Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.068784

ARTICLE

Automatic Detection of Health-Related Rumors: A Dual-Graph Collaborative Reasoning Framework Based on Causal Logic and Knowledge Graph

Ning Wang, Haoran Lyu* and Yuchen Fu

College of Information Science and Engineering, Hunan Institute of Engineering, Xiangtan, 411228, China

*Corresponding Author: Haoran Lyu. Email: lvhaoran37@gmail.com

Received: 06 June 2025; Accepted: 19 September 2025; Published: 10 November 2025

ABSTRACT: With the widespread use of social media, the propagation of health-related rumors has become a significant public health threat. Existing methods for detecting health rumors predominantly rely on external knowledge or propagation structures, with only a few recent approaches attempting causal inference; however, these have not yet effectively integrated causal discovery with domain-specific knowledge graphs for detecting health rumors. In this study, we found that the combined use of causal discovery and domain-specific knowledge graphs can effectively identify implicit pseudo-causal logic embedded within texts, holding significant potential for health rumor detection. To this end, we propose CKDG—a dual-graph fusion framework based on causal logic and medical knowledge graphs. CKDG constructs a weighted causal graph to capture the implicit causal relationships in the text and introduces a medical knowledge graph to verify semantic consistency, thereby enhancing the ability to identify the misuse of professional terminology and pseudoscientific claims. In experiments conducted on a dataset comprising 8430 health rumors, CKDG achieved an accuracy of 91.28% and an F1 score of 90.38%, representing improvements of 5.11% and 3.29% over the best baseline, respectively. Our results indicate that the integrated use of causal discovery and domain-specific knowledge graphs offers significant advantages for health rumor detection systems. This method not only improves detection performance but also enhances the transparency and credibility of model decisions by tracing causal chains and sources of knowledge conflicts. We anticipate that this work will provide key technological support for the development of trustworthy health-information filtering systems, thereby improving the reliability of public health information on social media.

KEYWORDS: Health rumor detection; causal graph; knowledge graph; dual-graph fusion

1 Introduction

The Internet now serves as a primary source of health information, with nearly 90% of individuals turning to online searches when confronted with health concerns [1]. While this ease of access has reduced barriers to obtaining medical consultation [2], shortcomings within the regulatory framework for information have facilitated the rapid dissemination of health-related misinformation [3]. The World Health Organization has cautioned that such misinformation poses a significant threat to global public health security [3,4].

Existing rumor detection methods can be broadly categorized into three approaches: (1) semantic modeling using deep language models; (2) propagation network analysis that leverages the structure of social networks; and (3) methods that combine external knowledge reasoning or causal-driven approaches. The first two categories have made significant progress, yet they remain insufficient in health-related scenarios.

Deep language models such as BERT and RoBERTa are capable of capturing rich semantic patterns; however, they lack the ability to verify the misuse of biomedical terminology or the logical consistency of medical statements [5,6]. Graph neural networks like TextGCN and Compare-GCN are effective at exploiting structural clues inherent in the online propagation of rumors, but they largely overlook the internal logic of the rumor content [7,8]. To address these shortcomings, recent work has begun exploring knowledge-driven and causal-driven methods [9–12]. Knowledge-based approaches enhance text representations by linking entities such as diseases, symptoms, and treatments to biomedical knowledge graphs, thereby introducing domain-specific prior knowledge. While these methods improve the identification of health-related concepts, they inherently model static associations and are unable to reveal the fictitious causal chains characteristic of many health rumors [13]. In contrast, causal-driven methods employ causal discovery or causal inference techniques to uncover underlying causal mechanisms and distinguish genuine causal relationships from spurious correlations [14]. However, in health-related contexts, these methods face two major challenges: (1) rumor texts are typically noisy and unstructured, rendering causal inference unreliable; and (2) without the support of domain knowledge, the inferred causal links may not be consistent with biomedical rationality. Consequently, a purely causal-based approach may misinterpret or overfit the rumor patterns when unconstrained by external knowledge. Although recent studies have attempted to optimize causal reasoning models for rumor detection [11], such efforts remain limited, with most research focusing on general social rumors rather than those specific to the health domain. Moreover, some recent studies have explored hybrid or multi-graph inference strategies that integrate heterogeneous information sources [10,13,15]. While these methods demonstrate the benefits of graph integration, they typically focus on combining social propagation graphs or multimodal graphs rather than jointly leveraging causal structures and biomedical knowledge. Thus, the feasibility and potential of unifying causal discovery with domain-specific knowledge for health rumor detection remain uncertain. In view of the recent progress in knowledge-driven and causal-driven approaches, we attempt to explore the feasibility of health rumor detection by integrating knowledge graphs with causal discovery methods.

Here, we propose a health rumor detection framework based on dual-graph fusion, termed CKDG. This framework establishes a collaborative reasoning architecture that integrates a causal graph with a knowledge graph, jointly modeling the implicit causal logic within text and an external medical knowledge verification mechanism, effectively mitigates the limitations of existing methods in capturing pseudo-causal relationships and in the appropriate usage of specialized terminology. Specifically, CKDG initially extracts medical entities from the text and constructs a weighted causal graph to quantify the strength of causal relationships among these entities. It then integrates the causal graph with a medical knowledge graph through a tailored knowledge-guided credibility estimation approach, wherein each causal edge is cross-validated against pre-validated knowledge within the medical graph. Finally, by fusing EGRET with a hierarchical pooling strategy, CKDG dynamically combines causal strength signals and knowledge verification cues, enabling multi-granularity reasoning that spans both local paths and global semantics.

In summary, our main contributions are as follows:

- We propose a dual-channel fusion framework, Causal Graph-Knowledge Graph, which, for the first time, jointly incorporates implicit causal logic and external domain knowledge into health rumor detection.
- We design a knowledge-guided causal strength estimation method, introducing an entity alignment strategy and a knowledge validation mechanism, which significantly correct causal direction misjudgments and enhance the discovery of implicit associations.
- We introduce an edge aggregated graph attention network and a hierarchical pooling strategy to achieve efficient fusion of complex causal topologies and knowledge validation signals, thereby improving the model's robustness.

- On a dataset comprising 8430 health rumor instances, CKDG, when combined with the Chinese Medical Knowledge Graph (CMeKG), achieves an accuracy of 91.28% (an improvement of 5.11% over the best-performing baseline) and an F1 score of 90.38% (an improvement of 3.29% over the best-performing baseline), validating the effectiveness of dual-graph collaborative reasoning.

2 Results Work

Rumor detection methods can be categorized from multiple perspectives. In this work, we focus on text-based approaches, incorporating causal graphs and knowledge graphs as key features. Therefore, we primarily discuss causality-based methods and approaches that leverage external knowledge.

2.1 Causality-Based Methods

Causal relationships characterize the logical connections between events or entities, and the causal semantics inherent in natural language have garnered significant scholarly attention. Extensive investigations have explored causal relations in natural language, yielding promising applications in text classification, word embeddings, and beyond. For example, reference [16] demonstrated that integrating the robustness and interpretability of models developed for natural language processing tasks. Similarly, reference [17] employed a causal perspective to construct graphs capturing user interactions and rumor propagation, while reference [18] applied causal intervention and reasoning approaches to mitigate biases between textual features and news labels. In the realm of fake news detection, reference [19] utilized counterfactual reasoning to address entity bias stemming from the uneven distribution of entities within texts. Collectively, these studies substantiate the feasibility and necessity of leveraging causal relationships to advance rumor detection methodologies.

2.2 Knowledge-Based Methods

Knowledge-based rumor detection methods can generally be categorized into two primary approaches. One approach focuses on fact verification by extracting and comparing triplet structures derived from knowledge graphs [7]. A unified framework has been developed to integrate fact-based and feature-based models by incorporating external fact-checking sources through model preference learning [20]. Another framework has been introduced to jointly model propagation dynamics and the structural relationships among entities in knowledge graphs [21]. Additionally, entity linking techniques have been employed to detect fake news by leveraging entity descriptions from external knowledge bases [22].

The second line of research emphasizes integrating social media responses with news content by leveraging external knowledge to identify key entities [23]. Additionally, several studies have investigated the incorporation of commonsense knowledge graphs, which encode complex conceptual relationships through structured representations [24]. These graphs have been shown to enhance performance in various downstream tasks, including open-domain conversations, question answering, sentiment analysis, and stance detection [25–27]. Furthermore, commonsense knowledge has proven effective in supporting tasks such as sarcasm detection, meme interpretation, and emotion recognition [28,29].

From extensive studies, it is evident that considerable progress has been achieved in the field of rumor detection. However, current approaches have not sufficiently addressed the relationships among entities or the impact of external knowledge on rumor detection. Indeed, knowledge serves as a paradigmatic standard for human judgment and plays a pivotal role in evaluating health-related rumors.

3 Preliminaries

Causal reasoning mainly concerns (1) the existence of causal relations and (2) their strength. To capture both, CKDG integrates causal discovery, edge strength estimation, and a graph attention model that fuses causal signals with domain verification.

3.1 Causal Discovery via Greedy Fast Causal Inference (GFCI)

Causal discovery corresponds to the first class of causal reasoning problems. From a graph-based perspective, it refers to inferring causal structures from observational data. In CKDG, this translates to identifying directional links among biomedical entities extracted from rumor texts. To achieve this, we adopt the GFCI algorithm [30], a hybrid method that integrates constraint-based conditional independence testing and score-based structure search. Unlike conventional approaches that assume causal sufficiency, GFCI allows for the presence of latent confounders by generating a Partial Ancestral Graph (PAG), which flexibly represents plausible causal structures. Each PAG corresponds to a Markov equivalence class of DAGs compatible with the observed independence constraints, and may contain uncertain or bidirectional edges. This makes it particularly suitable for noisy and incomplete data derived from real-world texts.

3.2 Causal Strength Estimation

Causal strength estimation corresponds to the second class of causal reasoning problems, which aims to quantify the influence of each learned causal relationship—whether it is strong or weak. In our CKDG framework, we adopt the Average Treatment Effect (ATE) as the metric for causal strength. For a directed edge $T \rightarrow Y$, ATE measures the expected change in Y when the treatment variable T changes, while controlling for observed confounding variables, i.e., $\text{ATE} = \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)]$. In practice, since interventions are not feasible, we approximate these expectations using Propensity Score Matching (PSM): each sample is matched to a counterpart with a similar propensity score but opposite treatment status, enabling counterfactual estimation. The resulting value serves as the edge weight, reflecting the reliability and importance of each causal connection.

3.3 Edge Aggregated Graph Attention Network (EGRET)

To effectively leverage both node-level and edge-level information in downstream classification, we utilize an EGRET [31]. EGRET extends traditional Graph Attention Networks (GAT) by incorporating edge features not only in attention score computation, but also in the message aggregation process. Formally, the attention coefficient α_{ji} from node j to i is computed as a function of both node embeddings and edge attributes. Furthermore, the edge feature vectors are linearly transformed and aggregated alongside node messages, enabling a richer representation that captures structural, semantic, and credibility cues. This design is particularly well-suited to our dual-graph setting, where causal strength and domain verification are encoded as edge features.

4 Methodology

Our proposed health rumor detection framework, termed Causal-Knowledge Dual Graph (CKDG), identifies health-related misinformation by jointly modeling causal reasoning and biomedical knowledge validation within the text. The overall architecture is illustrated in Figs. 1 and 2.

