Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083509

ARTICLE

Interpretable Multimodal Post-Traumatic Stress Disorder Detection via Heterogeneous Graph Attention Networks on Real-World Clinical Data

Engin Seven[1,*], **Eylem Yucel**[1] **and Munevver Yildirim**[2]

[1]Department of Computer Engineering, Istanbul University-Cerrahpasa, Istanbul, Türkiye
[2]Department of Psychology, Demiroglu Science University, Istanbul, Türkiye

*Corresponding Author: Engin Seven. Email: engin.seven@ogr.iuc.edu.tr

Received: 05 April 2026; Accepted: 20 July 2026; Published: 15 September 2026

ABSTRACT: Objective, interpretable decision support for Post-Traumatic Stress Disorder (PTSD) screening remains a challenge in computational psychiatry, where existing methods either rely on costly neuroimaging or lack the diagnostic transparency required for clinical accountability. This study presents Multimodal HetGAT-PTSD, a heterogeneous graph attention network (HetGAT) that integrates unstructured clinical narratives with structured item-level responses from the PTSD Checklist for DSM-5 (PCL-5). The model operates under a graph topology constrained by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria to ensure structural alignment between clinical theory and graph-based learning. For each patient, a 25-node directed heterogeneous graph is constructed, connecting a BERTurk-encoded narrative node (a pre-trained Transformer-based language model) to 20 PCL-5 symptom nodes and 4 DSM-5 cluster nodes via typed edges that encode clinically grounded relation types. Two stacked graph attention layers with Gated Recurrent Unit (GRU)-style gating propagate multimodal evidence across the graph, and attention-based pooling produces a patient-level representation. The framework was evaluated on 418 real-world Turkish clinical records from a psychiatric training hospital using stratified 5-fold cross-validation. Multimodal HetGAT-PTSD achieved 91.16 ± 4.25% accuracy, 90.34 ± 4.58% F1-score, and 94.27 ± 4.63% AUC-ROC, significantly outperforming the questionnaire-only Multi-Layer Perceptron (MLP) baseline (ΔAUC-ROC = +0.0942, $p = 0.020$, Cohen's $d = 2.532$). The model deliberately trades approximately 6% accuracy relative to unconstrained text-only models ($p = 0.174$, corrected paired t-test) in exchange for mechanistic transparency through its DSM-5-constrained architecture. Explanation stability analysis confirmed perturbation robustness of 0.986 and cross-fold attention similarity of 0.984. The architecture provides a two-tier audit trail comprising edge-level attention weights and node-level pooling scores, enabling clinicians to trace diagnostic evidence from narrative content through PCL-5 items to DSM-5 symptom clusters without post-hoc explanation modules. These preliminary, single-site findings suggest that the framework represents a DSM-5-grounded interpretable research prototype for PTSD decision support that operates on routinely available clinical data, pending external validation on independent, multi-center cohorts.

KEYWORDS: PTSD detection; heterogeneous graph attention networks; multimodal fusion; clinical natural language processing; explainable AI; clinical decision support

1 Introduction

Post-Traumatic Stress Disorder (PTSD) is a debilitating psychiatric condition triggered by life-threatening trauma, disrupting fear conditioning and emotion regulation. Current diagnosis relies on clinical interviews and self-report instruments such as the PCL-5 (PTSD Checklist for DSM-5). Psychiatric disorders exhibit a dynamic symptom network structure in which symptoms causally reinforce one another [1],

and susceptibility to clinician bias and patient underreporting further underscores the need for objective, data-driven computational tools.

Graph Neural Networks (GNNs) have emerged as powerful architectures for relational clinical data. Huynh et al. achieved 97.76% PTSD detection accuracy by combining Generative Adversarial Networks (GANs) with Graph Convolutional Networks (GCNs) on Resting-state fMRI (rs-fMRI) data [2]. In comparison, Alahmadi et al. reported 96.60% accuracy using stacked deep learning on rs-fMRI scans [3]. However, the high acquisition cost and limited scalability of neuroimaging restrict routine clinical adoption. Unstructured patient narratives and standardized clinical scale responses offer a scalable, accessible alternative that integrates naturally into existing workflows [4,5]. Integrating free-text narratives with numerical Likert-scale responses within a single model requires heterogeneous data fusion. Graph-based multimodal fusion has proven effective for this purpose: Liu et al. [6] demonstrate that GNN-based multimodal fusion substantially improves depression detection, and heterogeneous graphs with distinct node and edge types have demonstrated improved disease prediction by modeling clinically meaningful structural heterogeneity in diagnostic graph architectures [7].

Building on these insights, we propose a Multimodal Heterogeneous Graph Attention Network for interpretable PTSD detection. The patient narrative serves as the central node, connected via typed directed edges to 20 PCL-5 symptom nodes and 4 Diagnostic and Statistical Manual of Mental Disorders (DSM-5) cluster nodes. A Graph Attention mechanism [8] dynamically weights symptom–narrative relationships, providing discriminative power and clinical transparency.

The main contributions of this paper are as follows:

- **Multimodal graph fusion:** HetGAT-PTSD integrates free-text clinical narratives (a pre-trained Transformer-based language model encoded via frozen BERTurk [9]) with structured PCL-5 item scores, achieving competitive diagnostic performance without neuroimaging.
- **Domain-guided structural regularization:** The graph topology explicitly embeds DSM-5 diagnostic organization through typed relations (Text → Question → Cluster), acting as a structural regularizer against spurious text patterns.
- **Real-world clinical validation:** The framework is validated on 418 Turkish psychiatric interviews and PCL-5 records, demonstrating cross-linguistic applicability in authentic clinical settings.
- **Multi-hop clinical interpretability:** Learned attention weights (α) trace how narrative evidence is routed through PCL-5 items to DSM-5 clusters, enabling clinician-inspectable relative importance indicators. This framework is further reinforced by explanation stability assessments and error analysis, ensuring that the model's diagnostic logic is both robust and clinically transparent.

The remainder of this paper is organized as follows. Section 2 reviews related work on neuroimaging-based, Natural Language Processing (NLP)-based, and graph-based PTSD detection. Section 3 describes the dataset, the domain-guided heterogeneous graph construction, the HetGAT-PTSD architecture, and the training protocol. Section 4 presents experimental results, ablation studies, statistical comparisons, and interpretability analysis. Section 5 discusses findings in the context of clinical applicability and limitations. Section 6 concludes the paper.

2 Related Work

PTSD is a diagnostically challenging disorder owing to its symptom complexity, high inter-patient variability, and frequent comorbidity with conditions such as traumatic brain injury and depression [10]. Accordingly, the field is shifting from traditional clinical scales toward deep learning architectures capable of processing multimodal clinical data.

2.1 Neuroimaging-Based Deep Learning

Resting-state fMRI has yielded high PTSD detection accuracies—97.76% via GAN-augmented GCNs [2] and 96.60% with stacked Convolutional Neural Networks [3]—and has enabled symptom trajectory prediction [11]. However, these results warrant cautious interpretation given small sample sizes and limited external validation [12], with systematic reviews confirming persistent bias risk and reporting deficiencies [4,13]. Beyond methodological concerns, the high cost and equipment requirements of neuroimaging restrict routine clinical use, and Rudin [14] argues that structurally interpretable architectures should be preferred over post-hoc explanations in high-risk settings. These constraints motivate models built on more accessible data sources.

2.2 Graph-Based Multimodal Fusion and Heterogeneous Architectures

Graph Neural Networks provide a natural framework for modeling non-Euclidean clinical structures [8,15]. Recent studies confirm their effectiveness in multimodal healthcare applications, including heterogeneous clinical data modeling [7]. Relation-specific message passing [15] and Graph Attention Networks [8] form the methodological basis, with recent architectures demonstrating diagnostic utility and interpretability in neuroimaging [16,17]. Graph-based multimodal fusion with heterogeneous node and edge types has been shown to improve disease prediction in clinical settings [7,18]. Text-attributed graph frameworks [19] demonstrate that pre-trained language representations integrate naturally into graph structures, as further confirmed in clinical settings where language model outputs are grounded in graph-structured medical knowledge [18,20]. These findings motivate positioning patient narratives as a context-providing central node within a heterogeneous graph that interacts with structured PCL-5 symptom nodes via type-dependent relations.

2.3 Positioning of the Current Study

While Electronic Health Record (EHR)- and scale-based approaches offer greater accessibility, recent evidence shows that EHR-specific bias mechanisms can inflate apparent model performance beyond true clinical utility [21]. This highlights the need for methods that couple accessible data sources with explainable and clinically grounded representation learning, rather than merely reporting high aggregate metrics.

This study addresses this gap by integrating two routinely obtainable clinical data sources: unstructured patient narratives and PCL-5 item-level responses, within a single heterogeneous graph formulation, where type-dependent attention mechanisms provide both diagnostic performance and traceable, DSM-5-grounded interpretability. This approach aligns with international guidelines for trustworthy clinical AI deployment [22,23] and empirical evidence on the role of explainability in building clinician trust [24].

3 Materials and Methods

3.1 Dataset and Participants

The dataset comprises clinical interview narratives from 418 patients diagnosed with or evaluated for PTSD at Istanbul Bakirkoy Prof. Dr. Mazhar Osman Mental Health and Neurological Diseases Training and Research Hospital. Ethical approval was obtained from the Hamidiye University Scientific Research Ethics Committee (No: 29.02.2024-25936), and all records were anonymized in compliance with the Turkish Personal Data Protection Law. Key dataset characteristics stratified by diagnostic group are summarized in Table 1.

