Journal of Banking and Financial Dynamics
Vol. 9, No. 8, 1-11, 2025
ISSN(E): 2576-6821
DOI: 10.55220/2576-6821.v9.605

Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends

Tiejiang Sun[1]
Mengdie Wang[2]
Xu Han[3]

[1]*School of Information Engineering, Chang'an University, Xi'an 710064, China.*
[2]*School of Taxation and Public Administration, Shanghai Lixin University of Accounting and Finance, Shanghai 201620, China.*
[3]*School of Business, Renmin University of China, Beijing 100872, China.*
(Corresponding Author)

Abstract

Insurance fraud represents a significant financial burden globally, with annual losses exceeding $200 billion across healthcare, auto, and life insurance sectors. Traditional rule-based fraud detection systems have proven inadequate against increasingly sophisticated fraudulent schemes, prompting widespread adoption of deep learning (DL) approaches. This comprehensive review systematically examines the application of DL techniques to insurance fraud detection, analyzing 57 peer-reviewed studies published between 2019 and 2025. We evaluate the effectiveness of various architectures including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), Graph Neural Networks (GNNs), and hybrid models across healthcare, auto, and life insurance domains. Our analysis reveals that ensemble methods combining CNNs with LSTMs achieve accuracies ranging from 89.6% to 98%, while GNN-based approaches demonstrate superior performance in detecting collusive fraud networks with accuracies exceeding 84%. The review identifies critical challenges including severe class imbalance with fraud rates of 0.03-3%, model interpretability requirements, and limited availability of labeled datasets. We examine emerging trends including explainable artificial intelligence (XAI) frameworks, attention mechanisms, generative adversarial networks (GANs) for synthetic data generation, and federated learning approaches for privacy-preserving fraud detection. This review contributes to understanding the current state-of-the-art in DL for insurance fraud detection while highlighting critical research gaps and future directions in model interpretability, cross-domain transfer learning, and real-time detection systems.

Keywords: Auto insurance fraud, Convolutional neural networks, Deep learning, Explainable AI, Graph neural networks, Healthcare fraud, Imbalanced datasets, Insurance fraud detection, LSTM.

1. Introduction

Insurance fraud constitutes one of the most pervasive financial crimes globally, imposing substantial economic burdens on insurance companies, governments, and ultimately, honest policyholders through increased premiums. Conservative estimates suggest that fraudulent insurance claims account for 3-10% of total healthcare expenditures alone, translating to approximately $105 billion annually in the United States healthcare sector [1]. When considering all insurance sectors including auto insurance at $45 billion, life insurance at $74.7 billion, and property and casualty insurance, the total annual global losses attributable to insurance fraud exceed $200 billion [2]. These staggering figures underscore the critical importance of developing effective fraud detection mechanisms that can adapt to increasingly sophisticated fraudulent schemes. The Coalition Against Insurance Fraud estimates that insurance fraud costs American consumers an additional $400 to $700 per year in increased premiums, demonstrating how fraud impacts not just insurance companies but all policyholders [3].

Traditional approaches to insurance fraud detection have relied predominantly on rule-based systems and manual auditing processes conducted by fraud investigators. These conventional methods utilize predetermined heuristic rules and statistical thresholds to flag potentially fraudulent claims for human review [4]. While such approaches have provided valuable early-stage fraud detection capabilities, they suffer from several fundamental limitations that severely constrain their effectiveness in contemporary fraud detection scenarios. Rule-based systems exhibit high false positive rates, often flagging legitimate claims while missing sophisticated fraudulent schemes that fall outside predefined rule parameters [5]. Moreover, these systems require extensive domain expertise for rule formulation and frequent manual updates to adapt to evolving fraud patterns, making them resource-intensive and reactive rather than proactive [6]. The manual auditing process is particularly tedious and inefficient when confronted with the massive volumes of claims data generated by modern insurance systems, where human experts must sift through numerous records to identify suspicious or fraudulent behaviors [7].

The rapid digitalization of insurance processes and the proliferation of electronic health records, telematics data from vehicles, and digital transaction systems have generated unprecedented volumes of insurance-related data [8]. This data explosion, combined with the increasing sophistication of fraudulent schemes involving coordinated networks of perpetrators, has exposed the inadequacy of traditional detection methods. Fraudsters have evolved to exploit the limitations of rule-based systems through adaptive strategies including camouflage behavior, where fraudulent actors establish connections primarily with legitimate entities to avoid detection, and temporal manipulation, concentrating fraudulent activities within short timeframes to minimize exposure [9]. These sophisticated fraud patterns are particularly challenging for traditional methods to detect, as they involve complex relational dependencies and temporal dynamics that exceed the analytical capabilities of rule-based systems [10].

Deep learning (DL) has emerged as a transformative paradigm for addressing these challenges, offering several distinct advantages over traditional machine learning (ML) and rule-based approaches. Unlike conventional methods that rely on manually engineered features and domain-specific expert knowledge, DL models can automatically learn hierarchical representations and complex patterns directly from raw or minimally processed data [11]. This capability is particularly valuable in insurance fraud detection, where fraudulent patterns often involve subtle, nonlinear relationships across multiple variables that are difficult to capture through manual feature engineering [12]. The automatic feature learning capacity of DL eliminates the time-consuming and expertise-intensive process of feature engineering, allowing models to discover previously unknown fraud indicators from data.

Recent advances in DL architectures have demonstrated remarkable success across diverse insurance fraud detection applications. Convolutional Neural Networks (CNNs), originally developed for computer vision tasks, have been successfully adapted to extract spatial patterns from structured insurance claims data by treating tabular data as two-dimensional matrices where convolutional filters can identify local feature interactions [13]. Recurrent Neural Networks (RNNs) and their advanced variants, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, have proven highly effective for modeling temporal dependencies in sequential transaction data and identifying anomalous temporal patterns indicative of fraud [14]. These architectures maintain internal memory states that enable them to capture long-term dependencies in sequential data, making them particularly suitable for analyzing claim submission patterns, treatment histories, and transaction sequences. More recently, Graph Neural Networks (GNNs) have emerged as a powerful tool for detecting fraud networks and collusive behavior by explicitly modeling relational structures among policyholders, healthcare providers, and claims [15]. The ability of GNNs to process graph-structured data enables them to identify fraudulent patterns embedded in the network topology that would be invisible to traditional methods analyzing claims in isolation [16].

The application of DL to insurance fraud detection has accelerated dramatically in recent years, reflecting both technological advancements and growing institutional recognition of the value these methods provide. A systematic analysis of publication trends reveals a steep increase from 2022 onwards, with particularly pronounced growth between 2023 and 2024 [17]. This surge reflects multiple converging factors including advancements in DL architectures specifically designed for handling imbalanced datasets and relational data structures, increasing availability of large-scale insurance datasets for research purposes, growing regulatory pressure and financial incentives for improved fraud detection capabilities, and demonstrated superiority of DL approaches over traditional methods in rigorous empirical evaluations. Among DL techniques, LSTM networks have exhibited the most sustained growth trajectory, with applications increasing sharply from 2022 to 2024 as shown in Figure 1 [17]. This trend is attributable to the inherently sequential nature of insurance fraud datasets, where temporal patterns of claim submissions, payment histories, and service utilization sequences provide crucial signals for distinguishing fraudulent from legitimate behavior [18]. Multilayer Perceptrons (MLPs) and CNNs have maintained steady application rates due to their versatility in learning complex feature interactions and their computational efficiency for real-time deployment scenarios where rapid decision-making is essential [19].

