Towards Semantic Understanding of GNN Layers Embedding with Functional Semantic Activation Mapping

Kislay Raj [a,*] and Alessandra Mileo [a]
[a] *INSIGHT Centre for Data Analytics & School of Computing, Dublin City University, Ireland, Dublin*
E-mail: kislay.raj2@mail.dcu.ie

Abstract. Graph Neural Networks (GNNs) have demonstrated significant potential in learning representations from complex graph structured data. However, most existing explainability methods focus on instance level explanations, which aim to clarify why a model made a specific prediction for a given input. While valuable for understanding individual outcomes, such methods do not capture the broader, model level behaviours that emerge across all inputs leaving global interpretability as an open challenge. In addition, the impact of GNN architecture on the quality of deep representations is poorly understood, with the optimal layer configuration still typically determined empirically by trial and error. In this paper, we extend our previous work on Functional Semantic Activation Mapping (FSAM) to investigate how varying the number of GNN layers affects both representation quality and predictive performance. Across multiple datasets, increasing depth can improve accuracy, but does not necessarily enhance semantic coherence. In some cases, performance gains coincide with a decline in semantic quality, indicating that spurious patterns may drive correct predictions, suggesting correct predictions for incorrect reasons. FSAM layer-wise activation tracking allowed us to track neuron activations across layers, revealing that deeper layers can reduce neuron specialisation and lead to class misclassifications. Community analysis further indicates that certain misclassified classes share neurons in overlapping communities, highlighting a loss of class-specific representations at greater depths. Our findings demonstrate a critical trade-off that increased depth can compromise interpretability without commensurate gains in meaningful semantic learning. Instead of chasing accuracy alone, we need frameworks that assess whether models are learning coherent patterns; this is where FSAM proves invaluable, as it is emerging as a vital model level diagnostic tool for architectural analysis.

Keywords: Explainable AI, Graph Neural Network, Graph Analysis, Neuro-Symbolic AI

1. Introduction

GNNs [8, 22, 31] have shown remarkable performance in node classification, link prediction, and graph classification tasks. GNNs leverage structural information and node features to capture complex relationships within a graph. However, explaining GNN predictions remains a challenge due to the complex topological nature of graphs and how this is represented in GNN embeddings. Unlike traditional neural networks, GNNs operate on graph structures, which might suggest better interpretability, but understanding how these relationships are learned within the layers remains ambiguous. In current research on the explainability of GNN, most local methods [32] focus on generating small subgraphs and identifying which nodes and edges contribute to a specific prediction. However, they do not explicitly show how information is processed within the network layers. Although useful, they fail to

*Corresponding author. E-mail: kislay.raj2@mail.dcu.ie.

comprehensively understand how the GNN behaves across different layers and do not offer a global understanding of model behaviour.

In our previous work [15], we introduced the FSAM framework to address this gap. FSAM explains how the entire network behaves across its layers by producing, for each layer, a semantic graph whose vertices correspond to neurons and whose edges represent statistically significant coactivation patterns. These graphs are termed semantic because the neuron groups (communities) often correspond to higher-level concepts learned by the model. For example, in a citation network, one community might activate predominantly for papers on machine learning, another for data mining, with connections between them reflecting shared topical structure. By comparing these semantic graphs across layers, FSAM reveals how class-specific or concept-specific representations emerge or merge as network depth increases. In this extended work, we apply FSAM to explore a central question in GNN design: *To what extent does increasing the number of layers enhance the model's ability to represent meaningful patterns? Furthermore,* does higher predictive accuracy necessarily imply more faithful or discriminative internal representations?

The oversmoothing phenomenon has been well studied in the literature and is a well-known problem in GNNs [25, 9, 16], which occurs when we add more layers of information to a GNN architecture. Our findings indicate that oversmoothing reduces FSAM quality, as evidenced by the degradation of the model's ability to represent meaningful, class-specific features across layers. Instead of merely confirming that oversmoothing occurs, FSAM provides a better-structured way to detect and quantify the effects of oversmoothing at the neuron level. It tracks when and where neurons begin to lose their class-specific activations, offering a more profound insight into the impact of model depth on representation quality. In this sense, FSAM is more likely to be a diagnostic tool for global-level model behaviour, indicating where oversmoothing might occur. Furthermore, our research has practical implications. We have observed cases where GNN performance improves without a corresponding enhancement in FSAM quality; this suggests that the model may make correct predictions, but not necessarily for the right reasons, as it could rely on less meaningful or oversmoothed features. Therefore, FSAM offers an interpretability driven analysis of oversmoothing which provides insight into the reasons behind the model's predictions. Our findings elucidate the trade-off between model depth and interpretability, empirically demonstrating how excessive layering can degrade semantic coherence while maintaining superficial accuracy metrics. The FSAM framework emerges as an essential diagnostic tool offering researchers the unprecedented capability to (i) quantify the progressive loss of neuron specialisation across layers, (ii) identify where correct classifications stem from flawed reasoning patterns, and (iii) establish optimal depth thresholds before semantic collapse occurs, establishing a new paradigm for assessing both what GNNs predict and how they derive these predictions - a crucial distinction for deploying graph networks in high-stakes real-world applications.

Since this paper relies on capturing the GNN's behaviour through activation analysis with FSAM, our first contribution is to extend FSAM validation beyond our previous experiments on CORA [20] and CiteSeer [12]. To achieve this, we conducted additional experiments on four different datasets: PubMed [3], Amazon Computers [11], Amazon Photos [11], and Coauthor [21]. These datasets with their distinct topological complexities allow us to comprehensively evaluate FSAM's approach and determine how well the resulting activation graph reflects the GNN's behaviour and how effectively the network learns the semantic structure of the input data.

The contributions of this work can be summarised as follows. Firstly, we extend the FSAM approach by conducting experiments on a broader range of datasets to validate that the activation analysis and graphs generated by FSAM consistently reflect the network behaviour. This includes community analysis in different datasets that demonstrates the ability of FSAM to capture the semantic structure between classes reliably. Secondly, we extend our experimental analysis to confirm that the functional activation graph generated by FSAM aligns with the network's behaviour as the number of layers changes. By testing networks with different depths (from 1 to 4 layers) and comparing the correlation between misclassifications and class similarity, we show that improvements in network accuracy are reflected in the FSAM graph, and the FSAM structure also captures any decline in accuracy. This analysis emphasises FSAM's ability to represent network behaviour across different layer configurations accurately. Third, we conduct a detailed layer-by-layer analysis to assess how different GNN layer configurations affect the model's performance in node classification tasks. Specifically, we examine how varying the number of layers influences FSAM and the corresponding community structure and verify whether improvements in accuracy align with better FSAM graphs and, on the other hand, decreases in accuracy correlate with a decline in FSAM quality. This analysis demonstrates

that while additional layers may enhance performance, deeper layers can lead to oversmoothing; neuron activation overlap, and ultimately diminish the model's ability to differentiate between classes. As part of our detailed, comprehensive analysis, we also identify a few interesting cases where accuracy improves without a corresponding improvement in FSAM's semantic quality. These instances reveal situations where the GNN achieves better predictions, but not necessarily due to a better embedding of the semantic structure in the input data. It highlights FSAM's potential in identifying cases where a model makes accurate predictions for the wrong reasons.

2. State of the Art

The interest in neuro-symbolic AI has increased steadily, driven by the need for interpretable and accountable machine learning systems, especially in domains that require transparent decision making. Research has focused on integrating neural learning with symbolic reasoning, an essential step to enhance the explainability of deep learning models. This integration is crucial for high-stakes domains where accuracy and interpretability are essential. GNNs have shown exceptional performance in handling graph-structured data in a range of fields, such as social networks [28], molecular structures [6], and citation networks [26]. However, despite their success in learning complex relationships, GNNs remain largely opaque with respect to how specific predictions are made, particularly when compared to models in other domains like image and text analysis. The challenge lies in interpreting the internal representations learned by GNNs, particularly about prior knowledge.

Most existing methods for GNN explainability focus on *local explanations*, identifying key input features, nodes, or edges influencing individual predictions. These techniques are broadly divided into several categories: **Gradient/Feature-based methods** [14], which use gradient information or hidden feature map values to assess feature importance; **Perturbation-based methods** [29], which modify graph structures and monitor how these perturbations affect model outputs; **Decomposition methods** [18, 14], which break down the prediction score into contributions from different neurons or layers, propagating these contributions backwards through the network; and **Surrogate methods** [7, 23], which train interpretable models to approximate the GNN's behaviour by sampling the input graph's neighbourhood and constructing an explanation based on the simplified model.

In addition to these local approaches, recent research has investigated rule based explanations for GNNs, where the learned model is partially or entirely translated into a set of human interpretable logical rules [4]. Such methods explicitly connect neural computation with symbolic reasoning, enabling formal reasoning about the model's predictions. Although promising in their transparency, such approaches often suffer from generating overly complex rules for deeper architectures or large-scale graphs, which can limit practical interpretability. Our FSAM framework addresses a different, yet complementary, aspect of the explainability problem. Instead of enumerating potentially unwieldy rule sets, FSAM provides a global, structured, and layer-wise semantic representation of how neurons interact and form communities throughout the network. This perspective captures overall behaviour and the flow of information across layers. These features are typically absent from both rule-based and local explanation approaches. By offering a consolidated view of the model's semantic structure, FSAM facilitates reasoning about its behaviour in the context of prior knowledge and domain-specific concepts.

It is worth stating that the use of *global explanations* is less researched in the context of GNN. One of the methods is **XGNN** [33], which creates synthetic graphs tailored to class predictions to explain the behaviour of the GNN. However, XGNN's assumption that a single synthetic graph can represent an entire class of graphs is unrealistic, as many relationships exist within real-world datasets. Empirical studies and follow-up methods have pointed out this shortcoming. For example, in molecular graph classification, several different substructures (motifs) can cause a molecule to be mutagenic [1]. XGNN tends to identify one dominant motif and fails to identify other motifs, thus lacking multi-modal explanations. Note that generating one graph per class erases any combinatorial aspect that the GNN could have learned; if the model has a concept for a class as a conjunction or disjunction of several patterns, the XGNN explanation will be partial. Another concrete example comes from chemical ring structures: MAGE [30] found that prior model-level explainers like XGNN often fail to identify certain valid substructures (e.g. rings in molecules), leading to questionable interpretability. It happens because XGNN builds graphs edge-by-edge ("atom-by-atom"), which can struggle to capture a complex motif that requires adding a set of edges together (closing a

ring). XGNN's inability to capture multiple modes or diverse substructures within the same class means that its explanations can be incomplete or biased toward one pattern. In contrast, studies show that more nuanced approaches can uncover a richer set of class-specific patterns. In recent research, GLGExplainer [1] addresses these gaps by providing global explanations as Boolean combinations of learned graphical concepts. Unlike XGNN, GLGExplainer aggregates local explanations into interpretable concepts, which are then combined into logic formulas. It learns these concepts purely from local explanations, without grounding them in human-interpretable semantics (e.g., chemical functional groups in molecules). However, its reliance on local explanations, discrete clustering, and lack of conceptual grounding limit its ability to fully align with the internal reasoning of the GNN.

Although these approaches provide some information about the final predictions, they do not explain how the intermediate layers participate in the learned representations, as they are not suitable for explaining the relationship between the internal structure of the model and prior knowledge or domain knowledge, which limits the ability of users to trust and understand the model's decisions. Furthermore, the related work in SOTA is presented in Table 1.

Table 1

Comparison of existing explainability methods for GNNs. The columns represent various aspects of each method: - **TYPE**: Indicates whether the method is **instance level** (Local) or **model level** (Global). - **LEARNING**: Specifies whether the method uses **backward** or **forward** propagation for explanation. - **TASK**: The tasks the method is designed to explain, such as **GC** (Graph Classification) or **NC** (Node Classification). - **TARGET**: The target to be explained, including **N** (node), **E** (edge), **NF** (node features), or **Subgraph**. - **Layer wise interpretability**: Describes whether the method tracking semantic changes across layers (✓) or (✗). Abbreviations for task, target, layer wise interpretability are explained in the caption.

