Sciences du Nord Nature Science and Technology
Vol. 1, No. 01, June 2024: pp. 51-60
E-ISSN: 3046-6504
P-ISSN: 3047-4922
DOI: 10.71238/snnst.v2i02.181

Risk Graph-NLP: Graph-Assisted Language Modeling for Discovering Hidden Compliance Relationships in Enterprise Documents

Varsha Shah [1]
[1] Independent Researcher, Seattle, WA, USA varshachaudhari17@gmail.com
Coresponding Author: varshachaudhari17@gmail.com

ARTICLE HISTORY

Received June, 2024 **Revised June, 2024** **Accepted June, 2024**

ABSTRACT

Natural language processing is becoming a key tool for enterprise companies to deal with compliance obligations spread throughout contracts, policies and regulatory filings. Plain token-based pipelines do not easily bring to the surface relationships that cross clauses, sections, and/or files, resulting in a lack of compliance risk visibility and auditability. In this paper, we present a synthesis of a framework, RiskGraph-NLP, that proposes to combine contextual language models with graph neural networks to represent enterprise documents as heterogeneous compliance graphs, encompassing clauses, named entities, obligations, and regulations. The framework is grounded in three key research trends: transformer-based language representation, document-level relation extraction, and knowledge graph embedding, as well as graph-based fraud detection, with the goals of overcoming the limitations of sequential language models. The framework is inspired by three research directions, namely, transformer-based representation of language, document-level relation extraction, and knowledge graph embedding, along with graph-based fraud detection, with the aims of addressing the weaknesses of sequential language models. The synthesis shows that graph-assisted encoders are structurally more appropriate than sentence-bound models to show indirect or latent compliance dependencies, such as an obligation that is not explicitly referenced in the regulated text but is only implicitly referenced by an intermediate cross-reference. The topic is also explored from the architectural perspective, with a focus on the pros and cons of convolutional and attention-based graph encoders, as well as the ethical and regulatory implications of automated compliance inference in legal and financial applications. The paper concludes that hybrid language-graph architectures appear to be an exciting, but still developing, path for enterprise compliance analytics and provides methodological priorities for future empirical validation.

Keywords: *Compliance Analytics; Document-Level Relation Extraction; Enterprise Risk Management; Graph Neural Networks; Heterogeneous Graphs; Knowledge Graphs; Legal Text Mining; Natural Language Processing; Regulatory Technology; Transformer Models*

INTRODUCTION

The number and extent of regulations and laws related to data protection, financial behaviour, and corporate governance have grown significantly, and enterprises are tasked with balancing obligations that span between contracts, internal policies, and external statutory documents. An obligation to comply might appear in one clause, be modified in another section, and be coupled with a regulation that is several pages further away, leaving the full picture only as a result of the various pieces that are scattered in isolation. However, standard text-processing pipelines, which work on documents as linear strings, have a limited structure in the sense that they are unable to support such distributed dependencies because of the required attention windows and sentence boundaries [1], [2].

The transformer architecture has significantly improved the representation of language, through the use of bidirectional reasoning over the sequences of tokens [3], [4]. On this base, domain-adapted versions pre-trained on legal or financial corpora have shown that the usage of specific terminology and conventions significantly influences downstream tasks with regulatory and contractual text [5]. Meanwhile, progress in the field of document-level relation extraction has demonstrated that many useful relations in enterprise text have scope beyond a single sentence window (e.g., [1], [2].

An alternative representational paradigm which is complementary to this is the graph neural network (GNN), where entities are represented as nodes and relationships as edges, enabling information to spread out over several hops of the structured neighborhood [6]. Compliance dependencies are naturally relational, relating parties, obligations, clauses, and statutory provisions, which makes a graph-structured representation seem natural to this problem of discovering non-explicit relationships in a single text passage. We present the conceptual underpinnings of a hybrid approach that combines a contextual language encoder with graph-based reasoning to identify hidden compliance relationships between enterprise documents, and we summarize the methodological building blocks, comparative evidence, and open challenges of this approach.

