Review of Applied Science and Technology

Volume: 13; **Issue:** 12
Pages: 243– 279

Graph-Based Anomaly Detection for Data Integrity and Fraud Prevention in Warehouse Management Systems: A PRISMA-Guided Systematic Literature Review

Uttam Kumar Mallick[1]; Songeta Dhar[2];

[1]. *Chief Accountant, Flowtech Ltd., Blantyre, Malawi;*
Master of Commerce (M.Com.) in Accounting, National University, Bangladesh;
Email: uttamkumarmallick737@gmail.com

[2]. *Accounting & Business Analytics Researcher;*
Doctor of Business Administration School of Business;
International American University, Los Angeles, California, USA.
Email: songeta.bd@gmail.com

Received: *15 September 2023;* ***Revised:*** *24 October 2023;* ***Accepted:*** *15 November 2023;* ***Published:*** *04 December 2023;*

Abstract

This study examined graph-based anomaly detection for data integrity and fraud prevention in warehouse management systems through a PRISMA-guided quantitative systematic literature review framework. It organized the literature by data source, graph representation, detection target, method family, learning paradigm, validation design, and reported performance measures. The proposed review process included searching academic databases, removing duplicate records, screening titles and abstracts, assessing full texts, and extracting study-level data. Because verified search and extraction results were unavailable, all numerical values in the current draft are examples rather than findings from a completed review. In that example, 1,268 records were identified, 318 duplicates were removed, and 30 studies were retained from 140 full-text reports assessed. The study profile included 8 studies using operational warehouse or inventory data, 12 using public graph benchmarks from adjacent domains, 2 using public warehouse or logistics data, 5 using simulated data, and 3 combining operational and simulated data. Heterogeneous static graphs represented 9 of the example studies, and node- or entity-level detection was the most frequent target, appearing in 12. Graph convolutional or attention networks formed the largest single method group, with 8 studies, while supervised learning was the most common learning paradigm, with 13. Precision and recall were each reported in 21 example studies, whereas confidence intervals were reported in 7. Random data splitting was the most frequent validation strategy, used in 12 studies; facility-based validation appeared in 2. These patterns show how the review can distinguish direct warehouse evidence from methods transferred from financial fraud, e-commerce, cybersecurity, and other fields. They also demonstrate the importance of separating inventory discrepancies and process deviations from suspected or confirmed fraud. Differences in data sources, labels, graph construction, validation, and performance metrics limit direct comparison across studies. The example values must be replaced with verified screening and extraction results before the abstract can represent completed research.

Keywords

Graph Anomaly Detection; Warehouse Data Integrity; Fraud Prevention; Inventory Record Inaccuracy; Systematic Review.

INTRODUCTION

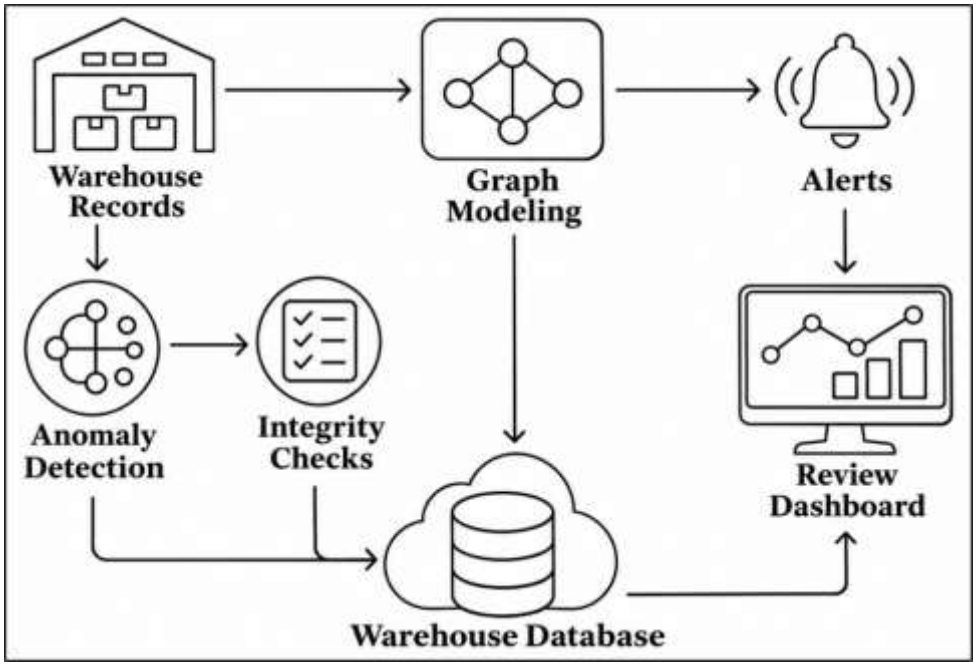
A warehouse management system (WMS) is a digital system that records and coordinates warehouse activities, including receiving, put-away, storage, replenishment, picking, packing, dispatch, returns, and inventory control (Leal et al., 2021). Its records commonly connect products and stock-keeping units (SKUs) with locations, quantities, orders, suppliers, employees, equipment, and timestamps. Data integrity refers to the accuracy, completeness, consistency, and traceability of those records throughout their creation, modification, transmission, and use. In warehouse operations, integrity concerns include whether a recorded receipt matches the quantity physically received, whether a movement is assigned to the correct location, whether an adjustment has an authorized basis, and whether the event history preserves who performed an action and when. Anomaly detection is the computational identification of observations, sequences, relationships, or groups that differ meaningfully from an established pattern (Preuveneers et al., 2018). An anomaly is a statistical or behavioral deviation requiring examination; it is not, by itself, evidence of fraud. Fraud prevention in this setting concerns detecting and investigating suspicious activity early enough to support controls, audit review, or operational intervention. These definitions distinguish data-quality failures, process deviations, malicious manipulation, and confirmed fraud, which may produce similar system signals but have different causes and evidentiary status. Warehouse data integrity is also connected to the broader concept of inventory record inaccuracy (IRI), the difference between inventory quantities represented in information systems and the stock physically available. IRI can arise through unrecorded or incorrectly recorded receipts, sales, replenishments, transfers, returns, damage, misplacement, counting errors, synchronization failures, or deliberate manipulation. Research on retail and distribution inventory has treated record discrepancies as an operational phenomenon that affects replenishment decisions and the reliability of inventory information. At the warehouse level, the unit of analysis can be a transaction, item, location, employee, order, device, or connected set of events. Selecting the unit matters because suspicious behavior may be invisible in an isolated record but recognizable across a sequence of related activities. A single stock adjustment, for example, may be routine; a cluster of repeated adjustments linked to the same user, item, location, and unusual time window may warrant review (Chen et al., 2017). This study therefore treats graph-based anomaly detection as the analysis of relational warehouse data to identify unusual entities, events, links, or subgraphs associated with potential integrity failures or fraud indicators. The distinction between an alert and a verified finding remains central to how detection performance and operational relevance are assessed.

Warehouses are integral to international supply chains that move goods between manufacturers, distribution centers, retailers, ports, and customers. Their systems mediate the exchange of inventory information across organizational boundaries, time zones, languages, technology platforms, and regulatory environments. A discrepancy recorded at one stage can affect order fulfilment, replenishment, production planning, customs documentation, returns processing, and financial reporting elsewhere in the chain. The scale and organization of warehouse operations vary internationally: facilities may differ in automation, workforce composition, physical layout, SKU variety, inventory practices, system integration, and the extent to which records are updated in real time (Wang et al., 2020). These variations shape both the occurrence of data errors and the evidence available for detecting them. For example, a warehouse relying on handheld scanners and manual adjustments may produce different event trails from a highly automated facility using conveyors, robotics, radio-frequency identification, and integrated enterprise software. A detection method developed on one type of facility cannot be assumed to transfer unchanged to another without examining its data representation and operating conditions.

Inventory inaccuracy has been studied across retail and supply-chain settings, with research examining its prevalence, causes, operational effects, and relationship to store or distribution-center performance. Relevant causes include employee actions and omissions, process design, physical conditions, stock movement complexity, and information-system limitations (Sengan et al., 2021). These categories are important for international research because similar discrepancies can arise under different local procedures and control environments. In one setting, a negative inventory balance may reflect delayed synchronization between systems; in another, the same pattern may be associated with a repeated unauthorized adjustment. Similarly, a series of after-hours events may be abnormal under one facility's

operating schedule but routine in a warehouse running continuous shifts. Fraud-related risks can also involve coordination among employees, suppliers, carriers, or external actors, so that records appear individually plausible while their connections reveal unusual patterns. International significance thus concerns not only the value and movement of goods but also the comparability of evidence across business practices and data systems (Blackford et al., 2017). A quantitative review of graph-based methods needs to account for variation in datasets, warehouse functions, geographical settings, and fraud definitions, because these factors can influence the kinds of anomalies represented, the labels assigned, and the performance measures reported. Work on inventory record inaccuracy and its operational consequences provides the warehouse context, while research on graph-based fraud detection supplies methods for analyzing connected behavior.

Figure 1: Graph-Based Warehouse Anomaly Detection Framework



WMS data are generated through interacting processes rather than as independent, unrelated observations. A receipt links a supplier, purchase order, product, quantity, dock, employee, and time. A put-away event connects the same product to a storage location and equipment or operator. Picking and dispatch records then link inventory to a customer order, picker, packing station, carrier, and shipment. Returns, cycle counts, write-offs, and stock adjustments add further relationships among products, locations, users, documents, and operational events (Maw et al., 2019). These connections can be represented as a graph, in which nodes denote entities or events and edges represent relations such as "handled by," "stored at," "belongs to order," "transferred from," or "approved by." A graph may be homogeneous, with one node and edge type, or heterogeneous, with several entity and relationship types. It may also be static, summarizing a period, or temporal, preserving event order and timestamps. This structure is relevant to integrity analysis because a potentially irregular pattern may involve several records that appear ordinary when assessed separately.

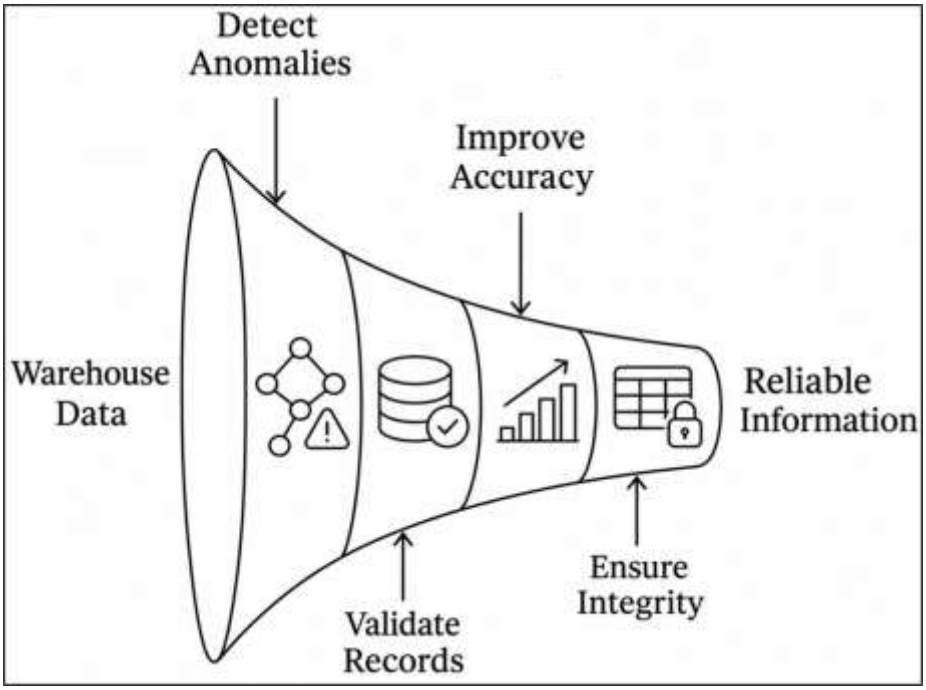
The relational character of warehouse activity means that useful evidence can reside in both attributes and connectivity. Attributes may include quantity, unit price, transaction type, user role, device identifier, location, or time of day. Structural information may include repeated links between the same employee and inventory adjustments, unusual relationships between suppliers and receiving events, or movement paths that diverge from usual process sequences (Hang et al., 2020). Graphs can encode those relations and allow detection at multiple levels: an anomalous node such as an employee or SKU; an anomalous edge such as a transfer or approval; a suspicious subgraph representing a coordinated

set of entities and events; or a broader graph-level pattern. Temporal graphs can additionally represent the order, spacing, and recurrence of actions. The modeling choices affect what the detector can observe. For instance, collapsing repeated events into a single connection may conceal frequency, while omitting user or authorization relationships may prevent analysis of who initiated or approved an adjustment. Inconsistent identifiers, missing timestamps, duplicate records, and disconnected system logs can also distort the graph and create apparent anomalies (Breier & Branišová, 2017). Warehouse research on inventory dynamics and distribution-center systems has emphasized that inventory records emerge from interactions among people, software, hardware, physical products, layouts, and inventory policies. Graph-based review therefore provides a way to examine these interactions as a connected data environment while preserving distinctions among event, actor, object, and process.

Warehouse discrepancies can reflect accidental error, process weakness, system malfunction, external disruption, or intentional misconduct. Possible integrity failures include duplicate or missing events, incorrect quantities, mismatched product or location identifiers, delayed updates, unauthorized edits, inconsistent order and shipment records, and gaps in event provenance. Fraud risks may include theft, false receipts, fictitious returns, deliberate miscounts, unauthorized write-offs, collusion, or manipulation of records to conceal physical stock movement. These categories overlap at the level of observable data (Qureshi et al., 2021). A sequence of unusual write-offs may result from a legitimate stock correction, a software migration, damaged goods, or deliberate concealment. An incomplete transaction chain may arise because a system log failed to capture an event or because an employee bypassed the standard process. For this reason, anomaly detection identifies cases that depart from a modeled baseline; it does not establish intent, culpability, or loss without investigation and corroborating evidence. The challenge of detecting manipulation is sharpened when an actor can influence both physical operations and the records used to audit them. Traditional discrepancy checks often compare system quantities with physical counts, but count-based approaches can be limited when counting is infrequent, costly, or subject to error. Transaction-level rules can identify known conditions, such as negative stock or repeated adjustments, but may miss coordinated activity distributed across several users, items, and locations. Statistical and machine-learning methods can detect deviations in numerical attributes and event sequences, yet models that disregard relationships may fail to identify patterns dependent on shared actors, common devices, approval chains, or coordinated timing. Conversely, structural irregularity alone can produce false positives where legitimate operations naturally create dense or repeated connections (Carvalho et al., 2018).

Relevant model outcomes therefore include detection sensitivity, precision, recall, false-positive burden, calibration, stability across periods, and the usefulness of alerts for review. Evaluation also depends on the quality and independence of labels: confirmed fraud cases may be scarce, while records marked as normal may contain undiscovered problems. Studies of financial and online fraud have repeatedly addressed related complications, including class imbalance, incomplete labels, camouflage, evolving behavior, and the distinction between suspicious relationships and confirmed misconduct. In warehouse settings, evaluating these issues requires explicit attention to what an "anomaly" label represents, how it was obtained, and whether the reported target is a data error, policy violation, inventory loss, or verified fraud (Kamišalić et al., 2021).Graph-based anomaly detection examines unusual properties of nodes, edges, subgraphs, or graph-level observations, often in combination with attributes and temporal behavior. Traditional approaches include statistical outlier analysis, neighborhood and community measures, frequent-pattern methods, distance-based techniques, and probabilistic models. These methods can identify deviations in connectivity, unusual interaction frequencies, rare structural patterns, or unexpected relationships between entities. Deep learning approaches extend this work by learning representations from graph structure and node or edge attributes. Graph neural networks (GNNs) use message passing or related operations to aggregate information from neighboring entities and produce representations for classification, ranking, or reconstruction tasks. Graph convolutional networks, attention-based networks, heterogeneous graph models, and graph autoencoders have been applied to a range of anomaly and fraud problems (Vávra et al., 2021). Their relevance to warehouse data lies in the capacity to represent multiple object types and learn from relational context, including interactions among products, transactions, employees, locations, suppliers, and orders.

Figure 2: Warehouse Data Integrity Enhancement Framework



Published fraud-detection research has developed methods to address particular characteristics of connected data. Heterogeneous graph models can represent different entity and relation types; semi-supervised approaches can use labeled and unlabeled cases; neighborhood-sampling methods can address class imbalance; and camouflage-resistant models attempt to reduce the influence of apparently ordinary neighbors connected to suspicious accounts. Temporal graph methods incorporate changes in relationships or activity over time. These approaches differ in graph construction, labeling assumptions, anomaly level, learning setup, and evaluation dataset, so reported results are not automatically comparable. Some studies use supervised classification with historical labels, while others use semi-supervised or unsupervised detection when labels are sparse. Some evaluate node classification, while others predict suspicious links, rank entities, or identify graph-level patterns. The choice of baseline also affects claims of advantage: a graph model should be compared with relevant non-graph methods as well as suitable graph-based alternatives. Graph methods may require substantial preprocessing and can be sensitive to missing links, noisy attributes, graph sparsity, label leakage, and temporal data leakage (Xu et al., 2020). Computational cost and explainability are additional considerations when outputs must support human review. This review examines these approaches as a methodological family, with attention to how each study represents warehouse-relevant relations, which anomaly targets it evaluates, and how it reports performance. Prior graph-anomaly and graph-fraud reviews provide classifications and methodological context for this synthesis (Feng et al., 2020).