Method	TYPE	LEARNING	TASK	TARGET	LAYER-WISE INTERPRETABILITY
SA [2, 14]	instance level	✗	GC/NC	N/E/NF	✗
Guided BP [2]	instance level	✗	GC/NC	N/E/NF	✗
CAM [14]	instance level	✗	GC	N	✗
Grad-CAM [14]	instance level	✗	GC	N	✗
GNNExplainer [29]	instance level	✓	GC/NC	E/NF	✗
PGExplainer [10]	instance level	✓	GC/NC	E	✗
GraphMask [17]	instance level	✓	GC/NC	E	✗
ZORRO [5]	instance level	✗	GC/NC	N/NF	✗
Causal Screening [24]	instance level	✗	GC/NC	E	✗
SubgraphX [32]	instance level	✓	GC/NC	Subgraph	✗
LRP [2, 19]	instance level	✗	GC/NC	N	✗
Excitation BP [14]	instance level	✗	GC/NC	N	✗
GNN-LRP [18]	instance level	✗	GC/NC	Walk	✗
GraphLime [7]	instance level	✓	NC	NF	✗
RelEx [34]	instance level	✓	NC	N/E	✗
PGM-Explainer [23]	instance level	✓	GC/NC	N	✗
XGNN [33]	model level	✓	GC	Subgraph	✗
FSAM (Our Work)	model level	✓	NC	N/NF	✓

Our previous work addresses this gap by introducing the **FSAM** approach, which provides a global explanation of GNNs by extracting deep representations in the form of semantic graphs. FSAM focuses on capturing the global structure of the GNN, along with the semantic relationships between neurons across different layers. This method explains which components contribute to predictions and reveals how information is processed throughout the network, offering a more transparent view of the GNN's behaviour. Unlike traditional input optimisation methods used for image classifiers [13], which cannot be applied to graph adjacency matrices without losing crucial structural information, FSAM is specifically designed to preserve the discrete properties of graph structures. One of the key aspects of FSAM over the existing model level methods is that it can project the learned representations of the GNN into the semantic space, which helps us to map the internal processes of the model to higher-order symbolic representations and, in turn, helps compare the model's decisions with prior knowledge and information. Although

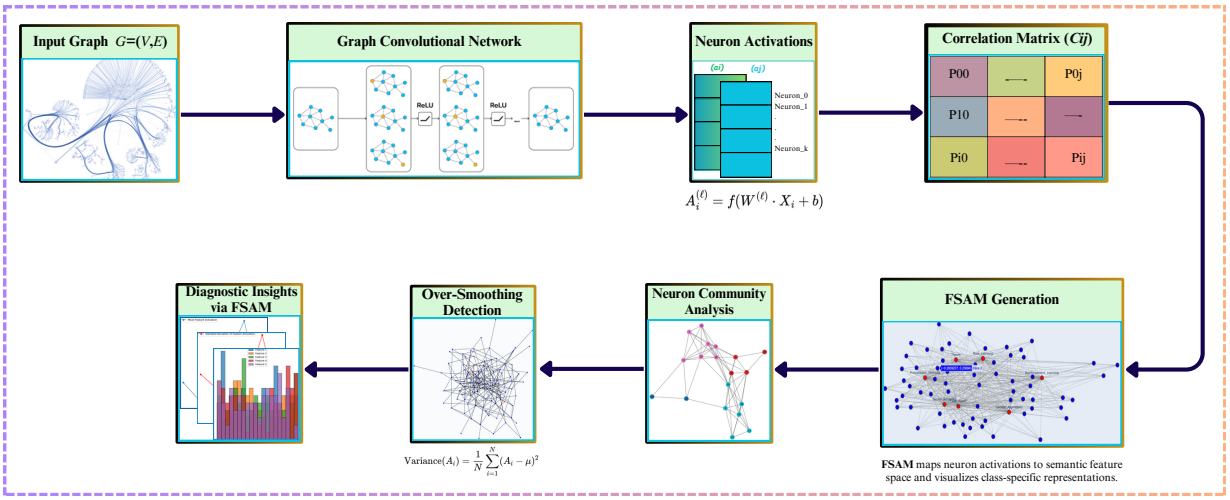

Fig. 1. Overall proposed system architecture

FSAM does not generate an explanation directly, it explains how the model works on graph data at each layer. In future work, this framework could help us develop explanations consistent with human understanding, which would be a substantial step toward neurosymbolic AI.

3. Overall Methodology: Generating the Semantic Graph

The primary aim of this paper is to improve the interpretability of GNNs by representing their internal mechanisms as semantic graphs. In our extended study, we hypothesised that adding more layers to GNNs does not necessarily increase their capacity for knowledge representation. Our FSAM method clarifies GNN decisions by focusing on how different layers contribute to, or sometimes reduce, model performance due to oversmoothing. FSAM identifies neuron groups involved in decision-making, termed *activation neurons*, and constructs a semantic graph to visualise their relationships. FSAM tracks neurone activations and visualises activation relationships across layers, providing valuable information on the network's decision-making process and semantic coherence. Figure 1 illustrates the proposed system architecture, highlighting the steps to optimise layer depth and improve GNN performance. This section presents the mathematical formulation for generating the semantic graph, integrated with insights from our extended experiments.

Step 1: Compute activation values and matrix: We begin by computing the activation values of all neurons in the GNN in a transductive node classification setting. We assume a single fixed attributed graph.

$$G = (V, E, X), \quad |V| = n,$$

where V is the set of nodes, $E \subseteq V \times V$ is the set of undirected edges (optionally weighted with $e_{v,w} \in \mathbb{R}$), and $X \in \mathbb{R}^{n \times d}$ is the node feature matrix. Node features $x_v \in \mathbb{R}^d$ may be binary (e.g., bag-of-words encodings in citation networks) or real-valued (e.g., normalised numerical attributes in product/coauthor graphs). All node features are known during training, but class labels $y_v \in \{1, \ldots, C\}$ are provided only for a subset $V_{\text{train}} \subset V$; the objective is to predict labels for $V \setminus V_{\text{train}}$. **Inputs:** the fixed graph (V, E), node features X, and partial labels for V_{train}. **Outputs:** per-node probability distributions over C classes, final predicted labels $\hat{\ell}_v$, and intermediate activations for FSAM analysis.

The GNN processes (G, X) to produce, at each layer i, an *activation matrix*

$$A^{(i)} = [a_1^{(i)}, a_2^{(i)}, \ldots, a_{h_i}^{(i)}] \in \mathbb{R}^{n \times h_i},$$

where h_i is the number of neurons (feature dimensions) in layer i. Each column $a_k^{(i)} \in \mathbb{R}^n$ contains the activation values of neuron k across all n nodes. In our extended analysis, we observe that beyond a certain depth, additional neurons contribute diminishing amounts of class-discriminative information due to oversmoothing, which can degrade performance.

Step 2: Graph embedding: To capture the behaviour of *neurons*, we use Graph Convolutional Networks (GCNs) [8], which transform each node's *ego-graph* (the node and its immediate neighbours) to a latent Euclidean space that encodes structural and feature information. The embedding for node v in layer ℓ is given by:

$$h_v^{(\ell)} = \mathrm{ReLU}\left(W^{(\ell)} \sum_{w \in N(v) \cup \{v\}} \frac{e_{v,w}}{\sqrt{d_v d_w}} h_w^{(\ell-1)} \right).$$

Notation: let $H^{(\ell)} \in \mathbb{R}^{|V| \times d_\ell}$ be the node embedding matrix whose row $h_v^{(\ell)}$ is the embedding of node v; we set $A^{(\ell)} := H^{(\ell)}$. Thus, the activation of neuron/feature i at node v is $a_{v,i}^{(\ell)} = h_{v,i}^{(\ell)}$, and the neuron i activation profile across nodes is the *column* $a_{\cdot,i}^{(\ell)}$ of $A^{(\ell)}$.

where d_v is the degree of v (including self-loops if present), $W^{(\ell)}$ are trainable weight matrices and ReLU is non-linear activation. At $\ell = 0$, we set $h_v^{(0)} = x_v$. The final layer outputs $\{h_v^{(L)}\}_{v \in V}$ are passed through a softmax classifier to yield per-node class probabilities; intermediate activations $\{A^{(\ell)}\}_{\ell=1}^{L}$ are retained for FSAM analysis.

Step 3: Compute edge weights: We compute edge weights within the coactivation matrix to analyse the relationships between neurons using Spearman's rank correlation coefficient, an appropriate metric for capturing monotonic and non-linear associations among activation patterns. The Spearman coefficient ρ_{ij} for neurons i and j is:

$$\rho_{ij} = \frac{\mathrm{cov}(\mathrm{rank}(a_i), \mathrm{rank}(a_j))}{\sigma_{\mathrm{rank}(a_i)}\, \sigma_{\mathrm{rank}(a_j)}},$$

where cov denotes covariance, $\mathrm{rank}(\cdot)$ returns the rank transformed activation vector across nodes, and $\sigma_{\mathrm{rank}(a_i)}$ is the standard deviation of the ranked values. Although neuron activations can be non-monotonic functions of each other due to signed weights and non-linearities, Spearman's ρ remains appropriate for our setting, as it measures the strength and direction of any consistent rank order relationship (increasing or decreasing) between nodes, without requiring strict monotonicity.

This measurement quantifies neuron–neuron relationships independently of the classifier head and will be used in subsequent sections to analyse representation structure across layers.

We correlate intermediate layer activations with the predicted class of the final layer to characterise the model's decision boundary. Let $c \in \{1, \ldots, C\}$, $\hat{\ell}_v = \arg\max_c p^{(L)}(y = c \mid v)$ and define the one–vs–rest indicator $\hat{y}_v^{(c)} = \mathbb{I}[\hat{\ell}_v = c]$. Set $G_c = \{v : \hat{y}_v^{(c)} = 1\}$ and $G_{\neg c} = \{v : \hat{y}_v^{(c)} = 0\}$ with sizes $n_c = |G_c|$, $n_{\neg c} = |G_{\neg c}|$, and $n = n_c + n_{\neg c}$. For layer ℓ and neuron (feature) $j \in \{1, \ldots, d_\ell\}$, let $a_{\cdot j}^{(\ell)} \in \mathbb{R}^n$ be the activation column of neuron j across nodes. We compute the point–biserial correlation *per layer, per neuron, per class* as

$$r_{pb}^{(\ell)}(j, c) = \frac{\bar{a}_{j|c}^{(\ell)} - \bar{a}_{j|\neg c}^{(\ell)}}{s_{j,\mathrm{pooled}}^{(\ell)}} \sqrt{\frac{n_c\, n_{\neg c}}{n^2}},$$

where

$$\bar{a}_{j|c}^{(\ell)} = \frac{1}{n_c} \sum_{v \in G_c} a_{v,j}^{(\ell)}, \qquad \bar{a}_{j|\neg c}^{(\ell)} = \frac{1}{n_{\neg c}} \sum_{v \in G_{\neg c}} a_{v,j}^{(\ell)},$$

and the pooled standard deviation is

$$s^{(\ell)}_{j,\text{pooled}} = \sqrt{\frac{(n_c - 1)\,(s^{(\ell)}_{j|c})^2 + (n_{\neg c} - 1)\,(s^{(\ell)}_{j|\neg c})^2}{n_c + n_{\neg c} - 2}},$$

with $(s^{(\ell)}_{j|c})^2$ and $(s^{(\ell)}_{j|\neg c})^2$ the sample variances of $a^{(\ell)}_{v,j}$ over G_c and $G_{\neg c}$, respectively. This quantity is computed for every (ℓ, j, c) triple.

Step 4: Thresholding and visualising: To visualise this behaviour, we constructed coactivations graphs for each layer and applied dynamic thresholding techniques [27] to filter out statistically insignificant edges. These visualisations reveal how additional layers alter the semantic structure of the learned representation. Deeper layers contribute minimal new discriminative information while increasing the overlap between classes, thereby reducing separability. FSAM captures and presents this progression, allowing both the diagnosis of oversmoothing and the design of more interpretable and efficient GNN architectures.

4. Key Contributions and Findings

This study explores how varying the number of layers in GNNs impacts both model performance and the quality of knowledge representation. Building on our previous research, we employ FSAM to systematically assess how well different layer configurations capture the underlying structure of input data. As our contributions address a central question, Do additional layers enhance the model's interpretability and accuracy, or do they introduce complexity that impairs representation quality?

The following contributions highlight the key findings of this study: **Contribution 1**: Extended the validation of FSAM on multiple datasets, as presented in Section 5.3. To evaluate FSAM's generalisability, we apply it to various datasets in Section 5.2. Using Jaccard correlation graphs Fig. (5, 6, 7), we analyse the ability of FSAM to capture semantic structure, defined here as the organisation of neurones into communities that reflect class-specific activation patterns and their relationships. Our results show that FSAM reliably mirrors changes in network behaviour, with variations in model accuracy typically accompanied by corresponding shifts in the clarity and separation of these communities. We also find cases where accuracy improves without stronger semantic alignment, underscoring the diagnostic value of FSAM to detect potential mismatches between accuracy and representational quality.

Contribution 2: In Section 5.3, we study how FSAM captures the behaviour of the network in GNNs with one to four layers. By examining the correlation between misclassifications and neuron community structures, we confirm that FSAM accurately reflects evolving network dynamics as depth increases. Here, coherent FSAM representations refer to graphs in which neuron communities are compact, well separated, and semantically aligned with class boundaries. When accuracy declines often due to oversmoothing in deeper layers, FSAM visualisations reveal reduced semantic clarity, with communities becoming more mixed and less separable. These patterns highlight the robustness of FSAM as a tool for interpreting and diagnosing GNN behaviour across architectural variations.