Table 1: Dataset characteristics and PCL-5 score statistics stratified by diagnostic group (PTSD vs. Non-PTSD). Values represent mean ± standard deviation unless stated otherwise.

Characteristic	PTSD (n = 200)	Non-PTSD (n = 218)
Class distribution	47.8%	52.2%
PCL-5 Total Score		
Mean ± SD	34.47 ± 6.44	27.43 ± 2.94
Median	36.0	28.0
DSM-5 Cluster Scores (Mean ± SD)		
B (Intrusion)	9.23 ± 2.92	7.12 ± 2.50
C (Avoidance)	3.90 ± 1.77	2.94 ± 1.69
D (Neg. Cognition/Mood)	11.88 ± 3.42	9.68 ± 2.56
E (Arousal/Reactivity)	9.46 ± 3.04	7.70 ± 2.56
Clinical Narrative Length (tokens)		
Mean ± SD	283.6 ± 70.9	234.1 ± 37.8

Note: PTSD: Post-Traumatic Stress Disorder; SD: Standard Deviation; PCL-5: PTSD Checklist for DSM-5. ± symbol indicates the variation around the mean value.

Each patient record comprises two complementary modalities:

- **Structured PCL-5 item scores:** The PTSD Checklist for DSM-5 (PCL-5) is a 20-item self-report instrument assessing symptom severity over the past month. Each item is rated on a Likert scale from 0 ("Not at all") to 4 ("Extremely").
- **Unstructured clinical narratives:** Free-text interview transcripts in which patients describe their traumatic experiences and psychological state in daily life.

Ground-truth labels were assigned by board-certified psychiatrists through a two-stage diagnostic process: (1) supervised administration of the PCL-5 questionnaire, and (2) a structured clinical interview based on the DSM-5 criteria. The final diagnosis reflected the clinician's integrated judgment of both sources.

3.2 Domain-Guided Heterogeneous Graph Modeling

Rather than relying on post-hoc explanations of unconstrained models [14,25], our approach encodes the DSM-5 diagnostic structure directly into the graph topology as a structural prior [26,27]. For each patient, a directed heterogeneous graph $G = (V, E)$ is constructed with 25 nodes across three semantic levels: a single text node ($\mathrm{h}_{\mathrm{text}} \in \mathbb{R}^{\mathrm{D}}$) encoding the clinical narrative; 20 *question nodes* ($h_{q_i} \in \mathbb{R}^{\mathrm{D}}$) corresponding to PCL-5 items q_1–q_{20}; and 4 cluster nodes representing DSM-5 symptom clusters (Intrusion, Avoidance, Negative Cognition/Mood, Arousal/Reactivity), initialized via mean-pooling of constituent question embeddings followed by a learnable projection.

The edge set E encodes four relation types: Text → Question edges model the influence of narrative content on individual symptom assessments; bidirectional Question → Question (intra-cluster) edges capture within-cluster symptom co-occurrence [27]; Question → Cluster edges aggregate item-level evidence toward cluster representations; and bidirectional Cluster → Cluster edges permit cross-cluster interaction, reflecting empirical symptom overlap across DSM-5 clusters [26].

A natural question is whether the DSM-5 structural constraints could be imposed via attention masking within a standard Transformer, rather than through a heterogeneous graph. While binary attention masks

can restrict which tokens attend to which, they cannot encode relation-specific transformations: in HetGAT-PTSD, each edge type (Text → Question, Question → Question, Question → Cluster, Cluster → Cluster) is processed by a dedicated weight matrix W_t, enabling the model to learn distinct message functions for clinically different interactions. A masked Transformer would apply the same parametric transformation regardless of relation type. Furthermore, the three-level hierarchical structure (narrative → symptom items → diagnostic clusters) emerges naturally from multi-hop message passing in the graph, whereas a flat Transformer architecture would require additional architectural modifications to support such hierarchical routing. These properties make the heterogeneous graph formulation a more natural fit for encoding the structured, multi-relational nature of DSM-5 diagnostic reasoning.

3.3 Multimodal HetGAT-PTSD Architecture

The proposed architecture consists of (i) modality-specific node encoders, (ii) two stacked heterogeneous graph attention layers, (iii) attention-based graph pooling, and (iv) a binary classifier. The overall architecture of the proposed Multimodal HetGAT-PTSD framework is illustrated in Fig. 1. The graph topology is fixed across patients; heterogeneity refers to the coexistence of distinct node types and relation-specific message functions (W_t), not per-sample topological variation. While the edge structure is shared, node representations are patient-specific: each patient's PCL-5 scores and clinical narrative produce unique node embeddings that propagate through the graph.

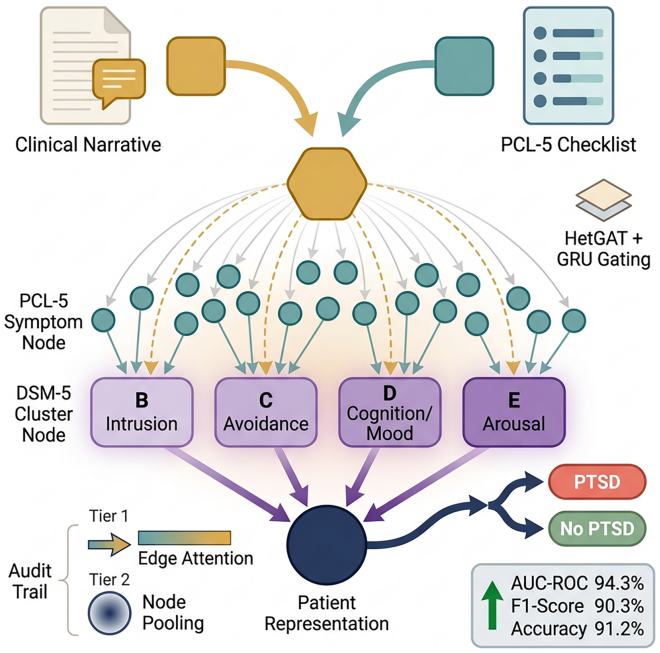

Figure 1: Overall architecture of the Multimodal HetGAT-PTSD framework, illustrating the end-to-end pipeline from multimodal input encoding to DSM-5-constrained graph construction, heterogeneous graph attention, and binary classification.

Text encoder: The clinical narrative $x^{(n)}$ is encoded using a frozen BERTurk model [9]. The [CLS] token $c \in \mathbb{R}^{768}$ is projected to a 128-dimensional intermediate representation via a two-layer MLP (Gaussian Error Linear Unit (GELU) activation, layer normalization), then linearly mapped to the graph hidden dimension D = 64. See Eq. (1):

$$\mathbf{h}_{\text{text}} = \mathbf{W}_{\text{proj}} \cdot \text{MLP}(\mathbf{c}) \in \mathbb{R}^{64} \tag{1}$$

PCL-5 item encoder: Each raw item score $s_i^{(n)} \in \{0, 1, 2, 3, 4\}$ is normalized to $\tilde{s}_i = s_i/4 \in [0,1]$ and encoded via a two-layer MLP (f_q: $1 \to 32 \to 32$), then projected to D = 64 and combined with a learnable cluster embedding. See Eq. (2):

$$\mathbf{h}_{q_i} = \mathbf{W}^{(Q)}{}_{\text{proj}} \left[f_q(\tilde{s}_i) + \mathbf{e}_{\kappa(i)} \right] \in \mathbb{R}^{64} \tag{2}$$

Cluster node initialization: Each cluster node aggregates its constituent question embeddings via mean pooling followed by a learnable projection f_c, yielding $\mathrm{h}_{ck} \in \mathbb{R}^{\mathrm{D}}$ for cluster $k \in \{1, \ldots, 4\}$.

Heterogeneous graph attention layers: Two stacked attention layers propagate the information across the graph. For each directed edge ($j \to i$) of type t, a type-specific message and attention coefficient are computed as defined in Eq. (3):

$$\mathbf{m}^{(t)}{}_{ij} = \varphi\left(\mathbf{W}_t \left[\mathbf{h}_j \| \mathbf{h}_i\right]\right), \alpha_{ij} = \frac{\exp\left(\mathbf{w}^\top \mathbf{m}^{(t)}{}_{ij}\right)}{\sum_{k \in N(i)} \exp\left(\mathbf{w}^\top \mathbf{m}^{(t)}{}_{ik}\right)} \tag{3}$$

where $\mathbf{W}_t$ is a relation-specific weight matrix [15] and $\varphi(\cdot)$ is a nonlinear activation function. Each node aggregates incoming messages as $\mathbf{a}_i = \sum_{j \in N(i)} \alpha_{ij} \mathbf{m}^{(t)}{}_{ij}$ and updates its state via a Gated Recurrent Unit (GRU) style gated mechanism with layer normalization and dropout. Because relation-specific matrices W_t process clinically distinct edge types independently and graph edges are grounded in DSM-5 nosology, the resulting attention weights {α_ij} constitute clinician-inspectable importance indicators, structurally aligned with the spirit of the Right for the Right Reasons criterion for model trustworthiness [28] and the contextual explainability requirements identified for clinical ML adoption [29], within the broader framework of trustworthy AI in healthcare [25]. We note that these weights constitute relative importance indicators rather than causal explanations [30,31]. Furthermore, our integration of [28] is architectural, constraining which evidence the model can use, rather than a literal implementation of Ross et al.'s training-time explanation-penalty mechanism.