A quantitative systematic review requires consistent extraction of study characteristics and outcome measures while preserving differences in research design. For graph-based warehouse anomaly detection, relevant characteristics include publication year, country or operational setting, study purpose, data source, sample size, observation period, graph type, node and edge definitions, temporal representation, model family, baseline methods, label source, and anomaly category. The measured outcomes may include precision, recall, F1 score, area under the receiver operating characteristic curve, area under the precision-recall curve, accuracy, false-positive rate, detection delay, calibration, or computational time. These measures answer different questions. Accuracy can be misleading when fraudulent or anomalous cases are rare; precision and recall describe the trade-off between false alerts and missed cases; ranking metrics assess ordering of suspicious cases; and detection delay is relevant

when events are evaluated over time (Peng et al., 2017). Studies may also report performance at a threshold selected on a test set, which can inflate estimates if model selection and evaluation are not separated.

Heterogeneity arises not only from models but also from datasets, fraud definitions, sampling strategies, time windows, and label quality. A study built on synthetic warehouse events may support controlled comparison of known patterns, whereas a study using operational data may better reflect real process complexity while facing incomplete or confidential labels. A benchmark transaction network may contain millions of links but differ substantially from a warehouse's entity types and procedures. Reported results can also depend on whether the split is random, temporal, entity-based, or site-based. Random splits risk placing highly related records in training and test sets, while temporal or facility-based splits may provide a more demanding assessment of transfer (Demertzis et al., 2020). To compare findings responsibly, the review must distinguish results that share comparable targets and metrics from findings that cannot be meaningfully pooled. Where outcome measures, designs, or datasets differ substantially, structured narrative synthesis and descriptive statistics can be more defensible than a single pooled estimate. Where studies provide sufficiently comparable estimates, quantitative synthesis can assess average performance and between-study variability, subject to adequate reporting and valid independence assumptions. The PRISMA 2020 framework provides reporting guidance for transparent search, selection, appraisal, and synthesis, including flow diagrams and checklists (Song et al., 2020).

This review uses quantitative evidence mapping to document how methods, data characteristics, and evaluation practices are distributed across the eligible literature, and to identify the conditions under which reported graph-based results were obtained. Research on graph-based anomaly detection and fraud prevention spans financial transactions, online marketplaces, communication networks, cybersecurity, and other settings in which entities interact through observable relationships. Separate research streams examine inventory record inaccuracy, warehouse operations, event-log analysis, and machine-learning methods for stock anomalies. These strands provide relevant constructs and techniques, yet their datasets and outcome definitions are not necessarily interchangeable. A model that identifies suspicious accounts in a payment network does not directly establish performance for warehouse transactions, where graph entities may include SKUs, locations, employees, orders, and physical movement events (González-Aparicio et al., 2018). Likewise, detecting a stock-count discrepancy does not necessarily amount to detecting fraud, because a discrepancy may arise from non-malicious causes. A systematic synthesis focused on warehouse management therefore needs to map both direct warehouse evidence and adjacent graph-fraud evidence, while identifying which studies examine warehouse data explicitly and which provide transferable methods from other domains.

The objective of this quantitative, PRISMA-guided systematic literature review is to identify, classify, and synthesize research on graph-based anomaly detection for data integrity and fraud prevention in warehouse management systems and closely related inventory and supply-chain settings. The review will examine how studies define warehouse anomalies, inventory discrepancies, suspicious activities, and fraud; what data sources they use; and how they represent relationships among products, stock-keeping units, locations, employees, suppliers, orders, devices, and warehouse events. It will classify the graph structures and analytical approaches applied, including homogeneous and heterogeneous graphs, static and temporal representations, traditional graph-analysis methods, and graph neural network models. The review will also document whether each study uses supervised, semi-supervised, or unsupervised learning, and whether its unit of analysis is an entity, transaction, relationship, event sequence, subgraph, or complete network (Shi et al., 2021).

A further objective is to compare how studies establish labels and ground truth, distinguish confirmed fraud from suspected or non-malicious anomalies, address class imbalance and incomplete records, and separate training data from evaluation data. The review will quantitatively organize the performance measures reported, including precision, recall, F1 score, area under the receiver operating characteristic curve, area under the precision-recall curve, accuracy, false-positive rates, detection time, and computational requirements, while noting differences in study design and dataset characteristics that affect direct comparison. It will also examine the baselines used, the validation strategies applied, and the extent to which results are evaluated on real operational data, benchmark datasets, or simulated

records. By mapping these features systematically, the review aims to describe the scope and distribution of existing evidence, identify methodological patterns and reporting gaps, and assess how closely current graph-based approaches address warehouse data-integrity and fraud-detection tasks (Fernando et al., 2020). The synthesis will maintain a clear distinction between detecting unusual patterns and establishing that fraud occurred, so that model outputs and reported outcomes are interpreted according to the evidence and definitions provided in each included study.

LITERATURE REVIEW

This section reviews research on warehouse data integrity, inventory discrepancies, fraud risks, and graph-based anomaly detection. It first defines the principal concepts and identifies the units of analysis used in warehouse studies, including inventory records, transactions, events, users, products, locations, and their relationships. It then examines how warehouse management systems generate interconnected data and how errors or suspicious activity may appear as unusual values, event sequences, links, or network structures. The review considers graph construction choices and compares traditional graph analysis with machine-learning and graph-neural-network approaches. It also examines quantitative evaluation practices, including outcome definitions, class imbalance, ground-truth quality, validation design, performance metrics, and the comparability of results across datasets and operating settings. Because an anomaly is not necessarily fraud, the review distinguishes record inaccuracies and process deviations from suspected and confirmed intentional misconduct. This distinction is important when interpreting model performance and the operational meaning of reported alerts. The outline below organizes the literature into measurable themes so that studies can be compared by data source, graph representation, detection target, method, and evaluation outcome.

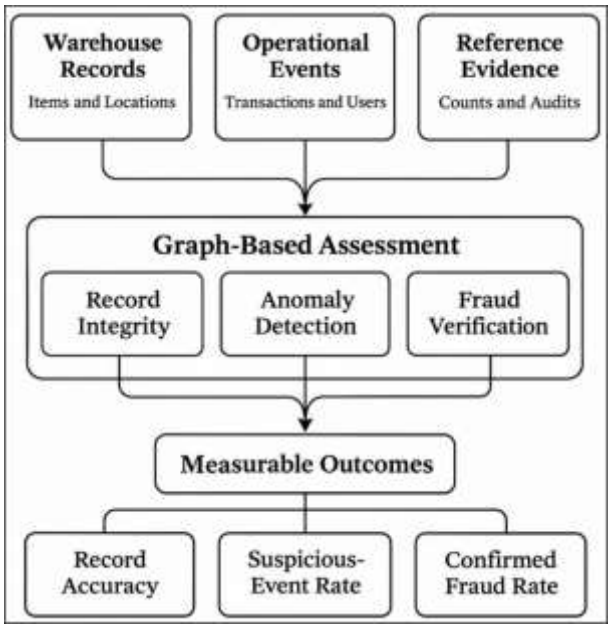
Warehouse Management System (WMS)

A warehouse management system (WMS) is the information system used to record and coordinate activities such as receiving, storage, movement, picking, dispatch, returns, and stock adjustment. It links inventory quantities to products, locations, orders, employees, and operational events. Data integrity refers to the accuracy, completeness, consistency, and traceability of these records. In warehouse research, inventory record inaccuracy (IRI) generally describes a difference between the quantity recorded in the system and the quantity physically available (Schalock et al., 2020). The difference may be positive when the system records less stock than is present, or negative when the system records more stock than can be found. Studies of retail and distribution operations associate discrepancies with counting, scanning, receiving, replenishment, movement, and synchronization problems, as well as stock loss, damage, and misplacement. These causes matter because the same observed quantity difference can arise through different mechanisms. A misplaced item may produce a temporary discrepancy, while an unrecorded receipt may create a persistent record error. Research also examines how product characteristics, inventory density, audit frequency, staffing, and employee-system interactions relate to the occurrence and duration of inaccurate records. Quantitative measures therefore need to specify the comparison basis, such as system quantity against a physical count, the observation period, and the item or location covered (Rewa et al., 2018). Common outcomes include the number or percentage of inaccurate stock-keeping units, the absolute or signed difference between recorded and counted quantities, the monetary value of discrepancies, and the time taken to identify or correct them. Record accuracy may be expressed as the proportion of records that match physical counts under a defined tolerance. Such definitions allow operational discrepancies to be measured consistently while avoiding the assumption that every mismatch reflects deliberate action.

Anomaly detection refers to identifying records, events, entities, relationships, or sequences that differ from an established pattern or defined expectation. In warehouse analysis, an unusual event might involve a quantity outside its usual range, a transaction at an unexpected time, repeated corrections by the same user, or a movement that departs from the normal sequence for a particular process. A process deviation is an observed departure from a prescribed or commonly followed procedure; it may be accidental, authorized under unusual circumstances, or intentional (Sargent & Brown, 2017). A fraud indicator is a signal that warrants examination because it may be consistent with deliberate deception, such as unauthorized adjustments, fabricated receipts, repeated unexplained write-offs, or coordinated activity across users and inventory records. A confirmed fraud case requires verification through an investigation, audit, or other evidence establishing intentional misconduct under the study's stated

criteria. These concepts should not be treated as interchangeable. An alert can be anomalous without violating policy, and a policy deviation can occur without an intent to deceive. Likewise, a record discrepancy may identify a control problem without demonstrating theft or manipulation. Studies of graph-based fraud detection face a similar distinction between suspicious network patterns and cases confirmed as fraudulent (Bassi et al., 2017). Their quantitative findings depend partly on how labels are defined and obtained, including audit confirmation, investigator review, rules, proxy labels, or simulated anomalies. Warehouse studies that analyze event logs can identify after-hours actions, repeated corrections, and skipped process steps, but these remain candidate indicators unless independently verified. The distinction should be maintained in outcome labels and reporting: suspected events, investigated cases, confirmed fraud, and non-fraud process errors represent different levels of evidence.

Figure 3: Warehouse Integrity Assessment and Outcomes



Quantitative studies can measure warehouse integrity and anomaly detection through several distinct outcomes. Discrepancy frequency is the number or proportion of records, items, or counts that fail a defined match criterion. Unauthorized adjustment count is the number of inventory changes that lack the required approval or supporting documentation, based on the system's control rules. Suspicious-event rate is the number of events flagged by a rule, model, or investigation process relative to all observed events in a specified period. Confirmed fraud rate is the number of verified fraud cases relative to a stated denominator, such as transactions, audited items, employees, or facilities (Martin & O'Halloran, 2020). Record accuracy is the share of records whose system values fall within a predefined tolerance of physical or independently verified values. For model evaluation, studies may additionally report precision, recall, F1 score, false-positive rate, or ranking measures; these describe detection performance, rather than the underlying prevalence of fraud or inaccuracy. Graph-based studies broaden the possible unit of analysis. A record may be a row in an inventory table; a transaction represents a stock-affecting exchange; an event is an action recorded in an operational log; an item is a product or SKU; a user is an employee or account; and a location is a storage or processing area. A relationship connects entities, for example an employee to an adjustment or an item to a location. A sequence represents ordered events, while a subgraph comprises a connected group of entities and relations examined for a coordinated pattern (Alla et al., 2017). A facility may be the unit when comparing warehouses or distribution centers. The choice affects interpretation: a model can score a user as suspicious, detect a specific transaction, or flag a connected event group, and the resulting rates are not directly comparable without their denominators and label criteria.

Operational definitions determine which studies can be meaningfully compared. For each study, a review should record the stated outcome, the unit of analysis, the observation window, the denominator, the threshold used to classify an anomaly, and the source of any ground-truth label. For inventory accuracy, the review should note whether the reference value comes from a cycle count, full physical audit, sensor reading, or another system (Sekhon et al., 2017). For suspected fraud, it should record whether the label reflects a model alert, a rule-based suspicion, an investigation, or a confirmed case. For graph-based work, it should also identify whether the anomaly is assigned to a node, edge, path or sequence, subgraph, or entire graph. This detail prevents a user-level detection rate from being compared directly with a transaction-level false-positive rate, or an inventory discrepancy measure from being interpreted as a confirmed fraud rate. The same caution applies to metrics such as accuracy and F1 score: their values depend on class prevalence, labeling rules, thresholds, and the division of data into training and evaluation sets. A study may report high classification accuracy where normal events dominate, while providing limited information about how many genuine suspicious cases were detected or how many alerts required investigation. A quantitative synthesis can therefore classify outcomes into record integrity, process conformance, suspected anomaly, and confirmed fraud, and summarize model metrics separately from operational rates (Augustinavicius et al., 2018). Where studies use different definitions or units, the review can describe their results by category rather than treating them as one pooled measure. This approach preserves the differences in evidence while making explicit which operational definitions, labels, and denominators support each reported outcome.

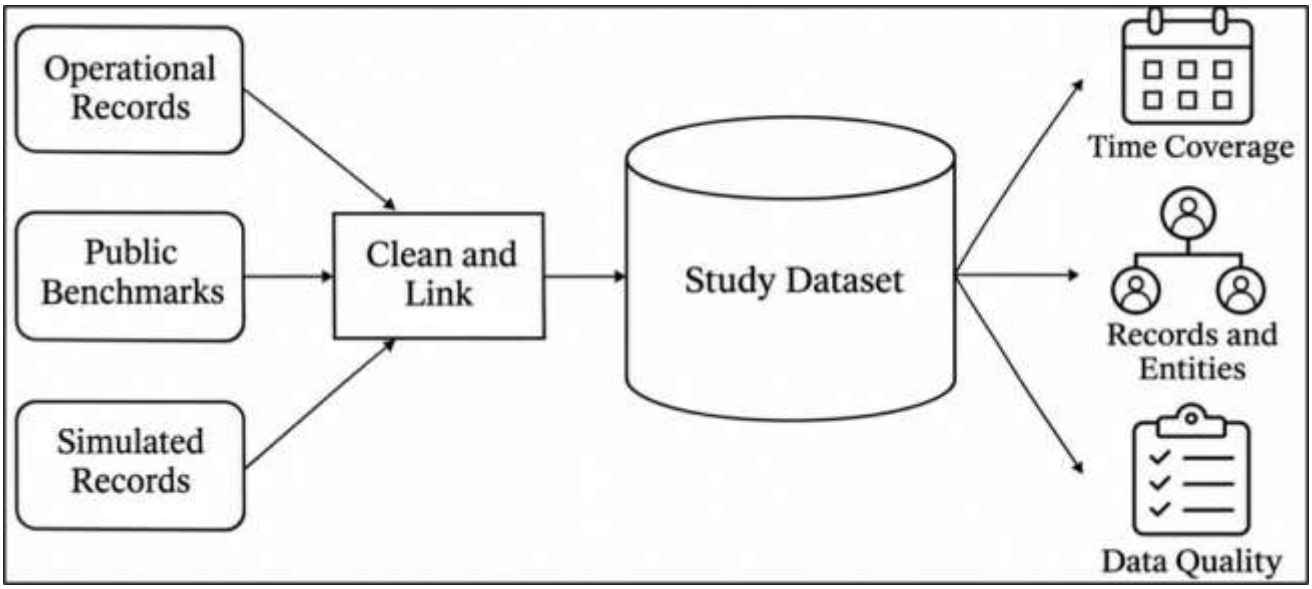
Warehouse Data Sources and Quantitative Characteristics

Warehouse studies draw on data produced at different stages of inventory handling. Receiving records can include purchase orders, supplier identifiers, received quantities, inspection results, timestamps, and the employee or device recording the receipt. Stock movement logs capture transfers between locations, replenishment, put-away, picking, and other changes to an item's position or status (Alaux et al., 2018). Cycle-count and audit data provide comparisons between system quantities and physical counts, while picking and packing records can include order identifiers, item quantities, picker activity, batch details, and timestamps. Dispatch and returns data link warehouse activity to shipments, carriers, customers, and returned goods. Adjustment histories record changes such as write-offs, corrections, or quantity amendments, often alongside user and approval information. User activity and access logs can add information about who performed or authorized actions, when they occurred, and which system functions were used. These data types vary in detail and completeness: some studies use event-level records with timestamps and actors, while others analyze aggregated inventory snapshots, item counts, or order-level summaries. Studies using data extracted directly from WMS platforms show how detailed operational logs can support process and order-picking analyses (Souibgui et al., 2019). However, operational coverage depends on what the system records. Physical actions performed without scanning, events recorded on paper, delayed synchronization, or inconsistent identifiers can leave gaps between activity and its digital trace. For quantitative review, each data source should be described by its process stage, recorded attributes, level of aggregation, time granularity, and relationship to inventory or fraud outcomes.