Contribution 3: Section 5.4 and Section 5.3 present a layer-wise analysis of how increasing the number of GNN layers impacts model performance. We focus on class-specific accuracy and oversmoothing, where neuron activations become overly similar. Our experiments reveal that, while additional layers may enhance performance, they lead to a decline in discriminative power beyond a certain depth. This contribution supports optimising layer depth in GNN architectures, highlighting the trade-offs between model complexity and representation quality.

Contribution 4: In Section 5.5, we discuss semantic divergences in the FSAM graphs. A key finding from our analysis is the ability of FSAM to identify cases where accuracy trends and FSAM quality diverge. Specifically, we highlight instances where model accuracy improves, but the quality of the FSAM graph decreases, indicating cases where the network achieves correct predictions without fully capturing the semantic structure of the input data. Conversely, we also find situations where accuracy decreases, but the quality of the FSAM graph improves, potentially due to richer insights gained from misclassifications. These cases emphasise FSAM's diagnostic potential in detecting "right for the wrong reasons" scenarios, offering a nuanced understanding of the network's semantic

alignment with the data.
Overall, these contributions extend our previous work, providing a detailed methodology for assessing GNN layer depth and performance. Our findings position FSAM as a valuable framework for balancing layer depth with interpretability and accuracy, ultimately enhancing the understanding and optimisation of GNN architectures across various datasets.

5. Experiment

5.1. Experimental Set-up

To evaluate the proposed FSAM framework, we conducted experiments in a setting *transductive node classification* using six benchmark data sets (Section 5.2). All experiments were implemented in PyTorch Geometric, with the GCNConv layer as the graph convolution operator. The base GCN architecture consisted of multiple graph convolution layers, each followed by a ReLU activation and dropout for regularisation. A final fully connected layer mapped the hidden representations to the C output classes, followed by a log-softmax output.

Model configurations:- We varied the number of GCN layers from 1 to 5 to assess the effect of network depth on both classification accuracy and semantic graph structure. All experiments used:

- **Hidden channels:** 32
- **Dropout rate:** 0.5 after each convolution layer
- **Optimiser:** Adam
- **Learning rate:** 0.01
- **Weight decay:** 5×10^{-4}
- **Loss:** Negative Log-Likelihood (NLL)
- **Epochs:** 200

Training and evaluation used the standard `train_mask` and `test_mask` splits provided with each dataset. All runs were repeated with a fixed random seed for reproducibility.

Evaluation metrics:-

- **Classification Accuracy:** the proportion of correctly classified nodes.
- **Macro-F1 Score:** the unweighted mean F1 score for all classes, ensuring balanced evaluation in datasets with class imbalance.
- **Layerwise Spearman Correlation:** mean absolute Spearman ρ between neuron activations, to measure coactivation growth indicative of oversmoothing.
- **Point Biserial Feature Class Correlation:** mean r_{pb} across neurones to quantify the loss of class-specific discriminative power with depth.

Our aim was to assess whether increasing GCN depth leads to more meaningful internal representations or, conversely, to oversmoothing and loss of discriminative capacity. FSAM was used to generate semantic graphs per layer, allowing a direct visual and quantitative comparison of neuron relationships across depths.

5.2. Datasets

We used six benchmark datasets covering diverse graph domains:

- **Cora** [20] and **CiteSeer** [12]: citation networks with nodes as publications and edges as citation links. Cora contains 2,708 nodes (7 categories); CiteSeer contains 3,312 nodes (6 categories).
- **PubMed** [3]: biomedical citation network with 19,717 publications in 3 disease categories.
- **Amazon Computers** and **Amazon Photos** [11]: product co-purchase graphs. Amazon Computers has 13,752 products; Amazon Photos has 7,650 products.

– **Coauthor CS** [21]: co-authorship network with 18,333 authors in computer science, labelled by research area.

These datasets span a wide range of domains, from academic publications and biomedical research to product co-purchases and academic co-authorships.

5.3. Extended Validation of FSAM Across Layer Configurations

Our second contribution involves validating the ability of the FSAM approach to capture GNN behaviour in varying layer configurations reliably. Through systematic experiments on several datasets, as detailed in Section 5.2, we analysed each GNN configuration (from 1 to 4 layers) to assess the alignment between model accuracy, misclassification patterns, and community structures represented by FSAM graphs.

Table 2
Final accuracy and Pearson correlation of models with different depths (1–4 layers) across three datasets.

Layer	Amazon Photos		CoauthorCS		Amazon Computers	
	Accuracy	**Pearson Correlation**	**Accuracy**	**Pearson Correlation**	**Accuracy**	**Pearson Correlation**
1	0.95±0.038	0.681	0.98±0.01	0.589	0.89±0.038	0.683
2	0.96±0.053	0.650	0.97±0.045	0.756	0.91±0.036	0.630
3	0.94±0.042	0.752	0.96±0.035	0.819	0.88±0.038	0.785
4	0.93±0.048	0.780	0.95±0.032	0.834	0.86±0.042	0.917

Note: Pearson's correlation is used here to measure the linear association between FSAM-derived metrics and model accuracy in different layer configurations. This differs from the Spearman correlation in Section 3, which is applied to neuron level activation patterns where nonlinear monotonic relationships are expected.

Table 3
Layer-wise Analysis with Statistical Validation

Layer	Amazon Photos		Coauthor CS		Amazon Computers	
	Acc. (95% CI)	**Corr. (p)**	**Acc. (95% CI)**	**Corr. (p)**	**Acc. (95% CI)**	**Corr. (p)**
1	0.95 [0.91, 0.99]	0.681 (0.041*)	0.98 [0.97, 0.99]	0.589 (0.038*)	0.89 [0.85, 0.93]	0.683 (0.025*)
4	0.93 [0.88, 0.98]	0.780 (0.008*)	0.95 [0.92, 0.98]	0.834 (0.002*)	0.86 [0.82, 0.90]	0.917 (0.001*)
Key Comparisons						
L1 vs L4	p=0.072 (Δ=-0.02)		p=0.011* (Δ=-0.03)		p=0.029* (Δ=-0.03)	
Corr. ↑	p=0.037*		p=0.003*		p<0.001*	

*Significant at α=0.05. Corr. ↑ tests Pearson increase from L1→L4.

In Table 2, we present the results for the Amazon Photo dataset, illustrating the progression of layerwise accuracy across configurations and highlighting how FSAM captures the relationship between classification errors and community structures.

In Layer 1, the model achieves an accuracy of 95% with a Pearson correlation of 0.681. This positive correlation suggests that class-specific representations are moderately well separated, with fewer overlapping nodes in the FSAM graph, leading to lower misclassification rates. The FSAM graph at this layer reveals distinct class representations, demonstrating effective differentiation early in the network. Adding a second layer improves accuracy to 96%, while the Pearson correlation decreases slightly to 0.650. This layer further strengthens class-specific separation without significant overlap in neuron activations. FSAM visualisations at this stage show that, while additional depth aids in correct predictions, it does not compromise the integrity of class distinctions, reflecting the model's enhanced capacity to maintain semantic coherence. In Layer 3, the accuracy begins to decline, dropping to 94%, while the Pearson correlation rises to 0.752. This increased correlation value indicates a greater overlap in neurone activations, signaling a loss of distinctiveness among class-specific features. Here, FSAM reveals that oversmoothing begins to emerge, with class representations blurring as neuron activations overlap. This finding aligns with our

previous work, which observed that classes with high node overlap in the FSAM graph tend to cause more mistakes, highlighting the need for improved class separation strategies. At Layer 4, the accuracy decreases further to 93%, and the Pearson correlation reaches 0.780, confirming substantial activation overlap and diminished distinctiveness in class representations. FSAM visualisations reveal extensive overlap between neuron communities, indicating that deeper layers contribute to oversmoothing. These observations suggest that overlapping nodes between similar classes might be prime targets for tuning, as reducing this overlap could improve the model's ability to distinguish these classes effectively.

These findings reinforce FSAM's effectiveness in tracing the network's behaviour across varying depths. Although initial layers improve accuracy with minimal activation overlap, additional layers increase the correlation between overlapping nodes and misclassification errors. This positive correlation between class similarity and mistake counts underscores FSAM's diagnostic potential, providing insight into where the network's performance could be optimised by minimising activation overlaps between similar classes, ultimately aiding in balancing depth and semantic clarity within GNNs.

Similarly, for the Coauthor CS dataset (Table 2), our findings strongly support the hypothesis that FSAM effectively captures layerwise shifts in network behaviour.

In the first layer, with a high accuracy of 98% and a low Pearson correlation of 0.589, neuron activations remain largely distinct, allowing for clear class separations. As we add layers, accuracy decreases slightly (97% at Layer 2) while correlation increases (0.756), indicating a gradual increase in activation overlap. By the third layer, accuracy drops further to 96%, with a higher Pearson correlation of 0.819, signalling the onset of oversmoothing as neuron activations increasingly overlap, thus blurring class distinctions. In the fourth layer, with an accuracy of 95% and a correlation of 0.834, this trend persists, showing that additional depth now undermines the model's ability to separate classes effectively.

These findings illustrate that FSAM consistently mirrors the evolving behaviour of the network across layers, accurately capturing the interaction between model accuracy and neuron overlap and confirming its usefulness in diagnosing the point at which further layers no longer benefit performance

In the Amazon Computers dataset (Table 2), we apply the same methodology, analysing how variations in accuracy between layers relate to the structures of the underlying graphs of FSAM. In the first layer, with an accuracy of 90% and a Pearson correlation of 0.683, the FSAM graph captures a balanced representation of the behaviour of the network. This correlation level suggests that neuron activations are distinct enough to preserve class separations effectively, reflecting that the FSAM captures clear distinctions among classes without excessive overlap.

When a second layer is added, the accuracy increases slightly to 91%, while the Pearson correlation decreases to 0.630. This reduction in correlation and improved precision indicate that neuron activations remain well separated, supporting the continued effectiveness of the model in distinguishing between classes. The FSAM graph here effectively aligns with the improved class distinction, reinforcing the model's structural clarity.

However, in the third layer, accuracy decreases to 88%, and the Pearson correlation increases to 0.785. This shift indicates an increase in the activation of overlapping neurons, suggesting a decline in class distinction, likely attributable to oversmoothing. The FSAM graph reflects this change, capturing the network's diminished ability to maintain distinct class representations as neuron activations converge.

In the fourth layer, the accuracy slightly recovers to 89%, but the Pearson correlation increases to 0.917. This high correlation signals significant overlap among neuron activations, indicating that further depth contributes little to class separation. Here, the FSAM graph reveals that, despite achieving correct classifications, the model no longer fully preserves the semantic structure of class-specific features. This scenario, where the model's predictions remain accurate without robust semantic alignment, highlights FSAM's diagnostic capability in identifying when a network may be "right for the wrong reasons".

These experiments effectively demonstrate FSAM's capacity to represent network behaviour across diverse configurations. Specifically, the FSAM activation graph tends to exhibit a more substantial alignment with the semantic structure as the accuracy improves. The initial layers, such as the second, achieve higher accuracy with low correlation, showing adequate class distinction. Beyond this point, additional layers lead to diminished accuracy and increased neuron overlap, confirming FSAM's reliability in capturing the balance between model accuracy and class separation. These findings attest to FSAM's robustness and consistency in representing GNN behaviour across different depths. Furthermore, these findings support Contribution 4, where we identify instances in which the FSAM

graph quality declines even as accuracy improves, underscoring FSAM's value in diagnosing subtle discrepancies in the network's semantic coherence with the data. As in the table 2 it shows raw layer-wise metrics, while Table 3 provides statistical validation. Together, they confirm that: (1) FSAM reliably captures GNN behaviour (all correlations significant, $p < 0.05$); (2) accuracy declines are dataset-dependent (significant in CoauthorCS/Amazon Computers); and (3) semantic patterns strengthen with depth (Δcorr. up to +0.328, $p < 0.001$).
These findings support our mathematical formulation of FSAM, in which we defined layer-wise activation matrices $A^{(i)}$ and correlation-based measures to quantify semantic structure and class separation. The experimental findings confirm the theoretical expectation that, in a transductive node classification setting, oversmoothing manifests itself as an increase in neuron-neuron correlation values (ρ_{ij}) alongside a decline in predictive accuracy. The observed layer-wise patterns of initial accuracy gains with low correlation, followed by accuracy plateaus or declines with increasing correlation, are entirely consistent with our hypothesis that deeper layers erode class-discriminative activation patterns.