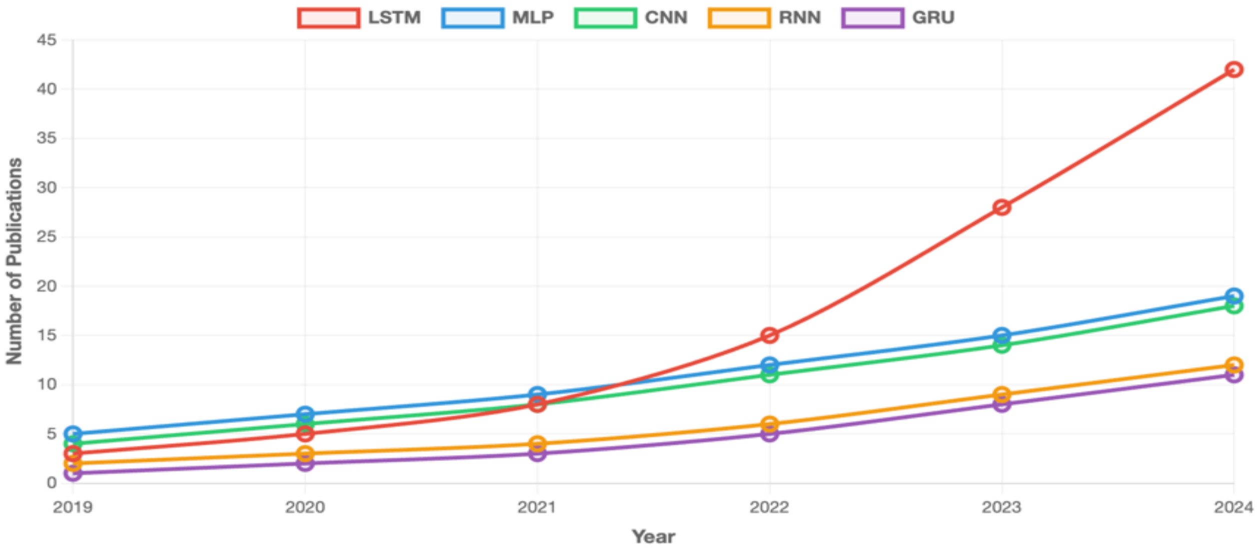

Figure 1. Yearly Trends of Deep Learning Algorithm Application in Fraud Detection (2019–2024)
Source: Chen et al. (2025) systematic review. The graph shows LSTM demonstrating the steepest growth trajectory, particularly from 2022-2024, reflecting the sequential nature of fraud detection data.

Despite these advances, several critical challenges continue to impede the widespread adoption and effectiveness of DL for insurance fraud detection. Chief among these is the severe class imbalance problem, where fraudulent cases typically represent only 0.03-3% of total claims in real-world datasets [20]. This extreme imbalance can cause DL models to exhibit bias toward the majority class, achieving high overall accuracy by simply predicting all cases as legitimate while failing to detect actual fraud cases that represent the minority class of primary interest [21]. Model interpretability represents another significant concern, as many high-performing DL architectures operate as black boxes, providing predictions without transparent reasoning about which factors drove particular decisions [22]. This lack of transparency conflicts with regulatory requirements in many jurisdictions that mandate explainable decision-making for actions affecting individuals, such as claim denials or fraud investigations [23]. Furthermore, practitioners including fraud investigators and claims adjusters require understandable explanations to validate model decisions, investigate flagged cases effectively, and maintain trust in automated systems.

Data quality and availability constitute additional major obstacles to effective DL deployment in insurance fraud detection. Many studies rely on private proprietary datasets that cannot be shared for research purposes due to confidentiality agreements and competitive considerations, hindering reproducibility and comparative evaluation of different approaches [24]. Publicly available datasets often suffer from inconsistent labeling where ground truth fraud labels may be uncertain or incomplete, missing values across important variables, and limited coverage of real-world fraud scenarios including emerging fraud types not represented in historical data [25]. Furthermore, privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe impose stringent constraints on the collection, storage, and sharing of insurance-related personal data, creating legal and ethical barriers to data-intensive DL research [26]. These regulations require extensive anonymization and de-identification procedures that may remove information valuable for fraud detection while still maintaining sufficient detail for model training.

This comprehensive review aims to address these challenges and provide a systematic synthesis of the current state-of-the-art in DL applications for insurance fraud detection. Through analysis of 57 peer-reviewed studies spanning 2019 to 2025, we examine the evolution and comparative effectiveness of different DL architectures across healthcare, auto, and life insurance fraud detection domains, techniques for addressing the class imbalance problem including sampling methods and cost-sensitive learning, publicly available datasets and their characteristics with discussion of their suitability for different research objectives, performance metrics and evaluation methodologies appropriate for imbalanced fraud detection scenarios, emerging trends including explainable AI frameworks and graph-based fraud network detection, and critical research gaps with promising directions for future work that could advance both theoretical understanding and practical deployment of DL fraud detection systems.

2. Literature Review

The application of computational methods to insurance fraud detection has evolved progressively from simple statistical models to sophisticated DL architectures over the past two decades. Early fraud detection systems relied primarily on statistical anomaly detection and rule-based expert systems that encoded domain knowledge from fraud investigators into explicit decision rules [27]. These approaches provided interpretable decisions that could be explained to stakeholders and audited for compliance, but suffered from limited adaptability to new fraud patterns and high maintenance costs as fraud schemes evolved and rules required constant updating. The transition from these traditional methods to ML-based approaches began in the early 2000s with the application of classical algorithms including logistic regression, decision trees, and support vector machines to fraud detection tasks [28]. These ML methods demonstrated improved performance over rule-based systems by learning patterns from labeled historical data rather than relying solely on predefined rules, yet they still required substantial manual feature engineering to transform raw claim data into representations suitable for modeling.

The emergence of DL has fundamentally transformed the landscape of insurance fraud detection by enabling end-to-end learning from raw data without extensive feature engineering. Recent systematic reviews have documented a surge in DL applications for financial fraud detection broadly, with insurance fraud representing a major application domain [17]. The review documented that traditional ML approaches remain dominant with 94 studies employing supervised methods, while DL techniques are experiencing rapid adoption with 41 studies using unsupervised methods and 12 using hybrid approaches that combine multiple paradigms [29]. The distribution across financial sectors shows credit card and banking fraud attracting the most research attention, though insurance fraud detection represents a substantial and growing portion of the literature with particular focus on healthcare and auto insurance domains.

