Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Review Article

Adversarially Robust and Explainable Graph Neural Networks for Fraud Detection in Dynamic Blockchain Networks: A Systematic Solution-Based Literature Review

Oluwaseun Adeniyi Ojerinde [1,*], Ramatu Abubakar [1], Benjamin Ilunuamie Alenoghena [2], and Kehinde Hussein Lawal [2]

[1] Department of Computer Science, Federal University of Technology, Minna, Niger State 920101, Nigeria; e-mail : o.ojerinde@futminna.edu.ng; ramatu.abubakar@st.futminna.edu.ng

[2] Department of Software Engineering, Federal University of Technology, Minna, Niger State 920101, Nigeria; e-mail : benjamin.alen@futminna.edu.ng; kehinde.lawal@futminna.edu.ng

* Corresponding Author : Oluwaseun Adeniyi Ojerinde

Abstract: Blockchain technology has revolutionized digital financial systems through decentralized, transparent, and secure transactions. However, the increasing sophistication of blockchain systems has coincided with the emergence of complex illicit activities, including ransomware transactions, Ponzi schemes, phishing, and money laundering. Traditional fraud-detection techniques may be inadequate for blockchain transaction networks because these networks are complex, dynamic, and highly interconnected. Graph Neural Networks (GNNs) have emerged as a promising approach because of their ability to capture relationships within graph structures and their evolving behavioural patterns. This paper presents a systematic literature review of adversarially robust and explainable fraud-detection solutions based on GNNs for dynamic blockchain networks. The review was conducted in accordance with the PRISMA guidelines and critically examined studies published between 2018 and 2026. The study provides an integrated synthesis of graph representation learning, dynamic graph learning, adversarial robustness, explainable artificial intelligence (XAI), trustworthy AI, benchmark datasets, evaluation strategies, and regulatory compliance within the context of blockchain fraud detection. The review indicates that recent advances in Graph Convolutional Networks, Graph Attention Networks, GraphSAGE, Graph Transformers, and Temporal Graph Neural Networks have demonstrated promising improvements in fraud-detection performance across various experimental settings. However, challenges remain in achieving scalability, robustness against adversarial attacks, explainability, standardized evaluation, and real-time deployment. The review identifies major research gaps in blockchain fraud detection and examines the strengths, limitations, applicability, and maturity of existing techniques. Finally, based on the synthesized findings, a research roadmap is proposed for developing blockchain fraud-detection systems with improved explainability, scalability, robustness, and trustworthiness using GNNs. The findings provide actionable insights and recommendations for researchers and practitioners developing blockchain security solutions.

Keywords: Adversarial Defence; Adversarial Robustness; Blockchain Fraud Detection; Dynamic Blockchain Networks; Explainable Artificial Intelligence; Graph Neural Networks; Temporal GNN; Trustworthy AI.

Received: July, 24th 2026
Revised: June, 31st 2026
Accepted: September, 1st 2026
Published: September, 16th 2026
Curr. Ver.: September, 16th 2026

1. Introduction

1.1 Background and Problem Context

Blockchain technology has developed from its initial application in cryptocurrency into a broader infrastructure for decentralised financial services, digital assets, smart contracts, and other distributed applications. However, the transparent and decentralised characteristics of

blockchain networks do not eliminate fraudulent and illicit activities [1]. Cryptocurrency ecosystems continue to experience illicit transactions, scams, money laundering, ransomware-related payments, and other forms of financial abuse. As blockchain transaction volumes increase, fraud detection has consequently become an important research problem requiring methods capable of identifying suspicious behavior within complex and highly interconnected transaction environments [2].

Traditional machine-learning approaches have been extensively applied to fraud detection, but many of these approaches represent transactions primarily through manually engineered or independently processed features. Such representations may be inadequate when fraudulent behavior depends on relationships among multiple entities rather than on the characteristics of individual transactions [3]. In blockchain environments, relationships among addresses, transactions, contracts, and other entities can provide important contextual information for identifying suspicious behaviour. Consequently, graph-based learning has become increasingly relevant because blockchain activity can naturally be represented through interconnected entities and their relationships.

Graph Neural Networks (GNNs) provide a framework for learning representations directly from graph-structured data by aggregating information from connected nodes and their neighborhoods. Their ability to exploit relational information has made them increasingly attractive for financial fraud detection, where fraudulent activities may involve groups of interacting entities rather than isolated transactions. A systematic review by [4], for example, identified GNNs as an increasingly important approach to financial fraud detection and organized the literature according to graph structures, GNN architectures, learning strategies, and application contexts. For blockchain fraud detection, however, predictive capability alone is insufficient. Blockchain transaction networks continuously evolve as new transactions and entities appear and existing relationships change. Fraud-detection models must therefore account for temporal changes in graph structure and transaction behavior. Recent research on dynamic temporal GNNs similarly identifies the ability to model evolving relationships as a major advantage of GNN-based anomaly detection, while also highlighting persistent challenges involving scalability, limited real-world datasets, and interpretability [5], [6].

In addition to temporal evolution, blockchain fraud-detection models face security and trust challenges. Because GNNs rely on graph structures and node or edge attributes, malicious actors may potentially manipulate the information used by the model to influence its predictions. This creates a need for adversarially robust learning, in which models are evaluated and designed to maintain reliable performance when exposed to malicious perturbations. At the same time, the increasing complexity of graph-based deep-learning models creates an explainability requirement [7]. Fraud investigators and other stakeholders may need to understand why a transaction or entity has been classified as suspicious before relying on the prediction for investigation or decision-making.

Taken together, these developments indicate that blockchain fraud detection is no longer concerned solely with predictive accuracy. The literature increasingly addresses three interconnected requirements: dynamic adaptability, to account for evolving transaction relationships and behavior; adversarial robustness, to maintain reliable performance under malicious manipulation; and explainability, to make model decisions interpretable to investigators and other stakeholders. These three dimensions provide the central analytical focus of this review.

1.2 Research Gap and Need for Robust, Explainable Dynamic Fraud Detection

The development of GNN-based blockchain fraud detection has addressed important limitations of conventional feature-based approaches, but the existing literature remains fragmented across different methodological and technical objectives. One important dimension concerns the dynamic nature of blockchain transaction networks. Transactions are continuously generated, new entities enter the network, relationships between entities change, and behavioral patterns associated with fraud may evolve over time [8]. A model trained on a static representation may therefore become less effective when the underlying distribution of transaction behavior changes. Recent systematic evidence on dynamic temporal GNNs indicates that temporal modelling has become an important research direction for anomaly detection, while scalability and the limited availability of realistic dynamic datasets continue to constrain practical deployment [6], [9].

A second-dimension concerns adversarial robustness. Fraud detection models operate in environments in which malicious actors may have incentives to evade detection. For graph-

based systems, this threat may involve manipulation of node attributes, graph connections, transaction relationships, or training data. Consequently, strong predictive performance under conventional test conditions does not necessarily indicate reliable performance under adversarial conditions. Robust fraud detection therefore requires evaluation beyond standard accuracy-based comparisons, including assessment of how model predictions change when the underlying graph or its attributes are deliberately manipulated [5], [10].

The third-dimension concerns explainability. GNNs can capture complex relational patterns, but their predictions may be difficult for human users to interpret. This is particularly important in fraud detection because model predictions may trigger investigation, account restrictions, or other consequential actions. Recent systematic research on XAI for fraud detection found that explanation methods are widely used, but explanation quality is frequently not evaluated directly; instead, predictive performance is often used as an indirect indicator of explanation effectiveness [11].

Importantly, these dimensions are interconnected rather than independent. Dynamic graph learning concerns how fraud-related relationships evolve; adversarial robustness concerns whether models remain dependable when those relationships or their associated features are deliberately manipulated; and explainability concerns how the resulting decisions can be interpreted [12], [13]. Considering only one of these dimensions may therefore provide an incomplete assessment of the practical readiness of GNN-based fraud-detection approaches.

Existing reviews provide important foundations, but they generally examine these dimensions from different perspectives. References [4] and [14] systematically reviewed GNN-based financial fraud detection, with emphasis on architectures, graph representations, and learning strategies. [6], [15] focused specifically on dynamic temporal GNNs for cybersecurity anomaly detection and identified scalability, dataset availability, and interpretability as continuing challenges. [11] in turn, examined methodological challenges in XAI for fraud detection and highlighted the need for direct evaluation of explanation quality. Collectively, these reviews demonstrate the growing maturity of the individual research areas while also indicating that the available evidence remains distributed across separate methodological concerns.

The gap addressed by the present review is therefore not simply the absence of another review of GNN-based fraud detection. Rather, an evidence and synthesis gap remains in understanding how dynamic blockchain fraud detection, adversarial robustness, and explainability are addressed collectively across the existing literature. This review addresses this gap by examining these dimensions together and synthesizing the available evidence in terms of the strengths and limitations of existing approaches, their trade-offs, research maturity, evaluation practices, deployment challenges, and unresolved evidence gaps.

1.3. Novelty and Contributions of This Review

The present review builds on previous research while adopting a more focused analytical perspective. Previous reviews have established the importance of GNNs for financial fraud detection, dynamic GNNs for evolving graph-based anomaly detection, and XAI for improving transparency in fraud-detection systems [4], [6], [10], [16]. However, these research streams have largely developed around different methodological objectives, leaving the evidence fragmented across temporal adaptability, adversarial robustness, and explainability.

The contribution of this review therefore extends beyond identifying the topical coverage of previous studies. It provides an integrated synthesis of these research dimensions by comparing solution categories, identifying trade-offs, assessing research maturity, and linking evidence gaps to future research needs. The review makes the following contributions:

- It develops a solution-oriented taxonomy that organizes the literature into foundational GNN-based fraud detection, dynamic graph learning, adversarially robust GNNs, explainable GNNs, and supporting datasets and evaluation approaches.
- It provides a cross-study comparative synthesis of major approaches in terms of their strengths, limitations, scalability, computational requirements, robustness, explainability, and suitability for different blockchain fraud-detection scenarios.
- It integrates dynamic learning, adversarial robustness, and explainability as the three core analytical dimensions of the review, enabling comparison beyond conventional predictive-performance measures. These dimensions are also aligned with the primary themes defined in the review methodology.

- It analyses trade-offs among predictive effectiveness, temporal modelling, adversarial robustness, explainability, scalability, and computational cost, while assessing the relative maturity of different solution categories.
- It identifies persistent evidence gaps involving realistic dynamic datasets, adversarial evaluation, standardized explainability assessment, scalability, real-time processing, reproducibility, and practical deployment.
- It develops a gap-driven research roadmap that connects the identified limitations and evidence gaps with future directions for adaptive, robust, explainable, scalable, and trustworthy blockchain fraud-detection systems.

This structure allows the review to move beyond a descriptive catalogue of GNN approaches towards a synthesis of how the different solution dimensions relate to one another and where the existing evidence remains insufficient. It also provides the analytical basis for the research roadmap developed later in the review.

Table 1. Comparison of Previous Reviews and the Scope of the Present Review

Ref.	Review scope	Dynamic blockchain fraud detection	Adversarial robustness	Explainability / XAI	Integrated analysis of the three themes	Research gaps and roadmap	Main limitation
[17]	AI-based fraud detection in DeFi	Partial	✗	Partial	✗	✓	DeFi-focused; lacks integrated analysis of dynamic GNNs, robustness, and explainability.
[4]	GNNs for financial fraud detection	Partial	✗	✗	✗	✓	Strong GNN-focused review, but limited blockchain-specific analysis of robustness and explainability.
[10]	Trustworthy GNNs	✗	✓	✓	✗	✓	Broad trustworthy-GNN perspective without a specific focus on dynamic blockchain fraud detection.
[8]	GNNs for financial fraud detection	Partial	Partial	Partial	✗	✓	Financial-fraud focus without an integrated treatment of the three core themes.
Ours	Adversarial robust and explainable GNNs for dynamic blockchain fraud detection	✓	✓	✓	✓	✓	Integrated cross-study synthesis, trade-off and maturity analysis, evidence-gap analysis, and a gap-driven research roadmap.

Table 1 compares the scope of the present review with previous reviews addressing GNN-based fraud detection, trustworthy GNNs, and AI-based DeFi fraud detection. Previous reviews provide important foundations, but they generally focus on individual research dimensions rather than their integration within dynamic blockchain fraud detection. The distinction of the present review therefore lies primarily in its analytical scope. It considers dynamic blockchain fraud detection, adversarial robustness, and explainability as interconnected requirements and synthesizes the literature according to their trade-offs, research maturity, evidence gaps, and practical implications. This integrated perspective provides the basis for the research roadmap developed later in the review.

2. Review Methodology

2.1. Systematic Literature Review Protocol

This study adopted a systematic literature review (SLR) methodology to identify, evaluate, classify, and synthesize research on Graph Neural Network (GNN)-based approaches for fraud detection in blockchain and cryptocurrency transaction networks. The review

followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure a transparent and systematic study-selection process [18], [19].

The review covers literature published between 2018 and June 2026. The starting year was selected to capture the development and application of graph-based deep-learning methods to fraud, anomaly, and transaction-network problems. The search covered both blockchain-specific fraud-detection studies and closely related methodological research on dynamic graph learning, adversarial robustness, and explainability where relevant to blockchain transaction analysis. The review process comprised identification, screening, eligibility assessment, quality assessment, and final inclusion. Potentially relevant records were retrieved from the selected bibliographic databases and additional sources using predefined search terms and Boolean operators. Duplicate records were identified based on bibliographic information, including publication title, authorship, publication year, and DOI where available. Ambiguous matches were manually checked before removal.

Following deduplication and predefined pre-screening exclusions, the remaining records underwent title and abstract screening. Potentially relevant studies were then assessed in full text against the predefined inclusion and exclusion criteria. Eligibility was determined based on relevance to blockchain or cryptocurrency fraud, illicit activity, anomaly detection, or transaction analysis; the use or methodological relevance of graph-based learning; and the availability of sufficient methodological information for assessment. Where multiple exclusion criteria applied, a single primary reason was assigned according to the predefined decision hierarchy. Studies meeting the eligibility requirements underwent quality assessment, and those satisfying the predefined quality requirements and providing sufficient evidence for qualitative synthesis were retained in the final review. The overall selection procedure follows the PRISMA 2020 reporting structure [18], [19].

The review was structured around three interconnected analytical themes: (i) dynamic blockchain fraud detection, (ii) adversarial robustness, and (iii) explainability. GNN architectures, graph representation learning, datasets, and optimisation techniques were treated as supporting methodological dimensions. This structure maintains alignment among the review methodology, research questions, literature synthesis, and research roadmap.