5.4. Comparison of Mistakes Across Communities for Each Dataset

In our analysis, communities are defined within individual GNN layers, where each neuron is a node in the FSAM coactivation graph and edges are weighted by Spearman's rank correlation of their activation patterns. Community detection (Louvain method) is applied separately to each layer's graph, producing clusters of neurons whose activations are strongly correlated for the same input graph. Although detection is performed per layer, we track the corresponding communities across depths by matching neurones with similar activation profiles, enabling us to observe how these clusters evolve. Communities are relevant here because they represent semantically related functional groups of neurons; changes in their size, cohesion, or separation with depth reveal how layer depth influences the organisation of learned representations, class-specific accuracy, and the onset of oversmoothing. This evolution is quantified in Table 4 and supports our findings in Table 2. The structure of the community is delineated in **Layer 1**, with minimal overlap of neurons between different fields. The community **C0** groups Machine Learning, Data Mining, NLP, and AI, while separate clusters represent **C1** for Theory, Programming Languages, and Software Engineering, **C2** for HCI, Robotics, Computer Vision, Computer Graphics, and Computer Networking, and **C3** for Databases and Information Retrieval. The mistake count 1318 reflects a relatively low level of classification errors, indicating that the network maintains well-defined boundaries between these communities. This structure aligns with high accuracy and low overlap in neuron activations, captured effectively by the FSAM graph. Upon analysing class-wise accuracy for this dataset in **Layer 1**, we observed that **C2**—comprising Human-Computer Interaction, Robotics, Computer Vision, and Computer Graphics —unexpectedly includes Computer Networking. Although the model placed Computer Networking within this group, **C2** is primarily centred on theoretical foundations and methodologies for software optimisation, suggesting that Computer Networking may not belong to this group. Upon analysing class-wise accuracy for this dataset in **Layer 1**, we observed that **C2**—comprising Human-Computer Interaction, Robotics, Computer Vision, and Computer Graphics —unexpectedly includes Computer Networking. Although the model placed Computer Networking within this group, **C2** is primarily centred on theoretical foundations and methodologies for software optimisation, suggesting that Computer Networking may not belong to this group. Our class accuracy representation Fig. 2 graph supports this observation, yet further evaluation is necessary to confirm the optimal alignment of community structures within the network.

In **Layer 2**, we observe an evolution in the community structure with HCI merging into Community **C0** (Machine Learning, Data Mining, NLP, AI, HCI), signaling the onset of activation overlap as fields with closer semantic ties cluster together. The mistake count increases to 1388, indicating a slight accuracy decline as neuron activations overlap between certain communities. Here, we represent 'Accuracy decline changes **A«** and **C»**' and 'Major community shift', respectively. This trend is captured in the FSAM, showing an increased correlation in activations, reflecting the blending of previously distinct class representations.
By **Layer 3**, further integration within the community structure occurs, with Theory joining Community **C0**, and a more refined clustering among Programming Languages and Software Engineering in Community **C1**. Mistakes continue to increase to 1410, signifying increased misclassifications as class boundaries blur. This layer also corresponds to a higher Pearson correlation, indicating substantial overlap in neuron activations. The FSAM graph effectively captures this oversmoothing, showing that the distinctiveness among communities is diminishing with

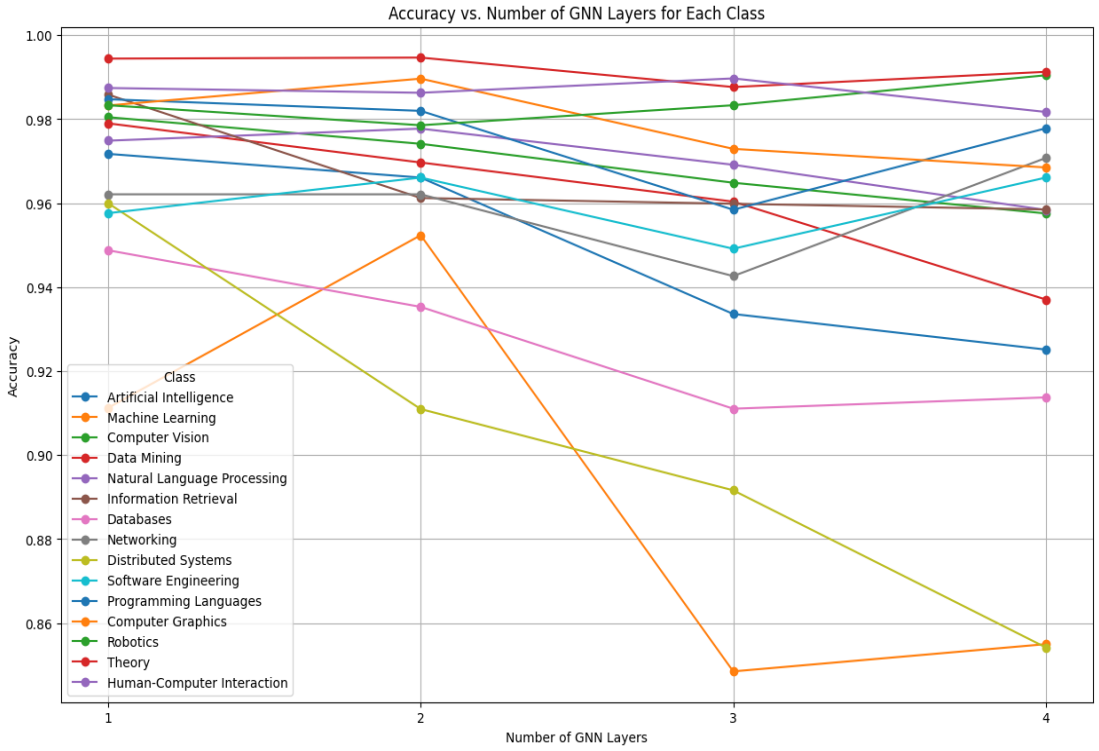

Fig. 2. Layer wise accuracy per class contribution of CoAuthorCs Dataset

deeper layers.

These results demonstrate that the CoauthorCS community structure evolves as layers are added, with previously distinct class groupings merging in response to overlapping neuron activations. This trend highlights the limitations of deeper layers in maintaining class specificity and supports FSAM's capability to capture the network's shifting behaviour across layers. The increase in misclassification and Pearson's correlation values illustrates how FSAM serves as a diagnostic tool, accurately reflecting the trade-off between layer depth and community coherence, thus validating the results shown in Table 2.

The analysis of errors between communities in the Amazon Photos dataset, as outlined in Table 2, provides valuable information on how community structures evolve across layers and impact model performance. This breakdown demonstrates FSAM's capability to capture structural shifts as the network depth increases, highlighting changes in how the model perceives class similarities.

In **Layer 1**, communities are separated, with distinct groups: **C0** (Cameras, Lenses, Camera Bags), **C1** (Memory Cards, Flashes, Batteries), and **C2** (Accessories, Tripods). The mistake count here is moderate, indicating that the model retains effective class distinction at this initial layer, with minimal overlap in neuron activations across communities.

As we progress to **Layer 2**, the network restructures communities, with **C0** narrowing its focus to Cameras and Lenses. In contrast, **C1** broadens to encompass Camera Bags, Memory Cards, Flashes, and Batteries. This reorganisation corresponds to a slight reduction in mistakes, suggesting that the network's representation has improved in distinguishing between these communities, with FSAM accurately reflecting the adjusted relationships among class representations.

However, the model's performance deteriorates in **Layer 3**, significantly increasing the mistake count. Communities become less distinct, as seen with **C0** now containing Memory Cards, Lenses, Flashes, Batteries, and Camera Bags. This expansion points to an increased overlap in neuron activations, aligning with a higher misclassification rate, which FSAM effectively captures by illustrating blurred boundaries between communities.

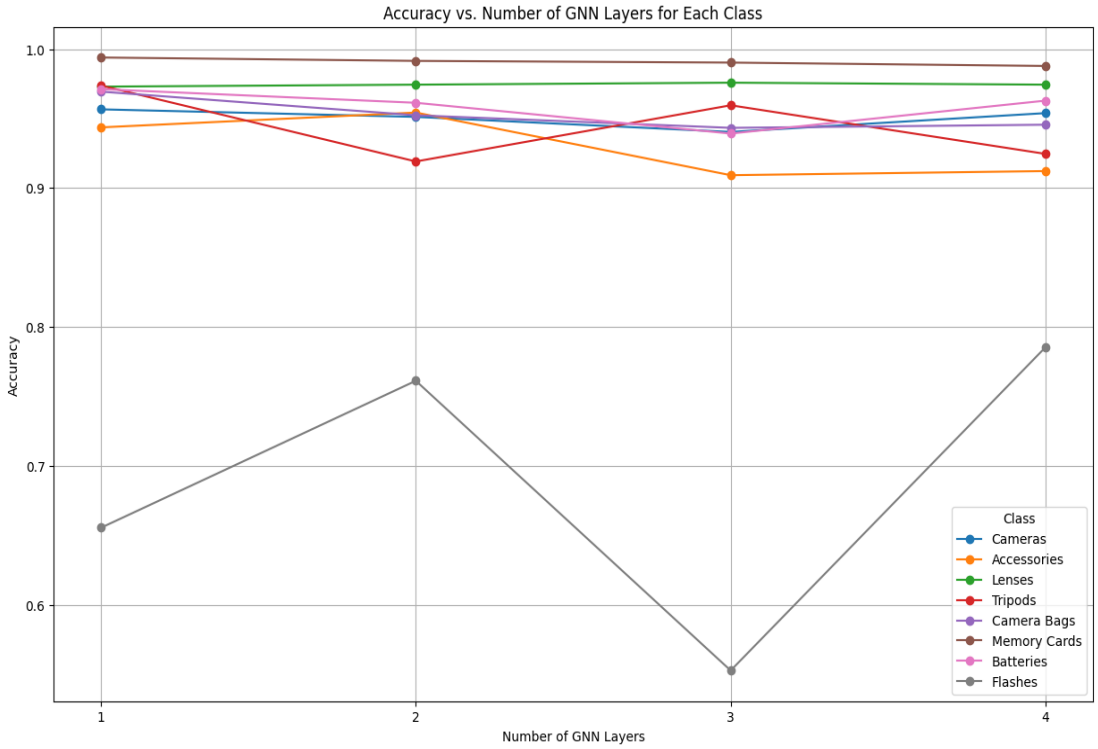

Fig. 3. Layer wise accuracy per class contribution of AmazonPhoto Dataset

By **Layer 4**, the network exhibits signs of oversmoothing, where distinctions between communities become less clear. Although the mistake count decreases slightly, this improvement may be misleading as FSAM reveals considerable overlap among communities. In this layer, **C0** isolates to represent Cameras alone, while **C1** groups Flashes, Tripods, and Camera Bags, and **C2** encompasses a diverse mix of Accessories, Memory Cards, Batteries, and Lenses. Indicates that, although errors may decrease, the underlying community distinctions are weakened, suggesting that the model may achieve accuracy without a robust semantic foundation.

This layerwise community analysis, as detailed in Table 2, demonstrates that FSAM not only reflects accuracy trends but also captures the nuanced structural shifts within the model as depth increases, reinforcing its utility in diagnosing when additional layers may lead to diminished class coherence.

In Table 4, the Amazon Computers dataset's analysis effectively captures effective shifts in network behaviour across different layers, especially in cases where accuracy trends diverge from FSAM correlation trends. Changes in community structures and pattern of mistakes in layers demonstrate this.

In **Layer 1**, the FSAM community structure exhibits clear distinctions: **C0** groups components like "Mice" and "Speakers," **C1** includes more complex devices such as "Desktops" and "Laptops," and **C2** contains "Monitors" and "Electronics." The mistake count in this layer is relatively moderate (452), indicating that the network maintains distinct activations with reasonable classification performance. This structured community alignment suggests a strong class separation in the network's internal representation.

At **Layer 2**, there is a noticeable change in the structure of the community. Products such as "Keyboards" and "Mice" migrate from **C1** to **C0**, as denoted by the significant labels $^{A»}$ and $^{C»}$. Interestingly, the accuracy improves in this layer and the error count decreases to 410. Although this reflects enhanced model performance, it also marks a case where accuracy improvements do not entirely align with FSAM's correlation trends. The slight decline in FSAM correlation indicates that the model may be achieving correct classifications without fully distinct semantic representations, an instance of potentially achieving the "right answer for the wrong reason." This scenario suggests that the network's internal representation might not be entirely aligned with the semantic structure of the input data,

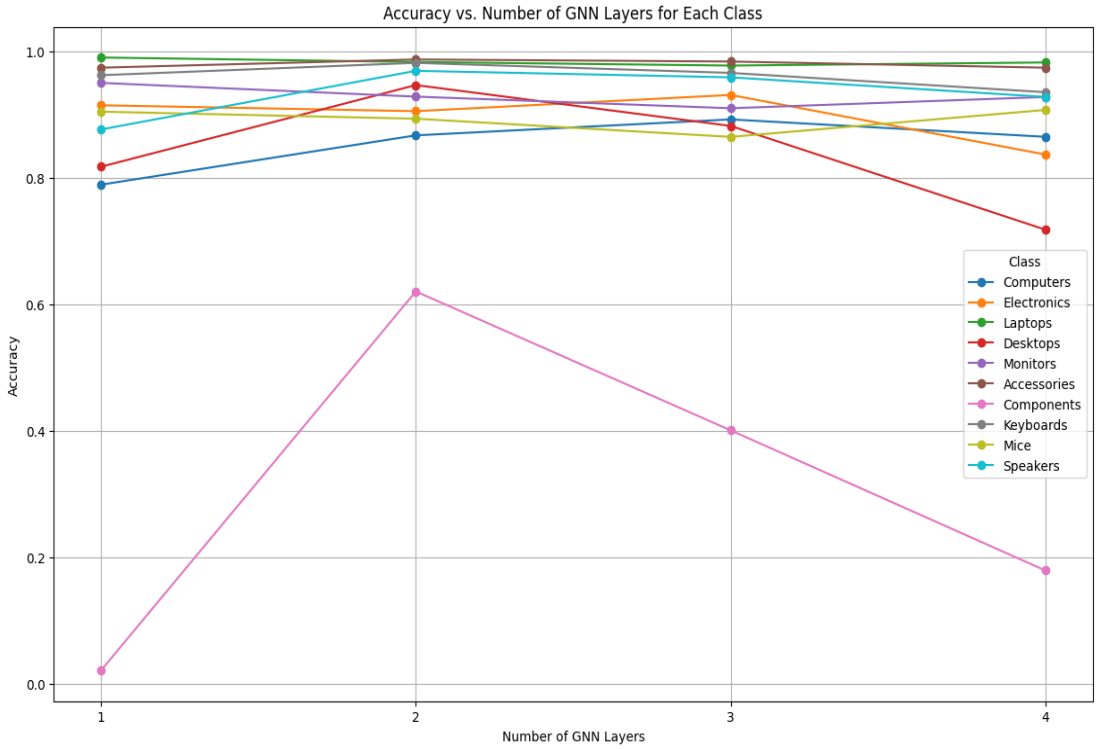

Fig. 4. Layer wise accuracy per class contribution of Computers Dataset

even as its accuracy improves.

