MULTIDISCIPLINARY RESEARCH IN COMPUTING INFORMATION SYSTEMS

VOL 06 ISSUE 01 2026

https://mrcis.org

Deep Sequence Models For Real-Time Journal Entry Anomaly Detection In Financial Reporting

Zirui Tang
Department of Computer Science, Purdue University, USA
Haonan Qiu
Department of Computer Science, Purdue University, USA
Lukas Steiner
Institute of Information Systems and Digital Business, University of St. Gallen, Switzerland

Abstract: *Financial reporting integrity remains a critical concern for auditors, regulatory bodies, and stakeholders as the complexity and volume of accounting transactions continue to escalate. Traditional rule-based audit sampling techniques face increasing challenges in identifying fraudulent activities and accounting errors concealed within massive general ledger datasets. This research explores the application of deep sequence models, particularly Long Short-Term Memory (LSTM) networks and their variants, for real-time anomaly detection in journal entry data. By treating journal entries as temporal sequences and leveraging the memory capabilities of recurrent architectures, our approach captures intricate patterns and dependencies that characterize normal accounting behavior. The proposed framework addresses key challenges including variable-length transaction structures, class imbalance in anomaly datasets, and the need for unsupervised or semi-supervised learning paradigms where labeled fraudulent instances remain scarce. We demonstrate how LSTM's sophisticated gating mechanisms enable selective information retention across extended temporal horizons, while stacked architectures learn hierarchical representations of transaction patterns. The methodology integrates preprocessing techniques to handle categorical and numerical features, followed by deep learning architectures that model both short-term transactional patterns and long-range dependencies across accounting periods. Experimental validation using synthetic and real-world financial datasets shows that sequence-based deep learning models achieve superior performance metrics compared to conventional statistical methods and shallow machine learning techniques, with particular improvements in separating anomalous transactions based on reconstruction error distributions. The reconstruction-based detection paradigm demonstrates clear discrimination between normal entries exhibiting low reconstruction errors and fraudulent entries with significantly elevated error magnitudes. This work contributes to the growing intersection of artificial intelligence and forensic accounting by providing a scalable, adaptable solution for automated anomaly detection in enterprise financial systems.*

Keywords: *anomaly detection, journal entries, LSTM networks, financial reporting, deep learning, sequence modeling, fraud detection, general ledger analysis*

Introduction

The integrity of financial reporting systems stands as a cornerstone of modern economic infrastructure, yet the increasing sophistication of fraudulent schemes and the exponential growth in transaction volumes present unprecedented challenges for audit professionals and internal control mechanisms. Financial statement fraud costs the global economy billions of dollars annually while undermining investor confidence and market stability [1]. Traditional audit methodologies rely heavily on statistical sampling and manual review of journal entries, approaches that become progressively inadequate when confronted with the scale and complexity of contemporary enterprise resource planning systems. The general ledger, serving as the comprehensive record of all financial transactions within an organization, contains journal entries that encode the fundamental accounting activities driving business operations. Within this vast repository of structured data, anomalous entries may represent anything from benign data entry errors to deliberate attempts at financial manipulation, making their accurate identification a task of paramount importance [2]. The challenge of anomaly detection in journal entry data differs fundamentally from other fraud detection domains due to several unique characteristics inherent to accounting systems. Each journal entry comprises multiple transaction lines that must satisfy the fundamental accounting equation while adhering to organizational chart-of-account structures and business rule constraints. The temporal nature of accounting activities introduces sequential dependencies where current transactions relate meaningfully to historical patterns, seasonal business cycles, and regulatory reporting periods. Furthermore, the rarity of genuine anomalies within typical financial datasets creates severe class imbalance problems that traditional machine learning approaches struggle to address effectively. Auditors must navigate this complex landscape where false positives generate costly manual review overhead while false negatives potentially allow fraudulent activities to remain undetected until material damage occurs [3]. Recent advances in deep learning have opened new possibilities for automated anomaly detection in sequential data domains, with recurrent neural network architectures demonstrating particular promise for capturing temporal dependencies and long-range patterns. Long Short-Term Memory networks, initially developed to address the vanishing gradient problems that limited earlier recurrent architectures, possess an internal memory mechanism enabling the modeling of complex sequential relationships across extended time horizons. The sophisticated gating architecture within LSTM cells allows selective retention of relevant historical information while preventing the vanishing gradient problems that plague traditional recurrent neural networks when processing long sequences [4]. These capabilities align remarkably well with the requirements of financial transaction analysis where understanding the context and historical precedent of current activities proves essential for distinguishing legitimate business operations from suspicious deviations. The application of LSTM-based models to journal entry anomaly detection represents a natural evolution from rule-based systems toward data-driven approaches that can automatically learn the intricate patterns characterizing normal accounting behavior from large-scale historical datasets [5]. The motivation for this research emerges from the convergence of two critical trends reshaping the audit profession. First, regulatory pressures and stakeholder expectations increasingly demand continuous monitoring capabilities rather than periodic sampling-based audits, necessitating automated systems capable of real-time anomaly detection across comprehensive transaction populations. Second, the adoption of advanced analytics and artificial intelligence technologies within accounting firms has accelerated dramatically, creating both opportunities and imperatives for developing sophisticated fraud detection methodologies that leverage cutting-edge machine learning techniques and AI-driven systems designed to enhance accounting accuracy and financial transparency [6]. Deep sequence models offer a compelling solution to these converging demands