Healthcare insurance fraud has received extensive research attention due to both the magnitude of financial losses involved and the availability of large-scale public datasets from government healthcare programs. A systematic literature review specifically examining fraud detection in healthcare claims using ML identified important patterns in research approaches and methodologies [1]. Their analysis revealed that studies focused on fraud detection by healthcare providers represent the most prevalent category, followed by fraud committed by patients, with relatively fewer studies examining fraud by insurance carriers or complex conspiracy frauds involving multiple parties. The review identified 30 studies utilizing private data sources and the remainder using publicly available datasets, highlighting ongoing challenges with data accessibility that limit reproducibility and comparative evaluation. Geographic distribution of research shows strong concentration in the United States with 96 studies, followed by China with 11 studies and Australia with 5 studies, reflecting both the scale of healthcare systems in these countries and the availability of research datasets [1].

Auto insurance fraud detection has similarly attracted substantial research interest, with particular focus on claim severity prediction and identification of exaggerated or fabricated claims [30]. Recent work presented a systematic review of data mining techniques applied in automobile insurance fraud detection, documenting the effectiveness of various classification algorithms and clustering approaches in identifying fraudulent claims [31]. The review emphasized that ensemble methods combining multiple algorithms consistently outperform individual models across diverse datasets and fraud scenarios. Research in this domain has progressively incorporated richer data sources including telematics data from vehicle sensors, geographic information about accident locations, and social network analysis of relationships among claimants, repair shops, and medical providers [32].

The theoretical foundations underlying DL applications to fraud detection draw from multiple disciplines including statistical learning theory, information theory, and game theory perspectives on adversarial behavior. Educational data mining and knowledge discovery methodologies provide frameworks for extracting meaningful patterns from complex datasets where fraudulent behaviors represent rare anomalies within predominantly legitimate activity [33]. The fundamental premise is that fraudulent behaviors, while deliberately designed to mimic legitimate activity, inevitably exhibit detectable statistical irregularities that can be learned by sufficiently flexible models trained on appropriate features [34]. However, the adversarial nature of fraud detection creates unique challenges not present in many other ML applications, as fraudsters actively adapt their strategies in response to detection systems, leading to a continuous arms race between fraud techniques and detection capabilities [35].

DL architectures offer particular advantages for this adversarial setting through their capacity for continuous learning and adaptation. Transfer learning approaches enable models trained on one insurance type or geographic region to be adapted for different contexts, reducing the need for large labeled datasets in every new deployment scenario [36]. Ensemble methods that combine predictions from multiple diverse models provide robustness against concept drift and adversarial manipulation attempts [37]. Recent advances in meta-learning and few-shot learning show promise for rapid adaptation to emerging fraud types based on limited examples [38]. These capabilities are particularly valuable given that new fraud schemes constantly emerge and labeled examples of novel fraud types are typically scarce in the period immediately following their introduction.

3. Deep Learning Architectures and Techniques

Convolutional Neural Networks represent one of the foundational DL architectures that has been successfully adapted from its original image processing applications to insurance fraud detection. CNNs employ specialized layers that apply learnable filters to input data, automatically discovering relevant local patterns and feature combinations without manual specification [39]. In insurance fraud detection contexts, CNNs treat structured tabular claim data as two-dimensional matrices where convolutional operations can identify spatial patterns across related variables [13]. The hierarchical feature learning capability of CNNs allows them to progressively build complex representations from simple features, with early layers detecting basic patterns and deeper layers combining these into sophisticated fraud indicators. Recent implementations have demonstrated that CNN-based models can achieve competitive performance with traditional ML approaches while offering greater flexibility and reducing the burden of manual feature engineering [30]. Xia, Zhou, and Zhang developed a CNN-LSTM hybrid model for auto insurance fraud detection that achieved 89.6% accuracy and 90.7% precision by automatically learning feature representations, significantly reducing the complexity and expert knowledge requirements associated with traditional feature engineering approaches [40]. The model demonstrated particular effectiveness in capturing subtle patterns in claim amounts, service provider relationships, and temporal characteristics that human-designed features might overlook.

Abakarim, Lahby, and Attioui proposed a CNN-based fraud detection model enhanced with ensemble bagging techniques that achieved 98% accuracy through the combination of multiple CNN models trained on different subsets of the data [41]. The ensemble approach provides robustness by aggregating predictions from diverse models, reducing the risk of overfitting to idiosyncrasies in the training data and improving generalization to new cases. The computational efficiency of CNNs makes them particularly attractive for real-time fraud screening applications where decisions must be made rapidly during claims processing. Modern CNN architectures can process thousands of claims per second on standard hardware, enabling their deployment as automated screening tools that flag suspicious cases for detailed human investigation [42]. However, CNNs face limitations when dealing with sequential temporal patterns and long-range dependencies in claim histories, motivating the development of hybrid architectures that combine CNNs with recurrent networks.

Recurrent Neural Networks and their advanced variants including LSTM and GRU networks are specifically designed to process sequential data by maintaining internal memory states that capture temporal dependencies [43]. Standard RNNs suffer from the vanishing and exploding gradient problem when processing long sequences, which limits their ability to capture long-term dependencies crucial for fraud detection where patterns may span multiple claims over extended time periods [44]. LSTMs address this fundamental limitation through a sophisticated gating mechanism comprising input gates that control what information enters memory, forget gates that determine what information to discard, and output gates that regulate what information to output from memory [45]. This architecture enables LSTMs to selectively retain relevant information over extended sequences while discarding irrelevant details, making them highly effective for modeling temporal patterns in insurance fraud scenarios.

Lai et al. employed LSTM networks to analyze brain injury insurance claims, achieving 74.33% accuracy in predicting fraudulent claims by capturing complex temporal patterns in medical treatment sequences that traditional methods struggled to identify [46]. The study demonstrated LSTM's capability to learn that certain sequences of treatments, the timing between procedures, and the progression of claimed symptoms contain subtle indicators of fabricated or exaggerated injuries. Research has shown that LSTM models exhibit sustained growth in application across financial fraud detection domains, with particularly sharp increases from 2022 to 2024, driven by the inherently sequential nature of transaction and claims data where temporal context provides critical information [17]. GRU networks represent a simplified variant of LSTMs that combine the forget and input gates into a single update gate and merge the cell state and hidden state, reducing computational complexity while maintaining competitive performance [47]. While GRUs require fewer parameters and train faster than LSTMs, empirical studies in fraud detection have generally shown LSTMs achieving slightly higher accuracy, particularly for complex sequential patterns requiring long-term memory capabilities, though the performance difference is often modest and context-dependent [48].