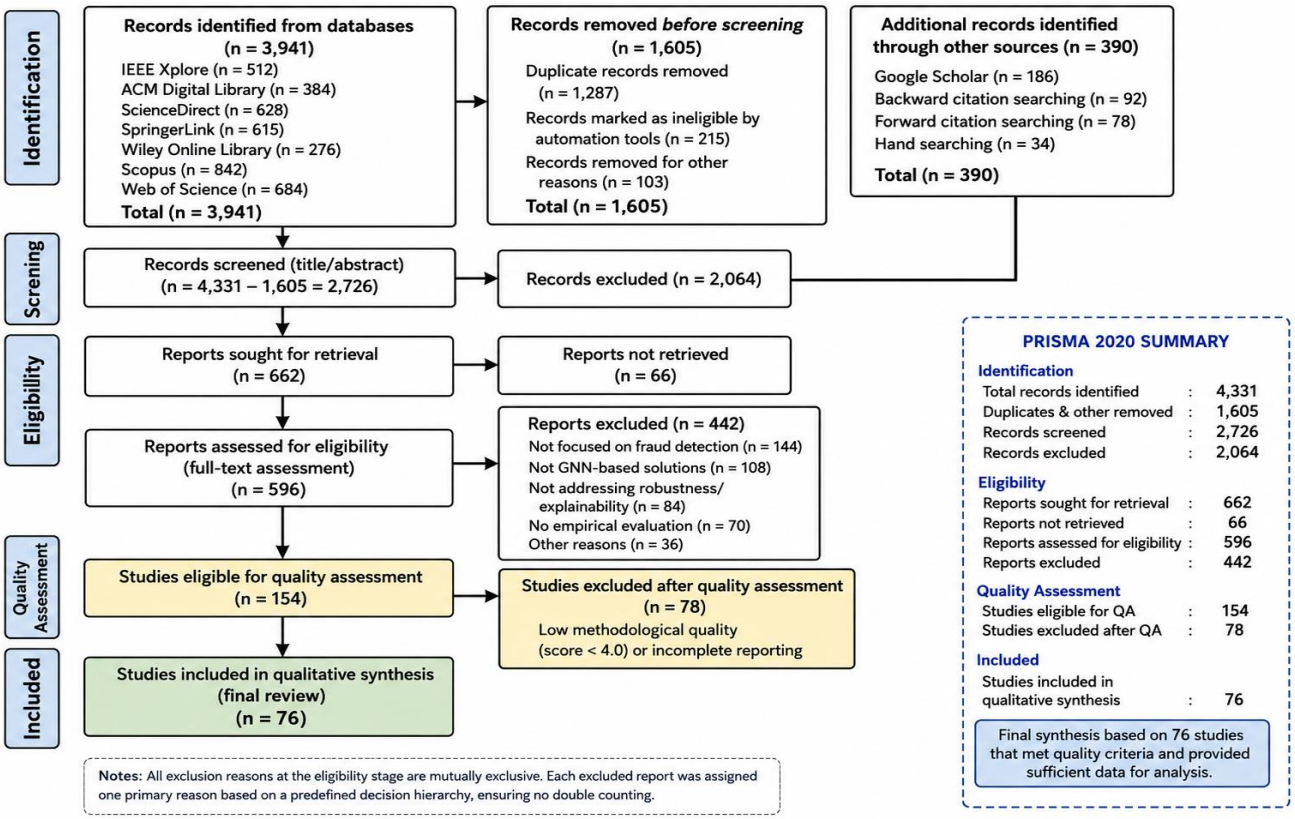

Figure 1. PRISMA 2020 flow diagram of the literature identification, screening, eligibility assessment, and study inclusion process.

Figure 1 summarizes the study-selection process from identification to final inclusion. At the eligibility stage, exclusion reasons were assigned using mutually exclusive primary categories so that each excluded report was counted once. The resulting counts provide a traceable progression from the 4,331 identified records to the 76 studies retained for qualitative synthesis [18], [19].

2.2. Research Questions

The research questions were formulated to reflect the central objective of the review while maintaining a clear distinction between the three core themes and supporting methodological issues. The following questions guided the search, screening, extraction, and synthesis processes:

- RQ1: What GNN-based approaches have been proposed for fraud detection in blockchain and cryptocurrency transaction networks?
- RQ2: How are dynamic and temporal graph-learning techniques used to model evolving transaction relationships and changing fraud patterns?
- RQ3: What adversarial threats and defence mechanisms have been investigated to improve the robustness of GNN-based fraud detection?
- RQ4: What explainability techniques have been applied to GNN-based fraud detection, and what limitations affect their practical usefulness?
- RQ5: What trade-offs emerge across existing approaches in terms of detection effectiveness, robustness, explainability, scalability, computational requirements, and suitability for dynamic blockchain environments?
- RQ6: What evidence gaps and future research directions emerge towards integrated, adaptive, adversarially robust, and explainable GNN-based blockchain fraud-detection systems?

2.3. Search Strategy

A structured search strategy was developed to identify studies relevant to GNN-based blockchain fraud detection and the principal analytical themes of the review, namely dynamic and temporal graph learning, adversarial robustness, and explainability. Searches were conducted across IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, Scopus, and Web of Science. Additional records were identified through Google Scholar, backward and forward citation searching, and targeted hand searching. The literature search commenced in December 2025 and was updated through June 2026 to capture recent developments in the field.

The search covered publications published between 2018 and June 2026. The initial database search was structured around three core conceptual groups: (i) blockchain and cryptocurrency, (ii) fraud and illicit activity, and (iii) graph neural networks and graph learning. These groups were used to retrieve studies broadly relevant to GNN-based fraud detection in blockchain and cryptocurrency transaction networks. The general search structure was:

("blockchain" OR "cryptocurrency" OR "Bitcoin" OR "Ethereum") AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR "GNN" OR "graph learning")

The retrieved studies were subsequently screened and analyzed with particular attention to a fourth conceptual group covering (iv) dynamic and temporal graph learning, adversarial robustness, and explainability. These aspects were used to classify and organize relevant studies during screening, data extraction, and thematic synthesis rather than being treated as separate Boolean blocks in the initial database search. This approach enabled the review to capture studies addressing the principal analytical themes while maintaining a focused and reproducible initial search strategy.

The general search structure was translated into database-specific formulations to accommodate differences in field operators, Boolean syntax, phrase searching, and indexing conventions. The three core search groups were maintained across the selected databases, while the fourth conceptual group informed the subsequent screening, data extraction, and thematic synthesis. This approach ensured consistency in the initial retrieval process while allowing the review to examine dynamic and temporal learning, adversarial robustness, and explainability as interconnected analytical dimensions [4], [6], [10].

2.3.1. Database-Specific Search Procedures

Searches were conducted from December 2025 to June 2026 across the selected bibliographic databases. The general search structure was adapted to the indexing and query syntax supported by each database while preserving the same conceptual components: (i) blockchain and cryptocurrency, (ii) fraud and illicit activity, (iii) graph learning, and (iv) dynamic learning, adversarial robustness, and explainability. Database-specific adaptations included field restrictions, Boolean operators, phrase-search conventions, and truncation or wildcard syntax where supported. These adaptations were intended to maintain conceptual equivalence across databases without changing the substantive inclusion scope.

Table 1. This is a table. Tables should be placed in the main text near to the first time they are cited.

Database	Search period	Database-specific search query	Adaptation applied
IEEE Xplore	Dec. 2025–June 2026	("blockchain" OR "cryptocurrency" OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning")	Boolean operators and exact-phrase searching adapted to IEEE Xplore search syntax.
ACM Digital Library	Dec. 2025–June 2026	(("blockchain" OR cryptocurrency OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning"))	Parentheses and Boolean operators adapted to ACM Digital Library.
ScienceDirect	Dec. 2025–June 2026	("blockchain" OR cryptocurrency OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning")	Phrase searching and Boolean combinations adapted to ScienceDirect.
SpringerLink	Dec. 2025–June 2026	blockchain OR cryptocurrency OR Bitcoin OR Ethereum AND "fraud detection" OR "illicit transaction" OR "financial fraud" AND "graph neural network" OR GNN OR "graph learning"	Search terms were adapted to SpringerLink's broader keyword-search interface.
Wiley Online Library	Dec. 2025–June 2026	("blockchain" OR "cryptocurrency" OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning")	Boolean and phrase-search syntax adapted to Wiley's search interface.
Scopus	Dec. 2025–June 2026	TITLE-ABS-KEY((blockchain OR cryptocurrency OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning"))	TITLE-ABS-KEY field restriction used to search titles, abstracts, and keywords.
Web of Science	Dec. 2025–June 2026	TS=((blockchain OR cryptocurrency OR Bitcoin OR Ethereum) AND ("fraud detection" OR "illicit transaction" OR "financial fraud" OR "money laundering") AND ("graph neural network" OR GNN OR "graph learning"))	TS topic-field syntax used to search title, abstract, author keywords, and related indexed terms.

Figure 2 presents the dominant terminology identified across the retrieved literature. The prominence of terms related to blockchain, fraud detection, graph neural networks, transaction analysis, anomaly detection, robustness, and explainability provides an exploratory overview of the research vocabulary. This analysis is used to contextualize the evidence base and does not determine study eligibility or research effectiveness.

Figure 3 complements the keyword-frequency analysis by illustrating relationships among frequently occurring terms in the retrieved literature. The resulting patterns provide an exploratory view of how blockchain fraud detection, graph learning, adversarial robustness, temporal modelling, and explainability are connected within the research corpus. These patterns support the thematic organization of the subsequent synthesis but do not replace the formal study-selection and quality-assessment procedures.

Figure 2. Keyword distribution of the retrieved literature on GNN-based blockchain fraud detection.

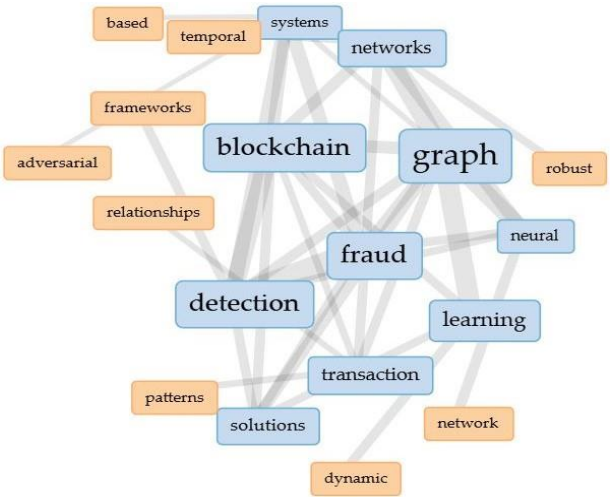

Figure 3. Textual relationship analysis of the retrieved literature based on the review search corpus.

2.4. Inclusion and Exclusion Criteria

Studies were considered eligible when they satisfied the predefined inclusion criteria and did not meet any of the exclusion criteria presented below.

Studies were included if they:

- were published between 2018 and June 2026;
- were written in English;
- were peer-reviewed journal articles, conference papers, or other scholarly publications meeting the review's evidence requirements;
- addressed blockchain, cryptocurrency, or transaction-network fraud, illicit activity, or anomaly detection;
- investigated GNNs, graph learning, graph representation learning, or closely related graph-based methods; or
- investigated dynamic graph learning, adversarial robustness, or explainability with direct methodological relevance to GNN-based fraud or anomaly detection in transaction networks; and
- provided sufficient methodological information to support qualitative comparison.

Studies were excluded if they:

- were published outside the 2018–June 2026 period;
- were not written in English;
- were editorials, news articles, opinion pieces, patents, theses, or non-scholarly web content;
- focused exclusively on conventional fraud-detection methods without graph-based relevance;

- addressed blockchain applications unrelated to fraud, illicit activity, anomaly detection, or transaction analysis;
- lacked sufficient methodological information for meaningful assessment; or
- were duplicate records identified across databases.

At the full-text eligibility stage, exclusion reasons were assigned using mutually exclusive primary categories to ensure numerical traceability. When a report met more than one exclusion criterion, the first applicable reason in the predefined decision hierarchy was recorded as the primary reason for exclusion. Each excluded report was therefore counted only once, allowing the exclusion categories reported in Figure 1 to be traced directly to the total number of reports excluded at the eligibility stage. The reported exclusion categories sum to the 392 reports excluded at this stage.

Eligibility decisions were made sequentially at the title-and-abstract and full-text stages. During title and abstract screening, records were retained when the available information indicated potential relevance to the review scope. Potentially eligible records were then assessed in full text against all predefined criteria. A study was retained when it addressed blockchain, cryptocurrency, or transaction-network fraud, illicit activity, or anomaly detection; investigated GNNs, graph learning, or a methodologically relevant graph-based approach; and provided sufficient methodological information for qualitative comparison. Reports that failed one or more mandatory eligibility requirements were excluded. When multiple exclusion criteria applied, only one primary exclusion reason was recorded according to the predefined decision hierarchy. This procedure ensured consistent eligibility decisions and prevented double counting of excluded reports.

2.5. Quality Assessment Criteria

The methodological quality of the included studies was evaluated using a structured assessment framework developed for this review. The assessment examined whether each study clearly reported its research objectives, data source, graph construction procedure, model architecture, experimental design, evaluation measures, comparison baselines, and study limitations. For studies addressing adversarial robustness, the assessment additionally considered the clarity of the threat model and robustness evaluation. For studies addressing explainability, the assessment considered whether the explanation method and its evaluation procedure were adequately reported. These criteria allowed methodological quality to be assessed in relation to the specific contribution and design of each study rather than assuming identical experimental requirements across all studies.

The scores were summed to obtain a maximum quality score of 9.0 per study. The resulting scores were used to assess reporting completeness and methodological quality and to support the selection of studies for the final evidence synthesis. Table 3 illustrates the application of the quality-assessment criteria to a sample of studies. The assessment covers general methodological reporting as well as the specific dimensions of adversarial robustness and explainability where applicable.

The quality levels were classified according to the cumulative scores obtained from the nine assessment criteria. As summarised in Table 4, studies scoring 7.0 or above were classified as High Quality, those scoring from 4.0 to below 7.0 as Medium Quality, and those scoring below 4.0 as Low Quality. This classification provided a consistent basis for determining the methodological strength of the studies considered for the final synthesis.

The nine assessment criteria used in the quality-assessment procedure are presented in Table 5. These criteria cover the clarity of the research design, adequacy of methodological reporting, dataset and model description, evaluation procedures, and discussion of limitations. QA5 and QA6 capture the review-specific dimensions of adversarial robustness and explainability, respectively, where these dimensions are relevant to the scope of the assessed study.