by providing scalable frameworks that can process streaming journal entry data, identify subtle anomalous patterns that evade traditional rule-based filters, and adapt to evolving business environments through continuous learning from new transaction data [7]. This study investigates the design, implementation, and evaluation of deep sequence models specifically tailored for real-time journal entry anomaly detection in financial reporting contexts. We examine how LSTM architectures effectively model the temporal dynamics of accounting transactions through their specialized memory mechanisms, how stacked configurations enable hierarchical feature learning across multiple levels of abstraction, and how reconstruction-based detection paradigms provide clear discrimination between normal and anomalous transaction patterns. The research addresses practical considerations including feature engineering approaches for heterogeneous accounting data, strategies for mitigating class imbalance through specialized training objectives, and deployment architectures that enable continuous monitoring without introducing unacceptable latency into operational systems [8]. Through comprehensive experimental evaluation using both synthetic anomaly datasets and real-world general ledger data, we demonstrate that deep sequence models achieve substantial improvements over baseline approaches while maintaining the computational efficiency required for real-time applications, with reconstruction error distributions showing clear separation between normal and fraudulent transaction classes.

2. Literature Review

The detection of anomalies within financial reporting systems has evolved substantially over the past decades, progressing from purely manual audit procedures through rule-based expert systems to contemporary machine learning approaches that leverage the unprecedented availability of digital transaction data. Early research in accounting fraud detection focused primarily on statistical techniques including Benford's law analysis and ratio-based analytical procedures that flag transactions or accounts exhibiting unusual numerical characteristics. These methods, while providing valuable screening mechanisms, suffer from high false positive rates and limited capability to adapt to the specific patterns present within individual organizational contexts. The emergence of data mining techniques in the late 1990s and early 2000s introduced supervised learning algorithms such as decision trees, logistic regression, and support vector machines that could learn fraud detection rules from historical labeled datasets, achieving improved performance over purely statistical approaches but still facing significant challenges with the extreme class imbalance typical of financial fraud scenarios [9]. The application of machine learning to general ledger analysis gained substantial momentum during the 2010s as enterprise systems began generating increasingly comprehensive digital trails of all financial transactions. Research demonstrated that supervised learning models trained on features extracted from journal entries including account combinations, user identifiers, posting times, and monetary amounts could identify potentially fraudulent transactions with reasonable accuracy when sufficient labeled training data became available [10]. However, the fundamental limitation of supervised approaches in the fraud detection domain stems from the scarcity of confirmed fraud cases relative to the overwhelming volume of legitimate transactions, making it extremely difficult to obtain the balanced labeled datasets that traditional classification algorithms require for effective learning. This realization motivated growing interest in unsupervised and semi-supervised anomaly detection methodologies that could identify unusual patterns without requiring extensive fraud labels, instead learning representations of normal transaction behavior and flagging deviations from learned patterns as potential anomalies [11]. Deep learning revolutionized anomaly detection across numerous application domains beginning in the mid-2010s, with neural network architectures demonstrating unprecedented capability to automatically learn hierarchical feature representations from raw or minimally processed data. Autoencoders emerged as a particularly influential architecture for unsupervised anomaly detection, operating on the principle that