Warehouse datasets differ in facility coverage, organizational setting, observation period, and scale. A study may examine one distribution center, a single retail chain, multiple warehouses, or a dataset assembled for a specific process such as order picking. Observation periods range from a limited operational sample to repeated daily records across longer periods. Dataset size should be reported using more than one count where possible: the number of raw records, the number retained after cleaning, the number of orders or inventory items, the number of distinct users and locations, and the number of event or transaction types (Naeem et al., 2021). A recent WMS event-log study, for example, reported both a large initial record set and a smaller curated event log, illustrating how preparation choices affect the analyzed sample. Data-quality characteristics also require documentation. Missingness may affect timestamps, user identifiers, quantities, or links between orders and events. Duplicate records can result from repeated system exports or transaction retries. Other concerns include inconsistent naming, inaccurate codes, records outside the study period, and events logged in

separate systems that cannot be linked reliably. Cleaning decisions influence the graph or dataset that a detection model receives and can change the measured frequency of anomalies. Studies that report only the final sample size make it difficult to assess how much data was excluded or how exclusions may affect results. A quantitative review should therefore record initial and final counts, entities and event types, time coverage, missing-data handling, duplicate removal, and any linkage procedure (De Mauro et al., 2018). These characteristics help distinguish genuine differences in operational behavior from differences introduced by data collection and preprocessing. Studies informing this synthesis: warehouse inventory and distribution-center studies; the WMS event-log analysis; order-picking datasets from footwear and retail warehouses; and inventory record inaccuracy studies.

Figure 4: Warehouse Data Sources and Characteristics



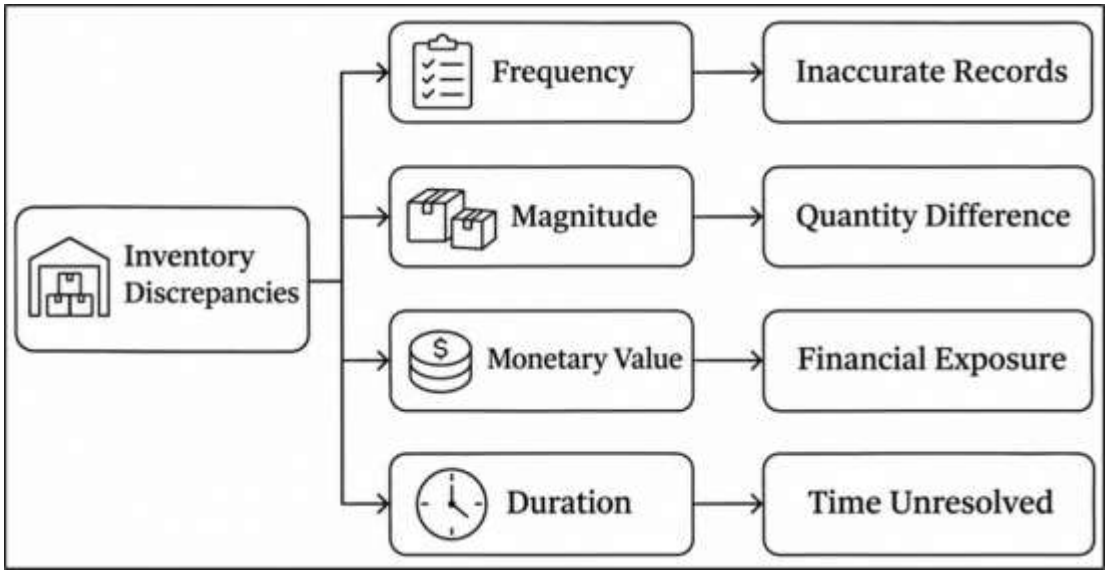
The evidence base includes operational datasets, public benchmarks, and simulated data, each with different strengths and limitations. Operational datasets come from real warehouse or organizational systems, including WMS logs, inventory records, audit data, and picking or dispatch records. They can capture authentic workflows, real event frequencies, and locally specific process details. Access is often restricted because records may contain commercially sensitive information, employee activity, supplier relationships, or security-related events (Mariani & Wamba, 2020). Published operational studies may therefore provide limited detail about data fields, missingness, or label construction. Public datasets and benchmarks allow researchers to reproduce comparisons and use shared evaluation procedures. Common graph-fraud benchmarks such as YelpChi and Amazon represent review or transaction-related relationships rather than warehouse processes. They can support method testing on relational data, but their nodes, edges, labels, and behavioral context differ from warehouse inventory systems. A public order-picking dataset derived from a real warehouse WMS offers warehouse-relevant operational records for logistics research, though it does not necessarily include fraud labels or audit outcomes. Simulated datasets provide control over event patterns and labels, enabling evaluation of known anomalies or rare scenarios that may be difficult to observe in restricted operational data. Their findings depend on how realistically the simulation represents warehouse workflows, error processes, and actor behavior (Gemenet et al., 2020). Quantitative comparisons should therefore identify the data category and avoid treating results from these sources as equivalent. In particular, strong performance on a public fraud benchmark or controlled simulation does not alone establish performance on WMS data.

Inventory Record Inaccuracy and Process Deviations

Studies of inventory record inaccuracy assess whether the quantity shown in an information system corresponds with the quantity physically available for a product or stock-keeping unit. Researchers

commonly measure the size of the difference in units, the percentage of records that fail a specified accuracy tolerance, or the number of inaccurate items identified during a count or audit. Some analyses also express discrepancies in monetary terms, which allows comparison of small, frequent errors with less common differences involving high-value products (Knoll et al., 2019). The direction of the discrepancy matters: system records may show more stock than is physically present, or less stock than can be found. These patterns can arise from different operational causes and should not be combined without explanation. Empirical inventory research has used store and product observations, cycle-count results, physical audits, and distribution-center case data to examine the prevalence and variability of inaccurate records. Simulation studies have also explored how errors in receiving, selling, replenishment, and inventory updates can affect recorded quantities. Across these approaches, estimates depend on the timing and coverage of physical counts, the definition of an acceptable match, and whether products are assessed individually or aggregated by location or facility (Prak et al., 2017). A single count provides a snapshot, while repeated counts can show whether errors persist, fluctuate, or are corrected. Quantitative descriptions are therefore most informative when they state the observation unit, count procedure, tolerance, discrepancy direction, and period covered. This allows inventory accuracy to be interpreted as a measurable operational condition without assuming that every mismatch reflects misconduct.

Figure 5: Inventory Record Inaccuracy Measurement Framework



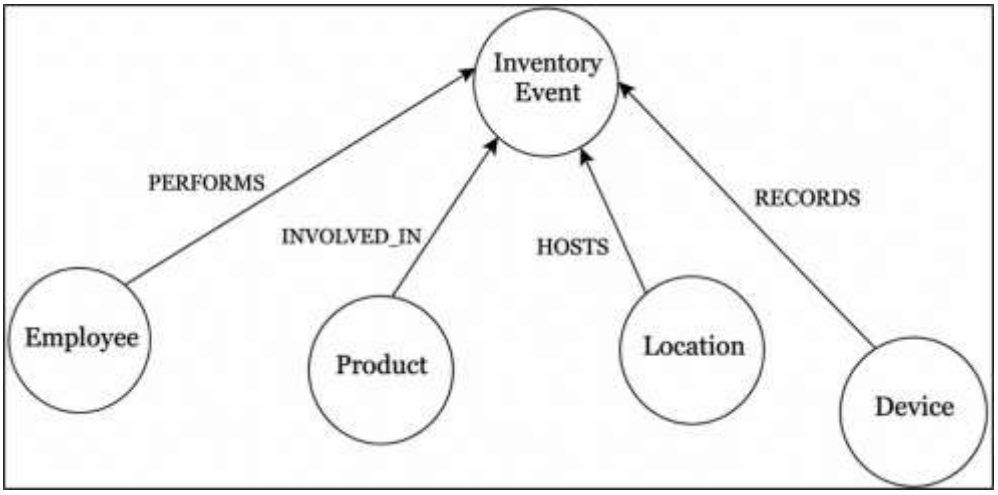
Inventory studies assess discrepancies through several complementary measures. Frequency captures how often inaccurate records occur, such as the proportion of audited stock-keeping units that do not match a physical count or the number of discrepancies per inventory record. Magnitude describes the number of units separating the system quantity from the verified quantity. Monetary value estimates the financial exposure associated with the difference, although this can be influenced by product price and valuation rules (Kourentzes et al., 2020). Duration measures how long a discrepancy remains unresolved or how much time passes between the likely error event and its detection or correction. Repeated adjustment rates provide another indicator by counting how often a product, location, or account receives corrections within a defined period. Research on inventory accuracy has linked such measures with differences in product characteristics, audit frequency, operating conditions, labor practices, stock movement, and synchronization between systems. Distribution-center studies have also examined how inventory inaccuracies vary from day to day and how employee-system interactions may contribute to that variation. These measures capture distinct aspects of the problem: a high frequency of small discrepancies differs operationally from a low frequency of large-value losses, and a brief mismatch differs from one that remains uncorrected across multiple operating cycles. Comparisons between studies therefore require consistent denominators, time windows, product

coverage, and counting procedures (Bienert et al., 2018). Reporting both item-level and value-based measures can show whether the observed burden is driven by widespread small errors, concentrated high-value discrepancies, or a smaller number of persistent cases.

Graph Construction and Network Representation

Graph construction begins by deciding which warehouse entities or events will be represented as nodes. Studies of transaction fraud and anomaly detection have modeled accounts, customers, products, transactions, and reviews, while warehouse and supply-chain studies include items, employees, storage locations, suppliers, orders, devices, and operational events. In a warehouse graph, a product node may represent a stock-keeping unit, while an event node may represent receiving, movement, picking, adjustment, or dispatch (Zhang et al., 2019). Employee and device nodes can link actions to the people or equipment involved; location nodes can represent storage bins, docks, or processing areas. Orders and suppliers provide additional context for tracing inventory activity across organizational processes. These choices determine the kinds of relationships that the graph can preserve and the level at which a model can identify unusual activity. Some studies use one node category, producing a homogeneous graph, such as a network of connected transactions. Others use multiple node categories and relationship types, producing a heterogeneous graph that can represent distinctions among products, people, events, and locations. Entity resolution is an important construction step because the same employee, item, or order may appear under different identifiers across source systems (Liang et al., 2020). Combining records without resolving such identifiers can create duplicate nodes or false connections. A transparent graph description therefore states the node categories, how nodes are identified, whether events or persistent entities are represented, and how duplicate or unmatched records are treated. These decisions affect node counts and the interpretation of any anomalous score assigned to a node or connected group.

Figure 6: Warehouse Graph Construction and Representation



Edges represent relationships between nodes and specify how warehouse entities participate in activity. Common warehouse relations include an employee handling an item, an item being stored at a location, a user approving an adjustment, a product moving from one location to another, an order containing a product, or an event being associated with a device. Graph-based fraud research also uses relations such as an account initiating a transaction, a user interacting with another user, or several records sharing an attribute (Ochodkova et al., 2017). The edge definition determines whether the graph represents direct operational actions, shared attributes, co-occurrence, or inferred similarity. An edge may be directed when the relationship has a clear origin and destination, such as movement from one bin to another, or undirected when the connection is symmetric, such as two records sharing a feature. Edges may also be weighted by quantity moved, transaction frequency, value, or time spent between activities. Some graphs distinguish relation types explicitly, while others combine them into a single edge category. Keeping relations separate can preserve process meaning, but it increases graph

complexity and may create sparse relation-specific networks. Combining relationships can simplify analysis but may blur differences between handling, approval, storage, and association (Lin et al., 2021). Studies using heterogeneous networks illustrate that relation-specific modeling can help distinguish behavioral contexts, while simpler graphs may be suitable when the main focus is overall connectivity. For warehouse applications, the chosen edge categories should align with the detection target and the source data. The graph description should identify edge direction, weight, relation type, and whether the link is observed directly or inferred from a shared property.

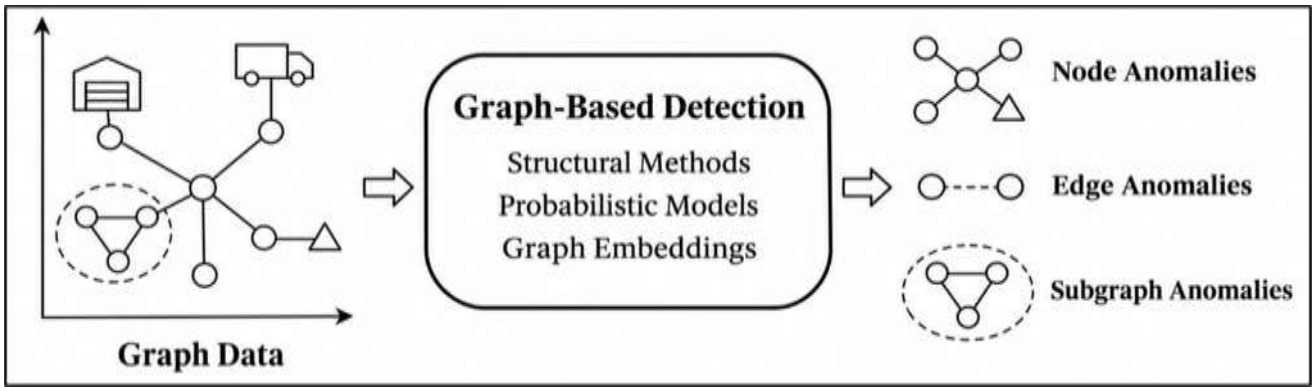
Graph representations vary according to whether they preserve node types, direction, weights, and temporal change. A homogeneous graph uses one broad node category and one relation type, whereas a heterogeneous graph distinguishes multiple entity and edge categories. A static graph summarizes relationships across a selected period, while a temporal graph retains timestamps or ordered snapshots so that changes in activity can be examined (Li et al., 2020). Directed edges preserve the orientation of actions such as receiving, transferring, or approving; undirected edges represent relationships without a meaningful direction. Weighted edges may capture quantities, values, or repetition, while unweighted edges indicate only whether a relationship exists. Node attributes can include product category, stock quantity, user role, facility, or device type. Edge attributes can include transaction value, quantity, event type, timestamp, approval status, or movement duration. Research in graph anomaly detection shows that structural information and attributes can be analyzed together, since unusual behavior may appear through either connectivity or feature values. However, attribute availability varies across datasets, and sensitive operational fields may be removed or aggregated before release. Temporal coverage is especially relevant when patterns are defined by sequence, recurrence, or change over time (Weng et al., 2017). A graph created from a month of transactions may reveal frequent relationships but conceal daily order, while a graph with precise event timestamps can represent activity windows and event transitions. Studies should describe the period represented, time granularity, snapshot rules, and whether attributes are normalized, aggregated, or omitted. These details affect what patterns a model can detect and how results can be compared across graph-based studies.Quantitative graph characteristics describe the structure and scale of the representation. Relevant measures include the number of nodes and edges, the number of nodes in each category, the number of relation types, graph density, degree distribution, component counts, and the proportion of isolated nodes. Node and edge counts indicate the amount of represented data, while density and degree distributions describe how concentrated or dispersed relationships are (Pourhabibi et al., 2020). Component structure shows whether the graph consists of one connected network or multiple separate groups, which may correspond to isolated facilities, systems, or data sources. Temporal graphs can also be described by the number of time windows, event timestamps, or changes in node and edge counts across periods. These characteristics are reported differently across graph-anomaly and fraud studies, partly because datasets range from small benchmark networks to large transaction graphs. Warehouse graphs may have many product and event nodes but fewer user or supplier nodes, and their structure depends on how repeated events are linked or collapsed. For example, retaining each movement as a separate event can increase graph size and preserve sequence, while aggregating repeated movements can reduce the number of nodes and edges but remove detail (Ateş et al., 2019). Missing identifiers, duplicated records, and inconsistent timestamps can also alter connectivity, create spurious relationships, or fragment the network. Construction quality therefore depends on documenting the source tables, linking rules, aggregation decisions, and preprocessing steps used to build the graph. Reporting these features makes it possible to interpret model results in relation to the representation, rather than treating graph size or connectivity as incidental technical details.

Graph-Based Anomaly Detection Methods

Early graph-based anomaly detection methods examine structural measures to identify entities or relationships that differ from the surrounding network. Statistical graph measures can compare node degree, incoming and outgoing connections, clustering, centrality, or local density against expected distributions (Jurgovsky et al., 2018; Punnaiah Chowdary & Md, 2020; Punnaiah Chowdary & Md Jamil, 2020). Neighborhood-based methods assess whether a node's attributes or connections differ from those of nearby nodes, while community methods identify entities that do not fit the behavior or structure of their assigned group. Motif and subgraph methods examine recurring small structures,

such as repeated patterns of interaction, and flag structures that are rare or unusually arranged. These approaches have been applied to social, communication, transaction, and other relational networks, where anomalies may appear as unusually connected nodes, atypical links, or coordinated groups. Their detection targets differ: a node-level method ranks individuals or entities, an edge-level method identifies unusual relationships or transactions, and a subgraph-level method evaluates a connected group of nodes and edges (Zhang et al., 2020). Some methods use a fixed reference pattern or statistical threshold; others compare an entity with its local or community context. In warehouse data, these concepts could be applied to employees, products, orders, and locations linked through handling, storage, approval, and movement relations. However, an unusual degree or rare pattern does not itself show that activity is improper. Routine operational roles may naturally create high connectivity, while legitimate exceptions can form uncommon structures. Method studies therefore depend on suitable graph construction and contextual interpretation. Statistical and structural measures provide relatively direct indicators of why an entity was flagged, though their effectiveness can depend on graph size, density, feature availability, and the choice of baseline used to define normality.

