Towards Semantic Understanding of GNN Layers embedding with Functional-Semantic Activation Mapping

Kislay Raj [a,*] and Dr. 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 shown significant potential in learning representations from complex graph-structured data. However, explainability for GNN still relies mostly on identifying parts of the input graph that is relevant for a particular instance (local explanations) while the ability to understand the model global explanation is still underinvestigated. In addition to that, the impact of the GNN structure on the quality of the deep representation is not well understood and determining the best configuration of layers for a given problem is still a trial-and-error process. In this paper, we extend our previous work on Functional-Semantic Activation Mapping (FSAM) to investigate how changing the number of GNN layers affects the quality of the deep representation and as a result its performance. Through experiments on multiple datasets, we observe that while adding layers may enhance accuracy, it does not consistently lead to improved semantic representations; in some cases, performance increases while semantic quality declines, suggesting correct predictions for incorrect reasons. FSAM 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 emphasise that adding layers does not always improve the GNN's deep representation and may, in fact, hinder its ability to learn meaningful semantic distinctions. This work underlines the importance of assessing GNNs beyond accuracy alone, advocating for a deeper analysis of the effects of layer depth, specifically in relation to performance versus semantic coherence in model interpretations.

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

1. Introduction

Graph Neural Networks (GNNs) [1, 2, 3] have demonstrated remarkable performance across a wide array of tasks, including node classification, link prediction, and graph classification, across diverse datasets such as citation networks, molecular data, and social networks. GNNs leverage both the structural information and node features of graph data, enabling them to capture complex relationships in the deep representation. However, despite these advances, explaining GNN predictions remains an open challenge, primarily due to the intricate topological nature of graphs and the opaque way this is reflected or learned in GNN embeddings. Unlike traditional neural networks, where inputs have fixed structures (e.g., grids in images), GNNs operate on graph structures as input. This might

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

suggest there is a potential for GNN to have better interpretability as input features have meaningful, graph-like relations rather than pixel-based data. However, the challenge lies in understanding how these relationships are learned and embedded within the GNN layers. This ambiguity complicates the interpretation of learned embeddings, making it essential to understand model behaviour through detailed activation analysis.

Current research in GNN explainability has focused on generating local explanations by identifying small sets of nodes and edges from the input data that have influenced the outcome for a specific test data sample [4]. These type of explanations based on input features provide insights into the decision-making process of GNNs by highlighting the critical parts of the input graph which most affected the final predictions [5]. While effective at the local level, these methods do not provide a holistic view of the model's overall behaviour. A global understanding of GNNs remains underexplored, particularly in terms of how individual layers contribute to the final representation.

In our previous work [6], we introduced **Functional-Semantic Activation Mapping (FSAM)**, a method aimed at improving the global interpretability of GNNs by constructing functional-semantic graphs that capture neuron activations across all layers of the network. FSAM links neural activations to human-interpretable concepts (the output classes), thereby providing a deeper understanding of how different parts of the network contribute to the final predictions. Through this, we laid the groundwork for analysing the relationship between relevance of neuron activations and output classes within GNNs.

In this extended work, we address a critical question in GNN design: to what extent does adding an additional layer enhance the model's ability to represent network behaviour? And does improved performance always correspond to better deep representations?

Our findings indicate that adding layers does not necessarily yield a better deep representation; in some cases, even a single GNN layer can achieve strong performance and effectively capture the relations in the input data. As anticipated, our thorough evaluation of GNNs across multiple datasets shows that adding layers eventually leads to over-smoothing, where node embeddings become overly similar, reducing model accuracy and increasing misclassifications. This degradation in performance aligns with FSAM quality, particularly when the semantic structure of the input data is poorly preserved, highlighting how over-smoothing can compromise both classification performance and representation integrity.

However, there are cases where improved GNN performance does not correlate with enhanced quality in the activation graph generated by FSAM. These cases are particularly insightful: they reveal instances where GNN performance increases even as neuron activations fail to adequately reflect the input data's semantic structure, suggesting that the GNN achieves the correct outcome, albeit for the wrong reasons.

Given the approach in this paper relies on capturing the GNN's behaviour through activation analysis with FSAM, our first contribution is to extend the FSAM validation beyond our previous experiments on CORA [7] and CiteSeer [8]. To do that, we carry on additional experiments on four datasets: PubMed [9], Amazon Computers [10], Amazon Photos [10], and Coauthor [11]. These datasets, with their distinct topological complexities, allow us to comprehensively evaluate the FSAM's approach and how well the resulting activation graph reflects the behaviour of the GNN and indicates how well the semantic structure of the input data has been learned by the network.

The contribution of this work can be summarised as follows: **Firstly**, we extend the FSAM approach by conducting experiments across a broader range of datasets to validate that the activation analysis and graphs generated by FSAM consistently reflect the network's behaviour. This contribution includes community analysis across the additional datasets to demonstrate FSAM's capability to reliably capture the semantic structure between classes. **Secondly**, we extend our experimental analysis to validate that the functional activation graph generated by FSAM aligns with the network's behaviour as the number of layers changes. By examining networks with varying depths (from 1 to 4 layers) and comparing the correlation between misclassifications and class similarity, we demonstrate that improvements in network accuracy are reflected in the FSAM graph, and similarly, any decline in accuracy is captured by the FSAM structure. This analysis reinforces FSAM's capability to accurately represent network behaviour across different layer configurations. **Thirdly**, we conduct a detailed layer-by-layer analysis to assess the impact of varying GNN layer configurations on the model's performance in node classification tasks. Specifically, we evaluate how different numbers of layers influence the FSAM and corresponding community structures, verifying whether improvements in accuracy align with a better FSAM graphs and, conversely, whether decreases in accuracy correlate with a decline in FSAM quality. This analysis demonstrates that while additional layers may initially enhance performance, deeper layers can lead to over-smoothing and neuron activation overlap, ultimately

diminishing the model's discriminative power. In our 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 likely due to a better embedding of the semantic structure in the input data. We believe this indicates a potential for FSAM's in identifying situations in which a model makes accurate predictions for the wrong reasons.
This work demonstrates that FSAM not only serves as a reliable tool for aligning accuracy with the GNN's internal representations but also offers unique insights when discrepancies between accuracy and FSAM quality arise. These findings underscore the importance of optimising GNN layer depth to balance model complexity with representation quality, providing a valuable resource for researchers and practitioners aiming to improve GNN interpretability and performance in real-world applications.

2. State of the art (SOTA)

Interest in neurosymbolic AI has been steadily increasing, driven by the need for interpretable and accountable machine learning systems, especially in domains requiring transparent decision-making. Research has focused on integrating neural learning with symbolic reasoning, a vital step for enhancing the explainability of deep learning models. This integration is particularly important for high-stakes domains where both accuracy and interpretability are essential. GNNs have shown exceptional performance in handling graph-structured data across a range of fields, such as social networks [12], molecular structures [13], and citation networks [14]. However, despite their success in learning complex relationships, GNNs remain largely opaque in terms of 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 in relation to 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** [15], which use gradient information or hidden feature map values to assess feature importance; **Perturbation-based methods** [16], which modify graph structures and monitor how these perturbations affect model outputs; **Decomposition methods** [17, 15], which break down the prediction score into contributions from different neurons or layers, propagating these contributions backwards through the network; and **Surrogate methods** [18, 19], 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.
Although these methods provide valuable insights, they predominantly focus on instance-level predictions, failing to capture the GNN's global behaviour or how information is processed through the layers of the network. Local methods often highlight specific features without offering a comprehensive view of the entire decision-making process, which is essential for understanding the model's behaviour in relation to prior knowledge and domain-specific concepts. Additionally, these methods tend to provide explanations that are difficult for humans to interpret, as they do not reveal the underlying relationships between the GNN's learned representations and the data's inherent structure. The non-grid structure of graphs further complicates the application of techniques traditionally used in image or text domains, making direct adaptation infeasible [4].
Global explanations are comparatively underexplored in GNN research. One notable method is **XGNN** [20], which generates synthetic graphs optimised for class predictions to explain the behaviour of the GNN. However, XGNN's assumption that a single synthetic graph can represent an entire class oversimplifies the complex relationships within real-world datasets. Such approaches, while offering some insight into the final predictions, fail to account for how intermediate layers contribute to the learned representations. They often focus on the GNN's output rather than explaining the interplay between graph components—nodes, edges, and their interactions—across layers. This approach is inadequate for providing human-understandable insights into how the model's internal workings relate to prior knowledge or domain-specific information, making it difficult for users to trust and interpret the model's decisions. Moreover, relevant SOTA research is summarized in Table. 1.