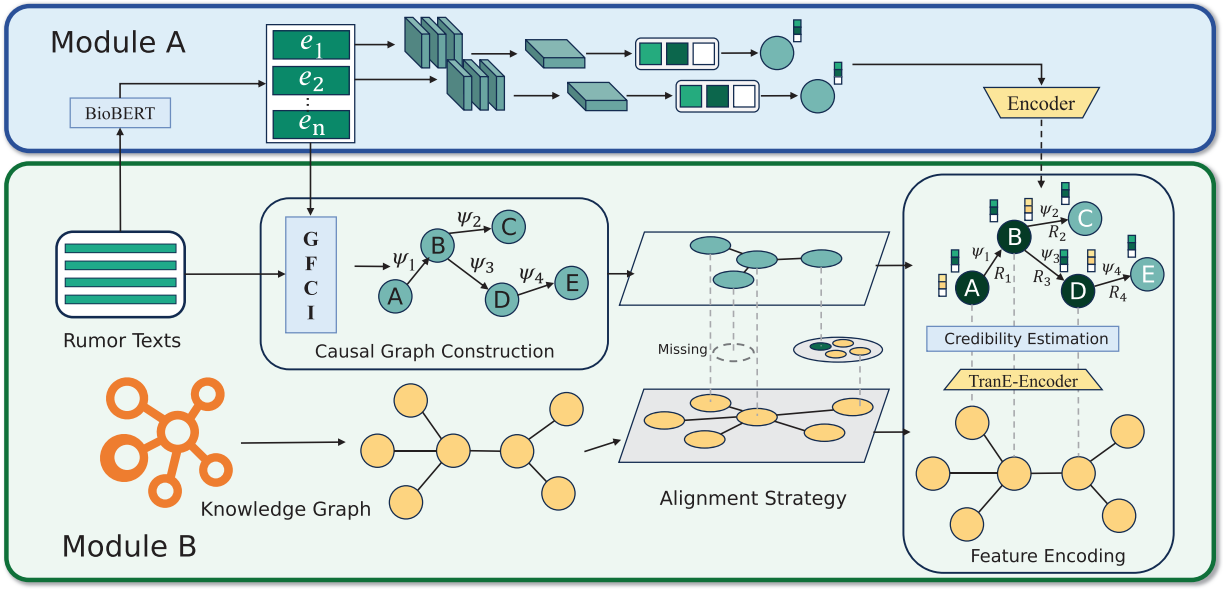

Figure 1: Causal-knowledge dual graph integration framework. The left part shows the causal graph construction process from rumor texts using BioBERT and GFCI. The right part depicts the knowledge verification pipeline that aligns causal edges with entities in the medical knowledge graph, evaluates credibility based on alignment confidence and directionality, and enhances the dual-graph structure

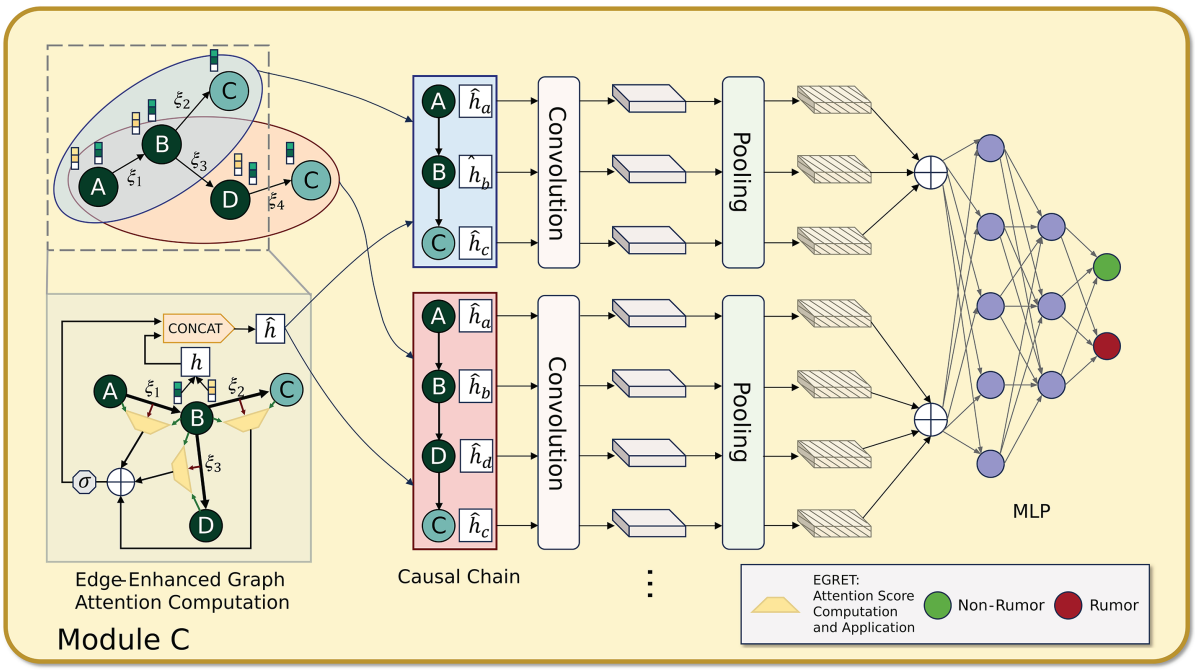

Figure 2: Edge-enhanced dual graph fusion classifier framework. The classifier utilizes node features from BioBERT and knowledge graph embeddings, along with fused edge features comprising causal strength and credibility. Through an EGRET-based graph attention mechanism, these features are jointly processed to compute attention weights and node embeddings. A hierarchical pooling module aggregates graph semantics, followed by an MLP for rumor classification

Fig. 1 illustrates the complete architecture of the CKDG framework, which consists of two collaborative modules. In the left part (Module A), BioBERT is used to extract biomedical entities from rumor texts, followed by the construction of a causal graph through the GFCI algorithm and constraint-based filtering. In the right part (Module B), these causal edges are verified against a domain-specific knowledge graph using alignment strategies and semantic similarity. A credibility score is then computed for each edge based on its alignment quality and consistency with medical knowledge, which will be integrated into the final dual-graph representation.

4.1 Problem Formulation

Given a health-related statement x composed of natural language text, our goal is to determine whether x constitutes a rumor. Formally, we define the rumor detection task as a binary classification problem, where the input is a textual instance $x \in \mathcal{X}$, and the output is a label $y \in \{0, 1\}$ indicating whether x is a health-related rumor ($y = 1$) or not ($y = 0$).

To achieve reliable prediction beyond shallow semantics, we construct a dual-graph representation $\mathcal{G} = \{G_c, G_k\}$, where $G_c = (V_c, E_c, \psi)$ is a causal graph derived from text, with nodes V_c representing extracted biomedical entities, edges E_c representing directed causal relations, and ψ denoting estimated causal strengths. $G_k = (V_k, E_k)$ is an external medical knowledge graph used for validation.

Our task is thus transformed into learning a function $f : \mathcal{X} \to \{0, 1\}$, parameterized by a dual-graph attention model, which jointly considers (1) node-level features from BioBERT and knowledge embeddings, and (2) edge-level features including causal strength ψ and knowledge-based credibility R. The goal is to enhance prediction accuracy by incorporating both implicit causal logic and external domain knowledge.

4.2 Causal Graph Construction and Strength Estimation

As outlined in Section 3, we model both the existence and strength of causal relations. we next elaborate on their instantiation in the CKDG framework.

4.2.1 Causal Graph Construction

To construct the nodes of the causal diagram, semantic parsing and entity extraction are performed on health rumor texts. Given the characteristics of health rumors, this task is classified as a biomedical named entity recognition (BioNER) task. General pre-trained language models (e.g., BERT) exhibit significant limitations in recognizing specialized terminology due to a lack of domain adaptation. Therefore, we employ the BioBERT model [32], which has been re-trained on the PubMed corpus. Through a domain adaptive strategy, BioBERT improves the F1 score in the BioNER task by 0.62% compared to BERT ($p < 0.05$), making it well-suited for processing health rumors. Subsequently, we establish the edges of the graph. Considering that health-related rumors are filled with unverified, fragmented, and often deliberately misleading information, many confounding factors interfere with causal inference. For example, take the research question of whether smoking causes lung cancer [33]. In this scenario, we examine the possible causal relationship between *smoking* (T) and *lung cancer* (Y), where *age* acts as a confounding factor (c). Although comparing *lung cancer* rates between smokers and non-smokers may seem straightforward, However, *age* influ-ences both *smoking* and *lung cancer*. Specifically, older individuals are more likely to smoke and also face a significantly higher risk of developing cancer. Ignoring the effect of *age* may lead to misattributing its influence on *lung cancer* to *smoking*, thereby exaggerating the estimated causal effect. Inspired by Liu et al. [30], we adopt the GFCI algorithm [34] to uncover causal relationships among various factors. GFCI integrates scoring-based and constraint-based algorithms, combining the robust performance of scoring-based methods with the

advantage of bypassing the constraint-based approach's dependence on the assumption of causal sufficiency (i.e., the absence of unobserved confounders). Specifically, GFCI does not rely on the assumption of no latent confounders, making it well-suited for noisy, unstructured textual data such as health rumors. Detailed comparative experiments can be found in Section 5.2.4.

However, automatically generated causal graphs may include noisy edges that are inconsistent with established medical logic. To enhance the reliability of the causal graph, we introduce two constraint-based rules to filter out such edges:

a) **Medical Logic Constraint:** Certain entities and relations in the medical domain follow well-defined logical structures [35]. For example, we prohibit edges from symptom-type entities pointing to disease, drug, or treatment entities, as such directions contradict the inherent etiology-symptom logic in medical knowledge bases.
b) **Temporal Constraint:** Given that causes generally precede effects in time, and symptom descriptions are typically documented in chronological order, we use the temporal order of textual mentions as a constraint. If factor A appears after factor B in most cases, an edge from A to B is disallowed.

It is important to note that the temporal constraint does not imply the existence of an edge from B to A, as temporal precedence is not a sufficient condition for causality.

Table 1 summarizes four types of uncertain relations found in the PAG output by GFCI, which pose challenges for causal strength estimation and downstream tasks. Drawing on the probabilistic causal discovery framework of Liu et al. [30], we generate Q candidate causal graphs (Fig. 3). For each sampled graph, we retain deterministic edges $\rightarrow$, remove bidirectional edges $\leftrightarrow$, and handle uncertain edges probabilistically: $\circ\!\!\rightarrow$ is retained or removed with 0.5 probability, while $\circ\!\!-\!\!\circ$ is retained, removed, or reversed with equal 1/3 probability. The quality of each sampled graph G_q is evaluated based on its goodness-of-fit to the data $\mathbf{X}$, using the Bayesian Information Criterion (BIC), denoted as $\mathrm{BIC}(G_q, \mathbf{X})$.

Table 1: Summary of PAG edge types

Edge	Meaning
A $\rightarrow$ B	A causes B.
A $\leftrightarrow$ B	There is an unobserved confounder of A and B.
A $\circ\!\!\rightarrow$ B	Either A causes B, or unobserved confounder.
A $\circ\!\!-\!\!\circ$ B	Either A causes B, or B causes A, or unobserved confounder.

This multi-graph sampling strategy allows us to handle uncertainty in PAG edge types and obtain a robust estimation of possible causal structures before strength computation.

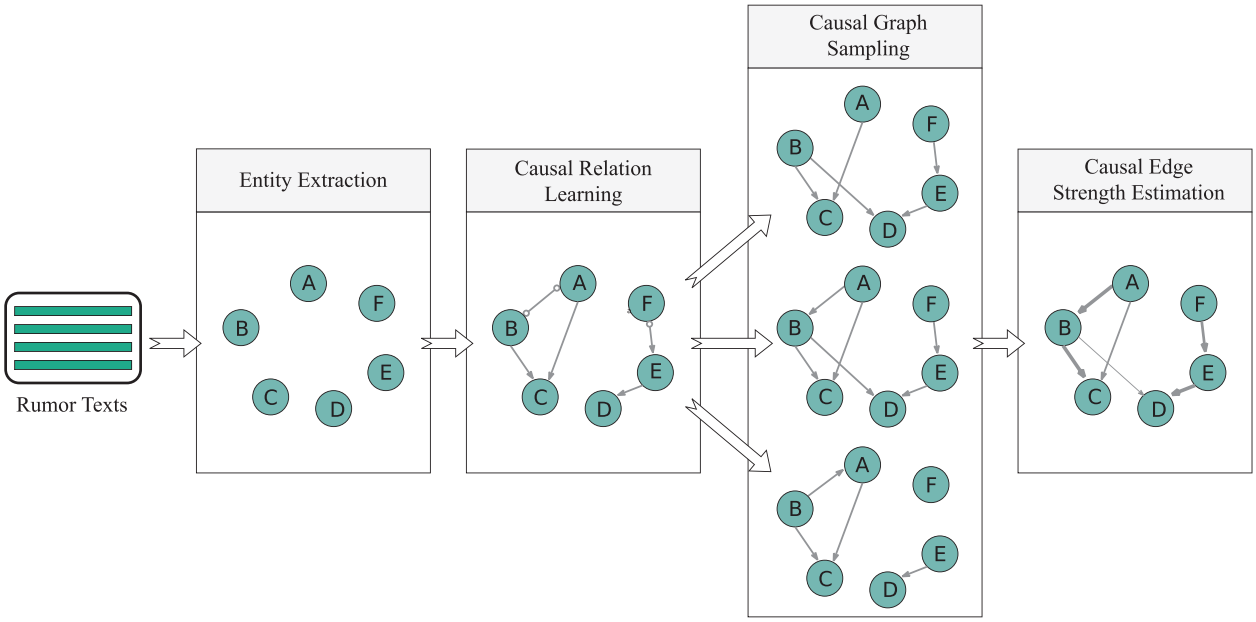

Figure 3: Overall pipeline of causal reasoning in CKDG. In the entity extraction phase, each circle represents a biomedical entity identified from the input rumor text. In the causal relation learning phase, directed and uncertain edges between entities are derived using the GFCI algorithm to construct a PAG, which includes four types of edges: $\rightarrow$, $\leftrightarrow$, $\circ\!\!\rightarrow$, and $\circ\!\!-\!\!\circ$. In the causal graph sampling stage, the PAG is transformed into multiple directed acyclic graphs (DAGs) by resolving uncertain edges probabilistically. Finally, in the causal edge strength estimation phase, each edge's influence is estimated using propensity score matching to quantify its causal effect. The shaded nodes and bold edges in the final graph indicate high-confidence causal factors that guide downstream rumor classification

4.2.2 Causal Graph Strength Estimation

Once the causal structure is established, the next step is to assign edge weights to reflect how influential each causal factor is. Given the inherently noisy nature of the resulting causal graph, we refine the sampled causal graph by quantifying the strength of learned causal associations. Edges demonstrating substantial causal effects are assigned higher weights, while those corresponding to non-causal relationships or exhibiting negligible effects receive weights approaching zero. Intuitively, this strength characterizes the magnitude of the expected change in outcome Y in response to variations in the treatment variable T. For example, in the causal diagram from "*smoking* $(T) \rightarrow$ *lung cancer* (Y)", the causal strength captures the average effect of smoking on the incidence of lung cancer after controlling for confounding factors such as age. Formally, we operationalize the edge strength for T to outcome Y in graph G as the average treatment effect $\psi^G_{T,Y}$, while is estimated through propensity score matching:

$$\hat{\psi}^G_{T,Y} = [\sum_{i:t_i=1} (y_i - y_j) + \sum_{i:t_i=0} (y_j - y_i)]/N \tag{1}$$

Here, $j = \underset{k:t_k \neq t_i}{\operatorname{argmin}} |L(z_i) - L(z_k)|$ denotes the instance in the opposite treatment group that is most similar to instance i. The variables t_i, y_i, z_i represent the treatment assignment, outcome, and confounder variables of instance i, respectively. $L(z)$ is a propensity score function based on confounding factors.