The integration of CNNs and LSTMs into hybrid architectures represents a significant advancement in insurance fraud detection, leveraging the complementary strengths of spatial feature extraction through convolutional layers and temporal dependency modeling through recurrent layers [49]. These hybrid models have consistently demonstrated superior performance compared to standalone architectures across multiple insurance fraud detection benchmarks, achieving state-of-the-art results on challenging datasets [40]. Reddy et al. proposed a multi-contextual modeling approach integrating CNN and bidirectional LSTM for financial fraud detection, achieving effective capture of both spatial and sequential dependencies in transaction patterns [50]. The bidirectional architecture allows the model to consider both past and future context when processing each time step, enhancing its ability to detect subtle fraudulent patterns that may only become apparent when examining complete claim sequences rather than processing them in strict temporal order. A comprehensive evaluation analyzing hybrid models for risk assessment in insurance companies demonstrated that CNN-LSTM architectures outperform standalone models in accurately assessing and categorizing fraud risk levels across diverse policyholder populations and claim types [51]. The CNN component extracts relevant features from structured claim data including amounts, service codes, provider characteristics, and geographic information, while the LSTM component models how these features evolve over time, capturing patterns like gradually escalating claim frequencies or systematic shifts in claim types that may indicate fraud.

Graph Neural Networks have emerged as a particularly powerful architecture for insurance fraud detection by explicitly modeling relational structures and network connections among entities involved in insurance claims [15]. Unlike traditional neural networks that process independent samples, GNNs operate on graph-structured data where nodes represent entities such as policyholders, healthcare providers, or claims, and edges represent relationships such as shared providers, co-occurrence in claims, or social connections [52]. This graph-based representation enables GNNs to capture complex fraud patterns that involve collusion among multiple actors, referral networks directing patients to complicit providers, or organized fraud rings coordinating their activities across many claims. Hong et al. proposed a multi-channel heterogeneous graph structure learning approach to detect health insurance fraud, utilizing diverse graph-based features from different claim aspects to capture complex relationships and patterns that substantially improved detection accuracy [53]. The multi-channel architecture processes multiple views of the data simultaneously, including patient-provider relationships, diagnosis-procedure associations, and temporal claim sequences, integrating these diverse perspectives to identify fraudulent patterns that might be invisible when examining any single view in isolation.

The effectiveness of GNNs for fraud detection stems from their ability to propagate information across the graph through message passing mechanisms, where each node aggregates information from its neighbors to update its representation [54]. This enables the model to identify suspicious patterns such as fraudsters who primarily connect with legitimate entities to camouflage their behavior, yet can still be detected through subtle differences in their network positions compared to truly legitimate actors. Research demonstrated that GNN-based models significantly outperformed baseline classifiers in credit card fraud detection by leveraging the transaction graph connecting users and merchants, a finding that generalizes to insurance fraud where relationships among claimants, providers, and referral sources contain valuable fraud signals [55]. GNNs have proven particularly effective for detecting organized fraud rings and collusion networks, achieving accuracies exceeding 84% in healthcare fraud detection tasks where traditional feature-based methods struggle because individual claims may appear legitimate when examined in isolation [16]. The heterogeneous graph structures common in insurance data, where multiple types of nodes and edges coexist, require specialized GNN architectures that can handle different relationship types and node attributes simultaneously [53].

Autoencoders and Variational Autoencoders (VAEs) represent unsupervised and semi-supervised DL architectures that have shown promise for fraud detection by learning compressed representations of normal behavior and using reconstruction error as an anomaly score [56]. Traditional autoencoders consist of an encoder network that compresses input data into a lower-dimensional latent representation and a decoder network that attempts to reconstruct the original input from this compressed form [57]. The fundamental insight for fraud detection is that autoencoders trained on predominantly legitimate claims will learn to accurately reconstruct normal patterns, while fraudulent claims that deviate from learned normal behavior will exhibit high reconstruction errors that can be thresholded to identify anomalies [58]. VAEs extend this framework by learning a probabilistic distribution in the latent space rather than fixed encodings, enabling both reconstruction and generation of new samples that resemble the training data [59]. This generative capability can be leveraged to create synthetic fraudulent samples for training purposes, addressing the class imbalance problem by augmenting the minority fraud class.

Generative Adversarial Networks (GANs) provide another approach to addressing class imbalance through synthetic data generation, consisting of a generator network that creates fake samples and a discriminator network that attempts to distinguish real from fake samples [60]. The adversarial training process where the generator tries to fool the discriminator while the discriminator tries to detect fakes leads both networks to improve, ultimately producing a generator capable of creating highly realistic synthetic fraudulent transactions that can augment training datasets [61]. Self-attention GANs leverage attention mechanisms to identify crucial features and patterns within extensive transaction datasets, fostering improved understanding and refined identification of fraud [62]. The self-attention mechanism allows the model to focus on the most relevant features for fraud detection rather than treating all input dimensions equally, improving both performance and interpretability by highlighting which factors drive fraud predictions.

4. Datasets and Performance Evaluation

The availability and characteristics of datasets fundamentally determine both the feasibility of DL research and the practical applicability of resulting models in real-world insurance fraud detection systems. Publicly available datasets enable reproducible research, facilitate comparative evaluation of different approaches, and lower barriers to entry for researchers lacking access to proprietary insurance data, yet such datasets remain limited in number and often lack the complexity and scale of real-world insurance operations [25]. The Medicare dataset represents one of the most widely used public resources for healthcare fraud detection research, containing claims data from the Centers for Medicare and Medicaid Services covering physician services, prescription drugs, and durable medical equipment [63]. The Medicare Part B dataset describes services and procedures that healthcare professionals provide to Medicare fee-for-service beneficiaries, including provider-level attributes such as National Provider Identifier, credentials, and address, along with claims information describing procedure codes, charge amounts, payment amounts, and service locations [64]. The Medicare Part D Prescriber dataset contains prescription drug information including drug names, costs, and prescriber characteristics, enabling analysis of potentially fraudulent prescribing patterns.

The DE-SynPUF dataset represents a synthetic version of Medicare claims data specifically designed to address privacy concerns while maintaining realistic statistical properties for research purposes [65]. This dataset consists of 66,773 insurance claim records covering the period 2008-2010 and has been employed in multiple fraud detection studies to develop and evaluate ML and DL approaches [25]. However, the synthetic nature of the data raises questions about how well findings generalize to real fraud patterns, as the data generation process may not fully capture the complex behavioral characteristics of actual fraudsters. Auto insurance fraud datasets are less commonly available in the public domain compared to healthcare data, with most research relying on proprietary datasets from specific insurance companies [30]. The limited public datasets that do exist often contain fewer features and smaller sample sizes compared to healthcare datasets, constraining the complexity of models that can be effectively trained and evaluated.

Table 1 presents a comprehensive comparison of commonly used datasets in insurance fraud detection research, including their characteristics, availability, and typical applications. The table demonstrates the significant variation in dataset size, fraud rate, and feature richness across different insurance domains, highlighting the challenges researchers face in selecting appropriate datasets for different research objectives.

Table 1. Comparison of commonly used dataset in insurance fraud detection research.