Our previous work addresses this gap by introducing the **Functional-Semantic Activation Mapping (FSAM)** approach, which provides a global explanation of GNNs by extracting deep representations in the form of semantic

Table 1
Relevant Work Summarized

Method	TYPE	LEARNING	TASK	TARGET	BLACK-BOX	FLOW	DESIGN
SA [21, 15]	Instance-level	✗	GC/NC	N/E/NF	✗	Backward	✗
Guided BP [21]	Instance-level	✗	GC/NC	N/E/NF	✗	Backward	✗
CAM [15]	Instance-level	✗	GC	N	✗	Backward	✗
Grad-CAM [15]	Instance-level	✗	GC	N	✗	Backward	✗
GNNExplainer [16]	Instance-level	✓	GC/NC	E/NF	✓	Forward	✓
PGExplainer [22]	Instance-level	✓	GC/NC	E	✗	Forward	✓
GraphMask [23]	Instance-level	✓	GC/NC	E	✗	Forward	✓
ZORRO [24]	Instance-level	✗	GC/NC	N/NF	✓	Forward	✓
Causal Screening [25]	Instance-level	✗	GC/NC	E	✓	Forward	✓
SubgraphX [4]	Instance-level	✓	GC/NC	Subgraph	✓	Forward	✓
LRP [21, 26]	Instance-level	✗	GC/NC	N	✗	Backward	✗
Excitation BP [15]	Instance-level	✗	GC/NC	N	✗	Backward	✗
GNN-LRP [17]	Instance-level	✗	GC/NC	Walk	✗	Backward	✓
GraphLime [18]	Instance-level	✓	NC	NF	✓	Forward	✗
RelEx [27]	Instance-level	✓	NC	N/E	✓	Forward	✓
PGM-Explainer [19]	Instance-level	✓	GC/NC	N	✓	Forward	✓
XGNN [20]	Model-level	✓	GC	Subgraph	✓	Forward	✓
Our Work	M/I-level	✓	NC/GC	N/E/NF	✓	Forward	✓

graphs. FSAM focuses on capturing the global structure of the GNN, along with the semantic relationships between neurons across different layers. This method not only explains which components contribute to predictions but also 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 [28], which cannot be applied to graph adjacency matrices without losing important structural information, FSAM is specifically designed to preserve the discrete properties of graph structures. Furthermore, soft masking techniques [29], which are effective in image domains, compromise the integrity of graph structures when adapted to GNNs. By taking both nodes and edges into account, FSAM ensures that the model's inner workings can be interpreted in relation to the underlying graph structure and the relationships embedded within it.

One key advantage of FSAM over existing methods is its ability to map the learned representations of the GNN into a human-interpretable semantic space. This allows us to link the model's internal mechanisms to higher-level symbolic concepts, facilitating the validation of model decisions against domain knowledge and prior information. By explaining how the model processes graph data at each layer, FSAM enhances both transparency and accountability, providing explanations that are intuitive and accessible to non-experts. This comprehensive approach to GNN interpretability marks a significant advancement towards neurosymbolic AI, where models are not only accurate but also provide clear, explainable reasoning aligned with human understanding.

3. Overall Methodology: Generating the Semantic Graph

The primary aim of this paper is to enhance the interpretability of GNNs by representing their internal mechanisms as semantic graphs. In our extended study, we hypothesise that adding more layers to GNNs does not necessarily increase their capacity for knowledge representation. Our FSAM method clarifies GNN decisions retrospectively, focusing on how different layers contribute to, or in some cases reduce, model performance due to over-smoothing. FSAM identifies neuron groups involved in decision-making, termed *activation neurons*, and constructs a semantic graph to visualise their relationships. This section presents the mathematical formulation for generating the semantic graph, integrated with insights from our expanded experiments.

3.1. Mathematical Formulation

The process begins with computing the activation values for each neuron. Given an input graph $G = (V, E)$, where V represents nodes and E represents edges, the GNN processes this structure to produce an activation matrix $A = [a_{i1}, a_{i2}, \ldots, a_{in}]$ for each layer i. Here, n represents the number of neurons in layer i, corresponding either to the number of nodes in G or the output dimensionality of that layer. Our extended analysis suggests that after a certain number of layers, additional neurons contribute less meaningful information due to over-smoothing, resulting in decreased model performance.

To capture the behaviour of neurons within the GNN, we calculate neuron activations using Graph Convolutional Networks (GCNs), which classify nodes by embedding ego-graphs in Euclidean space. The embedding for a node v at layer ℓ is computed as:

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

where $e_{v,w}$ represents the edge weight between nodes v and w, $N(v)$ includes v and its neighbours, d_v and d_w denote the degrees of nodes v and w, ReLU is the activation function, and $W^{(\ell)}$ are the learned parameters. Here, activation values correspond to node embeddings in the input graph. Through our findings, we validate the hypothesis that increasing layers can lead to reduced neuron specialisation and heightened activation overlap across classes.

To analyse the relationships between neurons, we compute edge weights within the co-activation matrix using Spearman's correlation coefficient, an ideal metric for capturing monotonic relationships and non-linear associations among activation patterns across layers. The Spearman correlation coefficient ρ_{ij} for neurons i and j is defined as:

$$\rho_{ij} = \frac{\text{cov}(\text{rank}(a_i), \text{rank}(a_j))}{\sigma_i \sigma_j}$$

where cov represents covariance, $\text{rank}(a_i)$ and $\text{rank}(a_j)$ are ranks of the activation values a_i and a_j, and σ_i and σ_j are their standard deviations. This measurement quantifies neuron relationships and highlights over-smoothing; as layers increase, activations from different classes increasingly overlap, diminishing model performance. Our observations confirm that co-activations escalate beyond a certain depth, validating our hypothesis.

Additionally, we employ the point-biserial correlation coefficient to evaluate the relationship between input features and output classes. This coefficient measures the correlation between binary input variables and continuous outputs, calculated as:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_{pooled}} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

where $\bar{X}_1$ and $\bar{X}_0$ denote the mean activations for the two groups, s_{pooled} is the pooled standard deviation, and n is the total sample count. This calculation, when applied across layers, reveals diminishing feature-class correlations as layer depth increases, further supporting our hypothesis regarding over-smoothing effects.

Finally, semantic graphs generated through FSAM depict the relationships between neurons across layers. We visualise these graphs using thresholding techniques to identify the most influential neurons in decision-making. In our extended analysis, we perform a layerwise comparison to observe how increased layers affect the semantic structure. We find that additional layers beyond a certain threshold do not yield significant new information. Instead, the overlap between neuron activations for different classes intensifies, undermining class-specific representation and confirming our hypothesis that over-smoothing impairs the model's ability to distinguish classes effectively.

This extended analysis substantiates our hypothesis that beyond an optimal point, adding layers fails to enhance the GNN's knowledge capacity. The FSAM framework thus proves to be an insightful tool, not only for visualising these limitations but also for guiding the design of more efficient GNN architectures.

4. Key Contributions and Findings

In this extended work, we explore how varying the number of layers in GNNs impacts both model performance and the quality of knowledge representation. We build on our previous research, employing FSAM to systematically assess how well different layer configurations capture the underlying structure of input data. Our contributions are structured to answer a central question: do additional layers enhance the model's interpretability and accuracy, or do they introduce complexity that impairs representation quality? We present the following contributions:

Contribution 1: In the subsection 5.2 We are extending FSAM Validation Across Diverse Datasets.To evaluate FSAM's generalisability, we apply it to multiple datasets 5.1, we use Jaccard correlation graphs to analyse FSAM's ability to consistently capture semantic relationships across varied data. This cross-domain validation confirms that FSAM reliably mirrors the network's behaviour, as changes in model accuracy are generally reflected in FSAM graph quality. However, we also observe rare cases where accuracy improves without a corresponding enhancement in FSAM's semantic alignment. Such instances suggest that while the model produces correct predictions, it may be doing so for reasons that diverge from the semantic structure of the input data. This contribution highlights FSAM's value in diagnosing potential misalignments in GNN predictions.

