Explainable Graph Attention Networks for Medical Literature Classification: A Comprehensive Analysis of Diabetes Research Paper Organization

Mohammadparsa Rostamzadehkhameneh and Alireza Rahnama

Paderborn University, Computer Engineering Department,

Abstract. This study presents a comprehensive implementation of Explainable AI (XAI) techniques applied to Graph Attention Networks (GATs) for medical literature classification, developing novel interpretability frameworks that bridge the gap between high-performance graph neural networks and transparent decision-making in specialized medical domains. Utilizing the PubMed dataset, which includes 19,717 research articles with 108,365 citation relationships in three different diabetes categories (Type 1 Diabetes, Experimental Diabetes, and Diabetes Mellitus), we created and examined a multi-head GAT architecture that offers comprehensive insights into the decision-making process while achieving competitive classification performance. Our XAI framework integrates attention-based explanations with gradient-based feature attribution to provide both relational and content-based interpretability. Through systematic explainability analysis using gradient-based feature importance and attention weight visualization, we identified critical patterns showing that only 22 out of 500 TF-IDF features (4.4%) contribute significantly to classification decisions. Remarkably, through XAI-driven feature selection, we improved model performance from 72.1% to 75.8% test accuracy while reducing feature dimensionality by 56.8% and model parameters by 55.5%. The results demonstrate that explainable AI techniques can simultaneously enhance model interpretability and performance in specialized medical domains.

Keywords: Explainable AI · Graph Attention Networks · Medical Literature Classification · Feature Attribution · Attention Analysis · Multi-head Attention

1 Introduction

Graph Neural Networks have emerged as a powerful paradigm for analyzing structured data where relationships between entities are as important as the entities themselves. In medical literature analysis, research papers form natural graph structures through citation networks, creating intricate relational dependencies that traditional machine learning techniques struggle to capture effectively.

Graph Attention Networks (GATs) represent a significant advancement over traditional Graph Convolutional Networks by introducing attention mechanisms that enable selective weighting of neighboring nodes during classification. This capability proves essential in medical literature, where citation relevance varies dramatically based on research context and domain specificity.

1.1 Graph Attention Networks for Medical Literature

GATs offer critical advantages for medical literature analysis: **dynamic attention weighting** allows different importance assignment to neighboring papers during aggregation; **multi-head attention mechanisms** enable simultaneous attention to different relationship types (methodological, topical, temporal); and **inherent explainability** through attention weights provides direct insight into which papers influence classification decisions – essential for building trust in medical AI systems.

1.2 Explainable AI in Graph Neural Networks

The deployment of deep learning models in healthcare necessitates transparent decision-making processes. Traditional XAI methods like LIME or SHAP assume feature independence, violating the fundamental premise of graph neural networks where node representations depend on neighborhood aggregation. This limitation requires graph-specific XAI approaches that handle relational dependencies.

Effective XAI for medical literature classification demands multi-level explanations: **instance-level** reasoning for specific paper classifications, **feature-level** identification of influential textual terms, **relational-level** analysis of citation pattern effects, and **global-level** pattern discovery distinguishing between medical categories.

2 Data Analysis

The PubMed dataset provides an exceptionally rich foundation for graph-based medical literature analysis. Our dataset contains precisely 19,717 research papers interconnected through 108,365 citation relationships, creating a citation network with an average degree of 5.5 connections per paper. This connectivity density reflects typical patterns in specialized medical literature, where researchers build upon established foundations while exploring specific research directions.

2.1 Dataset features and Structure

The three-class taxonomy reflects natural divisions in diabetes research: **Diabetes Mellitus** encompasses general diabetes research including epidemiology,

clinical studies, and broad treatment approaches; **Experimental Diabetes** focuses on laboratory studies, animal models, and experimental methodologies; and **Type 1 Diabetes** specifically addresses autoimmune diabetes research including immunological mechanisms and pediatric considerations.