Moving to **Layer 3**, accuracy begins to decline, with a further reduction in mistake count to 397. FSAM's community structure reveals additional overlap within **C0**, now encompassing "Speakers," "Laptops," and "Keyboards" in close association, which suggests diminished class distinctions. The corresponding increase in Pearson's correlation in this layer implies a more significant overlap in neuron activations, indicative of oversmoothing. While the network's classification ability is maintained, the underlying activations are less reflective of clear semantic boundaries, indicating a potential alignment misalignment.

By **Layer 4**, accuracy decreases, and the mistake count rises to 406. FSAM reveals that **C0** now includes a mix of "Desktops," "Speakers," and "Laptops," signifying an even greater overlap between distinct product categories. The increase in Pearson correlation and decreased accuracy indicates that additional layers now degrade the model's class-separation capability, aligning with FSAM's observation of blurred distinctions in class-specific representations. This combined result demonstrates that the added depth diminishes the network's ability to maintain semantic coherence within the deeper layers.

These findings substantiate our hypothesis by demonstrating the ability of FSAM to capture alignment and divergence between accuracy and semantic quality in GNNs. As seen in **Layer 2**, where accuracy improves but FSAM correlation declines, FSAM provides critical insight by identifying potential misalignments in the network's internal representations. Conversely, in **Layer 4**, where both the accuracy and quality of FSAM degrade, FSAM effectively reflects the reduced class-specific representation, underscoring its utility as a diagnostic tool to evaluate GNN behaviour in layers.

5.5. Layerwisewise Class Similarity and Misclassification Analysis

Our analysis reveals FSAM's unique ability to detect discrepancies between accuracy metrics and semantic understanding. In particular, we observe two critical scenarios: (1) cases where improving accuracy coincides with deteriorating FSAM graph quality, suggesting that the model achieves correct predictions without proper semantic

Table 4

Community Structure and Mistakes Across Layers for Each Dataset, where **A»**, **A«**, and **C»** represent 'accuracy increased', 'accuracy decreased' and 'community improvement', respectively. Mistakes are presented as both absolute counts and percentages.

Dataset	**Layer**	**Community (Classes)**	**Mistakes** (Absolute)	**Mistakes (%)**
CoauthorCS	1	**C0**: Machine Learning, Data Mining, NLP, AI; **C1**: Theory, Programming Languages, Software Engineering; **C2**: HCI, Robotics, Computer Vision, Computer Graphics, Computer Networking; **C3**: Databases, Information Retrieval.	1318	5.2%
	2	**C0**: Machine Learning, Data Mining, NLP, AI, HCI; **C1**: Theory, Programming Languages, Software Engineering; **C2**: Robotics, Computer Vision, Computer Graphics, Computer Networking; **C3**: Databases, Information Retrieval.	$1388^{A«,C»}$	5.5%
	3	**C0**: NLP, AI, HCI, Machine Learning, Data Mining, Theory; **C1**: Programming Languages, Software Engineering; **C2**: Robotics, Computer Vision, Computer Graphics, Computer Networking; **C3**: Databases, Information Retrieval.	1410	5.6%
	4	**C0**: AI; **C1**: Networking, Computer Graphics, Information Retrieval, Distributed Systems, Databases; **C2**: Machine Learning, Theory, HCI, Data Mining, NLP, Computer Vision, Robotics, Programming Languages; **C3**: Software Engineering.	1542	6.1%
Amazon Photos	1	**C0**: Cameras, Lenses, Camera Bags; **C1**: Memory Cards, Flashes, Batteries; **C2**: Accessories, Tripods	452	4.7%
	2	**C0**: Cameras, Lenses; **C1**: Camera Bags, Memory Cards, Flashes, Batteries; **C2**: Accessories, Tripods	$410^{A»,C»}$	4.2%
	3	**C0**: Memory Cards, Lenses, Flashes, Batteries, Camera Bags; **C1**: Accessories, Tripods, Cameras	497	5.1%
	4	**C0**: Cameras; **C1**: Flashes, Tripods, Camera Bags; **C2**: Accessories, Memory Cards, Batteries, Lenses	406	4.2%
PubMed	1	**C0**: Cardiovascular Disease, Diabetes; **C1**: Breast Cancer	46	3.7%
	2	**C0**: Cardiovascular Disease, Diabetes; **C1**: Breast Cancer	38	3.0%
	3	**C0**: Breast Cancer; **C1**: Cardiovascular Disease, Diabetes	$32^{A»,C»}$	2.5%
	4	**C0**: Breast Cancer; **C1**: Cardiovascular Disease, Diabetes	62	4.9%
Cora	1	**C0**: Case-Based, Neural Networks, Genetic Algorithms; **C1**: Theory; **C2**: Reinforcement Learning, Probabilistic Methods	220	6.1%
	2	**C0**: Case-Based, Genetic Algorithms; **C1**: Reinforcement Learning, Rule Learning, Probabilistic Methods; **C2**: Neural Networks, Theory	356	9.9%
	3	**C0**: Neural Networks, Theory; **C1**: Case-Based, Rule Learning, Genetic Algorithms; **C2**: Reinforcement Learning, Probabilistic Methods	$364^{A»,C»}$	10.0%
	4	**C0**: Case-Based, Neural Networks, Probabilistic Methods, Theory; **C1**: Genetic Algorithms, Reinforcement Learning, Rule Learning	378	10.4%
AmazonComputers	1	**C0**: Components, Mice, Speakers; **C1**: Desktops, Laptops, Keyboards, Computers, Accessories; **C2**: Monitors, Electronics	452	5.6%
	2	**C0**: Keyboards, Components, Mice, Speakers ; **C1**: Desktops, Laptops, Computers, Electronics; **C2**: Monitors, Accessories	$410^{A»,C»}$	5.0%
	3	**C0**: Speakers, Laptops, Keyboards, Components, Mice, Accessories; **C1**: Desktops, Monitors, Electronics, Computers	397	4.9%
	4	**C0**: Desktops, Speakers, Laptops, Keyboards, Computers, Components, Accessories; **C1**: Monitors, Electronics, Mice	406	5.0%

grounding, and (2) situations where reduced accuracy accompanies enhanced FSAM quality, potentially indicating more meaningful learning from misclassifications. These findings demonstrate FSAM's value in model-level diagnostic tools to identify when models are "right for the wrong reasons," offering crucial insights into the semantic alignment between networks and their training data.

To systematically validate these observations regarding layer depth and class similarity effects, we present multiple visualisation strategies that capture fundamental aspects of model behaviour.

- **Per-Class Accuracy vs. Number of GCN Layers**: By examining class-specific accuracy across layers, this figure reveals which classes experience increased misclassifications as layer depth grows, underscoring the model's reduced capacity to maintain distinct representations for these categories as shown in figure 2 , 3, and 4.
- **FSAM Graphs Showing Neuron Activations for Specific Classes**: These graphs show neuron activation patterns within each class, allowing us to track the ability of the GNN to capture class-specific characteristics across layers. They reveal the points at which neuron activations overlap, indicating where class boundaries lose distinctiveness, as shown in figure 13, 14, 15, and 16.
- **Community Structures Highlighting Class Groupings**: This visualisation illustrates the community structures of neuron activations, clustering classes based on coactivation. These clusters indicate the relationship between certain classes and provide insight into the knowledge organisation of GNN, revealing where class separability degrades with additional layers, as detailed in table 4.
- **Jaccard Coefficient vs. Number of Mistakes at Layer 3** presents the Jaccard similarity between misclassifications for class pairs, demonstrating a positive correlation between high similarity in neuron activation overlaps and error rates. This relationship supports our observation that classes with more significant overlap in FSAM exhibit more frequent misclassifications, as shown in the figure 5, 6, and 7.

Our extended analysis revealed a positive correlation between class similarity and the number of mistakes involving them, as illustrated with examples from the datasets **CoauthorCS** and **Amazon Photos** datasets. Table 2 shows that class pairs with higher overlap in the FSAM graph exhibit more misclassifications. Our findings in the **CoauthorCS dataset** reveal that Layer 1 achieves optimal performance, as demonstrated in. Adding further layers decreases accuracy, corroborated by our FSAM graph analysis. The Jaccard similarity at Layer 2 aligns with this trend, indicating that increased depth introduces more overlap in neuron activations, which diminishes the model's ability to distinguish between closely related fields such as *Machine Learning* and *Data Mining*. Grouped within the same community, these fields are prone to misclassification due to their inherent similarity.

A similar trend appears in the **Amazon Photos dataset**, where accuracy increases from Layer 1 to Layer 2 but declines with additional layers. This pattern, shown in the Table 2, 4 is consistent with our Jaccard similarity analysis at Layer 3. In this layer, product categories such as *Memory Cards* and *Accessories* show high Jaccard similarity, resulting in frequent misclassifications due to overlapping neuron activations. This finding indicates that the GNN model faces challenges in distinguishing between these similar classes, as they share substantial overlap of activation within the same community.

These findings suggest that adding layers beyond an optimal depth does not necessarily improve knowledge representation. Instead, it introduces an oversmoothing effect in which neuron activations for different classes become increasingly indistinct, reducing the model's ability to differentiate between them. Our correlation analysis substantiates this effect, which shows that pairs of classes with significant overlap in the FSAM graph tend to experience higher misclassification rates.

Our analysis of community structures aligns with this observation, allowing us to identify classes that the GNN perceives as similar based on FSAM patterns. By examining the Jaccard similarity coefficient, which quantifies the overlap in neuron activations for each pair of classes, we evaluated the impact of these similarities on GNN decision making. In the Amazon Photos dataset, for example, product categories such as *Memory Cards* and *Accessories* displayed high Jaccard similarity, leading to frequent misclassifications. For a detailed analysis of the misclassification rates per class pair for each community across different layers, we have presented the results for all layers in Figures 8, 9, 10, and 11. Furthermore, we have conducted more experiments to calculate the time for each layer in different datasets as detailed in Table 5

5.6. Actionable Insights for GNN Performance Improvement

These insights suggest that tuning efforts should reduce overlap in the coactivation graph for similar classes to enhance the GNN's ability to differentiate between them. Targeting overlapping nodes can potentially decrease misclassification rates and improve overall model accuracy. This comprehensive evaluation supports our hypothesis that increasing layers does not necessarily yield better performance and, in some instances, may decrease the discriminative power of the model due to overlapping neuron activations. Insights from our FSAM analysis suggest that overlap in coactivations for similar classes is a critical issue that affects the GNN's ability to differentiate between those classes. By targeting and reducing this overlap, we can improve the model's ability to distinguish between similar courses, thus reducing misclassification rates and improving overall accuracy. FSAM provides a framework for identifying the overlapping nodes in the coactivation graph. By addressing these overlaps, we can potentially reduce the adverse effects of oversmoothing, thus enhancing model performance and ensuring that the GNN can retain useful discriminative features across layers. Although FSAM does not directly generate explanations, the insights gained from mapping neuron activations and analysing coactivation overlap can pave the way for explanation-generation methods. These insights could ultimately contribute to more human-interpretable explanations of GNN decision making, especially in safety critical applications.