Following the full-text eligibility assessment, 154 studies satisfied the predefined inclusion criteria and were subjected to quality assessment using the nine criteria (QA1–QA9). Each criterion was scored as Yes (1), Partially (0.5), or No (0), producing a maximum cumulative score of 9.0 per study. The quality assessment served as a secondary selection stage to evaluate methodological rigour, reporting completeness, and the suitability of eligible studies for inclusion in the final evidence synthesis. The resulting distribution of studies across the three quality levels and the corresponding selection decisions are presented in Table 6. Studies classified as High Quality, with cumulative scores of 7.0 or above, were retained for the final

qualitative synthesis. Studies classified as Medium or Low Quality were excluded after the quality-assessment stage.

Table 3. Illustrative Examples of Quality Assessment Application

No.	Study	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9	Total Score	Quality Level
1	[20]	✓	✓	✓	✓	~	~	✓	✓	✓	7.5	High
2	[21]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High
3	[22]	✓	✓	✓	✓	~	~	✓	✓	~	6.5	High
4	[23]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High
5	[24]	✓	✓	✓	✓	✓	~	✓	✓	~	7.5	High
6	[25]	✓	✓	✓	~	~	~	✓	✓	~	5.5	Medium
7	[26]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High
8	[27]	✓	✓	✓	✓	✓	✓	✓	✓	✓	8.5	High
9	[25]	✓	✓	✓	✓	~	~	✓	✓	~	6.5	Medium
10	[12]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High
11	[28]	✓	✓	✓	✓	~	✓	✓	✓	~	7.5	High
12	[1]	✓	✓	✓	~	✓	~	✓	✓	~	6.5	Medium
13	[29]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High
14	[30]	✓	✓	✓	✓	~	✓	✓	✓	~	7.5	High
15	[31]	✓	✓	✓	✓	✓	✓	✓	✓	✓	9.0	High

Note: ✓ = Yes (1); ~ = Partially (0.5); ✗ = No (0).

Table 4. Quality-Level Classification.

Quality level	Score range
High Quality	≥ 7.0
Medium Quality	4.0 ≤ Score < 7.0
Low Quality	< 4.0

Table 5. Quality Assessment Criteria.

Code	Criterion
QA1	Clear objectives
QA2	Methodology adequately described
QA3	Dataset clearly identified
QA4	GNN model clearly described
QA5	Adversarial robustness addressed
QA6	Explainability addressed
QA7	Evaluation metrics reported
QA8	Results adequately validated
QA9	Limitations and future work discussed

Table 6. Quality Assessment Results and Final Study Selection

Quality level	Score range	Number of studies	Decision
High Quality	≥ 7.0	76	Included in final synthesis
Medium Quality	4.0–6.5	50	Excluded after quality assessment
Low Quality	< 4.0	28	Excluded after quality assessment
Total	—	154	—

As shown in Table 6, 76 of the 154 studies subjected to quality assessment met the High-Quality threshold and were retained for the final qualitative synthesis. The remaining 78 studies, comprising 50 Medium-Quality and 28 Low-Quality studies, were excluded after the quality-assessment stage. This quality-based selection constituted the final filtering stage of the review process and corresponds to the final stage of the PRISMA flow diagram.

2.6. Data Extraction and Synthesis

A structured data-extraction framework was developed to ensure consistency and comparability across the included studies. The extracted information covered publication year, research objective, blockchain or cryptocurrency platform, fraud or illicit-activity type, dataset, graph representation, GNN architecture, learning paradigm, temporal or dynamic mechanism, adversarial threat model, defence mechanism, explainability technique, evaluation metrics, baseline methods, scalability considerations, computational requirements, reported limitations, and future research directions [1], [2], [31]. This framework enabled the review to capture both the methodological characteristics of individual studies and the factors relevant to their practical applicability.

The extracted studies were subsequently organized into the major analytical categories of the review: GNN-based fraud detection, dynamic graph learning, adversarial robustness, and explainability [31]. Supporting dimensions, including graph representation, optimization, datasets, and evaluation metrics, were retained to provide methodological context and facilitate cross-study comparison but were not treated as independent central themes. This organization allowed related approaches to be examined within a common analytical structure while preserving differences in their methodological assumptions and application settings.

The synthesis employed a thematic and comparative approach rather than simply summarizing individual studies. Studies were compared in terms of their research objectives, methodological assumptions, strengths, limitations, computational requirements, scalability, robustness, explainability, and applicability to evolving blockchain environments. The analysis also examined recurring trade-offs across the literature, including predictive performance versus computational cost, model complexity versus interpretability, static versus adaptive learning, and robustness versus model utility [32]. These comparisons were used to identify areas of convergence, methodological differences, and limitations that may affect the practical deployment of GNN-based fraud detection systems.

Research maturity was further assessed by considering the extent to which proposed approaches had progressed beyond evaluation on benchmark datasets towards more realistic dynamic, adversarial, and operational conditions. This distinction is important because recent evidence indicates that dynamic GNN research continues to face challenges related to scalability, realistic datasets, and interpretability [6], [33], [34]. Similarly, broader research on trustworthy GNNs identifies robustness and explainability as important and interconnected dimensions of trustworthy graph learning [10]. Accordingly, the synthesis considered not only reported predictive performance but also the extent to which studies addressed the practical requirements of robustness, adaptability, interpretability, and scalability. The resulting cross-study synthesis is presented in Section 4, covering research trends, methodological trade-offs, strengths and limitations, research maturity, and evidence gaps. The identified gaps and limitations subsequently inform the future research roadmap presented in Section 5.

3. Solution-Based Literature Review

The literature on blockchain and financial fraud detection has progressively shifted from conventional machine-learning approaches towards graph-based representation learning, temporal graph learning, adversarially robust models, and explainable artificial intelligence (XAI) [35], [36]. This transition reflects the relational nature of blockchain transaction environments, in which fraudulent behaviour may emerge from interactions among multiple accounts, transactions, and other entities rather than from the attributes of an individual transaction alone. Recent systematic evidence shows that GNNs have become increasingly prominent in financial fraud detection because they can exploit these relational structures and support node-, edge-, and graph-level anomaly detection [4]. At the same time, broader research on GNNs identifies persistent challenges related to scalability, robustness, explainability, and deployment under realistic conditions [10], [37]. These challenges are particularly relevant to blockchain environments, where transaction graphs continuously evolve, fraudulent actors

may adapt their behavior, and detection decisions may need to withstand deliberate manipulation while remaining interpretable to investigators and other stakeholders.

The reviewed solutions are therefore organized into five analytical areas that operationalize the three core themes established in the review methodology: GNN-based fraud detection; graph representation and dynamic learning; adversarial robustness and defence; explainable and trustworthy GNNs; and benchmark datasets and evaluation approaches. The first four areas describe the principal solution mechanisms identified across the literature, whereas datasets and evaluation approaches provide the supporting basis for determining whether those solutions are reliable, scalable, robust, and applicable to evolving blockchain environments. This organization enables the review to move beyond an architecture-by-architecture description by examining what problem each solution addresses, what methodological advantages it provides, what limitations remain, and how the different solution dimensions interact in the development of trustworthy blockchain fraud-detection systems. The cross-study comparisons presented in the following subsections therefore focus not only on reported detection performance but also on adaptability, robustness, explainability, computational requirements, scalability, and practical suitability.

3.1. GNN-Based Solutions for Blockchain Fraud Detection

Graph Neural Networks provide a natural framework for modelling blockchain transaction networks because they learn representations by propagating information across connected entities. In a transaction graph, nodes may represent wallets, accounts, addresses, or other entities, while edges represent transactions or interactions, with relevant attributes incorporated as node or edge features. This representation allows fraud-detection models to consider both the characteristics of individual entities and the relational context in which transactions occur. Such relational modelling is particularly important when suspicious behavior is distributed across multiple connected entities rather than expressed through the attributes of a single transaction. The ability to exploit this graph structure is consequently one of the principal reasons GNNs have become increasingly important in financial fraud detection [4], [38].

The reviewed literature shows an architectural progression from relatively simple neighborhood aggregation towards increasingly adaptive, expressive, heterogeneous, and scalable graph-learning mechanisms. Early approaches commonly relied on Graph Convolutional Networks (GCNs), which aggregate information from neighboring nodes to construct node representations. GCNs provide a relatively straightforward foundation for exploiting local transaction relationships and have therefore served as an important baseline for subsequent graph-learning approaches [39], [40]. Their limitations, however, become more apparent as graph depth increases or when the underlying transaction network contains complex structural dependencies, since repeated neighborhood aggregation can introduce information loss or over-smoothing. Graph Attention Networks (GATs) address part of this limitation by assigning different attention weights to neighboring nodes, allowing the model to emphasize relationships that are more informative for a given prediction. This selective aggregation is particularly relevant to fraud detection because transaction relationships are unlikely to contribute equally to the identification of suspicious behavior. Recent financial fraud research has consequently continued to investigate attention-based mechanisms for representing complex transaction relationships [41].

GraphSAGE provides a different response to the scale of transaction networks by using neighborhood sampling rather than aggregating information from every neighbor. Its inductive formulation and sampling mechanism make it particularly relevant to large or evolving graphs in which exhaustive neighborhood aggregation may become computationally expensive. This provides an important scalability advantage for blockchain applications, although sampling introduces a corresponding trade-off: structurally or behaviorally informative neighbors may be omitted when they are not selected during aggregation [8], [42]. Graph Isomorphism Networks (GINs), in contrast, emphasize the expressive capability of graph aggregation and are designed to distinguish structural patterns that may be difficult for less expressive architectures to separate. More recent fraud-detection studies have also explored combinations of graph-representation mechanisms to capture complementary structural information, indicating a broader movement towards architectures tailored to the characteristics of specific fraud environments rather than relying on a single generic GNN architecture [26], [43].

The increasing complexity of blockchain transaction ecosystems has also motivated heterogeneous and long-range graph-learning approaches. Heterogeneous GNNs are particularly relevant when a transaction environment contains multiple entity and relationship types, such as wallets, transactions, smart contracts, exchanges, and decentralized applications. By explicitly preserving these semantic distinctions, heterogeneous representations can retain information that may be lost when the entire ecosystem is reduced to a homogeneous graph [20]. Evidence from financial fraud research demonstrates the value of combining heterogeneous connectivity with graph learning, particularly in transaction networks characterized by complex and imbalanced relationships [22]. Graph Transformer architectures provide another extension by employing attention mechanisms to model broader dependencies within graph structures. Their capacity to capture relationships beyond immediate neighborhoods may be useful when fraudulent behavior involves multiple intermediate entities; however, the additional representational capacity also increases computational requirements and therefore introduces a significant scalability consideration for large transaction graphs [44].

Hybrid GNN approaches extend this architectural progression by combining complementary learning mechanisms rather than relying on a single aggregation strategy. These approaches may integrate graph learning with attention, sequential learning, neighborhood sampling, or adaptive aggregation to capture different aspects of transaction structure and behavior. Recent financial fraud studies, for example, have combined GNNs with self-attention and neighborhood sampling to model irregular transactions and customer relationships while improving computational efficiency [42]. Other approaches have incorporated adaptive aggregation to address contextual and camouflage-related fraud behaviors [41]. Taken together, these studies indicate that architectural development is increasingly driven by specific limitations of blockchain and financial transaction graphs, including scale, heterogeneity, irregularity, structural complexity, and the ability of fraudulent entities to conceal their relationships within legitimate activity.

Table 7. Comparative Characteristics of GNN Architectures for Blockchain Fraud Detection

Architecture	Main capability	Strength	Main limitation	Scalability	Suitable scenario
GCN	Local structural aggregation	Simple and effective baseline	Over-smoothing and neighbourhood dependence	Moderate	Basic transaction-network detection
GAT	Attention-based aggregation	Emphasizes informative relationships	Higher computational cost	Moderate	Selective relationship modelling
GraphSAGE	Sampled neighborhood aggregation	Inductive and scalable	Sampling may lose informative structural information	High	Large and evolving graphs
GIN	Expressive structural learning	Strong structural discrimination	Higher model complexity	Moderate	Complex structural fraud patterns
Heterogeneous GNN	Multiple node and edge types	Preserves semantic relationships	More complex graph construction and modelling	Moderate	Multi-entity blockchain ecosystems
Graph Transformer	Long-range attention	Captures broader dependencies	High computational requirements	Moderate–Low	Multi-hop fraud relationships
Hybrid GNN	Combines complementary mechanisms	Flexible representation learning	Increased architectural complexity	Model-dependent	Complex fraud environments

Note: The comparative characteristics in Table 7 should be interpreted as a synthesis of the reviewed literature rather than as universal properties of every implementation within each architectural family. Actual scalability and performance depend on graph size, graph construction, implementation, optimization strategy, hardware resources, and evaluation protocol [42], [44], [45].

The comparative evidence therefore does not support a universal ranking of GNN architectures for blockchain fraud detection. Instead, the suitability of an architecture depends on the characteristics of the transaction graph and the requirements of the detection task. GCNs provide a relatively simple and interpretable baseline for local structural aggregation; GATs offer more selective relationship modelling; GraphSAGE provides an attractive option for large graphs through neighborhood sampling; GINs emphasize structural discrimination; heterogeneous GNNs preserve semantic distinctions among multiple entity and relationship types; Graph Transformers offer greater capacity for modelling broader dependencies; and hybrid architectures provide flexibility by combining complementary mechanisms. These advantages are accompanied by corresponding limitations involving over-smoothing, sampling-

induced information loss, graph-construction complexity, model complexity, or computational cost. Recent blockchain anomaly-detection experiments comparing GCN, GAT, GraphSAGE, and specialized GNN models further demonstrate that performance differences vary according to the dataset and evaluation metrics [45]. The evidence therefore favors a task- and context-dependent selection of GNN architectures, rather than treating any individual architecture as universally superior.

The main comparative characteristics of the reviewed GNN architectures are summarized in Table 7. The table is intended to complement the cross-study discussion by relating each architectural family to its principal capability, strength, limitation, scalability profile, and potential suitability for blockchain fraud-detection scenarios. Figure 4 should provide a high-level taxonomy of the architectural progression discussed in this subsection, linking each GNN family to the principal problem it addresses and its associated trade-off. In particular, the figure should distinguish the progression from local structural aggregation to attention-based, sampled, expressive, heterogeneous, long-range, and hybrid representation learning. Its purpose is to visually support the comparative synthesis rather than merely illustrate the internal structure of individual architectures.