4.3 Knowledge Graph-Driven Causal Validation

Automatically inferred causal graphs may contain spurious or implausible edges. To enhance their reliability, we validate them using domain-specific knowledge.

This section aims to validate whether the causal relations identified in the causal graph G_c are consistent with domain knowledge, by comparing them with a domain-specific knowledge graph G_k. The knowledge graph G_k consists of a set of entities V_k and a set of relations R_k. By aligning the entities and relations in the causal graph with those in the knowledge graph, we categorize edge alignments into four verification scenarios—positive, reverse, unrelated, and missing—and design scoring strategies accordingly.

4.3.1 Entity Alignment

The objective of entity alignment is to match each entity e_i in the causal graph with a corresponding entity e_i' in the knowledge graph. We define a semantic similarity matrix $\Theta \in \mathbb{R}^{|V_c|\times|V_k|}$, where $\Theta_{e_i,e_i'}$ denotes the semantic similarity between entity e_i in the causal graph and entity e_i' in the knowledge graph. The alignment process consists of three steps, as illustrated in Fig. 4:

a) **Exact Matching:** Entities with identical names are directly matched using a hash table. For these, $\Theta_{e_i,e_i'} = 1$.
b) **Fuzzy Matching:** For unmatched entities, semantic similarity scores $\theta_{e_i,e_i'}$ are computed using vector representations. Entity pairs with similarity scores exceeding a predefined threshold θ are retained, and their corresponding $\Theta_{e_i,e_i'}$ values are set to $\theta_{e_i,e_i'}$.
c) **Unaligned Tagging:** Entities that remain unmatched after the above steps are labeled as "Missing".

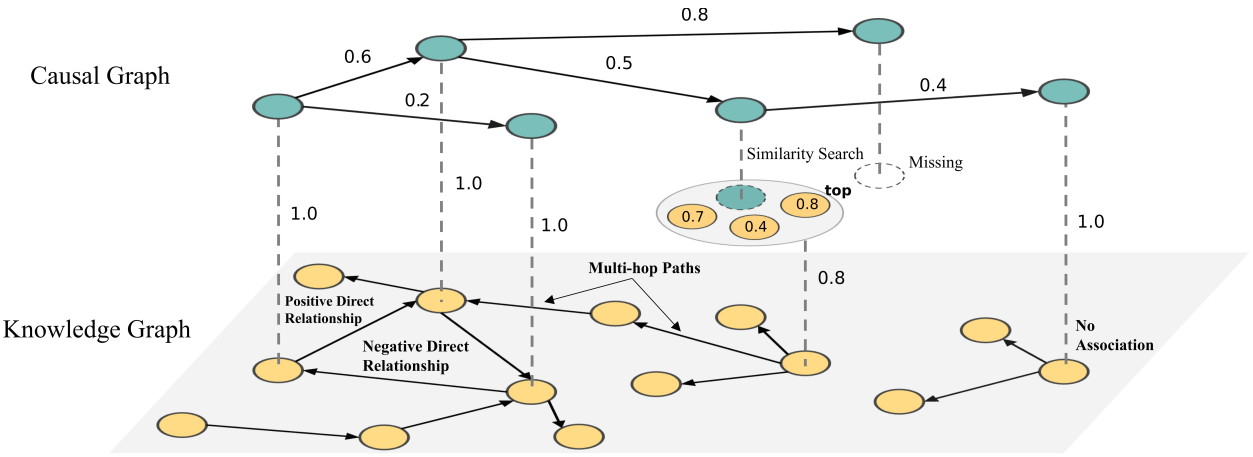

Figure 4: Illustration of the edge-level alignment between the causal graph and knowledge graph. Three alignment strategies—direct matching, multi-hop reasoning, and semantic similarity—are applied to verify the validity of causal edges. Each edge is then assigned to one of four verification scenarios based on its correspondence in the knowledge graph: positive, negative, ambiguous, or missing

4.3.2 Knowledge-Guided Causal Credibility Estimation

After completing entity alignment, we estimate the credibility R of the causal strength for each edge $e_i \to e_j$ in the causal graph by verifying it against the knowledge graph. To achieve this, we have designed a knowledge-guided gating mechanism. The core idea is that the credibility of textual causal claims depends on their consistency with established domain knowledge, and is modulated by the alignment confidence between the textual entities and their corresponding entities in the normative knowledge graph. This

mechanism functions as a "gate" or "filter": it amplifies causal signals that are consistent with knowledge, suppresses those that contradict it, and penalizes assertions that cannot be validated due to missing knowledge or alignment failures. Formally, we design a credibility function $\mathrm{R}(\cdot)$. Consider the minimal model illustrated in Fig. 5, which includes four entities: e_i and e_j from the causal graph, and e_i' and e_j', which are the aligned counterparts in the knowledge graph. We define $\phi(e_i' \to e_j')$ as the external domain knowledge constraint, $\psi(e_i' \to e_j')$ as the original causal strength estimate $\hat{\psi}^{G}_{e_i,e_j}$, and $\Theta_{e,e'}$ as the semantic similarity score resulting from entity alignment.

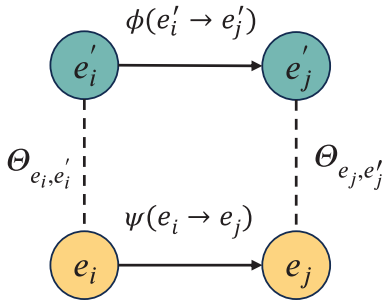

Figure 5: Minimal model in the causal credibility estimation task

Following the idea of gated mechanisms, we construct the credibility estimation function $\mathrm{R}(\cdot)$ guided by two design principles: knowledge dominance and directionality. Under the principle of knowledge dominance, the external domain constraint ϕ acts as a prior and should play a leading role in modulating the credibility. In particular, when $\phi = 0$, the credibility should be completely suppressed. Simultaneously, the function must capture the correlation—positive or negative—between the direction of the knowledge path and the causal direction.

Starting with a multiplicative gating structure, we formulate the function as $\mathrm{R} = \psi \cdot \mathrm{g}(\phi, \Theta_{e,e'})$, where the gating function $\mathrm{g}(\cdot)$ must reflect the dominant role of ϕ. To address potential semantic drift during entity alignment, we use the arithmetic mean of the alignment confidence scores for both endpoints, $\Theta_{e_i,e_i'}$ and $\Theta_{e_j,e_j'}$, to obtain an overall alignment confidence. This helps mitigate the impact of misalignment at a single endpoint. Regarding directionality, we allow ϕ to take negative values when the knowledge path contradicts the causal direction, and apply normalization at the end. The gating function is then expressed analytically as: $\mathrm{g}(\phi, \Theta) = \mathrm{avg}(\Theta) \cdot \phi$.

Accordingly, the complete credibility estimation function becomes:

$$\mathrm{R}(\cdot) = \psi(e_i \to e_j) \cdot \frac{\Theta_{e_i,e_i'} + \Theta_{e_j,e_j'}}{2} \cdot \phi(e_i' \to e_j') \tag{2}$$

To better illustrate how the credibility function R modulates the original causal strength ψ using structured knowledge, we provide an intuitive example in Fig. 6. In this figure, each case corresponds to a textual causal claim $e_i' \to e'j$ with estimated strength $\psi = 0.85$, and shows how knowledge-based consistency ϕ and semantic alignment Θ influence the final credibility R.

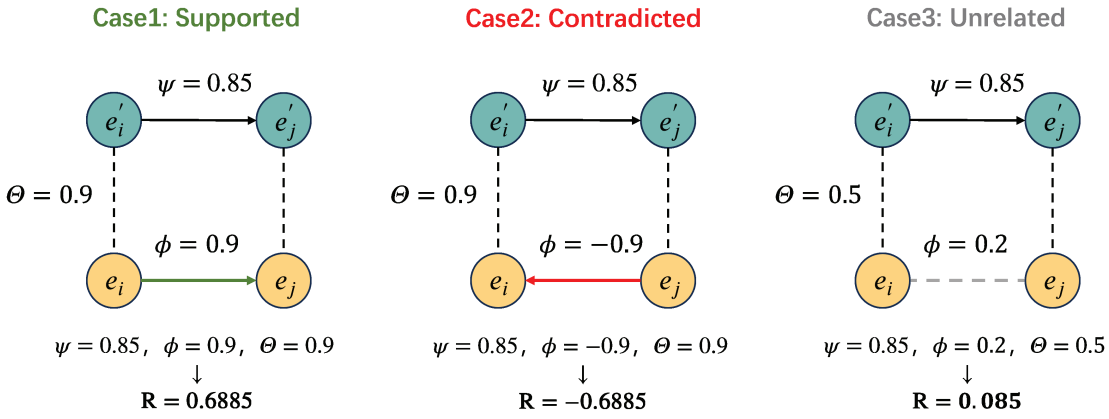

Figure 6: Illustration of three typical cases in knowledge-guided credibility estimation. Case 1 (Support): when the causal claims align with biomedical knowledge (e.g., "Vitamin C can boost immunity"), a positive trustworthiness score is obtained. Case 2 (Contradiction): when the claims contradict established knowledge (e.g., "High doses of antibiotics can cure viral infections"), a negative trustworthiness score results. Case 3 (Irrelevant): when the claims lack clear support in the knowledge graph (e.g., "*Ginkgo biloba* cures all diseases"), the trustworthiness score approaches zero

In Case 1, the knowledge graph supports the same causal direction (i.e., $e_i \to e_j$ exists), yielding a positive ϕ and a high alignment score Θ, thereby preserving most of the original causal strength in R.

In Case 2, a contradictory path exists in the knowledge graph (i.e., $e_j \to e_i$), causing $\phi < 0$ and effectively reversing the impact of ψ, which results in a negative credibility.

In Case 3, no valid path is found between e_i and e_j, and alignment is weak, so ϕ and Θ are both low, leading to a suppressed R even though ψ is high.

These examples demonstrate how our credibility estimation acts as a selective gate to suppress unreliable or unsupported causal edges, and to retain the ones consistent with external knowledge.

After designing the credibility function R, we further examine the determination of $\phi(e_i' \to e_j')$ within the knowledge graph.

Direct Relations: For a causal edge $e_i \xrightarrow{r} e_j$, if there exists a triple $(e_i', r, e_j') \in R_k$ in the knowledge graph where r represents a direct causal chain, we define a directionality consistency score:

$$\phi_{dir}(r) = \begin{cases} \phi(r) \cdot \lambda_{pos}, & \text{if } r = e_i \to e_j \\ -\phi(r) \cdot \lambda_{neg}, & \text{if } r = e_j \to e_i \end{cases} \tag{3}$$

When λ_{pos} and λ_{neg} are penalty factors reflecting the credibility difference between positive and reverse validations.

Multi-Hop Paths: When a direct connection between two entities does not exist in the knowledge graph but a multi-hop path is available (e.g., $P = e_i' \xrightarrow{r_1} e_1' \xrightarrow{r_2} \cdots \xrightarrow{r_n} e_j'$), it is essential to evaluate the semantic coherence and directional consistency of the path. To efficiently search for multi-hop paths, we design a bidirectional breadth-first search (BFS) algorithm:

a) Path Search: The search starts simultaneously from both e_i' and e_j', expanding along relational edges in the knowledge graph until the two search frontiers intersect or the maximum number of hops L_{max} is reached.
b) Path Filtering: Only paths with consistent directionality (e.g., $e_i' \to \cdots \to e_j'$ or its reverse) are retained.

Intuitively, multi-hop paths can reveal underlying indirect causal mechanisms, such as "virus infection → immune response → fever," and thus it is necessary to quantify their contributions while ensuring semantic coherence. In this context, the strength of a path $\phi(e'_i \to e'_j)$ is modeled using a semantic decay function as follows:

$$\phi_{path} = \prod_{k=1}^{n} \sigma(r_k) \cdot \gamma^{|P|} \cdot \left(1 - \frac{|P|}{L_{max}}\right) \tag{4}$$

Among these, $\sigma(r_k) \in [0,1]$ denotes the weight of relationship type r_k, γ is the decay factor related to path length, $|P|$ represents the number of hops in the path, and L_{max} indicates the maximum allowable hops. This design reflects the intuitive rule of multi-hop path strength: on one hand, the overall strength is jointly determined by the reliability of all relationships within the path—if any edge is unreliable, it diminishes the entire path's strength; on the other hand, longer paths incur greater semantic uncertainty, and thus are subjected to an exponential decay factor of $\gamma^{|P|}$, further penalized by $1 - |P|/L_{max}$ to prevent overestimation of excessively long paths. If the path direction is consistent with the causal direction, ϕ_{path} takes on a positive value; otherwise, it is negative. This design also lays the theoretical foundation for subsequent interpretability.