Table 5
Execution Times for Different Datasets and Number of Layers

Dataset	**Layer 1**	**Layer 2**	**Layer 3**	**Layer 4**	**Layer 5**
AmazonComputers	1.2	1.24	1.23	1.19	1.49
AmazonPhoto	0.77	0.97	1.07	1.02	0.96
CoauthorCS	0.79	0.94	0.92	0.9	0.89
Cora	0.79	0.94	0.91	0.89	0.89
CiteSeer	0.81	0.94	0.91	0.89	0.88
PubMed	0.78	0.95	1.09	1.04	1.01

5.7. Scalability and Computational Complexity

While FSAM offers valuable insights into GNN's semantic coherence, its application to large-scale graphs faces computational challenges, particularly in higher-dimensional feature space, as it requires more memory and processing power as it monitors neuronal activations across multiple layers and conducts correlation-based analyses as graph size and model depth increase. The complexity of tracking activations and conducting semantic correlation analysis increases non-linearly with the increase in nodes and edges. More complex GNN architectures exacerbate this issue by introducing additional activation patterns, increasing memory usage and bottleneck, and prolonging the computation duration. Optimisations such as node and edge sampling, parallelisation techniques (e.g., GPU acceleration, distributed computing), and approximate correlation methods can mitigate these challenges by enhancing efficiency. Future research should explore lightweight FSAM variants, such as selective neuronal tracking layerwise approximations, to enhance scalability while preserving interpretability. The practical implementation of FSAM relies on resolving these challenges for large-scale graph datasets.

6. Conclusion and Future Work

In this extended study, we have worked to deepen the understanding of how GNNs behave using FSAM to examine the link between model depth, performance, and semantic representation. Through experiments on several datasets, we found that FSAM consistently captures meaningful semantic relationships across different contexts, reinforcing its reliability as a tool for interpreting network behaviour. Our findings also indicate that adding more layers to GNNs does not always lead to better performance or richer knowledge representation. In these FSAM

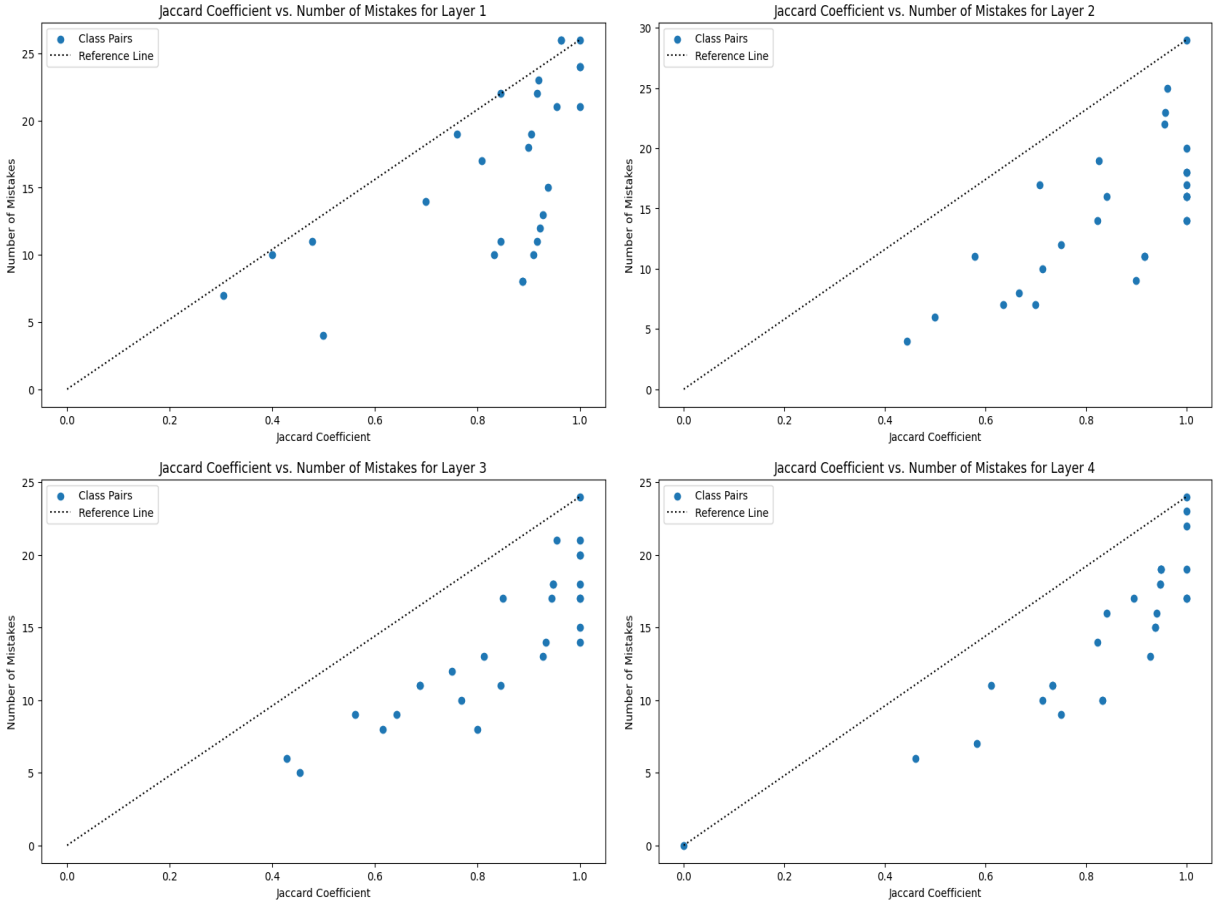

Fig. 5. Jaccard Similarity between different layers for AmazonPhoto dataset

graphs, nodes represent neurons, and weighted edges indicate the strength of their coactivation relationships, reflecting correlations in activation patterns across layers. This layered view of the GNN function shows how neurons contribute to specific class predictions and influence overall model decisions. Our experiments confirmed that the FSAM graph structure aligns well with the knowledge stored in GNNs, especially in distinguishing closely related classes. In all data sets, FSAM consistently highlighted key neurons and communities within the GNN that are central to specific class predictions, providing valuable insights into the model decision making process. We used community detection in FSAM graphs to see how the GNN naturally groups classes based on activation patterns. Our analysis showed that courses with high overlap in the FSAM graph are more likely to be misclassified, suggesting that focusing on these overlapping nodes could help fine-tune the model and improve accuracy. This ability to identify cases where accuracy may be achieved 'for the wrong reasons' where predictions are correct but lack deep semantic alignment highlights the diagnostic power of FSAM. The FSAM graphs and community detection further clarify how the GNN organises knowledge, revealing class groups with high activation overlap that the GNN treats as similar. This overlap is often associated with higher misclassification rates, supporting strategies to reduce this overlap and improve the model's ability to distinguish between classes.

For future work, we will explore whether similar interpretability degradation occurs when increasing hidden-layer dimensionality rather than depth, and how FSAM patterns evolve throughout training. This includes analysing loss trajectories and probing connections to phenomena such as grokking. We also aim to develop adaptive depth GNN architectures that self-tune based on graph structure, refine FSAM's class-level evaluation metrics, and integrate FSAM insights with graph topology for richer and context-aware explanations. Advancing FSAM in these direc-

tions could transform it from a post hoc diagnostic into a proactive design tool, guiding the creation of GNNs that are not only accurate but also efficient, interpretable, and semantically aligned.

7. Acknowledgement

This work was conducted with the financial support of the Science Foundation. Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

[1] Steve Azzolin et al. “Global explainability of gnns via logic combination of learned concepts”. In: *arXiv preprint arXiv:2210.07147* (2022).
[2] Federico Baldassarre and Hossein Azizpour. “Explainability techniques for graph convolutional networks”. In: *arXiv preprint arXiv:1905.13686* (2019).
[3] T. O. Botari et al. “Gene expression-based classification of diffuse large B-cell lymphoma”. In: *Nature* (2002), pp. 261–268.
[4] David J Tena Cucala and Bernardo Cuenca Grau. “Bridging max graph neural networks and datalog with negation”. In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 21. 1. 2024, pp. 950–961.
[5] Thorben Funke, Megha Khosla, and Avishek Anand. “Hard masking for explaining graph neural networks”. In: *Advances in neural information processing systems* (2020).
[6] Robert Geirhos et al. “Generalisation in humans and deep neural networks”. In: *Advances in neural information processing systems* 31 (2018).
[7] Qiang Huang et al. “Graphlime: Local interpretable model explanations for graph neural networks”. In: *IEEE Transactions on Knowledge and Data Engineering* (2022).
[8] Thomas N Kipf and Max Welling. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).
[9] Meng Liu, Hongyang Gao, and Shuiwang Ji. “Towards deeper graph neural networks”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 338–348.
[10] Dongsheng Luo et al. “Parameterized explainer for graph neural network”. In: *Advances in neural information processing systems* 33 (2020), pp. 19620–19631.
[11] Julian McAuley et al. “Image-based recommendations on styles and substitutes”. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2015, pp. 43–52.
[12] Galen Namata et al. “Query-driven active surveying for collective classification”. In: *10th International Workshop on Mining and Learning with Graphs (MLG)* (2012).
[13] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. “Feature visualization”. In: *Distill* 2.11 (2017), e7.
[14] Phillip E Pope et al. “Explainability methods for graph convolutional neural networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10772–10781.
[15] Kislay Raj and Alessandra Mileo. “Towards Understanding Graph Neural Networks: Functional-Semantic Activation Mapping”. In: *International Conference on Neural-Symbolic Learning and Reasoning*. Springer. 2024, pp. 98–106.
[16] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. “A survey on oversmoothing in graph neural networks”. In: *arXiv preprint arXiv:2303.10993* (2023).
[17] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. “Interpreting graph neural networks for nlp with differentiable edge masking”. In: *arXiv preprint arXiv:2010.00577* (2020).
[18] Thomas Schnake et al. “Higher-order explanations of graph neural networks via relevant walks”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.11 (2021), pp. 7581–7596.
[19] Robert Schwarzenberg et al. “Layerwise relevance visualization in convolutional text graph classifiers”. In: *arXiv preprint arXiv:1909.10911* (2019).
[20] Prithviraj Sen et al. “Collective classification in network data”. In: *AI magazine* 29.3 (2008), pp. 93–93.
[21] Oleksandr Shchur et al. “Pitfalls of Graph Neural Network Evaluation”. In: *Relational Representation Learning Workshop, NeurIPS 2018*. 2018.
[22] Petar Velickovic et al. “Graph attention networks”. In: *stat* 1050.20 (2017), pp. 10–48550.
[23] Minh Vu and My T Thai. “Pgm-explainer: Probabilistic graphical model explanations for graph neural networks”. In: *Advances in neural information processing systems* 33 (2020), pp. 12225–12235.
[24] Xiang Wang et al. “Causal screening to interpret graph neural networks”. In: (2020).
[25] Zonghan Wu et al. “A comprehensive survey on graph neural networks”. In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.
[26] Weiting Xi et al. “A Graph Partitioning Algorithm Based on Graph Structure and Label Propagation for Citation Network Prediction”. In: *International Conference on Knowledge Science, Engineering and Management*. Springer. 2023, pp. 289–300.

[27] Xiaoran Yan et al. “Weight thresholding on complex networks”. In: *Physical Review E* 98.4 (2018), p. 042304.
[28] Pinar Yanardag and SVN Vishwanathan. “Deep graph kernels”. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 1365–1374.
[29] Zhitao Ying et al. “Gnnexplainer: Generating explanations for graph neural networks”. In: *Advances in neural information processing systems* 32 (2019).
[30] Zhaoning Yu and Hongyang Gao. “Mage: Model-level graph neural networks explanations via motif-based graph generation”. In: *arXiv preprint arXiv:2405.12519* (2024).
[31] Hao Yuan and Shuiwang Ji. “Structpool: Structured graph pooling via conditional random fields”. In: *Proceedings of the 8th International Conference on Learning Representations*. 2020.
[32] Hao Yuan et al. “On explainability of graph neural networks via subgraph explorations”. In: *International conference on machine learning*. PMLR. 2021, pp. 12241–12252.
[33] Hao Yuan et al. “Xgnn: Towards model-level explanations of graph neural networks”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 430–438.
[34] Yue Zhang, David Defazio, and Arti Ramesh. “Relex: A model-agnostic relational model explainer”. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 1042–1049.