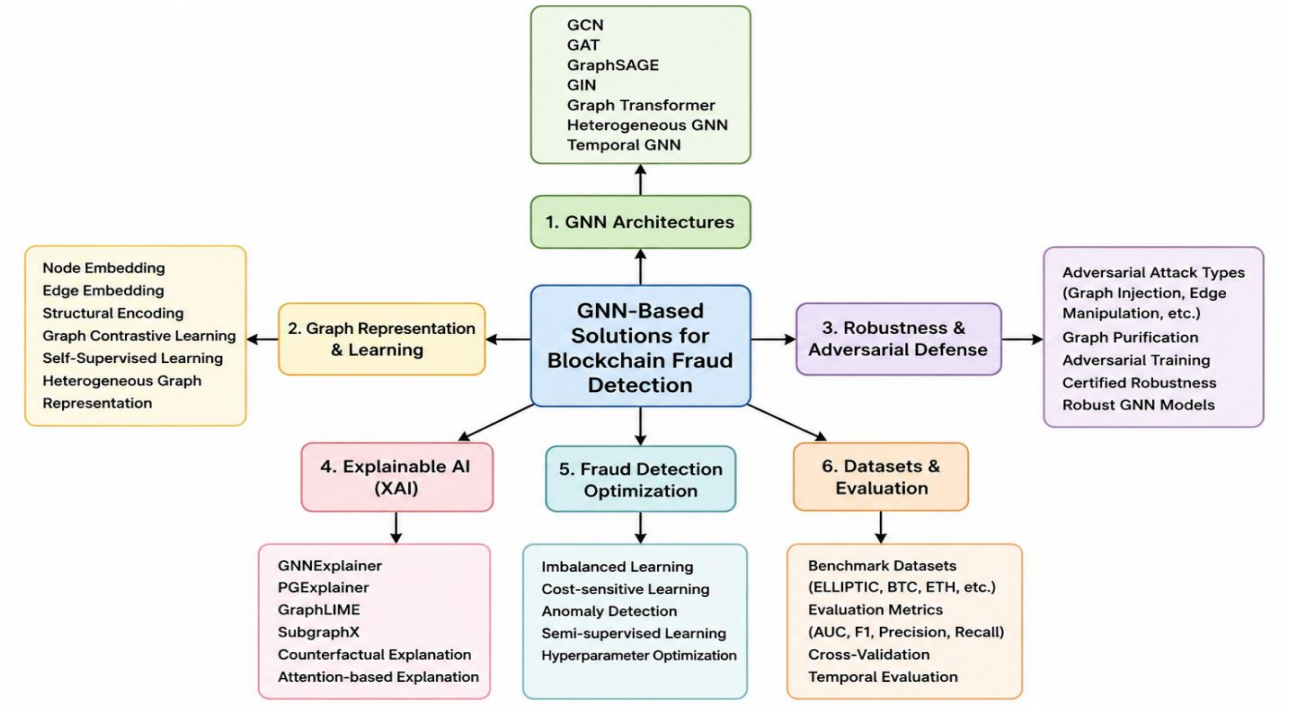

Figure 4. Taxonomy of GNN-based solutions for blockchain fraud detection

3.2. Graph Representation and Dynamic Blockchain Learning

The effectiveness of GNN-based fraud detection depends substantially on how the underlying transaction network is represented. Graph representation learning transforms transaction relationships and entity characteristics into representations that can support classification, anomaly detection, or link prediction. In financial fraud detection, accounts or addresses can be represented as nodes and transactions as edges, with transaction amounts, timestamps, and other attributes incorporated as node or edge features [4], [34]. The choice of graph representation therefore directly influences the structural and behavioral information available to the detection model.

Graph construction may be homogeneous or heterogeneous. Homogeneous representations simplify the learning process by using common node and edge types, whereas heterogeneous graphs preserve multiple entity and relationship types and can provide richer semantic information. This additional expressiveness, however, introduces greater graph-construction and modelling complexity. Recent financial fraud research demonstrates the potential of combining heterogeneous connectivity, neighborhood sampling, and attention mechanisms to address the structural complexity and class imbalance of transaction networks [22]. Thus, graph representation is not simply a preprocessing step but an important component of the

detection solution, particularly when fraudulent behavior involves interactions among different types of entities.

The challenge becomes more pronounced as blockchain transaction networks continuously evolve. New addresses and transactions are generated over time, existing relationships change, and fraudulent entities may adapt their behavior in response to detection mechanisms. A static graph representation can consequently become outdated as the transaction environment changes. Dynamic graph learning addresses this limitation by incorporating temporal information into graph representation and prediction [28]. As illustrated in Figure 5, the evolution from relatively small transaction networks to increasingly complex and continuously changing graphs creates interconnected challenges involving continuous data generation, scalability, temporal dependencies, concept drift, real-time detection, adversarial adaptation, and interpretability.

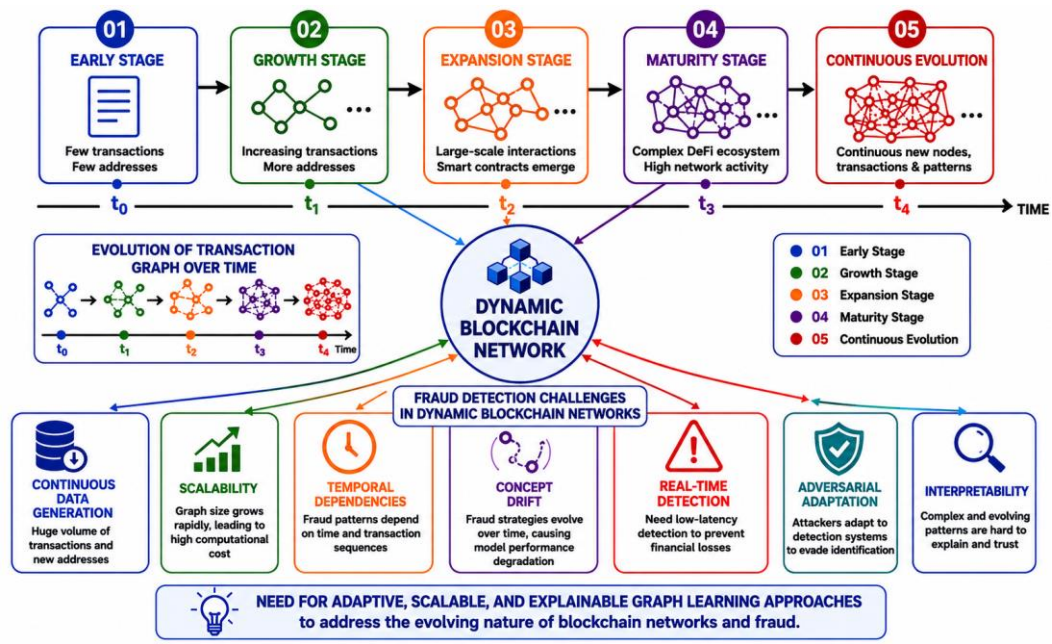

Figure 5. Dynamic graph representation and temporal learning for evolving blockchain transaction networks.

The reviewed studies provide evidence of the transition from static graph representations towards temporal and dynamic learning. Study [7] investigated temporal graph convolutional networks combined with LSTM-based learning for anti-money-laundering detection using Bitcoin data. Similarly, [1] proposed feature-gated temporal graph learning for illicit Bitcoin transaction detection, while [28] developed CoSemiGNN using dynamic graph neural networks for blockchain fraud detection. Collectively, these studies indicate that temporal modelling enables fraud-detection systems to account for changes in transaction relationships and behavioral patterns over time. However, the reviewed evidence also suggests that dynamic approaches remain less mature than conventional static GNN models, particularly with respect to robustness evaluation and explainability.

Recent financial fraud research further supports the integration of temporal and structural information. DyHDGE, for example, was proposed as a dynamic heterogeneous transaction graph embedding approach that jointly models temporal and structural information in financial transaction networks. This direction is particularly relevant to blockchain fraud detection because suspicious activities may become apparent only when sequences of related transactions are examined over time rather than when individual transactions are considered independently. Nevertheless, dynamic modelling introduces additional computational and memory requirements, as continuously updating graph representations can become increasingly expensive as transaction volume and graph complexity grow.

More importantly, the use of temporal information does not automatically guarantee adaptation to evolving fraud behavior. A model may incorporate timestamps while continuing to rely on assumptions derived from historical data distributions. Changes in fraud strategies

may therefore introduce concept drift that cannot be addressed simply by adding a temporal component. Effective dynamic blockchain fraud detection consequently requires mechanisms that can update graph representations while responding to changes in behavioral patterns, maintaining computational efficiency, and supporting timely detection.

Overall, the literature indicates that graph representation and dynamic learning address complementary limitations of conventional fraud detection. Heterogeneous representations provide richer information about diverse entities and relationships, while temporal mechanisms capture how those relationships evolve. However, increasing representational and temporal complexity also increases computational demands and does not, by itself, resolve adversarial manipulation or explainability. The evidence therefore points towards a need for adaptive graph-learning solutions that can remain informative and computationally feasible while operating on evolving transaction networks. This requirement provides a direct basis for considering adversarial robustness as a further dimension of blockchain fraud-detection solutions.

3.3. Adversarial Robustness and Defence

The deployment of GNNs in security-sensitive applications introduces an important concern regarding adversarial manipulation. Because GNN predictions depend on graph topology, node attributes, relationships, and training data, deliberate changes to these components can affect the learned representations and ultimately alter fraud-detection decisions. Research on adversarial GNNs distinguishes several attack settings, including graph modification and graph injection, as well as poisoning, evasion, and backdoor scenarios [46]. These vulnerabilities are particularly relevant to fraud detection because malicious actors may deliberately manipulate transaction relationships or observable characteristics to reduce the likelihood of detection.

Adversarial attacks can occur at different stages of the learning pipeline. During training, poisoning attacks seek to influence the model by modifying training data or graph structures. During inference, evasion attacks attempt to alter the input graph or fraudulent behavior so that malicious entities are classified as legitimate. Graph injection attacks introduce new nodes or relationships, whereas graph modification attacks alter existing structures [47]. These attack types affect different components or stages of the detection process and therefore require defence mechanisms that are aligned with the assumed threat model.

The corresponding defence literature has developed several solution categories. Training-stage strategies include adversarial training, robust aggregation, and modifications to message-passing mechanisms, whereas inference-stage approaches can include graph purification, anomaly filtering, and other mechanisms designed to identify or suppress suspicious graph components. These approaches indicate that GNN robustness cannot be addressed solely at the architectural level because both graph structure and message-passing processes influence how perturbations propagate through the model [48], [49]. The effectiveness of a defence therefore depends on the characteristics of the attack as well as the assumptions under which the model is trained and evaluated.

Adversarial training is particularly relevant because exposing a model to perturbed examples during learning can encourage representations that are less sensitive to specific perturbations. However, its effectiveness remains dependent on the attack model used during training. A model optimized against one perturbation strategy may not necessarily remain robust against a different or adaptive attack. This limitation highlights the importance of explicitly defining the threat model and evaluating robustness against attack conditions that are representative of the intended deployment environment [48]. Consequently, reporting performance against a single predefined perturbation provides only limited evidence of general adversarial resilience.

For blockchain fraud detection, the principal attack types identified in the reviewed literature include node injection, edge manipulation, feature perturbation, evasion, and poisoning attacks, as illustrated in Figure 6. Node injection introduces artificial entities into the transaction graph, while edge manipulation adds, deletes, or modifies transactional relationships. Feature perturbation alters attributes such as transaction amounts, frequencies, or timestamps to disguise malicious behavior. Evasion attacks target the inference process by modifying fraudulent behavior or transactions to bypass an established detection model, whereas poisoning attacks compromise the learning process through malicious or mislabelled training data [50]. These attack types can produce reduced detection accuracy, increased false

negatives, successful evasion of fraudulent activities, loss of trust in detection systems, financial losses, and regulatory or compliance risks.

The reviewed evidence also reveals a limitation in the alignment between adversarial GNN research and blockchain fraud detection. Study [5] examined the robustness of GNN defence mechanisms, while [48] investigated adversarial training for GNNs and highlighted important limitations and challenges associated with existing defence strategies. More recent studies, including [47] and [43], proposed robust GNN architectures and defence mechanisms against graph adversarial attacks. However, these contributions are predominantly situated in general graph-learning environments rather than specifically in dynamic blockchain transaction networks. The resulting gap is therefore not the absence of adversarial GNN research, but the limited empirical integration and evaluation of these robustness solutions under the dynamic, heterogeneous, and fraud-specific characteristics of blockchain transaction graphs.

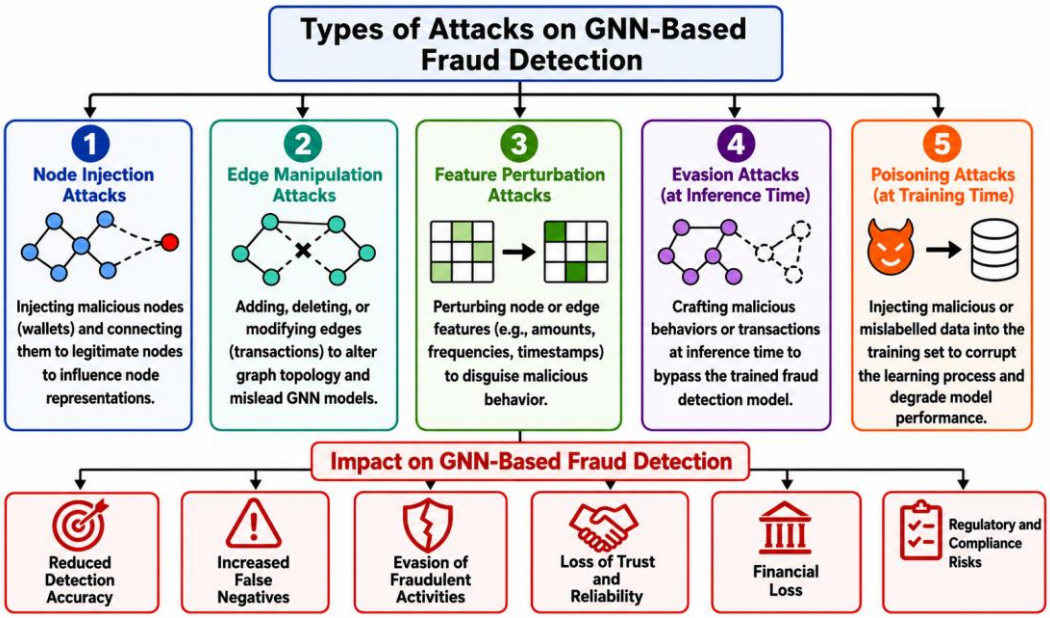

Figure 6. Adversarial attack surfaces and defence strategies in GNN-based blockchain fraud detection.

The literature further indicates a trade-off between robustness and computational efficiency. Additional defence mechanisms can increase training time, memory requirements, and model complexity. A defence that improves resistance to perturbations but introduces substantial computational overhead may therefore be unsuitable for high-volume or near-real-time transaction monitoring. Adversarial robustness should consequently be assessed together with clean-data performance, adversarial performance, computational cost, and detection latency. More broadly, trustworthy-GNN research treats robustness as one component of a wider trustworthiness framework that also includes explainability, privacy, and fairness [10].