LITERATURE REVIEW

2.1. Contextual Language Representation

Using self-attention to relate every token in a sequence to every other one, instead of recurrent and convolutional sequence encoders that required fixed path length, the introduction of the transformer architecture greatly enhanced translation quality and training efficiency [3]. On top of this architecture, masked language modeling and next-sentence prediction bidirectional pre-training yielded representations that successfully generalised to a large variety of language-understanding tasks in the downstream setting [4]. Most of the domain-specific and structure-aware extensions that followed have been based on these basic models.

The vocabulary and context association learned from general-purpose corpora (such as web and news text) may not be the same as those learned from specialized ones (such as statutory or contractual text). It has been demonstrated that lexical and syntactic specialization can significantly impact the quality of downstream inference related to compliance, with continued pre-training on domain-specific corpora (such as for legal text and financial communications) outperforming generic pre-trained baselines on domain benchmarks [5], [7].

2.2. Structural and Layout-Aware Encoding

Enterprise documents often contain non-linear structure such as multi-column layouts, headers, and tables, which are often lost by plain-text extraction pipelines. This subject has been addressed as a way to maintain the structural information of the text side by side with a two-dimensional position of the layout, allowing downstream models to use spatial information to interpret form-like or semi-structured documents [8]. This layout-aware encoding is especially important for compliance documents, where obligations may be presented in schedules, annexes, and tabular disclosures, and not merely as continuous prose.

2.3. Entity and Relation Extraction Across Document Spans

Named entity recognition, covering organized parties, regulatory bodies, and defined terms, is a key property required to build any graph-based model of documents [9]. In addition to recognizing entities, relation extraction techniques have gradually shifted from sentence-level classifiers to document-level formulations that are able to link mentions occurring in different sentences or paragraphs. Other methods for document-level relation extraction build intermediate graphs between mentions and entities, and propagate this graph to find relations across the entire document instead of a local context [1], [2]. This approach shows empirically that learning relationships by explicitly modeling the graph structure leads to better recall of relations that would otherwise require long-range dependency modeling beyond what can be practically captured by purely sequential encoders.

2.4. Graph Neural Networks and Knowledge Graph Reasoning

Graph convolutional networks (GCNs) extend the existing convolution operation to irregularly structured graph data by combining and manipulating local neighborhood features of a node [6]. Graph attention networks go beyond this approach by learning attention coefficients for each adjacent node instead of using a fixed structural normalization, so that the model can attend more to informative relational paths [10]. Moreover, in enterprise contexts where the number of clauses, parties, or documents is constantly growing, inductive frameworks such as GraphSAGE generalize neighborhood aggregation to unseen nodes during inference, a capability that is crucial in such scenarios. A survey of this literature shows that graph neural architectures can be broadly classified into convolutional, attentional, and recurrent types, each with its own trade-offs in terms of expressiveness, scalability, and inductive capability [6], [11].

Relational and heterogeneous graph architectures generalize the base GNN formulation, capturing type-specific transformations and attention, when compliance relationships are naturally heterogeneous, with multiple node and edge types (parties, obligations, provisions, etc.) [12], [13]. Complementary work on knowledge graph embedding (KGE) aims to encode entities and relations in continuous vector space for the purpose of link prediction and relation inference over large-scale relational data sets, providing another, complementary way to extract undocumented relationships between entities of interest for compliance [14], [15]. Similarly, graph-based propagation has been found to be superior to sequential classification techniques for text classification, where the graph is constructed over documents and words and can benefit from the global word co-occurrence structure [16].

2.5. Graph-Based Applications in Regulatory and Financial Domains

Relationships among transactions have been shown to contain discriminative signals beyond what can be captured using only transactional features: in the financial forensics domain, transaction networks are modeled as graphs, and patterns indicative of illegal activity such as money laundering have been detected [17]. Inductive graph representation learning has also been used to tackle fraud detection problems that require generalization to unseen transactions or new accounts [20]. More pertinent to contractual compliance, automated checking of data processing agreements against data protection regulation has gone beyond mapping the data-processing clause set to the regulation; it has applied NLP techniques and structured requirement modeling to check the conformity of each clause, thereby offering a precedent for treating regulatory compliance as a structured matching problem rather than merely a classification problem [18].

METHODS

THE RISKGRAPH-NLP FRAMEWORK

Combining the methodological principles discussed above, RiskGraph-NLP is envisioned as a multi-layered system capable of mapping unstructured enterprise text into a heterogeneous compliance graph and applying graph-assisted encoding to extract relationships that are not explicit in any particular clause. The conceptual processing pipeline is shown in Figure 1.