8. Appendix

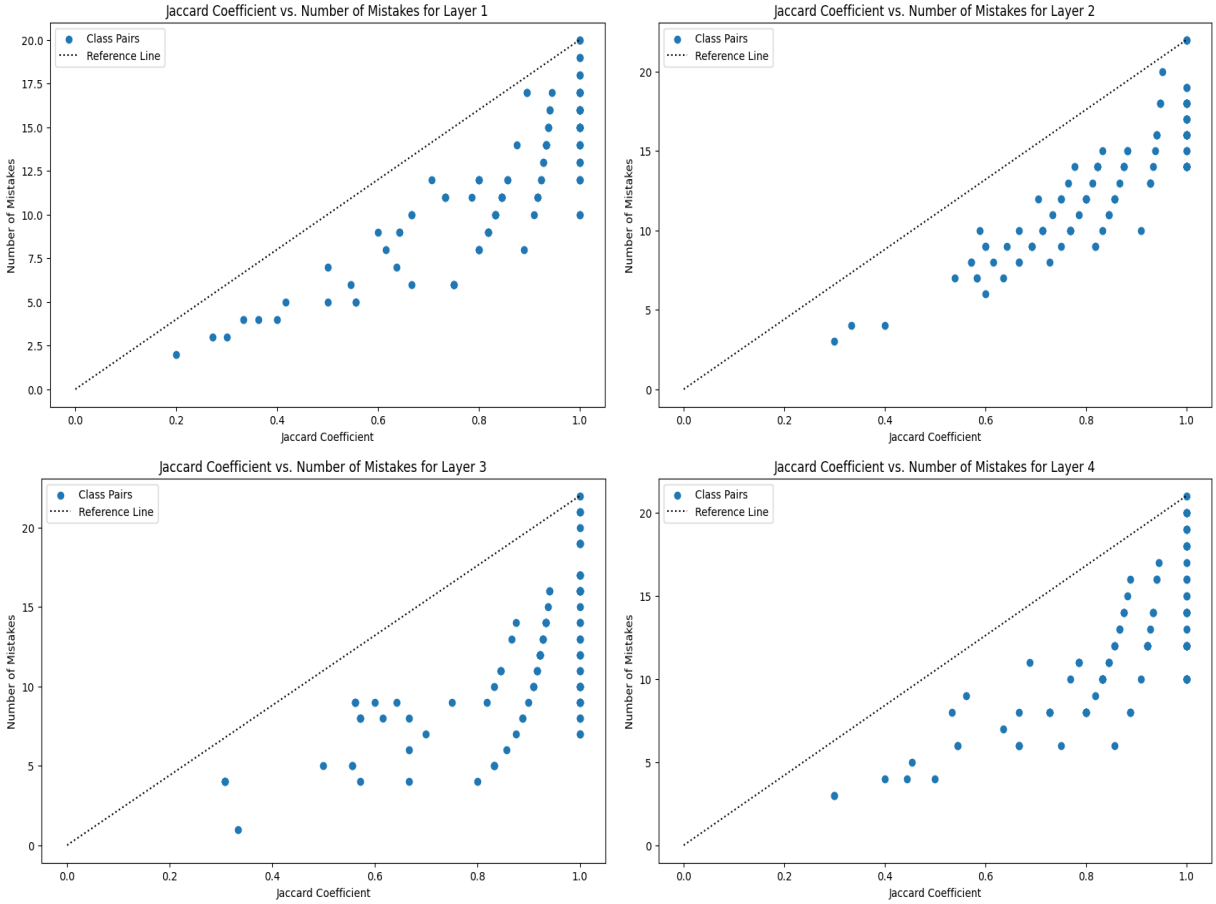

Fig. 6. Jaccard Similarity between different layers for CoauthorCs

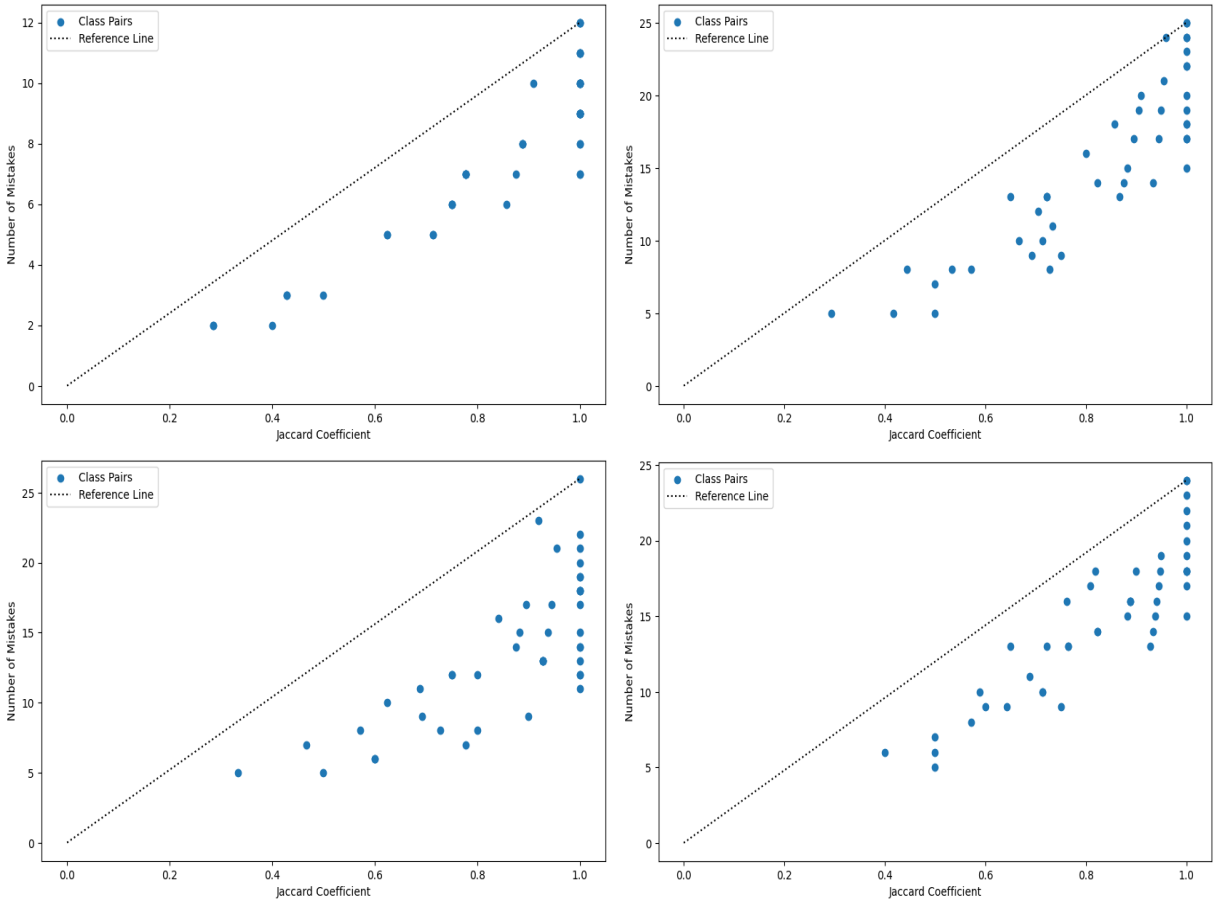

Fig. 7. Jaccard Similarity between different layers for Computers

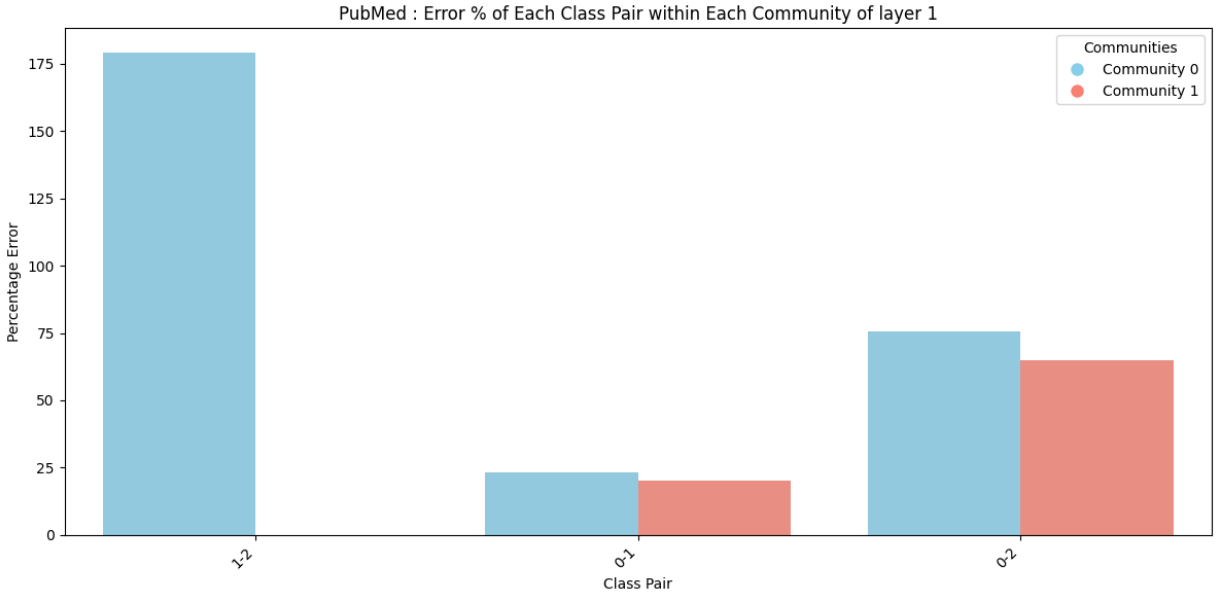

Fig. 8. Per-class misclassification rates in each community between classes pair in layer 1

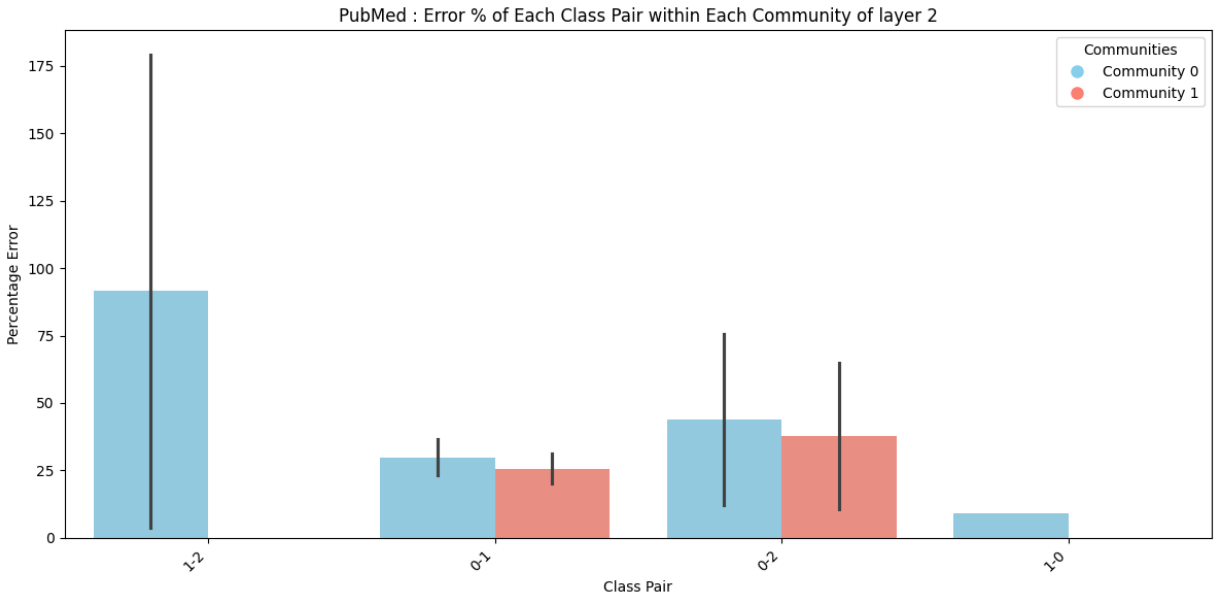

Fig. 9. Per-class misclassification rates in each community between classes pair in layer 2

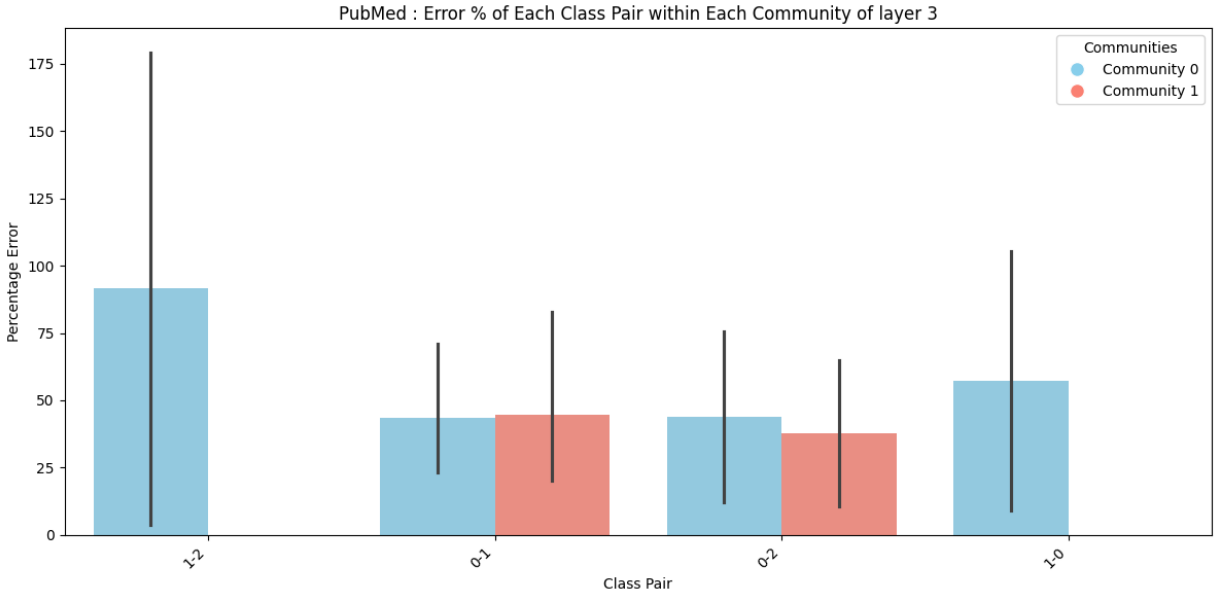

Fig. 10. Per-class misclassification rates in each community between classes pair in layer 3