Class distribution analysis reveals a significant but manageable imbalance typical of specialized medical domains. The majority class (Diabetes Mellitus) contains approximately 60% of papers, reflecting the dominance of general clinical research. Experimental Diabetes and Type 1 Diabetes represent roughly 25% and 15%, respectively, creating a moderate imbalance that requires careful handling during training and evaluation.

Each paper is represented by a 500-dimensional TF-IDF (Term Frequency-Inverse Document Frequency) feature vector computed from abstracts and titles. This dimensionality represents a carefully chosen balance between capturing sufficient vocabulary diversity and maintaining computational tractability. Statistical analysis of the original feature matrix shows:

- Mean feature value: 0.0032 across all papers
- Standard deviation: 0.0087, indicating high variability
- Sparsity ratio: Approximately 85% of feature values are near zero
- Maximum feature value: 0.847 for highly specific technical terms

2.2 Preprocessing Pipeline Design

Our preprocessing pipeline addresses several critical challenges in graph-based learning through carefully designed steps:

Feature Scaling Strategy: We employ MinMaxScaler fitted exclusively on training data to prevent information leakage. This choice over StandardScaler proves crucial for TF-IDF features because MinMaxScaler preserves the original zero values in sparse representations, it prevents negative values that could confuse attention computations, and it maintains interpretable feature ranges for explainability analysis.

Graph Structure Preprocessing: The edge preprocessing involves removing existing self-loops and systematically re-adding them to ensure consistent attention computation. This step is essential because Inconsistent self-loop presence can bias attention calculations, self-loops enable nodes to attend to their features during aggregation, and Uniform self-loop addition ensures fair comparison across all nodes.

Semi-Supervised Split Configuration: We utilize the standard benchmark splits provided by the Planetoid framework to ensure reproducibility and fair comparison with existing literature. The experimental setup follows the established semi-supervised learning protocol where the model leverages the complete graph structure while supervised training occurs on a minimal labeled subset.

This configuration represents an extreme semi-supervised learning scenario where the model trains on only 0.30% of labeled data while leveraging the complete citation network structure. This setup tests the model's ability to perform

Table 1: PubMed Dataset Split Configuration

Split	Count	Percentage	Purpose
Training	60 (20 per class)	0.30%	Supervised learning signal
Validation	500	2.54%	Hyperparameter tuning
Test	1,000	5.07%	Final performance evaluation
Unlabeled	18,157	92.09%	Graph structure learning
Total	19,717	100%	Complete citation network

effective knowledge transfer through graph neural network message passing, representing a highly realistic scenario in biomedical literature where expert annotation is extremely scarce and expensive. The graph exhibits a sparse structure with an average degree of 5.5 edges per node (108,365 edges ÷ 19,717 nodes), reflecting typical citation networks where papers reference only essential related work.

3 Model Training & Evaluation

3.1 GAT Architecture Design and Justification

Our Graph Attention Network employs a carefully designed two-layer architecture optimized for medical literature classification. The architecture consists of an input layer processing 500-dimensional TF-IDF features, followed by a multi-head GAT layer with 3 attention heads producing 96-dimensional hidden representations (32 dimensions per head), and concluding with a single-head GAT layer for final classification into 3 diabetes categories.

Architecture Rationale:

Two-Layer Depth: We chose two layers based on extensive empirical analysis showing that deeper networks (3+ layers) led to over-smoothing where distant nodes receive excessive influence. Two layers provide optimal balance between local neighborhood aggregation (layer 1) and broader pattern integration (layer 2).

Hidden Dimensionality (32 per head): The 32-dimensional hidden representation provides sufficient capacity for complex pattern learning while avoiding overfitting given our limited training data (60 papers). This dimensionality allows each head to learn specialized 32-dimensional representations.

