
| Random seed | 2025 (deterministic computation) |
| Loss function | Weighted cross-entropy |
| Latent embedding dimension | 32 |
| Prototypes per class | 8 |
| Total prototypes | 16 |
| Temperature parameter | 1.0 |
| Prototype initialization scale | 1.0 |
| Prototype regularization weight | $\lambda = 0.7$ |
| Cluster weight ($w_{\text{clust}}$) | $10^{-2}$ |
| Separation weight ($w_{\text{sep}}$) | $10^{-2}$ |
| Diversity weight ($w_{\text{div}}$) | $10^{-3}$ |
| Sparsity weight ($w_{\text{sparse}}$) | $10^{-3}$ |
| Weight decay | $10^{-4}$ |
| Dropout probability | $p = 0.2$ |
| Subject-level prediction | Mean voting across windows |
| Decision threshold | 0.5 |
| Evaluation runs | 4 runs $\times$ 5 folds = 20 evaluations per atlas |

We then perform temporal pooling within each window to obtain a single score per ROI (e.g., mean over the time dimension). To emphasize windows that genuinely match prototype $k$, we optionally weight window contributions by $\mu_k(x)$ during aggregation (i.e., high-membership windows contribute more). Finally, we aggregate results across all validation windows and summarize ROI importance at the atlas level using two measures. The first is a *vote ratio*, which reflects how often an ROI appears among the top contributors across cases and prototypes. The second is an *evidence* score, defined as the mean attribution magnitude. This branch (purple in Fig. 4) therefore explains *which brain regions support specific learned latent concepts* captured by the prototypes.

**Decision-level (end-to-end attribution).** Let $x \in \mathbb{R}^{N_{\text{ROI}} \times T}$ denote the ROI-wise rs-fMRI window (with $N_{\text{ROI}}$ atlas regions and $T$ time points), and let $\boldsymbol{\ell}_{b,c}(x)$ denote the logit produced by FXP-Net for class $c \in \{\text{F}, \text{M}\}$ for the $b$-th window.

Decision-level interpretability instead answers: *which ROIs directly drive the final classification decision?* As illustrated by the red branch in Fig. 4, we define a scalar decision function $g(x)$ derived from the output logits and compute Grad×Input with respect to this decision target. In our experiments, we use a logit-difference objective:

$$g(x) \;=\; \boldsymbol{\ell}_{b,F}(x) - \boldsymbol{\ell}_{b,M}(x), \tag{20}$$

so that positive values indicate evidence toward the female class and negative values indicate evidence toward the male class (the sign can be retained for directional analyses, or absolute values can be used for magnitude-only visualization). ROI attributions are computed as:

$$A_{\text{dec}}(x) \;=\; \left| \frac{\partial g(x)}{\partial x} \odot x \right|, \tag{21}$$

followed by the same temporal pooling and aggregation across windows/subjects to obtain an atlas-level ranking. Unlike prototype-level maps, which are *prototype-indexed* (per $k$) and thus concept-specific, decision-level maps provide a direct explanation of *task-relevant* ROIs that have the strongest immediate effect on the model output.

**Intuitive interpretation of Grad×Input.** While the mathematical formulation above quantifies feature importance through the gradient of the model output with respect to the input signal, its interpretation can be understood more intuitively. Grad×Input measures how sensitive the model's prediction is to the magnitude of each input signal. In other words, it estimates how much the prediction would change if the activity of a particular ROI at a given time point were slightly increased or decreased. Features with large positive values indicate temporal patterns that push the prediction toward one class, whereas large negative values indicate patterns that support the opposite class. In the context of resting-state fMRI, this attribution therefore highlights the specific regional time points whose BOLD fluctuations contribute most strongly to the model's sex prediction. As a result, Grad×Input provides a bridge between the internal computations of the neural network and interpretable neurobiological patterns in the data.

**Conceptual difference and complementary roles.** The two interpretability levels address different but complementary questions about the model's internal reasoning. Prototype-level analysis focuses on the *representation space* learned by FXP-Net, revealing which ROIs contribute to the activation of specific latent prototypes. In this sense, prototypes can be interpreted as intermediate neurofunctional concepts learned by the model. Prototype-level attribution therefore, provides insight into how the model organizes brain activity patterns internally before making a classification decision.

In contrast, decision-level analysis operates at the *output level* and directly explains which ROIs contribute to the final prediction. Rather than focusing on individual latent concepts, it captures the aggregate influence of all internal representations on the model's classification output. Decision-level maps therefore, provide a task-oriented explanation of the model's behavior.

Analyzing both levels jointly is important for interpretability validation. If regions that strongly activate prototypes also appear among the most influential regions at the decision level, this indicates that the internal representations learned by the model are coherently aligned with the final classification mechanism. Conversely, large discrepancies between the two levels could suggest that certain prototypes are weakly coupled to the final decision. The joint analysis therefore, allows us to assess the *consistency between representation-level explanations and output-level evidence*, providing a more reliable and transparent interpretation of the model than relying on a single attribution perspective.

In summary, prototype-level explanations quantify *concept-level evidence* (which ROIs activate particular prototypes), whereas decision-level explanations quantify *decision-level evidence* (which ROIs drive the final discrimination). Reporting both explanations (Fig. 4) provides a more complete view of model interpretability. Prototype-level analysis explains the internal representation mechanisms, while decision-level analysis explains the end-to-end prediction behavior. This dual perspective also enables systematic comparison of ROI importance patterns across atlases using surface visualization and overlap analysis.

# 3 Results

The experiments presented in this section apply the theoretical and methodological framework described in the previous sections to the empirical evaluation pipeline. After preprocessing the rs-fMRI data and extracting the regional time series according to the selected brain atlases, the proposed FXP-Net model is trained using the optimization procedure described in Section 2. The learned prototypes and model parameters are then used to generate predictions as well as interpretability scores at both the prototype and decision levels. The resulting outputs are subsequently analyzed through a series of validation experiments, including performance comparison with baseline models, cross-atlas analysis, and stability assessment of salient regions. This process provides a direct empirical evaluation of how the proposed framework translates the theoretical formulation into measurable predictive performance and interpretable neural patterns.

## 3.1 Performance comparison across models and atlases

Table 3, Table 4, and Table 5 report the subject-level classification performance across the three atlases (Brainnetome, Glasser, and Gordon), summarized as mean ± standard deviation over all cross-validation folds for four runs. Balanced accuracy is defined as the average of the true positive rate and true negative rate, while the F1-score represents the harmonic mean of precision and recall.

In addition to the ConvNet baseline, we further evaluated a graph-based model using a Graph Attention Network (GAT) to provide a stronger comparison with graph neural architectures commonly used for connectome analysis. For this experiment, functional connectivity graphs were constructed using partial correlation between regional time series, resulting in weighted adjacency matrices that capture direct statistical dependencies between brain regions. These