Unrelatedness Test: When no explicit relation exists between the knowledge graph entities corresponding to a candidate causal edge, its credibility is dynamically adjusted based on the global coverage of the knowledge graph, as defined by the following formula:

$$\phi_{null} = \alpha \cdot \left(1 - \frac{\log\left(1 + \mathrm{Deg}(e'_i) \cdot \mathrm{Deg}(e'_j)\right)}{\log(N_k)}\right) \tag{5}$$

where $\mathrm{Deg}(\cdot)$ denotes the degree of an entity, N_k is the total number of entities in the knowledge graph, and α is a base penalty coefficient. This formulation reflects the idea that the absence of observed connections in densely linked subgraphs (e.g., between two common disease entities) is more significant for falsification than in sparse subgraphs (e.g., between a rare drug and a symptom). Specifically, a low ϕ_{null} value indicates that entities e'_i and e'_j reside in densely connected regions of the knowledge graph. In such cases, the absence of even multi-hop paths between them suggests a potentially weak or non-existent association. Conversely, a high ϕ_{null} value suggests that the entities lie in sparsely connected regions of the knowledge graph, where missing links may be due to incomplete data rather than a truly weak association between the entities.

Missing Entity Detection: The annotation of missing entities is performed by an upstream task. Here, we implement a penalty function for specific missing entities:

$$\phi_{missing} = \exp(-\beta \cdot \psi) \cdot \mathbb{I}_{missing} \tag{6}$$

where β denotes the penalty coefficient for missing entities, and $\mathbb{I}_{missing}$ is an indicator function that evaluates to 1 when there exists an unaligned entity at either endpoint of a causal edge.

Integrated Credibility Computation: Considering that both a direct causal relation and multiple multi-hop paths may exist simultaneously between two entities, we integrate the final causal edge strength $\phi(e'_i \to e'_j)$ by combining the following scenarios and adjusting for evidence inconsistency:

$$\phi(e_i \to e_j) = \frac{\sum_{k \in K} \phi_k}{1 + \mathrm{Var}(\{\phi_k\})} \cdot [1 + \log(1 + \psi)] \tag{7}$$

Here, $K = \{\phi_{\mathrm{dir}}, \phi_{\mathrm{path}}\}$ denotes the set of credibility scores derived from different validation scenarios. The logarithmic enhancement term $\log(1 + \psi)$ amplifies the influence of edges with high causal strength.

In contrast, the variance term $\mathrm{Var}(\cdot)$ serves as a conflict-aware penalty: when direct relations and multi-hop paths yield contradictory credibility scores (e.g., $\phi_{\mathrm{dir}} > 0$ but $\phi_{\mathrm{path}} < 0$), the resulting variance increases, thereby downweighting the final credibility $\phi(e_i \to e_j)$. This mechanism effectively suppresses unreliable edges arising from conflicting validation signals and ensures robust fusion of heterogeneous evidence.

Subsequently, the resulting ϕ is fed into the normalization function $\mathrm{R}(\cdot)$ to obtain the final credibility score $\hat{\mathrm{R}} = (\mathrm{R}(\cdot) + 1)/2$. This credibility score is then incorporated into the original causal strength estimation graph to construct an enhanced structural representation $G_c = (V_c, E_c, \hat{\psi}, \hat{\mathrm{R}})$, Where $\hat{\mathrm{R}} = \{\mathrm{R}(e_i \to e_j) | e_i \to e_j \in E_c\}$.

This refined graph integrates both data-driven causal discovery and knowledge-informed validation, forming a more trustworthy basis for downstream classification.

4.4 Edge-Enhanced Dual-Graph Fusion Classifier

In our graph $G_c = (V_c, E_c, \hat{\psi}, \hat{\mathrm{R}})$, both node features and edge features are present. However, the standard GAT framework does not explicitly incorporate edge features into its computational process. To overcome this limitation and fully exploit the structural information encoded by edge features, we adopt a variant of GAT known as EGRET [31]. Unlike conventional GAT models, EGRET is capable of effectively utilizing the structural information encoded in edge features during both the aggregation phase and the attention score computation phase, thereby systematically enhancing the model's ability to exploit edge feature information.

Fig. 2 details the EGRET-based dual-graph fusion classifier used in CKDG. It encodes both node features (from BioBERT and aligned TransE embeddings) and edge features (causal strength and credibility). The edge-enhanced attention mechanism enables the model to incorporate structural and semantic cues from both the causal graph and knowledge validation during message passing. The attention-weighted features are aggregated into a graph-level representation using a hierarchical pooling mechanism, which is then passed to a MLP classifier for final prediction.

4.4.1 Feature Encoding

Node Feature Encoding: As shown in Module A in Fig. 1, We utilize the final hidden states of BioBERT as the initial feature representation $h_i^{(0)}$ for each node $v_i \in V_c$ in the causal graph. This design leverages BioBERT's strong semantic understanding of biomedical texts [36]. To further enhance the semantic representation of nodes, we incorporate aligned knowledge graph entities $e_i' \in G_k$ by embedding them using TransE and integrating the embeddings with the BioBERT representations as follows:

$$h_i^{(0)} = \mathrm{BioBERT}(v_i) + W_i \cdot \mathrm{TransE}(e_i') \tag{8}$$

Here, $\mathrm{TransE}(e_i') \in \mathbb{R}^d$ denotes the TransE embedding of the entity e_i', and W_i is a learnable parameter matrix used to project the knowledge graph embedding into the semantic space of BioBERT.

Edge Feature Encoding: To reduce computational complexity in subsequent processes, we encode the causal strength ψ_{ij}, estimated from the confidence-weighted causal graph, together with the knowledge verification reliability score R_{ij} into a unified edge feature ξ_{ij}. The dimensions of ξ_{ij} correspond to those of ψ_{ij} and R_{ij}, respectively. We design a fusion and normalization strategy as follows:

$$\xi_{ij} = \sigma\left(\alpha \cdot \psi_{ij} + (1 - \alpha) \cdot R_{ij}\right) \tag{9}$$

where α is a learnable parameter, and σ denotes the sigmoid function, which dynamically balances the two sources of information.

4.4.2 Edge-Enhanced Graph Attention Computation

Following the approach of EGRET, the attention score for each directed edge e_{ij} is computed not only based on the node features h_i and h_j, but also incorporates the edge feature ξ_{ij} from node i to node j. Eqs. (10) and (11) illustrate how the attention score is calculated by jointly considering node and edge information.

$$d_{ij} = \text{LeakyReLU}\left(W_a\left[W_v h_i^{(0)} \| W_v h_j^{(0)} \| W_e \xi_{ij}\right]\right) \tag{10}$$

By combining node embeddings with edge features, the attention mechanism is guided to focus on semantically consistent and structurally validated relations. Here, d_{ij} denotes the unnormalized attention score, and $\|$ represents the concatenation operation. The matrices W_a, W_v and W_e are learnable parameters that linearly transform the node and edge features, respectively. Subsequently, the attention scores $\{d_{ij}|j \in N_i\}$ are normalized using a softmax function over the neighbors N_i of the central node i, yielding the attention distribution $\{\alpha_{ij}|j \in N_i\}$:

$$\alpha_{ij} = \text{softmax}_i(d_{ij}) = \frac{\exp(d_{ij})}{\sum_{k\in N_i}\exp(d_{ki})} \tag{11}$$

To fully exploit the informative capacity of the edge features ξ_{ij}, they are incorporated into the aggregation process alongside the neighbor node features $\{h_j|j \in N_i\}$. Specifically, the updated representation for the central node i is computed as:

$$\hat{h}_i = \sigma\left(\sum_{j\in N_i}\alpha_{ij}W_v h_j + \sum_{j\in N_i}\alpha_{ij}W_e\xi_{ij}\right)\|h_i \tag{12}$$

Before aggregation, the edge features ξ_{ij} are linearly transformed using ξ_{ij}. The function $\sigma(\cdot)$ denotes a non-linear activation function. The result is then concatenated with the original node representation h_i to produce the final node representation $\hat{h}_i$ which serves as the output of the edge-enhanced graph attention layer.

4.4.3 Classifier Design

To map the structural information of the graph into the classification space, we employ a MLP as the classifier. Given the hierarchical nature of graph-level semantics in the task of health rumor detection, node-level embeddings must first be aggregated into a graph-level representation. Inspired by the work of Lee et al. [37], we incorporate an attention pooling mechanism to dynamically capture the varying contributions of individual nodes. Specifically, for all node embeddings $\{\hat{h}_i\}_{i=1}^{V_c}$ in the causal graph G_c, attention weights are computed using a learnable query vector $q \in \mathbb{R}^d$ as follows:

$$\eta_i = \text{softmax}(\tanh(W_p\hat{h}_i + b_p)\cdot q) \tag{13}$$

where $W_p \in \mathbb{R}^{d\times d}$ and $b_p \in \mathbb{R}^d$ are trainable projection parameters. The hyperbolic tangent function enhances the non-linear representational capacity. The resulting graph-level representation h_G is obtained via a weighted sum of node embeddings:

$$h_G = \sum_{i=1}^{|V_c|}\eta_i\cdot\hat{h}_i \tag{14}$$

This mechanism assigns greater importance to nodes with high causal strength, aligning with the task requirement of focusing on core causal chains in rumor detection. The graph-level embedding h_G

is passed through a MLP with Batch Normalization and GeLU activation. The first layer reduces the dimensionality from d to $d/4$, and the second layer projects the feature into a 2-class probability space for rumor classification. The final output is computed as:

$$y_{\text{pred}} = \text{Softmax}(W_2 \cdot \text{GeLU}(\text{BN}(W_1 h_G + b_1) + b_2)) \tag{15}$$

where $W_1 \in \mathbb{R}^{d/4 \times d}$ and $W_2 \in \mathbb{R}^{2 \times d/4}$ are learnable parameters. BN denotes the batch normalization operation. This architectural design enhances the model's capability to abstract hierarchical features and discern complex causal logic, while simultaneously mitigating the risk of overfitting.

4.4.4 Loss Function

The training objective of the model is jointly optimized via a multi-task loss, which consists of the primary classification loss, a knowledge consistency constraint, and a sparsity regularization term. The primary loss adopts the Label Smoothing Cross-Entropy (LS-CE) loss to mitigate the overconfidence problem caused by class imbalance by introducing a smoothing factor ε:

$$\mathscr{L}_{cls} = -\sum_{c \in \{0,1\}} \left[(1-\varepsilon) y_c \log p_c + \varepsilon \cdot \frac{1}{2} \log p_c \right] \tag{16}$$

where y_c denotes the ground-truth label, and p_c represents the predicted probability. To further enhance the semantic alignment between causal paths and the knowledge graph, we introduce a knowledge consistency loss term $\mathscr{L}_{kg}$, which enforces the consistency between the attention weights a_{ij} and the causal credibility scores $\hat{\text{R}}_{ij}$. This consistency is measured using the Kullback-Leibler (KL) divergence:

$$\mathscr{L}_{kg} = \frac{1}{|E_c|} \sum_{e_{ij} \in E_c} \hat{\text{R}}_{ij} \cdot \log \left(\frac{\hat{\text{R}}_{ij} + \varepsilon}{\alpha_{ij} + \varepsilon} \right) \tag{17}$$

where $\varepsilon = 1e{-}8$ is a small constant added for numerical stability. This loss encourages the model to rely more heavily on high-confidence edges verified by external knowledge during the decision-making process. In addition, L2 regularization is applied to the parameters of the MLP to control model complexity: $\mathscr{L}_{reg} = \lambda(\|W_1\|_2^2 + \|W_2\|_2^2)$. Finally, the overall loss function is a weighted sum of the three components:

$$\mathscr{L}_{total} = \lambda_1 \mathscr{L}_{cls} + \lambda_2 \mathscr{L}_{kg} + \lambda_3 \mathscr{L}_{reg} \tag{18}$$

The weight coefficients $\lambda_1, \lambda_2, \lambda_3$ are determined via Pareto frontier analysis to balance classification performance and the robustness of knowledge-based reasoning.

4.5 Limitations and Discussion

Although the CKDG framework demonstrated outstanding performance in experiments, it still presents several inherent limitations that delineate the scope of the current study while also pointing to directions for future improvements.

Language and knowledge graph dependencies: The current CKDG is designed exclusively for Chinese health rumor detection, relying on language-specific resources such as the BioBERT model fine-tuned on Chinese biomedical texts and the CMeKG knowledge graph. This design choice inevitably limits its direct applicability to other languages. The entity extraction and semantic alignment modules are highly sensitive to linguistic features, and the knowledge validation process is also constrained by the coverage and structural patterns of CMeKG.