Taken together, the evidence suggests that adversarial robustness should be treated as an integrated design and evaluation dimension rather than as an additional defence module applied after model development. The central challenge for blockchain fraud detection is to achieve resilience across relevant attack types while retaining detection effectiveness, computational feasibility, and interpretability. This consideration also connects adversarial robustness with the subsequent discussion of explainable and trustworthy GNNs, where the ability to understand and validate model decisions becomes essential for practical deployment.

3.4. Explainable and Trustworthy GNNs

Although predictive performance is important, fraud-detection systems also need to explain why a transaction or entity has been classified as suspicious. This requirement has increased interest in explainable GNNs and broader trustworthy-AI approaches. Explainable AI aims to make model decisions more understandable by identifying the nodes, edges, or

substructures that contribute to a model's prediction, thereby supporting transparency, interpretation, and informed decision-making [76]. More broadly, trustworthy GNN research considers explainability alongside other dimensions such as robustness, privacy, and fairness to improve the reliability and practical acceptance of GNN-based systems [51].

Several approaches have been developed to explain GNN predictions, including GNNExplainer, PGExplainer, SubgraphX, gradient-based approaches, and surrogate-based methods. These methods can identify influential features, nodes, edges, or subgraphs that contribute to a model's prediction. A systematic evaluation framework developed by [52] demonstrates that different GNN explanation methods can be benchmarked using datasets with known explanation structures, highlighting the importance of evaluating explanation quality rather than presenting explanations without validation.

The relevance of explainability is particularly strong in financial and blockchain fraud detection, where investigators may need to understand the relationships underlying a suspicious prediction. An explanation may, for example, identify influential transaction edges, neighboring addresses, or behavioral features that contribute to a fraud score. Such information can assist human analysts in prioritizing investigations and assessing whether a model's prediction is consistent with observable transaction behavior.

However, explainability also introduces methodological challenges. Different explanation techniques may produce different explanations for the same prediction, while an explanation that appears visually plausible may not necessarily reflect the actual reasoning process of the model. [52], [53] therefore emphasize the difficulty of evaluating GNN explanations and the importance of systematic benchmarks and evaluation procedures. Explainability should consequently be assessed not only in terms of whether an explanation can be generated, but also in terms of whether the explanation provides reliable and meaningful evidence for the corresponding prediction.

For dynamic blockchain graphs, explanations must additionally account for temporal behavior. A static explanation may identify an important transaction relationship without showing how that relationship evolved over time. A transaction or entity that appears influential at one point may become less relevant as new transactions and relationships emerge. Consequently, explainable blockchain fraud detectors should consider temporal explanations that identify not only influential graph components but also relevant changes in transaction behavior over time.

The reviewed literature indicates that explainability remains an emerging but insufficiently integrated research direction. While [53] focuses on evaluating explainability for GNNs and [52] reviews explainable GNN approaches, studies such as [16] and [54] demonstrate the increasing application of explainable AI to transaction-fraud analysis. Nevertheless, the reviewed evidence suggests that explainability is frequently addressed separately from graph robustness and dynamic graph learning. Few studies integrate dynamic modelling, adversarial robustness, and explanation mechanisms within a unified blockchain fraud-detection framework. This indicates that the main challenge is not simply developing additional explanation techniques, but integrating explanation into models that must simultaneously operate under evolving and potentially adversarial transaction conditions.

Therefore, the literature suggests that XAI should be integrated into the design and evaluation of blockchain fraud-detection systems rather than being added only after the predictive model has been developed. Such integration would allow explainability to be considered alongside predictive performance, temporal adaptability, adversarial robustness, and computational requirements, providing a more comprehensive basis for developing trustworthy GNN-based fraud-detection solutions.

3.5 Benchmark Datasets and Evaluation Approaches

The conclusions drawn from GNN-based fraud-detection research depend heavily on the datasets, graph-construction strategies, and evaluation protocols used. Financial fraud studies employ a variety of public and domain-specific datasets, including transaction networks and heterogeneous financial graphs. However, differences in dataset composition, graph construction, class imbalance, feature availability, and label quality make direct comparison between studies difficult [4]. Consequently, reported performance needs to be interpreted in relation to the characteristics of the dataset and the experimental conditions under which each model was evaluated.

Class imbalance is particularly important because fraudulent entities or transactions may constitute only a small proportion of the overall network. Recent GNN-based fraud research has therefore explored imbalance-aware approaches, including neighborhood sampling and specialized learning mechanisms designed to improve the representation of minority fraud classes [22]. Other work has demonstrated that fraud graphs may also exhibit disassortative structures and edge-level imbalance, challenging assumptions made by [25]. These characteristics indicate that the effectiveness of a fraud-detection model depends not only on its architecture but also on how the underlying graph structure and class distribution are represented during learning.

Evaluation should therefore not rely exclusively on accuracy. Precision, recall, F1-score, ROC-AUC, PR-AUC, false-positive rate, and other appropriate measures provide a more informative assessment of fraud-detection performance, particularly when the classes are highly imbalanced. The choice of evaluation metrics should also reflect the practical consequences of classification errors, since missed fraudulent activities and excessive false positives may have different operational implications.

Temporal evaluation is also important for blockchain fraud detection because transaction networks evolve continuously. Randomly splitting transactions into training and testing sets may allow information associated with future transaction behavior to influence model development, producing an evaluation that does not accurately represent deployment conditions. Time-aware evaluation, in which earlier transactions are used for training and later transactions are reserved for testing, provides a more realistic approximation of model performance in an evolving blockchain environment. This approach is particularly important for determining whether a model can generalize to behavioral patterns that were not available during training.

Adversarially robust models require additional evaluation under both clean and perturbed conditions. Performance on clean data provides evidence of conventional detection capability, whereas evaluation under adversarial perturbations indicates how effectively the model can withstand deliberate manipulation. Similarly, explainable models should be evaluated using measures of explanation fidelity, stability, and consistency rather than relying only on visual examples. Research on GNN explainability has emphasized the need for benchmark datasets and systematic metrics for this purpose [52].

Computational efficiency and scalability should also be considered when evaluating blockchain fraud-detection solutions. A model may achieve strong predictive performance on a relatively small benchmark dataset while becoming computationally impractical when applied to larger and continuously evolving transaction networks. Evaluation should therefore consider computational requirements and scalability alongside predictive performance, particularly for applications requiring timely or near-real-time detection.

The literature consequently points towards multidimensional evaluation, in which predictive performance, temporal generalization, adversarial robustness, explainability, computational efficiency, and scalability are considered together. This perspective provides a more meaningful basis for comparing GNN-based fraud-detection solutions because strong performance under one experimental setting does not necessarily imply that a model will maintain its effectiveness when applied to a larger, evolving, imbalanced, or adversarial blockchain environment. Evaluation should therefore reflect the conditions under which the proposed solution is expected to operate rather than relying solely on conventional benchmark performance.

4. Cross-Study Discussion and Findings

The synthesis of the reviewed literature indicates that blockchain fraud detection has evolved from relatively static, feature-oriented approaches towards graph-based, temporal, adversarially aware, and explainable learning systems. The studies reviewed in Section 3 demonstrate substantial progress in exploiting transaction relationships through GNNs, modelling evolving transaction behavior through dynamic graph learning, improving resilience against graph manipulation, and providing explanations for model predictions. However, these capabilities remain unevenly developed and are rarely evaluated together within a single framework. Recent reviews of financial fraud detection using GNNs similarly identify substantial progress in graph-based fraud detection while noting continuing limitations in dynamic learning, deployment, evaluation, and real-world applicability [4], [54].

The principal finding of this review is therefore that the central challenge has shifted from developing GNNs that can detect fraudulent patterns to developing GNN-based systems that can maintain reliable detection as blockchain structures evolve, withstand adversarial manipulation, provide meaningful explanations, and remain computationally feasible at scale. This finding provides the basis for the comparative synthesis presented below. Table 7 summarizes the distribution of key methodological and evaluation characteristics identified across the 76 included studies. The categories are non-mutually exclusive because a single study may address multiple methodological characteristics, such as temporal learning and adversarial robustness.

Table 7. Distribution of key methodological and evaluation characteristics across the 76 studies.

Methodological or evaluation characteristic	Number of studies	Percentage (%)
GCN-based models	8	10.5%
GAT-based models	2	2.6%
GraphSAGE-based models	1	1.3%
GIN-based models	1	1.3%
Graph Transformer-based models	1	1.3%
Heterogeneous/Hybrid GNNs	5	6.6%
Temporal/Dynamic GNNs	15	19.7%
Adversarial defence mechanisms	9	11.8%
Adversarial training	6	7.9%
Explainability/XAI methods	6	7.9%
Temporal evaluation strategies	11	14.5%
Concept-drift/adaptive learning	6	7.9%

Note: Categories are non-mutually exclusive; therefore, a study may be counted in more than one category.

4.1. Evolution of Research Trends

The chronological distribution of the reviewed studies suggests a shift in research emphasis. Earlier work, such as [3], primarily investigated static graph convolutional approaches for identifying illicit Bitcoin transactions. Subsequent studies expanded graph modelling towards temporal and evolving structures, including [7], [12], and more recent dynamic blockchain-focused approaches such as [28]. In parallel, robustness research has developed primarily within the broader GNN security literature, including [5], [48] and [47]. Explainability research has similarly expanded through studies such as [53] and [52]. Taken together, the reviewed evidence suggests that these research streams have developed largely in parallel, with limited integration within a single adversarially robust and explainable framework for dynamic blockchain fraud detection.

The transition towards GNN-based approaches reflects the recognition that fraudulent behavior is often relational. Fraudulent entities may not appear anomalous when examined independently but can exhibit suspicious patterns through their connections, transaction communities, or interaction structures. The systematic review by [4] found that GNN research in financial fraud detection has increasingly considered homogeneous, heterogeneous, static, temporal, and dynamic graphs, demonstrating the expansion of graph learning beyond simple static node classification. More recent work reinforces this progression. For example, HHLN-GNN incorporates heterogeneous and homogeneous connectivity, neighborhood sampling, and self-attention to address imbalance and hidden transaction relationships, while other recent models incorporate context encoding, adaptive aggregation, or metapath-based representations [22], [41].

The next transition is from static to dynamic graph learning. [55], [56] observe that much GNN research has historically focused on static graphs even though many real-world networks evolve continuously. Dynamic GNNs address this limitation by integrating temporal dependencies into graph representation learning, but scalability, heterogeneous information, and the availability of diverse dynamic datasets remain unresolved challenges [57]. The most recent direction is the movement towards trustworthy graph learning. Rather than optimizing

predictive performance alone, recent research increasingly considers robustness, explainability, privacy, and fairness as complementary dimensions of GNN reliability [58].

4.2. Comparative Findings Across Solution Categories

The cross-study analysis reveals that the reviewed solution categories address different but interconnected dimensions of blockchain fraud detection. Rather than identifying a single superior approach, the evidence suggests that GNN architectures, dynamic graph learning, adversarial robustness, explainability, and trustworthy AI address complementary requirements. Their effectiveness therefore needs to be interpreted in relation to the characteristics of the transaction network, the intended deployment environment, and the evaluation protocol.

4.2.1. GNN-Based Fraud Detection

The strongest advantage of GNN-based methods is their ability to exploit relationships between transaction entities. This provides a structural perspective that conventional transaction-level models may not capture adequately. Recent reviews report substantial use of GCN-, GAT-, heterogeneous-, and Transformer-based architectures in financial fraud detection [59], [60]. However, the evidence does not support a universal claim that one architecture consistently outperforms all others. Performance depends on graph construction, dataset characteristics, fraud type, class imbalance, feature representation, and evaluation protocol. Recent models have achieved improvements through heterogeneous connectivity, adaptive aggregation, or metapath-based learning, but these improvements are generally reported under specific experimental conditions rather than across a common benchmark [41], [61]. Architecture selection should therefore be considered in relation to the characteristics of the target transaction network rather than according to a general performance hierarchy.

4.2.2. Dynamic Graph Learning

Dynamic GNNs offer an important advantage because they can model changes in graph topology and temporal behavior. This is particularly relevant to blockchain environments, where transactions are continuously generated and relationships among entities evolve over time. Recent dynamic-GNN research identifies temporal modelling as an important development while also highlighting computational scalability and the limited availability of dynamic benchmark datasets as continuing challenges [62].

The evidence further indicates that temporal modelling should not automatically be interpreted as genuine adaptation. A model may incorporate timestamps while continuing to rely on assumptions derived from historical data distributions. Changes in fraud strategies may consequently introduce concept drift that cannot be addressed simply by adding a temporal component. Dynamic learning should therefore be considered together with time-aware evaluation, adaptive mechanisms, scalability, and detection latency.

4.2.3. Adversarial Robustness

Adversarial defence addresses the possibility that malicious actors deliberately manipulate graph structures or node information to influence model predictions. The broader GNN security literature demonstrates vulnerabilities to structural and feature perturbations and identifies defence mechanisms that can be implemented during training or or inference [43], [63], [64]. These findings are particularly relevant to blockchain fraud detection because the transaction graph itself can constitute an attack surface.

A major limitation is that robustness is strongly dependent on the assumed threat model. A defence that is effective against one perturbation strategy may provide limited protection against another. Consequently, adversarial performance should be interpreted in relation to the attack assumptions, perturbation budget, attack surface, and evaluation protocol. Adversarial training and other defence mechanisms may also introduce additional computational requirements, affecting their suitability for high-volume or near-real-time transaction monitoring [47], [65].

4.2.4. Explainability

Explainable GNNs address the interpretability requirements associated with fraud investigation and decision support. Explanation methods can identify influential features, nodes, edges, or subgraphs underlying a prediction. Recent surveys show rapid development in GNN explanation methods but also substantial differences in explanation objectives and

evaluation criteria [43], [48]. An explanation should therefore not be considered reliable simply because it produces a visually intuitive subgraph.

Explanation quality requires systematic assessment of properties such as correctness, robustness, usability, and computational complexity [63], [66]. This is particularly important in fraud detection, where explanations may support investigative decisions rather than merely illustrate model behavior. The dynamic nature of blockchain transaction networks introduces an additional challenge because relevant explanatory evidence may change as new transactions and relationships emerge. Explainability should therefore also be considered in relation to temporal changes and computational requirements.

4.2.5. Trustworthy AI

Trustworthy AI provides a broader perspective by connecting robustness and explainability with additional properties such as privacy and fairness. [10] demonstrates that these dimensions are interrelated and should be considered when GNNs are deployed in high-stakes applications. From this perspective, a blockchain fraud-detection model should not be evaluated solely according to whether it correctly classifies fraudulent transactions.