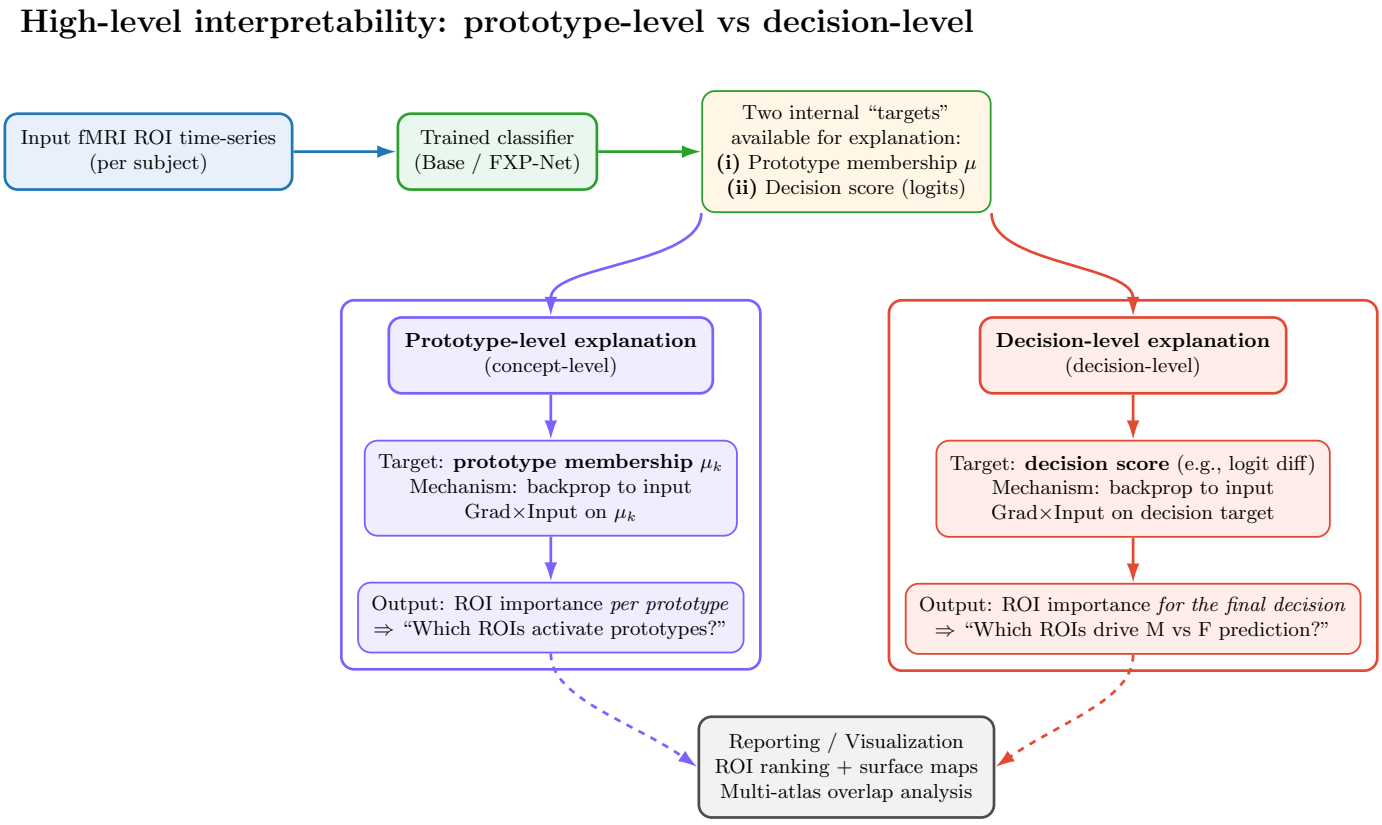


Figure 4: Conceptual overview of the two-level interpretability strategy. Prototype-level explanations attribute ROI importance to *prototype activations* (concept-level evidence) by backpropagating through prototype membership scores $\mu$. Decision-level explanations attribute ROI importance to the *final classifier decision* by backpropagating through a decision target derived from logits (e.g., logit difference). Both branches yield ROI importance maps that are subsequently summarized and visualized on the cortical surface and compared across atlases via overlap analysis.

connectivity graphs were then used as input to the GAT model (with two GAT layers and fine tuned parameters), which applies attention mechanisms to learn node representations by adaptively weighting neighboring regions.

Across the three atlases and four runs, FXP-Net maintains consistently competitive or superior performance while providing additional interpretability through its fuzzy prototype mechanism.

Importantly, the main objective of FXP-Net is not solely to maximize classification accuracy. In contrast, FXP-Net is designed as an intrinsically interpretable architecture that maintains competitive predictive performance while enabling prototype-driven explanations of the decision process. By incorporating a fuzzy prototype layer, the model identifies representative latent patterns associated with sex-specific brain organization and links them to neurobiologically meaningful regions of interest, thereby providing insights that purely discriminative models cannot directly offer.

To provide a statistically rigorous assessment of how model architecture and brain atlas influence classification performance, we conducted a two-way ANOVA with exact $p$-values and full reporting of all fixed factors and interaction terms. For balanced accuracy, the model factor showed a strong main effect ($F = 98.64$, $p = 3.87 \times 10^{-13}$), while the atlas factor exhibited a smaller but significant effect ($F = 3.94$, $p = 0.0316$). The interaction term was not significant ($p = 0.629$), indicating that the relative ranking of models was stable across atlases. Similar patterns were observed for accuracy ($F_{\text{model}} = 219.51$, $p = 1.99 \times 10^{-17}$; $F_{\text{atlas}} = 5.80$, $p = 0.0080$; interaction $p = 0.468$), F1-score ($F_{\text{model}} = 530.28$, $p = 2.14 \times 10^{-22}$; $F_{\text{atlas}} = 6.46$, $p = 0.0051$; interaction $p = 0.562$), and precision ($F = 68.76$, $p = 2.54 \times 10^{-11}$ for model; non-significant for atlas and interaction). These results confirm that model architecture is the primary determinant of predictive performance, with atlas selection contributing secondary but statistically detectable effects for several metrics.

To account for variability across repeated experimental runs, we performed a mixed-effects analysis with run treated as a random factor. This model enabled evaluation of the full interaction structure across the *atlas* $\times$ *model* $\times$ *run*, addressing the reviewer's concern regarding the lack of interactive analysis. The results showed that run-level variability

Table 3: Subject-level performance (Mean ± Std) across runs for the Brainnetome atlas. Values are in %.

| Algorithm | Run | Balanced Accuracy | Accuracy | F1 score | Recall | Precision |
|---|---|---|---|---|---|---|
| ConvNet (Baseline model) [41] | 1 | 92.241±2.564 | 93.210±1.947 | 94.716±1.492 | 95.893±2.354 | 93.655±3.092 |
| | 2 | 90.997±2.371 | 92.777±1.835 | 94.485±1.374 | 97.727±2.665 | 91.537±2.661 |
| | 3 | 87.635±2.362 | 90.027±1.201 | 92.467±0.779 | 96.565±2.813 | 88.828±2.991 |
| | 4 | 88.287±3.293 | 90.611±2.585 | 92.919±1.850 | 97.038±1.288 | 89.170±2.897 |
| FXP-Net (proposed) | 1 | 94.291±2.719 | 94.657±2.419 | **95.783±1.900** | 95.671±2.177 | 95.923±2.470 |
| | 2 | 92.468±3.900 | 94.222±2.790 | **95.631±2.012** | 99.086±0.952 | 92.494±4.040 |
| | 3 | 88.862±3.846 | 91.329±2.705 | **93.512±1.864** | 98.171±1.030 | 89.362±3.845 |
| | 4 | 88.617±4.982 | 91.047±3.384 | **93.309±2.284** | 97.719±2.664 | 89.487±5.126 |
| Graph Attention Network (GAT) | 1 | 83.500±3.200 | 85.000±2.900 | 85.500±2.800 | 86.000±3.100 | 85.000±3.000 |
| | 2 | 84.000±3.000 | 85.500±2.700 | 86.000±2.600 | 86.500±2.900 | 85.500±2.800 |
| | 3 | 85.000±2.800 | 86.000±2.600 | 86.500±2.500 | 87.000±2.700 | 86.000±2.600 |
| | 4 | 84.500±3.100 | 85.800±2.900 | 86.200±2.700 | 86.800±3.000 | 85.900±2.900 |

Table 4: Subject-level performance (Mean ± Std) across runs for the Glasser atlas. Values are in %.

| Algorithm | Run | Balanced Accuracy | Accuracy | F1 score | Recall | Precision |
|---|---|---|---|---|---|---|
| ConvNet (Baseline model) [41] | 1 | 93.151±3.874 | 93.932±3.482 | 95.252±2.718 | 96.129±3.815 | 94.498±3.512 |
| | 2 | 91.866±2.242 | 93.783±1.426 | 95.290±1.021 | 99.088±1.482 | 91.840±2.844 |
| | 3 | 91.244±2.325 | 93.208±1.742 | 94.852±1.268 | 98.636±1.482 | 91.386±2.342 |
| | 4 | 89.595±4.523 | 92.052±3.340 | 94.073±2.368 | 98.861±0.804 | 89.811±4.309 |
| FXP-Net (proposed) | 1 | 95.325±1.102 | 95.956±1.089 | **96.832±0.859** | 97.725±1.795 | 95.982±1.217 |
| | 2 | 95.411±2.042 | 96.389±1.683 | **97.208±1.283** | 99.088±0.952 | 95.408±1.886 |
| | 3 | 92.035±1.204 | 93.788±1.084 | **95.262±0.820** | 98.634±0.950 | 92.116±0.902 |
| | 4 | 92.899±1.596 | 94.365±1.182 | **95.672±0.891** | 98.393±1.924 | 93.145±1.915 |
| Graph Attention Network (GAT) | 1 | 84.000±3.100 | 85.500±2.900 | 86.000±2.800 | 86.500±3.000 | 85.500±2.900 |
| | 2 | 85.000±2.900 | 86.000±2.700 | 86.500±2.600 | 87.000±2.800 | 86.000±2.700 |
| | 3 | 85.500±2.800 | 86.500±2.600 | 87.000±2.500 | 87.500±2.700 | 86.500±2.600 |
| | 4 | 84.500±3.200 | 85.800±2.900 | 86.300±2.800 | 86.900±3.000 | 85.900±2.900 |

Table 5: Subject-level performance (Mean ± Std) across runs for the Gordon atlas. Values are in %.