Dataset Name	Domain	Size (Records)	Fraud Rate	Features	Access	Primary Applications
Medicare Part B	Healthcare	~1.2 million	0.03-0.5%	45+	Public	Provider fraud detection, service pattern analysis, prescription fraud
Medicare Part D	Healthcare	~900,000	0.05-0.8%	38+	Public	Prescription fraud, opioid abuse detection, prescriber behavior analysis
CMS DE-SynPUF	Healthcare	66,773	1-3%	25-30	Public	Synthetic data for privacy-preserving research, algorithm benchmarking
Auto Insurance (Kaggle)	Auto	~15,000	5-8%	33	Public	Claim severity prediction, fabrication detection, exaggeration analysis
Private Auto Claims	Auto	100,000+	2-6%	50-80	Private	Collision fraud, staged accident detection, inflated repair costs
European Credit Card	Credit/Financial	284,807	0.172%	30 (PCA)	Public	Transaction fraud, anomaly detection, imbalanced learning techniques
IEEE-CIS Fraud	Credit/Financial	590,540	3.5%	434	Public	E-commerce fraud, device fingerprinting, behavioral analysis
Health Insurance Claims	Healthcare	50,000-500,000	1-4%	40-100	Private	Diagnosis coding fraud, unbundling, upcoding detection
Life Insurance	Life/Annuity	Variable	0.5-2%	20-50	Private	Application fraud, beneficiary fraud, premium evasion

Note: Data compiled from systematic reviews by du Preez et al. (2025), Hamid et al. (2024), Chen et al. (2025), and individual dataset documentation. Fraud rates vary by institution, detection methods, and data collection periods. PCA indicates features have been transformed through Principal Component Analysis for privacy protection. Size estimates reflect typical available versions; actual operational datasets may be significantly larger.

Evaluating the performance of fraud detection models requires careful selection of metrics appropriate for the extreme class imbalance characteristic of fraud datasets, where traditional accuracy measures can be misleading or meaningless [66]. Accuracy defined as the ratio of correct predictions to total predictions provides an intuitive overall performance measure but fails catastrophically on imbalanced data where a naive model predicting all cases as legitimate can achieve very high accuracy while completely missing all fraud cases [67]. For example, in a dataset where only 1% of claims are fraudulent, a model that predicts all claims as legitimate achieves 99% accuracy despite zero fraud detection capability, illustrating why accuracy is inappropriate as a primary metric for fraud detection. Precision defined as the proportion of positive predictions that are actually positive addresses the question of how many flagged cases are truly fraudulent, directly relating to operational efficiency as high precision minimizes wasted investigation effort on false alarms [68]. Precision is critical in contexts where investigation resources are limited and false positives create substantial costs through unnecessary investigations, damaged relationships with legitimate policyholders, or delayed claims processing.

Recall or sensitivity defined as the proportion of actual fraud cases correctly identified addresses the complementary question of how many true frauds the model successfully detects, directly relating to financial protection as high recall minimizes losses from undetected fraud [69]. Recall is paramount in contexts where missing fraud cases creates severe consequences including major financial losses, regulatory penalties for inadequate fraud prevention, or erosion of trust if fraud becomes widespread. The fundamental trade-off between precision and recall represents a core challenge in fraud detection system design, as increasing the sensitivity of detection by flagging more cases as suspicious inevitably increases false positives and reduces precision, while increasing precision by being more selective about what to flag inevitably misses more true frauds and reduces recall. The F1 score defined as the harmonic mean of precision and recall provides a balanced metric that accounts for both concerns, achieving its maximum value of one only when both precision and recall are perfect and generally tracking the lower of the two metrics [70].

Figure 2 illustrates the performance comparison of different deep learning architectures across key evaluation metrics including accuracy, precision, recall, and F1 score based on aggregated results from recent studies. The visualization demonstrates that hybrid CNN-LSTM models achieve the most balanced performance across all metrics, while GNN-based approaches excel particularly in precision for fraud network detection tasks.

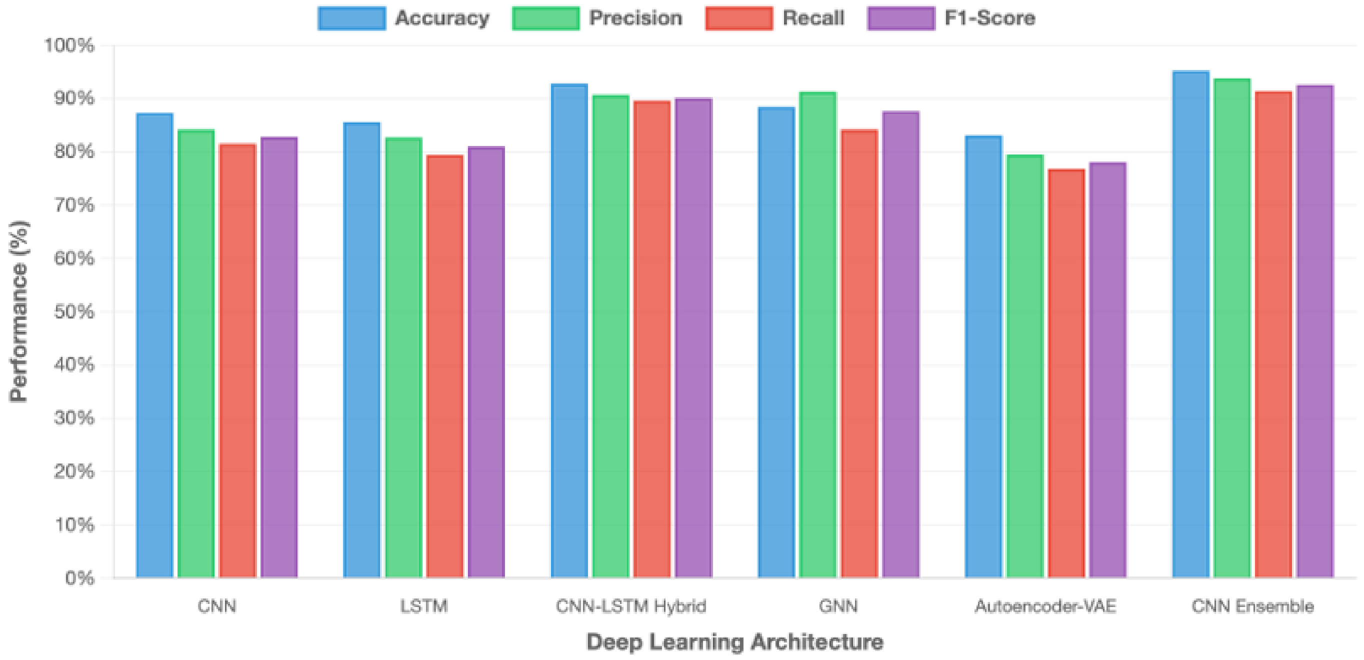

Figure 2. Performance comparison of deep learning architectures for insurance fraud detection.
Source: Data aggregated from recent studies: Xia et al. (2022) for CNN-LSTM hybrid, Hong et al. (2024) for GNN, Abakarim et al. (2023) for CNN ensemble, Lai et al. (2022) for LSTM, and systematic review by Chen et al. (2025). Metrics represent average performance across multiple datasets. Hybrid CNN-LSTM models demonstrate the most balanced performance across all metrics, while GNN excels in precision for fraud network detection tasks.