3.4 Graph Readout and Classification

After two message-passing layers, a patient-level graph representation is obtained via attention-based pooling is obtained via attention-based pooling (Eq. (4)):

$$\beta_v = \text{softmax}\left(g(\mathbf{h}_v)\right), \mathbf{z}_{\text{patient}} = \sum_{v \in V} \beta_v \mathbf{h}_v \tag{4}$$

where $g(\cdot)$ is a learnable scoring function. The pooling weights $\{\beta_v\}$ assign differential importance to the text, question, and cluster nodes, providing a sample-specific decomposition of the diagnostic evidence. Together with the edge-level attention weights $\{\alpha_{ij}\}$ from Eq. (3), these form a two-tier audit trail that allows clinicians to trace which symptom–narrative interactions drove message passing and which node types contributed most to the final diagnosis. A feed-forward classifier then maps $\mathbf{z}_{\text{patient}}$ to a probability $P = \sigma(f_{clf}(\mathbf{z}_{\text{patient}}))$, and the model is trained end-to-end using class-weighted binary cross-entropy with $\mathrm{w}^+ = 1.3$ to penalize false negatives.

3.5 Error Analysis Methodology

To characterize failure modes, we categorized validation predictions into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) following standard classification evaluation

practice [32]. We then compared HetGAT errors with BERTurk predictions to identify unique errors (one model correct, other wrong) vs. shared errors (both models wrong). Additionally, we analyzed DSM-5 cluster profiles for error cases to identify comorbidity patterns.

3.6 Explanation Stability Assessment

To validate explanation reproducibility, we assessed stability using five complementary metrics following established Explainable Artificial Intelligence (XAI) evaluation practice [33]: (1) perturbation robustness by injecting Gaussian noise ($\sigma \in \{0.01, 0.05, 0.10\}$) into text embeddings and measuring attention stability via Kendall's τ, cosine similarity, and Jaccard overlap [34]; the three metrics were averaged within each noise level, and the resulting three noise-level scores were averaged to yield the single reported robustness score (0.986); (2) cross-fold weight similarity via pairwise cosine similarity between attention vectors across the five cross-validation (CV) folds; and (3) top-10 feature overlap via Jaccard similarity to assess convergence on core diagnostic features [35]. (4) cross-fold Kendall's τ rank correlation to assess whether the precise ordering of attended items is preserved across folds; and (5) rank discrepancy, quantifying the average positional difference in top-10 rankings across fold pairs.

3.7 Training Protocol and Evaluation Metrics

We perform stratified 5-fold cross-validation [36,37], preserving class proportions in each fold. Since each patient contributes a single record, the protocol is inherently patient-level and prevents subject-level leakage. Model parameters (excluding the frozen BERT encoder) are optimized using AdamW with an initial learning rate of 5×10^{-4} and weight decay of 10^{-4}. The learning rate was dynamically adjusted throughout training via cosine annealing scheduling (η_min = 10^{-5}), which gradually reduces the learning rate from its initial value toward the specified minimum to stabilize late-stage convergence. Gradient clipping (max norm = 1.0) and 4-head attention in each Graph Attention Network (GAT) layer were also applied. The objective is the class-weighted binary cross-entropy, with w^+ penalizing false negatives. Early stopping (patience = 8, validation F1) selects the best model per fold. Predictions use a fixed threshold of 0.5. We report Accuracy, Precision, Recall, F1-score, ROC-AUC, and AUC-PR [32,38]. All hyperparameters were set a priori and kept fixed across all five folds (learning rate = 5×10^{-4}, weight decay = 1×10^{-4}, hidden dimension = 64, attention heads = 4, batch size = 16, early stopping patience = 8, gradient clipping = 1.0, classification threshold = 0.5). No hyperparameter tuning was performed on the validation folds used for performance evaluation.

All experiments were conducted on an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB VRAM. The proposed model comprises 371,170 trainable parameters (excluding the frozen BERTurk encoder with 110,617,344 parameters). Total 5-fold training time was 2860 s, compared to 1001 s for BERTurk-only (A2) and 170 s for the PCL-5 MLP baseline (A1). At inference, the model processes a single patient in 24.3 ms, confirming its suitability for real-time clinical decision support.

4 Experiments and Results

4.1 Overall Classification Performance

We compared the proposed Multimodal HetGAT-PTSD against single-modal and multimodal baselines using stratified 5-fold cross-validation. Results are summarized in Table 2.

Table 2: Ablation summary for A1, A2, A3, and A8 (mean and standard deviation across folds). Best value per metric among these variants is shown in bold.

Variant	Setting	Acc	Precision	Recall	F1	AUC-ROC	AUC-PR
A1	PCL-5 MLP Only	0.7967 ± 0.0283	0.7768 ± 0.0365	0.8100 ± 0.0374	0.7922 ± 0.0271	0.8485 ± 0.0212	0.8711 ± 0.0272
B1	LR (PCL-5 Total)	0.9330 ± 0.0123	1.0000 ± 0.0000	0.8600 ± 0.0255	0.9245 ± 0.0147	0.8777 ± 0.0319	0.9302 ± 0.0156
A2	BERTurk Only	**0.9713 ± 0.0143**	0.9708 ± 0.0232	**0.9700 ± 0.0292**	**0.9699 ± 0.0150**	**0.9940 ± 0.0055**	**0.9943 ± 0.0050**
A3	Concat Fusion (No Graph)	0.9665 ± 0.0158	**0.9729 ± 0.0431**	0.9600 ± 0.0339	0.9651 ± 0.0155	0.9922 ± 0.0053	0.9921 ± 0.0057
A8	Multimodal HetGAT-PTSD (Full)	0.9116 ± 0.0425	0.9459 ± 0.0504	0.8650 ± 0.0464	0.9034 ± 0.0458	0.9427 ± 0.0463	0.9583 ± 0.0300

Note: Values represent the mean ± standard deviation. Bold text indicates the superior performance in the respective metric. AUC-ROC: Area Under the Receiver Operating Characteristic Curve; AUC-PR: Area Under the Precision-Recall Curve; PCL-5: PTSD Checklist for DSM-5; MLP: Multilayer Perceptron; Acc: Accuracy; LR: Logistic Regression.

The clinical baseline (A1) achieved a 79.67% accuracy and an 84.85% AUC-ROC. The pure language model (A2) and early concatenation fusion (A3) yielded the highest raw accuracies at 97.13% and 96.65%, respectively. The proposed Multimodal HetGAT-PTSD (A8) achieved 91.16 ± 4.25% accuracy, 90.34 ± 4.58% F1-score, and 94.27 ± 4.63% AUC-ROC. Although lower than A2 and A3 (accuracy gap: −5.97% and −5.49%, respectively), these differences do not reach statistical significance at N = 418 (p = 0.174 and p = 0.202, corrected paired t-test; see Section 4.3)—a result consistent with limited statistical power in a small cohort, as the corrected paired t-test intentionally overestimates variance to prevent Type I errors, and should not be interpreted as evidence of equivalence. The large effect sizes (Cohen's $d > 1.0$) further confirm a practically meaningful performance gap in favor of A2 and A3. These differences reflect an intentional accuracy–interpretability trade-off: unlike A2 and A3, A8 encodes the DSM-5 diagnostic domain knowledge as a structural regularizer, enabling mechanistic transparency and a verifiable decision pathway essential for psychiatric decision support. Critically, neither A2 nor A3 can associate predictions with specific DSM-5 symptom clusters, rendering them unsuitable where clinical justification is as important as predictive accuracy. Furthermore, A8 significantly outperforms A1 by 11.49 percentage points in accuracy and 9.42 percentage points in AUC-ROC (p = 0.020), confirming that the proposed architecture captures complementary information beyond structured questionnaire scores alone. Calibration analysis yielded a Brier score of 0.1402, an expected calibration error (ECE) of 0.1737, and a maximum calibration error (MCE) of 0.3965, indicating reasonable overall calibration with localized deviations in high-confidence regions. To further assess class-wise reliability, we examined per-class performance and prediction stability.

As shown in the ROC curves (Fig. 2a), the model achieved a mean AUC of 93.70 ± 4.60% across five folds. The mean average precision of 93.80 ± 2.80% (Fig. 2b) further confirms robust performance across precision-recall trade-offs. The confusion matrix (Fig. 2c) shows that the model achieved 88.41% precision and 94.50% recall for the non-PTSD class, and 93.51% precision and 86.50% recall for the PTSD class. The higher recall for non-PTSD (94.50%) and higher precision for PTSD (93.51%) reflect a clinically desirable tendency to minimize false-positive diagnoses, where overdiagnosis carries significant psychological and therapeutic consequences.