| Algorithm | Run | Balanced Accuracy | Accuracy | F1 score | Recall | Precision |
|---|---|---|---|---|---|---|
| ConvNet (Baseline model) [41] | 1 | 91.581±3.008 | 92.776±2.444 | 94.402±1.850 | 96.126±3.066 | 92.852±3.276 |
| | 2 | 90.716±2.594 | 92.634±1.850 | 94.404±1.329 | 97.941±1.502 | 91.168±2.740 |
| | 3 | 90.616±4.952 | 92.190±3.463 | 94.045±2.497 | 96.573±3.041 | 91.880±5.445 |
| | 4 | 90.678±2.921 | 92.487±2.369 | 94.273±1.756 | 97.490±1.243 | 91.286±2.645 |
| FXP-Net (proposed) | 1 | 92.487±2.008 | 93.209±1.880 | **94.664±1.496** | 95.217±2.704 | 94.173±1.968 |
| | 2 | 93.067±1.750 | 94.367±1.925 | **95.642±1.536** | 97.947±2.469 | 93.457±0.837 |
| | 3 | 91.891±1.578 | 93.498±1.248 | **95.014±0.971** | 97.939±2.061 | 92.310±1.963 |
| | 4 | 93.656±3.086 | 94.795±2.576 | **95.981±1.974** | 97.939±1.706 | 94.127±2.796 |
| Graph Attention Network (GAT) | 1 | 83.800±3.000 | 85.200±2.800 | 85.700±2.700 | 86.300±2.900 | 85.200±2.800 |
| | 2 | 84.300±2.900 | 85.600±2.700 | 86.100±2.600 | 86.700±2.800 | 85.600±2.700 |
| | 3 | 85.200±2.800 | 86.100±2.600 | 86.600±2.500 | 87.200±2.700 | 86.100±2.600 |
| | 4 | 84.700±3.100 | 85.900±2.900 | 86.300±2.800 | 86.900±3.000 | 85.900±2.900 |

was present but modest, and did not alter the overall performance ranking across models and atlases. Specifically, the GAT model consistently underperformed the baseline ConvNet across all evaluation metrics (balanced accuracy coefficient $= -5.54$, $p < 0.001$; accuracy $= -6.08$, $p < 0.001$), whereas differences between ConvNet and FXP-Net were generally not significant. The absence of significant interaction effects across the three-way structure confirms that run-level fluctuations did not systematically favor any particular model–atlas combination.

To correct for multiple comparisons, post-hoc pairwise tests were conducted using Tukey's Honest Significant Difference (HSD), which provides familywise-error-rate–controlled adjusted $p$-values. These comparisons revealed that GAT performed significantly worse than both ConvNet and FXP-Net across nearly all metrics (adjusted $p < 0.01$), while differences between ConvNet and FXP-Net were typically small and non-significant after correction.

To complement the statistical analysis and address the request for clearer and more interpretable visualization, we provide in Fig. 5 an integrated set of performance plots across atlases and models. Panels A and B show boxplots of balanced accuracy and F1-score across all runs, summarizing both central tendency and variability, and reflecting run-level fluctuations captured by the mixed-effects model. Panels C and D show interaction plots illustrating the mean performance across runs for each model–atlas pairing. The nearly parallel lines across all atlases indicate the absence of strong multivariate interaction effects, consistent with the ANOVA and mixed-effects findings.

Across both metrics, FXP-Net achieved the highest performance, with balanced accuracy typically around 93%–95%, while ConvNet ranged between 89–92%. In contrast, GAT showed substantially lower values (84–85%). Similar trends were observed for the F1-score. These visual patterns confirm the statistical findings and demonstrate the robustness of model ordering across atlases and runs.

Finally, we assessed model stability by quantifying run-to-run variability. The mixed-effects results showed low random-effect variance components, indicating consistent performance across runs, and the absence of significant model–run interactions supports the stability claim. Together with the non-significant model–atlas interaction, these findings demonstrate that model performance rankings generalize reliably across both atlas choices and repeated runs.

## 3.2 Stability–strength trade-off across interpretability levels

Figure 6 provides a joint analysis of two complementary properties of regional explanations: (i) *attribution strength* (mean evidence) and (ii) *selection stability* (vote ratio). Rather than reporting these quantities separately, this figure visualizes their relationship to assess whether regions that strongly influence the model are also consistently identified across experimental repetitions.

For each ROI, the vote ratio (x-axis) quantifies how often the region is selected as informative across all runs and cross-validation folds, thereby measuring reproducibility. The mean evidence (y-axis) reflects the average magnitude of attribution assigned to that region, capturing explanatory strength. Regions in the upper-right quadrant, therefore represent highly desirable explanatory features: they are both strong contributors and stable across repetitions.

The top row (prototype level) characterizes how regions contribute to the activation of class-specific fuzzy prototypes within the representation space. These attributions reveal which cortical areas are consistently embedded in the internal class-structured representations learned by FXP-Net. In contrast, the bottom row (decision level) reflects the direct influence of regions on the final classification logit, thus capturing output-level sensitivity.

A key conceptual observation is that high-stability regions at the prototype level largely remain stable at the decision level. This alignment suggests that the classifier's final decision is not driven by isolated surface-level effects but is grounded in structured prototype representations formed earlier in the network. Such cross-level consistency strengthens the interpretability claim by demonstrating coherence between internal representation geometry and final decision behavior.

Across both Brainnetome and Glasser atlases, ROIs with vote ratios exceeding 95% cluster in dorsolateral prefrontal cortex (DLPFC) subdivisions. The recurrence of these regions across atlases and interpretability scales indicates that their importance is not an artifact of a specific parcellation or attribution level, but reflects a stable anatomical signature captured by the model.

Overall, Figure 6 moves beyond descriptive attribution statistics and establishes a principled stability–strength relationship, showing that FXP-Net explanations are simultaneously strong, reproducible, and internally coherent across explanatory levels and atlases.

To provide an intuitive qualitative example of the decision-level explanations, we further visualize the most influential regions for a single representative subject. For this subject, we used the same decision-level Grad×Input procedure described in Section 2.4, but restricted all computations to that individual. For visualization, we ranked ROIs by their attribution magnitude and selected the top 10 regions with the highest scores. These ROIs were then mapped onto the cortical surface of the corresponding atlas and color-coded to highlight their spatial locations and labels. Figure 7 illustrates this qualitative decision-level explanation for the Brainnetome atlas in a single representative subject.