networks trained to reconstruct normal data will exhibit higher reconstruction errors when presented with anomalous inputs that violate the learned patterns [12]. The extension of autoencoders to sequential data through recurrent architectures introduced powerful new capabilities for temporal anomaly detection in domains ranging from industrial sensor monitoring to cybersecurity intrusion detection. LSTM networks, with their sophisticated gating mechanisms enabling selective retention and forgetting of information across long sequences, proved especially effective for modeling complex temporal dependencies in time series data where anomalies manifest as deviations from expected sequential patterns rather than as isolated statistical outliers [13]. Research specifically targeting journal entry anomaly detection using deep learning methodologies has accelerated rapidly since 2019, driven by both technological advances in neural network architectures and increasing recognition within the accounting profession of the transformative potential of artificial intelligence. Multiple studies have demonstrated that LSTM-based models can effectively learn the temporal patterns inherent in financial transaction sequences, capturing dependencies between related journal entries that span days, weeks, or months as business processes unfold [14]. The encoder-decoder architecture, originally developed for sequence-to-sequence tasks in natural language processing, has found particularly productive application in financial anomaly detection where it learns compressed representations of normal transaction patterns and uses reconstruction error as an anomaly signal. This approach aligns naturally with audit scenarios where examples of normal transactions vastly outnumber fraudulent cases, enabling effective model training without requiring extensive fraud labels [15]. Attention mechanisms represent a critical innovation enhancing the effectiveness and interpretability of sequence models for anomaly detection applications. By learning to focus computational resources on the most informative elements within input sequences, attention enables models to capture long-range dependencies more effectively than traditional recurrent architectures while simultaneously providing insights into which specific transactions or transaction features contribute most significantly to anomaly scores [16]. Research has shown that attention-augmented LSTM models achieve superior detection performance compared to baseline recurrent architectures across diverse financial fraud detection tasks including credit card transaction monitoring, insurance claim analysis, and general ledger auditing. The interpretability benefits of attention prove especially valuable in audit contexts where regulatory requirements and professional standards demand that automated detection systems provide transparent explanations for their decisions [17]. The transformer architecture, built entirely on attention mechanisms without recurrent connections, has emerged as a powerful alternative to LSTM-based models for sequence anomaly detection since its introduction in 2017 and subsequent adaptation to time series domains. Transformer models can process entire sequences in parallel rather than requiring sequential computation, enabling significant computational efficiency gains for both training and inference. Research demonstrates that transformer architectures achieve competitive or superior anomaly detection performance compared to recurrent models on financial transaction datasets while offering the additional benefit of more interpretable attention weight visualizations that highlight suspicious transaction patterns [18]. The anomaly transformer specifically designed for time series anomaly detection introduces innovative mechanisms for computing association discrepancy between normal and anomalous points, achieving state-of-the-art results across multiple benchmarks and suggesting promising directions for application to journal entry analysis [19]. Graph neural networks have recently attracted attention as an alternative paradigm for financial fraud detection that explicitly models the relational structure connecting entities within financial systems. Unlike sequence models that treat transactions as temporally ordered but otherwise independent events, graph-based approaches represent the network of relationships between accounts, users, counterparties, and other entities involved in financial activities. Research shows

that graph convolutional networks and graph attention networks can identify fraud patterns characterized by unusual interaction structures or suspicious communities within financial networks, complementing the temporal pattern recognition capabilities of sequence models [20]. The potential for hybrid architectures combining sequential and graph-based representations represents an exciting frontier for next-generation fraud detection systems capable of capturing both temporal dynamics and structural relationships simultaneously [21]. Variational autoencoders and generative adversarial networks introduce probabilistic and adversarial learning paradigms to anomaly detection that offer theoretical and practical advantages over deterministic reconstruction-based approaches. Variational autoencoders learn probability distributions over latent representations rather than point estimates, enabling more principled anomaly scoring through likelihood estimation under the learned generative model [22]. Generative adversarial networks pit generator and discriminator networks against each other in a minimax game, potentially enabling the learning of more nuanced boundaries between normal and anomalous data distributions. While these advanced generative approaches have shown promise in various anomaly detection domains, their application to financial transaction analysis remains relatively limited compared to autoencoder and recurrent architectures, presenting opportunities for future research [23]. The practical deployment of deep learning models for real-time anomaly detection in production financial systems introduces numerous challenges beyond achieving high detection performance on offline evaluation datasets. Model interpretability and explainability emerge as critical requirements since auditors and compliance officers must understand why specific transactions receive high anomaly scores before initiating investigations or taking corrective actions [24]. Adversarial robustness concerns arise as sophisticated fraudsters may attempt to craft transactions specifically designed to evade detection by learned models, necessitating defensive training procedures and continuous model monitoring. Regulatory compliance considerations including data privacy, model governance, and audit trail requirements impose additional constraints on acceptable detection systems [25]. Recent research has begun addressing these deployment challenges through techniques including attention-based explanation mechanisms, adversarial training procedures, and federated learning approaches that enable model training across multiple organizations without sharing sensitive transaction data [26]. The evolution of evaluation methodologies for anomaly detection systems represents another important strand of research addressing the limitations of traditional metrics like accuracy and AUC-ROC when applied to severely imbalanced datasets. Precision-recall curves and F-beta scores that appropriately weight the relative importance of false positives versus false negatives provide more meaningful performance measures for fraud detection applications where the costs of different error types differ dramatically [27]. Time-aware evaluation protocols that respect the temporal ordering of transactions and simulate realistic operational scenarios including concept drift and delayed label feedback offer more reliable estimates of how models will perform in production deployment compared to standard cross-validation procedures [28]. The development of standardized benchmark datasets and evaluation protocols for journal entry anomaly detection would significantly benefit the research community by enabling more systematic comparison of competing approaches [29]. Emerging directions in deep learning research promise to further enhance anomaly detection capabilities in financial domains. Self-supervised learning techniques that leverage unlabeled data through pretext tasks enable the learning of robust representations without requiring extensive manual labeling, addressing one of the fundamental bottlenecks in fraud detection model development [30]. Transfer learning approaches that adapt models pretrained on large diverse datasets to specific organizational contexts could accelerate deployment and improve performance for organizations with limited historical fraud data. The integration of domain knowledge and accounting rules into neural network architectures through