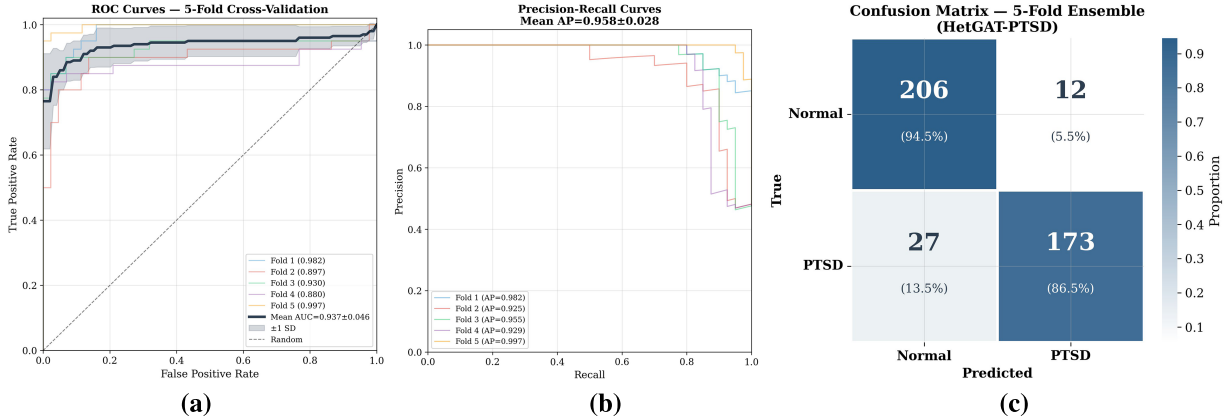

Figure 2: Performance visualization: (**a**) ROC curves; (**b**) Precision–recall curves; (**c**) Confusion matrix across all 5-fold predictions (aggregate accuracy: 91.16%). Cell-derived metrics may differ marginally from fold-mean values in Table 2, as per standard cross-validation reporting convention.

A t-SNE projection of the final graph-level embeddings (Fig. 3) reveals clear spatial separation between PTSD and non-PTSD representations, confirming that the message-passing mechanism produces a discriminative latent space from the multimodal inputs.

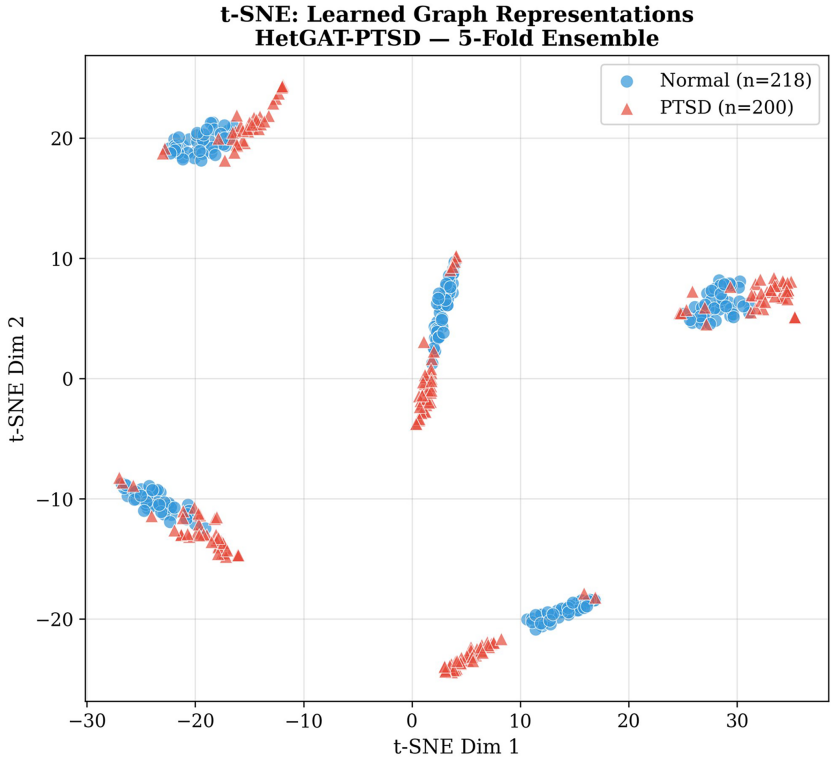

Figure 3: t-SNE visualization of the learned graph-level embeddings for PTSD classification. The clear spatial separation between PTSD and non-PTSD clusters confirms the discriminative latent space produced by the message-passing mechanism.

4.2 Ablation Studies on Graph Topology

To isolate the contribution of individual architectural components, we conducted ablation studies on the graph structure, edge heterogeneity, domain knowledge injection, gating, and network depth. Results are reported in Table 3.

Table 3: Component ablation analysis of the Multimodal HetGAT-PTSD model. Delta values denote the performance difference between the full model and the ablated variant (Δ = Full − Ablated). Positive values indicate full-model superiority; negative values reflect a deliberate design trade-off.

Component Removed	Model	ΔAUC-ROC	ΔF1	Interpretation
—	HetGAT-PTSD (Full)	—	—	Reference
Graph Structure	Concat Fusion (A3)	−0.0495	−0.0617	Performance–interpretability trade-off
Edge Heterogeneity	Homo GAT (A4)	+0.0039	−0.0151	Typed edges improve AUC; reduce F1
DSM-5 Cluster Nodes	w/o Cluster (A5)	+0.0133	+0.0027	Domain knowledge beneficial
GRU Gating	w/o GRU (A6)	+0.0104	+0.0030	Gating mechanism beneficial
2nd GAT Layer	1-Layer GAT (A7)	−0.0041	−0.0360	Second layer necessary for two-hop interpretability

Note: Negative Δ values indicate that the ablated variant achieves higher raw metrics, reflecting a deliberate accuracy–interpretability trade-off in the full model. AUC-ROC: Area Under the Curve-Receiver Operating Characteristic; AUC: Area Under the Curve; GRU: Gated Recurrent Unit; GAT: Graph Attention Network; DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition.

- **Graph Structure (A3):** A3 achieves higher raw metrics than the full model (ΔAUC-ROC = −0.0495, ΔF1 = −0.0617), consistent with the accuracy–interpretability trade-off in clinical AI. Operating without structural constraints, A3 cannot distinguish symptom types or DSM-5 clusters, rendering its decision pathway clinically unverifiable. A8 accepts this marginal cost to provide a transparent, domain-grounded inference pathway auditable by clinicians.
- **Edge Heterogeneity (A4):** Collapsing typed edges into a single type produced a mixed effect: heterogeneous edges marginally improved AUC-ROC (ΔAUC-ROC = +0.0039), while the homogeneous variant achieved a slightly higher F1 (ΔF1 = −0.0151). This indicates that clinically typed edges primarily benefit ranking-based discrimination (AUC) at a minor cost to class-balanced classification (F1), while preserving semantic relation types required for interpretability. While AUC-ROC and Recall are the primary metrics for clinical screening evaluation, F1 was used for model selection during training as it balances precision and recall at a fixed threshold.
- **DSM-5 Cluster Nodes (A5):** Removing cluster nodes decreased both metrics (ΔAUC-ROC = +0.0133, ΔF1 = +0.0027), validating their role as a structural regularizer that embeds domain knowledge into the graph topology and underpins the two-tier audit trail.
- **GRU Gating (A6):** Disabling the GRU update gate reduced both metrics (ΔAUC-ROC = +0.0104, ΔF1 = +0.0030), confirming its contribution to stable node-state updates across heterogeneous message types.
- **Network Depth (A7):** A7 achieves marginally higher raw metrics (ΔAUC-ROC = −0.0041, ΔF1 = −0.0360). The second attention layer does not improve discriminative performance; rather, it

enables two-hop propagation along the Narrative → Question → Cluster pathway, and is retained as an interpretability requirement rather than a raw-performance optimizer.

4.3 Statistical Significance of Model Comparisons

Standard paired t-tests applied to k-fold cross-validation results violate the independence assumption because training sets overlap across folds. The corrected paired t-test proposed by Nadeau and Bengio [39] adjusts the variance estimate to account for this dependency, providing a more conservative and statistically valid comparison. Among alternatives, McNemar's test requires a single fixed test set rather than cross-validated folds, and the Wilcoxon signed-rank test lacks sufficient power with only five observations. We therefore adopted the corrected paired t-test as the most suitable option for our experimental design, while acknowledging that statistical power remains inherently limited with five folds, which is why we additionally report Cohen's d to quantify practical effect size.

We conducted pairwise comparisons using the corrected paired t-test of Nadeau and Bengio [39], which accounts for the dependency structure in k-fold cross-validation. Results are reported in Table 4.

Table 4: Statistical significance of pairwise model comparisons using the corrected paired t-test across 5-fold cross-validation. ΔAUC = AUC (A8) − AUC (Baseline); positive values indicate A8 outperforms the baseline. Cohen's d quantifies practical effect size.

Comparison	ΔAUC	t	p-Value	Cohen's d (AUC)
A8 vs. A1 (PCL-5 MLP)	+0.0942	+3.775	0.020*	+2.532 (Large)
A8 vs. A2 (BERTurk)	−0.0513	−1.650	0.174	−1.107 (Large)
A8 vs. A3 (Concat Fusion)	−0.0495	−1.524	0.202	−1.022 (Large)

Note: *$p < 0.05$. A8 = Multimodal HetGAT-PTSD (proposed model); A1 = PCL-5 MLP baseline; A2 = BERTurk only; A3 = Concat Fusion (No Graph).