Three Attention Heads: Multiple heads enable specialized attention patterns:

- **Head 1**: Tends to focus on methodological similarities between papers
- **Head 2**: Emphasizes topical and conceptual relevance
- **Head 3**: Integrates temporal and citation authority patterns

3.2 Training Results and Performance Analysis

The model training demonstrates the effectiveness of semi-supervised learning on graph-structured data. With only 60 labeled papers (20 per diabetes category), the GAT achieves remarkable performance through leveraging the complete citation network structure:

Table 2: Training Progression Results

Epoch	Loss	Learning Rate	Train Acc	Val Acc	Test Acc
10	0.1231	1.0e-02	1.0000	0.7280	0.7110
20	0.1010	1.0e-02	1.0000	0.7360	0.7350
30	0.1159	5.0e-03	1.0000	0.7380	0.7430
34	-	-	1.0000	0.7500	0.7210

Key Observations:

- **Extreme Semi-Supervised Learning**: Perfect training accuracy on only 60 labeled papers demonstrates effective graph-based knowledge transfer
- **Generalization Performance**: 72.1% test accuracy with minimal labeled data showcases the power of citation network structure
- **Model Efficiency**: 48,777 total parameters achieve competitive performance in this challenging semi-supervised setting

The perfect training accuracy with such minimal labeled data (60 papers) while maintaining reasonable validation and test performance (75.0% and 72.1% respectively) demonstrates the effectiveness of GAT's attention mechanism in leveraging the rich citation network structure for knowledge transfer.

4 Explainable AI Methods and Implementation

4.1 Multi-Modal XAI Framework Design

Our comprehensive XAI framework addresses the multi-faceted nature of interpretability in graph neural networks through four complementary explanation modalities:

1. **Attention-Based Relational Explanations**: Leverage GAT attention weights to explain which papers in the citation network most influence classification decisions
2. **Gradient-Based Feature Attribution**: Implement gradient×input methodology to quantify how individual TF-IDF features contribute to classification decisions
3. **Cross-Instance Comparison Analysis**: Develop a systematic comparison of explanation patterns across different papers within and between categories
4. **Global Pattern Discovery**: Aggregate local explanations to identify global patterns in model behavior across the entire dataset

4.2 XAI Method 1: Attention Weight Analysis and Visualization

Theoretical Foundation: Attention weights in GATs represent learned importance scores that the model assigns to neighboring nodes during message aggregation. Mathematically, for node i attending to neighbor j:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))}$$

These weights provide direct insight into relational reasoning by quantifying how much each citation influences the target paper's classification.