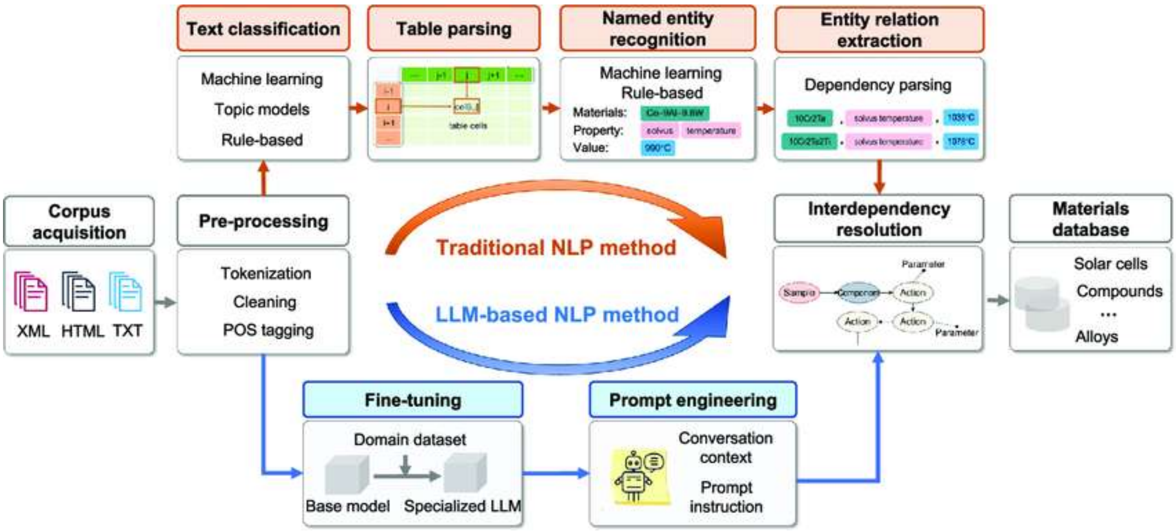

Figure 1. Conceptual processing pipeline of the RiskGraph-NLP framework.

The pipeline starts with layout-aware text encoding, where the document text is encoded together with its structural position to preserve information typically conveyed by a document's tables, headers, and section hierarchy [8]. Parties, defined terms, obligations, and regulatory references are then recognized in a named entity recognition phase [9], and candidate relations are found in a document-level relation extraction phase that includes multi-sentence and multi-section relations [1], [2]. The resulting entities and relations are then combined into a compliance graph, in which the nodes are clauses, obligations, named entities, and regulatory provisions, and the typed edges are relations such as imposes, binds, or grounded in, as shown schematically in Figure 2.

Figure 2. Schematic node and edge typology of the heterogeneous compliance graph.

The dashed edge in Figure 2 represents the class of relationship of principal interest to the framework: one that is not explicitly stated in any single clause, but may be implied by traversing an intermediate obligation and a cross-reference anchor. Because this type of relationship requires reasoning over several graph hops, a graph neural encoder is applied to the constructed compliance graph. For the convolutional formulation, node representations are updated by following the layer-wise propagation rule of graph convolutional networks [19]:

$$H(l+1) = \sigma(\tilde{D}^{(-1/2)}\ \tilde{A}\ \tilde{D}^{(-1/2)}\ H(l)\ W(l))$$

Where Ã is the adjacency matrix of the compliance graph with self-connections, $\tilde{D}$ is the corresponding degree matrix, H(l) is the node representation at layer l, W(l) is a trainable layer-specific weight matrix, and σ is a nonlinear activation function [17]. Alternatively, an attention-focused formulation may be preferred, in which the contribution of each neighboring node j to node i is given by a learned coefficient rather than a fixed structural normalization [10]:

$$\alpha ij = softmax\_j(LeakyReLU(a^T [W hi \parallel W hj]))$$

In the above, hi and hj represent the input representations of nodes i and j, W is a shared linear transformation, a is a learnable attention vector, and ‖ is the concatenation operator [10]. Consistently, the attention coefficients in the resulting output of the framework assign greater weight to, for instance, an obligation-to-provide edge that comes from an explicit statutory citation, compared to a weaker structural co-occurrence edge — an outcome consistent with the differential treatment of relational evidence in relational graph convolutional formulations [12].