A8 significantly outperforms the PCL-5 MLP baseline (A1) in AUC-ROC (Δ = +0.0942, t = 3.775, p = 0.020, Cohen's d = 2.532). Differences relative to A2 and A3 are not statistically significant (p = 0.174 and p = 0.202, respectively), consistent with the conservative nature of the corrected t-test under 5-fold cross-validation on 418 patients [39]. The large effect sizes (Cohen's d = −1.107 and −1.022, both Large) indicate a negative effect direction—that is, A8 underperforms both A2 and A3 on raw AUC-ROC—even though the difference does not reach statistical significance at this sample size. Crucially, A8's primary advantage over A2 and A3 is mechanistic transparency: it provides a two-tier audit trail comprising edge-level attention weights $\{\alpha_{ij}\}$ and node-level pooling weights $\{\beta_v\}$, enabling clinicians to trace the diagnostic pathway from narrative evidence through PCL-5 items to DSM-5 clusters without any external explanation module.

4.4 Explanation Stability

Stability was assessed via four complementary metrics (Table 5): perturbation robustness (0.986), cross-fold cosine similarity (0.984), top-10 Jaccard overlap (0.551), and Kendall's τ rank correlation (0.027 ± 0.543). Following established practice for correlation-based interpretability metrics [33], we adopt thresholds of >0.7 for correlation metrics, >0.6 for overlap metrics, and <0.3 for discrepancy metrics as indicative of reliable explanations. By this criterion, perturbation robustness and cosine similarity indicate strong stability. In contrast, the near-zero Kendall's τ indicates that the precise rank ordering of attended symptom items is not consistently preserved across folds—even though the magnitude distribution (cosine similarity) and top-k membership (Jaccard overlap) are. We interpret this as evidence that the model reliably identifies which symptom clusters matter, without committing to a single fixed internal ranking among closely weighted

items. Moderate top-10 overlap reflected PTSD's clinical heterogeneity [26]. The text-to-question attention weights, pooled across all five folds ($n = 100$), remained non-uniform ($\sigma = 0.125$, max/min ratio = 5.1×) rather than diffuse, indicating that the model consistently concentrates evidence on a subset of PCL-5 items rather than distributing attention uniformly across all 20 questions.

Table 5: Explanation stability metrics.

Metric	Value
Perturbation Robustness	0.986
Cross-Fold Cosine Similarity	0.984 ± 0.012
Top-10 Jaccard Overlap	0.551 ± 0.139
Cross-Fold Kendall's τ	0.027 ± 0.543
Rank Discrepancy	0.429 ± 0.173

We acknowledge that explanation stability, as measured here, is a necessary but not sufficient condition for faithfulness; stable attention distributions do not guarantee that weights reflect the model's true reasoning process [30,40]. However, two architectural properties of HetGAT-PTSD mitigate this concern: the graph topology restricts attention to DSM-5-predefined pathways rather than arbitrary input regions, and relation-specific weight matrices (W_t) constrain message passing to semantically distinct channels. These design choices narrow the gap between attention weights and model reasoning compared to unconstrained architectures, though a formal faithfulness evaluation via counterfactual analysis remains for future work.

4.5 Error Analysis

HetGAT-PTSD achieved 91.16% accuracy with an exceptionally low false positive rate (2.9%, 12 cases), critical for psychiatric screening, where unnecessary diagnoses can cause patient harm. Comparison with BERTurk revealed that only 16.7% of false positives and 0% of false negatives were shared errors, indicating that the majority of HetGAT's errors (37 of 39, 94.9%) were correctly classified by the simpler BERTurk-only baseline. Critically, HetGAT produced 37 unique errors (8.9% of total), demonstrating that DSM-5-constrained reasoning introduces a measurable additional error cost in exchange for mechanistic transparency that black-box models cannot offer (Table 6).

Table 6: Error analysis showing HetGAT's unique errors represent only 8.9% of total cases.

Category	N Cases	%
True Positives	173	41.4%
True Negatives	206	49.3%
False Positives	12	2.9%
Shared with BERTurk	2	16.7%
Unique to HetGAT	10	83.3%
False Negatives	27	6.5%
Shared with BERTurk	0	0.0%
Unique to HetGAT	27	100.0%
Total Unique Errors	**37**	**8.9%**

Comorbidity analysis of misclassified cases revealed distinct symptom profiles (Table 7). False positives exhibited elevated Negative Cognition/Mood (D, mean 10.2) and Arousal (E, mean 8.0) scores alongside high overall PCL-5 severity (29.8 ± 1.5), consistent with subthreshold presentations that mimic PTSD without meeting full diagnostic criteria. False negatives showed markedly lower and more variable overall severity (22.6 ± 7.5) with comparatively muted Avoidance (C, mean 2.7) scores, suggesting the model under-detects cases with atypical or attenuated avoidance presentations.

Table 7: DSM-5 symptom cluster profiles of misclassified cases (false positives and false negatives).

Error Type	N Cases	PCL-5 Total	Intrusion (B)	Avoidance (C)	Neg. Cognition (D)	Arousal (E)
FP	12	29.8 ± 1.5	7.0	4.6	10.2	8.0
FN	27	22.6 ± 7.5	6.3	2.7	7.9	5.7

Note: FP: False Positive; FN: False Negative; PCL-5: PTSD Checklist for DSM-5.

4.6 Clinical Interpretability

A key advantage of the proposed architecture is that the multi-hop attention weights are directly inspectable, enabling clinicians to trace how narrative evidence propagates through PCL-5 items to DSM-5 clusters. Together with the graph-level pooling weights, these form a two-tier audit trail conceptually aligned with the spirit of the Right for the Right Reasons framework [28]. Unlike Ross et al.'s original formulation, which explicitly penalizes reliance on annotated incorrect input regions during training, our approach enforces this alignment structurally: the DSM-5-constrained graph topology restricts attention pathways at the architecture level, so that whatever evidence the model uses is, by construction, routed through clinically grounded symptom–narrative associations. The model not only produces correct diagnoses but does so via symptom–narrative associations that are grounded in DSM-5 nosology.

Multi-Hop Attention Analysis (Fig. 4a): At the first hop, the model assigns the highest text-to-question attention weights (approximately 0.6) to items in the Avoidance cluster (C), specifically q6 and q7, in both diagnostic groups. The PTSD group exhibits a more selective distribution concentrated on specific symptom nodes, whereas the control group displays a broader baseline distribution. At the second hop (Fig. 4b), the graph topology constrains each question to route exclusively to its DSM-5 cluster; the attention mechanism therefore learns patient-specific importance weights along each predefined pathway rather than discovering these mappings *de novo*. High attention coefficients concentrate within the Avoidance (C) and Intrusion (B) blocks for the PTSD group, and the differential analysis reveals that attention intensity varies substantially between diagnostic groups, confirming that the model adapts its hierarchical weighting to the individual clinical profile. This constitutes mechanistic transparency: the inference pathway is not only constrained by domain knowledge but is also individually traceable per patient.

Cluster-to-Cluster Attention (Fig. 5): The cluster interaction heatmaps reveal strong connectivity between Intrusion (B) and Avoidance (C), with peak attention weights of approximately 0.221, consistent with the clinical understanding that intrusive memories trigger avoidant behaviors. The PTSD cohort shows consistently higher weights for interactions involving Negative Cognition/Mood (D) and Arousal (E), and the differential heatmap highlights an intensified Intrusion → Avoidance pathway in PTSD cases. This pattern suggests that the model captures a symptom cascading effect in which the activation of one cluster reinforces another, consistent with network models of interconnected PTSD symptom structure [1,27]. Although the categorical DSM-5 framework does not formalize this relational property, our DSM-5-constrained cluster topology can approximate it, forming a clinically verifiable component of the audit trail.