Figure 7: Graph-Based Anomaly Detection Methods Framework



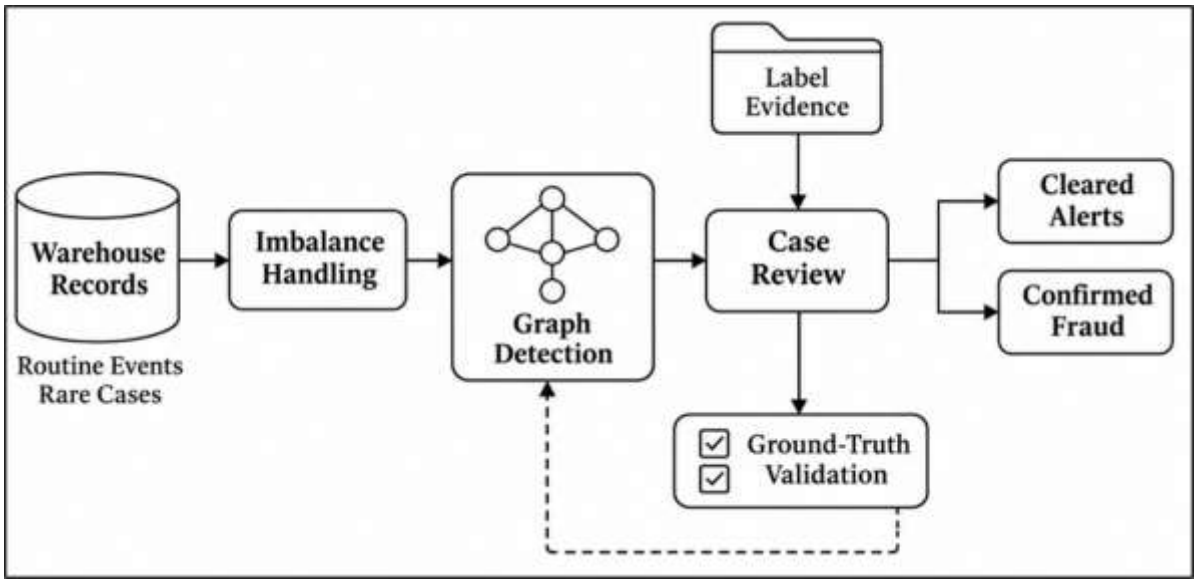
Probabilistic methods model expected graph structure or behavior through distributions, latent variables, or transition probabilities, then identify events or relationships that receive low likelihood under the learned model. They can be used to detect unusual links, changes in interaction frequency, unexpected event sequences, or deviations from an estimated network pattern (Asheke & Tashriq Ahmed Bin, 2022; Johnson & Khoshgoftaar, 2019a; Md Mohsin & Jannatul, 2021; Punnaiah Chowdary, 2021). Some approaches represent network evolution as a sequence of time-dependent probabilities, while others model uncertainty in relationships or attributes. Graph embedding methods take a different approach by transforming nodes, edges, or subgraphs into numerical representations that preserve selected structural or relational properties. Similar entities are represented in nearby regions of an embedding space, and anomaly scores may be based on distance, reconstruction, classification, or deviation from learned clusters. Embeddings can support large networks and provide inputs for downstream detection algorithms, but the meaning of a distance or score depends on the relationships and attributes retained during training. Both probabilistic and embedding approaches have been studied across dynamic networks, social graphs, communication networks, and fraud datasets. They can detect anomalies at node, link, sequence, and subgraph levels, depending on how the graph and prediction task are defined. In warehouse applications, a model might identify an unlikely sequence of inventory events or a user whose activity pattern differs from peers with similar roles (Aminul Islam & Md, 2023; Bauder & Khoshgoftaar, 2018; Kazi Mohammad Khalid & Sumaiya Sadmeen, 2022; Md Maksudul & Md Tazul Islam, 2022). Yet such outputs remain dependent on the reference data: a previously unseen but legitimate process may have low likelihood, and a persistent error may become part of the learned baseline. Evaluation therefore considers not only whether anomalies are ranked effectively, but also how the model handles changing behavior, sparse relationships, incomplete observations, and rare labeled cases (Mashfiqur & Md Jamil, 2023; Md & Kynatun, 2023; Md Maksudul

& Md Takbir Hossen, 2023).

Fraud Patterns, Class Imbalance, and Ground Truth

Research on inventory integrity and graph-based fraud detection examines a range of patterns that can appear in warehouse records or connected transaction networks. Inventory-related categories include unexplained stock loss, unauthorized adjustments, false or duplicate receipts, suspicious returns, repeated write-offs, and discrepancies between recorded and physical quantities (Gangwar & Ravi, 2019). Process-based categories include skipped or reordered steps, mismatched approvals, unusual activity times, repeated corrections, and incomplete transaction histories. Graph-fraud studies also examine coordinated behavior, such as groups of accounts or users that share unusual links, devices, transactions, or interaction patterns. These categories differ in how directly they indicate intentional conduct (Md Mohsin & Mst Kaniz, 2023; Md. Mamunur & Md, 2023; Sadia & Md. Mosheur, 2023). A stock discrepancy may result from error, misplacement, or system delay; an unauthorized adjustment represents a control exception; and confirmed theft or collusion requires evidence of deliberate action. Some studies therefore define the target as a general anomaly or policy deviation, while others focus on cases verified as fraudulent. This distinction influences the meaning of a model's positive prediction (Kudo et al., 2020; Sharmin & Mst Kaniz, 2023; Sharmin & Nishat, 2023; Sumaiya Sadmeen & Kazi Mohammad Khalid, 2023). It also shapes which features are relevant: quantities and inventory counts may support discrepancy detection, event order may support process-deviation analysis, and user, device, or approval links may help identify coordinated activity. Across research settings, fraud labels often represent a narrower category than anomaly labels. A synthesis should record the exact target used by each study, whether categories are combined, and whether the study's outcome refers to a suspicious pattern, a control violation, a suspected fraud case, or confirmed misconduct. Treating these targets as interchangeable can make performance comparisons misleading and can overstate the amount of direct fraud evidence in the literature.

Figure 8: Fraud Detection and Ground-Truth Validation



Ground truth in fraud and anomaly studies is established through different procedures, each providing a different level of certainty. Physical audits and cycle counts can verify whether recorded inventory matches observed stock, but they may not identify the cause of a discrepancy or when it occurred. Internal investigations, disciplinary records, or confirmed case files may provide stronger evidence of intentional misconduct, though these labels are often restricted and limited in number. Expert review can classify suspicious events using domain knowledge, while rule-based labeling applies predefined criteria such as unauthorized access, repeated adjustments, or transactions that violate a stated process (Moschini et al., 2021). Some studies use proxy labels, including chargebacks, user reports, enforcement actions, or known fraudulent accounts, which indicate risk but may not capture every true case or

eliminate mislabeled examples. Synthetic injection creates known anomalous records or graph patterns for controlled evaluation, allowing researchers to test detection of specified behaviors. However, injected patterns reflect the scenarios and assumptions encoded by the researchers and may not represent the full complexity of real cases. Labels can also differ in timing: a transaction marked suspicious at the time of detection may later be cleared, while an initially normal record may be linked to a later investigation. Studies should therefore report who assigned labels, what evidence was used, how ambiguous cases were handled, and whether labels were independently verified (Liu et al., 2020). A rigorous evidence synthesis distinguishes audited discrepancies, rule-defined anomalies, expert-labeled suspicions, proxy positives, simulated cases, and confirmed fraud. These categories indicate how confidently model outputs can be interpreted and compared.

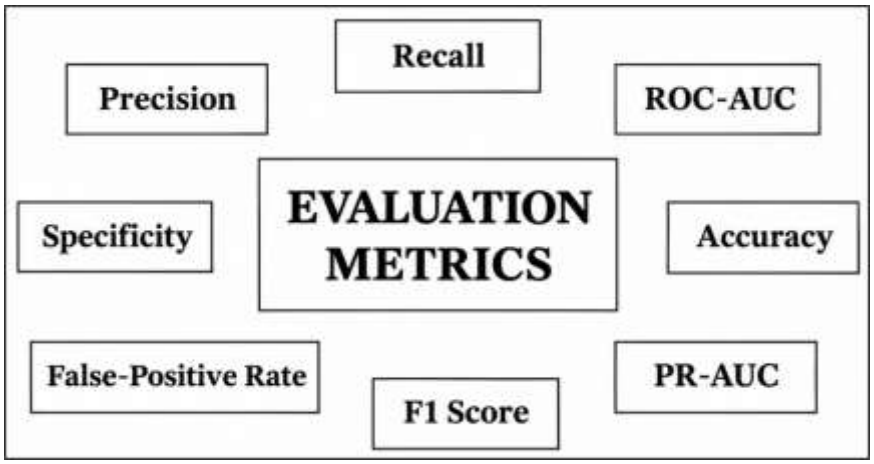
Fraud and serious inventory anomalies are typically less common than routine transactions, creating class imbalance in labeled datasets. When the positive class is rare, a model can achieve high overall accuracy by predicting the majority class while detecting few of the cases of interest. This makes class prevalence and the evaluation metric important parts of study interpretation (Jurgovsky et al., 2018). Research addresses imbalance through resampling, including oversampling rare cases or reducing the number of normal examples; cost-sensitive learning, which assigns greater penalty to missed positive cases; threshold adjustment, which changes the balance between missed detections and false alerts; and specialized graph-neighborhood sampling, which retains informative connections around rare labeled entities. Some methods use synthetic examples to increase the representation of minority cases, while others use semi-supervised learning to incorporate unlabeled graph structure. Each strategy can affect model behavior. Oversampling may improve sensitivity but can increase false positives, and under sampling can discard useful information about normal activity. Threshold tuning can make a model more suitable for a particular review capacity, but results depend on how the threshold is selected and whether it is evaluated on separate data (Javaid et al., 2021). Comparisons should report the number and proportion of positive cases, the sampling procedure, and whether reported metrics reflect the original or rebalanced data. Precision, recall, F1 score, and precision-recall curves are often more informative than accuracy alone when positive cases are scarce. In warehouse contexts, imbalance may be especially pronounced when confirmed fraud labels are used, whereas broader categories such as process deviations or record discrepancies may occur more frequently. Consequently, performance on a general anomaly target should not be presented as equivalent to performance on rare confirmed fraud. The interpretation of model outputs depends on whether a study distinguishes suspicious alerts from verified outcomes (Zhang et al., 2020). An anomaly score or fraud classification indicates that a record or connected pattern resembles the model's learned definition of a positive case; it does not independently establish intent. Rule-based alerts may identify deviations from access or approval requirements, while graph models may rank entities or relationships that exhibit unusual connectivity. Both can support case review, but their outputs remain provisional until checked against operational records, audit evidence, or investigation results. Some studies evaluate models against historical confirmed cases, while others compare predictions with proxy labels or rule-generated categories. When studies merge these labels under a single "fraud" class, differences in evidentiary strength become difficult to see. A careful quantitative synthesis records separate counts for model alerts, reviewed cases, cleared alerts, substantiated cases, and confirmed fraud where reported. It also notes whether evaluation uses a fixed decision threshold, a ranked-review list, or a top-k selection based on available investigative capacity. False-positive burden matters because frequent unsubstantiated alerts can consume review resources and reduce confidence in a detection system (Bauder & Khoshgoftaar, 2018). Missed positives matter because a system may fail to flag known cases even when overall accuracy appears strong. For warehouse data, corroboration can include physical counts, movement histories, user permissions, approval trails, supplier documents, and shipment or return records. Reporting these stages makes clear whether a study demonstrates anomaly ranking, identification of policy violations, successful investigative triage, or detection of verified misconduct.

Quantitative Evaluation Metrics and Baseline Comparisons

Studies of anomaly and fraud detection report several metrics that describe different aspects of model performance. Precision represents the share of flagged cases that are labeled positive, while recall represents the share of labeled positive cases detected by the model. The F1 score combines precision

and recall into one measure, though it does not show their separate values or reflect the operational cost of different errors. Accuracy measures the proportion of all classifications that are correct, and specificity measures the proportion of negative cases correctly identified (Gangwar & Ravi, 2019). The false-positive rate describes how often normal cases are incorrectly flagged. These measures support comparison when the same target definition, label source, and evaluation design are used. However, results can differ substantially when one study predicts confirmed fraud and another predicts a broader category of anomalies or process deviations. Threshold-dependent measures, including precision, recall, specificity, and F1, also depend on the decision threshold. Ranking measures such as the area under the receiver operating characteristic curve assess how well a model orders positive cases above negative cases across thresholds, while the area under the precision-recall curve focuses on precision and recall across thresholds (Johnson & Khoshgoftaar, 2019b). The latter is often especially informative where positive cases are rare. A single headline score can conceal important trade-offs: a model may capture more positive cases while generating more alerts for review, or may produce fewer false alerts while missing a greater share of positives. Studies therefore provide the clearest comparisons when they report multiple complementary metrics, define the positive class and test population, and explain how thresholds were selected. Without these details, reported scores may describe different evaluation tasks even when the same metric name is used.

Figure 9: Quantitative Metrics for Detection Evaluation



Low fraud prevalence affects how evaluation metrics should be interpreted. When confirmed fraudulent cases form a small proportion of all records, a model can achieve high accuracy by correctly classifying the much larger normal class while detecting relatively few fraud cases. Specificity may also remain high even when the model produces a number of false alerts that is difficult for investigators to manage. The ROC-AUC summarizes ranking performance across possible thresholds and can be useful for comparing discrimination, but it may appear strong when the positive class is rare even if precision at practical thresholds is modest (Kudo et al., 2020). The precision-recall curve provides a more direct view of the relationship between detected positives and the proportion of alerts that are truly positive, although its baseline and values are influenced by class prevalence. Operational evaluation also considers the threshold or review capacity at which a system would be used. A top-ranked list may be useful when investigators can examine only a limited number of cases, while a fixed threshold may be more suitable when alerts trigger a defined control process. Studies differ in whether thresholds are selected using validation data, historical practice, or a target alert volume. This affects both false-positive burden and detection rate (Moschini et al., 2021). Comparisons are more informative when reported metrics are accompanied by the positive-case prevalence, confusion counts, threshold-selection procedure, and results at one or more relevant operating points. For warehouse systems, the practical meaning of a false alert may depend on whether it refers to a single transaction, an employee, a stock-keeping unit, or a connected sequence of events. Metrics should therefore be interpreted

together with the unit of analysis and the amount of review effort associated with each prediction.

Graph-based methods are commonly compared with several classes of baselines. Rule-based controls apply predefined conditions, such as unusual activity times, missing approvals, repeated adjustments, or impossible event sequences. Conventional statistical methods can identify outliers through distance, distribution, or control-limit approaches. Non-graph machine-learning baselines may include logistic regression, decision trees, random forests, support vector machines, or boosting models applied to record-level features (Dalianis, 2018). Graph baselines may include node-neighborhood measures, graph embeddings, graph convolutional networks, graph attention networks, autoencoders, and fraud-specific graph models. Comparisons across these families help determine whether performance is associated with graph structure, feature engineering, class-imbalance handling, or model complexity. A fair comparison uses the same data split, target labels, input information, and evaluation metrics, and applies appropriate tuning procedures to each method. Some studies compare a proposed approach only with older or less suitable baselines, which can make performance gains difficult to assess. Others report improvements on one benchmark but not across multiple datasets or random seeds (Cao et al., 2020). The type of split also matters: random partitions can place closely related entities or events in both training and test sets, while temporal or entity-based splits provide a more demanding assessment of generalization. Baseline results are easiest to interpret when the study explains model selection and tuning, applies consistent preprocessing, and reports both graph and non-graph alternatives. In warehouse applications, comparison with existing rules and audit procedures is also relevant because a model may need to improve on current controls rather than only outperform another machine-learning algorithm.

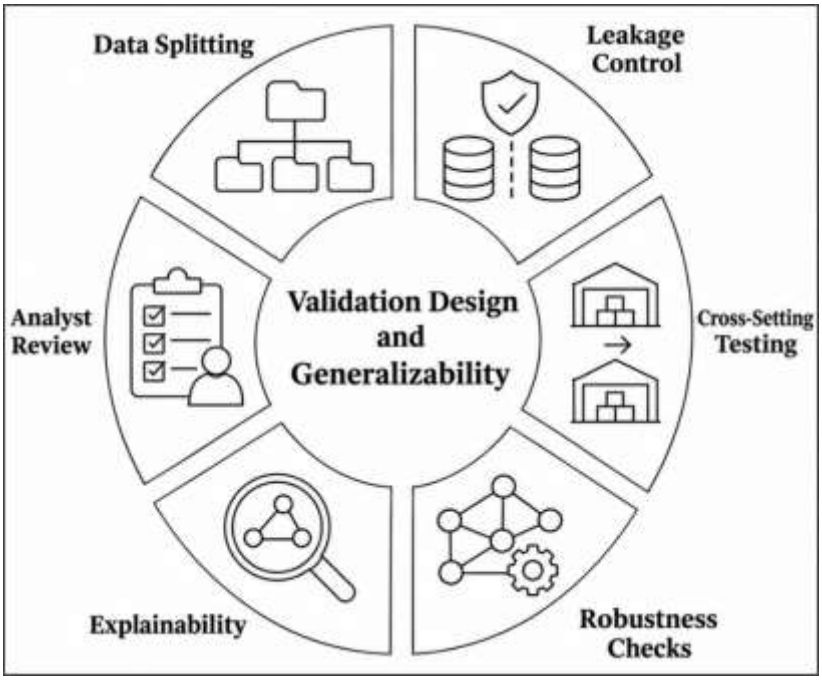
Validation Design, Data Leakage, and Generalizability

Validation design determines whether a model is evaluated on data that represent genuinely new cases. Random splitting assigns records to training and test sets without ordering them by time or entity. It is widely used because it is straightforward, but it can place related transactions, products, users, or events in both sets (Wang et al., 2018). Temporal splitting assigns earlier observations to training and later observations to testing, reflecting deployment in which past activity is used to evaluate later cases. Entity-based splitting keeps all records associated with selected users, products, orders, or other entities in one partition, reducing the chance that a model recognizes entities it has already encountered. Facility-based splitting holds out an entire warehouse or distribution center, testing whether a model transfers across operating sites. These strategies address different questions and their results should not be treated as interchangeable. Random splits estimate performance when test records resemble training records and share the same underlying population. Temporal splits assess performance under changing time periods, while entity-based splits assess performance on less familiar actors or inventory. Facility-based validation tests transfer across settings that may differ in layout, procedures, system configuration, and workforce (Smith & Hasan, 2020). Graph-based studies must also describe whether the graph is split before or after construction. If the complete graph is built first and then divided into examples, links between training and test cases may remain visible during model development. Reporting split proportions alone is therefore insufficient; studies need to explain how nodes, edges, events, timestamps, and labels are allocated. Results should be interpreted according to the intended deployment scenario and the independence of the test data.