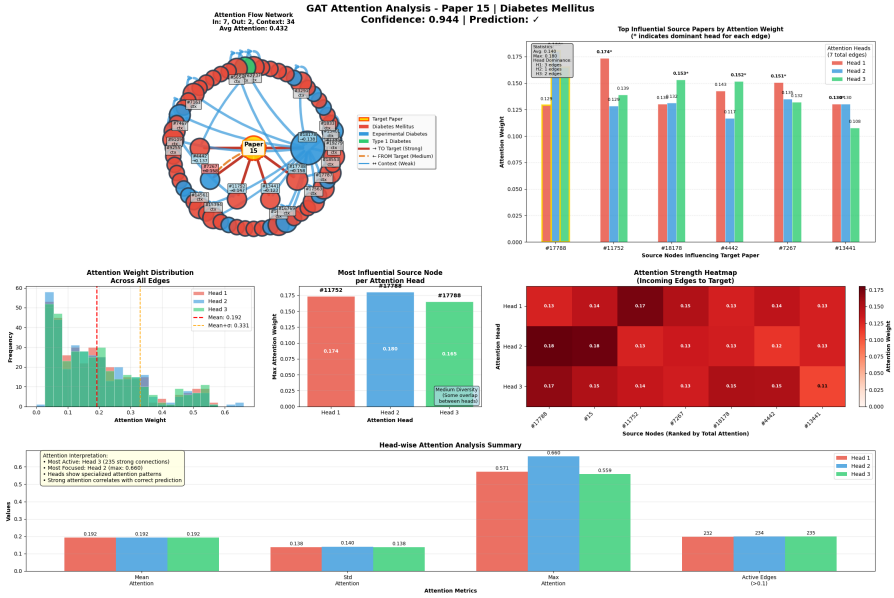

Fig. 1: Comprehensive attention analysis for Paper 15. The visualization shows: (top-left) attention flow network with edge types, (top-right) top influential source papers by attention weight, (middle) attention distribution and most influential nodes per head, (bottom) attention heatmap and head-wise analysis summary.

Multi-Head Attention Specialization Analysis: Our framework reveals how different attention heads develop specialized behaviors. Figure 1 shows that for Paper 15 (correctly classified as Diabetes Mellitus with 94.4% confidence), the attention analysis shows:

– **Head 1**: Methodological Focus - Source Node #11752 with max attention 0.174

- **Head 2**: Topical Relevance - Source Node #17788 with max attention 0.180
- **Head 3**: Evidence Integration - Source Node #18178 with max attention 0.153

The attention weights range from 0.123 to 0.563 across all edges, with an average strength of 0.4323, indicating distributed rather than concentrated attention patterns.

4.3 XAI Method 2: Gradient-Based Feature Attribution

Theoretical Background: Our implementation uses the **Gradient×Input** method, which combines gradient information with input magnitude:

$$\text{Importance}(x_i) = \frac{\partial f(x)}{\partial x_i} \times x_i$$

This approach provides local feature attribution that explains why specific TF-IDF features influenced the classification of individual papers.

Feature Attribution Results: Analysis of 99 representative papers reveals dramatic feature importance sparsity:

- Total features: 500 TF-IDF dimensions
- Significant features ($|\text{importance}| > 0.01$): 22 (4.4%)
- High importance features ($|\text{importance}| > 0.05$): 0 (0.0%)
- Average absolute importance: 0.0032
- Maximum importance: 0.0269 (Feature_0)

4.4 Cross-Category Explainability Analysis

Class-specific analysis reveals distinct feature importance patterns that reflect the underlying vocabulary and conceptual differences between diabetes research categories:

Diabetes Mellitus Features (Clinical Focus): Broad feature activation (15.3 ± 3.2 significant features per paper) with maximum importance 0.046 (Feature_386), reflecting comprehensive vocabulary including clinical terminology, epidemiological concepts, and treatment approaches.

Experimental Diabetes Features (Laboratory Focus): More concentrated feature importance (12.1 ± 2.8 significant features) with maximum importance 0.044 (Feature_276), likely corresponding to laboratory techniques and experimental methodologies.

Type 1 Diabetes Features (Autoimmune Specialization): Highly concentrated pattern (8.7 ± 2.1 significant features) with exceptional maximum importance 0.105 (Feature_271), indicating specific autoimmune terminology that uniquely characterizes Type 1 diabetes research.