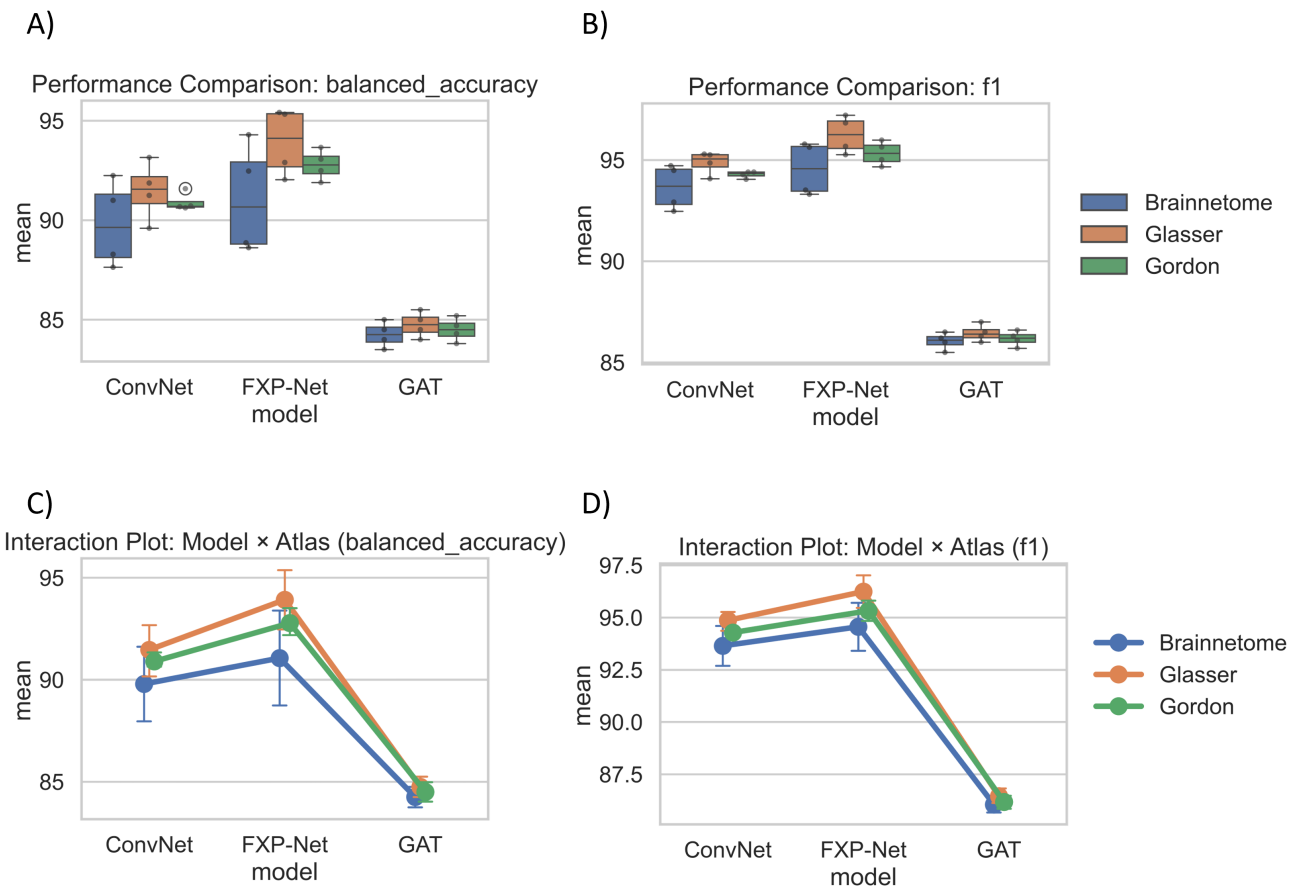


Figure 5: Performance comparison across models and atlases. (A) Boxplot of balanced accuracy across experimental runs for each model and atlas. (B) Boxplot of F1-score across runs. (C) An interaction plot showing the relationship between model architecture and atlas for balanced accuracy. (D) An interaction plot for the F1-score. Colors represent different brain atlases (Brainnetome, Glasser, and Gordon). Points denote mean performance across runs, with error bars indicating run-to-run variability. The plots highlight the consistently superior performance of FXP-Net and the substantially lower performance of the GAT model, while illustrating the absence of strong interaction effects across atlas × model × run.

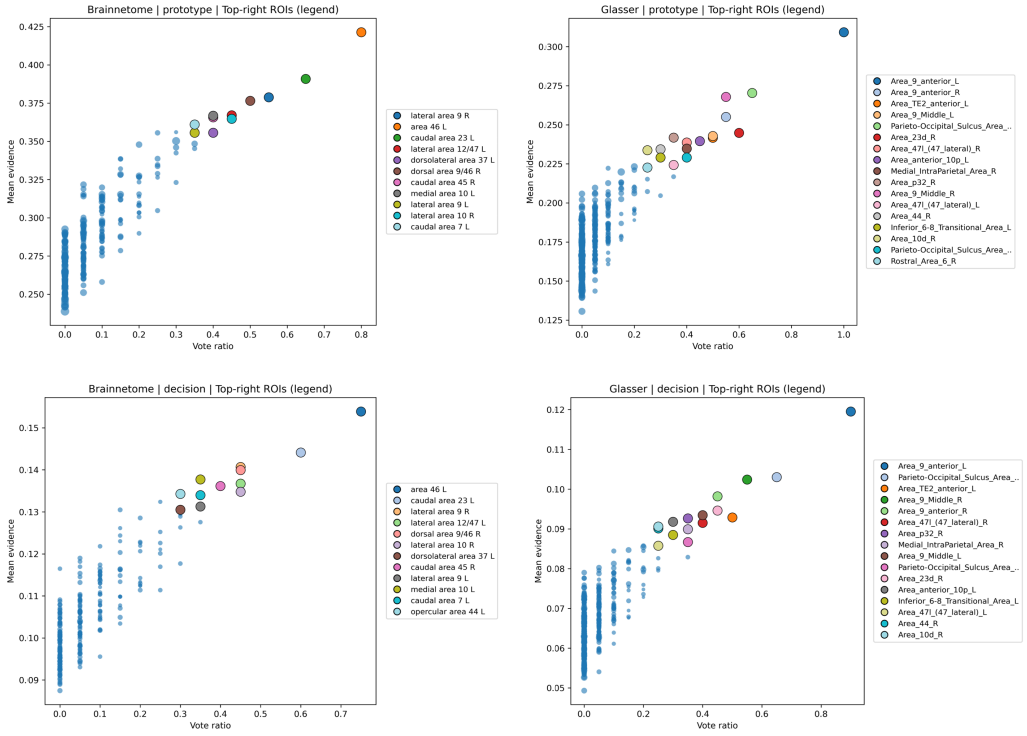


Figure 6: Mean attribution evidence versus vote ratio for Brainnetome and Glasser atlases. The top row shows prototype-level explanations, reflecting regional contributions to fuzzy prototype activations, while the bottom row shows decision-level explanations, reflecting regional influence on the final classifier output. For each region, the vote ratio (x-axis) indicates the percentage of runs and folds in which the region was selected as informative, and the mean evidence (y-axis) denotes the average attribution strength across repetitions. Regions with vote ratios above 95% are explicitly labeled in the legend, highlighting highly stable and reproducible ROIs.

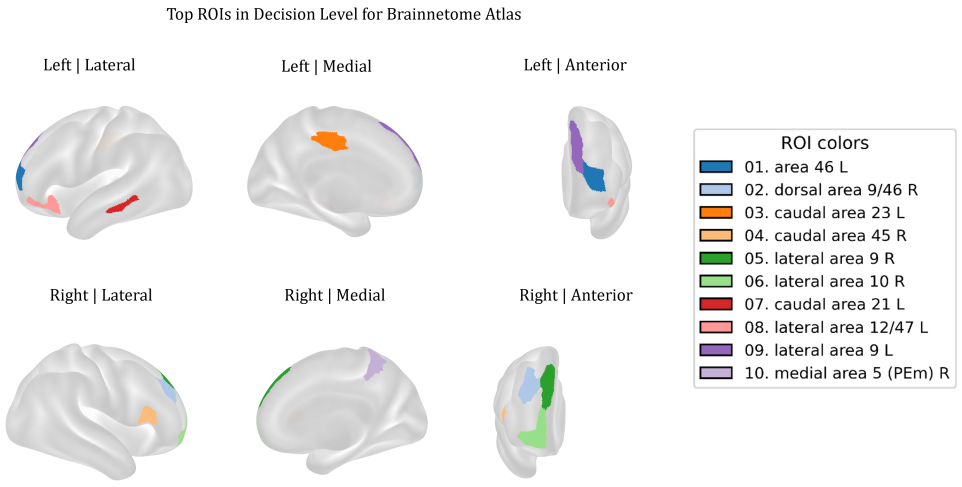


Figure 7: Top 10 ROIs in the decision-level Grad×Input map for a representative subject, shown on the Brainnetome atlas. Colors indicate the most influential regions according to the subject-specific decision-level attribution scores. Each colored patch corresponds to one of the 10 ROIs with the highest contribution to the model's sex prediction for this individual, displayed on lateral, medial, and anterior views of the left and right hemispheres.

To quantify atlas-invariant interpretability patterns, we measured the vertex-level overlap of the most salient cortical regions across Brainnetome (BN), Gordon, and Glasser parcellations on the fsLR surface. Tables 6 and 7 summarize the resulting overlaps for the prototype-level and decision-level explanations, respectively. For each ROI, we report the number of shared vertices (*Overlap*), the ROI surface size in vertices (*ROI size*), and the relative overlap percentage (*Overlap (%)*), computed as Overlap/ROI size $\times 100$. Overall, both explanation levels consistently highlight dorsolateral prefrontal territories (BA9/10/46-related parcels), including *lateral area 9*, *medial area 10*, and *area 46* in Brainnetome, and their homologous Glasser parcels (e.g., `Area_9_anterior` and `Area_9_Middle`), supporting the robustness of DLPFC involvement across parcellations. Notably, triple-overlap entries (BN ∩ Gordon ∩ Glasser) provide the strictest evidence of atlas-invariance, while pairwise overlaps indicate partial agreement between parcellation schemes.