Contribution 2: In the subsection 5.3 We are examining FSAM's Reflection of Network Behaviour Across Layer Configurations. We investigate FSAM's consistency across GNNs with varying layer depths, from one to four layers. By analysing how misclassifications and neuron community structures correlate within FSAM graphs, we validate that FSAM reliably captures network dynamics as layer configurations change. This analysis demonstrates that FSAM effectively reflects the network's evolving behaviour across layers, with improvements in accuracy typically mirrored by more coherent FSAM representations. Conversely, when accuracy declines—often due to over-smoothing in deeper layers—FSAM graphs capture this reduction in semantic clarity, underscoring FSAM's robustness as a tool for representing network behaviour across different layer depths.

Contribution 3: In the subsection 5.4, 5.3 We discuss a Layer-Wise Analysis of Knowledge Representation through a comprehensive, layer-by-layer analysis, we assess how increasing the number of GNN layers impacts the model's performance, focusing on class-specific accuracy and the phenomenon of over-smoothing, where neuron activations become overly similar. Our experiments reveal that while adding layers may initially enhance performance, beyond a certain depth, additional layers result in a decline in discriminative power. This contribution provides empirical support for optimising layer depth in GNN architectures, illustrating the trade-offs between model complexity and representation quality.

Contribution 4: In the subsection 5.5 We discuss the semantic Divergences in FSAM Graphs A key insight from our analysis is FSAM's ability to identify instances where accuracy trends and FSAM quality diverge. Specifically, we highlight cases where model accuracy improves, yet FSAM graph quality declines, revealing instances where the network may achieve correct predictions without fully capturing the semantic structure of the input data. Conversely, we also identify scenarios where accuracy decreases, but FSAM graph quality improves, potentially due to richer insights from misclassifications. These cases underscore FSAM's diagnostic potential in detecting "right for the wrong reasons" scenarios, providing a nuanced understanding of the network's semantic alignment with the data.

Overall, these contributions extend our prior findings, offering a detailed methodology for assessing GNN layer depth and performance. Together, these insights position FSAM as a valuable framework for balancing layer depth with interpretability and accuracy, ultimately enhancing the understanding and optimisation of GNN architectures across varied datasets.

5. Experimental Results and Validation of Contributions

We conducted experiments to evaluate how semantic graphs capture the behavior of GNNs across different layer depths, focusing on whether additional layers contribute meaningful knowledge or lead to over-smoothing. Using semantic graphs, we mapped neuron relationships within hidden layers and correlated these with output classes, identifying key neurons that influence model predictions. This approach demonstrated the effectiveness of semantic graphs in extracting knowledge from trained GNNs. We utilized six benchmark datasets to assess the impact of

layer depth on model performance and knowledge representation, testing our hypothesis that deeper layers may not always provide additional knowledge and could hinder class differentiation.

5.1. Datasets

For our extended experiments, we used six benchmark datasets to study how GNNs behave in different contexts. These datasets include **Cora** [7], **CiteSeer** [8], **PubMed** [9], **Amazon Computers** [10], **Amazon Photos** [10], and **Coauthor** [11].

The **Cora** and **CiteSeer** datasets are citation networks where nodes represent academic publications, and edges represent citation links between them. In **Cora**, there are 2,708 publications divided into seven categories: *Neural Networks*, *Rule Learning*, *Reinforcement Learning*, *Probabilistic Methods*, *Theory*, *Genetic Algorithms*, and *Case-Based Reasoning*. **CiteSeer** contains 3,312 publications in six categories: *Agents*, *Artificial Intelligence*, *Database*, *Information Retrieval*, *Machine Learning*, and *Human-Computer Interaction*. These datasets allow us to explore how GNNs classify papers based on their citation connections.

The **PubMed** dataset [9] is another citation network, focused on biomedical publications. It contains 19,717 publications, each classified into three categories: *Diabetes*, *Cardiovascular Disease*, and *Breast Cancer*. This dataset challenges the GNN's ability to handle complex medical literature classification, testing its ability to distinguish between closely related categories in the biomedical domain.

The **Amazon Computers** and **Amazon Photos** datasets [10] are product co-purchase networks, where nodes represent products and edges indicate products frequently bought together. The **Amazon Computers** dataset includes 13,752 products, covering categories like *desktops*, *laptops*, and *computer accessories*. The **Amazon Photos** dataset contains 7,650 products related to *cameras*, *photography accessories*, and *digital media*. These datasets evaluate the GNN's capacity to model product relationships and predict their respective categories.

Finally, the **Coauthor** dataset [11] represents a co-authorship network of academic authors. We used the **Coauthor CS** variant, which includes 18,333 nodes, representing authors in the field of computer science. The classification task is to assign authors to areas of expertise such as *Machine Learning*, *Artificial Intelligence*, and *Data Mining*. This dataset evaluates the GNN's ability to model relationships between authors and their research areas.

These datasets span a wide range of domains, from academic publications and biomedical research to product co-purchases and academic co-authorships. By testing GNNs across these varied graph structures, we can better understand how adding layers affects model performance and knowledge representation.

5.2. Analysing the Relationship between Semantic Graphs

As shown in Figure 1, the red dots represent the predicted classes, while blue nodes denote the activated neurons. Our analysis began by mapping each activated neuron from layer 1 to the final predicted class, extending this mapping progressively through each subsequent layer up to the output layer. This layer-wise mapping approach enabled a deeper understanding of the model's behaviour across layers and allowed us to evaluate the effects of increasing model depth.

In our extended experiments, we examined the relationships within the semantic graphs across multiple datasets, observing each layer's contribution to class-specific predictions and knowledge representation. As shown in Figure 2, we visualised the semantic graph for the Coauthor dataset, with blue nodes representing hidden layer neurons and red nodes indicating class labels. The predicted classes associated with activated neurons demonstrate alignment across layers, particularly in the earlier layers. Notably, in this model, layer 1 achieves an optimal accuracy of 98%. However, As shown in Table 2, beyond this depth—specifically after the second layer—we observe a decline in alignment, indicated by drops in accuracy and an increase in misclassification overlap as additional layers are added. This degradation corresponds with shifts observed in the FSAM graphs, as presented in the layer-wise semantic graphs in Appendix 8, where neuron activation patterns begin to lose their distinctiveness.

In contrast, the Amazon Photo dataset exhibited a distinct pattern, as outlined in Table 2, the addition of a second layer initially enhanced accuracy, a result corroborated by the FSAM semantic graph, which accurately reflected the network's structure up to this depth. Beyond the second layer, however, the FSAM graph (see Appendix 8) indicated

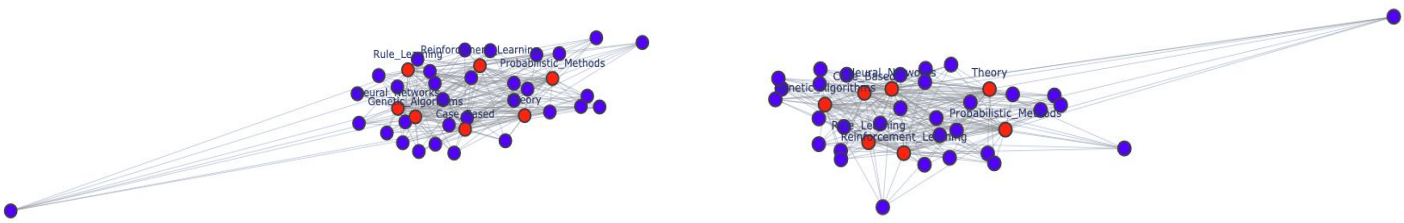

Fig. 1. Using Plotly.js, we visualised the semantic graph for the Cora dataset. Blue nodes, red nodes represent hidden layer neurons and classes, respectively. The left graph maps from layer 1 to the final class, the right graph from layer 2 to the final class.