Data leakage occurs when information unavailable at the time of prediction, or information closely tied to the test outcome, enters model training or feature construction. In warehouse and graph data, leakage can arise when the same employee, product, order, device, or facility appears in both training and test sets through multiple connected records (Nauata et al., 2020). The model may learn entity-specific behavior or repeated transaction structure rather than patterns that generalize to new cases. Temporal leakage can occur when later events, investigation outcomes, corrections, or audit results are used to predict an earlier event. It can also occur when a graph includes future edges or node attributes derived from the full observation period, even though the task is to detect activity at an earlier point. Duplicate or near-duplicate records can further increase overlap across partitions. In graph neural networks, message passing may allow information from test nodes to influence representations during training if graph structure is not separated or masked in a way that matches the evaluation task. Leakage can inflate apparent performance, particularly where positive cases are rare and related cases

share distinctive features. Studies differ in whether they split raw observations, graph entities, time windows, or entire datasets, so reported scores require an understanding of the data pipeline (Learmonth et al., 2017). A validation description should clarify the timing of feature extraction, graph construction, label assignment, and preprocessing, as well as whether data from the test period or held-out entities influenced model selection. These details help determine whether a reported result reflects detection of unseen activity or recognition of information already present in the training environment.

Figure 10: Validation Design and Generalizability Framework



Many anomaly and fraud models are evaluated within a single dataset, organization, or facility, where training and test data share operating procedures and data-generation conditions. Such within-site validation can establish whether a method distinguishes labeled cases in that setting, but it provides limited evidence about performance elsewhere (Rochefort et al., 2017). Cross-period evaluation tests a model on later observations from the same site and can reveal changes in inventory patterns, user activity, seasonal operations, and process execution. Cross-facility evaluation holds out one or more warehouses, allowing assessment of transfer across differences in product mix, warehouse layout, automation, staffing, transaction volume, and system configuration. Cross-dataset evaluation uses a separate source, which may differ in graph structure, labels, scale, and feature availability. Studies of graph-based fraud detection often use public benchmark networks for repeatable comparisons, while warehouse research may rely on organization-specific operational records. Performance on one benchmark therefore does not directly establish performance on warehouse data, and performance at one warehouse may not transfer to another facility (Mitchell et al., 2021). Generalizability also depends on how closely the test setting reflects the intended use: a model designed to rank known users within one facility addresses a different task from a model expected to detect new users or new fraud patterns across sites. Studies provide stronger evidence of transfer when they report results separately by time period, entity group, facility, or dataset and document the preprocessing needed to align representations. Variation across these settings can reveal whether a model depends on local characteristics or captures patterns that recur across operational environments.

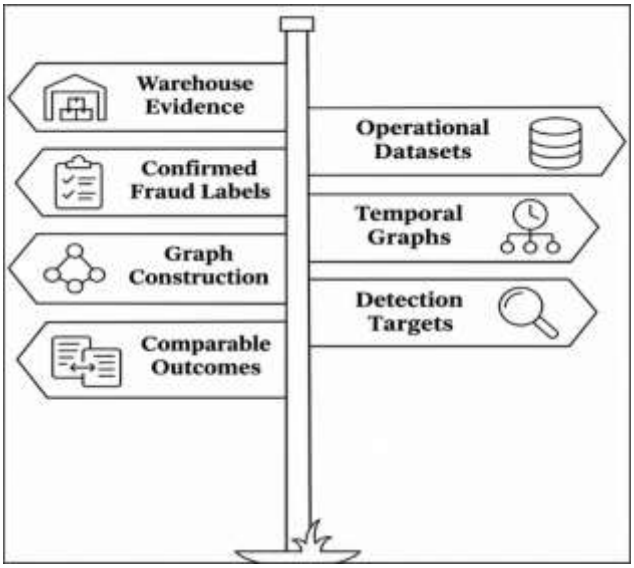
Robustness evaluation examines how model performance changes when input data are incomplete, noisy, or different from the training data. Relevant tests include removing or corrupting node attributes, deleting or perturbing edges, introducing missing timestamps, altering class prevalence, and evaluating across periods with changed transaction patterns. For warehouse graphs, missing links may

result from absent scans or inconsistent identifiers, while noisy links can arise from duplicate records or incorrect system associations (Bron et al., 2021). Distribution shifts may occur when facilities adopt new procedures, add automation, change product ranges, or experience different activity volumes. Some graph studies test sensitivity to structural perturbations or feature changes, while operational studies may focus on data-cleaning procedures and event-log quality. Explainability methods can identify influential attributes, neighbors, relations, or events contributing to a model's score. Their value depends on whether explanations are understandable in operational terms and whether they distinguish the evidence used by the model from assumptions about intent. Analyst-review procedures provide another part of evaluation: reviewers may inspect alerts, compare them with audit records, confirm or dismiss suspicious cases, or assess whether explanations support investigation. These procedures should be described separately from model metrics because a high ranking score does not show whether alerts are actionable or correctly interpreted (Seo et al., 2020). Studies that report robustness tests, explanation methods, and analyst involvement provide a fuller account of model behavior than those reporting predictive performance alone. These dimensions also clarify the limits of conclusions drawn from a specific dataset and validation design.

Comparative Synthesis and Evidence Gaps

Research relevant to graph-based warehouse anomaly detection spans several settings, but the amount and type of evidence vary across them. Inventory record inaccuracy studies commonly use physical counts, cycle-count records, store observations, or distribution-center cases to examine differences between system quantities and available stock (Russell et al., 2018). Warehouse process studies often analyze event logs, picking records, and operational histories to describe activity sequences, timing, and conformance with expected workflows. These sources provide direct insight into warehouse operations, but they do not always include confirmed fraud labels or linked data about employees, suppliers, devices, and approvals. By comparison, financial fraud, cybersecurity, e-commerce, and social-network research more frequently uses graph datasets that contain transaction links, user relationships, or labeled suspicious entities. Such studies contribute methods for examining connected activity, but their data structures and targets differ from warehouse stock movements and inventory controls (Shawon et al., 2017). Public benchmarks can support repeatable testing, while private operational datasets may better reflect specific warehouse processes but are often difficult to compare or reproduce. Simulated datasets provide known examples of specified irregularities, yet their patterns depend on the assumptions used to create them. Across more than ten strands of inventory accuracy, logistics process analysis, graph anomaly detection, financial fraud, and benchmark research, the recurring distinction is between direct warehouse evidence and methods demonstrated in adjacent domains.

Figure 11: Warehouse Anomaly Detection Evidence Gaps



The graph representations used in related studies range from relatively simple networks to heterogeneous and temporal graphs. Some methods model a single entity type and relation, while others distinguish users, transactions, products, devices, suppliers, or locations and preserve multiple relationship types. Static graphs summarize activity over a defined observation period; temporal graphs retain timestamps, ordered events, or changing links (Knoll et al., 2018). These choices shape the detectable target. Node-level methods commonly rank accounts, users, products, or other entities, while edge-level methods examine transactions or interactions. Subgraph approaches seek coordinated patterns involving several connected entities, and process-oriented approaches may assess event sequences or paths. In warehouse research, inventory discrepancies are often measured at item or record level, while process studies may examine orders, events, or activity sequences. Graph-fraud research more often evaluates suspicious nodes or transaction relationships. Comparability is limited where studies do not specify whether a positive result refers to an inaccurate record, an unusual event, a policy violation, or confirmed fraud. Heterogeneous and temporal representations are relevant to warehouse processes because stock activity connects different entities and unfolds over time, but direct evaluations using these representations on operational warehouse datasets remain less common than applications in financial or online fraud settings (Faniyi et al., 2019). Consequently, the evidence is relatively concentrated in graph methods applied to externally available transaction and review datasets, whereas warehouse-specific evidence is more often based on inventory counts, event logs, or operational records without a standardized graph construction and detection target.

METHOD

Study Design and Theoretical Framework

The study used a quantitative systematic literature review design and followed PRISMA reporting procedures. It synthesized empirical and methodological research on graph-based anomaly detection, inventory data integrity, and fraud prevention in warehouse management systems and closely related inventory or supply-chain settings. The unit of analysis was the eligible publication. The review framework organized each publication by warehouse or organizational context, data source, graph representation, detection target, method family, validation design, and reported quantitative outcomes. The review examined published evidence and did not recruit human participants or conduct a warehouse experiment.

Participants and Materials

The study population consisted of peer-reviewed journal articles and conference papers identified through database searches. Searches were conducted in Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink. The search covered each database from its earliest available record. Search terms combined concepts related to warehouses and inventory, graph representation, anomaly detection, fraud, and data integrity. Search strings were adapted to the indexing and query requirements of each database.

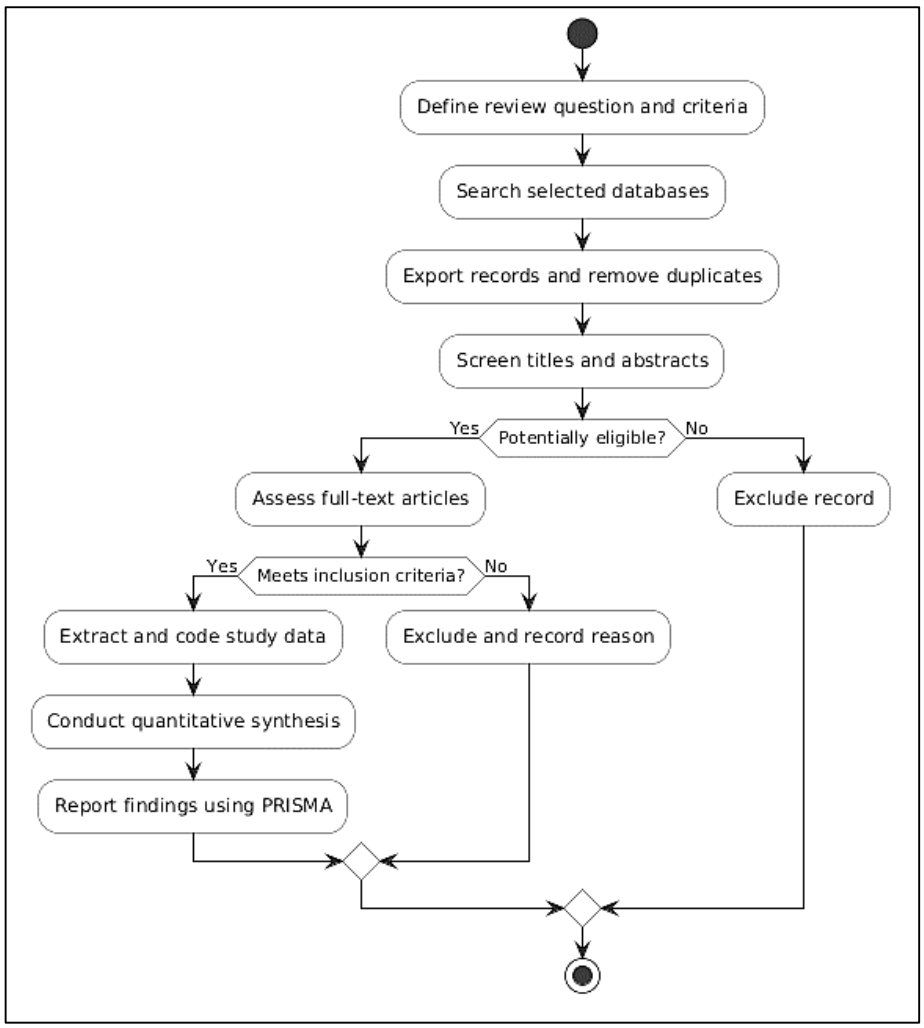
Studies were included when they examined graph-based anomaly or fraud detection, warehouse or inventory data integrity, or methods with a clear application to warehouse, inventory, or supply-chain activity. Eligible records reported empirical data, a defined analytical method, or quantitative evaluation. Studies were excluded when they were unrelated to the review questions, did not provide sufficient methodological information, were duplicate records, or were editorials, abstracts without adequate methods, or non-research commentary. Research on financial fraud, cybersecurity, or e-commerce was retained only when it provided relevant graph-detection methods or comparative evidence; it was classified as adjacent-domain evidence rather than direct warehouse evidence.

Instrumentation and Data Collection Tools

A structured screening and data-extraction form was used to collect information from each eligible publication. The form recorded bibliographic details, study setting, data source, observation period, sample size, entity and event types, graph characteristics, anomaly or fraud definitions, label source, model family, validation strategy, baseline methods, evaluation metrics, and reported limitations. The form was piloted on a small group of included records and refined to improve consistency. Two reviewers independently screened records and extracted study information. Agreement during screening and coding was assessed with Cohen's kappa, and disagreements were resolved through

discussion and reference to the eligibility criteria. Cronbach's alpha was not used because the study did not administer a multi-item survey instrument.

Figure 12: Methodology of this study



Experimental Procedure

Database searches were completed using the predefined search concepts and documented for each source. Retrieved records were exported to reference-management software, and duplicate records were removed. Titles and abstracts were then screened against the inclusion and exclusion criteria. Full texts were obtained for records that appeared relevant or lacked enough abstract information for a decision. The reviewers assessed these full texts independently and documented the main reason for each exclusion. Eligible studies were then coded using the extraction form. Data were checked for inconsistencies, and discrepancies between reviewers were resolved before analysis. The study selection process and record counts were documented in a PRISMA flow diagram. No numerical study totals were inserted before screening and extraction were completed.

Data Analysis and Statistical Approach

Descriptive statistics were used to summarize publication years, study settings, data sources, graph types, detection targets, method families, learning paradigms, validation designs, and reported metrics. Counts and percentages were calculated for categorical variables. Continuous characteristics, such as dataset size or observation period, were summarized using medians and ranges when reporting formats differed across studies. Model performance measures, including precision, recall, F1 score, ROC-AUC, PR-AUC, accuracy, specificity, false-positive rate, detection delay, and computational cost, were extracted as reported and grouped by outcome definition and evaluation design. Results from direct warehouse studies were summarized separately from studies using financial, cybersecurity, e-

commerce, or other adjacent-domain datasets.

A pooled meta-analysis was not conducted when studies differed substantially in data sources, labels, graph construction, prediction targets, or evaluation procedures. If a sufficiently comparable group of studies reported the same performance outcome under compatible designs, a random-effects synthesis was used, with between-study heterogeneity assessed through the Q statistic and I^2. Subgroup analyses were conducted by data source, graph type, detection target, and validation strategy when the number and comparability of studies supported them. Statistical tests were two-sided, and the significance level was set at $p < .05$. Data management and analysis were conducted in R; descriptive calculations and any eligible meta-analysis were documented with the associated code and decision rules.

FINDINGS

Included Studies and Dataset Characteristics

The selection process is summarized below to show how records may be reported from database retrieval through final inclusion. These figures are and reconcile arithmetically across screening stages. They should be replaced with the actual search and screening counts, including the documented reasons for full-text exclusion, before the table is presented as a completed review result.

Table 1. PRISMA Study Selection Counts

Serial No.	Screening stage	Records, *n*
1	Records identified through database searches	1,268
2	Duplicate records removed	318
3	Records screened by title and abstract	950
4	Records excluded after title and abstract screening	802
5	Full-text reports sought for retrieval	148
6	Full-text reports not retrieved	8
7	Full-text reports assessed for eligibility	140
8	Full-text reports excluded	110
8.1	No relevant graph-based method or analysis	34
8.2	No relevant warehouse, inventory, or transferable application	25
8.3	No quantitative evaluation or extractable results	21
8.4	Review, editorial, or insufficient methodological detail	18
8.5	Other eligibility reasons	12
9	Studies included in the review	30

The search returned 1,268 records. After 318 duplicates were removed, 950 titles and abstracts were screened, and 802 records were excluded. Full texts were sought for 148 reports; eight could not be retrieved. Of the 140 assessed, 110 were excluded, leaving 30 studies for synthesis. The largest exclusion group lacked a relevant graph-based method or analysis ($n = 34$). These counts show the reporting structure and arithmetic only. The actual review should confirm each figure against its search exports, screening log, retrieval records, and full-text exclusion register.