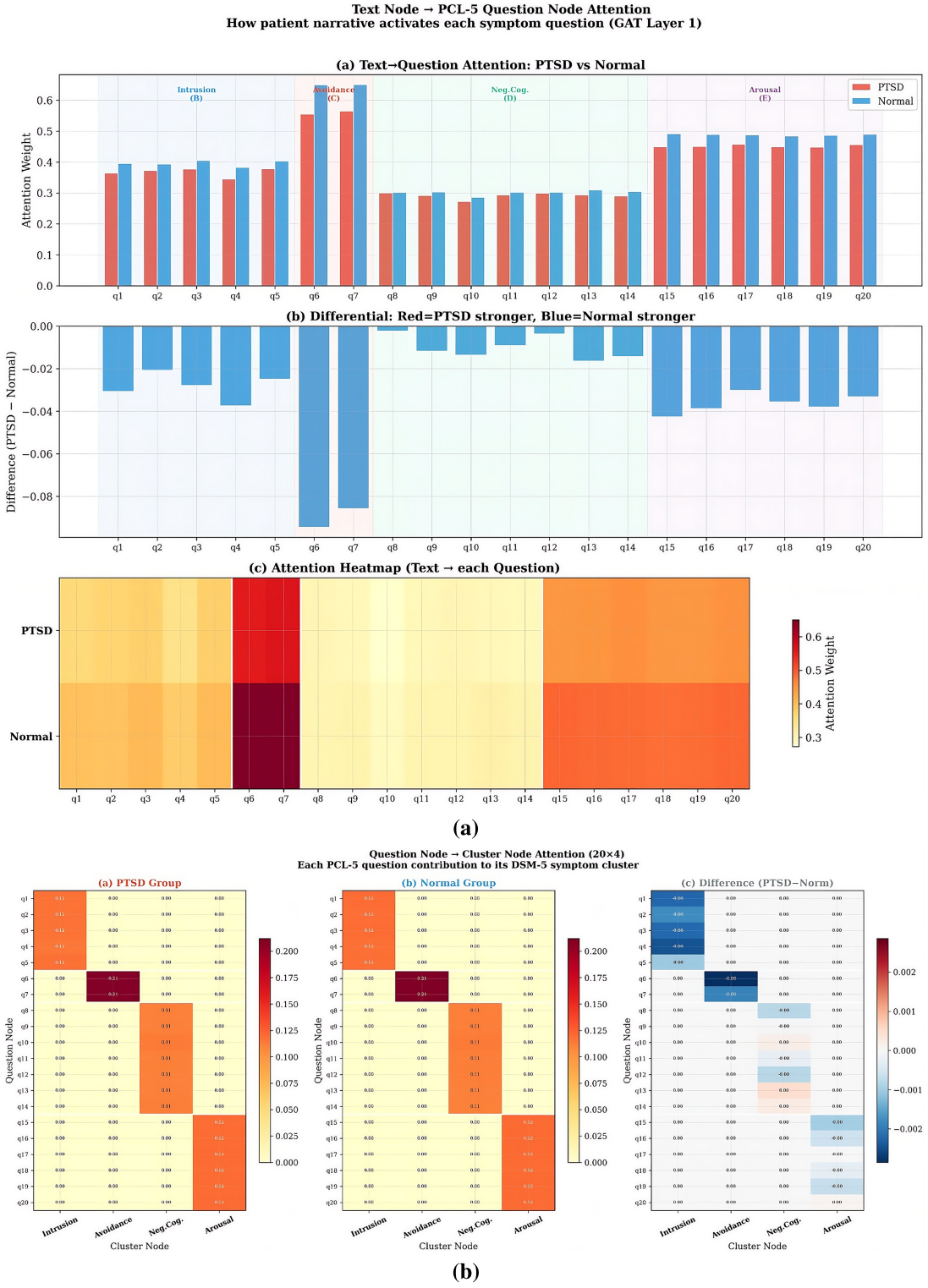

Figure 4: Multi-hop attention maps in HetGAT-PTSD: (**a**) Text-to-question attention; (**b**) Question-to-cluster attention.

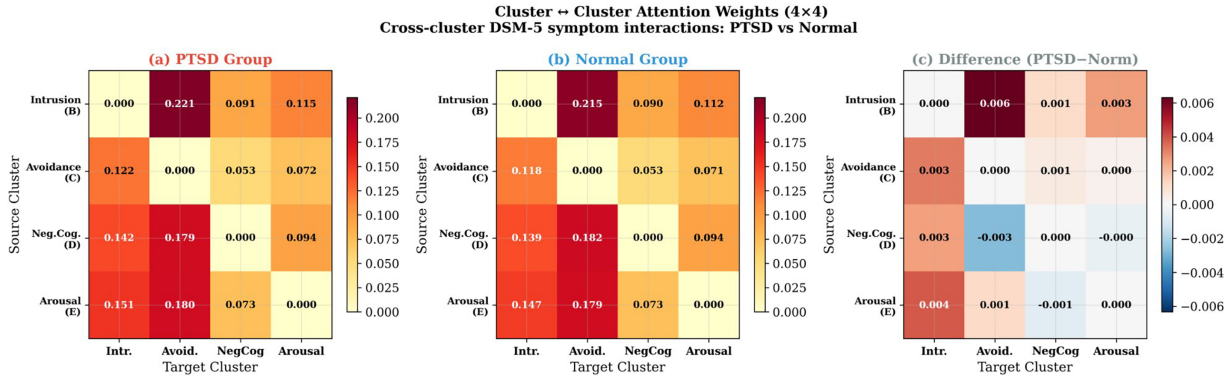

Figure 5: Global interpretability analysis of DSM-5 cluster interaction patterns: Illustrating the relational dependencies between diagnostic clusters within the heterogeneous graph topology.

Global Node Importance (Fig. 6): The Text Node holds the highest mean attention weight (approx. 0.048) in the PTSD cohort, confirming the narrative's role as the primary information anchor. Among PCL-5 items, q6 and q7 remain the most influential, consistent with the local attention maps. Differential analysis shows a significant positive deviation for the Text Node and q7 in the PTSD group, indicating selective prioritization of narrative expressions and avoidant symptoms for pathology detection. These node-level weights constitute the second tier of the audit trail, recording which modality and symptom items drove the final diagnostic representation.

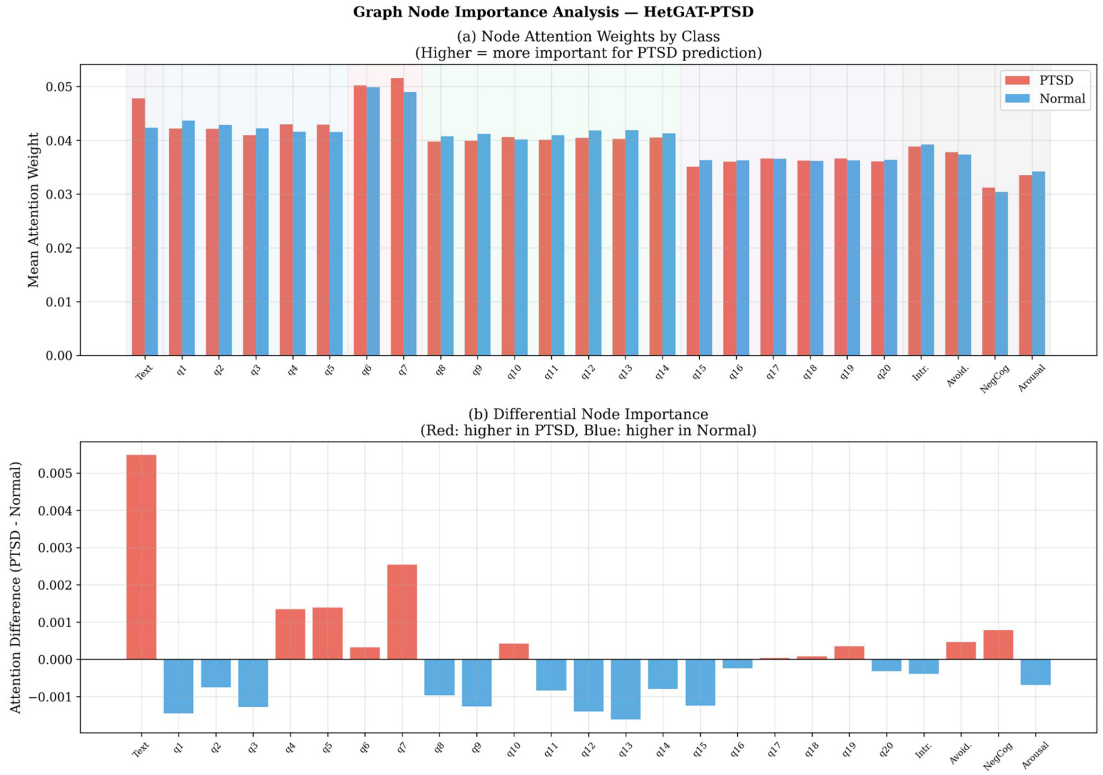

Figure 6: Aggregated node-importance scores: A global summary identifying the most influential PCL-5 items and DSM-5 clusters in the automated PTSD-detection process.

4.7 Case Study: Patient-Level Diagnostic Pathways

To illustrate the dynamic, patient-specific nature of the attention mechanism, we compare two PTSD-positive patients from the test cohort.

Case 1—Hyperarousal-dominant (Patient #42): This patient frequently expressed themes of insomnia and hypervigilance. The model assigned the highest text-to-question attention to q16 (Hypervigilance) and q20 (Sleep disturbance), which at the second hop routed predominantly to the Arousal and Reactivity cluster node. The resulting cluster-level attention weights provide a directly interpretable record of the diagnostic pathway for this patient.

Case 2—Avoidance-dominant (Patient #87): This patient described deliberate avoidance of trauma-associated places and refusal to discuss the traumatic event. Accordingly, the model shifted its attention to q6 and q7, strongly activating the Avoidance cluster node. These cases demonstrate that HetGAT-PTSD produces patient-specific feature weighting within a fixed DSM-5-constrained topology via the Narrative → PCL-5 Item → DSM-5 Cluster pathway. The per-patient audit trail generated by the two-tier attention mechanism allows clinicians to verify that the model attends to clinically coherent symptom–narrative associations, thereby functioning as an interpretable decision-support tool rather than an opaque classifier [25,29].

5 Discussion

5.1 Comparison with Related Work

Table 8 situates the proposed Multimodal HetGAT-PTSD within the broader landscape of automated PTSD detection. To the best of our knowledge, no prior study has addressed PTSD detection using Turkish clinical narratives combined with PCL-5 item-level graph modeling. We compare against the most methodologically similar approaches across three paradigms: neuroimaging-based, NLP/text-based, and multimodal clinical methods. Neuroimaging-based approaches report high accuracy but require fMRI acquisition environments that limit their routine deployment (Table 8). Text- and interview-based NLP methods operate on more accessible data, yet lack structured clinical scale integration or DSM-5-guided interpretability [5,41,42]. The proposed model uniquely combines accessible clinical data within a DSM-5-constrained heterogeneous graph, achieving a 94.27% AUC-ROC with multi-hop attention-based interpretability absent from all compared systems.

Table 8: Comparison of representative PTSD-detection approaches. Acc = Accuracy; N/A = not reported. [†] fMRI-based methods require specialized neuroimaging hardware.