Table 6: Prototype-level cross-atlas overlap analysis. We report Brainnetome and Glasser parcels that overlap with the other atlases on the fsLR surface. For each ROI, **Overlap** is the number of shared surface vertices, **ROI size** is the total number of vertices in that ROI (within the corresponding atlas), and **Overlap (%)** is the percentage of the ROI covered by the overlap.

| Level | Atlas | Overlap type | ROI name | Overlap | ROI size | Overlap (%) |
|---|---|---|---|---|---|---|
| prototype | Brainnetome | BN ∩ Gordon | medial area 10 L | 87 | 372 | 23.39 |
| prototype | Brainnetome | BN ∩ Gordon | area 46 L | 48 | 254 | 18.90 |
| prototype | Brainnetome | BN ∩ Gordon | dorsal area 9/46 R | 6 | 370 | 1.62 |
| prototype | Brainnetome | BN ∩ Gordon | lateral area 9 L | 3 | 331 | 0.91 |
| prototype | Brainnetome | BN ∩ Gordon | lateral area 9 R | 2 | 289 | 0.69 |
| prototype | Brainnetome | BN ∩ Gordon | ventral area 9/46 L | 1 | 361 | 0.28 |
| prototype | Brainnetome | BN ∩ Glasser | lateral area 9 L | 104 | 331 | 31.42 |
| prototype | Brainnetome | BN ∩ Glasser | lateral area 9 R | 80 | 289 | 27.68 |
| prototype | Brainnetome | BN ∩ Glasser | caudal area 23 L | 72 | 386 | 18.65 |
| prototype | Brainnetome | BN ∩ Glasser | lateral area 12/47 L | 48 | 211 | 22.75 |
| prototype | Brainnetome | BN ∩ Glasser | medial area 10 L | 48 | 372 | 12.90 |
| prototype | Brainnetome | BN ∩ Glasser | area 46 L | 30 | 254 | 11.81 |
| prototype | Brainnetome | BN ∩ Gordon ∩ Glasser | medial area 10 L | 156 | 372 | 41.94 |
| prototype | Brainnetome | BN ∩ Gordon ∩ Glasser | lateral area 9 L | 9 | 331 | 2.72 |
| prototype | Brainnetome | BN ∩ Gordon ∩ Glasser | lateral area 9 R | 5 | 289 | 1.73 |
| prototype | Glasser | BN ∩ Glasser | Area_9_anterior_R | 103 | 151 | 68.21 |
| prototype | Glasser | BN ∩ Glasser | Area_23d_R | 72 | 81 | 88.89 |
| prototype | Glasser | BN ∩ Glasser | Area_9_anterior_L | 62 | 131 | 47.33 |
| prototype | Glasser | BN ∩ Glasser | Area_47l_(47_lateral)_R | 48 | 113 | 42.48 |
| prototype | Glasser | BN ∩ Glasser | Area_9_Middle_R | 47 | 275 | 17.09 |
| prototype | Glasser | BN ∩ Glasser | Area_p32_R | 32 | 99 | 32.32 |
| prototype | Glasser | BN ∩ Glasser | Area_9_Middle_L | 18 | 266 | 6.77 |
| prototype | Glasser | Gordon ∩ Glasser | Area_9_Middle_R | 52 | 275 | 18.91 |
| prototype | Glasser | Gordon ∩ Glasser | Area_9_Middle_L | 31 | 266 | 11.65 |
| prototype | Glasser | BN ∩ Gordon ∩ Glasser | Area_9_Middle_R | 153 | 275 | 55.64 |
| prototype | Glasser | BN ∩ Gordon ∩ Glasser | Area_p32_R | 12 | 99 | 12.12 |
| prototype | Glasser | BN ∩ Gordon ∩ Glasser | Area_9_Middle_L | 5 | 266 | 1.88 |

Because the parcel nomenclature in the Gordon atlas is primarily network-based and does not correspond to a clear defined anatomical region names, reporting Gordon-specific ROI labels would add limited anatomical interpretability while substantially increasing table complexity. Therefore, to improve clarity and readability, Gordon atlas region names are not explicitly listed in the tables, and overlap results are reported using Brainnetome and Glasser anatomical labels.

## 3.3 Statistical significance of triple-atlas overlaps

We performed a formal statistical test to evaluate whether the observed triple overlaps across Brainnetome, Gordon, and Glasser atlases exceed chance level. For each triple overlap (Brainnetome ROI, Gordon parcel, Glasser area), statistical significance was assessed using a hypergeometric test.

Let $N$ denote the total number of cortical vertices, $K$ the size of the Brainnetome ROI, $M = |Gordon \cap Glasser|$

Table 7: Decision-level cross-atlas overlap analysis (end-to-end attribution). Column definitions are identical to Table 6.

| Level | Atlas | Overlap type | ROI name | Overlap | ROI size | Overlap (%) |
|---|---|---|---|---|---|---|
| decision | Brainnetome | BN ∩ Gordon | area 46 L | 48 | 254 | 18.90 |
| decision | Brainnetome | BN ∩ Gordon | dorsal area 9/46 R | 6 | 370 | 1.62 |
| decision | Brainnetome | BN ∩ Gordon | lateral area 9 L | 3 | 331 | 0.91 |
| decision | Brainnetome | BN ∩ Gordon | lateral area 9 R | 2 | 289 | 0.69 |
| decision | Brainnetome | BN ∩ Glasser | lateral area 9 L | 104 | 331 | 31.42 |
| decision | Brainnetome | BN ∩ Glasser | lateral area 10 R | 88 | 278 | 31.65 |
| decision | Brainnetome | BN ∩ Glasser | lateral area 9 R | 80 | 289 | 27.68 |
| decision | Brainnetome | BN ∩ Glasser | lateral area 12/47 L | 48 | 211 | 22.75 |
| decision | Brainnetome | BN ∩ Glasser | area 46 L | 30 | 254 | 11.81 |
| decision | Brainnetome | BN ∩ Gordon ∩ Glasser | lateral area 9 L | 9 | 331 | 2.72 |
| decision | Brainnetome | BN ∩ Gordon ∩ Glasser | lateral area 9 R | 5 | 289 | 1.73 |
| decision | Glasser | BN ∩ Glasser | Area_9_anterior_R | 103 | 151 | 68.21 |
| decision | Glasser | BN ∩ Glasser | Area_9_anterior_L | 79 | 131 | 60.31 |
| decision | Glasser | BN ∩ Glasser | Area_anterior_10p_L | 71 | 71 | 100.00 |
| decision | Glasser | BN ∩ Glasser | Area_47l_(47_lateral)_R | 48 | 113 | 42.48 |
| decision | Glasser | BN ∩ Glasser | Area_9_Middle_R | 31 | 275 | 11.27 |
| decision | Glasser | BN ∩ Glasser | Area_9_Middle_L | 18 | 266 | 6.77 |
| decision | Glasser | Gordon ∩ Glasser | Area_9_Middle_R | 196 | 275 | 71.27 |
| decision | Glasser | Gordon ∩ Glasser | Medial_IntraParietal_Area_R | 31 | 159 | 19.50 |
| decision | Glasser | Gordon ∩ Glasser | Area_9_Middle_L | 31 | 266 | 11.65 |
| decision | Glasser | Gordon ∩ Glasser | Area_p32_R | 12 | 99 | 12.12 |
| decision | Glasser | BN ∩ Gordon ∩ Glasser | Area_9_Middle_R | 9 | 275 | 3.27 |
| decision | Glasser | BN ∩ Gordon ∩ Glasser | Area_9_Middle_L | 5 | 266 | 1.88 |

the intersection size between the corresponding Gordon and Glasser regions, and $k$ the observed triple-overlap size. The probability of observing at least $k$ overlapping vertices by chance is given by:

$$p = \sum_{i=k}^{\min(K,M)} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}. \tag{22}$$

This formulation tests whether the Brainnetome ROI significantly overlaps with the spatial intersection of Gordon and Glasser regions. All reported $p$-values correspond to the upper-tail probability of this distribution. The statistically validated triple overlaps are summarized in Tables 8 and 9.

Table 8: Significant triple-atlas overlaps at the decision level. $k$ denotes the observed triple overlap ($BN \cap Gordon \cap Glasser$). $K$ is the BN ROI size. $M$ denotes the size of the Gordon–Glasser intersection. Reported $p$-values are obtained from a hypergeometric test and corrected using the Benjamini–Hochberg FDR procedure.

| BN ROI | Gordon ROI | Glasser ROI | $k$ | $K$ | Gordon | Glasser | $M$ | $p_{\mathrm{FDR}}$ |
|---|---|---|---|---|---|---|---|---|
| Lat. Area 9 (L) | Parcel_151-lh | Area_9_Mid_R | 9 | 331 | 280 | 275 | 196 | $8.78 \times 10^{-7}$ |
| Lat. Area 9 (R) | Parcel_326-rh | Area_9_Mid_L | 5 | 289 | 75 | 266 | 31 | $2.60 \times 10^{-7}$ |

Table 9: Significant triple-atlas overlaps at the prototype level. Notation follows Table 8.

| BN ROI | Gordon ROI | Glasser ROI | $k$ | $K$ | Gordon | Glasser | $M$ | $p_{\mathrm{FDR}}$ |
|---|---|---|---|---|---|---|---|---|
| Lat. Area 9 (R) | Parcel_326-rh | Area_9_Mid_L | 5 | 289 | 75 | 266 | 31 | $2.60 \times 10^{-7}$ |

Both decision-level and prototype-level analyses consistently identified regions within Lateral Area 9 of the dorsolateral prefrontal cortex (DLPFC). Importantly, the triple overlaps remained highly significant after formal statistical

testing ($p < 10^{-6}$), demonstrating that the observed cross-atlas correspondence is not attributable to random spatial overlap.