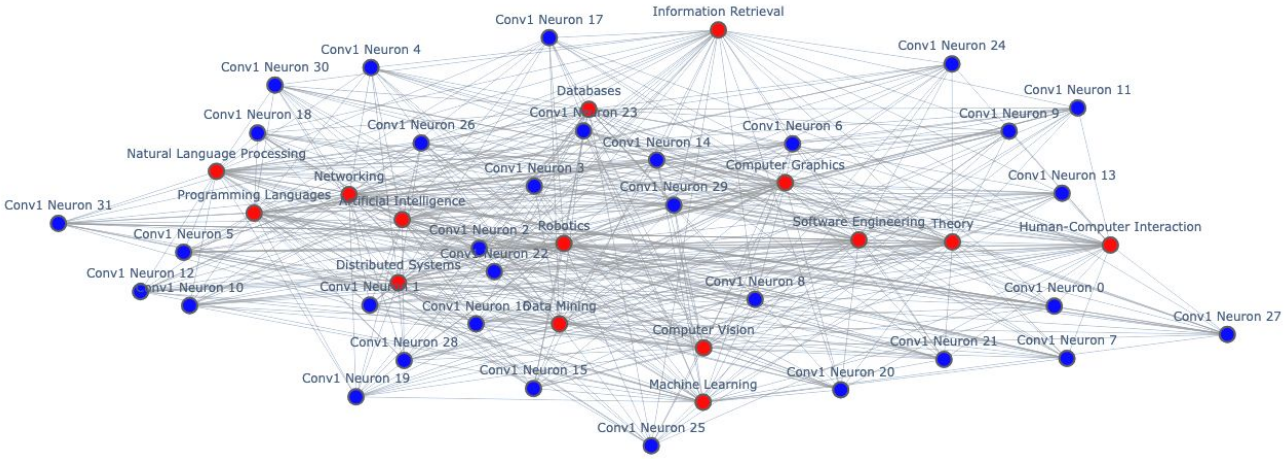

Fig. 2. Using Plotly.js, we visualised the semantic graph for the Coauthor dataset, with blue nodes representing hidden layer neurons and red nodes indicating class labels. This semantic graph was constructed across four distinct layers, allowing for a clear, layer-by-layer comparison of neuron-class relationships.

a divergence in representations from the network's behaviour, suggesting that further layers did not contribute to improved model performance.

For the Amazon Computers dataset, we constructed FSAM semantic graphs, as shown in Figure 4. In the initial two layers, accuracy improved steadily, which was corroborated by the FSAM semantic graphs in Table 2, reflecting well-aligned neuron activations and class distinctions that support accurate predictions. This indicates that the GNN is effectively leveraging layer depth to enhance representational capacity up to this point. However, in Layer 3, we observed a decrease in accuracy alongside a corresponding decline in the distinctiveness of FSAM representations, suggesting an onset of over-smoothing where neuron activations begin to overlap across classes. Interestingly, while accuracy increased again in Layer 4, the FSAM graph no longer exhibited a clear alignment with network behaviour. This divergence highlights a critical insight: the model's predictions may improve quantitatively, yet the qualitative alignment between FSAM activations and semantic coherence deteriorates. Such instances underscore the importance of balancing layer depth to maintain meaningful semantic representation without compromising interpretability.

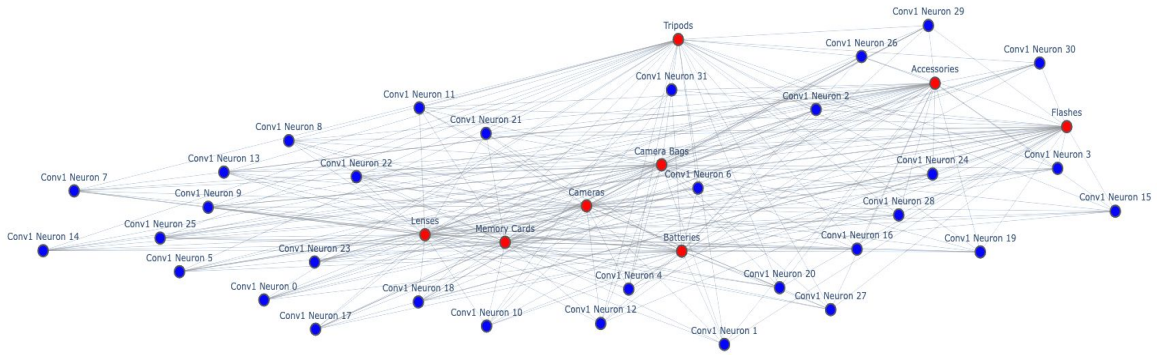

Fig. 3. Using Plotly.js, we visualised the semantic graph for the Amazon photo dataset, with blue nodes representing hidden layer neurons and red nodes indicating class labels. This semantic graph was constructed across four distinct layers, allowing for a clear, layer-by-layer comparison of neuron-class relationships.

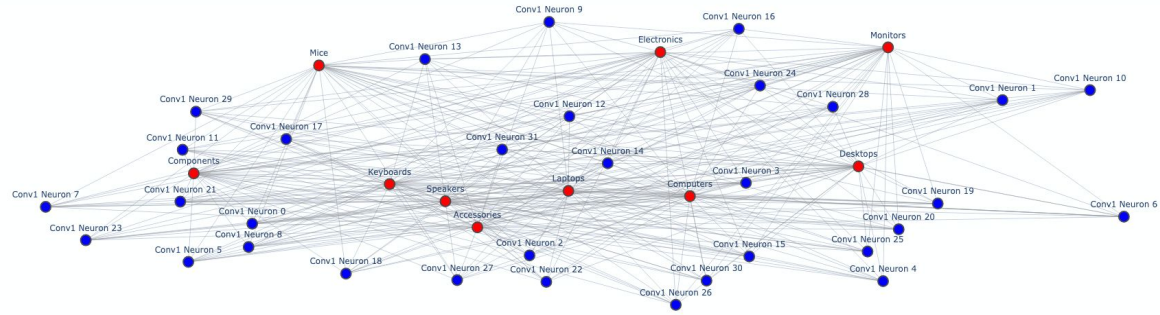

Fig. 4. Using Plotly.js, we visualised the semantic graph for the Amazon computer dataset, with blue nodes representing hidden layer neurons and red nodes indicating class labels. This semantic graph was constructed across four distinct layers, allowing for a clear, layer-by-layer comparison of neuron-class relationships.

5.3. Extended Validation of FSAM Across Layer Configurations

Our second contribution involves validating the FSAM approach's ability to reliably capture GNN behaviour across varying layer configurations. Through systematic experiments on several datasets, as detailed in 5.1, we analysed each GNN configuration (from 1 to 4 layers) to assess the alignment between model accuracy, misclassification patterns, and the community structures represented by FSAM graphs.

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

At 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

Table 2
Layer-wise Accuracy and Pearson Correlation for Various Datasets

Layer	Amazon Photos		CoauthorCs		Amazon Computers	
	Accuracy	**Pearson Correlation**	**Accuracy**	**Pearson Correlation**	**Accuracy**	**Pearson Correlation**
1	95%	0.681	98%	0.589	90%	0.683
2	96%	0.650	97%	0.756	91%	0.630
3	94%	0.752	96%	0.819	88%	0.785
4	93%	0.780	95%	0.834	89%	0.917

96%, while the Pearson correlation slightly decreases 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, accuracy begins to decline, dropping to 94%, while the Pearson correlation rises to 0.752. This increased correlation value indicates heightened overlap in neuron activations, signaling a loss of distinctiveness among class-specific features. Here, FSAM reveals that over-smoothing begins to emerge, with class representations blurring as neuron activations start to 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, 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 are contributing to over-smoothing. 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. While the initial layers enhance accuracy with minimal activation overlap, further layers lead to increased correlation between overlapping nodes and misclassification errors. This positive correlation between class similarity and mistake counts underscores FSAM's diagnostic potential, providing insights 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 layer-wise shifts in network behaviour.

At 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 rises (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 over-smoothing 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 network's evolving behaviour across layers, accurately capturing the interplay between model accuracy and neuron overlap, and reinforcing 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 across layers relate to the underlying FSAM graph structures. 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 network's behaviour. 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, accuracy increases slightly to 91%, while Pearson correlation decreases to 0.630. This reduction in correlation, alongside improved accuracy, indicates that neuron activations remain well-separated, supporting the model's continued effectiveness in distinguishing between classes. The FSAM graph here effectively aligns with the improved class distinction, reinforcing the model's structural clarity.

However, by the third layer, accuracy decreases to 88%, and the Pearson correlation rises to 0.785. This shift marks an increase in overlapping neuron activations, suggesting a decline in class distinction, likely attributable to over-smoothing. 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, accuracy slightly recovers to 89%, yet the Pearson correlation further 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 demonstrate FSAM's capacity to accurately represent network behaviour across diverse configurations. Specifically, as accuracy improves, the FSAM activation graph tends to exhibit stronger alignment with the semantic structure. Initial layers, such as the second, achieve higher accuracy with low correlation, showing effective 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.