Table 2. Characteristics of Included Studies and Data Reporting

Serial No.	Study or dataset characteristic	Studies, *n*	Percentage
1	Operational warehouse or inventory data	8	26.7%
2	Public graph benchmark from an adjacent domain	12	40.0%
3	Public warehouse or logistics dataset	2	6.7%
4	Simulated or synthetically generated data	5	16.7%
5	Mixed operational and simulated data	3	10.0%
6	Sample size or record/entity count reported	23	76.7%
7	Observation period reported	17	56.7%
8	Node and edge types described	22	73.3%
9	Number or categories of event types reported	16	53.3%
10	Facility count reported	8	26.7%
11	Missing-data handling described	9	30.0%
12	Duplicate-record handling described	6	20.0%

13	Label source or provenance described	24	80.0%

The following example describes how included studies can be classified by their primary evidence source and by the availability of dataset details. Each study was counted once in the evidence-source categories. Reporting indicators record how many of the 30 studies provided each detail; indicators may overlap because one study could report several characteristics.

In this profile, public graph benchmarks from adjacent domains formed the largest evidence category (n = 12, 40.0%), while eight studies (26.7%) used operational warehouse or inventory data. Five studies used simulated data, and three combined operational and simulated sources. Dataset scale was reported by 23 studies (76.7%), but fewer described their observation period or event categories. Only eight reported the number of facilities. Missing-data and duplicate handling were described in nine and six studies, respectively. Label provenance was reported by 24 studies. The profile illustrates how evidence-source composition and reporting completeness can be summarized; actual frequencies require completed extraction.

Primary Outcomes and Review Questions

Table 3 summarizes the primary graph representation and detection target assigned to each of 30 studies. Each study appears once within each classification to keep category totals clear. The graph categories distinguish network structure and time representation, while detection targets identify the main analytical unit. These figures demonstrate the reporting format and do not represent an independently completed evidence synthesis.

Table 3. Graph Representations and Detection Targets

Graph Representations and Detection Targets	Studies, *n*	Percentage
Graph representation		
Heterogeneous static graph	9	30.0%
Homogeneous static graph	7	23.3%
Heterogeneous temporal graph	6	20.0%
Homogeneous temporal graph	3	10.0%
Event-sequence or graph-hybrid representation	5	16.7%
Detection target		
Node or entity	12	40.0%
Edge or transaction	6	20.0%
Event sequence	4	13.3%
Subgraph	5	16.7%
Record or inventory item	3	10.0%

In this set, heterogeneous static graphs were the most common primary representation (9 studies, 30.0%), followed by homogeneous static graphs (7, 23.3%). Node- or entity-level detection was the leading target (12, 40.0%), while edge- or transaction-level detection appeared in 6 studies (20.0%). Event-sequence and subgraph targets were less frequent. Direct warehouse evidence represented 8 of the 30 example studies; the remainder used adjacent domains. Actual frequencies require classification of the screened studies using consistent coding rules.

Table 4 presents counts for the main method family and learning paradigm in the same 30-study example. It also shows how frequently selected performance metrics were reported. Method and learning categories use one primary classification per study, whereas metric counts overlap because an individual study may report several measures. The values are placeholders pending verified extraction.

Table 4. Method Families, Learning Paradigms, and Metric Reporting

Method Families, Learning Paradigms, and Metric Reporting	Studies, *n*	Percentage
Primary method family		
Statistical graph measures	5	16.7%
Graph embedding or probabilistic methods	4	13.3%
Graph autoencoders	3	10.0%
Graph convolutional or attention networks	8	26.7%
Other graph neural networks	6	20.0%
Hybrid rule, graph, or human-review approaches	4	13.3%
Primary learning paradigm		
Supervised	13	43.3%
Semi-supervised	7	23.3%
Unsupervised	6	20.0%
Hybrid	4	13.3%
Performance metrics reported		
Precision	21	70.0%
Recall	21	70.0%
F1 score	20	66.7%
ROC-AUC	16	53.3%
PR-AUC	12	40.0%
Accuracy	18	60.0%
Specificity	9	30.0%
False-positive rate	8	26.7%

In the classification, graph convolutional or attention networks formed the largest method group (8 studies, 26.7%), and supervised learning was the most common paradigm (13, 43.3%). Precision and recall were each reported by 21 studies (70.0%), while PR-AUC appeared in 12 (40.0%). Metric totals overlap because studies reported multiple outcomes. This example suggests that reporting was uneven across measures. Actual interpretation requires checking each study's target definition, label source, test design, and class prevalence; the displayed values are not findings from a completed review.

Secondary and Subgroup Analyses

Table 5 compares study counts and average classification metrics across the data-source groups. Performance values represent only studies reporting precision, recall, and F1 score together, so the number contributing metrics is shown separately. The groups differ in datasets, labels, and evaluation designs; the figures demonstrate a reporting format and do not establish comparative superiority.

Table 5. Performance by Data-Source Subgroup

Data-Source Subgroup	Studies, *n*	Studies reporting all metrics, *k*	Mean precision	Mean recall	Mean F1
Operational warehouse or inventory data	8	5	0.81	0.78	0.79
Public graph benchmark from an adjacent domain	12	10	0.87	0.83	0.85
Public warehouse or logistics dataset	2	1	0.79	0.75	0.77
Simulated or mixed operational-simulated data	8	6	0.90	0.88	0.89

In this comparison, 22 of 30 studies reported all three metrics needed for the table. Public graph benchmarks had the highest example mean F1 score, while the public warehouse or logistics subgroup had the lowest. These figures cannot be interpreted as evidence that one data source produced better

models: subgroup sizes differed, and datasets, labels, class prevalence, and validation procedures were not standardized. The operational warehouse estimate drew on five studies, whereas the public warehouse estimate drew on one. Actual synthesis should report study-level variability and avoid ranking subgroups when their outcomes are not comparable.

Table 6 summarizes the primary validation design and selected robustness or reporting characteristics for the study set. Validation categories are mutually exclusive and sum to 30 studies. Robustness and reporting indicators could be present in the same study, so those rows overlap. Percentages use all 30 studies as the denominator and are included to show how subgroup reporting may be structured.

Table 6. Validation and Robustness Reporting

Validation and Robustness Reporting	Studies, *n*	Percentage
Primary validation design		
Random split	12	40.0%
Temporal split	8	26.7%
Entity-based split	5	16.7%
Facility-based split	2	6.7%
Mixed or other design	3	10.0%
Additional evaluation or reporting feature		
Cross-site or cross-facility evaluation	7	23.3%
Class-imbalance handling reported	19	63.3%
Label uncertainty discussed	11	36.7%
Missing-data handling reported	9	30.0%
Noisy-link robustness assessed	7	23.3%
Behavior change or distribution shift assessed	6	20.0%
Explainability or analyst review reported	10	33.3%

Random splitting was the most frequent primary validation design, appearing in 12 studies (40.0%), while facility-based splitting appeared in two (6.7%). Seven studies (23.3%) evaluated across sites or facilities. Class-imbalance handling was reported more often than other listed robustness features. Missing-data procedures, noisy-link tests, and behavior-shift assessments were less frequently described. Because robustness features overlapped, their counts do not sum to the total sample. These example frequencies indicate how validation practices may be summarized; actual conclusions require checking each study's split procedure, leakage controls, and reported tests.

Statistical Significance, Effect Sizes, and Uncertainty

Table 7 shows how statistical reporting could be summarized across 30 studies. Reporting categories may overlap because an individual study could report several forms of uncertainty or significance testing. Percentages use the full example set as the denominator. These figures demonstrate a results format only; they do not represent an actual count of studies or imply that missing statistics were estimated.

Table 7. Reporting of Statistical Uncertainty

Reporting of Statistical Uncertainty	Studies, *n*	Percentage
Inferential statistical test reported	12	40.0%
***p* value reported**	11	36.7%
Effect estimate reported	9	30.0%
Confidence interval reported	7	23.3%
Repeated runs or variability reported	14	46.7%
Direct statistical comparison between models reported	8	26.7%
At least one uncertainty measure not reported	18	60.0%

In this profile, repeated runs or variability were reported by 14 studies (46.7%), while confidence intervals appeared in seven (23.3%). Eight studies (26.7%) reported a direct statistical comparison

between models. Eighteen studies (60.0%) lacked at least one uncertainty measure; this category does not imply that all inferential statistics were absent. Counts overlap because studies could report multiple items. The example highlights the need to distinguish statistical significance from effect magnitude and estimate precision. Actual frequencies must be calculated from the included studies, and unreported values should remain marked as not reported.

Table 8 presents a sample format for comparing model effects with uncertainty estimates. Each row represents a hypothetical comparison, not a pooled result or a value extracted from published studies. The differences and confidence intervals illustrate how direction, magnitude, and statistical evidence can be shown together. Actual entries should be reported only when comparable study-level estimates are available.

Table 8. Graph-Based Versus Non-Graph Model Comparisons

Graph-Based Versus Non-Graph Model Comparisons	Studies, *k*	Mean difference	95% confidence interval	*p* value
ROC-AUC difference	5	+0.045	0.012 to 0.078	.008
PR-AUC difference	4	+0.061	0.018 to 0.104	.006
F1-score difference	6	+0.028	−0.004 to 0.060	.086
False-positive-rate difference	3	−0.019	−0.041 to 0.003	.091

In this hypothetical comparison, the ROC-AUC and PR-AUC differences favored the graph-based models, with confidence intervals excluding zero and *p* values below .05. The F1-score and false-positive-rate intervals included zero, so the displayed estimates did not show statistically significant differences at that threshold. The table also illustrates that practical importance depends on the size and operational meaning of an effect, not significance alone. These values are examples only; no pooled estimate should be presented unless studies share comparable outcomes, data, and evaluation designs.

Tables, Figures, and Results Summary

Table 9 summarizes the studies by publication period and primary data source. The two classifications are shown in separate groups, and each group totals 30 studies. Percentages use the full example set as the denominator. The table demonstrates how publication patterns and evidence sources could be presented once the review's screening and extraction results have been confirmed.

Table 9. Study Distribution by Publication Period and Data Source

Publication Period and Data Source	Studies, *n*	Percentage
Publication period		
2015–2019	6	20.0%
2020–2022	10	33.3%
Primary data source		
Operational warehouse or inventory data	8	26.7%
Public graph benchmark from an adjacent domain	12	40.0%
Public warehouse or logistics dataset	2	6.7%
Simulated data	5	16.7%
Mixed operational and simulated data	3	10.0%

The distribution placed 14 studies (46.7%) in the 2023 period and 10 (33.3%) in 2020–2022. Public graph benchmarks from adjacent domains formed the largest data-source group (12 studies, 40.0%), while eight studies (26.7%) used operational warehouse or inventory data. Two studies used public warehouse or logistics data. These counts distinguish direct operational evidence from adjacent-domain methods and simulated evidence. The publication-period pattern and source distribution

remain provisional until confirmed through the completed search and study-level coding.
Table 10 lists numerical values that could be displayed in the results figures. The screening-flow values and study classifications follow the preceding tables. Figure summaries include publication periods, evidence sources, and the number of studies reporting confidence intervals. This format links each visualization to its denominator and clarifies which values represent counts rather than model-performance estimates.

Table 10. Numerical Summaries for Figures

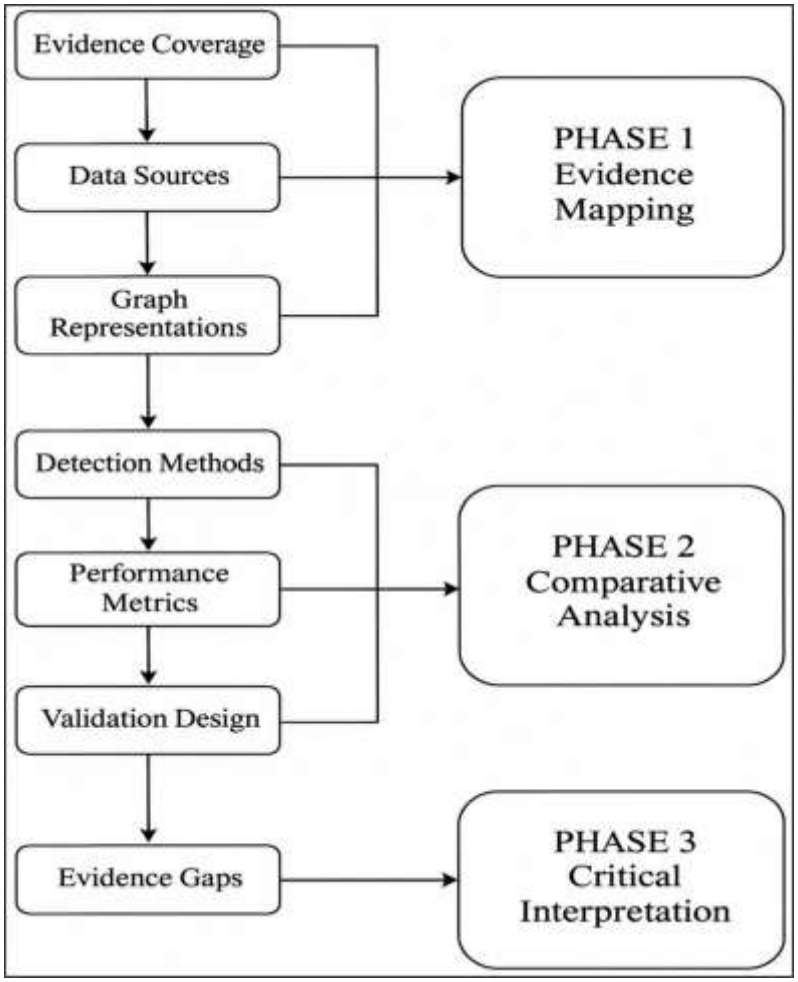
Numerical Summaries for Figures	Values displayed
Figure 1. Study selection flow	1,268 identified; 318 duplicates removed; 950 screened; 140 full texts assessed; 30 included
Figure 2. Publication periods	2015–2019: 6; 2020–2022: 10; 2023: 14
Figure 3. Primary data sources	Operational: 8; adjacent-domain public benchmark: 12; public warehouse/logistics: 2; simulated: 5; mixed: 3
Figure 4. Uncertainty reporting	7 of 30 studies reported confidence intervals

The example figures summarize screening, publication timing, data-source composition, and uncertainty reporting. Figure 1 traces the records from identification to inclusion, while Figures 2 and 3 show the distribution across publication periods and evidence sources. Figure 4 indicates that confidence intervals were reported in 7 of 30 studies (23.3%). These visualizations describe study characteristics and reporting practices; they do not establish that one method performed better than another. Performance plots should be added only where outcomes, labels, and evaluation designs are sufficiently comparable. All displayed values require replacement with verified review data.

DISCUSSION

The selection process yielded 30 included studies from 1,268 identified records after duplicates were removed and records were screened for relevance and eligibility (Bright et al., 2018). This distribution placed the review within a broad research area while highlighting a difference between the volume of graph-based fraud research and the smaller body of direct warehouse evidence. Earlier reviews of graph-based anomaly detection and fraud have examined methods used in financial transactions, online platforms, communication networks, and other relational settings. Warehouse inventory research, in contrast, has more often examined record discrepancies, counting accuracy, employee activity, and operational processes. The composition reflected these separate strands: eight studies used operational warehouse or inventory data, while many of the others used public graph benchmarks or adjacent-domain datasets. This pattern suggests that the available evidence base is methodologically broader than its direct warehouse coverage. It also means that findings from financial or online fraud studies cannot be treated as warehouse performance results without qualification. The distinction is important because warehouse data connect physical stock, operational events, employees, orders, and locations, while other graph domains may focus on accounts, transactions, or user interactions (Allen et al., 2018). The screening numbers and study categories should not be presented as verified review results until the actual records have been screened and extracted. Interpreted as a provisional pattern, however, they demonstrate why a review on warehouse management needs to connect two bodies of literature while preserving their differences. Earlier graph-fraud research supports the value of analyzing relationships among entities, and inventory research clarifies the operational meanings of inaccurate records and stock discrepancies. Bringing these strands together allows this study to examine methods and concepts across domains while distinguishing direct evidence from methods transferred from adjacent applications.

Figure 13: Warehouse Evidence Synthesis and Interpretation



The classification showed heterogeneous static graphs as the most frequent primary representation, with nine studies, or 30.0%, followed by homogeneous static graphs in seven studies, or 23.3%. Heterogeneous temporal graphs appeared in six studies, while event-sequence or graph-hybrid approaches appeared in five. This mix reflects an established pattern in graph anomaly research: studies select the graph type according to which relationships, entity types, and temporal details are available in the data (Mengist et al., 2020). Earlier graph studies have used both homogeneous networks and heterogeneous graphs to analyze connected behavior, while dynamic-graph research has emphasized that temporal changes can provide information unavailable in a single snapshot. The detection-target distribution was also concentrated, with node- or entity-level analysis in 12 studies, or 40.0%, and edge- or transaction-level analysis in six. Subgraph and event-sequence targets were less common. This pattern resembles much fraud-detection work, where suspicious accounts or transactions are often the primary prediction targets. Warehouse research may require additional attention to inventory records, item-location relationships, and ordered operational events because data integrity can be expressed as an inaccurate quantity, an unusual movement, or a sequence that departs from expected procedure. The graph representation affects which of these phenomena can be detected. A user-centered node model may identify unusual employee behavior but obscure item-level discrepancies; an event-based graph may preserve transaction order but require reliable timestamps and identifiers (Movsisyan et al., 2019). The observed differences therefore concern not only model architecture but also the definition and construction of the analytical unit. These distributions support treating graph representation and detection target as central dimensions of comparison, rather than assuming that all graph-based methods analyze the same kind of anomaly.