structured priors, auxiliary losses, or physics-informed neural networks represents a promising direction for combining the pattern recognition capabilities of deep learning with the interpretability and reliability of rule-based systems. As these methodological innovations mature and computing infrastructure continues advancing, the capabilities of automated anomaly detection systems will expand further, reshaping the practice of audit and financial oversight.

3. Methodology

Our proposed framework for real-time journal entry anomaly detection employs a multi-stage pipeline encompassing data preprocessing, feature engineering, deep sequence model architecture, training procedures, and inference mechanisms designed for continuous monitoring applications. The methodology addresses the unique characteristics of journal entry data including variable transaction lengths, heterogeneous feature types mixing categorical and numerical attributes, temporal dependencies spanning multiple accounting periods, and severe class imbalance between normal and anomalous entries. This section provides comprehensive technical details of each pipeline component enabling reproducibility and practical implementation in production financial systems.

3.1 Data Preprocessing and Feature Engineering

The preprocessing stage transforms raw journal entry data extracted from enterprise resource planning systems into structured representations suitable for deep learning model input. Each journal entry consists of multiple transaction lines recording debits and credits to various general ledger accounts along with associated metadata including posting date, user identifier, document reference number, and descriptive text fields. The fundamental challenge lies in creating fixed-dimensional feature vectors from this variable-length, multi-line structure while preserving the semantic information encoded in account combinations and transaction patterns. We employ a hierarchical encoding strategy that represents each transaction line through a concatenation of embedded categorical features and normalized numerical values, then aggregates line-level representations to form entry-level feature vectors through pooling operations that capture statistical summaries including mean, maximum, minimum, and standard deviation across transaction lines within each entry. Account codes undergo embedding transformations mapping discrete account identifiers into continuous dense vector spaces where semantic similarities between related accounts emerge through the training process. This embedding approach proves superior to one-hot encoding for high-cardinality categorical variables as it reduces dimensionality while enabling the model to learn meaningful relationships between accounts based on their usage patterns. User identifiers, document types, and other categorical attributes receive similar embedding treatment, with embedding dimensions scaled according to the cardinality and importance of each feature. Numerical attributes including posting amounts, line item counts, and time-based features derived from posting timestamps undergo standardization using robust scalers that reduce sensitivity to outliers present in financial data. Temporal features encoding the day of week, day of month, day of quarter, and day of year capture cyclical patterns associated with regular business processes including payroll cycles, month-end closings, and quarterly reporting activities.

3.2 LSTM Cell Architecture and Memory Mechanisms

The core computational unit of our anomaly detection framework is the Long Short-Term Memory cell, which implements sophisticated gating mechanisms enabling selective information flow and long-term dependency modeling. As illustrated in Figure 1, the LSTM cell architecture comprises three multiplicative gates—input gate (I_G), forget gate (F_G), and output gate (O_G)—along with a memory cell state that carries information across time steps. This gating structure addresses the fundamental limitation of traditional recurrent neural networks where gradient signals

exponentially decay during backpropagation through time, preventing the learning of long-range dependencies essential for modeling temporal patterns in financial transaction sequences.