For clauses and obligations, underlying node features are initialized from contextual language representations rather than static embeddings, consistent with the objective of bidirectionally encoding language in pre-trained models [1]:

$$L(MLM) = -\ \Sigma\_i \log P(xi \mid x \setminus i\ ;\ \theta)$$

where xi denotes a masked token, x \ i denotes the surrounding context, and θ denotes the model parameters [4]. Initializing graph nodes with such contextually pre-trained representations, optionally continued on domain-specific legal or financial corpora [5], [7], allows the graph encoder to operate over semantically informed rather than purely lexical node features, combining the complementary strengths of sequence-level and structure-level modeling identified throughout the reviewed literature.

The final step of the framework is a scoring function over all pairs of candidate nodes in the encoded graph, determining the likelihood that the relationship between two nodes is in fact a compliance relationship, similar to the link-prediction task in knowledge graph embedding [14], [15] and the anomaly-scoring formulations used in graph-based financial forensics [15], [20]. Relationships with scores above a calibrated threshold are returned for compliance review, transforming a manual document reconciliation task into a ranked discovery problem.

COMPARATIVE SYNTHESIS OF METHODOLOGICAL COMPONENTS

Table 1 summarizes the principal technique families synthesized into the RiskGraph-NLP framework, situating each within the broader literature reviewed in Section 2.

Table 1. Core Technique Families Synthesized into the RiskGraph-NLP Framework

Technique Family	Representative Work	Core Mechanism	Primary Relevance to Compliance Discovery
Contextual Language Models	[4], [3]	Bidirectional self-attention over token sequences	Encodes clause-level semantics and long-range lexical dependencies
Domain-Adapted Language Models	[5], [7]	Continued pre-training on legal/financial corpora	Improves sensitivity to domain-specific obligation and risk language
Document Layout Models	[8]	Joint text-and-layout embedding	Preserves structural cues (tables, headers) lost in plain-text pipelines
Named Entity Recognition	[9]	Sequence labeling / span classification	Identifies parties, obligations, and regulatory anchors as graph nodes
Document-Level Relation Extraction	[1], [2]	Graph-based reasoning over multi-sentence spans	Surfaces relations that cross sentence and paragraph boundaries
Graph Neural Networks	[19], [10], [21]	Neighborhood aggregation / attention over graph structure	Propagates evidence across the compliance graph to expose indirect links
Knowledge Graph Embedding	[14], [12]	Relational embedding of heterogeneous triples	Supports link prediction for undocumented compliance dependencies

In addition to the qualitative synthesis, the most representative quantitative results from the reviewed literature are summarized in Table 2. These figures are not obtained from a common set of experiments, cannot be directly compared row to row, and are intended only to represent the empirical strength established for each methodological component prior to their combination within the proposed framework.

Table 2. Representative Reported Outcomes for Constituent Methods

Model	Benchmark Context	Reported Outcome	Source
BERT (base/large)	GLUE natural-language-understanding suite	GLUE score of approximately 80.5, exceeding prior state-of-the-art contextual embedding methods	[4]
Transformer (base)	WMT 2014 English-to-German / English-to-French translation	BLEU scores of approximately 28.4 and 41.8, respectively, at reduced training cost	[3]
GCN	Cora / Citeseer / Pubmed citation networks	Node classification accuracy of approximately 81.5%, 70.3%, and 79.0%	[19]
GAT	Cora / Citeseer / Pubmed citation networks	Node classification accuracy of approximately 83.0%, 72.5%, and 79.0%	[10]
LEGAL-BERT	Legal text classification and sequence-labeling corpora	Consistent gains over generic BERT on domain-specific legal benchmarks	[5]
Graph-based fraud model	Bitcoin transaction network (Elliptic dataset)	Improved illicit-transaction detection relative to non-graph baselines	[17]

The two key graph neural architectures employed by the proposed encoder, graph convolutional networks and graph attention networks, deliver similar yet distinguishable results on standard citation-network datasets, with the attention-based model showing slight improvements on two of the three datasets [19], [10].