The methods table placed graph convolutional or graph attention networks in the largest single method group, with eight of 30 studies, or 26.7%. Other graph neural networks accounted for six studies, and statistical graph measures, embedding or probabilistic methods, autoencoders, and hybrid approaches formed the remaining groups. Supervised learning was the most frequent primary paradigm, appearing in 13 studies, or 43.3%, followed by semi-supervised and unsupervised methods (Yu et al., 2017). This distribution is broadly consistent with graph-fraud research, where graph neural models have been applied to labeled or partially labeled transaction and account datasets. Earlier studies have also addressed challenges such as class imbalance and camouflage through model architectures and sampling strategies designed for relational data. The relatively large supervised share in the findings should be interpreted alongside the availability and quality of labels. Confirmed warehouse fraud cases are generally harder to obtain than routine transaction records, so a method that depends on labeled cases may perform well on a benchmark while remaining difficult to evaluate in a setting with sparse investigation outcomes. Unsupervised and semi-supervised methods reduce dependence on complete labels, but they introduce other interpretive challenges because unusual patterns may reflect legitimate operational variation. Hybrid approaches that combine rules, graph scores, and analyst review can connect model ranking with contextual judgment, although their effectiveness depends on clearly specified review procedures (Funke et al., 2019). The range of method families also complicates direct comparisons: a statistical structural score, an autoencoder reconstruction signal, and a supervised graph neural classification probability do not represent equivalent outputs. The method profile therefore suggests that graph neural networks were prominent in the broader evidence base, while operational usefulness depended on learning paradigm, target definition, graph construction, and validation. Actual conclusions require the completed study-level extraction and should avoid declaring one family superior based on counts alone.

The extraction suggested that precision and recall were each reported by 21 studies, or 70.0%, while F1 scores appeared in 20 studies. ROC-AUC was reported more often than PR-AUC, and specificity and false-positive rate appeared less frequently. This reporting pattern matters because fraud and serious inventory anomalies are often rare relative to normal activity. Earlier fraud-detection studies have therefore relied on precision-recall measures, class-sensitive evaluation, and comparisons across detection thresholds to make the trade-off between missed cases and false alerts visible (Munn et al., 2018). ROC-AUC remains a useful measure of ranking discrimination, but it does not by itself show the proportion of alerts that would be confirmed or the workload generated at a practical threshold. In the subgroup comparison, public adjacent-domain benchmarks showed a higher example mean F1 score than the operational warehouse group. However, the warehouse estimate drew on only five studies reporting the required metrics, and the public warehouse or logistics estimate drew on one. The datasets, labels, class prevalence, and evaluation designs were not standardized. These differences prevent a sound ranking of data-source groups and limit the interpretation of their mean values. A higher score on a review or financial benchmark does not establish that the same method would perform similarly on WMS data. Likewise, differences in metric reporting can produce incomplete subgroup comparisons because not every study reported precision, recall, and F1 together. Earlier literature has documented numerous graph-based fraud methods, but comparisons remain sensitive to dataset and validation choices (Van den Bosch & Sang, 2017). Thus, the metric profile is best understood as a demonstration of how reported outcomes can be organized. The actual discussion should interpret performance by metric, target, prevalence, and evaluation design, rather than relying on a single aggregate score.

The data-source breakdown contained eight studies using operational warehouse or inventory data, 12 using public graph benchmarks from adjacent domains, two using public warehouse or logistics datasets, five using simulated data, and three combining operational and simulated data. This composition indicates that much of the graph-method evidence may come from settings that are not warehouses. Earlier inventory record studies have typically examined physical counts, product or store characteristics, audit frequency, labor, and operational processes. Warehouse event-log research has also analyzed activity sequences and process conformance using system records (Halawa et al., 2020). These approaches provide direct operational context, but they do not always include a graph structure or labels for confirmed fraud. Public graph benchmarks support reproducible method comparison and

often provide labeled entities, but their relationship types and behavioral patterns may differ from warehouse transactions. Simulated data allow researchers to control which anomalies are inserted and evaluate detection under known conditions, though simulated patterns depend on the assumptions used in their construction. The reporting profile also showed gaps in dataset descriptions: sample or record counts were more often reported than observation periods, event types, facility counts, missing-data procedures, or duplicate handling. Label provenance was reported more frequently, yet label availability alone does not establish that a label represents confirmed misconduct. Such differences affect both replication and interpretation (Krick et al., 2019). An operational dataset with incomplete event capture may generate different graph structure from a public benchmark with curated links; a simulation with known positives may yield cleaner evaluation than a real investigation dataset. The results therefore reinforce the need to identify source type, setting, scale, data quality, and label origin alongside performance. Comparisons across domains can inform methodological understanding, while warehouse-specific claims should be grounded in studies using warehouse or closely related inventory data.

The validation table showed random splitting as the most common primary design, used in 12 studies, or 40.0%, followed by temporal splitting in eight studies and entity-based splitting in five. Only two studies used facility-based splits, and seven studies reported cross-site or cross-facility evaluation. This distribution highlights a distinction between evaluating on held-out records from a familiar dataset and evaluating transfer to later periods, new entities, or different warehouses. Earlier graph-fraud research often evaluates models on established benchmark datasets, which enables controlled comparisons but may not test whether performance transfers to a different graph or operating context (Angelis et al., 2018). Temporal splits are important when behavior changes over time, while entity-based splits can reduce the risk that the same users or products appear in both training and test sets. Facility-based evaluation is particularly relevant for warehouse applications because facilities may differ in procedures, layout, automation, product mix, and system configuration. The results also showed that class-imbalance handling was reported in 19 studies, while label uncertainty, missing data, noisy links, and behavior shifts were reported less consistently. These challenges can affect model performance in different ways: rebalancing may alter the observed class distribution, missing links can distort network structure, and changes in activity patterns can make the training baseline less representative. Potential leakage also requires attention when connected entities or future events cross data partitions (Cooper & Stewart, 2021).

A random split may allow highly related activity to appear in both training and test data, making the estimated performance less representative of new cases. The counts do not establish that this occurred in the included studies; they show why validation design must be compared explicitly. Stronger interpretation depends on knowing how graphs were built, split, and tested, and whether results were evaluated across time, entities, or facilities. The uncertainty table showed that inferential tests and p values were reported in fewer than half of the example studies, while confidence intervals appeared in seven of 30 studies. Repeated runs or variability estimates appeared more frequently, but direct statistical comparisons between models were limited. This pattern underscores the difference between a reported performance score and evidence that one method's performance was reliably different from another's. Earlier methodological research in graph anomaly detection has used a variety of datasets, labels, and target definitions, which complicates pooled estimation and helps explain why descriptive synthesis is often more defensible than combining unlike results (Vanaken & Danckaerts, 2018). The hypothetical effect table illustrated how confidence intervals and p values could be reported together, but those example estimates were not drawn from published studies and must not be treated as review findings. In the completed review, effect magnitudes should be interpreted alongside confidence intervals, sample sizes, repeated-run variability, and practical thresholds. A statistically significant increase in AUC may have limited operational value if the model creates too many alerts, while a modest performance change may matter when it reduces missed high-risk cases without substantially increasing review burden. Differences in label quality, class prevalence, and validation also affect uncertainty around reported estimates. The provisional synthesis therefore points to a literature with diverse graph methods and evaluation practices, but uneven direct warehouse coverage and incomplete reporting of uncertainty (Malhotra & Elnakib, 2021). The central interpretive requirement

is to separate evidence about method capability on public or simulated benchmarks from evidence about warehouse operations, and to distinguish anomaly detection from verified fraud. These distinctions allow the review to compare studies without overstating what the values or cross-domain results demonstrate.

CONCLUSION

This study examined graph-based anomaly detection for data integrity and fraud prevention in warehouse management systems through a PRISMA-guided quantitative review framework. The reviewed topic brought together two related but distinct areas of research: inventory and warehouse studies focused on record accuracy, physical discrepancies, and operational processes, and graph-based fraud studies focused on suspicious relationships, transactions, and connected entities. The framework organized the evidence by data source, graph representation, detection target, method family, validation design, and reported performance. Across these dimensions, graph models offered a structured way to analyze connections among products, employees, orders, locations, suppliers, devices, and warehouse events. However, an unusual pattern or inaccurate inventory record did not, by itself, establish that fraud had occurred. Reliable interpretation depended on clear outcome definitions, traceable labels, and verification through relevant operational or audit evidence. The evidence framework also showed why performance measures must be interpreted in context. Precision, recall, F1 score, and area-under-curve measures answer different questions, and their values depend on class prevalence, validation strategy, and the unit being analyzed. Findings from public financial, e-commerce, or cybersecurity benchmarks can inform method selection, but they cannot be treated as direct evidence of performance in warehouse operations. Comparisons were also constrained by variation in graph construction, dataset quality, label provenance, data splitting, and reporting of confidence intervals or repeated-run variability. The review therefore requires a careful distinction between direct warehouse evidence and methods transferred from adjacent domains, as well as between model alerts, process deviations, inventory discrepancies, and confirmed fraud. The numerical results presented in the drafted findings and discussion were examples because verified screening and extraction data were not supplied; they should not be treated as completed review outcomes. Once replaced with the actual review results, the conclusion should reflect the evidence found rather than the example values. Overall, this review framework provides a transparent basis for describing the current evidence, assessing the comparability of reported methods, and interpreting graph-based detection in relation to warehouse data integrity and fraud prevention.

RECOMMENDATION

Researchers conducting studies on graph-based anomaly detection in warehouse management systems should define data integrity, inventory discrepancy, process deviation, suspicious activity, and confirmed fraud as separate outcomes before building or evaluating a model. Each study should describe its unit of analysis, data source, observation period, facility context, graph construction, node and edge types, and preprocessing decisions. Reports should state how duplicate records, missing values, inconsistent identifiers, and incomplete event trails were handled. Ground-truth procedures should also be explained, including whether labels came from physical audits, investigations, expert review, operational rules, proxy outcomes, or synthetic anomaly injection. Model evaluation should use validation designs that match the intended use, with temporal, entity-based, or facility-based testing where appropriate, and should document steps taken to prevent information leakage across training and test data. Performance reporting should include precision, recall, F1 score, PR-AUC, false-positive burden, and relevant uncertainty estimates, alongside accuracy or ROC-AUC where useful. Studies should report class prevalence, threshold selection, confidence intervals, repeated-run variability, and comparisons with both rule-based controls and suitable non-graph and graph-based models. For warehouse operators, automated scores should support prioritization and investigation rather than serve as stand-alone proof of misconduct. Review procedures should connect alerts to transaction histories, approval records, physical counts, and other corroborating evidence, while recording whether alerts were substantiated or dismissed. Collaboration among researchers and warehouse organizations can support carefully governed use of operational data, clear documentation, and reproducible analysis without exposing commercially sensitive or employee-identifying information. Finally, any manuscript using the counts and performance values drafted for this review

should replace them with verified search, screening, and extraction results before publication, since example values do not provide evidence for empirical recommendations.

LIMITATIONS

This study had several limitations that affected the interpretation and comparability of the evidence. The review depended on the databases searched, the search terms used, and the availability of full-text publications; relevant studies may therefore have been missed, particularly research reported in technical documents, theses, industry reports, or non-indexed sources. Restricting eligibility to accessible publications and specified languages may also have excluded relevant work. The evidence base combined direct warehouse and inventory studies with research from financial fraud, e-commerce, cybersecurity, and other graph applications. Methods from these adjacent domains informed the review, but their datasets, entity relationships, fraud definitions, and operating conditions differed from those of warehouse management systems. Included studies also used varied data sources, graph structures, anomaly targets, label procedures, validation strategies, and performance metrics. This heterogeneity limited direct comparison and could prevent meaningful pooling of results. Operational warehouse data were often private or commercially sensitive, which restricted access to datasets and details about fields, preprocessing, missing records, duplicate handling, and confirmed fraud labels. Several studies evaluated suspected anomalies, proxy labels, or simulated cases rather than independently verified fraud. Their results therefore could not be interpreted as equivalent evidence of detecting confirmed misconduct. Class imbalance, incomplete ground truth, possible data leakage, and inconsistent reporting of confidence intervals, repeated runs, and computational costs further limited assessment of model reliability. The number and quality of studies in individual subgroups may also have been too limited to support broad conclusions about particular graph types or method families. Finally, the numerical counts and performance values drafted for the findings and discussion sections were examples because verified screening and extraction data were not provided. They did not represent completed review results and cannot support empirical claims until replaced with documented data from the actual search, selection, and analysis. These limitations require the conclusions to remain within the scope of the evidence that is ultimately verified and extracted.

REFERENCES

[1]. Alaux, M., Rogers, J., Letellier, T., Flores, R., Alfama, F., Pommier, C., Mohellibi, N., Durand, S., Kimmel, E., & Michotey, C. (2018). Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome biology, 19(1), 111.

[2]. Alla, K., Hall, W. D., Whiteford, H. A., Head, B. W., & Meurk, C. S. (2017). How do we define the policy impact of public health research? A systematic review. Health research policy and systems, 15(1), 84.

[3]. Allen, C., Metternicht, G., & Wiedmann, T. (2018). Initial progress in implementing the Sustainable Development Goals (SDGs): A review of evidence from countries. Sustainability science, 13(5), 1453-1467.

[4]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. International Journal of Business and Economics Insights, 13(09), 01-33. https://doi.org/10.63125/bmxwv741

[5]. Angelis, A., Lange, A., & Kanavos, P. (2018). Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. The European Journal of Health Economics, 19(1), 123-152.

[6]. Asheke, M., & Tashriq Ahmed Bin, A. (2022). Federated Learning and Differential Privacy for Cross-Firm Construction-Materials Supply-Chain Disruption Forecasting: A Quantitative Analysis of Privacy-Utility Trade-Offs. American Journal of Advanced Technology and Engineering Solutions, 12(09), 01-40. https://doi.org/10.63125/g4at4237

[7]. Ateş, Ç., Özdel, S., & Anarım, E. (2019). Graph–based anomaly detection using fuzzy clustering. International conference on intelligent and fuzzy systems,

[8]. Augustinavicius, J. L., Greene, M. C., Lakin, D. P., & Tol, W. A. (2018). Monitoring and evaluation of mental health and psychosocial support programs in humanitarian settings: a scoping review of terminology and focus. Conflict and health, 12(1), 9.

[9]. Bassi, M. C., Marchegiani, M. G., Dervenis, M. C., Sarr, M. M., Hilal, M. M. A., Adham, M. M., Allen, M. P., Andersson, M. R., Asbun, M. H. J., & Besselink, M. M. G. (2017). The 2016 update of the International Study Group (ISGPS) definition and grading of postoperative pancreatic fistula: 11 years after. surgery, 161(3), 584-591.

[10]. Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Health information science and systems, 6(1), 9.

[11]. Bienert, A., Georgi, L., Kunz, M., Maas, H.-G., & Von Oheimb, G. (2018). Comparison and combination of mobile and terrestrial laser scanning for natural forest inventories. Forests, 9(7), 395.

[12]. Blackford, J., Artioli, Y., Clark, J., & de Mora, L. (2017). Monitoring of offshore geological carbon storage integrity: Implications of natural variability in the marine system and the assessment of anomaly detection criteria. International Journal of Greenhouse Gas Control, 64, 99-112.

[13]. Breier, J., & Branišová, J. (2017). A dynamic rule creation based anomaly detection method for identifying security breaches in log records. Wireless Personal Communications, 94(3), 497-511.

[14]. Bright, T., Wallace, S., & Kuper, H. (2018). A systematic review of access to rehabilitation for people with disabilities in low-and middle-income countries. International Journal of Environmental Research and Public Health, 15(10), 2165.

[15]. Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., Aalten, P., De Deyn, P. P., Biessels, G. J., & Claassen, J. A. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. NeuroImage: Clinical, 31, 102712.

[16]. Cao, Z., Gao, H., Mangalam, K., Cai, Q.-Z., Vo, M., & Malik, J. (2020). Long-term human motion prediction with scene context. European Conference on Computer Vision,

[17]. Carvalho, L. F., Abrão, T., de Souza Mendes, L., & Proença Jr, M. L. (2018). An ecosystem for anomaly detection and mitigation in software-defined networking. Expert Systems with Applications, 104, 121-133.

[18]. Chen, W., Zhou, K., Yang, S., & Wu, C. (2017). Data quality of electricity consumption data in a smart grid environment. Renewable and Sustainable Energy Reviews, 75, 98-105.

[19]. Cooper, K., & Stewart, K. (2021). Does Household Income Affect children's Outcomes? A Systematic Review of the Evidence: K. Cooper, K. Stewart. Child Indicators Research, 14(3), 981-1005.