Generalization to multilingual and cross-domain scenarios: To enable CKDG to adapt to a wider range of environments, the following extensions can be considered. For multilingual adaptation, pre-trained cross-lingual models (such as mBERT or XLM-R) can be incorporated to support cross-lingual entity extraction and semantic similarity computation. Additionally, integrating multilingual or language-agnostic knowledge bases (such as UMLS or Wikidata) can provide the necessary domain information for knowledge verification in different languages. The causal discovery module (GFCI) relies solely on statistical correlation and is language-independent, so it can essentially remain unchanged. The primary challenge lies in achieving high-quality entity alignment and relation mapping across heterogeneous knowledge graphs.

Other considerations: Additionally, the computational complexity of causal discovery and graph inference is relatively high, which may affect real-time applications. Relevant experiments can be found in Section 5.2.8; moreover, the model relies on the coverage and accuracy of the underlying knowledge graph, and missing or erroneous knowledge could compromise the reliability of validation.

5 Experiments and Evaluation

5.1 Experimental Setup

5.1.1 Datasets

We conducted experiments using three datasets: the LTCR dataset[1] [38], the COVID-19 Health Rumors (CHR) dataset[2] [39], and a self-constructed multi-source health rumor dataset. The LTCR dataset comprises 2307 samples, including 1731 rumors, 516 non-rumors, and 60 unknown cases. The CHR dataset contains 408 samples, all of which are health-related rumors. To enhance the diversity and scale of the experimental data, we collected samples from multiple authoritative fact-checking platforms and constructed a novel health rumor dataset. The sources and statistics of each platform are detailed in Table 2. The veracity labels (rumor/non-rumor) were directly assigned by domain experts and editorial teams from these platforms, ensuring a high level of annotation accuracy and reliability. Additionally, to guarantee quality, we randomly selected 500 samples, which were independently reviewed by two authors alongside the original platform decisions. In the end, the self-built dataset comprises a total of 5775 samples, including 3954 rumors and 1821 non-rumors.

Table 2: Sources of the self-constructed dataset

Source	Counts
Weibo fact-checking	2189
Toutiao fact-checking	1816
China internet joint rumor debunking platform	776
Tencent fact-check	660
Dingxiang doctor	334

After merging the three datasets, we obtained a total of 8430 samples, consisting of 6093 rumors and 2337 non-rumors. We performed data preprocessing procedures including duplicate removal, word segmentation, and entity annotation based on BioBERT. The final dataset was split into training, validation, and test sets in a 7:2:1 ratio.

[1] https://github.com/Enderfga/DoubleCheck (accessed on 18 September 2025)
[2] https://github.com/Kelaxon/COVID19-Health-Rumor (accessed on 18 September 2025)

In addition, we conducted comparative experiments on the knowledge graph components using CMeKG and Disease-KB. CMeKG 2.0 is currently the largest publicly available Chinese medical knowledge graph, containing approximately 1.56 million medical knowledge triples. Disease-KB focuses on common disease-related information and includes approximately 310,000 relational entries.

5.1.2 Baseline Models

We compare our framework against a broader set of baseline models from four categories.

Shallow models:

- SVM-BOW [40] employs a support vector machine classifier on bag-of-words features as a naive lexical baseline.
- RFC [41] leverages handcrafted features including structural and linguistic statistics via random forest classification.
- CSI [42] introduces a hybrid approach combining user behavior scoring with textual modeling.

Deep text-based models:

- GRU [43] uses gated recurrent units to capture temporal evolution in rumor sequences.
- TextCNN [44] and BiLSTM [45] are classical CNN and bidirectional LSTM architectures, respectively, used to encode local n-gram features and global dependencies.
- BERT [36] and BERT-LSTM [46] represent pre-trained transformer-based encoders fine-tuned for classification, where BERT-LSTM appends a recurrent classifier on top of contextual embeddings.
- RoBERTa [47] is a robustly optimized variant of BERT that removes the next sentence prediction objective and dynamically changes masking patterns, often leading to stronger contextual encoding performance.

Graph-based models:

- TextGCN [48] constructs a document-word graph and uses GCN for text classification.
- Compare-GCN [49] captures inter-document semantic differences through GCN variants.
- CRNN [50] combines convolution and recurrent layers to capture both local and sequential features.
- Bi-GCN [51] performs bidirectional graph propagation over rumor trees, modeling both top-down and bottom-up rumor diffusion.

Knowledge-based models:

- KCNN [52] incorporates commonsense knowledge embeddings into convolutional text encoders.
- B-TransE [53] aligns textual features with TransE-encoded knowledge triples.
- K-BERT [54] integrates triples from external knowledge bases into the attention flow of BERT via soft position constraints.
- R-GCN [55] applies relational GCNs over multi-relational knowledge graphs to enrich node-level representations.

5.1.3 Implementation Details

We implemented our model using PyTorch 1.13.1 and conducted experiments on a server equipped with an Intel Xeon Platinum 8352V CPU (12 cores), 48 GB RAM, and dual NVIDIA RTX 3080 GPUs (20 GB VRAM each). Accuracy, Precision, Recall, and F1 score were used as evaluation metrics to assess model performance.

For causal graph construction and strength estimation, the significance level was set to 0.05, with the number of iterations $T = 200$. The tolerance threshold for unobserved confounding was set to 0.1. During

causal graph sampling, the number of candidate graphs was fixed at $Q = 100$ to balance computational efficiency and result stability. The BIC was employed for graph quality evaluation, with a regularization coefficient $\lambda = 0.01$ to prevent overfitting.

In knowledge graph-driven causal validation, the fuzzy matching threshold for entity alignment was set to $\theta = 0.85$. For multi-hop path reasoning, the maximum number of hops was set to $L_{max} = 3$, and the path decay factor $\gamma = 0.7$. The penalty coefficient for missing entities was set to $\beta = 1.2$, and tuned via cross-validation.

In the edge-enhanced dual-graph fusion classifier, node features were encoded using the hidden layer representations of BioBERT with dimensionality $d = 768$, and knowledge graph embeddings generated via the TransE algorithm had dimensionality $d_k = 256$. In the attention mechanism, the LeakyReLU negative slope was set to 0.2 to alleviate the vanishing gradient problem, and the number of attention heads $H = 4$ was used to enhance multi-view feature fusion. The weight coefficients of the loss function λ_1:λ_2:λ_3 were dynamically determined through Pareto frontier analysis. A path conflict relaxation factor $\gamma = 0.2$ was introduced to allow partial correction of errors in the knowledge graph.

The initial selection of these hyperparameter values was based on common practices in the relevant literature, preliminary experiments conducted on a reserved validation set, and the limitations imposed by our computational resources [10,11,13,14]. The sensitivity of these choices will be rigorously assessed in Section 5.2.8.

All models were trained using the Adam optimizer, with an initial learning rate $l_r = 3 \times 10^{-5}$, weight decay $wd = 0.01$, and dropout rate of 0.3. Layer normalization was applied to all graph attention layers to stabilize the training process. In the training strategy, the batch size was set to 16, and the number of training epochs was fixed at 100. Early stopping was triggered if the F1 score on the validation set did not improve for five consecutive epochs, in order to retain the best-performing model.

5.2 Experimental Results

5.2.1 Overall Performance

Table 3 summarizes the performance of CKDG and 18 baseline models across four categories. Overall, CKDG-CMeKG achieves the best results, with 91.28% accuracy and 90.38% F1-score, surpassing all shallow, deep text, graph-based, and knowledge-based baselines. The improvements over the strongest baseline model (K-BERT) reach 5.11% in accuracy and 3.29% in F1, demonstrating the effectiveness of our dual-graph reasoning mechanism.

Table 3: Performance comparison between CKDG and baseline methods on the fused dataset (%)

Type	Model	Accuracy	Precision	Recall	F1
Shallow	SVM-BOW	71.35	69.87	70.11	69.99
	RFC	72.17	70.45	71.23	70.84
	CSI	80.75	81.51	80.34	80.92
Deep-Text	GRU	77.26	76.82	77.98	77.40
	TextCNN	78.73	78.17	78.50	78.33
	BERT-LSTM	78.49	77.17	77.98	77.57
	BERT	80.93	80.67	82.12	81.39
	BiLSTM	81.45	81.19	81.54	81.36
	RoBERT	83.45	85.19	82.93	84.07

(Continued)

Table 3 (continued)

Type	Model	Accuracy	Precision	Recall	F1
Graph	TextGCN	79.43	79.33	79.02	79.17
	CRNN	81.63	82.49	82.28	82.38
	Compare-GCN	82.14	81.21	83.67	82.42
	Bi-GCN	85.42	84.75	86.57	85.65
Knowledge	KCNN	83.11	84.46	83.29	83.87
	B-TransE	84.25	83.91	85.33	84.61
	R-GCN	84.92	85.56	84.69	85.12
	K-BERT	86.17	86.32	87.87	87.09
CKDG	CKDG-DiseaseKB	88.56	87.76	87.49	87.62
	CKDG-CMeKG	**91.28**	**89.57**	**91.21**	**90.38**

Note: The bold values indicate the best results.

Traditional shallow classifiers such as SVM-BOW and RFC show limited performance, with F1-scores around 70%. CSI, which combines user behavior with textual features, achieves a relatively strong 80.92% F1, outperforming many deep models. However, such models still fall short in handling implicit semantics or domain-specific logic, making them less suitable for complex rumor detection scenarios.

Deep learning methods based on sequential and transformer architectures yield moderate improvements. Models like GRU, BiLSTM, and TextCNN show similar performance levels, while BERT and RoBERTa significantly outperform them, with RoBERTa reaching 84.07% F1 and 83.45% accuracy. This suggests that pre-trained language models are more effective at capturing contextual nuances. Nonetheless, these methods rely solely on text and are insensitive to external knowledge or causal structure, which limits their interpretability and robustness.

Graph-based approaches provide stronger performance, especially in modeling structural patterns within rumor propagation or entity interactions. Compare-GCN and CRNN both exceed 82% F1, while Bi-GCN attains 85.65%, outperforming all non-knowledge-based models. These results demonstrate the benefits of structural representation learning, although such models still lack domain validation and may be misled by noise in graph construction.

Introducing external knowledge further improves results. Knowledge-based models like KCNN, R-GCN, and B-TransE all achieve F1 scores above 84%. K-BERT, which directly integrates medical triples into the attention mechanism, obtains 87.09% F1, outperforming both text-only and graph-based methods. This supports the hypothesis that health rumor detection benefits significantly from background knowledge. However, these models do not reason over causal relationships, and their decision process remains pattern-driven.

The CKDG variants demonstrate clear superiority. Even using the smaller Disease-KB, CKDG already surpasses K-BERT by more than 0.5% in F1. When integrated with CMeKG, CKDG achieves the highest performance across all metrics. The combination of BioBERT-based entity representation, GFCI-derived causal structure, credibility-weighted edge features, and EGRET-based attention enables CKDG to dynamically integrate causal reasoning with domain verification, suppressing spurious correlations and improving generalization to complex health misinformation.

Taken together, these results highlight that shallow and deep models capture surface patterns, graph models capture structural propagation, and knowledge models introduce domain priors. CKDG unifies all three aspects and advances the state-of-the-art via fine-grained causal and knowledge fusion.

5.2.2 Ablation Study

In this subsection, we conduct experiments to evaluate the effectiveness of each module within CKDG. Using the CMeKG knowledge graph consistently, we report the average performance over five runs on the test set. As shown in Table 4, removing the knowledge graph validation module (-w/o KG) results in a 6.75% drop in accuracy and a 6.32% decrease in F1-score, which we attribute to the model's inability to filter out spurious causal relations that contradict established medical facts. This compromises its capacity to validate causal chains and renders it susceptible to data-driven pseudo-correlations, ultimately undermining detection reliability. Similarly, removing the causal inference module (-w/o Causal) leads to a 4.1% decrease in accuracy and a 2.43% decline in F1-score, highlighting the essential role of causal graph construction and strength estimation in capturing the latent causal logic within text; without it, the model struggles to differentiate between genuine and pseudo-causal relations implied in rumors. Moreover, excluding the edge-enhanced graph attention mechanism (-w/o EGRET) causes a 2.21% reduction in accuracy and a 2.58% drop in F1-score, demonstrating that dynamically integrating edge features and emphasizing edges with stronger causal strength can significantly enhance the model's capacity to handle complex causal topologies.

Table 4: Ablation study results on the CMeKG knowledge graph (%)

Variant model (CMeKG)	Accuracy	F1	Accuracy decrease	F1 score decrease
CKDG	91.28	90.38	–	–
- w/o Causal	87.18	87.95	4.10	2.43
- w/o KG	84.53	84.06	6.75	6.32
- w/o EGRET	89.07	87.80	2.21	2.58
- w/o Causal and KG	79.07	79.86	12.21	10.52

Notably, when both the causal inference and knowledge graph validation modules are removed (-w/o Causal and KG), the model experiences a 12.21% decrease in accuracy and a 10.52% decline in F1-score, indicating a non-linear performance degradation that strongly supports the effectiveness of CKDG's dual-graph fusion architecture.