The reviewed evidence suggests that trustworthiness is a multidimensional property emerging from the interaction of several capabilities. Dynamic learning contributes adaptability, adversarial defence contributes resilience, and explainability contributes transparency, while privacy and fairness address additional deployment risks. However, these properties are frequently investigated separately using different datasets, assumptions, and evaluation criteria. Consequently, evidence remains limited on whether a single GNN-based system can simultaneously maintain temporal adaptability, adversarial resilience, explanatory reliability, and computational feasibility.

The comparative findings therefore indicate that the five solution dimensions play complementary roles: GNN architectures provide the structural learning foundation; dynamic graph learning addresses temporal evolution; adversarial robustness addresses deliberate manipulation; explainability supports interpretation and investigation; and trustworthy AI provides an overarching perspective for considering these capabilities together. The main challenge identified across the literature is consequently not the absence of individual techniques, but their limited empirical integration under realistic and evolving blockchain conditions.

4.3. Major Trade-Offs Identified in the Literature

The reviewed evidence reveals several cross-cutting trade-offs that influence the design and deployment of blockchain fraud-detection solutions. These trade-offs indicate that improving one aspect of a GNN-based system may introduce additional computational, interpretability, or robustness requirements.

Accuracy versus scalability. More expressive architectures can capture complex transaction relationships but may require greater computational resources. Sampling-based approaches can reduce computational requirements, although aggressive sampling may discard informative neighborhood structures. Recent fraud-detection models therefore increasingly incorporate sampling, adaptive aggregation, and structured graph representations to balance representation quality and efficiency [67], [68].

Temporal adaptability versus computational cost. Dynamic GNNs can capture evolving transaction behavior, but continuously updating graph representations requires additional memory and computation. Scalability therefore remains an important challenge in dynamic-GNN research [8]. The benefits of temporal modelling consequently need to be considered alongside the computational requirements of continuously evolving transaction networks.

Robustness versus efficiency. Adversarial training and other defence mechanisms can improve resilience but may require additional training procedures and computational resources. Moreover, robustness depends strongly on the assumed threat model [47], [65]. A defence effective against one perturbation strategy may therefore not provide equivalent protection against other attacks.

Explainability versus complexity. Explanation methods can improve transparency, but generating reliable explanations for large and dynamic graphs may introduce additional computational overhead. The growing diversity of explanation methods also creates a standardization challenge because different approaches may produce different explanations for the same prediction [16], [53].

Rich representation versus interpretability. Heterogeneous GNNs and Graph Transformers can represent increasingly complex transaction relationships, but greater model complexity may make model behavior more difficult to interpret. This creates an important design challenge for trustworthy blockchain fraud detection, where both representational capacity and understandable evidence are required. These trade-offs indicate that blockchain fraud detection should be approached as a multi-objective problem rather than as an accuracy-maximization task. Predictive performance, temporal adaptability, adversarial robustness, explainability, scalability, and computational efficiency therefore need to be considered jointly when assessing the practical value of a proposed solution.

4.4. Strengths and Limitations of the Existing Evidence

The reviewed literature demonstrates several important strengths. GNN-based methods provide a strong foundation for modelling relational transaction behavior, while heterogeneous and adaptive architectures increasingly capture complex relationships. Dynamic GNNs provide a basis for modelling evolving transaction networks, and adversarial-learning research has established the importance of explicitly considering security threats. XAI and trustworthy-GNN research further contribute mechanisms for improving transparency and reliability.

However, several limitations constrain the comparability and practical interpretation of the existing evidence. First, dataset heterogeneity remains a major challenge. Studies differ in preprocessing, graph construction, feature engineering, class distributions, train-test strategies, and evaluation metrics. The financial-fraud GNN literature similarly identifies differences in graph types and learning settings as an important challenge for synthesis [4].

Second, dynamic evaluation remains limited. Many studies rely on static or historical datasets even though blockchain transaction networks evolve continuously. The shortage of diverse dynamic graph datasets has also been identified as an important limitation in dynamic-GNN research [22]. This limits the evidence available for determining whether proposed models can generalize to future transaction structures and behavioral patterns.

Third, adversarial evaluation is not yet sufficiently standardized. Robustness results are often associated with particular attack assumptions and perturbation strategies, making direct comparison between defence mechanisms difficult. Similarly, explainability lacks a universally accepted evaluation protocol, with explanation correctness, robustness, usability, and computational cost remaining active research questions [69], [70].

Finally, integrated evaluation remains limited. Most studies focus on one or two objectives, such as fraud-detection performance, temporal modelling, robustness, or explainability, rather than evaluating these dimensions simultaneously. Consequently, there is limited empirical evidence on how these capabilities interact within a single blockchain fraud-detection system.

Despite substantial methodological progress, the evidence remains fragmented across datasets, evaluation protocols, and research objectives. The principal limitation is therefore not a lack of proposed techniques, but the limited availability of comparable and integrated evidence demonstrating their effectiveness under evolving, imbalanced, adversarial, and operational blockchain conditions.

4.4. Strengths and Limitations of the Existing Evidence

The reviewed literature demonstrates several important strengths in the development of GNN-based blockchain fraud detection. GNNs provide a suitable foundation for modelling relational transaction behavior, while heterogeneous and adaptive architectures offer greater capacity to represent complex interactions among entities. Dynamic GNNs further extend this capability by incorporating temporal information into evolving transaction networks, and research on adversarial learning has established the importance of considering graph structure as a potential attack surface. At the same time, explainable GNN and trustworthy-AI research has introduced mechanisms for improving transparency, reliability, and the practical interpretation of model decisions. These developments indicate that the field has progressed beyond conventional static fraud-detection approaches towards solutions that increasingly consider the structural, temporal, security, and interpretability requirements of blockchain transaction environments.

Despite this progress, several limitations constrain the comparability and practical interpretation of the existing evidence. First, substantial heterogeneity exists across datasets and

experimental settings. Studies differ in preprocessing procedures, graph-construction strategies, feature engineering, class distributions, train-test strategies, and evaluation metrics, making direct comparison of reported results difficult. The broader financial-fraud GNN literature similarly identifies differences in graph types and learning settings as important challenges for synthesizing findings across studies [4]. Consequently, a reported improvement in predictive performance cannot always be interpreted as evidence that one methodological approach is generally superior to another, because the observed results may depend strongly on the underlying dataset and experimental configuration.

Second, dynamic evaluation remains limited in blockchain-specific fraud detection. Although transaction networks evolve continuously, many studies continue to rely on static or historical datasets and evaluation settings that do not fully reproduce the temporal conditions of operational deployment. The limited availability of diverse dynamic graph datasets has also been identified as an important challenge in dynamic-GNN research [22]. This limitation reduces the available evidence for determining whether proposed models can generalize to future transaction structures, newly emerging behavioral patterns, and changes in fraud strategies. Temporal modelling therefore represents an important development, but its practical value cannot be fully established without evaluation protocols that reflect the evolving nature of blockchain transaction networks.

Third, adversarial evaluation is not yet sufficiently standardized across the reviewed literature. Robustness results are often reported under particular attack assumptions, perturbation budgets, or threat models, making direct comparison between defence mechanisms difficult. A defence that performs well against one attack strategy may not provide equivalent protection against another or against an adaptive attacker. Similar concerns apply to explainability, where different explanation methods may produce different results and where explanation correctness, robustness, usability, and computational cost remain active evaluation challenges [69], [70]. These limitations indicate that robustness and explainability should be assessed using more systematic and comparable procedures rather than being inferred from individual demonstrations.

Finally, integrated evaluation remains limited. Most studies concentrate on one or two objectives, such as fraud-detection performance, temporal modelling, adversarial robustness, or explainability, rather than evaluating these dimensions simultaneously within the same experimental setting. As a result, there is limited empirical evidence regarding how temporal adaptability, robustness, explainability, scalability, and predictive performance interact when implemented as components of a single blockchain fraud-detection solution. The principal limitation of the current evidence is therefore not a lack of proposed techniques, but the limited availability of comparable and integrated evidence demonstrating their effectiveness under evolving, imbalanced, adversarial, and operational blockchain conditions.

4.5. Research Maturity and Evidence Gaps

The cross-study evidence indicates that the research areas considered in this review have reached different levels of methodological maturity. GNN-based fraud detection can be considered relatively mature, as multiple GNN architectures and learning strategies have been extensively investigated and a substantial body of financial-fraud research has already been synthesized in the literature [71]. The development of GCN, GAT, GraphSAGE, GIN, heterogeneous GNNs, and more recent hybrid architectures demonstrates that the basic capability of applying graph learning to relational fraud-detection problems is well established. However, maturity at the architectural level does not necessarily imply that these approaches have been sufficiently validated under the more complex conditions encountered in evolving blockchain transaction networks.

Dynamic graph learning is developing but remains less mature in blockchain-specific applications. Although dynamic GNN research is expanding, important challenges involving scalability, temporal datasets, heterogeneous graphs, and adaptive learning remain unresolved [72]. In particular, the incorporation of temporal information does not by itself demonstrate that a model can adapt effectively to concept drift or changes in fraudulent behavior. The evidence therefore suggests that future progress requires not only more temporal architectures, but also stronger evaluation under realistic and continuously evolving transaction conditions.

Adversarial robustness remains comparatively less mature in blockchain fraud detection. The broader GNN security literature has established a range of attack and defence

mechanisms, including different forms of graph manipulation and adversarial training, but their systematic application to evolving blockchain fraud environments remains limited [73]. The main evidence gap is consequently related to the transfer and empirical validation of robustness mechanisms under blockchain-specific characteristics, rather than to the absence of adversarial GNN research in general. Similarly, explainability is methodologically more established but remains less mature in application-specific evaluation. Numerous explanation techniques have been developed, yet the literature continues to address challenges concerning how explanations should be evaluated, compared, and validated in practical settings [6], [74]. These limitations become particularly relevant when explanations must account for evolving transaction relationships rather than static graph structures.

The least mature area identified by the synthesis is the integration of these capabilities into a single trustworthy fraud-detection framework. The literature increasingly recognizes robustness, explainability, privacy, and fairness as interconnected dimensions of trustworthy GNNs [10], but these properties are rarely evaluated simultaneously with temporal adaptability and scalability in blockchain fraud-detection systems. The central evidence gap is therefore not the absence of individual solutions, but the limited empirical integration of solutions that can jointly address evolving transaction behavior, adversarial manipulation, explainability, computational efficiency, and trustworthy decision-making. This maturity assessment consequently supports the broader conclusion of the review that future research should move from isolated methodological improvements towards integrated and systematically evaluated solutions for dynamic blockchain fraud detection.

4.6. Overall Discussion and Key Findings

The overall synthesis leads to five principal findings. First, graph learning provides an appropriate foundation for blockchain fraud detection because fraudulent behavior is often relational and cannot always be adequately represented through independent transaction features. Second, dynamic graph learning is increasingly important for environments in which transaction structures and fraud behaviors evolve over time. However, temporal modelling needs to be accompanied by efficient updating and mechanisms capable of responding to distributional changes. Third, adversarial robustness is an important consideration because graph structure itself can become an attack surface. Robustness therefore needs to be evaluated under explicit and diverse threat models rather than inferred from clean-data performance.

Fourth, explainability is increasingly important for practical fraud investigation and trustworthy decision-making, particularly when model predictions need to be interpreted by human analysts. However, generating an explanation alone is insufficient, and explanation quality should be assessed using systematic evaluation procedures. Fifth, the most significant gap identified across the reviewed evidence is not the absence of individual techniques, but the limited integration among them. Current studies provide evidence for GNN-based detection, dynamic learning, adversarial defence, and explainability, but these capabilities are rarely evaluated together under realistic blockchain conditions. There is therefore a need for solutions that jointly address temporal adaptation, adversarial robustness, explainability, scalability, and trustworthy decision-making.

The findings therefore support the central premise of this review: future blockchain fraud-detection research should move beyond isolated improvements in GNN architecture towards integrated, adaptive, and trustworthy graph-learning systems. Such systems should be evaluated not only according to predictive performance, but also according to their ability to operate on evolving transaction networks, withstand realistic adversarial conditions, provide meaningful explanations, and remain computationally feasible at scale. Figure 7 summarizes the cross-study synthesis by linking the identified strengths, methodological trade-offs, research maturity, and evidence gaps.

4.7. Implications for Research and Practice

For researchers, the findings suggest that future studies should evaluate GNN-based blockchain fraud detection using more consistent experimental protocols that incorporate temporal validation, adversarial testing, explanation evaluation, and computational-efficiency measures. Such evaluation would improve comparability between studies and reduce the current dependence on isolated performance claims. Greater attention should also be given to

realistic dynamic datasets and evaluation settings that reflect changes in transaction structures and fraud behaviors, as these factors are essential for determining whether a proposed solution can maintain its effectiveness beyond static benchmark conditions.

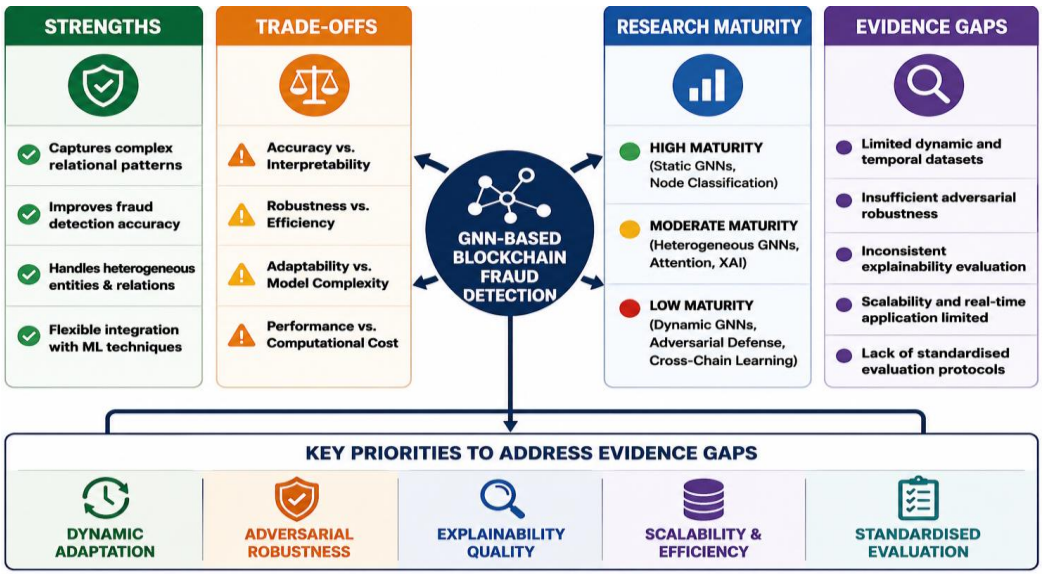

Figure 7. Cross-study synthesis of strengths, trade-offs, research maturity, and evidence gaps.