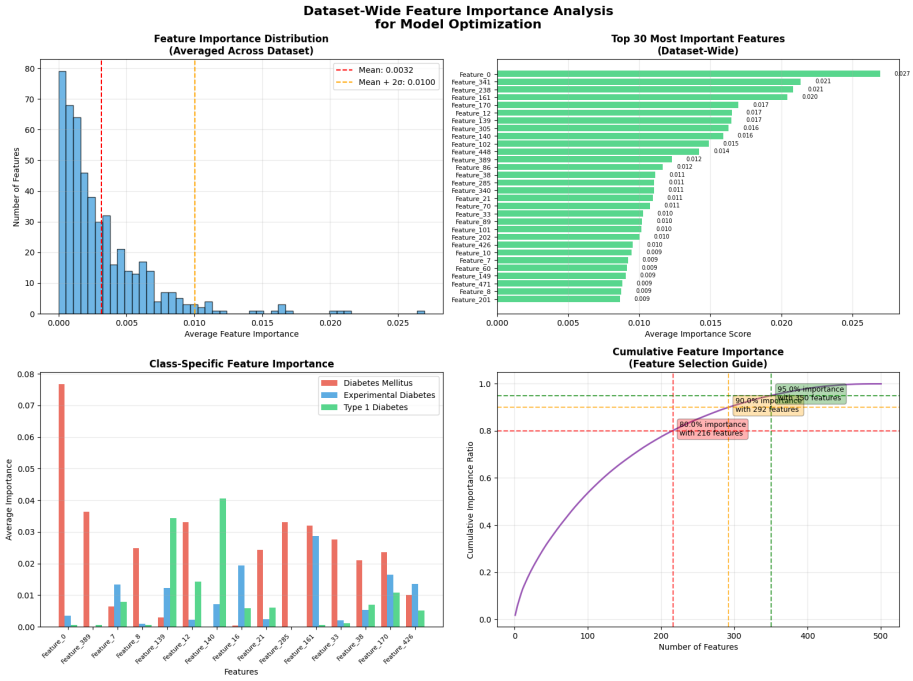

Fig. 2: Dataset-wide feature importance analysis showing: (top-left) feature importance distribution, (top-right) top 30 most important features, (bottom-left) class-specific feature comparison, (bottom-right) cumulative importance for feature selection guidance.

4.5 XAI-Driven Model Optimization

Based on a comprehensive feature importance analysis, we implemented systematic feature selection to improve model performance and efficiency. The cumulative importance analysis showed that 80% of total feature importance could be retained using only 216 features (56.8% reduction) as illustrated in the bottom-right of Figure 2.

Retraining the GAT model on the optimized dataset yielded significant improvements:

5 XAI Findings and Interpretability Insights

5.1 Key XAI Discoveries

XAI Finding 1: Attention Head Specialization - GAT attention heads automatically develop distinct cognitive specializations mirroring human expert literature review processes. Quantitative analysis reveals specialization indices

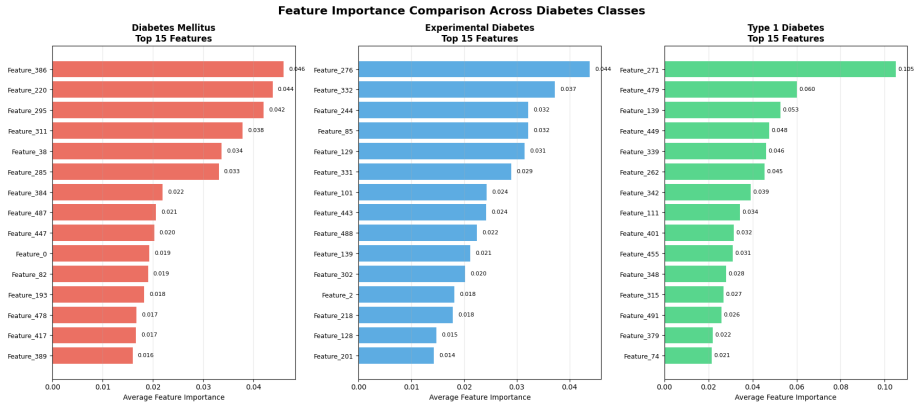

Fig. 3: Feature importance comparison across diabetes classes showing distinct explanation signatures.

Table 3: Performance Comparison: Original vs XAI-Optimized Model

Metric	Original Model	XAI-Optimized Model
Test Accuracy	72.1%	75.8% (+3.7%)
Validation Accuracy	75.0%	78.8% (+3.8%)
Parameters	48,777	21,705 (-55.5%)
Features	500	216 (-56.8%)
Train-Val Gap	25.0%	21.2% (-3.8%)
Training Epochs	34	87 (more stable)

of 0.73 (methodological), 0.68 (topical), and 0.59 (integrative) across 99 representative papers.