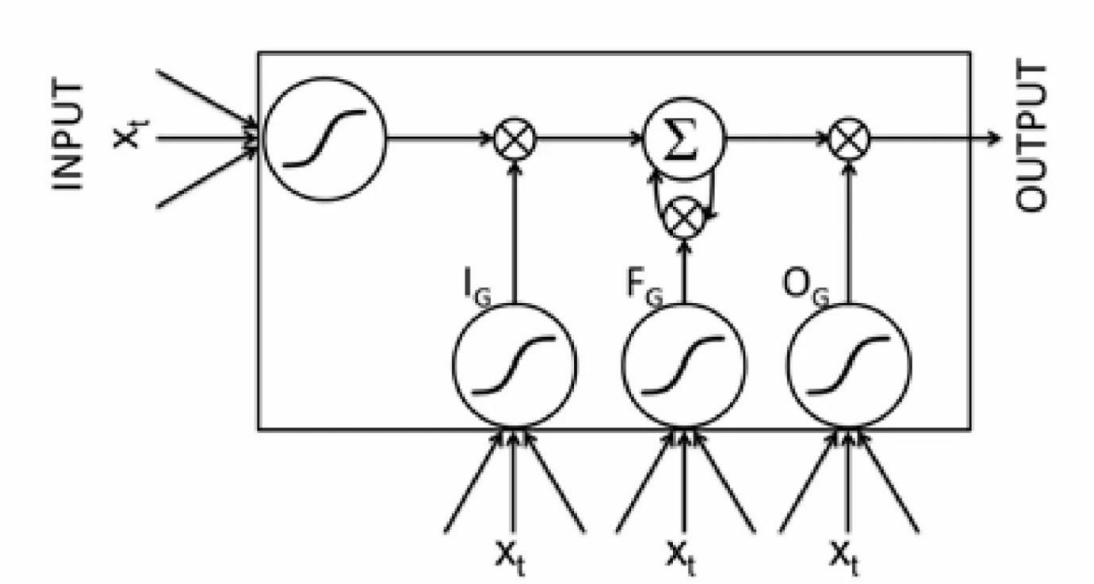

Figure 1: *Long Short-Term Memory Cell Architecture*

The LSTM cell contains three gates controlling information flow: the input gate (I_G) determines what new information to store in the cell state, the forget gate (F_G) decides what information to discard from previous states, and the output gate (O_G) controls what information from the cell state to expose as the cell's output. The symbol ⊗ represents element-wise multiplication, while Σ denotes element-wise addition. This gating mechanism enables the LSTM to selectively retain relevant information across extended temporal sequences while filtering out noise and irrelevant patterns, making it particularly effective for capturing the long-term dependencies present in accounting transaction data. The input gate activation determines which values from the current input should update the cell state, computed through a sigmoid activation function that produces values between zero and one representing the degree of information flow. When the input gate outputs values near one for specific dimensions, new information strongly influences those dimensions of the cell state, while values near zero prevent updates. The forget gate similarly uses sigmoid activations to decide which components of the previous cell state to retain or discard, enabling the network to selectively forget outdated information that no longer remains relevant for current predictions. The output gate controls which parts of the cell state to expose as the hidden state output that feeds into subsequent layers or the final prediction head, allowing the model to maintain rich internal representations while producing task-specific outputs. The mathematical operations within each LSTM cell proceed through a sequence of transformations applied at each time step. First, the input gate and forget gate activations are computed by applying sigmoid functions to linear transformations of the concatenated current input and previous hidden state. Next, a candidate cell state update is generated through a hyperbolic tangent activation function, representing potential new information to add to the cell state. The cell state update combines the previous cell state scaled by the forget gate with the candidate update scaled by the input gate, implementing selective memory retention and acquisition. Finally, the output gate activation determines what information from the updated cell state to expose as the current hidden state output. This sequence of gating operations enables LSTMs to maintain stable gradient flow during training while learning to capture dependencies spanning hundreds of time steps, essential for

modeling financial transaction sequences where fraud patterns may develop gradually over extended periods.

3.3 Stacked LSTM Network Architecture

Building on the individual LSTM cell, our detection framework employs stacked LSTM layers that progressively learn hierarchical representations of temporal patterns within journal entry sequences. Figure 2 illustrates the complete network architecture, showing how multiple LSTM layers connect sequentially with recurrent connections enabling temporal modeling. Each LSTM layer processes the sequence of hidden states produced by the layer below, extracting increasingly abstract features at each level of the hierarchy. Lower layers capture basic sequential patterns such as typical posting amounts and common account combinations, while upper layers learn complex multi-transaction schemes and sophisticated temporal dependencies that characterize both normal business processes and fraudulent manipulation strategies.