5.2.3 Impact of Different Knowledge Graphs

To assess the impact of different knowledge graphs, we evaluated the alignment performance using two distinct knowledge graphs, as presented in Table 5. A comparison with the results in Table 3 reveals that CKDG-CMeKG achieves a significantly higher alignment success rate (95.7% for exact and fuzzy matching combined) than CKDG-DiseaseKB (81.77%). We attribute this performance gap primarily to differences in the scale and coverage of the two knowledge graphs. CMeKG, as a large-scale medical knowledge graph, demonstrates clear advantages in both the breadth and depth of domain knowledge. In contrast, although DiseaseKB is tailored to a specific domain, its relatively limited size constrains its capacity to represent complex medical entities. These results underscore the substantial influence of knowledge graph selection on the performance of the CKDG framework.

Table 5: Entity alignment performance using different knowledge graphs (%)

Knowledge graph	Exact match rate	Fuzzy match rate	Unaligned rate
CMeKG	44.27	51.43	4.30
Disease-KB	37.17	44.60	18.23

5.2.4 Impact of Different Causal Discovery Algorithms

This experiment compares four causal discovery algorithms—GFCI, PC, PCMCI, and PCGCE. As shown in Table 6, CKDG-GFCI achieves a significantly higher accuracy (91.28%) and F1 score (90.38%) than other variants, highlighting the unique advantages of GFCI in handling unobserved confounders and nonlinear causal interactions in medical texts. The PC algorithm (CKDG-PC, F1 = 87.10%) fails to accommodate potential confounders in health rumors due to its reliance on the causal sufficiency assumption. PCMCI (CKDG-PCMCI, F1 = 88.08%) imposes temporal constraints that conflict with the non-temporal causal chains often present in textual data, thereby weakening its generalization capability. Although PCGCE (CKDG-PCGCE, F1 = 89.74%) enhances representation through latent space embedding, its strong dependence on prior knowledge limits its robustness in knowledge-sparse scenarios. The results demonstrate that GFCI, through a nonparametric framework and dynamic constraint rules, generates highly reliable causal paths under noisy conditions, offering optimal logical support for the model.

Table 6: Accuracy and F1 scores of CKDG using different causal discovery algorithms (%)

Variant model (CMeKG)	Accuracy	F1
CKDG-GFCI	91.28	90.38
CKDG-PC	87.19	87.10
CKDG-PCMCI	87.82	88.08
CKDG-PCGCE	89.53	89.74

5.2.5 Impact of Text Length on Model Performance

To examine the impact of text length on model performance, we compared CKDG-CMeKG with representative baselines from each model family based on Table 3: CSI (shallow), RoBERTa (deep text), Bi-GCN (graph-based), and K-BERT (knowledge-based). The test set was partitioned by character count into four groups: short (0–20), medium (21–50), long (51–100), and ultra-long (>100) texts, each containing 70 samples. Evaluation was repeated five times, and averaged results are shown in Fig. 7.

CKDG achieves the best performance across all text lengths, with 86.12% accuracy and 85.59% F1 on short texts, and 90.76% accuracy and 90.03% F1 on ultra-long texts. The relatively strong results on short texts can be attributed to the presence of compact but informative medical terms, enabling CKDG's dual-graph design to leverage localized causal links and knowledge validation. Despite this, its performance remains slightly lower compared to longer inputs, due to sparse entities and limited graph depth in short samples. On medium and long texts, CKDG shows substantial improvements as longer context allows for richer causal structure construction and credibility filtering. Notably, Bi-GCN benefits significantly from longer inputs, reaching 90.07% accuracy and 88.52% F1 on long texts—second only to CKDG—demonstrating the strength of bidirectional rumor tree modeling. However, its performance plateaus on ultra-long inputs due to information redundancy and graph over-smoothing.

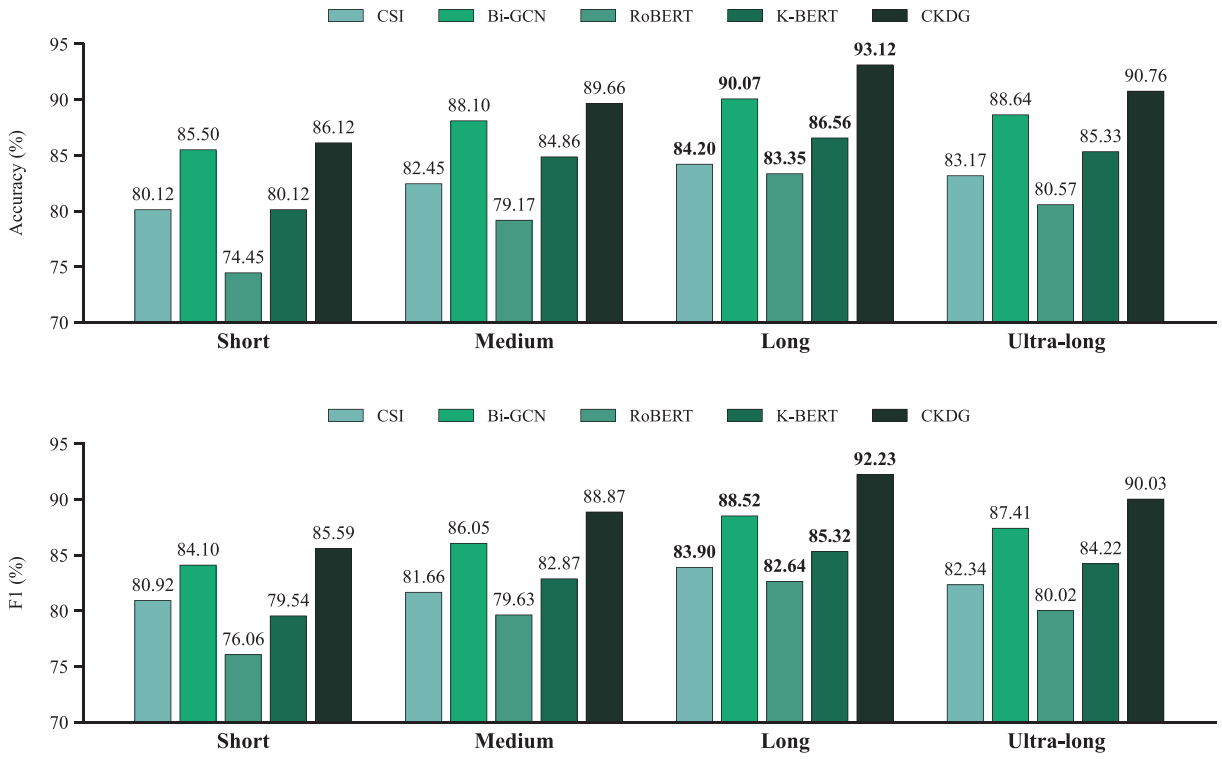

Figure 7: Accuracy and F1 scores of the best performing models for each category with different data volumes

K-BERT exhibits stable accuracy and F1 across all lengths (within a 5% range), confirming that incorporating structured knowledge contributes to robustness. Nonetheless, it lacks mechanisms for reasoning over causal inconsistencies, limiting its effectiveness in longer, more ambiguous inputs. RoBERTa shows relatively weaker performance throughout, particularly on short texts, likely due to its reliance on pure contextual embeddings without external knowledge or structure. CSI, despite being a shallow model, performs surprisingly well on short and medium texts, outperforming RoBERTa in F1. This suggests that behavior-aware heuristics still offer complementary cues in rumor detection, though their gains diminish on longer texts due to their limited semantic capacity.

5.2.6 Sensitivity Analysis

To evaluate the robustness of CKDG, we conducted sensitivity analysis on four key hyperparameters: entity alignment threshold (θ), path decay factor (γ), maximum hop number (L_{max}), and missing entity penalty coefficient (β). Fig. 8 illustrates their effects on performance, while Table 7 provides a concise summary. As shown, the model generally exhibits stable performance across a reasonable range of values, with clear optimal settings.

The sensitivity analysis reveals three key patterns in how hyperparameters affect CKDG's performance. First, the entity alignment threshold (θ) is the most critical parameter, with performance sharply declining when set outside the 0.80–0.90 range. This underscores the importance of precise entity matching for reliable knowledge validation. Second, both the path decay factor (γ) and missing entity penalty (β) exhibit moderate sensitivity, maintaining stable performance across broad ranges (0.65–0.75 for γ and 1.0–1.5 for β), indicating the framework's robustness to these parameter variations. Third, the maximum hop number (L_{max}) shows

minimal impact on F1 score beyond 3 hops, despite significantly increasing computational cost, making $L_{max} = 3$ the optimal efficiency-accuracy tradeoff.

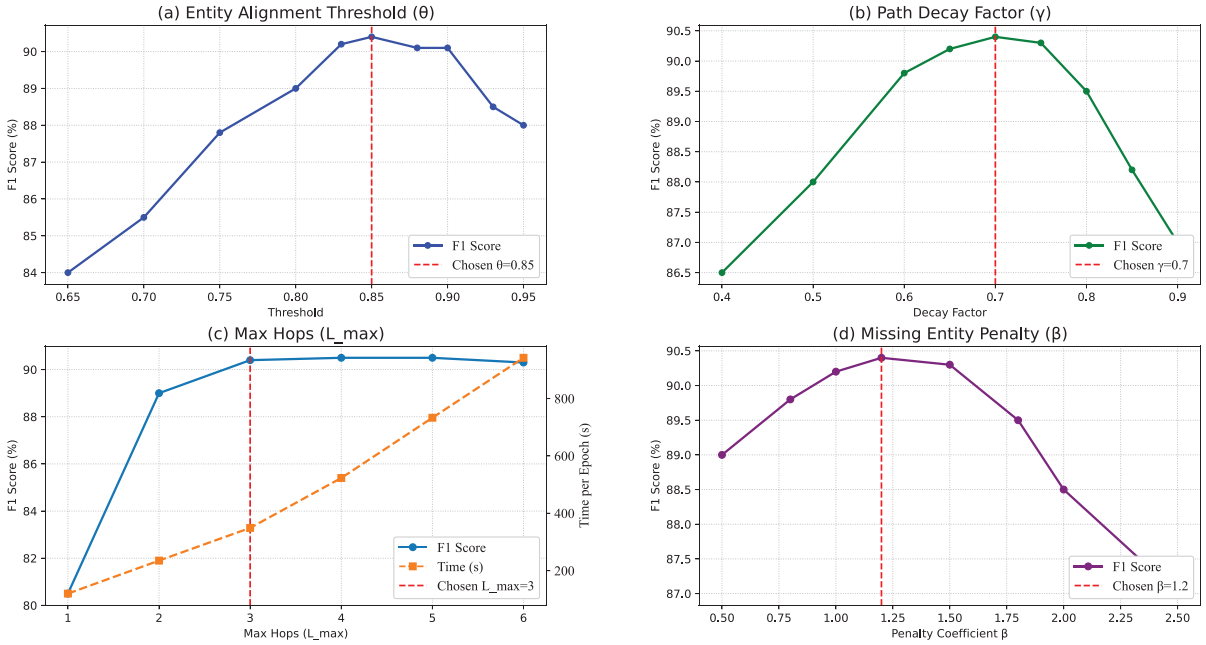

Figure 8: Sensitivity analysis of key hyperparameters. (a) Entity alignment threshold (θ). (b) Path decay factor (γ). (c) Maximum hop number ($L_{\max}$), showing F1-score (left axis) and computation time per epoch in seconds (right axis). (d) Missing entity penalty coefficient (β). The red dashed vertical line in each subplot indicates our chosen value

Table 7: Summary of key hyperparameter sensitivity analysis. The F1 range indicates the performance variation across tested values, while sensitivity reflects how strongly performance depends on parameter selection

Hyperparameter	Range tested	Optimal value	Sensitivity
Entity alignment threshold (θ)	0.65–0.95	0.85	High
Path decay factor (γ)	0.4–0.9	0.7	Medium
Maximum hop number (L_{max})	1–6	3	Low
Missing entity penalty (β)	0.5–2.5	1.2	Medium

Overall, these findings demonstrate that CKDG maintains robust performance across reasonable hyperparameter choices, with particularly stable behavior in knowledge integration parameters (γ, β). The framework shows sensitivity only to extreme parameter values, primarily in entity alignment quality. This robustness enhances the practical deployability of CKDG, as it reduces the need for extensive hyperparameter tuning in real-world applications while ensuring consistent detection performance.

5.2.7 Robustness Evaluation

To evaluate the practicality of CKDG in real-world scenarios, we conducted robustness tests under two categories of noise conditions: Lexical Noise (text content perturbation) and Edge Perturbation (graph topology perturbation). Specifically, lexical noise comprises three types: Shape Replacement, Random

Deletion, and Synonym Replacement[3]. For comparison, we selected three representative baseline models: TextGCN, K-BERT, and Bi-GCN. Noise was injected at rates of 0%, 5%, 10%, 15%, and 20%, with each condition run 5 times on the test set, and the average results were recorded. The final data are presented in Fig. 9.