The Area Under the Receiver Operating Characteristic curve (AUC-ROC) provides a threshold-independent performance measure by evaluating classification performance across all possible decision thresholds [71]. The ROC curve plots true positive rate against false positive rate as the classification threshold varies, and the area under this curve ranges from 0.5 for random classification to 1.0 for perfect classification. However, AUC-ROC can be misleading on highly imbalanced datasets because the false positive rate in the denominator is calculated using the large number of negative examples, potentially obscuring poor performance on the minority positive class. The Area Under the Precision-Recall curve (AUC-PR) addresses this limitation by plotting precision against recall, focusing directly on performance on the positive class and providing more informative evaluation for imbalanced datasets [72].

Empirical comparisons across studies reveal that hybrid CNN-LSTM architectures consistently achieve among the highest performance levels for insurance fraud detection tasks. Xia, Zhou, and Zhang reported that their CNN-LSTM model for auto insurance fraud detection achieved 89.6% accuracy, 90.7% precision, and 89.6% recall, substantially outperforming standalone CNN and LSTM models evaluated on the same dataset [40]. The hybrid architecture demonstrated particular strength in capturing both spatial feature patterns through the CNN component and temporal claim sequence patterns through the LSTM component. Abakarim, Lahby, and Attioui achieved 98% accuracy using a CNN-based model with ensemble bagging for fraud detection, demonstrating the substantial performance gains possible from ensemble methods that combine multiple models [41]. The ensemble approach provided robustness by aggregating predictions from diverse models trained on different data subsets, effectively reducing overfitting and improving generalization to new cases.

5. Explain Ability and Emerging Trends

The black box nature of many high-performing DL architectures poses significant challenges for practical fraud detection deployment where stakeholders require understandable explanations for algorithmic decisions affecting individuals [22]. Explainable AI (XAI) frameworks aim to make DL models more interpretable by providing insights into which features drive predictions, how the model arrives at particular decisions, and why certain cases receive high fraud scores [73]. The importance of XAI for fraud detection stems from multiple factors including regulatory requirements in many jurisdictions that mandate transparency in automated decision-making affecting individuals, practitioner needs for fraud investigators to understand why cases were flagged to conduct effective investigations, model debugging and validation to identify and correct biases or errors in model logic, and stakeholder trust building confidence among claims processors, policyholders, and regulators that automated systems make reasonable decisions.

SHapley Additive exPlanations (SHAP) represents one of the most widely adopted XAI techniques, providing a unified framework for interpreting model predictions based on game theory principles [74]. SHAP values quantify each feature's contribution to a particular prediction by calculating the expected change in model output when that feature is included versus excluded, considering all possible feature combinations. The additive nature of SHAP values enables intuitive interpretation where features with positive SHAP values push predictions toward fraud while features with negative SHAP values push toward legitimate, and the magnitude reflects the strength of influence. Research applying SHAP to fraud detection has demonstrated that the technique successfully identifies the most influential features distinguishing fraudulent from legitimate claims, providing actionable insights for investigators and enabling validation that models rely on sensible fraud indicators rather than spurious correlations [75]. Hosseini Chagahi et al. employed SHAP to identify the top ten most important features for distinguishing fraud from normal transactions in credit card fraud detection, using these features in their attention-based model to achieve high accuracy and robust generalization [76].

Local Interpretable Model-agnostic Explanations (LIME) provides an alternative XAI approach that explains individual predictions by approximating the complex DL model locally with a simpler interpretable model such as linear regression or decision tree [77]. LIME perturbs the input features of a specific instance and observes how predictions change, then fits an interpretable model to these perturbations weighted by proximity to the instance of interest. This local approximation reveals which features were most influential for that particular prediction without requiring access to the global model structure. LIME has been successfully applied to fraud detection to generate case-specific explanations that investigators can use to understand why particular claims received high fraud scores, though the technique faces limitations including potential instability where small changes in sampling of perturbations can produce different explanations.

Attention mechanisms integrated directly into neural network architectures provide inherent interpretability by learning to focus on the most relevant inputs for prediction [78]. Attention weights assigned to different features, time steps, or graph nodes indicate their relative importance for the model's decision, enabling interpretation of what the model considers most relevant. Farbmacher et al. developed an Explainable Attention Network specifically for fraud detection in claims management, where attention weights highlight the most critical features of fraudulent behavior enabling transparent decision-making [79]. The integration of attention provides superior interpretability compared to post-hoc explanation methods because the attention weights directly reflect what information the model actually used rather than attempting to reverse-engineer the reasoning process after the fact.

Federated learning has emerged as a promising approach to address data privacy concerns and enable collaborative model training across multiple insurance institutions without centralizing sensitive data [80]. In federated learning, each participating institution trains a local model on their private data, and only model parameters or gradients are shared with a central coordinator that aggregates updates to improve a global model. This approach enables institutions to benefit from larger effective training datasets and diverse fraud patterns while maintaining data privacy and regulatory compliance. Recent pilot implementations have demonstrated that federated learning can achieve comparable accuracy to centralized training while providing stronger privacy guarantees [81]. The technique is particularly valuable for insurance fraud detection where individual institutions may have limited labeled fraud examples, but collaborative learning across multiple institutions could substantially improve detection capabilities.

Quantum machine learning represents an emerging frontier that may offer computational advantages for certain fraud detection tasks, though practical applications remain primarily experimental at this stage [82]. Quantum algorithms leveraging superposition and entanglement could potentially accelerate training of complex models or enable more efficient exploration of high-dimensional feature spaces. Recent theoretical work has explored quantum graph neural networks for fraud detection, suggesting potential advantages for analyzing complex relationship structures, though significant technical challenges remain in scaling these approaches to real-world problem sizes. As quantum computing hardware continues to mature, this represents an area for continued monitoring and potential future application.

Blockchain integration with machine learning has been proposed as an approach to enhance data integrity and transparency in fraud detection systems [83]. Blockchain technology can provide immutable audit trails of claims data, model predictions, and investigation outcomes, enabling verification of system operations and facilitating regulatory compliance. Some implementations have explored using blockchain to securely share fraud intelligence across institutions while maintaining privacy protections. The combination of blockchain's transparency and immutability with ML's analytical power offers potential synergies, though practical implementations must carefully balance the benefits against the computational costs and complexity of blockchain systems.

6. Conclusion

This comprehensive review has examined the application of deep learning to insurance fraud detection through systematic analysis of 57 studies published between 2019 and 2025, revealing substantial progress in both methodological sophistication and empirical performance. The evidence demonstrates that DL approaches, particularly hybrid architectures combining CNNs with LSTMs and GNN-based models for relational fraud detection, consistently outperform traditional rule-based and classical ML methods across healthcare, auto, and life insurance domains. Hybrid CNN-LSTM models achieve accuracies ranging from 89.6% to 98% on standard benchmarks, representing substantial improvements over traditional approaches, while GNNs demonstrate particular effectiveness for detecting collusive fraud networks with accuracies exceeding 84%. The automatic feature learning capabilities of DL eliminate the need for extensive manual feature engineering, enabling models to discover subtle fraud indicators that human experts might overlook.