5.4. Comparison of Mistakes Across Communities for Each Dataset

In Table 3, we present a detailed comparison of mistakes across communities for each dataset at varying layer depths, structured around our core hypotheses. This analysis provides insights into the effects of layer depth on knowledge representation, class-specific accuracy, and GNN interpretability using FSAM.
The analysis of mistakes across communities within the CoauthorCS dataset reveals a progressive shift in community structure as the number of GNN layers increases, illustrating how layer depth impacts classification accuracy and neuron activation overlap. Each layer's community structure, represented by clusters of semantically related fields, highlights distinct groupings at lower layers, which gradually blend as network depth increases, thus validating our results presented in Table 2.
In **Layer 1**, the community structure is clearly delineated, with minimal neuron overlap between different fields. 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 of 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 in this cluster. 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 in this cluster. Our class accuracy representation Fig. 5 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 decline in accuracy as neuron activations

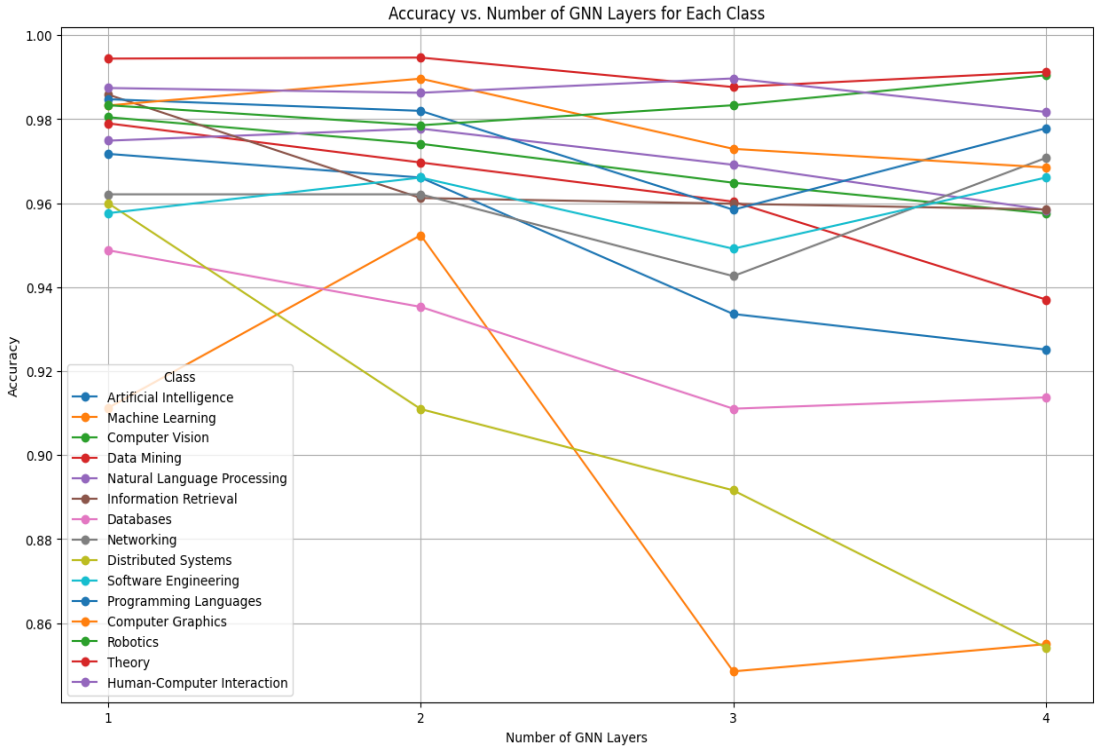

Fig. 5. Layer wise accuracy per class contributionof CoAuthorCs

begin to overlap between certain communities. Here we represent 'Prediction decline changes **A«** and **C»**' and 'Major community shift', respectively This trend is captured in the FSAM, showing increased correlation in activations that reflects 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 rise to 1410, signifying increased misclassifications as class boundaries blur. This layer also corresponds to higher Pearson correlation, indicating substantial overlap in neuron activations. The FSAM graph effectively captures this over-smoothing, showing that the distinctiveness among communities is diminishing with deeper layers.

Overall, these results demonstrate that as layers are added, the CoauthorCS community structure evolves, 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 in capturing the network's shifting behaviour across layers. The increase in misclassification and Pearson correlation values illustrates how FSAM serves as a diagnostic tool, accurately reflecting the trade-off between layer depth and community coherence, thereby validating the results as displayed in Table 2.

The analysis of mistakes across communities in the Amazon Photos dataset, as outlined in Table 2, provides valuable insights into 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 clearly 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, while **C1** broadens to encompass Camera Bags, Memory Cards, Flashes, and Batteries. This reorganisa-

tion 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, in **Layer 3**, the model's performance begins to deteriorate, with the mistake count increasing significantly. 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.

By **Layer 4**, the network exhibits signs of over-smoothing, 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. This indicates that, although mistakes may appear to lessen, the underlying community distinctions are weakened, suggesting that the model may be achieving accuracy without a robust semantic foundation.

This layer-wise 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 3, The analysis of the Amazon Computers dataset is effective in capturing effective shifts in network behaviour across different layers, especially in cases where accuracy trends diverge from FSAM correlation trends. This is demonstrated through the changes in community structures and mistake patterns across the layers.

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 shift in community structure. Products such as "Keyboards" and "Mice" migrate from **C1** to **C0**, as denoted by the significant labels $^{A»}$ and $^{C»}$. Interestingly, accuracy improves in this layer, and the mistake count decreases to 410. While 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, 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 correlation in this layer implies greater overlap in neuron activations, indicative of over-smoothing. 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 further decreases, and the mistake count rises to 406. FSAM reveals that **C0** now includes a mix of "Desktops," "Speakers," and "Laptops," signifying even greater overlap between distinct product categories. The increase in Pearson correlation and decrease in accuracy indicate 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 FSAM's ability to capture both 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 accuracy and FSAM quality degrade, FSAM effectively reflects the reduced class-specific representation, underscoring its utility as a diagnostic tool for evaluating layer-wise GNN behaviour.

Table 3

Community Structure and Mistakes Across Layers for Each Dataset, where **A»**, **A«**, and **C»** represent 'Prediction changes noticeable' and 'Major community shift', respectively.

Dataset	Layer	Community (Classes)	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
	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»
	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
	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
Amazon Photos	1	**C0**: Cameras, Lenses, Camera Bags; **C1**: Memory Cards, Flashes, Batteries; **C2**: Accessories, Tripods	452
	2	**C0**: Cameras, Lenses; **C1**: Camera Bags, Memory Cards, Flashes, Batteries; **C2**: Accessories, Tripods	410 A»,C»
	3	**C0**: Memory Cards, Lenses, Flashes, Batteries, Camera Bags; **C1**: Accessories, Tripods, Cameras	497
	4	**C0**: Cameras; **C1**: Flashes, Tripods, Camera Bags; **C2**: Accessories, Memory Cards, Batteries, Lenses	406
PubMed	1	**C0**: Cardiovascular Disease, Diabetes; **C1**: Breast Cancer	46
	2	**C0**: Cardiovascular Disease, Diabetes; **C1**: Breast Cancer	38
	3	**C0**: Breast Cancer; **C1**: Cardiovascular Disease, Diabetes	32 A»,C»
	4	**C0**: Breast Cancer; **C1**: Cardiovascular Disease, Diabetes	62
Cora	1	**C0**: Case-Based, Neural Networks, Genetic Algorithms; **C1**: Theory; **C2**: Reinforcement Learning, Probabilistic Methods	220
	2	**C0**: Case-Based, Genetic Algorithms; **C1**: Reinforcement Learning, Rule Learning, Probabilistic Methods; **C2**: Neural Networks, Theory	356
	3	**C0**: Neural Networks, Theory; **C1**: Case-Based, Rule Learning, Genetic Algorithms; **C2**: Reinforcement Learning, Probabilistic Methods	364 A»,C»
	4	**C0**: Case-Based, Neural Networks, Probabilistic Methods, Theory; **C1**: Genetic Algorithms, Reinforcement Learning, Rule Learning	378
Amazon Computers	1	**C0**: Components, Mice, Speakers; **C1**: Desktops, Laptops, Keyboards, Computers, Accessories; **C2**: Monitors, Electronics	452
	2	**C0**: Keyboards, Components, Mice, Speakers ; **C1**: Desktops, Laptops, Computers, Electronics; **C2**: Monitors, Accessories	410 A»,C»
	3	**C0**: Speakers, Laptops, Keyboards, Components, Mice, Accessories; **C1**: Desktops, Monitors, Electronics, Computers	397
	4	**C0**: Desktops, Speakers, Laptops, Keyboards, Computers, Components, Accessories; **C1**: Monitors, Electronics, Mice	406

5.5. Layer-Wise Class Similarity and Misclassification Analysis

A key insight from our analysis is FSAM's ability to identify instances where accuracy trends and FSAM quality diverge. Specifically, we highlight cases where model accuracy improves, yet FSAM graph quality declines, revealing instances where the network may achieve correct predictions without fully capturing the semantic structure of the input data. Conversely, we also identify scenarios where accuracy decreases, but FSAM graph quality improves,

potentially due to richer insights from misclassifications. These cases underscore FSAM's diagnostic potential in detecting "right for the wrong reasons" scenarios, providing a nuanced understanding of the network's semantic alignment with the data.