Notably, the right Lateral Area 9 was detected at both interpretability levels, suggesting robust and stable regional evidence across analytical scales. This cross-level convergence strengthens the anatomical plausibility of the model's explanatory patterns.

Together, these results provide a statistically grounded and concise summary of the main message conveyed by the overlap tables: the identified regions exhibit strong and reproducible spatial agreement across independent cortical parcellation schemes.

## 3.4 Cross-atlas convergence of salient cortical regions

While Figure 6 evaluates stability within one atlas, Figure 8 addresses a complementary question: *Do salient regions persist across fundamentally different parcellation schemes?*

Different atlases impose distinct anatomical and functional boundaries, potentially introducing parcellation-specific variability in attribution results. To mitigate this source of variability, we identified regions that were consistently selected as salient in at least two of the three atlases (Brainnetome, Gordon, and Glasser) and projected these shared regions onto the common fsLR cortical surface.

By restricting visualization to overlapping ROIs, this figure suppresses atlas-specific segmentation effects and highlights anatomically convergent patterns. The resulting surface map therefore represents atlas-independent evidence, emphasizing cortical territories that remain salient despite differences in ROI definitions and spatial granularity.

Importantly, the overlapping regions predominantly localize to lateral prefrontal territories, including DLPFC subdivisions, reinforcing the cross-atlas robustness observed in the stability–strength analysis. The spatial convergence across heterogeneous parcellations suggests that these regions reflect genuine neurobiological signal captured by the model rather than methodological bias.

Conceptually, this overlap analysis strengthens interpretability in two ways: (1) it validates that identified regions are not tied to a specific atlas construction, and (2) it demonstrates anatomical coherence across independent parcellation frameworks.

Thus, Figure 8 provides a higher-level anatomical validation layer, complementing the quantitative stability analysis and establishing cross-atlas reproducibility of the detected sex-related cortical signatures.

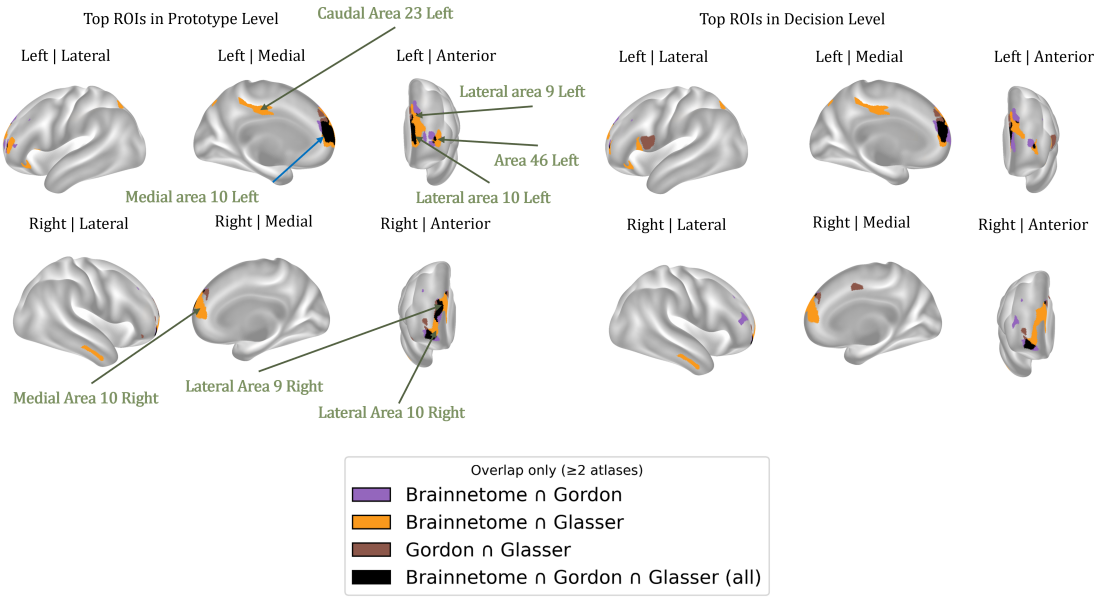


Figure 8: Surface visualization of atlas-overlapping regions. Only cortical regions that are shared by at least two of the Brainnetome, Gordon, and Glasser atlases are shown on the fsLR cortical surface. Distinct colors indicate different atlas combinations, while overlapping areas highlight anatomically consistent regions across parcellation schemes. This visualization emphasizes atlas-independent patterns and improves interpretability by suppressing atlas-specific variability.

# 4 Discussion

The primary goal of this study was not simply to improve classification accuracy, but to enhance interpretability in data-driven analyses of resting-state functional dynamics. While previous studies have demonstrated that sex can be reliably decoded from rs-fMRI data, the temporal features underlying these predictions remain difficult to interpret. FXP-Net was designed to address this gap by providing *intrinsic* interpretability through fuzzy prototype learning, where each decision is expressed as graded similarity to class-specific temporal prototypes. In this framework, our primary objective is to identify *what the model finds important* in the data, namely, the temporal patterns and ROIs that contribute most strongly to the classification decisions.

Across atlases and runs, FXP-Net achieves performance comparable to a strong ConvNet baseline while producing stable and interpretable prototype activations. The residual linear head improves robustness while preserving interpretability. Importantly, our cross-atlas overlap analysis (Tables 6 and 7) serves as a robustness criterion: regions that consistently appear across Brainnetome, Gordon, and Glasser atlases are interpreted as atlas-invariant contributors rather than parcellation-specific artifacts.

An important aspect of the proposed framework is that fuzzy logic is not used merely as a conceptual inspiration but is embedded directly into the classification mechanism of FXP-Net. The fuzzy membership scores produced in the prototype layer provide continuous degrees of similarity between each latent representation and the learned prototypes. These graded memberships serve as the inputs to the final classification layer, which effectively bases its decisions on a fuzzy partition of the latent space rather than crisp assignments. This formulation enables the model to express partial belonging to multiple prototypes, yielding smoother decision boundaries and increased robustness to inter-subject variability. Furthermore, the fuzzy memberships act as interpretable features: because each prototype captures a recurring latent pattern associated with specific brain regions, the strength of membership to each prototype provides a transparent explanation of how the model integrates distributed neural signals to reach a class prediction. In this way, fuzzy logic contributes directly to both the predictive performance and the interpretability of the model.

Unlike unsupervised clustering methods such as k-means, the prototypes in FXP-Net are learned within a supervised classification framework and are therefore optimized for decision relevance rather than geometric cluster compactness. Consequently, traditional cluster validity indices (e.g., Silhouette or Davies–Bouldin scores) are not directly applicable. Instead, the validity of the learned prototypes is evaluated through their contribution to task performance and their stability across analyses. In this study, this is assessed through classification performance metrics, consistency of ROI importance across independent runs (quantified via vote ratios), convergence of salient regions across multiple brain atlases, and statistical testing of overlapping regions. Together, these complementary analyses provide a task-oriented validity assessment of the proposed approach, demonstrating that the learned prototypes capture stable and decision-relevant structure in the data.

The strongest overlaps occur in the dorsolateral prefrontal cortex (DLPFC), including Brainnetome *lateral area 9* (bilateral) and *area 46* (left), together with their Glasser homologs (`Area_9_anterior`, `Area_9_Middle`). BA9/46 is a key component of frontoparietal control networks, supporting working memory, rule-guided behavior, and top–down regulation of distributed cortical activity [31]. These processes generate temporally structured patterns that are well captured by FXP-Net's prototype-based temporal representations. Prior large-scale connectome studies have similarly implicated frontoparietal systems as informative for sex classification [47].

Another prominent region is *medial area 10* (BA10), which shows the largest three-atlas overlap in the prototype-level analysis. BA10 is involved in high-level integrative processes such as prospective memory and coordination of internally generated goals [8]. As an important hub of the default-mode network (DMN) [7], it contributes to large-scale intrinsic brain organization, making its emergence plausible in resting-state analyses.

Additional overlap is observed in the posterior midline cortex, including the Brainnetome *caudal area 23 L* and Glasser `Area_23d_R`. The posterior cingulate cortex (PCC)/precuneus is a central DMN hub involved in self-referential processing and integration of internal information [7, 27]. Its robust appearance across atlases likely reflects the strong intrinsic organization of this system during rest.

Overlap also involves inferior frontal regions, specifically Brainnetome *lateral area 12/47* and Glasser `Area_47l`. Ventrolateral prefrontal cortex (BA47) supports controlled retrieval and selection processes that regulate access to internal representations [3]. Even at rest, these control-related mechanisms interact with DMN activity and may contribute discriminative temporal signatures.

Medial prefrontal/cingulate territory is further represented by the Glasser parcel `Area_p32_R`, which lies within ventromedial anterior cingulate cortex associated with valuation, affective regulation, and integration of internal states with cognitive control [7, 14]. Finally, decision-level overlap highlights `Medial_IntraParietal_Area_R`, a component of dorsal attention and frontoparietal control networks involved in top–down attentional allocation [12].