For practitioners, the findings suggest that model selection should not be based solely on reported accuracy or F1-score. A model intended for operational blockchain monitoring should also be considered in terms of detection latency, scalability, robustness against manipulation, explanation quality, and its ability to adapt to newly emerging transaction behaviors. These considerations are particularly important in security-sensitive applications, where a model with high predictive performance under controlled conditions may not necessarily provide the same level of effectiveness when deployed in a continuously evolving and potentially adversarial transaction environment.

Recent work demonstrates that research is already moving towards more real-time and adaptive systems. For example, reinforcement-learning and GNN combinations have been investigated for real-time financial fraud detection, while recent financial GNN studies increasingly incorporate adaptive aggregation, metagraphs, and temporal information [41]. These developments indicate a broader shift towards adaptive graph-learning solutions, although further evidence is still required to establish how such approaches perform when temporal adaptation, adversarial robustness, explainability, and scalability are evaluated together.

5. Future Research Directions and Research Roadmap

The findings synthesized in Section 4 indicate that future blockchain fraud-detection research should move beyond isolated improvements in GNN architecture towards systems that can adapt to evolving transaction patterns while maintaining robustness, explainability, scalability, and trustworthiness. Recent research on dynamic and trustworthy GNNs similarly highlights the need to address temporal evolution, scalability, robustness, and explainability in an integrated manner [10]. Accordingly, the following directions focus on the main evidence gaps identified in the review and provide a progressive roadmap for developing more adaptive and practically reliable GNN-based blockchain fraud-detection systems.

5.1 Dynamic and Adaptive Fraud Detection

Future research should priorities GNN models capable of learning from continuously evolving blockchain transaction networks. Static models may become less effective when transaction structures and fraud behaviors change over time. Dynamic GNNs, incremental learning, temporal memory, and concept-drift detection therefore represent important directions for maintaining detection performance in changing environments [45]. Future studies should also use chronological evaluation and continuously updated datasets to assess whether models can adapt to emerging fraudulent behaviors rather than relying only on randomly

divided historical datasets. Particular attention should be given to mechanisms that can update graph representations efficiently while maintaining detection performance as transaction volumes and behavioral patterns change.

5.2 Adversarial Robustness and Explainability

A second priority is the integration of adversarial robustness and explainability into dynamic blockchain fraud-detection models. Existing research demonstrates vulnerabilities in GNNs to graph and feature perturbations, while explainability research shows the importance of evaluating whether generated explanations are reliable and meaningful [52]. Future work should therefore evaluate models under both clean and adversarial conditions while providing explanations that identify the transaction relationships, features, or temporal patterns contributing to fraud predictions. Explanation quality should be assessed using measurable criteria rather than visual inspection alone. In addition, future studies should investigate whether explanations remain stable and meaningful when the underlying graph is dynamically updated or deliberately manipulated, as this would provide stronger evidence of the practical trustworthiness of the resulting fraud-detection systems.

5.3 Scalable and Reproducible Evaluation

Scalability and reproducibility remain important barriers to the practical deployment and comparison of blockchain fraud-detection solutions. Future research should investigate efficient graph sampling, temporal updating, large-scale graph processing, and heterogeneous or cross-chain learning to support increasingly complex transaction networks. At the same time, studies should adopt more consistent evaluation protocols covering temporal generalization, class imbalance, computational cost, adversarial robustness, and explanation quality. The development of standardized dynamic blockchain benchmarks would make it easier to compare competing approaches and determine whether reported improvements generalize beyond individual datasets. Such benchmarks should ideally provide clearly defined graph-construction procedures, temporal splits, fraud labels, evaluation metrics, and computational reporting requirements to improve the reproducibility and practical interpretation of experimental results.

5.4 Research Roadmap

The research gaps identified in Section 4 can be translated into a progressive roadmap for future research. As shown in Figure 9, the roadmap progresses from establishing a reliable data and graph foundation to adaptive learning, robust and secure learning, explainable AI, and finally trustworthy and scalable deployment. This progression reflects the increasing integration of capabilities identified throughout the review, rather than suggesting that each stage must be completed independently before the next stage can be investigated.

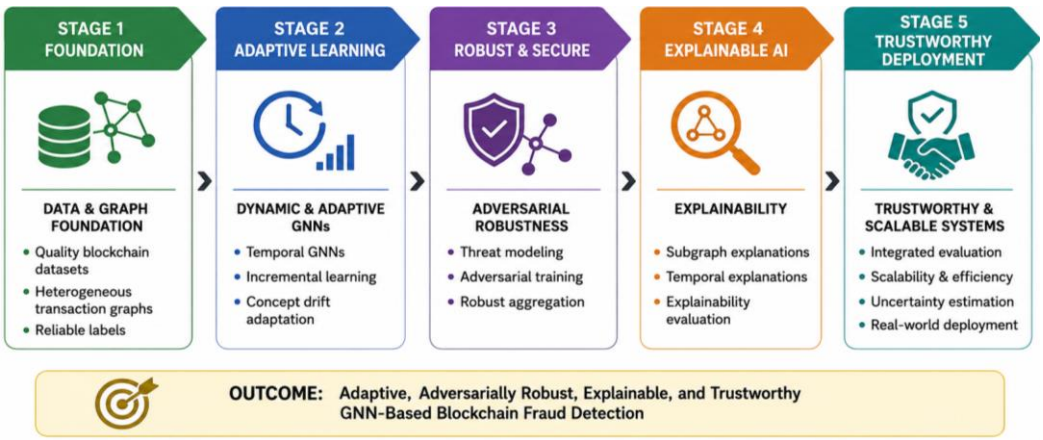

Figure 9. Research roadmap for adaptive, adversarially robust, explainable, and trustworthy GNN-based blockchain fraud detection.

- Stage 1: Foundation — Data and Graph Foundation. This stage focuses on establishing quality blockchain datasets, heterogeneous transaction graphs, and reliable fraud labels. A reliable data and graph foundation is necessary for developing and evaluating

subsequent GNN-based fraud-detection methods, particularly because differences in datasets and graph-construction strategies currently limit the comparability of reported results.

- Stage 2: Adaptive Learning — Dynamic and Adaptive GNNs. This stage addresses the temporal and evolving nature of blockchain transactions through Temporal GNNs, incremental learning, and concept-drift adaptation. The objective is to enable fraud-detection models to respond to changes in transaction behavior and graph structure while maintaining computational efficiency and timely detection.
- Stage 3: Robust and Secure — Adversarial Robustness. This stage incorporates threat modelling, adversarial training, and robust graph aggregation to improve the resilience of GNN-based fraud detection against adversarial perturbations and manipulation of graph structures. Evaluation at this stage should consider both clean-data performance and robustness under clearly defined and diverse attack conditions.
- Stage 4: Explainable AI — Explainability. This stage integrates subgraph explanations, temporal explanations, and systematic explanation evaluation to make fraud predictions more interpretable and to provide evidence that can support investigation and model assessment. Particular attention should be given to the reliability and stability of explanations when transaction relationships evolve over time.
- Stage 5: Trustworthy Deployment — Trustworthy and Scalable Systems. The final stage focuses on integrated evaluation, scalability and efficiency, uncertainty estimation, and real-world deployment. This stage represents the transition from individual research components towards operationally reliable fraud-detection systems that jointly consider predictive performance, temporal adaptability, robustness, explainability, and practical deployment requirements.

Overall, the roadmap connects the major research gaps identified in Section 4 and provides a progressive pathway from reliable data and graph construction to adaptive learning, adversarial robustness, explainability, and trustworthy deployment. For researchers, the roadmap provides a structured sequence for developing and evaluating future GNN-based blockchain fraud-detection systems. For practitioners, it provides broader criteria for assessing system readiness beyond conventional predictive-performance measures, including robustness, interpretability, scalability, uncertainty, and deployment considerations.

Practical Implications

Based on the results of this review, the proposed roadmap provides practical guidance for financial institutions, cryptocurrency exchanges, blockchain developers, and regulatory bodies. Rather than assessing fraud-detection systems solely according to predictive accuracy, practical deployment should consider whether a system can operate on large and evolving transaction networks, withstand adversarial manipulation, provide interpretable evidence for its decisions, and maintain reliable performance under changing conditions. These considerations can support the design and assessment of fraud-detection systems that are not only accurate, but also robust, explainable, scalable, and better aligned with the operational and regulatory requirements of blockchain-based financial environments.

6. Conclusions

As blockchain-based financial activity continues to expand, fraud-detection systems must increasingly address transaction networks that are dynamic, structurally complex, and potentially subject to adversarial manipulation [35], [61]. This systematic literature review examined the development of GNN-based solutions for fraud detection in dynamic blockchain networks, with particular attention to four interconnected requirements: dynamic adaptation, adversarial robustness, explainability, and trustworthy deployment. The synthesis shows that GNNs provide an appropriate foundation for modelling relational transaction behavior, while temporal and heterogeneous approaches extend their ability to represent evolving and complex blockchain interactions. However, the evidence does not support a universal advantage of any single GNN architecture; reported effectiveness remains dependent on the dataset, graph representation, fraud characteristics, class distribution, and evaluation setting.

The review further shows that advances in adversarial defence and explainable AI have expanded the scope of GNN-based fraud detection beyond predictive performance alone [10]. Nevertheless, these capabilities remain unevenly integrated into dynamic blockchain

applications. Important limitations persist in the availability of realistic dynamic benchmark datasets, scalability to large and continuously evolving transaction networks, and standardized evaluation protocols covering predictive performance, temporal generalization, adversarial robustness, explainability, and computational efficiency [2], [44]. These limitations restrict direct comparison across studies and leave insufficient empirical evidence regarding whether proposed solutions can maintain reliable performance under realistic operational conditions.

In addressing the objectives of this review, the findings were synthesized into a solution-oriented view of the field, highlighting the progression from conventional graph-based fraud detection towards dynamic, robust, explainable, and trustworthy GNN-based systems. The review also identified the principal trade-offs between predictive capability, scalability, temporal adaptability, robustness, and interpretability, and used these findings to develop a research roadmap for future work. The roadmap provides a progressive pathway from reliable data and graph foundations through adaptive learning, adversarial robustness, and explainability towards integrated and trustworthy deployment. Future research should therefore focus not only on developing new GNN architectures, but on establishing integrated solutions and evaluation frameworks that can demonstrate their reliability under evolving, imbalanced, adversarial, and large-scale blockchain conditions [20], [75].

Author Contributions: Conceptualization: O.A.O. and R.A.; Writing—original draft preparation: O.A.O.; Search for related works: O.A.O., R.A., B.I.A., and K.H.L.; Writing—review and editing: R.A., B.I.A., and K.H.L.; Supervision: R.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study did not create or generate new data. The findings are based on information obtained from publicly available scholarly literature and sources cited within the manuscript.

Acknowledgements: The authors acknowledge the use of artificial intelligence (AI)-based tools to assist with language refinement, grammar checking, paraphrasing, and improving the clarity and readability of the manuscript. AI tools were used as editorial support and did not replace the authors' responsibility for the literature search, study selection, analysis, synthesis, interpretation of the findings, or development of the research conclusions. The authors critically reviewed and verified all AI-assisted content and remain fully responsible for the accuracy, originality, and integrity of the manuscript.

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

References

[1] N. Han, R. Zhang, X. Liu, and H. Zhang, "Illicit Bitcoin transaction detection via feature-gated temporal graph learning," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-53783-y.

[2] Y. Elmougy and L. Liu, "Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3979–3990. doi: 10.1145/3580305.3599803.

[3] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*. https://doi.org/10.48550/arXiv.1908.02591

[4] S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," *Expert Syst. Appl.*, vol. 240, p. 122156, Apr. 2024, doi: 10.1016/j.eswa.2023.122156.

[5] F. Mujkanovic, S. Geisler, S. Günnemann, and A. Bojchevski, "Are Defenses for Graph Neural Networks Robust?," in *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0651.

[6] F. Ares-Robledo, H. Rifà-Pous, and R. Clirisó, "Graph neural networks for anomaly detection: a systematic review of dynamic temporal approaches," *Artif. Intell. Rev.*, vol. 59, no. 5, 2026, doi: 10.1007/s10462-026-11532-7.

[7] I. Alarab and S. Prakoonwit, "Graph-Based LSTM for Anti-money Laundering: Experimenting Temporal Graph Convolutional Network with Bitcoin Data," *Neural Process. Lett.*, vol. 55, no. 1, 2023, doi: 10.1007/s11063-022-10904-8.

[8] R. Huang, "FinGuard-GNN: Dynamic Graph Neural Network Framework for Financial Fraud Detection," *Front. Business, Econ. Manag.*, vol. 19, no. 3, 2025, doi: 10.54097/wh6hg844.

[9] A. Harper and M. D. Lee, "Scalable Blockchain Fraud Detection Using Spatial-Temporal Graph Neural Networks," *Front. Appl. Phys. Math.*, vol. 2, no. 1, 2025, doi: 10.71465/fapm141.

[10] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang & Suhang Wang, "A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability," *Mach. Intell. Res.*, 2024, doi: 10.1007/s11633-024-1510-8.

[11] H. Farrukh, S. Zafar, Z. U. Rehman, A. A. Shah, and N. Alshammry, "Blockchain-Based Fraud Detection: A Comparative Systematic Literature Review of Federated Learning and Machine Learning Approaches," *Electronics*, 2025, doi: 10.3390/electronics14244952.

[12] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao B. Schardl, Charles E. Leiserson, "EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs." 2019 doi.org/10.48550/arXiv.1902.10191

[13] Kulkarni, O., & Chandra, R. (2025). DynBERG: Dynamic BERT-based Graph neural network for financial fraud detection. arXiv. https://doi.org/10.48550/arXiv.2511.00047.

[14] V. Sivakumar, A. K. Verma, P. Kumar, N. Saw, and V. H. Krishnan, "Fraud Detection in Financial Transaction Using GNN," in *Proceedings of International Conference on Visual Analytics and Data Visualization (ICVADV 2025)*, 2025. doi: 10.1109/ICVADV63329.2025.10960835.

[15] Gu, W., Sun, M., Liu, B., Xu, K., & Sui, M. (2024). Adaptive spatio-temporal aggregation for temporal dynamic graph-based fraud risk detection. Journal of Computer Technology and Software, 3(5), 1–5. https://doi.org/10.5281/zenodo.13626101.