While the density and semantics of the edges in citation networks differ considerably from those in compliance graphs, this comparison suggests that attention-based neighborhood weighting is valuable whenever edges carry different evidentiary values a property a compliance graph is expected to exhibit, given its mixture of strong statutory citations and weaker structural co-occurrence edges.

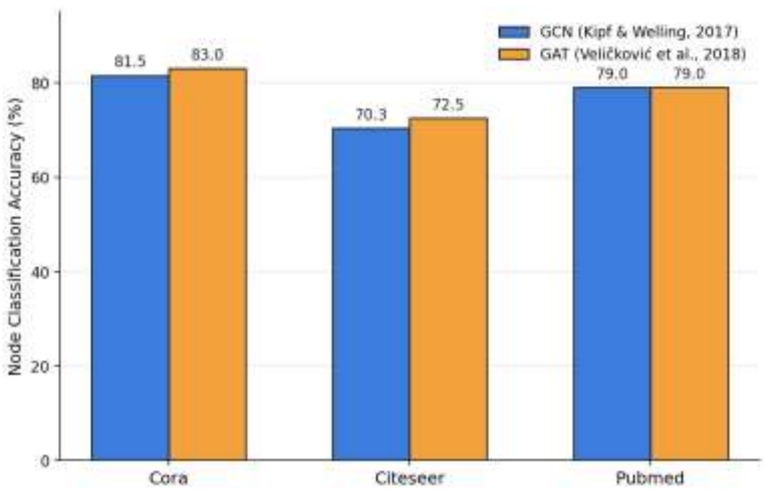

Figure 3. Reported node classification accuracy of GCN and GAT architectures on citation-network benchmarks, synthesized from [19] [10]

A further consideration is inductive capacity. A compliance graph encoder should generalize to nodes that were unavailable during training, since contract amendments and regulatory disclosures are added continuously over time. This scenario is unfavorable for transductive convolutional formulations that rely on the entire graph during training, but favors inductive frameworks in which aggregation functions are learned and can transfer to unseen neighborhoods [21]. This consideration favors an inductive, attention-based encoder for production deployment, even where transductive formulations report marginally better benchmark performance on a static citation graph.

RESULTS AND DISCUSSION

4.1. Structural Advantages over Sequential Baselines

The central idea drawn from the reviewed literature is that hidden compliance is, by definition, a multi-hop phenomenon: it relies on chains of reference that link a clause to an obligation, an obligation to a provision, and a provision to a party, sometimes through an intermediate cross-reference. Sequential language models, even with long attention windows, handle this only implicitly and imperfectly, since the relevant elements can be far apart in the sequence and unrelated to the specific dependency being encoded [3], [4]. Evidence that explicit graph construction over mentions substantially improves the extraction of exactly this kind of cross-sentence relation [1], [2] provides direct support for the structural premise of RiskGraph-NLP.

4.2. Domain Adaptation and Data Quality

The literature also shows that the advantage of graph-structural reasoning depends on the quality of node representations. In legal and financial text, domain-adapted pre-training has been shown to be particularly important because of differences in vocabulary and phrasing relative to general corpora [5], [7]. A graph built on poor entity disambiguation or inaccurate obligation-span classification, arising from named entity recognition errors, would propagate that noise through all downstream graph operations, since graph propagation aggregates information across a node's entire neighborhood [9], [19]. Thus, the reliability of relationship discovery at downstream stages is fundamentally limited by the quality of the upstream extraction stages.

4*.3. Ethical and Regulatory Considerations*

Discovering compliance relationships raises considerations that differ materially from ordinary information extraction. A false negative a legitimate obligation-to-provision dependency that is not surfaced carries substantially different consequences here than a comparable error in general-purpose text classification. Prior research on automated compliance checking of data processing agreements against data protection regulation has treated such tools as decision-support aids rather than as complete compliance checks, with automated reports informing, rather than replacing, human reviewers [18]. This precedent suggests that RiskGraph-NLP outputs should likewise be treated as prioritized leads for expert evaluation rather than compliance determinations, particularly since graph-based relationship and anomaly scores as used in financial forensics applications are designed to rank candidates rather than assert a binary judgment.