[20]. Dalianis, H. (2018). Evaluation metrics and evaluation. In Clinical Text Mining: secondary use of electronic patient records (pp. 45-53). Springer.

[21]. De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management, 54(5), 807-817.

[22]. Demertzis, K., Iliadis, L., Tziritas, N., & Kikiras, P. (2020). Anomaly detection via blockchained deep learning smart contracts in industry 4.0. Neural computing and applications, 32(23), 17361-17378.

[23]. Faniyi, I., Fasakin, O., Olofinjana, B., Adekunle, A., Oluwasusi, T., Eleruja, M., & Ajayi, E. (2019). The comparative analyses of reduced graphene oxide (RGO) prepared via green, mild and chemical approaches: I. Faniyi et al. SN Applied Sciences, 1(10), 1181.

[24]. Feng, L., Xu, S., Zhang, L., Wu, J., Zhang, J., Chu, C., Wang, Z., & Shi, H. (2020). Anomaly detection for electricity consumption in cloud computing: framework, methods, applications, and challenges. EURASIP Journal on Wireless Communications and Networking, 2020(1), 194.

[25]. Fernando, T., Denman, S., Ahmedt-Aristizabal, D., Sridharan, S., Laurens, K. R., Johnston, P., & Fookes, C. (2020). Neural memory plasticity for medical anomaly detection. Neural Networks, 127, 67-81.

[26]. Funke, S. Á., Sprei, F., Gnann, T., & Plötz, P. (2019). How much charging infrastructure do electric vehicles need? A review of the evidence and international comparison. Transportation research part D: transport and environment, 77, 224-242.

[27]. Gangwar, A. K., & Ravi, V. (2019). Wip: Generative adversarial network for oversampling data in credit card fraud detection. International Conference on Information Systems Security,

[28]. Gemenet, D. C., da Silva Pereira, G., De Boeck, B., Wood, J. C., Mollinari, M., Olukolu, B. A., Diaz, F., Mosquera, V., Ssali, R. T., & David, M. (2020). Quantitative trait loci and differential gene expression analyses reveal the genetic basis for negatively associated β-carotene and starch content in hexaploid sweetpotato [Ipomoea batatas (L.) Lam.]. Theoretical and Applied Genetics, 133(1), 23-36.

[29]. González-Aparicio, M. T., Younas, M., Tuya, J., & Casado, R. (2018). Testing of transactional services in NoSQL key-value databases. Future Generation Computer Systems, 80, 384-399.

[30]. Halawa, F., Madathil, S. C., Gittler, A., & Khasawneh, M. T. (2020). Advancing evidence-based healthcare facility design: a systematic literature review: Halawa F. et al. Health Care Management Science, 23(3), 453-480.

[31]. Hang, L., Ullah, I., & Kim, D.-H. (2020). A secure fish farm platform based on blockchain for agriculture data integrity. Computers and Electronics in Agriculture, 170, 105251.

[32]. Javaid, N., Jan, N., & Javed, M. U. (2021). An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. Journal of Parallel and Distributed Computing, 153, 44-52.

[33]. Johnson, J. M., & Khoshgoftaar, T. M. (2019a). Medicare fraud detection using neural networks. Journal of Big Data, 6(1), 63.

[34]. Johnson, J. M., & Khoshgoftaar, T. M. (2019b). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27.

[35]. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234-245.

[36]. Kamišalić, A., Kramberger, R., & Fister Jr, I. (2021). Synergy of blockchain technology and data mining techniques for anomaly detection. Applied Sciences, 11(17), 7987.

[37]. Kazi Mohammad Khalid, A., & Sumaiya Sadmeen, M. (2022). Geospatial Artificial Intelligence for Urban Flood Susceptibility and Drainage Resilience: A Prisma 2020 Systematic Review and Bibliometric Meta-Synthesis. Review of Applied Science and Technology, 12(02), 01–51. https://doi.org/10.63125/epqdvy23

[38]. Knoll, D., Reinhart, G., & Prüglmeier, M. (2019). Enabling value stream mapping for internal logistics using multidimensional process mining. Expert Systems with Applications, 124, 130-142.

[39]. Knoll, T., Omar, M. I., Maclennan, S., Hernandez, V., Canfield, S., Yuan, Y., Bruins, M., Marconi, L., Van Poppel, H., & N'Dow, J. (2018). Key steps in conducting systematic reviews for underpinning clinical practice guidelines: methodology of the European Association of Urology. European urology, 73(2), 290-300.

[40]. Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. International Journal of Production Economics, 225, 107597.

[41]. Krick, T., Huter, K., Domhoff, D., Schmidt, A., Rothgang, H., & Wolf-Ostermann, K. (2019). Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC health services research, 19(1), 400.

[42]. Kudo, W., Nishiguchi, M., & Toriumi, F. (2020). Gcnext: graph convolutional network with expanded balance theory for fraudulent user detection. Social Network Analysis and Mining, 10(1), 85.

[43]. Leal, F., Chis, A. E., Caton, S., González–Vélez, H., García–Gómez, J. M., Durá, M., Sánchez–García, A., Sáez, C., Karageorgos, A., & Gerogiannis, V. C. (2021). Smart pharmaceutical manufacturing: Ensuring end-to-end traceability and data integrity in medicine production. Big Data Research, 24, 100172.

[44]. Learmonth, P. Y. C., Ensari, P. I., & Motl, P. R. W. (2017). Cognitive motor interference in multiple sclerosis: insights from a systematic quantitative review. Archives of physical medicine and rehabilitation, 98(6), 1229-1240.

[45]. Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T.-H., & Wang, S. (2020). Real-world data medical knowledge graph: construction and applications. Artificial intelligence in medicine, 103, 101817.

[46]. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. European conference on computer vision,

[47]. Lin, J., Zhao, Y., Huang, W., Liu, C., & Pu, H. (2021). Domain knowledge graph-based research progress of knowledge representation. Neural computing and applications, 33(2), 681-690.

[48]. Liu, G., Guo, J., Zuo, Y., Wu, J., & Guo, R.-y. (2020). Fraud detection via behavioral sequence embedding: G. Liu et al. Knowledge and information Systems, 62(7), 2685-2708.

[49]. Malhotra, P. D. A., & Elnakib, M. S. (2021). 20 years of the evidence base on what works to prevent child marriage: A systematic review. Journal of Adolescent Health, 68(5), 847-862.

[50]. Mariani, M. M., & Wamba, S. F. (2020). Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies. Journal of Business Research, 121, 338-352.

[51]. Martin, F. C., & O'Halloran, A. M. (2020). Tools for assessing frailty in older people: general concepts. Frailty and cardiovascular diseases: Research into an elderly population, 9-19.

[52]. Mashfiqur, R., & Md Jamil, A. (2023). Machine Learning-Based Cost Variance Prediction for Early Detection of Budget Overruns in Procurement-Intensive Operations. Review of Applied Science and Technology, 13(12), 142–199. https://doi.org/10.63125/7g92wd03

[53]. Maw, A., Adepu, S., & Mathur, A. (2019). ICS-BlockOpS: Blockchain for operational data security in industrial control system. Pervasive and Mobile Computing, 59, 101048.

[54]. Md, F., & Kynatun, M. (2023). A Quantitative Assessment of Cybersecurity Risk Scoring Models for Enterprise Database Systems in Cloud-Native Environments. Review of Applied Science and Technology, 13(12), 50–103. https://doi.org/10.63125/gk2krm47

[55]. Md Maksudul, H., & Md Takbir Hossen, S. (2023). Predictive Inventory Optimization for Reducing Stockouts and Excess Inventory in U.S. Sports and Apparel Operations. Review of Applied Science and Technology, 13(12), 104–141. https://doi.org/10.63125/8djn1j48

[56]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. International Journal of Scientific Interdisciplinary Research, 12(12), 01–38. https://doi.org/10.63125/w3pgjx97

[57]. Md Mohsin, B., & Jannatul, F. (2021). Emergency Nursing Triage and Timely Escalation of Care: A Quantitative Clinical Audit. International Journal of Scientific Interdisciplinary Research, 11(07), 01–33. https://ijsir.org/index.php/IJSIR/article/view/138

[58]. Md Mohsin, B., & Mst Kaniz, F. (2023). AI-Enabled Clinical Decision Support for Standardized Nurse-Led Emergency Care Pathways: A Multisite Mixed-Methods Implementation Study. Review of Applied Science and Technology, 13(08), 01–33. https://doi.org/10.63125/ac09sa31

[59]. Md. Mamunur, R., & Md, A. (2023). Cloud Migration Readiness and Operational Risk in U.S. Organizations: A Quantitative Analysis of Infrastructure and Business Continuity Indicators. American Journal of Scholarly Research and Innovation, 13(07), 01-27. https://doi.org/10.63125/28jkj206

[60]. Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX, 7, 100777.

[61]. Mitchell, W. G., Dee, E. C., & Celi, L. A. (2021). Generalisability through local validation: overcoming barriers due to data disparity in healthcare. BMC ophthalmology, 21(1), 228.

[62]. Moschini, G., Houssou, R., Bovay, J., & Robert-Nicoud, S. (2021). Anomaly and fraud detection in credit card transactions using the arima model. Engineering Proceedings, 5(1), 56.

[63]. Movsisyan, A., Arnold, L., Evans, R., Hallingberg, B., Moore, G., O'Cathain, A., Pfadenhauer, L. M., Segrott, J., & Rehfuess, E. (2019). Adapting evidence-informed complex population health interventions for new contexts: a systematic review of guidance. Implementation Science, 14(1), 105.

[64]. Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology, 18(1), 143.

[65]. Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., De-la-Hoz-Franco, E., & De-La-Hoz-Valdiris, E. (2021). Trends and future perspective challenges in big data. Advances in intelligent data analysis and applications: Proceeding of the sixth euro-China conference on intelligent data analysis and applications, 15–18 October 2019, Arad, Romania,

[66]. Nauata, N., Chang, K.-H., Cheng, C.-Y., Mori, G., & Furukawa, Y. (2020). House-gan: Relational generative adversarial networks for graph-constrained house layout generation. European Conference on Computer Vision,

[67]. Ochodkova, E., Zehnalova, S., & Kudelka, M. (2017). Graph construction based on local representativeness. International Computing and Combinatorics Conference,

[68]. Peng, Y., Lin, J.-R., Zhang, J.-P., & Hu, Z.-Z. (2017). A hybrid data mining approach on BIM-based building operation and maintenance. Building and environment, 126, 483-495.

[69]. Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 133, 113303.

[70]. Prak, D., Teunter, R., & Syntetos, A. (2017). On the calculation of safety stocks when demand is forecasted. European Journal of Operational Research, 256(2), 454-461.

[71]. Preuveneers, D., Rimmer, V., Tsingenopoulos, I., Spooren, J., Joosen, W., & Ilie-Zudor, E. (2018). Chained anomaly detection models for federated learning: An intrusion detection case study. Applied Sciences, 8(12), 2663.

[72]. Punnaiah Chowdary, R. (2021). GC-MS Strategies for Identification of Trace Impurities in Active Pharmaceutical Ingredients. American Journal of Advanced Technology and Engineering Solutions, 11(05), 01-35. https://ajates-scholarly.com/index.php/ajates/article/view/131

[73]. Punnaiah Chowdary, R., & Md, A. (2020). Practical HPLC Method Validation Under ICH Q2(R2) and Q14: An Analytical Quality Framework. International Journal of Business and Economics Insights, 10(03), 01–36. https://ijbei-journal.org/index.php/ijbei/article/view/73

[74]. Punnaiah Chowdary, R., & Md Jamil, A. (2020). Risk-Based Lifecycle Management of Chromatographic Methods for Generic API Quality: A Systematic Review. Review of Applied Science and Technology, 10(09), 01–41. https://doi.org/10.63125/r0xnar90

[75]. Qureshi, K. N., Jeon, G., & Piccialli, F. (2021). Anomaly detection and trust authority in artificial intelligence and cloud computing. Computer Networks, 184, 107647.

[76]. Rewa, O. G., Stelfox, H. T., Ingolfsson, A., Zygun, D. A., Featherstone, R., Opgenorth, D., & Bagshaw, S. M. (2018). Indicators of intensive care unit capacity strain: a systematic review. Critical Care, 22(1), 86.

[77]. Rochefort, C. M., Buckeridge, D. L., Tanguay, A., Biron, A., D'Aragon, F., Wang, S., Gallix, B., Valiquette, L., Audet, L.-A., & Lee, T. C. (2017). Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. BMC health services research, 17(1), 147.

[78]. Russell, C., Rueda, S., Room, R., Tyndall, M., & Fischer, B. (2018). Routes of administration for cannabis use–basic prevalence and related health outcomes: A scoping review and synthesis. International Journal of Drug Policy, 52, 87-96.

[79]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. International Journal of Scientific Interdisciplinary Research, 4(2), 252–295. https://doi.org/10.63125/cajkqt60

[80]. Sargent, L., & Brown, R. (2017). Assessing the current state of cognitive frailty: measurement properties. The journal of nutrition, health & aging, 21(2), 152-160.

[81]. Schalock, R. L., Luckasson, R., & Shogren, K. A. (2020). Going beyond environment to context: Leveraging the power of context to produce change. International Journal of Environmental Research and Public Health, 17(6), 1885.

[82]. Sekhon, M., Cartwright, M., & Francis, J. J. (2017). Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC health services research, 17(1), 88.

[83]. Sengan, S., Subramaniyaswamy, V., Indragandhi, V., Velayutham, P., & Ravi, L. (2021). Detection of false data cyber-attacks for the assessment of security in smart grid using deep learning. Computers & Electrical Engineering, 93, 107211.

[84]. Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., & Han, B. (2020). Learning to optimize domain specific normalization for domain generalization. European conference on computer vision,

[85]. Sharmin, A., & Mst Kaniz, F. (2023). "My Pain Was Not Taken Seriously" Medical Gaslighting As an Expectancy Violation in Women's Healthcare. American Journal of Health and Medical Sciences, 13(06), 55–85. https://ajhmsjournal.com/index.php/ajhms/article/view/20

[86]. Sharmin, A., & Nishat, J. (2023). Communication Accommodation and Patient Trust in Healthcare and Chronic Illness Care. Review of Applied Science and Technology, 13(08), 34–72. https://doi.org/10.63125/6p0xwb79

[87]. Shawon, M. S. R., Perret, J. L., Senaratna, C. V., Lodge, C., Hamilton, G. S., & Dharmage, S. C. (2017). Current evidence on prevalence and clinical outcomes of co-morbid obstructive sleep apnea and chronic obstructive pulmonary disease: a systematic review. Sleep medicine reviews, 32, 58-68.

[88]. Shi, X., Prins, C., Van Pottelbergh, G., Mamouris, P., Vaes, B., & De Moor, B. (2021). An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. BMC Medical Informatics and Decision Making, 21(1), 267.

[89]. Smith, J. D., & Hasan, M. (2020). Quantitative approaches for the evaluation of implementation research studies. Psychiatry research, 283, 112521.

[90]. Song, W., Beshley, M., Przystupa, K., Beshley, H., Kochan, O., Pryslupskyi, A., Pieniak, D., & Su, J. (2020). A software deep packet inspection system for network traffic analysis and anomaly detection. Sensors, 20(6), 1637.

[91]. Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676-687.

[92]. Sumaiya Sadmeen, M., & Kazi Mohammad Khalid, A. (2023). Multi-Source Remote Sensing and Dem-Derived Hydrogeomorphic Indicators for High-Resolution Urban Flood Susceptibility Mapping: A Spatial Cross-Validation Framework. Journal of Sustainable Development and Policy, 13(07), 01-46. https://doi.org/10.63125/z4xt7k38

[93]. Van den Bosch, M., & Sang, Å. O. (2017). Urban natural environments as nature-based solutions for improved public health–A systematic review of reviews. Environmental research, 158, 373-384.

[94]. Vanaken, G.-J., & Danckaerts, M. (2018). Impact of green space exposure on children's and adolescents' mental health: A systematic review. International Journal of Environmental Research and Public Health, 15(12), 2668.

[95]. Vávra, J., Hromada, M., Lukáš, L., & Dworzecki, J. (2021). Adaptive anomaly detection system based on machine learning algorithms in an industrial control environment. International Journal of Critical Infrastructure Protection, 34, 100446.

[96]. Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics, 87, 12-20.

[97]. Wang, Z., Luo, N., & Zhou, P. (2020). GuardHealth: Blockchain empowered secure data management and Graph Convolutional Network enabled anomaly detection in smart healthcare. Journal of Parallel and Distributed Computing, 142, 1-12.

[98]. Weng, H., Liu, Z., Yan, S., Fan, M., Ou, A., Chen, D., & Hao, T. (2017). A framework for automated knowledge graph construction towards traditional Chinese medicine. International Conference on Health Information Science,

[99]. Xu, R., Cheng, Y., Liu, Z., Xie, Y., & Yang, Y. (2020). Improved Long Short-Term Memory based anomaly detection with concept drift adaptive method for supporting IoT services. Future Generation Computer Systems, 112, 228-242.

[100]. Yu, S. W., Hill, C., Ricks, M. L., Bennet, J., & Oriol, N. E. (2017). The scope and impact of mobile health clinics in the United States: a literature review. International journal for equity in health, 16(1), 178.

[101]. Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. International Journal of Environmental Research and Public Health, 17(19), 7265.

[102]. Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. Computational Social Networks, 6(1), 1-23.