To validate our findings on the influence of layer depth and class similarity on GNN performance, we generated several key figures to capture essential aspects of the model's behaviour:

- **Total Accuracy vs. Number of GCN Layers**: This illustrates how overall model accuracy changes with the addition of layers, providing a broad view of whether deeper architectures consistently improve performance or contribute to over-smoothing, thereby reducing accuracy, as shown in section 8.
- **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 section 8.
- **FSAM Graphs Showing Neuron Activations for Specific Classes**: These graphs display neuron activation patterns within each class, allowing us to track the GNN's ability to capture class-specific features across layers. They reveal the points at which neuron activations begin to overlap, indicating where class boundaries start to lose distinctiveness, as shown in section 8.
- **Community Structures Highlighting Class Groupings**: This visualisation illustrates the community structures of neuron activations, clustering classes based on co-activation. These clusters indicate relatedness among certain classes and provide insight into the GNN's knowledge organisation, revealing where class separability degrades with additional layers, as shown in section 8.
- **Jaccard Coefficient vs. Number of Mistakes at Layer 3**: This 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 greater overlap in FSAM exhibit more frequent misclassifications, as shown in section 8.

Our extended analysis revealed a positive correlation between class similarity and the number of mistakes involving them, as illustrated with examples from the **CoauthorCS** and **Amazon Photos** datasets. Table 2 shows that class pairs with higher overlap in the FSAM graph also exhibit more misclassifications. In the **CoauthorCS dataset**, our findings reveal that Layer 1 achieves optimal performance, as demonstrated in . Adding further layers results in decreased 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 further layers. This pattern, shown in the Table 2, 3 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 activation overlap within the same community.

These findings suggest that adding layers beyond an optimal depth does not necessarily improve knowledge representation. Instead, it introduces an over-smoothing effect, where neuron activations for different classes become increasingly indistinct, reducing the model's ability to differentiate between them. This effect is substantiated by our correlation analysis, 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 class pair, we assessed the impact of these similarities on the GNN's decision-making. In the Amazon Photos dataset, for instance, product categories such as *Memory Cards* and *Accessories* displayed high Jaccard similarity, leading to frequent misclassifications.

These insights suggest that tuning efforts should focus on reducing overlap in the co-activation graph for similar classes to enhance the GNN's ability to differentiate between them. By targeting overlapping nodes, we 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 certain cases, may diminish the model's discriminative power due to overlapping neuron activations.

6. Conclusion and Future Work

In this extended study, we've worked to deepen the understanding of how GNNs behave by 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 behavior. Our findings also indicate that adding more layers to GNNs doesn't always lead to better performance or richer knowledge representation.

In these FSAM graphs, nodes represent neurons, and weighted edges indicate the strength of their co-activation relationships, reflecting correlations in activation patterns across layers. This layered view of the GNN's function shows how neurons contribute to specific class predictions and influence overall model decisions. Our experiments confirmed that FSAM's graph structure aligns well with the knowledge stored in GNNs, especially in distinguishing closely related classes. Across datasets, FSAM consistently highlighted key neurons and communities within the GNN that are central to specific class predictions, providing valuable insights into the model's decision-making process.

Additionally, we used community detection in FSAM graphs to see how the GNN naturally groups classes based on activation patterns. Our analysis showed that classes 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 FSAM's diagnostic power. The FSAM graphs and community detection further clarify how the GNN organizes 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 propose a few directions. One is to develop methods that dynamically adjust GNN layer depth based on the properties of the input graph, allowing for model configuration without manual tuning. Another focus could be on further class-level analysis within FSAM to develop more holistic metrics for evaluation. We also plan to combine FSAM insights with contextual information from input graphs, aiming to create more detailed, context-aware explanations that enhance both local and global interpretability.

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] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).
[2] Petar Velickovic et al. "Graph attention networks". In: *stat* 1050.20 (2017), pp. 10–48550.
[3] Hao Yuan and Shuiwang Ji. "Structpool: Structured graph pooling via conditional random fields". In: *Proceedings of the 8th International Conference on Learning Representations*. 2020.
[4] Hao Yuan et al. "On explainability of graph neural networks via subgraph explorations". In: *International conference on machine learning*. PMLR. 2021, pp. 12241–12252.
[5] Xiang Wang et al. "Towards multi-grained explainability for graph neural networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18446–18458.
[6] 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.
[7] Prithviraj Sen et al. "Collective classification in network data". In: *AI magazine* 29.3 (2008), pp. 93–93.

[8] Galen Namata et al. “Query-driven active surveying for collective classification”. In: *10th International Workshop on Mining and Learning with Graphs (MLG)* (2012).
[9] T. O. Botari et al. “Gene expression-based classification of diffuse large B-cell lymphoma”. In: *Nature* (2002), pp. 261–268.
[10] 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.
[11] Oleksandr Shchur et al. “Pitfalls of Graph Neural Network Evaluation”. In: *Relational Representation Learning Workshop, NeurIPS 2018*. 2018.
[12] 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.
[13] Robert Geirhos et al. “Generalisation in humans and deep neural networks”. In: *Advances in neural information processing systems* 31 (2018).
[14] 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.
[15] 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.
[16] Zhitao Ying et al. “Gnnexplainer: Generating explanations for graph neural networks”. In: *Advances in neural information processing systems* 32 (2019).
[17] 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.
[18] Qiang Huang et al. “Graphlime: Local interpretable model explanations for graph neural networks”. In: *IEEE Transactions on Knowledge and Data Engineering* (2022).
[19] 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.
[20] 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.
[21] Federico Baldassarre and Hossein Azizpour. “Explainability techniques for graph convolutional networks”. In: *arXiv preprint arXiv:1905.13686* (2019).
[22] Dongsheng Luo et al. “Parameterized explainer for graph neural network”. In: *Advances in neural information processing systems* 33 (2020), pp. 19620–19631.
[23] 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).
[24] Thorben Funke, Megha Khosla, and Avishek Anand. “Hard masking for explaining graph neural networks”. In: *Advances in neural information processing systems* (2020).
[25] Xiang Wang et al. “Causal screening to interpret graph neural networks”. In: (2020).
[26] Robert Schwarzenberg et al. “Layerwise relevance visualization in convolutional text graph classifiers”. In: *arXiv preprint arXiv:1909.10911* (2019).
[27] 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.
[28] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. “Feature visualization”. In: *Distill* 2.11 (2017), e7.
[29] Jianbo Chen et al. “Learning to explain: An information-theoretic perspective on model interpretation”. In: *International conference on machine learning*. PMLR. 2018, pp. 883–892.