Study	Modality	Acc (%)	AUC-ROC	AUC-PR	Interpretable	Data Source
Huynh et al. [2] (2024)	fMRI + GANs/GCNs[†]	97.76	N/A	N/A	No	rs-fMRI
Alahmadi et al. [3] (2025)	fMRI + Stacked DL[†]	96.60	N/A	N/A	No	rs-fMRI
Schultebraucks et al. [41] (2022)	Audio + Video + Text	N/A	0.90	N/A	No	ED interviews
Sawalha et al. [5] (2022)	Text (sentiment)	80.40	0.80	N/A	No	AVEC-19
Chen et al. [42] (2025)	Text (SBERT + LR)	N/A	0.856	0.758	Partial	DAIC-WOZ
Ours (HetGAT-PTSD)	Text + PCL-5 Graph	91.16	0.9427	0.9584	Yes (XAI)	Turkish EHR

Note: Acc: Accuracy; AUC-ROC: Area Under the Curve-Receiver Operating Characteristic; AUC-PR: Area Under the Curve-Precision-Recall; fMRI: functional Magnetic Resonance Imaging; rs-fMRI: resting-state fMRI; GANs: Generative Adversarial Networks; GCNs: Graph Convolutional Networks; DL: Deep Learning; ED: Emergency Department; SBERT: Sentence-BERT; LR: Logistic Regression; DAIC-WOZ: Distress Analysis Interview Corpus-Wizard of Oz; AVEC-19: Audio/Visual Emotion Challenge 2019; XAI: Explainable Artificial Intelligence; HetGAT-PTSD: Heterogeneous Graph Attention Network for PTSD detection; PCL-5: PTSD Checklist for DSM-5; EHR: Electronic Health Record.

5.2 The Value of Domain Knowledge and Multimodal Fusion

The central contribution of HetGAT-PTSD is not predictive superiority but a structured, DSM-5-guided mechanism for inspecting model decisions: typed attention weights trace diagnostic evidence from narrative content through PCL-5 items to DSM-5 clusters, producing a patient-specific audit trail absent from all compared systems.

Although BERTurk (A2) achieved a higher raw accuracy (97.13%), this gap is not statistically significant (p = 0.174, corrected paired t-test [39]; see Section 4.3). Purely text-based models are susceptible to data memorization and lack mechanistic transparency [43], and recent work confirms that language model outputs must be grounded in structured domain knowledge to ensure diagnostic safety [20]—a principle embodied by routing representations through a DSM-5-constrained graph topology.

It is important to acknowledge that B1—the logistic regression on PCL-5 total score alone—surpasses HetGAT-PTSD on raw accuracy (93.30% vs. 91.16%) and F1 (92.45% vs. 90.34%). However, B1's AUC-ROC (87.77%) is substantially lower than HetGAT-PTSD (94.27%), suggesting that its accuracy advantage reflects the high separability of PCL-5 total scores in this cohort (PTSD mean: 34.47 ± 6.44 vs. Non-PTSD mean: 27.43 ± 2.94) rather than superior generalizable discrimination. More fundamentally, B1 operates on a single scalar summary and produces no symptom-level reasoning, no DSM-5 cluster attribution, and no patient-specific audit trail—distinctions essential for clinical accountability. HetGAT-PTSD's contribution is not raw predictive superiority over B1, but the provision of a structured, interpretable decision pathway that a scalar threshold model cannot offer.

Ablation studies (Table 3) quantitatively support this design. Removing the four DSM-5 cluster nodes (A5) consistently decreased performance, confirming their dual role as a structural regularizer and interpretability scaffold. Graph-based fusion yielded mechanistically interpretable representations that flat concatenation (A3) cannot offer, demonstrating that Graph Attention Networks effectively capture non-linear, non-Euclidean relationships in multimodal clinical data [18,44].

The model exhibits higher fold-to-fold variance (±4.25% accuracy) than the unimodal baselines, attributable to additional learnable components in a 418-patient cohort. This does not undermine reliability: mean AUC-ROC is 0.9427 ± 0.0463 with four of five folds exceeding 0.90. Stability analysis confirms perturbation robustness (0.986) and cross-fold weight similarity (0.984), both exceeding the >0.7 threshold for correlation-based reliability [33]. However, cross-fold Kendall's τ (0.027) indicates that exact rank ordering is less consistent than magnitude-based agreement (see Section 4.4). Furthermore, only 5.1% of HetGAT errors overall (2 of 39; 16.7% of false positives and 0% of false negatives) were shared with BERTurk; the remaining 37 unique errors (8.9% of total cases) represent a genuine accuracy cost of the graph-structured architecture, confirming the deliberate accuracy–interpretability trade-off already noted in Section 4.1: the model sacrifices some raw discriminative accuracy in exchange for a DSM-5-grounded, patient-specific audit trail unavailable from the text-only baseline.

In addition to the reported baselines, we evaluated a Bidirectional Long Short-Term Memory (BiLSTM) and a fine-tuned BERTurk model on clinical narratives. The BiLSTM baseline used a randomly initialized 256-dimensional embedding layer (not pretrained), a 2-layer bidirectional LSTM (hidden size 128), and masked mean pooling, trained with the same 5-fold stratified cross-validation protocol as the other baselines. The fine-tuned BERTurk model unfroze the last two encoder layers of the pretrained BERTurk backbone (all other layers frozen, identical to A2's frozen configuration) and was trained with AdamW (learning rate 5×10^{-4}, weight decay 1×10^{-4}), a linear warmup schedule (10% of training steps), batch size 16, for up to 40 epochs with early stopping (patience 8). Fine-tuning the last two encoder layers introduced training instability (Recall = 0.715 ± 0.329), attributable to the limited sample size (N = 418). The frozen encoder

already captures sufficient linguistic information for this task. These findings confirm that the primary contribution of HetGAT-PTSD is not incremental predictive gain over text-based models, but rather the provision of a DSM-5-grounded interpretive pathway that enables clinicians to trace diagnostic evidence through structured symptom clusters. The DSM-5-constrained topology is a deliberate design choice: the objective is not to discover novel latent pathology patterns, but to provide an interpretable screening tool that mirrors structured clinical reasoning. In clinical decision support, alignment with established diagnostic criteria is a requirement for regulatory compliance and clinical trust, not a limitation.

5.3 Clinical Implications

False-positive psychiatric diagnoses can lead to unnecessary treatments and social stigma, while false negatives deprive patients of urgent care. The model's precision (93.51%) ensures a low false-alarm rate, and its reliance on routinely collected data suggests potential applicability in resource-constrained settings, contingent on external validation and clinician review. Beyond accuracy, HetGAT-PTSD provides a traceable decision pathway auditable against clinical judgment—a property mandated by regulatory frameworks for AI-based clinical decision support [22,24].

5.4 Limitations and Future Directions

This study has several limitations. First, the dataset comprises 418 patients from a single institution; validation on external, multi-center cohorts is necessary to establish broader generalizability. Second, the frozen text encoder could be enhanced through parameter-efficient fine-tuning (e.g., Low-Rank Adaptation) to adapt language representations without data memorization. Third, incorporating temporal edges to track longitudinal patient visits could extend the framework into a prognostic tool for disease trajectory prediction. Fourth, the dataset lacks demographic variables (age and sex), precluding subgroup bias analysis; future studies should examine model fairness across patient subgroups. Fifth, the graph architecture is language-agnostic; only the text encoder (BERTurk) is language-specific. Replacing it with any equivalent pre-trained model (e.g., Bio-ClinicalBERT for English) would require no modification to the graph structure or training protocol. The framework relies on BERTurk, which limits direct applicability to Turkish-language data. However, this dependency is modular by design. BERTurk functions solely as a feature extractor that produces a fixed-dimensional [CLS] embedding, while the graph topology, edge types, and attention mechanism remain language-independent. Substituting BERTurk with encoders such as Bio-ClinicalBERT, CamemBERT, or multilingual XLM-RoBERTa requires modifications only to the text encoding module. Empirical validation of cross-linguistic transfer, including whether the observed attention patterns generalize across linguistic and cultural contexts, constitutes an important direction for future work. Sixth, the DSM-5-constrained topology limits the model's capacity to identify symptom patterns not captured by the current diagnostic taxonomy; future work will evaluate hybrid architectures combining constrained and unconstrained branches. Seventh, a structured usability evaluation with independent psychiatrists is needed to assess the clinical utility of the two-tier audit trail before any deployment consideration.

6 Conclusion

We proposed Multimodal HetGAT-PTSD, a heterogeneous graph attention network that integrates unstructured patient narratives with structured PCL-5 assessments under a DSM-5-constrained graph topology. The framework achieved 91.16% accuracy and 94.27% AUC-ROC on real-world Turkish clinical data, substantially outperforming the questionnaire-only baseline while providing multi-hop attention-based transparency that traces diagnostic evidence from narrative expressions through PCL-5 items to DSM-5 symptom clusters. By combining graph-based multimodal fusion with embedded medical domain

knowledge, the model offers a clinically grounded, interpretable, and accessible research prototype for PTSD decision support using routinely available clinical data. These findings are preliminary and based on a single site; external validation on multi-center cohorts and structured evaluation by independent clinicians are essential next steps.