However, significant challenges continue to impede both research advancement and practical deployment. The severe class imbalance characteristic of fraud datasets, where fraudulent cases represent only 0.03-3% of total claims, creates fundamental difficulties for learning algorithms that tend to optimize overall accuracy. While techniques including synthetic oversampling, cost-sensitive learning, and ensemble methods provide partial solutions, no approach fully resolves this challenge. Model interpretability remains a critical concern, as many high-performing DL architectures operate as black boxes providing predictions without transparent reasoning, conflicting with regulatory requirements and practitioner needs for explainable decisions. Current XAI techniques including SHAP, LIME, and attention mechanisms provide valuable but imperfect solutions to interpretability challenges.

Data availability represents another major obstacle, with limited public datasets constraining reproducible research and comparative evaluation. Privacy regulations impose stringent constraints on data collection and sharing, creating barriers to both research and deployment. The adversarial nature of fraud detection, where fraudsters actively adapt strategies in response to detection systems, requires continuous model updating and robust architectures resistant to concept drift. Future research should prioritize development of inherently interpretable DL architectures that provide transparent reasoning by design, creation of high-quality benchmark datasets through partnerships between researchers and insurance organizations, exploration of federated learning approaches enabling collaborative training across institutions without centralizing sensitive data, investigation of transfer learning techniques leveraging knowledge from data-rich domains to improve detection in data-scarce contexts, and development of real-time detection systems with low-latency inference suitable for online claims processing.

The integration of emerging technologies including quantum computing, blockchain for data integrity, and advanced privacy-preserving techniques offers promising directions for future work. Causal inference methods that understand not just correlations but actual causal relationships between variables and fraud could improve robustness and interpretability. Meta-learning approaches enabling rapid adaptation to new fraud types from limited examples represent important research directions given the constantly evolving fraud landscape. The development of comprehensive evaluation frameworks that go beyond technical performance metrics to assess real-world operational impact, cost-effectiveness, and fairness across different demographic groups will be essential for responsible deployment of DL fraud detection systems. As the field continues to mature, the focus must remain on developing systems that not only achieve high detection accuracy but also operate transparently, fairly, and in accordance with ethical principles and regulatory requirements.

References

Du Preez, A., Bhattacharya, S., Beling, P., & Bowen, E. (2025). Fraud detection in healthcare claims using machine learning: A systematic review. *Artificial Intelligence in Medicine, 160,* 103061.

Federal Trade Commission. (2022, February 22). *Consumer Sentinel Network Data Book 2021.* U.S. Federal Trade Commission.

Woodson, V. L. (2024). *Evaluating fraudulent auto insurance claims: The role of technology and law enforcement in detection and prevention* (Doctoral dissertation, Saint Leo University).

Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making, 24,* 112.

Vyas, S., & Serasiya, S. (2022). Fraud detection in insurance claim system: A review. In *Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy* (pp. 922–927). IEEE.

Bello, O. A., & Olufemi, K. (2024). Artificial intelligence in fraud prevention: Exploring techniques and applications, challenges and opportunities. *Computer Science and IT Research Journal, 5*(6), 1505–1520.

Vemulapalli, G. (2024). Fighting fraud with algorithms: AI solutions for claim detection and revolutionizing fraud detection in insurance. In *Artificial Intelligence and Machine Learning for Sustainable Development* (pp. 125–140). CRC Press.

Williamson, B. (2019). Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education. *British Journal of Educational Technology, 50*(6), 2794–2809.

Steffens, T. (2020). *Attribution of advanced persistent threats: How to identify the actors behind cyber-espionage.* Springer Nature.

Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021* (pp. 3168–3177).

Taye, M. M. (2023). Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers, 12*(5), 91.

Al Doulat, A., Ayo-Bali, O. E., & Shaik, S. (2025, July). Fraud detection in insurance claims using supervised machine learning models. In *2025 International Conference on Smart Applications, Communications and Networking (SmartNets)* (pp. 1–7). IEEE.

Badr, B. E., Altawil, I., Almomani, M., Al-Saadi, M., & Alkhurainej, M. (2023). Fault diagnosis of three-phase induction motors using convolutional neural networks. *Mathematical Modelling of Engineering Problems, 10*(5).

Toufik, G., Khaldi, Y., Pandey, P. S., & Abusal, Y. A. (2024). Advanced fraud detection in card-based financial systems using a bidirectional LSTM-GRU ensemble model. *Applied Computer Science, 20*(3).

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24.

Rahman, M. M. (2025). Data-driven graph neural network models for detecting fraudulent insurance claims in healthcare systems. *American Journal of Interdisciplinary Studies, 6*(1), 263–294.

Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. *arXiv preprint arXiv:2502.00201.*

Fursov, I., Kovtun, E., Rivera-Castro, R., Zaytsev, A., Khasyanov, R., Spindler, M., & Burnaev, E. (2022). Sequence embeddings help detect insurance fraud. *IEEE Access, 10,* 54326–54339.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1), 3.

Mwangi, E. (2024). Employing AI/ML to determine and mitigate fraud in the insurance industry. *SSRN.* https://doi.org/10.2139/ssrn.4907329

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning, 113*(7), 4845–4901.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education, 60*(7), 1048–1064.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94,* 335–343.

Becker, J., Efstathiades, A., Portmann, J., & Zeier Röschmann, A. (2024). *Data competences in the insurance industry.* ZHAW / Cognizant / AWS. https://doi.org/10.21256/zhaw-2501

Vallarino, D. (2025). Detecting financial fraud with hybrid deep learning: A mix-of-experts approach to sequential and anomalous patterns. *arXiv preprint arXiv:2504.03750.*

Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. *International Journal of Modern Education and Computer Science, 13*(1), 1.

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2019). Implementing AutoML in educational data mining for prediction tasks. *Applied Sciences, 10*(1), 90.

Subudhi, S., & Panigrahi, S. (2020). Use of optimized fuzzy c-means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University — Computer and Information Sciences, 32*(5), 568–575.

Schrijver, C. J., Tan, S., Boerkamp, B. J., & Simons, M. (2024). Automobile insurance fraud detection using data mining: A systematic literature review. *Expert Systems with Applications, 213,* 119153.

Chauhan, V., & Yadav, J. (2024). Bibliometric review of telematics-based automobile insurance: Mapping the landscape of research and knowledge. *Accident Analysis & Prevention, 196,* 107428.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.

Pawar, K., & Attar, V. (2019). Deep learning approaches for video-based anomalous activity detection. *World Wide Web, 22*(2), 571–601.

Ijiga, O. M., Idoko, I. P., Ebiega, G. I., Olajide, F. I., Olatunde, T. I., & Ukaegbu, C. (2024). Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention. *Journal of Science and Technology, 11*, 1–24.

Iman, M., Arabnia, H. R., & Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies, 11*(2), 40.