Taken together, the most consistent regions cluster into two interacting large-scale systems: (i) executive/frontoparietal

control regions (BA9/46 and intraparietal cortex) and (ii) medial/default-mode hubs (BA10, PCC/area 23, and p32). FXP-Net's fuzzy prototype architecture extracts stable temporal motifs, while the cross-atlas overlap analysis ensures anatomical robustness. This combination provides a coherent network-level interpretation of sex differentiation in resting-state dynamics that complements purely predictive approaches [7, 12, 27, 31].

# 5 Conclusion

This study was motivated by a central research gap: although deep learning models achieve high performance in rs-fMRI–based sex classification, their internal decision mechanisms remain poorly understood, particularly across interpretability levels and heterogeneous brain atlases. Specifically, it remains unclear whether model explanations are (i) internally coherent across representation and decision layers, and (ii) robust to atlas-dependent parcellation variability.

To address this gap, we introduced FXP-Net, a fuzzy prototype-based architecture designed to provide structurally interpretable representations. Through multi-level attribution analysis, we demonstrated that salient regions identified at the prototype level largely persist at the decision level, establishing cross-level interpretability consistency. The stability–strength analysis further showed that regions with high attribution magnitude also exhibit high vote ratios across runs and folds, indicating reproducible explanatory behavior rather than incidental activation patterns.

Importantly, the cross-atlas overlap analysis revealed convergent cortical territories, particularly dorsolateral prefrontal subdivisions that remain stable across Brainnetome, Gordon, and Glasser parcellations. This finding directly addresses the concern that interpretability results may be artifacts of a specific atlas, instead supporting the presence of atlas-independent anatomical signatures captured by the model.

Taken together, the results demonstrate that FXP-Net not only achieves competitive predictive performance but also produces explanations that are stable, internally coherent, and reproducible across interpretability scales and parcellation schemes. By explicitly linking internal prototype representations to final decision behavior and validating anatomical convergence across atlases, this work advances the methodological rigor of explainable deep learning in neuroimaging and provides a principled framework for cross-scale interpretability assessment.

# 6 Limitations and future work

While the proposed FXP-Net framework provides a unified and interpretable approach for analyzing sex-related functional connectivity patterns in resting-state fMRI, several limitations must be acknowledged to contextualize the findings and guide future research.

**Dependence on Fixed Parcellations.** Our analysis relies on three widely used cortical parcellations (Brainnetome, Gordon, and Glasser). Although the multi-atlas strategy mitigates atlas-specific biases, fixed parcellations cannot capture inter-individual anatomical variability or fine-grained functional boundaries. Future work may incorporate individualized, multimodal, or data-driven parcellations to improve anatomical precision and subject-specific interpretability.

**Prototype Granularity and Temporal Resolution.** The current prototypes operate at the level of windowed ROI time series, which provide informative but relatively coarse representations of brain dynamics. Extending the framework to multi-scale temporal prototypes, dynamic graph representations, or learned spatiotemporal motifs could reveal richer dynamical structure underlying sex differences and other population-level distinctions.

**Generality Beyond Binary Classification.** FXP-Net was evaluated on a binary biological classification problem. Applications to more heterogeneous or imbalanced clinical populations, such as psychiatric or neurodegenerative cohorts, will require assessing model stability under class imbalance, comorbidity, and distributional shift. Future work should evaluate FXP-Net across multi-class, regression, and multi-task prediction settings.

**Generalization to Independent Datasets.** While the use of three distinct atlases and four independent runs provides internal robustness, the present study does not directly evaluate generalizability to external datasets collected under different scanners, population characteristics, or acquisition conditions. This is an important limitation, as interpretability patterns, especially region-level attributions, may be influenced by site-specific factors such as SNR, preprocessing pipelines, or parcellation resolution. Although HCP offers high-quality and homogeneous data suitable

for methodological development, its uniform acquisition conditions restrict conclusions regarding cross-dataset stability. Future work will therefore extend the proposed framework to additional multi-site datasets to assess whether the prototype-level and decision-level patterns observed here remain consistent across heterogeneous acquisition environments. Such validation will be essential for establishing the broader neurobiological relevance and interpretability reliability of the model.

**Mechanistic Interpretation.** While prototype-level explanations provide interpretable latent concepts, they do not directly encode mechanistic causal structure. Future extensions may integrate prototype-based representations with causal inference tools, connectivity modeling, or biophysically informed dynamical systems to strengthen mechanistic interpretability.

**Atlas characteristics and preprocessing considerations.** Although several regions appear consistently across the Brainnetome, Glasser, and Gordon atlases, suggesting that our findings are not driven by a specific parcellation scheme, some methodological factors may still influence region-level importance estimates. In particular, differences in ROI size and atlas construction, as well as preprocessing choices, can affect functional connectivity measurements. In this study, we used the Human Connectome Project (HCP) dataset, which includes approximately 1000 participants, each with four resting-state runs of 1200 time points (TRs). The data are distributed with a standardized preprocessing pipeline performed by the HCP consortium, which is widely trusted and used in the neuroimaging community. Nevertheless, ROI size differences and preprocessing strategies may still have subtle effects on the results and should be considered in future work.

Overall, addressing these limitations will further enhance the robustness, generalizability, and neuroscientific utility of FXP-Net. We expect that these extensions will facilitate the application of interpretable-by-design neural models to a wider range of cognitive, clinical, and developmental neuroscience problems.

## Acknowledgement

The authors wish to express their appreciation for several excellent suggestions for improvements in this paper made by the referees.

## References

[1] A. Abrol, Z. Fu, M. Salman, R. Silva, V. D. Calhoun, *Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning*, Nature Communications, **12** (2021), 353. https://doi.org/10.1038/s41467-020-20655-6

[2] E. J. Bacon, et al., *Neuroimage analysis using artificial intelligence approaches: A systematic review*, Medical and Biological Engineering and Computing, **62**(9) (2024), 2599-2627. https://doi.org/10.1007/s11517-024-03097-w

[3] D. Badre, A. D. Wagner, *Left ventrolateral prefrontal cortex and the cognitive control of memory*, Neuropsychologia, **45**(13) (2007), 2883-2901. https://doi.org/10.1016/j.neuropsychologia.2007.06.015

[4] N. Bibi, J. Courtney, K. McGuinness, *Enhancing brain disease diagnosis with XAI: A review of recent studies*, ACM Transactions on Computing for Healthcare, **6**(2) (2025), 1-35. https://doi.org/10.1145/3709152

[5] I. D. Borlea, R. E. Precup, F. Dragan, A. B. Borlea, *Centroid update approach to K-means clustering*, Advances in Electrical and Computer Engineering, **17**(4) (2017). https://doi.org/10.4316/AECE.2017.04001

[6] R. Botvinik-Nezer, et al., *Variability in the analysis of a single neuroimaging dataset by many teams*, Nature, **582** (2020), 84-88. https://doi.org/10.1038/s41586-020-2314-9

[7] R. L. Buckner, J. R. Andrews-Hanna, D. L. Schacter, *The brain's default network: Anatomy, function, and relevance to disease*, Annals of the New York Academy of Sciences, **1124**(1) (2008), 1-38. https://doi.org/10.1196/annals.1440.011

[8] P. W. Burgess, I. Dumontheil, S. J. Gilbert, *The gateway hypothesis of rostral prefrontal cortex (area 10) function*, Trends in Cognitive Sciences, **11**(7) (2007), 290-298. https://doi.org/10.1016/j.tics.2007.05.004

[9] N. Çağman, S. Enginoğlu, *Fuzzy soft matrix theory and its application in decision making*, Iranian Journal of Fuzzy Systems, **9**(1) (2012), 109-119. https://doi.org/10.22111/ijfs.2012.229

[10] N. Çağman, S. Enginoğlu, F. Çıtak, *Fuzzy soft set theory and its applications*, Iranian Journal of Fuzzy Systems, **8**(3) (2011), 137-147. https://doi.org/10.22111/ijfs.2011.292

[11] H. S. Chiang, D. H. Shih, B. Lin, M. H. Shih, *An APN model for arrhythmic beat classification*, Bioinformatics, **30**(12) (2014), 1739-1746. https://doi.org/10.1093/bioinformatics/btu101

[12] M. Corbetta, G. L. Shulman, *Control of goal-directed and stimulus-driven attention in the brain*, Nature Reviews Neuroscience, **3**(3) (2002), 201-215. https://doi.org/10.1038/nrn755

[13] E. Dhamala, K. W. Jamison, M. R. Sabuncu, A. Kuceyeski, *Sex classification using long-range temporal dependence of resting-state functional MRI time series*, Human Brain Mapping, **41**(13) (2020), 3567-3579. https://doi.org/10.1002/hbm.25030