[16] A. A. J. Al-Hchaimi, M. A. Khalifa, and W. El-Shafai, "Explainable AI With Imbalanced Learning Strategies for Blockchain Transaction Fraud Detection," *Eng. Reports*, vol. 8, no. 1, 2026, doi: 10.1002/eng2.70545.

[17] T. Zhang and H. Ye, "Fraud Detection Based on Graph Neural Network," in *Frontiers in Artificial Intelligence and Applications*, 2023. doi: 10.3233/FAIA230858.

[18] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Systematic Reviews, 10, 89. https://doi.org/10.1186/s13643-021-01626-4.

[19] Prisma, "Transparent reporting of systematic reviews and meta-analyses." 2020.

[20] Y. Zhang and others, "A Survey on Privacy in Graph Neural Networks: Attacks, Preservation, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, 2024, doi: 10.1109/TKDE.2024.3454328.

[21] B. Wu, K. M. Chao, and Y. Li, "Heterogeneous graph neural networks for fraud detection and explanation in supply chain finance," *Inf. Syst.*, vol. 121, p. 102335, 2024, doi: 10.1016/j.is.2023.102335.

[22] Y. Liu, Z. Sun, and W. Zhang, "Improving fraud detection via hierarchical attention-based Graph Neural Network," *J. Inf. Secur. Appl.*, vol. 72, 2023, doi: 10.1016/j.jisa.2022.103399.

[23] A. Huang, X. Zhang, Y. Wang, S. Tsai, P. Zhou, and L. Chen, "Dynamic Calibration of Decision Thresholds for Financial Anomaly Detection: Verification With Payment Platform Information and Data," *J. Glob. Inf. Manag.*, vol. 33, no. 1, 2025, doi: 10.4018/JGIM.395852.

[24] J. Xia, L. Wu, J. Chen, B. Hu, and S. Z. Li, "SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation," in *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 2022. doi: 10.1145/3485447.3512156.

[25] P. Li, Y. Xie, X. Xu, J. Zhou, and Q. Xuan, "Phishing Fraud Detection on Ethereum Using Graph Neural Network," in *Communications in Computer and Information Science*, 2022. doi: 10.1007/978-981-19-8043-5_26.

[26] J. Chen and Y. Yang, "Real-time dynamic graph learning with temporal attention for financial fraud detection," *Front. Artif. Intell.*, vol. 9, 2026, doi: 10.3389/frai.2026.1774013.

[27] Y. Tian, G. Liu, J. Wang, and M. Zhou, "ASA-GNN: Adaptive Sampling and Aggregation-Based Graph Neural Network for Transaction Fraud Detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 3, 2024, doi: 10.1109/TCSS.2023.3335485.

[28] Y. Wang, Q. Zheng, X. Li, L. Wang, and L. Lin, "CoSemiGNN: Blockchain fraud detection with dynamic graph neural networks based on co-association of semi-supervised," *Expert Syst. Appl.*, vol. 298, p. 129853, 2026, doi: 10.1016/j.eswa.2025.129853.

[29] M. U. Khan, M. T. Shahid, M. Ashraf, M. Riaz, U. Rahman, and A. Iqbal, "Financial Fraud Detection with AI: A Machine Learning-Based Approach for Securing Digital Transactions," *Glob. Res. J. Nat. Sci. Technol.*, 2025, doi: 10.53762/grjnst.03.03.09.

[30] A. Singh and V. Pareek, "Blockchain for Cybersecurity: Strengthening Digital Defenses with Decentralization and Cryptography," in *Multimedia And Multimodal Intelligence For Sustainable Development*, 2026. doi: 10.1201/9781003634522-6.

[31] J. Gao, B. Ansu, Q. Xia, C. Akwaboah, I. A. Obiri, and G. M. Ntuala, "Enhanced Temporal Graph Networks for Fraud Detection in Transactional Blockchain Networks," in *Proceedings of the International Joint Conference on Neural Networks*, 2025. doi: 10.1109/IJCNN64981.2025.11228825.

[32] B. S. Richard, J. Gao, Q. Xia, K. Victor, E. B. Fabien, and M. M. Rossini, "EAGLE: Ensemble Adaptive Graph Learning for Enhanced Ethereum Fraud Detection," in *Lecture Notes in Computer Science*, 2026. doi: 10.1007/978-981-95-3543-9_19.

[33] Z. Sheng, L. Song, and Y. Wang, "Dynamic Feature Fusion: Combining Global Graph Structures and Local Semantics for Blockchain Phishing Detection," *IEEE Trans. Netw. Serv. Manag.*, vol. 22, no. 5, 2025, doi: 10.1109/TNSM.2025.3576130.

[34] M. Chen, R. Liu, Q. Song, G. Song, C. Li, and Q. Zeng, "HSTA: Ethereum Phishing Fraud Detection Model Based on Dynamic Graph Hybrid Spatio-Temporal Attention Mechanism," in *Communications in Computer and Information Science*, 2026. doi: 10.1007/978-981-95-4142-3_5.

[35] Y. Cui, X. Han, J. Chen, X. Zhang, J. Yang, and X. Zhang, "FraudGNN-RL: A Graph Neural Network With Reinforcement Learning for Adaptive Financial Fraud Detection," *IEEE Open J. Comput. Soc.*, vol. 6, pp. 426–437, 2025, doi: 10.1109/OJCS.2025.3543450.

[36] Lin, Z., Luo, Q., Wu, D., Shen, J., Li, L., Nong, X., & Qin, Z. (2026). Detecting illicit transactions in bitcoin: A wavelet-temporal graph transformer approach for anti-money laundering. Scientific Reports, 16(1), 1548. https://doi.org/10.1038/s41598-025-23901-3.

[37] A. Amores, G. M. Ramírez V, F. Gutiérrez, J. Díaz-Arancibia, and F. Moreira, "Navigating the Trust Landscape: Fraud Analysis of Ethereum Blockchain Networks Using A.I," in *Lecture Notes in Networks and Systems*, 2026. doi: 10.1007/978-3-032-01234-0_32.

[38] F. Johannessen and M. Jullum, "Finding Money Launderers Using Heterogeneous Graph Neural Networks." 2023.

[39] F. Wan and P. Li, "A Novel Money Laundering Prediction Model Based on a Dynamic Graph Convolutional Neural Network and Long Short-Term Memory," *Symmetry (Basel).*, vol. 16, no. 3, 2024, doi: 10.3390/sym16030378.

[40] E. Cengiz and M. Gök, "Anti Money Laundering in Bitcoin Network Using Chaotic Time Series and Graph Convolution Network," *J. Univers. Comput. Sci.*, vol. 31, no. 11, 2025, doi: 10.3897/jucs.135907.

[41] C. Lou, Y. Wang, J. Li, Y. Qian, and X. Li, "Graph neural network for fraud detection via context encoding and adaptive aggregation," *Expert Syst. Appl.*, vol. 261, p. 125473, 2025, doi: 10.1016/j.eswa.2024.125473.

[42] L. Hu, "GNN-Augmented RL for Fraud Detection in Decentralized Finance," *Appl. Comput. Eng.*, vol. 152, no. 1, 2025, doi: 10.54254/2755-2721/2025.22856.

[43] B. Deng, J. Chen, Y. Hu, C. Chen, and T. Zhang, "Mutual GNN-MLP distillation for robust graph adversarial defense," *Neural Networks*, vol. 189, 2025, doi: 10.1016/j.neunet.2025.107513.

[44] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-023-00876-4.

[45] J. Chen, M. Ma, H. Ma, and H. Zheng, "A Survey of Privacy Security for Graph Neural Networks," *J. Cyber Secur.*, vol. 10, no. 3, 2025, doi: 10.19363/J.cnki.cn10-1380/tn.2025.05.08.

[46] Wu, T., Cui, C., Xian, X., Qiao, S., Wang, C., Yuan, L., & Yu, S. (2025). Understanding the robustness of graph neural networks against adversarial attacks. Knowledge-Based Systems, 323, 113714. https://doi.org/10.1016/j.knosys.2025.113714.

[47] Q. Tao, J. Liao, E. Zhang, and L. Li, "A Dual Robust Graph Neural Network Against Graph Adversarial Attacks," *Neural Networks*, vol. 175, 2024, doi: 10.1016/j.neunet.2024.106276.

[48] L. Gosch, S. Geisler, D. Sturm, B. Charpentier, D. Zügner, and S. Günnemann, "Adversarial Training for Graph Neural Networks: Pitfalls, Solutions, and New Directions," in *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-2532.

[49] L. Shi and W. Liu, "A Closer Look at Curriculum Adversarial Training: From an Online Perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i13.29418.

[50] H. A. Prasetya, X. Liu, T. Murata, and A. Matono, "A multi-rounded adversarial scenario for graph-based promo fraud detection," *Soc. Netw. Anal. Min.*, vol. 16, no. 1, 2025, doi: 10.1007/s13278-025-01566-0.

[51] Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, Suhang Wang. A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability[J]. Machine Intelligence Research, 2024, 21(6): 1011-1061. DOI: 10.1007/s11633-024-1510-8.

[52] M. Nandan, S. Mitra, and D. De, "GraphXAI: a survey of graph neural networks (GNNs) for explainable AI (XAI)," *Neural Comput. Appl.*, 2025, doi: 10.1007/s00521-025-11054-3.

[53] C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik, "Evaluating explainability for graph neural networks," *Sci. Data*, vol. 10, no. 1, 2023, doi: 10.1038/s41597-023-01974-x.

[54] F. Ertam, "Near Real-Time Ethereum Fraud Detection Using Explainable AI in Blockchain Networks," *Appl. Sci.*, vol. 15, no. 19, 2025, doi: 10.3390/app151910841.

[55] M. Kamran, M. M. Rehan, W. Nisar, and M. W. Rehan, "ARCADE-Adversarially Robust Cost-Sensitive Anomaly Detection in Blockchain Using Explainable Artificial Intelligence," *Electronics*, vol. 14, no. 8, 2025, doi: 10.3390/electronics14081648.

[56] S. Ferretti, G. D'Angelo, and V. Ghini, "Enhancing Anti-Money Laundering Frameworks: An Application of Graph Neural Networks in Cryptocurrency Transaction Classification," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3552240.

[57] M. Z. Haider, T. Noreen, and M. Salman, "Towards Quantum-Ready Blockchain Fraud Detection via Ensemble Graph Neural Networks," in *2025 7th International Conference on Blockchain Computing and Applications (BCCA 2025)*, 2025. doi: 10.1109/BCCA66705.2025.11229725.

[58] A. Asiri and K. Somasundaram, "Graph convolution network for fraud detection in bitcoin transactions," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-95672-w.

[59] M. Z. Haider, T. Noreen, and M. Salman, "Blockchain Fraud Detection Using Ensemble Graph Neural Networks," *J. Artif. Intell. Res.*, vol. 2, no. 2, 2025, doi: 10.70891/jair.2025.080018.

[60] A. Laurent, "Graph Neural Networks for Blockchain Security: A Deep Learning Approach to Anomaly Detection," *Front. Interdiscip. Appl. Sci.*, vol. 2, no. 01, 2025, doi: 10.71465/fias.v2i01.18.

[61] Q. Wang, Y. Shen, and H. Dong, "Hypergraph-based contrastive learning for enhanced fraud detection," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1703135.

[62] V. Kale and S. Chakraborty, "Graph Neural Networks for Financial Fraud Detection: A Detailed Literature Review," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 12, no. 2, pp. 346–353, 2026, doi: 10.32628/CSEIT26121352.

[63] G. M. Ntuala and others, "A dual-layer GNN with economic penalty mechanisms for blockchain fraud detection," *Expert Syst. Appl.*, vol. 299, p. 130121, 2026, doi: 10.1016/j.eswa.2025.130121.

[64] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial Training Methods for Deep Learning: A Systematic Review," *Algorithms*, 2022, doi: 10.3390/a15080283.

[65] F. Guan, T. Zhu, W. Zhou, and K. K. R. Choo, "Graph neural networks: a survey on the links between privacy and security," *Artif. Intell. Rev.*, vol. 57, no. 2, 2024, doi: 10.1007/s10462-023-10656-4.

[66] A. Sajeeda and B. M. M. Hossain, "Exploring generative adversarial networks and adversarial training," *Int. J. Cogn. Comput. Eng.*, 2022, doi: 10.1016/j.ijcce.2022.03.002.

[67] Q. Z. Cai, C. Liu, and D. Song, "Curriculum adversarial training," in *IJCAI International Joint Conference on Artificial Intelligence*, 2018. doi: 10.24963/ijcai.2018/520.

[68] A. Mohan, K. P.V, P. Sankar, K. M. Manohar, and A. Peter, "Improving anti-money laundering in bitcoin using evolving graph convolutions and deep neural decision forest," *Data Technol. Appl.*, vol. 57, no. 3, 2023, doi: 10.1108/DTA-06-2021-0167.

[69] Xie, B., Chang, H., Zhang, Z., Wang, X., Wang, D., Zhang, Z., Ying, R., & Zhu, W. (2023). Adversarially robust neural architecture search for graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8143–8152). IEEE. https://doi.org/10.1109/CVPR52729.2023.00787.

[70] R. Weerawarna, S. J. Miah, and X. Shao, "Emerging advances of blockchain technology in finance: a content analysis," *Pers. Ubiquitous Comput.*, vol. 27, no. 4, 2023, doi: 10.1007/s00779-023-01712-5.

[71] T. Adam and F. Babič, "Identifying Illicit Activities in Blockchain Transaction Graph Networks," *Electronics*, vol. 14, no. 23, 2025, doi: 10.3390/electronics14234599.

[72] A. Ancelotti and C. Liason, "Review of blockchain application with Graph Neural Networks, Graph Convolutional Networks and Convolutional Neural Networks." 2024. doi: 10.2139/ssrn.4971275.

[73] G. Tong and J. Shen, "Financial transaction fraud detector based on imbalance learning and graph neural network," *Appl. Soft Comput.*, vol. 149, p. 110984, 2023, doi: 10.1016/j.asoc.2023.110984.

[74] S. J. Buu and H. J. Kim, "Disentangled Prototypical Graph Convolutional Network for Phishing Scam Detection in Cryptocurrency Transactions," *Electronics*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214390.

[75] M. T. Le, O. Harris, C. Bennett, and F. Greene, "Behavior Path Analysis for Blockchain Fraud Detection Using Graph Neural Architectures," *Front. Appl. Phys. Math.*, vol. 2, no. 1, 2025, doi: 10.71465/fapm229.

[76] Amara, K., Ying, R., Zhang, Z., Han, Z., Shan, Y., Brandes, U., Schemm, S., & Zhang, C. (2022). GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks. arXiv, doi.org/10.48550/arXiv.2206.09677