Acknowledgement: The authors would like to express their gratitude to the psychiatrists and clinical staff at Istanbul Bakirkoy Prof. Dr. Mazhar Osman Mental Health and Neurological Diseases Training and Research Hospital for their invaluable support in clinical evaluation and data curation. This study was part of the PhD thesis titled *Detection of Post-Traumatic Stress Disorder with Natural Language Processing Method*.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contributions to the paper as follows: Conceptualization, Engin Seven and Eylem Yucel; methodology, Engin Seven and Eylem Yucel; software, Engin Seven; validation, Munevver Yildirim; formal analysis, Engin Seven; data curation, Munevver Yildirim; writing—original draft preparation, Engin Seven; writing—review and editing, Engin Seven, Eylem Yucel and Munevver Yildirim; supervision, Eylem Yucel. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The clinical dataset is not publicly available due to patient privacy regulations (KVKK). De-identified feature matrices can be made available by the corresponding author upon reasonable request, subject to ethical approval. The model architecture, weights, and custom code are publicly available at https://github.com/Engin-SEVEN/Multimodal_HetGAT-PTSD, including synthetic data samples and execution instructions for reproducibility.

Ethics Approval: This study involved human subjects. Ethical approval was obtained from the Hamidiye University Scientific Research Ethics Committee (No.: 29.02.2024-25936). The study complied with the Turkish Personal Data Protection Law (KVKK) and the Declaration of Helsinki. All records were anonymized at the source; patient consent was waived by the ethics committee due to the retrospective nature of the study and use of fully de-identified data. The manuscript contains no identifiable personal data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kratzer L, Knefel M, Haselgruber A, Heinz P, Schennach R, Karatzias T. Co-occurrence of severe PTSD, somatic symptoms and dissociation in a large sample of childhood trauma inpatients: a network analysis. Eur Arch Psychiatry Clin Neurosci. 2022;272(5):897–908. doi:10.1007/s00406-021-01342-z.
2. Huynh N, Yan D, Ma Y, Wu S, Long C, Sami MT, et al. The use of generative adversarial network and graph convolution network for neuroimaging-based diagnostic classification. Brain Sci. 2024;14(5):456. doi:10.3390/brainsci14050456.
3. Alahmadi TJ, Khan AR, Ali H, Al-Ghofaily B, Alghanim A, Ayesha N. Post-traumatic stress disorder diagnosis using brain cellular resting-state functional magnetic resonance imaging with stacked deep learning framework. Open Biomed Eng J. 2025;19:e18741207333407. doi:10.2174/0118741207333407241002063548.
4. Wu Y, Mao K, Dennett L, Zhang Y, Chen J. Systematic review of machine learning in PTSD studies for automated diagnosis evaluation. npj Ment Health Res. 2023;2(1):16. doi:10.1038/s44184-023-00035-w.
5. Sawalha J, Yousefnezhad M, Shah Z, Brown MRG, Greenshaw AJ, Greiner R. Detecting presence of PTSD using sentiment analysis from text data. Front Psychiatry. 2021;12:811392. doi:10.3389/fpsyt.2021.811392.
6. Liu S, Zhou J, Zhu X, Zhang Y, Zhou X, Zhang S, et al. An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network. Patterns. 2024;5(12):101081. doi:10.1016/j.patter.2024.101081.

7. Shi G, Zhu Y, Liu W, Yao Q, Li X. Heterogeneous graph-based multimodal brain network learning. IEEE Trans Knowl Data Eng. 2025;37(8):4664–76. doi:10.1109/TKDE.2025.3569648.
8. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv:1710.10903. 2018.
9. Schweter S. BERTurk-BERT models for Turkish. Zenodo. 2020. doi:10.5281/zenodo.3770924.
10. Cowansage K, Nair R, Lara-Ruiz JM, Berman DE, Boyd CC, Milligan TL, et al. Genetic and peripheral biomarkers of comorbid posttraumatic stress disorder and traumatic brain injury: a systematic review. Front Neurol. 2025;16:1500667. doi:10.3389/fneur.2025.1500667.
11. Hinojosa CA, George GC, Ben-Zion Z. Neuroimaging of posttraumatic stress disorder in adults and youth: progress over the last decade on three leading questions of the field. Mol Psychiatry. 2024;29(10):3223–44. doi:10.1038/s41380-024-02558-w.
12. Zhu X, Kim Y, Ravid O, He X, Suarez-Jimenez B, Zilcha-Mano S, et al. Neuroimaging-based classification of PTSD using data-driven computational approaches: a multisite big data study from the *ENIGMA*-PGC PTSD consortium. NeuroImage. 2023;283:120412. doi:10.1016/j.neuroimage.2023.120412.
13. Chen Z, Liu X, Yang Q, Wang YJ, Miao K, Gong Z, et al. Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: a systematic review. JAMA Netw Open. 2023;6(3):e231671. doi:10.1001/jamanetworkopen.2023.1671.
14. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. doi:10.1038/s42256-019-0048-x.
15. Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: The semantic web. Berlin/Heidelberg, Germany: Springer; 2018. p. 593–607. doi:10.1007/978-3-319-93417-4_38.
16. Jemima DD, Selvarani AG, Lovenia JDL. Multi-view united transformer block of graph attention network based autism spectrum disorder recognition. Front Psychiatry. 2025;16:1485286. doi:10.3389/fpsyt.2025.1485286.
17. Zheng K, Yu S, Li B, Jenssen R, Chen B. BrainIB: interpretable brain network-based psychiatric diagnosis with graph information bottleneck. IEEE Trans Neural Netw Learn Syst. 2025;36(7):13066–79. doi:10.1109/TNNLS.2024.3449419.
18. Vaida M, Huang Z. Multimodal graph neural networks in healthcare: a review of fusion strategies across biomedical domains. Front Artif Intell. 2025;8:1716706. doi:10.3389/frai.2025.1716706.
19. Zhang DC, Yang M, Ying R, Lauw HW. Text-attributed graph representation learning: methods, applications, and challenges. In: Proceedings of the Companion Proceedings of the ACM Web Conference 2024; 2024 May 13–17; Singapore. doi:10.1145/3589335.3641255.
20. Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, et al. Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study. JMIR AI. 2025;4(140):e58670. doi:10.2196/58670.
21. Crow TM, Lin E, Harper KL, Crowe ML, Keane TM, Marx BP. Misleading results in posttraumatic stress disorder predictive models using electronic health record data: algorithm validation study. J Med Internet Res. 2025;27(3):e63352. doi:10.2196/63352.
22. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. doi:10.1136/bmj-2024-081554.
23. Mehta V, Komanduri A, Bhadouriya RS, Mehta V, Johnson MD, Shrestha P, et al. Evaluating transparency in AI/ML model characteristics for FDA-reviewed medical devices. npj Digit Med. 2025;8(1):673. doi:10.1038/s41746-025-02052-9.
24. Hassan R, Nguyen N, Finserås SR, Adde L, Strümke I, Støen R. Unlocking the black box: enhancing human-AI collaboration in high-stakes healthcare scenarios through explainable AI. Technol Forecast Soc Change. 2025;219:124265.

25. Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion. 2023;96(10):156–91. doi:10.1016/j.inffus.2023.03.008.
26. Bryant RA, Galatzer-Levy I, Hadzi-Pavlovic D. The heterogeneity of posttraumatic stress disorder in DSM-5. JAMA Psychiatry. 2023;80(2):189–91. doi:10.1001/jamapsychiatry.2022.4092.
27. Misitano A, Tarantino A, Geddo F, Oppo A, Forresi B. The network structure of PTSD symptoms in children and adolescents exposed to potentially traumatic events: a systematic review. Children. 2025;12(11):1516. doi:10.3390/children12111516.
28. Ross AS, Hughes MC, Doshi-Velez F. Right for the right reasons: training differentiable models by constraining their explanations. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017 Aug 19–26; Melbourne, Australia. doi:10.24963/ijcai.2017/371.
29. Nasarian E, Alizadehsani R, Acharya UR, Tsui KL. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework. Inf Fusion. 2024;108(1):102412. doi:10.1016/j.inffus.2024.102412.
30. Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. doi:10.18653/v1/2020.acl-main.386.
31. Wiegreffe S, Pinter Y. Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3–7; Hong Kong, China. doi:10.18653/v1/D19-1002.
32. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–37. doi:10.1016/j.ipm.2009.03.002.
33. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput Surv. 2023;55(13s):1–42. doi:10.1145/3583558.
34. Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada. doi:10.5555/3327757.3327875.
35. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montreal, QC, Canada.
36. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995 Aug 20–25; Montreal, QC, Canada.
37. Szeghalmy S, Fazekas A. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors. 2023;23(4):2333. doi:10.3390/s23042333.
38. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.
39. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn. 2003;52(3):239–81. doi:10.1023/A:1024068626366.
40. Jain S, Wallace BC. Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. doi:10.18653/v1/N19-1357.
41. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychol Med. 2022;52(5):957–67. doi:10.1017/S0033291720002718.
42. Chen F, Ben-Zeev D, Sparks G, Kadakia A, Cohen T. Detecting PTSD in clinical interviews: a comparative analysis of NLP methods and large language models. Pac Symp Biocomput. 2026;31:265–79. doi:10.1142/9789819824755_0019.

43. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med. 2023;6(1):120. doi:10.1038/s41746-023-00873-0.
44. Mienye ID, Viriri S. Graph neural networks in medical imaging: methods, applications and future directions. Information. 2025;16(12):1051. doi:10.3390/info16121051.