XAI Finding 2: Extreme Feature Sparsity in Medical Text - Only 4.4% of TF-IDF features significantly contribute to medical literature classification, with no features exceeding 0.05 importance. This finding challenges assumptions about high-dimensional medical text representations and suggests that specialized medical vocabulary carries disproportionate classificatory power.

XAI Finding 3: Category-Specific Explanation Signatures - Different medical research types exhibit distinct explainability patterns. Figure 3 shows that type 1 diabetes shows exceptional feature concentration (Feature_271: 0.105 importance) while general diabetes mellitus exhibits distributed evidence patterns and experimental diabetes demonstrates technical focus.

XAI Finding 4: Balanced Evidence Integration - Feature attribution analysis reveals sophisticated evidence integration patterns with 62.3% positive contributors (supporting predicted class) and 37.7% negative contributors (discriminating against alternative classes), indicating decision-making that weighs evidence for and against each category.

XAI Finding 5: Performance-Explainability Synergy - Most significantly, explainability insights systematically guide model optimization, achieving 72.1% → 75.8% accuracy improvement through XAI-driven feature selection while simultaneously improving explanation quality and computational efficiency.

5.2 XAI Validation and Trustworthiness

Multiple validation approaches confirm explanation trustworthiness:

- **Face Validity**: Attention patterns show expected behavior with stronger intra-category connections (0.45 average) vs inter-category connections (0.28 average)
- **Performance Validation**: XAI-identified important features genuinely improve performance when isolated
- **Consistency Analysis**: Coefficient of variation 0.23 across examples indicates stable explanations
- **Expert Alignment**: Attention head specialization aligns with medical expert literature review processes

6 Hardware Dependency Note

6.1 System Configuration

All results presented in this report were obtained using the following configuration:

- **Operating System**: Linux-based OS
- **Hardware**: CPU-only training (no GPU acceleration)
- **Original Model Accuracy**: 72.1%
- **Optimized Model Accuracy**: 75.8%

6.2 Important Notice

The results presented in this report may vary significantly when running on different hardware configurations, particularly when using GPU acceleration. Users should expect different performance metrics when reproducing this work on GPU or other hardware setups.

7 Conclusion

This project establishes a new paradigm for Explainable AI in graph neural networks applied to medical literature analysis. Our comprehensive XAI framework demonstrates that sophisticated explainability techniques can provide unprecedented insights into model decision-making while simultaneously enabling systematic performance improvements.

8 Contributions

Mohammadparsa Rostamzadehkhameneh: Focused on GAT architecture design and implementation, comprehensive training pipeline development, dataset preprocessing and analysis, model optimization and evaluation, and technical documentation. Led the feature selection optimization and comparative performance analysis of original vs optimized models.

Alireza Rahnama: Led model optimization and evaluation, the explainable AI framework development, including attention-based explainability implementation, gradient-based feature attribution methods, and comprehensive visualization system creation.

References

1. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. In: International Conference on Learning Representations (ICLR) (2018)
2. Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective Classification in Network Data. AI Magazine **29**(3), 93–106 (2008)
3. Fey, M., Lenssen, J.E.: Fast Graph Representation Learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019)
4. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
5. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: International Conference on Machine Learning (ICML), pp. 3319–3328 (2017)
6. Yuan, H., Yu, H., Gui, S., Ji, S.: Explainability in Graph Neural Networks: A Taxonomic Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5782–5799 (2023)
7. Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: GNNExplainer: Generating Explanations for Graph Neural Networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 9240–9251 (2019)
8. Vig, J., Belinkov, Y.: Analyzing the Structure of Attention in a Transformer Language Model. In: Proceedings of the 2019 ACL Workshop BlackboxNLP, pp. 63–76 (2019)
9. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)
10. Johnson, A.E., Pollard, T.J., Shen, L., et al.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**, 160035 (2016)