Azeem, A., Ismail, I., Mohani, S. S., Danyaro, K. U., Hussain, U., Shabbir, S., & Jusoh, R. Z. B. (2025). Mitigating concept drift challenges in evolving smart grids: An adaptive ensemble LSTM for enhanced load forecasting. *Energy Reports, 13*, 1369–1383.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5149–5169.

Suresh, S., & Mohan, S. (2020). ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis. *Neural Computing and Applications, 32*(20), 15989–16009.

Xia, P., Zhou, H., & Zhang, L. (2022). Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model. *International Journal of Ad Hoc and Ubiquitous Computing, 39*(1–2), 37–49.

Abakarim, Y., Lahby, M., & Attioui, A. (2023). A bagged ensemble convolutional neural networks approach to recognize insurance claim frauds. *Applied System Innovation, 6*(1), 20.

Azad, T., & William, P. (2024). Fraud detection in healthcare billing and claims. *International Journal of Science and Research, 13*(02), 3375–3395.

Nosouhian, S., Nosouhian, F., & Khoshouei, A. K. (2021). A review of recurrent neural network architecture for sequence learning: Comparison between LSTM and GRU. *Preprint (open access)*. https://doi.org/10.20944/preprints202107.0252.v1

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information, 15*(9), 517. https://doi.org/10.3390/info1509-0517

Wang, B., Kong, W., Guan, H., & Xiong, N. N. (2019). Air quality forecasting based on gated recurrent long short term memory model in Internet of Things. *IEEE Access, 7*, 69524–69534.

Mahveen, Z. (2025). Optimizing fraud detection in healthcare: A hybrid machine learning approach. *Manuscript in preparation.*

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270.

Heald, J. B., Wolpert, D. M., & Lengyel, M. (2023). The computational and neural bases of context-dependent learning. *Annual Review of Neuroscience, 46*(1), 233–258.

Dunka, V. (2023). AI-driven claims fraud detection using hybrid deep learning models: Integrating convolutional neural networks and recurrent neural networks for real-time fraud detection in insurance claims. *Essex Journal of AI Ethics and Responsible Innovation, 3*, 276–311.

Reddy, N. M., Sharada, K. A., Pilli, D., Paranthaman, R. N., Reddy, K. S., & Chauhan, A. (2023). CNN-bidirectional LSTM based approach for financial fraud detection and prevention system. In *2023 International Conference on Sustainable Computing and Smart Systems* (pp. 541–546). IEEE.

Chidananda, A. (2025). *Credit card fraud detection using hybrid deep learning CNN-LSTM and CNN-GRU models* (Doctoral dissertation, California State University, Northridge).

Monchka, B. A., Leung, C. K., Nickel, N. C., & Lix, L. M. (2022). The effect of disease co-occurrence measurement on multimorbidity networks: A population-based study. *BMC Medical Research Methodology, 22*(1), 165.

Hong, X., Wang, H., Zhang, Y., & Liu, J. (2024). Health insurance fraud detection based on multi-channel heterogeneous graph structure learning. *Heliyon, 10*(9), e30045.

Fan, X., Gong, M., Wu, Y., Qin, A. K., & Xie, Y. (2021). Propagation enhanced neural message passing for graph representation learning. *IEEE Transactions on Knowledge and Data Engineering, 35*(2), 1952–1964.

Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020, October). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 315–324).

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. (2024). Autoencoders and their applications in machine learning: A survey. *Artificial Intelligence Review, 57*(2), 28.

Chen, S., & Guo, W. (2023). Auto-encoders in deep learning—a review with new perspectives. *Mathematics, 11*(8), 1777.

Zavrak, S., & İskefiyeli, M. (2020). Anomaly-based intrusion detection from network flow features using variational autoencoder. *IEEE Access, 8*, 108346–108358.

Bhalodia, R., Lee, I., & Elhabian, S. (2020). DPVAEs: Fixing sample generation for regularized VAEs. In *Proceedings of the Asian Conference on Computer Vision.*

Ali-Gombe, A., & Elyan, E. (2019). MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial networks. *Neurocomputing, 361*, 212–221.

Liu, X., & Hsieh, C. J. (2019). Rob-GAN: Generator, discriminator, and adversarial attacker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11234–11243).

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning* (pp. 7354–7363). PMLR.

Hotz, A., Sprecher, E., Bastianelli, L., Rodean, J., Stringfellow, I., Barkoudah, E., … Berry, J. G. (2023). Categorization of a universal coding system to distinguish use of durable medical equipment and supplies in pediatric patients. *JAMA Network Open, 6*(10), e2339449.

Chan, G. K., Cummins, M. R., Taylor, C. S., Rambur, B., Auerbach, D. I., Meadows-Oliver, M., … Pittman, P. P. (2023). An overview and policy implications of national nurse identifier systems: A call for unity and integration. *Nursing Outlook, 71*(2), 101892.

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digital Health, 2*(1), e0000082.

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061.*

Khalil, A. A., Liu, Z., Fathalla, A., Ali, A., & Salah, A. (2024). Machine learning based method for insurance fraud detection on class imbalance datasets with missing values. *IEEE Access.* (Advance online publication)

Beneish, M. D., & Vorst, P. (2022). The cost of fraud prediction errors. *The Accounting Review, 97*(6), 91–121.

Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J. D. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy, 3*, e12.

Foody, G. M. (2023). Challenges in the real-world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLOS ONE, 18*(10), e0291908.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution, 10*(4), 565–577.

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2023). The ROC-AUC accurately assesses imbalanced datasets. *SSRN.* https://doi.org/10.2139/ssrn.4655233

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S., & Qiu, G. (2025). A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud. *Scientific Reports, 15*(1), 218.

Hosseini Chagahi, M., Mohammadi Dashtaki, S., Moshiri, B., & Piran, M. J. (2024). Explainable AI for fraud detection: An attention-based ensemble of CNNs, GNNs, and a confidence-driven gating mechanism. *arXiv preprint arXiv:2410.09069.*

Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525–541.

Li, A., Xiao, F., Zhang, C., & Fan, C. (2021). Attention-based interpretable neural network for building cooling load prediction. *Applied Energy*, *299*, 117238.

Farbmacher, H., Ihle, P., Schubert, I., & Winter, J. (2022). Explainable attention network for fraud detection in claims management. *Health Economics*, *31*(12), 2599–2615.

Dong, P., Quan, Z., Edwards, B., Wang, S. H., Feng, R., Wang, T., … Shah, P. (2024). Privacy-enhancing collaborative information sharing through federated learning – A case of the insurance industry. *arXiv preprint arXiv:2402.14983*.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems* (pp. 429–450).

Innan, N., Sawaika, A., Dhor, A., Dutta, S., Thota, S., Gokal, H., & Bennai, M. (2023). Financial fraud detection using quantum graph neural networks. *Quantum Machine Intelligence*, *5*(2), 27.

Sahu, A., Kumar, R., Behera, R. K., & Rath, S. K. (2024). Blockchain and machine learning based secure driver behavior-centric insurance model for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *25*(8), 9632–9645.