[14] A. Etkin, T. Egner, R. Kalisch, *Emotional processing in anterior cingulate and medial prefrontal cortex*, Trends in Cognitive Sciences, **15**(2) (2011), 85-93. https://doi.org/10.1016/j.tics.2010.11.004

[15] L. Fan, et al., *The human brainnetome Atlas: A new brain atlas based on connectional architecture*, Cerebral Cortex, **26**(8) (2016), 3508-3526. https://doi.org/10.1093/cercor/bhw157

[16] E. S. Finn, et al., *Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity*, Nature Neuroscience, **18** (2015), 1664-1671. https://doi.org/10.1038/nn.4135

[17] S. Gadgil, Q. Zhao, A. Pfefferbaum, et al., *Spatio-temporal graph convolution for resting-state fMRI analysis*, Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, **12267** (2020), 528-538. https://doi.org/10.1007/978-3-030-59728-3_52

[18] M. Gholami, M. Faramarzi, N. Alipour, M. Pakravan, *Node embedding extraction for causal brain graphs in fMRI data*, In Proceedings of the 29th International Computer Conference, Computer Society of Iran (CSICC), (2025), 1-6. https://doi.org/10.1109/CSICC65765.2025.10967434

[19] M. F. Glasser, et al., *The minimal preprocessing pipelines for the human connectome project*, NeuroImage, **80** (2013), 105-124. https://doi.org/10.1016/j.neuroimage.2013.04.127

[20] M. F. Glasser, et al., *A multi-modal parcellation of human cerebral cortex*, Nature, **536** (2016), 171-178. https://doi.org/10.1038/nature18933

[21] E. M. Gordon, et al., *Generation and evaluation of a cortical area parcellation from resting-state correlations*, Cerebral Cortex, **26**(1) (2016), 288-303. https://doi.org/10.1093/cercor/bhu239

[22] R. M. Hutchison, et al., *Dynamic functional connectivity: Promise, issues, and interpretations*, NeuroImage, **80** (2013), 360-378. https://doi.org/10.1016/j.neuroimage.2013.05.079

[23] M. Khosla, K. Jamison, G. H. Ngo, A. Kuceyeski, M. R. Sabuncu, *Machine learning in resting-state fMRI analysis*, Magnetic Resonance Imaging, **64** (2019), 101-121. https://doi.org/10.1016/j.mri.2019.05.031

[24] P. Y. Kim, J. Kwon, et al., *SwiFT: Swin 4D fMRI Transformer*, Advances in Neural Information Processing Systems (NeurIPS 2023), **36** (2023), 42015-42037. https://doi.org/10.48550/arXiv.2307.05916

[25] B. H. Kim, J. C. Ye, J. J. Kim, *Learning dynamic graph representation of brain connectome with spatio-temporal attention*, Advances in Neural Information Processing Systems (NeurIPS 2021), **34** (2021), 4314-4327. https://proceedings.neurips.cc/paper/2021/hash/229754a0ca4d715694200427845774a3-Abstract.html

[26] J. Kwon, J. Seo, H. Wang, T. Moon, S. Yoo, J. Cha, *Predicting task-related brain activity from resting-state brain dynamics with fMRI Transformer*, Imaging Neuroscience, **3** (2025). https://doi.org/10.1162/imag_a_00440

[27] R. Leech, D. J. Sharp, *The role of the posterior cingulate cortex in cognition and disease*, Brain, **137**(1) (2014), 12-32. https://doi.org/10.1093/brain/awt162

[28] M. Leeming, J. Suckling, *Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank*, NeuroImage, **241** (2021), 118409. https://doi.org/10.1016/j.neuroimage.2021.118409

[29] N. Leonardi, D. Van De Ville, *On spurious and real fluctuations of dynamic functional connectivity during rest*, NeuroImage, **104** (2015), 430-436. https://doi.org/10.1016/j.neuroimage.2014.09.007

[30] X. Li, X. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, et al., *BrainGNN: Interpretable brain graph neural network for fMRI analysis*, Medical Image Analysis, **74** (2021), 102233. https://doi.org/10.1016/j.media.2021.102233

[31] V. Menon, M. D'Esposito, *The role of PFC networks in cognitive control and executive function*, Neuropsychopharmacology, **47** (2022), 90-103. https://doi.org/10.1038/s41386-021-01152-w

[32] R. Nazari, M. Salehi, A. Shoeibi, *An explainable connectome convolutional transformer for multimodal autism spectrum disorder classification*, International Journal of Neural Systems, **35**(8) (2025), 2550043. https://doi.org/10.1142/S0129065725500431s

[33] J. Ning, *Neural network-based pattern recognition in the framework of edge computing*, Romanian Journal of Information Science and Technology, **2024**(1) (2024), 106-119. https://doi.org/10.59277/ROMJIST.2024.1.08

[34] M. Pakravan, *FuzzyCAL: A fuzzy logic enhanced causal attention GNN for robust cocaine use disorder classification*, Iranian Journal of Fuzzy Systems, **22**(6) (2025), 167-182. https://doi.org/10.22111/ijfs.2025.52840.9343

[35] M. Pakravan, *Uncertainty aware type II fuzzy graph modeling of resting state fMRI uncovers robust sex differences*, Journal of Neuroscience Methods, **431** (2026). https://doi.org/10.1016/j.jneumeth.2026.110745

[36] M. Pakravan, *Explainable graph attention network on resting state fMRI dynamic functional connectivity reveals cerebellar hub breakdown in chronic knee osteoarthritis pain*, Iranian Journal of Science and Technology, Transactions of Electrical Engineering, **50** (2026), 289-300. https://doi.org/10.1007/s40998-025-01012-z

[37] U. Pervaiz, R. Vidaurre, M. W. Woolich, M. E. Smith, *Optimising network modelling methods for fMRI*, NeuroImage, **211** (2020), 116604. https://doi.org/10.1016/j.neuroimage.2020.116604

[38] J. D. Power, et al., *Methods to detect, characterize, and remove motion artifact in resting-state fMRI*, NeuroImage, **84** (2014), 320-341. https://doi.org/10.1016/j.neuroimage.2013.08.048

[39] S. Ramezanzadeh, A. Memariani, S. Saati, *Data envelopment analysis with fuzzy random inputs and outputs: A chance-constrained programming approach*, Iranian Journal of Fuzzy Systems, **2**(2) (2005), 21-35. https://doi.org/10.22111/IJFS.2005.479

[40] S. J. Ritchie, et al., *Sex differences in the adult human brain: Evidence from 5,216 UK Biobank participants*, Cerebral Cortex, **28**(8) (2018), 2959-2975. https://doi.org/10.1093/cercor/bhy109

[41] S. Ryali, Y. Zhang, et al., *Deep learning models reveal replicable and behaviorally relevant sex differences in brain organization*, Proceedings of the National Academy of Sciences, **121**(9) (2024), e2310012121. https://doi.org/10.1073/pnas.2310012121

[42] M. A. Saket, M. Pakravan, *Cross-atlas identification of narrative hubs via multi-embedding graph models in fMRI data*, Neuroinformatics, **24** (2026), 29. https://doi.org/10.1007/s12021-026-09787-0

[43] S. S. Salehi, et al., *There is no single functional atlas even for a single individual: Functional parcel definitions change with task*, NeuroImage, **208** (2020), 116366. https://doi.org/10.1016/j.neuroimage.2019.116366

[44] E. Tjoa, C. Guan, *A survey on explainable artificial intelligence (XAI): Toward medical XAI*, IEEE Transactions on Neural Networks and Learning Systems, **32**(11) (2021), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

[45] D. C. Van Essen, et al., *The WU-Minn human connectome project: An overview*, NeuroImage, **80** (2013), 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

[46] C. Wang, A. U. Yaari, et al., *BrainBERT: Self-supervised representation learning for intracranial recordings*, arXiv preprint arXiv:2302.14367, (2023). https://doi.org/10.48550/arXiv.2302.14367

[47] S. Weis, K. R. Patil, et al., *Sex classification by resting-state brain connectivity*, Cerebral Cortex, **30**(2) (2020), 824-835. https://doi.org/10.1093/cercor/bhz129

[48] L. A. Zadeh, *Fuzzy sets*, Information and Control, **8**(3) (1965), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

[49] A. Zalesky, A. Fornito, L. Cocchi, L. L. Gollo, M. Breakspear, *Time-resolved resting-state brain networks*, Proceedings of the National Academy of Sciences, **111**(28) (2014), 10341-10346. https://doi.org/10.1073/pnas.1400181111

[50] L. Zhao, *Embedding human brain function via transformer*, Medical Image Computing and Computer-Assisted Intervention (MICCAI), **13431** (2022), 366-375. https://doi.org/10.1007/978-3-031-16431-6_35