8. Appendix

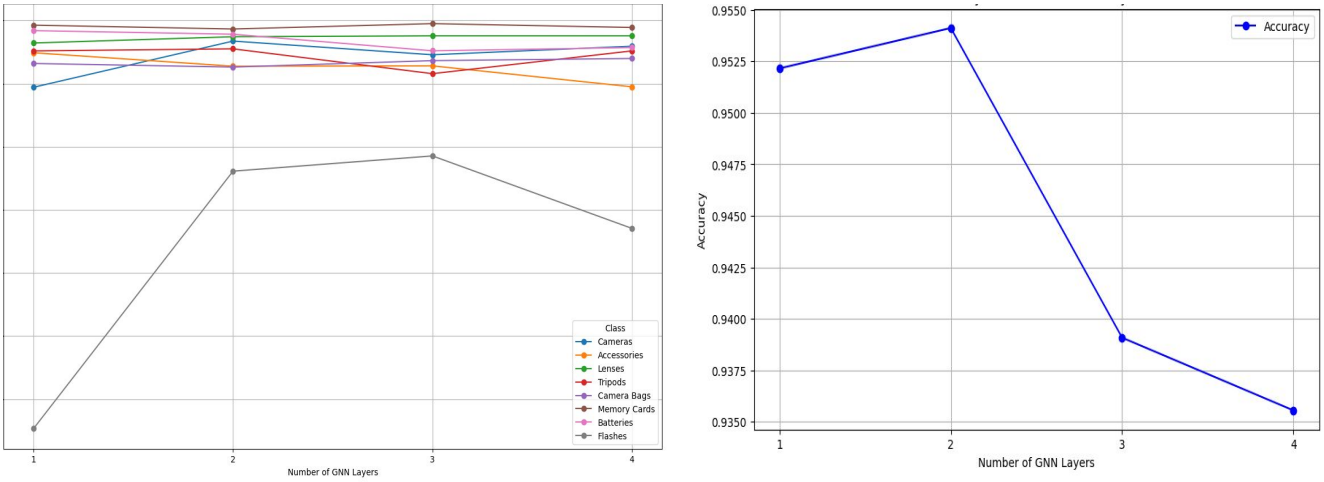

Fig. 6. Total Accuracy vs. Number of GCN Layers for Amazon photo

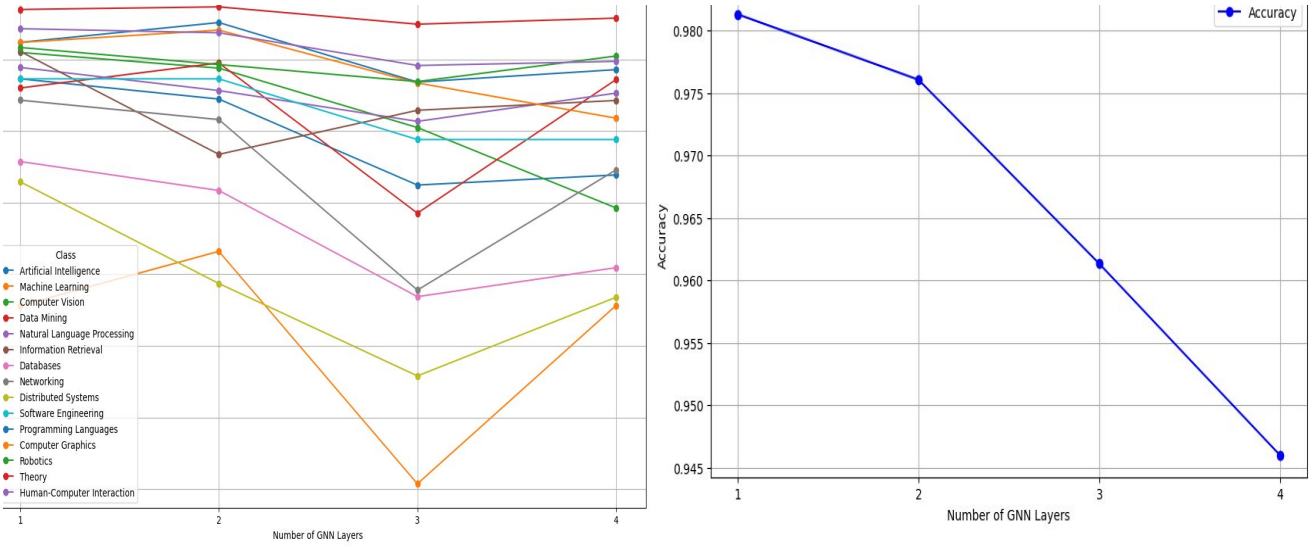

Fig. 7. Total Accuracy vs. Number of GCN Layers for CoauthorCs

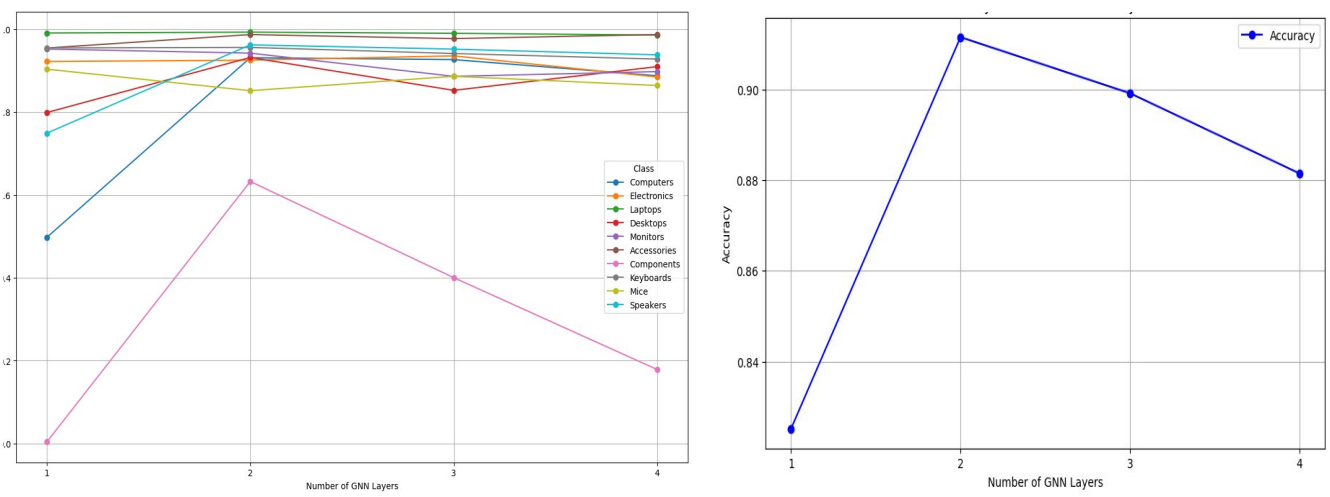

Fig. 8. Total Accuracy vs. Number of GCN Layers for Computers