4*.4. Limitations of the Present Synthesis*

The framework proposed in this paper is presented at the conceptual and architectural level and has been compiled from methodological precedents rather than validated through original experiments on enterprise compliance corpora. The quantitative results reported in Table 2 are drawn from diverse benchmarks citation networks, translation corpora, and financial transaction graphs none of which directly models the structure of enterprise compliance documents. The comparison in Section 4 should therefore be read as a hypothesis to be tested with the proposed architecture, not as direct evidence of its performance, and as a motivation for empirical validation on annotated compliance corpora.

CONCLUSION

We have brought together methodological advances in contextual language modeling, document-level relation extraction, graph neural networks, and knowledge graph reasoning into a conceptual framework, RiskGraph-NLP, capable of discovering hidden compliance relationships scattered across enterprise documents. The synthesis suggests that graph-structured representations provide a principled complement to sequential language encoders, since compliance dependencies often span multiple clauses, sections, and documents that are not reliably captured by local attention windows. Layout-aware encoding, domain-specific pre-training, and document-level relation extraction were identified as necessary upstream components whose accuracy determines the reliability of downstream graph reasoning, while attention-based and inductive graph architectures were identified as better suited to the heterogeneous, continuously evolving nature of enterprise compliance graphs. Positioning such a system as a tool for human compliance review, rather than as an autonomous determination system, was found to be consistent with precedent established for automated regulatory verification. Future work should focus on building annotated compliance-graph benchmarks, empirically comparing convolutional, attentional, and inductive encoders under realistic enterprise document distributions, and systematically analyzing the trade-off between hidden-relation recall and the reviewer workload induced by false positives.

REFERENCES

[1] F. Christopoulou, M. Miwa, and S. Ananiadou, "Connecting the Dots: Document-Level Neural Relation Extraction with Edge-Oriented Graphs," in *Proceedings of EMNLP-IJCNLP 2019*, 2019, pp. 4925–4936. doi: 10.18653/v1/D19-1498.

[2] S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou, "Inter-Sentence Relation Extraction with Document-Level Graph Convolutional Neural Network," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4309–4316. doi: 10.18653/v1/P19-1423.

[3] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008. doi: 10.48550/arXiv.1706.03762.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2898–2904. doi: 10.18653/v1/2020.findings-emnlp.261.

[6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 4–24, 2021, doi: 10.1109/TNNLS.2020.2978386.

[7] Y. Yang, M. C. S. Uy, and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications," 2020. doi: 10.48550/arXiv.2006.08097.

[8] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document

Image Understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, 2020, pp. 1192–1200. doi: 10.1145/3394486.3403172.

[9] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, pp. 50–70, 2022, doi: 10.1109/TKDE.2020.2981314.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018. doi: 10.48550/arXiv.1710.10903.

[11] Z. Zhang, P. Cui, and W. Zhu, "Deep Learning on Graphs: A Survey," *IEEE Trans. Knowl. Data Eng.*, pp. 249–270, 2022, doi: 10.1109/TKDE.2020.2981333.

[12] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," in *The Semantic Web: ESWC 2018*, 2018, pp. 593–607. doi: 10.1007/978-3-319-93417-4_38.

[13] X. Wang *et al.*, "Heterogeneous Graph Attention Network," in *The World Wide Web Conference*, 2019, pp. 2022–2032. doi: 10.1145/3308558.3313562.

[14] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications," *IEEE Trans. Knowl. Data Eng.*, pp. 2724–2743, 2017, doi: 10.1109/TKDE.2017.2754499.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.

[16] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 7370–7377. doi: 10.1609/aaai.v33i01.33017370.

[17] M. Weber *et al.*, "Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics," 2019. doi: 10.48550/arXiv.1908.02591.

[18] O. Amaral Cejas, M. I. Azeem, S. Abualhaija, and L. C. Briand, "NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR," *IEEE Trans. Softw. Eng.*, pp. 4282–4303, 2023, doi: 10.1109/TSE.2023.3288901.

[19] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, 2017. doi: 10.48550/arXiv.1609.02907.

[20] R. Van Belle, C. Van Damme, H. Tytgat, and J. De Weerdt, "Inductive Graph Representation Learning for Fraud Detection," *Expert Syst. Appl.*, 2022, doi: 10.1016/j.eswa.2021.116463.

[21] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 1024–1034. doi: 10.48550/arXiv.1706.02216.