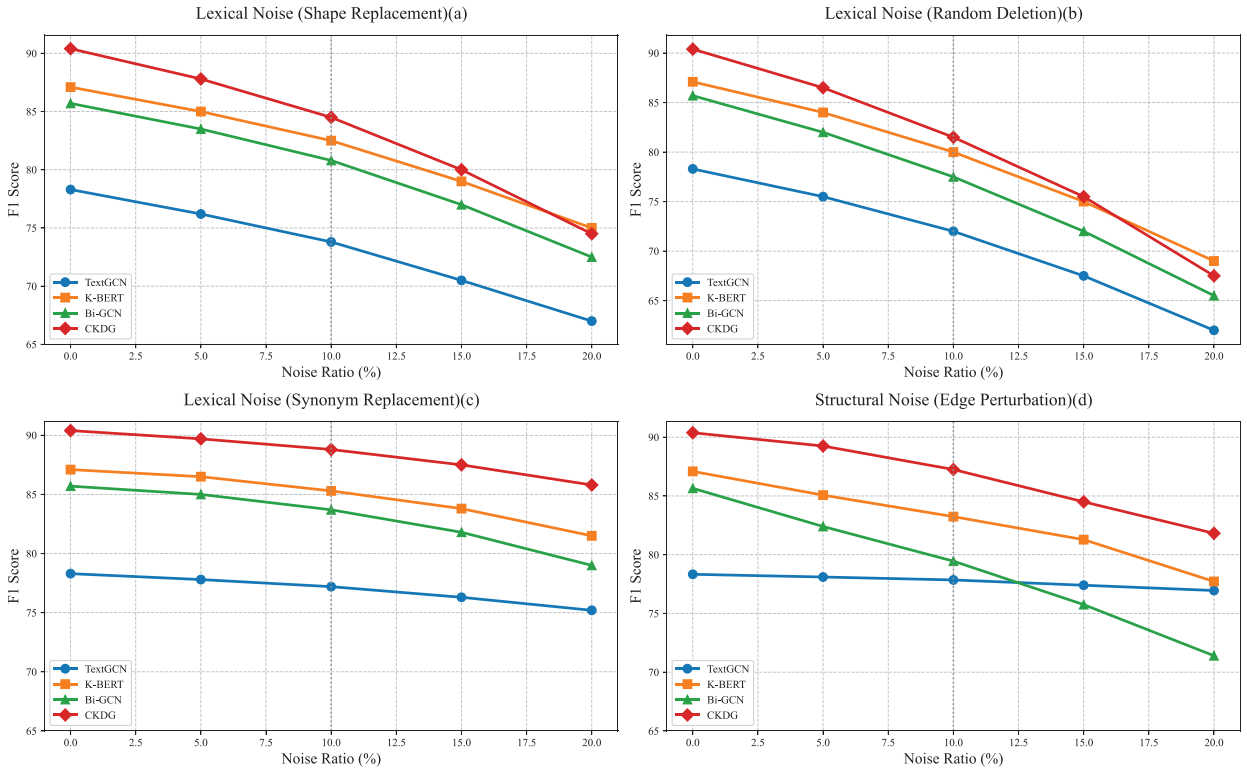

Figure 9: Robustness analysis under four types of noise: (a) shape substitution, (b) random deletion, (c) synonym substitution, and (d) structural noise. The *x*-axis represents the noise ratio, and the *y*-axis shows the F1-score

Experimental results indicate that different types of noise have distinct impacts on model performance. Lexical noise, shape replacement, and random deletion notably affect CKDG, with random deletion having the most significant impact. As the noise ratio increases, CKDG's performance drops sharply from 90.4% to 67.5%. This is primarily attributed to CKDG's reliance on BioBERT for extracting medical entities, where disruptions in the text directly impair the accuracy of entity recognition and subsequent graph construction. In contrast, synonym replacement has a relatively minor effect on CKDG, with performance declining only from 90.4% to 85.8%. This improvement is likely due to the integration of similarity calculations during the subsequent entity alignment step, which helps to mitigate the impact of changes in word meaning. For the baseline models, TextGCN, K-BERT, and Bi-GCN exhibit similar trends; however, CKDG consistently maintains a clear advantage across all noise levels, demonstrating its robustness in medical entity comprehension and graph structure reasoning.

In the Edge Perturbation experiment, CKDG's performance decreased from 90.38% to 81.82%, a smaller decline compared to that caused by lexical noise. We believe this is primarily attributable to the design of the GFCI causal discovery algorithm, which permits the existence of unobserved confounders, thus mitigating

[3] https://github.com/One-sixth/HIT-IR-Lab-Tongyici-Cilin-Extended (accessed on 18 September 2025)

the impact of graph structure perturbations on the model. In contrast, other baseline models, lacking a similar causal robustness mechanism, exhibit greater sensitivity to structural noise.

This experiment demonstrates that CKDG exhibits robust characteristics when confronted with both text content perturbations and graph structure disturbances: it is sensitive to highly disruptive lexical noise, such as random deletions, minimally affected by synonym substitutions, and possesses a certain tolerance for graph structure perturbations, making it appropriate for scenarios with stable input conditions.

5.2.8 Model-Level Efficiency Comparison

To comprehensively evaluate the computational characteristics of various rumor detection methods in practical deployment, we measured the inference efficiency and GPU memory usage of all models under uniform hardware(NVIDIA RTX 3080 GPU, 48 GB RAM) and consistent parameters(batch size = 16, sequence length = 128). Table 8 presents the average inference time per sample and the peak GPU memory consumption during the evaluation phase.

Table 8: Comparison of inference efficiency and memory consumption

Model	Inference time (ms/sample)	Peak GPU memory (MB)
CKDG	41.75	8245.33
K-BERT	11.92	4128.50
Bi-GCN	24.63	3845.75
RoBERTa	7.85	2987.20
CSI	4.20	187.50

The results exhibit efficiency levels corresponding to the complexity of the model architectures. CKDG imposes the highest computational demand, requiring 41.75 ms per inference with a peak memory usage of 8245.33 MB. This overhead arises from its multi-stage process, including BioBERT-based entity extraction, GFCI-based causal discovery, medical knowledge validation using CMeKG, and EGRET dual graph inference. The peak memory consumption is primarily due to the persistent memory occupation by the medical knowledge graph and intermediate graph structures. Among the compared models, CSI is the most efficient due to its shallow architecture and feature-based design (4.20 ms per item, 187.50 MB). Moreover, RoBERTa, serving as an optimized Transformer encoder, also demonstrates excellent performance (7.85 ms per item, 2987.20 MB). Graph-based models (Bi-GCN) and knowledge-enhanced models (K-BERT) exhibit moderate efficiency, as their graph computations and knowledge injection mechanisms incur measurable additional overhead compared to purely text-based models.

The computational characteristics of each model reflect the inherent trade-off between architectural complexity and operational efficiency. Although CKDG demands higher resources, it prioritizes detection accuracy through explicit causal reasoning and knowledge verification—capabilities that are absent in more efficient models. Therefore, CKDG is particularly suitable for application scenarios that require extremely high accuracy and can accommodate batch processing or offline analysis. Therefore, we recommend that in practical deployments, CKDG is most appropriate for high-risk areas such as medical rumor detection, public health monitoring, or policy-related debunking, where the demands for accuracy and interpretability outweigh the computational costs. For real-time or resource-constrained environments (e.g., social media stream filtering), lighter models (such as RoBERTa or CSI) might be more optimal. Consequently, when the risk of misclassification is extremely high and computation can be allocated through offline or batch inference, CKDG should be considered a priority.

5.3 Case Study

5.3.1 Reasoning Analysis

We selected a real-world health rumor case from the test set, as illustrated in Fig. 10. While baseline models such as Compare-GCN and RoBERT misclassified the instance as credible, our CKDG model successfully identified its pseudoscientific nature. As shown in Fig. 10, the causal graph construction module extracted four causal chains from the text and constructed a corresponding causal strength graph. Notably, the edge *ionized water* → *free radicals* was associated with a high estimated causal strength of 0.9, largely due to the phrase *removes*, suggesting a strong textual correlation. However, the knowledge verification module assigned this edge a low credibility score of 0.1. Semantic alignment revealed two critical contradictions: (1) biomedical knowledge bases report no evidence of ionized water affecting human free radical metabolism, and (2) the knowledge graph contains neither a direct link nor any valid multi-hop paths within a three-hop range (with a path decay factor $\gamma = 0.7$) to support the claim.

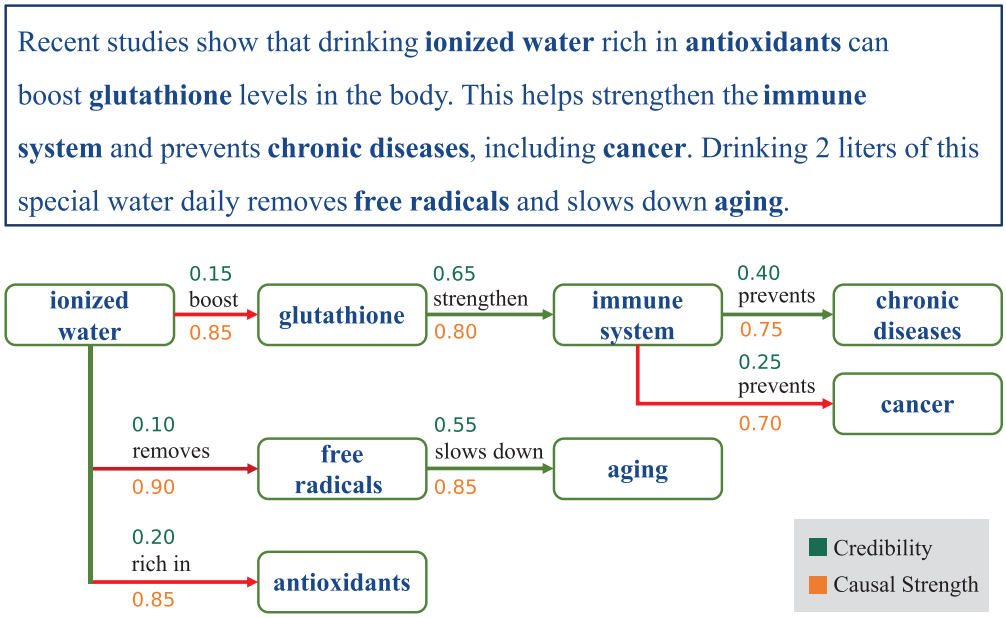

Figure 10: Rumor content and causal graph after attention calculation[4]

It is noteworthy that the downstream causal chain *glutathione* → *immune system* → *chronic disease* (with causal strengths of 0.82 and 0.79, respectively) aligns closely with the results of knowledge verification (with confidence scores of 0.85 and 0.83). However, as revealed by CKDG through EGRET, the attention weight $\alpha_{ij} = 0.12$ assigned to this subpath within the global causal topology is merely 0.12—significantly lower than its independently validated value. This discrepancy arises from the error propagation effect triggered by an upstream pseudo-causal edge: when the edge *ionized water* → *free radicals* is assigned a low confidence score due to failed knowledge verification, its downstream subpaths, although locally valid, compromise the overall coherence of the causal chain. By applying a path coherence constraint ($\rho < 0.7$) within its hierarchical pooling strategy, the model identifies such "locally valid but globally spurious" hybrid causal chains as high-risk paths, ultimately activating the anomaly detection mechanism of the classifier.

[4]The original data in this case was in Chinese. To facilitate reading, it has been translated into English.

5.3.2 Interpretability for End-Users

Beyond causal reasoning, CKDG can provide end-users (e.g., health professionals and regulators) with interpretable results by explicitly mapping low-credibility causal relations back to the original sentences. For example, in Fig. 11, the sentence *"Recent studies show that drinking ionized water rich in antioxidants can boost glutathione levels in the body."* is linked to two causal edges (*ionized water* → *antioxidants*, *ionized water* → *glutathione*). However, their credibility scores are only 0.20 and 0.15, respectively, and are not supported by biomedical knowledge graphs. Therefore, CKDG highlights this claim as suspicious and provides a cautionary note. Such interpretability enhancement transforms CKDG from a black-box classifier into a decision-support tool that helps professionals understand *why* a claim is problematic, rather than merely labeling it as a rumor.

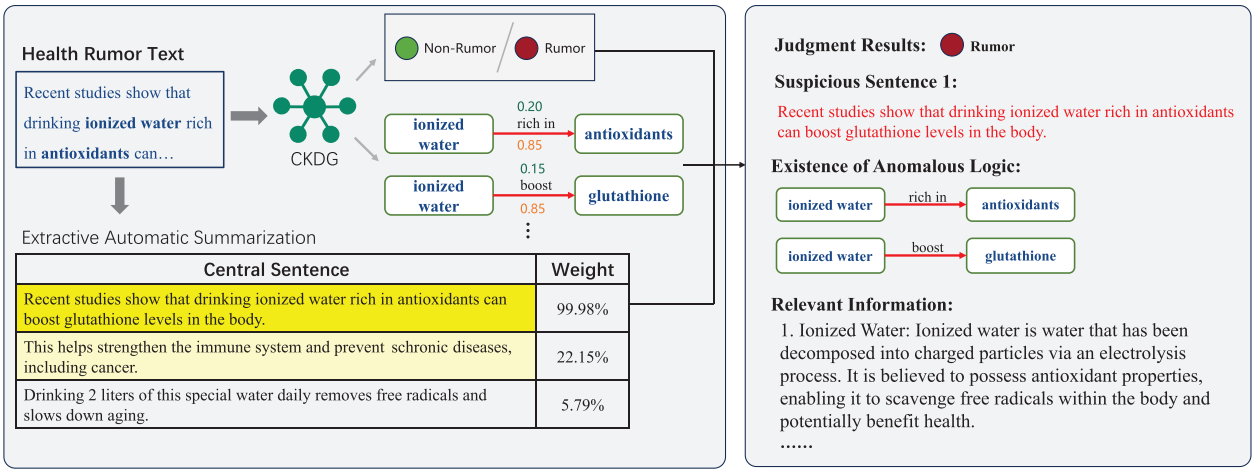

Figure 11: Interpretability enhancement in case study. CKDG maps low-credibility causal edges (red dashed arrows) back to their original sentences, providing end-users with transparent explanations and evidence

6 Conclusion

To address the limitations of existing health rumor detection methods, particularly their reliance on label-specific features and their difficulty in identifying the misuse of medical knowledge and pseudoscientific logic, this paper proposes a dual-graph fusion framework (CKDG) that integrates causal reasoning with domain knowledge graphs. By constructing an interaction mechanism between implicit textual causal graphs and medical knowledge graphs, and incorporating an Edge Aggregated Graph Attention Network alongside a hierarchical pooling strategy, the model robustly captures complex causal topologies and domain logic within health rumors.