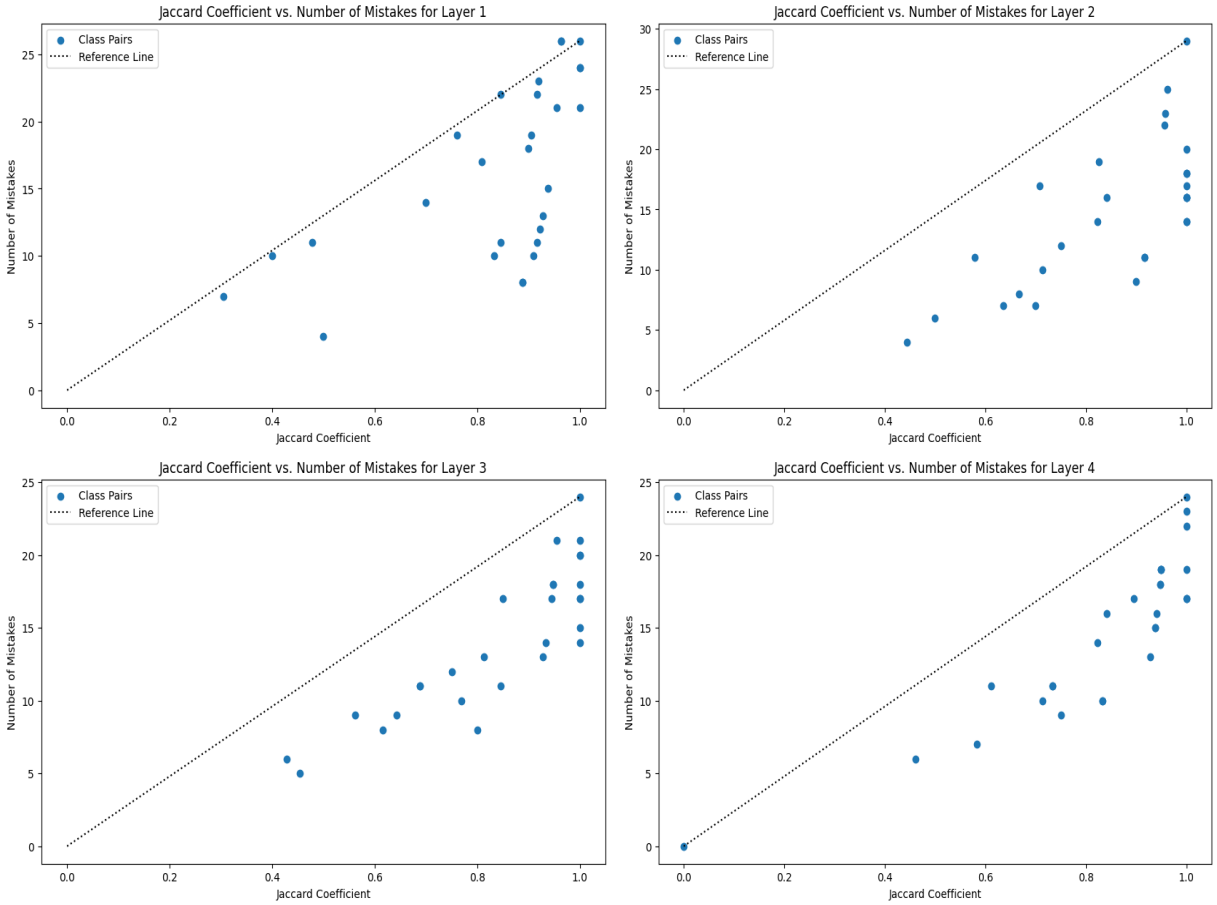

Fig. 9. Jaccard Similarity between different layers for AmazonPhoto dataset

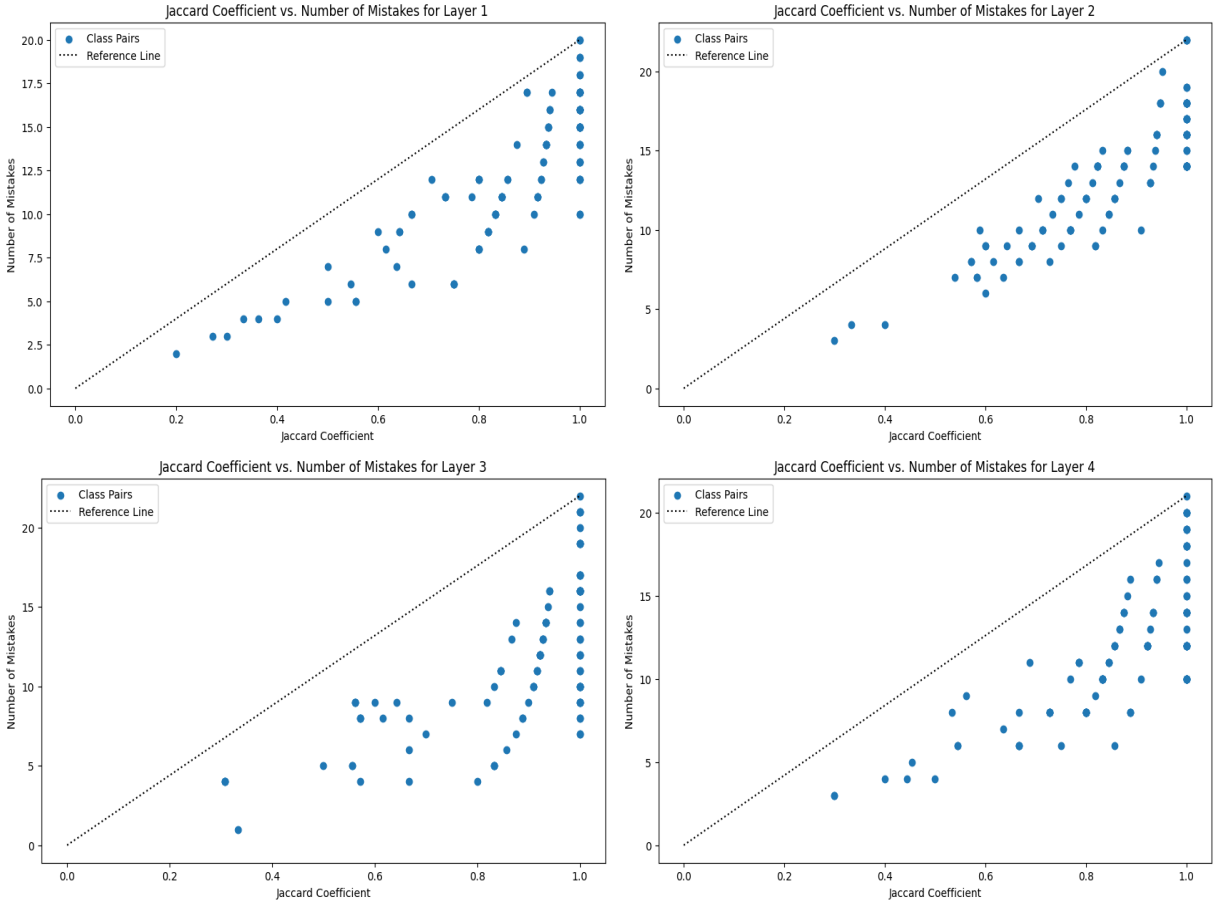

Fig. 10. Jaccard Similarity between different layers for CoauthorCs

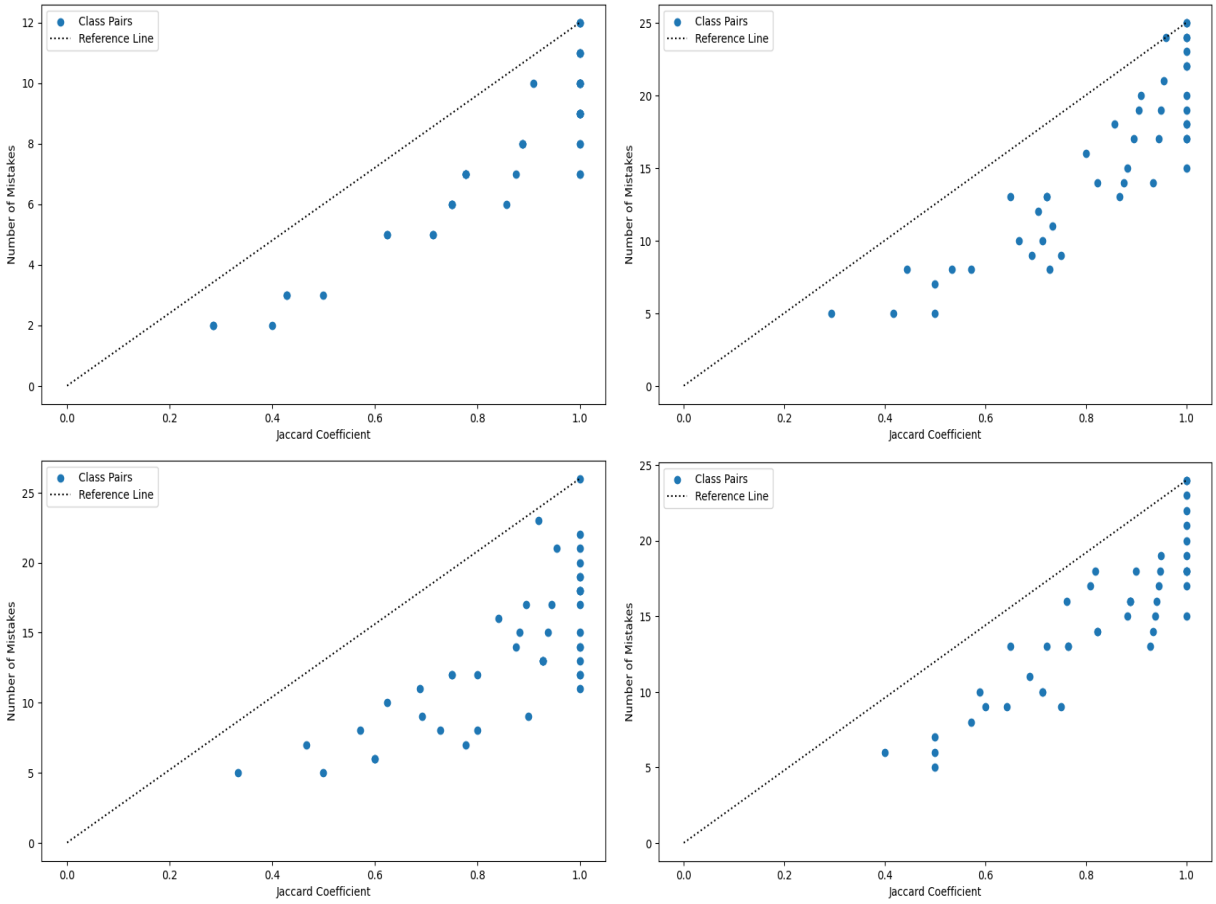

Fig. 11. Jaccard Similarity between different layers for Computers