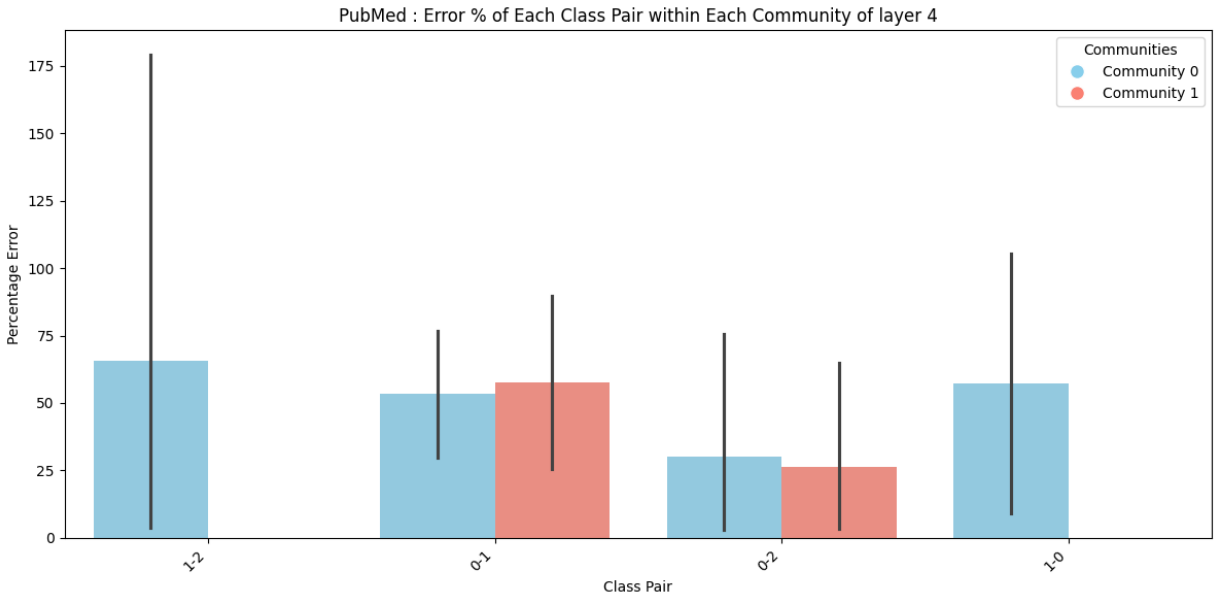

Fig. 11. Per-class misclassification rates in each community between classes pair in layer 4

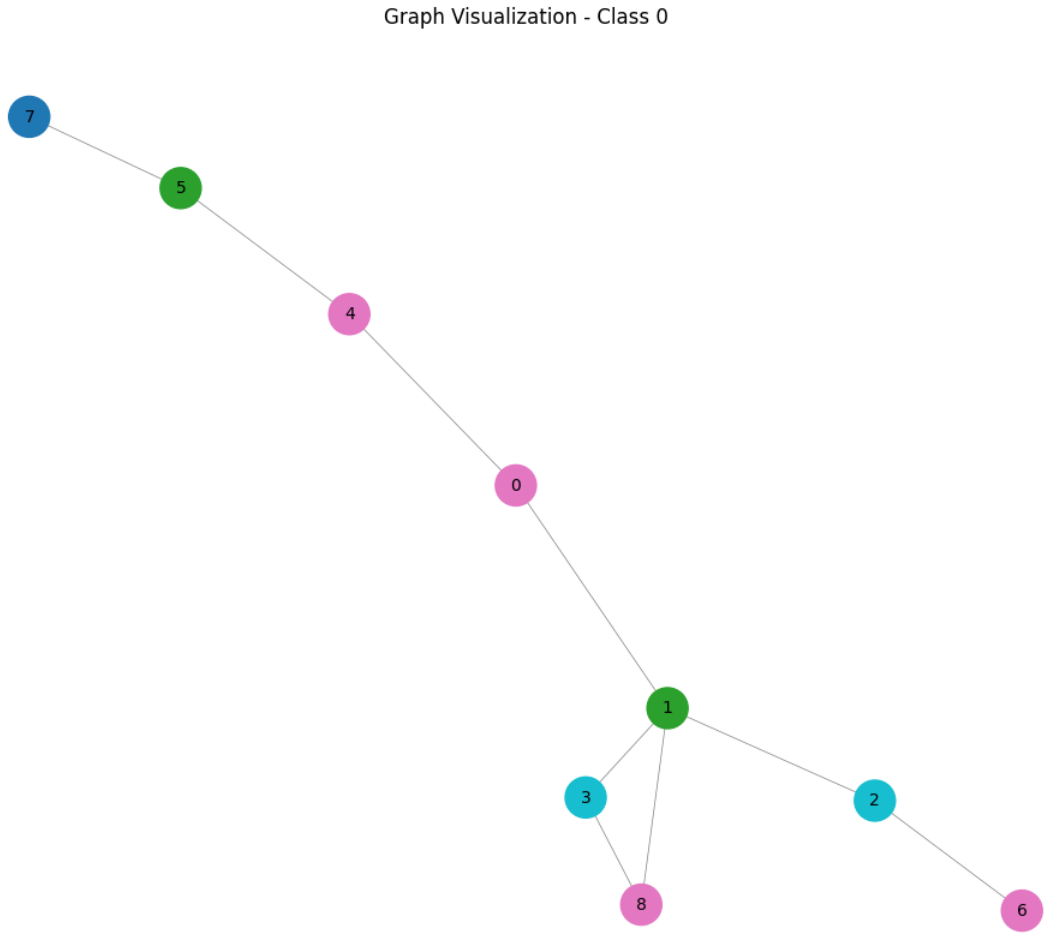

Fig. 12. XGNN visualisation for Class 0 of the Cora Dataset

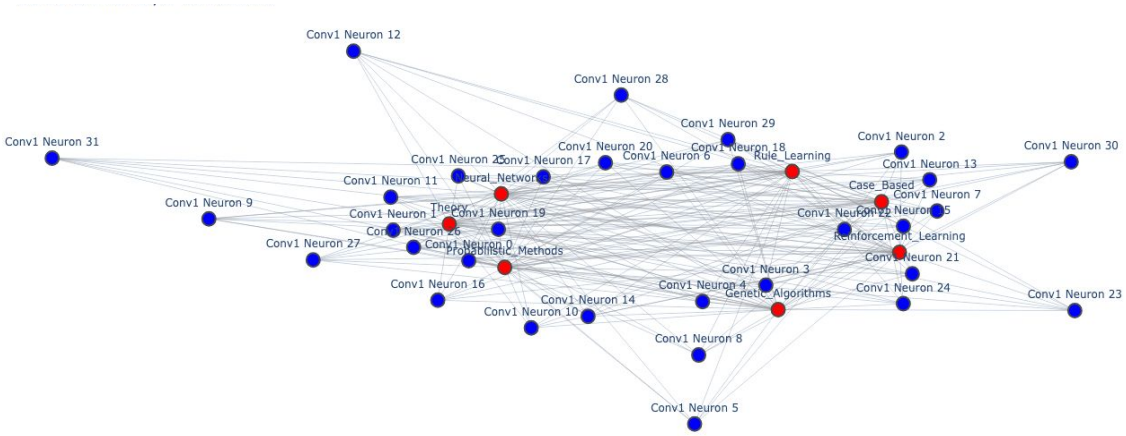

Fig. 13. Layer 1 to the Fully Connected FSAM Graph for the Cora Dataset

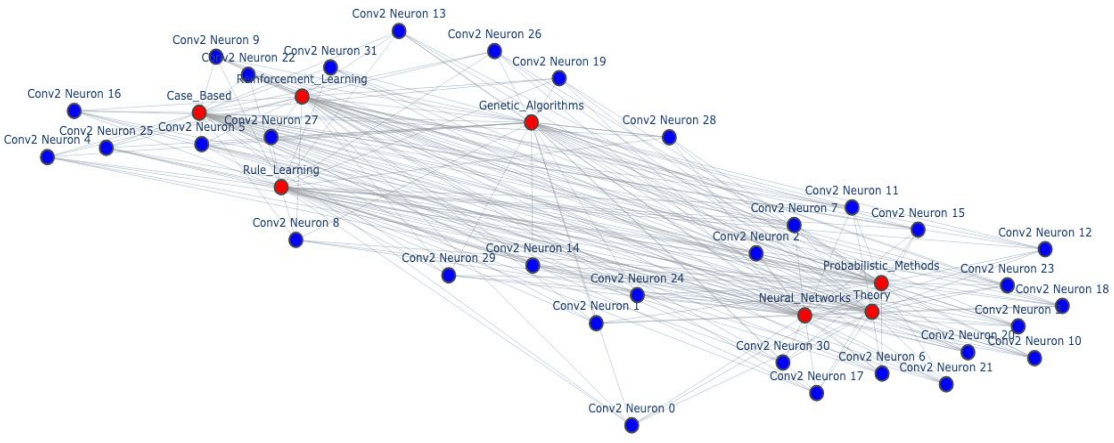

Fig. 14. Layer 2 to the Fully Connected FSAM Graph for the Cora Dataset

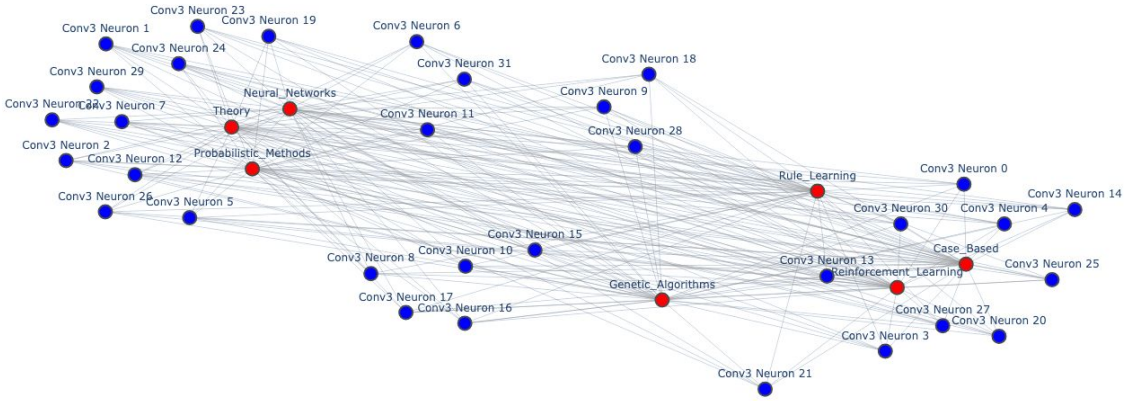

Fig. 15. Layer 3 to the Fully Connected FSAM Graph for the Cora Dataset

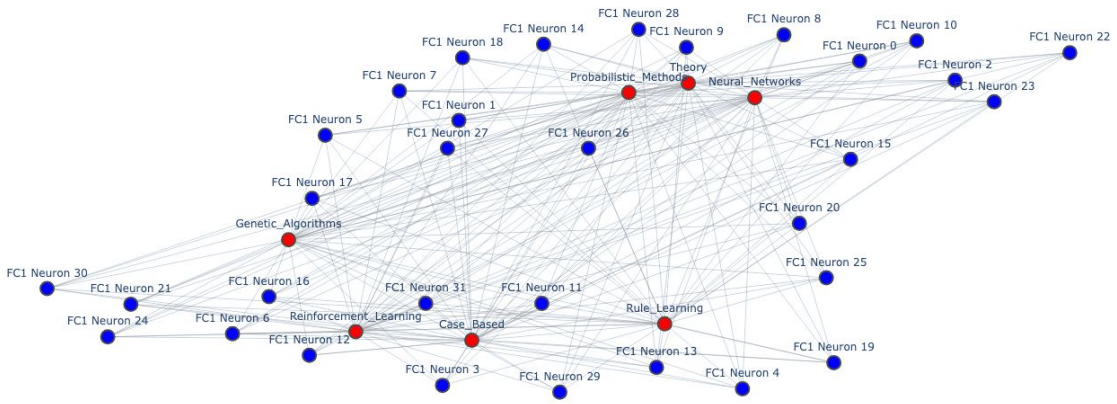

Fig. 16. Layer 4 to the Fully Connected FSAM Graph for the Cora Dataset