Experimental results on a dataset of 8430 health-related rumors demonstrate that CKDG, when combined with the CMeKG knowledge graph, achieves an accuracy of 91.28% and an F1 score of 90.38%, outperforming the best baseline by 5.11% and 3.29%, respectively. Ablation studies reveal that the knowledge verification and causal reasoning modules contribute accuracy gains of 6.75% and 4.1%, confirming the critical role of dual-graph synergy in suppressing spurious correlations.

Nonetheless, several limitations remain. First, the incomplete coverage of rare medical entities in current knowledge graphs may hinder the verification of unaligned concepts. Second, the construction of causal graphs relies on text extraction quality, posing challenges for handling ambiguous or redundant expressions. Third, the model's computational complexity restricts its applicability in real-time detection scenarios. Moreover, the current framework is tailored to Chinese health rumor texts and Chinese medical

knowledge graphs, which limits its generalizability to other languages and regions. To extend CKDG to multilingual settings, future work could incorporate pretrained cross-lingual language models such as mBERT or XLM-R to support entity extraction and alignment across languages. Additionally, integrating multilingual or language-agnostic knowledge bases like UMLS or Wikidata can provide broader domain coverage. The causal discovery module may also be adapted to handle semantic variations across languages by leveraging translation-invariant representations or universal embedding spaces. Future work will also explore lightweight graph reasoning algorithms, dynamic knowledge graph updates, multilingual adaptation, and multimodal data integration to further enhance model generalizability.

Acknowledgement: This work was supported by the College of Information Science and Engineering at Hunan Institute of Engineering and the HUGONG ACM Laboratory.

Funding Statement: This research was funded by the Hunan Provincial Natural Science Foundation of China (Grant No. 2025JJ70105) and the Hunan Provincial College Students' Innovation and Entrepreneurship Training Program (Project No. S202411342056). The article processing charge (APC) was funded by the Project No. 2025JJ70105.

Author Contributions: Conceptualization, Haoran Lyu; Formal analysis, Ning Wang and Haoran Lyu; Funding acquisition, Ning Wang; Investigation, Haoran Lyu; Methodology, Ning Wang and Haoran Lyu; Software, Haoran Lyu; Supervision, Ning Wang; Visualization, Yuchen Fu; Writing—original draft, Haoran Lyu; Writing—review & editing, Yuchen Fu. All authors will be updated at each stage of manuscript processing, including submission, revision, and revision reminder, via emails from our system or the assigned Assistant Editor. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analysed during the current study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. "Cyberchondriacs" on the Rise? The Harris Poll; 2010 [Internet]. [cited 2025 Aug 25]. Available from: https://www.fiercehealthcare.com/healthcare/cyberchondriacs-rise.
2. Chua AY, Banerjee S. To share or not to share: the role of epistemic belief in online health rumors. Int J Med Inform. 2017;108(6):36–41. doi:10.1016/j.ijmedinf.2017.08.010.
3. Yin M, Liu L, Cheng L, Li Z, Tu Y. A novel group multi-criteria sorting approach integrating social network analysis for ability assessment of health rumor-refutation accounts. Expert Syst Appl. 2024;238(2):121894. doi:10.1016/j.eswa.2023.121894.
4. Bhatnagar B, Office WP, Wadia R. Working together to tackle the "infodemic". 2022 [Internet]. [cited 2025 Aug 25]. Available from: https://www.who.int/europe/news-room/29-06-2020-working-together-to-tackle-the-infodemic-.
5. Guan R, Zhang H, Liang Y, Giunchiglia F, Huang L, Feng X. Deep feature-based text clustering and its explanation. IEEE Trans Knowl Data Eng. 2022;34(8):3669–80. doi:10.1109/icde55515.2023.00362.
6. Yang Z, Lin J, Guo Z, Li Y, Li X, Li Q, et al. Towards rumor detection with multi-granularity evidences: a dataset and benchmark. IEEE Trans Knowl Data Eng. 2024;36(11):7188–200. doi:10.1109/tkde.2024.3401700.
7. Fionda V, Pirrò G. Fact checking via evidence patterns. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden. p. 3755–61.
8. Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv. 2020;53(5):1–40. doi:10.1145/3395046.

9. Li J, Li R, Ni S, Kao HY. EPRD: exploiting prior knowledge for evidence-providing automatic rumor detection. Neurocomputing. 2024;563(8):126935. doi:10.1016/j.neucom.2023.126935.
10. Haque A, Abulaish M. CARES: a commonsense knowledge-enriched and graph-based contextual learning approach for rumor detection on social media. Expert Syst Appl. 2025;266(2):125965. doi:10.1016/j.eswa.2024.125965.
11. Yao X, Zhao Y, Gao N, Du H, Huang H. Causal related rumors controlling in social networks of multiple information. IEEE ACM Trans Netw. 2024;32(3):2085–98. doi:10.1109/tnet.2023.3337774.
12. Wang J, Qian S, Hu J, Dong W, Huang X, Hong R. End-to-end explainable fake news detection via evidence-claim variational causal inference. ACM Trans Inf Syst. 2025;43(4):1–26. doi:10.1145/3728462.
13. Zhang Y, Huang S. Automatic rumor recognition for public health and safety: a strategy combining topic classification and multi-dimensional feature fusion. J King Saud Univ—Comput Inf Sci. 2024;36(5):102087. doi:10.1016/j.jksuci.2024.102087.
14. Huang J, Cao D, Lin D. Leverage causal graphs and rumor-refuting texts for interpretable rumor analysis. In: ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 9946–50.
15. Yuan X, Shu A, Wu Y, Liu J, Wang M, Liu J. MSSCR: multi-scale semantic collaborative reasoning model for explainable multimodal rumor detection. Neurocomputing. 2025;653(6380):131042. doi:10.1016/j.neucom.2025.131042.
16. Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, et al. Causal inference in natural language processing: estimation, prediction, interpretation and beyond. Trans Assoc Comput Linguist. 2022;10(1):1138–58. doi:10.1162/tacl_a_00511.
17. Zhang W, Zhong T, Li C, Zhang K, Zhou F. CausalRD: a causal view of rumor detection via eliminating popularity and conformity biases. In: IEEE INFOCOM 2022—IEEE Conference on Computer Communications. 2022 May 2–5; Online. p. 1369–78.
18. Chen Z, Hu L, Li W, Shao Y, Nie L. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, ON, Canada: Association for Computational Linguistics; 2023. p. 627–38.
19. Sheng Q, Cao J, Li S, Wang D, Zhuang F. Generalizing to the future: mitigating entity bias in fake news detection. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2022. p. 2120–5.
20. Sheng Q, Zhang X, Cao J, Zhong L. Integrating pattern-and fact-based fake news detection via model preference learning. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021 Nov 1–5; Onine. p. 1640–50.
21. Sun M, Zhang X, Zheng J, Ma G. DDGCN: dual dynamic graph convolutional networks for rumor detection on social media. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36; Washington, DC, USA: AAAI; 2022. p. 4611–9.
22. Hu L, Yang T, Zhang L, Zhong W, Tang D, Shi C, et al. Compare to the knowledge: graph neural fake news detection with external knowledge. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021 Aug 1–6; Online. p. 754–63.
23. Yang SH, Chen CC, Huang HH, Chen HH. Entity-aware dual co-attention network for fake news detection. arXiv:2302.03475. 2023.
24. Speer R, Chin J, Havasi C. ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the 31th AAAI Conference on Artificial Intelligence. Vol. 31; 2017 Feb 4–9; San Francisco, CA, USA. p. 4444–51.
25. Bauer L, Wang Y, Bansal M. Commonsense for generative multi-hop question answering tasks. arXiv:1809.06309. 2018.
26. Xie Y, Liu Z, Ma Z, Meng F, Xiao Y, Miao F, et al. Natural language processing with commonsense knowledge: a survey. arXiv:2108.04674. 2021.

27. Zhang H, Li Y, Zhu T, Li C. Commonsense-based adversarial learning framework for zero-shot stance detection. Neurocomputing. 2024;563(3):126943. doi:10.1016/j.neucom.2023.126943.
28. Zhong P, Wang D, Miao C. Knowledge-enriched transformer for emotion detection in textual conversations. arXiv:1909.10681. 2019.
29. Sharma S, Arora U, Akhtar MS, Chakraborty T. MEMEX: detecting explanatory evidence for memes via knowledge-enriched contextualization. arXiv:2305.15913. 2023.
30. Liu X, Yin D, Feng Y, Wu Y, Zhao D. Everything has a cause: leveraging causal inference in legal text analysis. arXiv:2104.09420. 2021.
31. Mahbub S, Bayzid MS. EGRET: edge aggregated graph attention networks and transfer learning improve protein-protein interaction site prediction. Brief Bioinform. 2022;23(2):bbab578. doi:10.1101/2020.11.07.372466.
32. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. 2019;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
33. Pearl J, Mackenzie D. The book of why: the new science of cause and effect. London, UK: Penguin Books Limited; 2018.
34. Ogarrio JM, Spirtes P, Ramsey J. A hybrid causal search algorithm for latent variable models. In: Proceedings of the Eighth International Conference on Probabilistic Graphical Models; 2016. Vol. 52, p. 368–79.
35. Li L, Lian R, Lu H, Tang J. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022 Oct 12–17; Gyeongju, Republic of Korea. p. 2098–107.
36. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
37. Lee J, Lee I, Kang J. Self-attention graph pooling. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 3734–43.
38. Ma Z, Liu M, Fang G, Shen Y. LTCR: long-text Chinese rumor detection dataset. arXiv:2306.07201. 2023.
39. Yang W, Wang S, Peng Z, Shi C, Ma X, Yang D. Know it to defeat it: exploring health rumor characteristics and debunking efforts on Chinese social media during COVID-19 crisis. arXiv:2109.12372. 2021.
40. Ma J, Gao W, Wong KF. Rumor detection on twitter with tree-structured recursive neural networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, VIC, Australia. p. 1980–9.
41. Kwon S, Cha M, Jung K, Chen W, Wang Y. Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining; 2013 Dec 7–10; Dallas, TX, USA. p. 1103–8.
42. Ruchansky N, Seo S, Liu Y. CSI: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017 Nov 6–10; Singapore. p. 797–806.
43. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. arXiv:1611.06639. 2016.
44. Chen YC, Liu ZY, Kao HY. IKM at SemEval-2017 Task 8: convolutional neural networks for stance detection and rumor verification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017 Aug 3–4; Vancouver, BC, Canada. p. 465–9.
45. Augenstein I, Rocktäschel T, Vlachos A, Bontcheva K. Stance detection with bidirectional conditional encoding. arXiv:1606.05464. 2016.
46. Wang J, Wang X, Yu A. Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. Sci Rep. 2025;15(1):1118. doi:10.21203/rs.3.rs-4116981/v1.
47. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
48. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. Proc AAAI Conf Artif Intell. 2019;33(1):7370–7. doi:10.1609/aaai.v33i01.33017370.

49. Vashishth S, Sanyal S, Nitin V, Talukdar P. Composition-based multi-relational graph convolutional networks. arXiv:1911.03082. 2019.
50. Liu Y, Wu YF. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: AAAI'18: AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA. p. 354–61.
51. Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, et al. Rumor detection on social media with bi-directional graph convolutional networks. Proc AAAI Conf Artif Intell. 2020;34(1):549–56. doi:10.1609/aaai.v34i01.5393.
52. Wang H, Zhang F, Xie X, Guo M. DKN: deep knowledge-aware network for news recommendation. In: Proceedings of the 2018 World Wide Web Conference; 2018 Apr 23–27; Lyon, France. p. 1835–44.
53. Pan JZ, Pavlova S, Li C, Li N, Li Y, Liu J. Content based fake news detection using knowledge graphs. In: International Semantic Web Conference. Cham, Switzerland: Springer; 2018. p. 669–83.
54. Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, et al. K-BERT: enabling language representation with knowledge graph. Proc AAAI Conf Artif Intell. 2020;34(3):2901–8. doi:10.1609/aaai.v34i03.5681.
55. Schlichtkrull M, Kipf T, Bloem P, Berg RVD, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: The Semantic Web (ESWC 2018). Cham, Switzerland: Springer; 2017. p. 593–607.