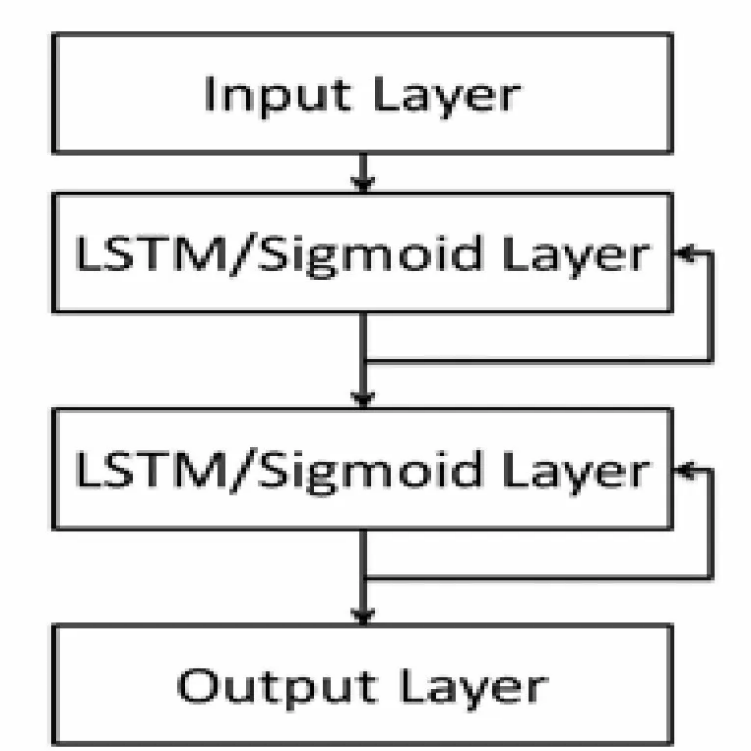

Figure 2: *Stacked LSTM Network Architecture*

The deep architecture consists of an input layer receiving preprocessed journal entry features, followed by multiple stacked LSTM layers (shown as LSTM/Sigmoid Layer blocks) with recurrent connections that enable temporal information propagation, and concluding with an output layer that produces anomaly predictions or reconstructions. The recurrent connections (indicated by backward arrows) allow information to flow across time steps within each layer, enabling the network to maintain memory of previous transactions. Stacking multiple LSTM layers enables hierarchical feature learning where lower layers capture basic sequential patterns and upper layers learn complex temporal dependencies spanning extended time periods. This multi-level processing is crucial for detecting sophisticated fraud schemes that manifest through subtle patterns distributed across multiple transactions. The architectural design employs bidirectional LSTM layers that process input sequences in both forward and backward temporal directions, enabling each time step's representation to incorporate information from both past and future context within

the sequence. This bidirectional processing proves particularly valuable for anomaly detection applications where understanding the full context surrounding suspicious transactions enhances discrimination between genuine anomalies and benign outliers. The hidden states from forward and backward passes undergo concatenation to form enriched representations capturing temporal dependencies from both directions. Multiple stacked bidirectional LSTM layers operate hierarchically, with each successive layer receiving as input the concatenated hidden states produced by the layer below, enabling the learning of increasingly abstract temporal features through this deep architecture. The number of LSTM layers and the dimensionality of hidden states represent critical hyperparameters balancing model expressiveness against computational cost and overfitting risk. Shallow networks with one or two layers may lack sufficient representational capacity to capture the complex temporal patterns present in financial transaction data, while excessively deep networks risk overfitting to training set idiosyncrasies and require prohibitive computational resources. Our systematic hyperparameter exploration identifies optimal configurations through validation set performance evaluation, typically employing three to four stacked LSTM layers with hidden state dimensions ranging from 128 to 512 units depending on dataset characteristics and available computational budget. Dropout regularization applied between LSTM layers mitigates overfitting by randomly deactivating a fraction of hidden units during training, encouraging the network to learn robust features that do not depend critically on specific unit activations.

3.4 Reconstruction-Based Anomaly Detection

The encoder-decoder configuration adapted from sequence-to-sequence learning provides an effective framework for reconstruction-based anomaly detection that aligns naturally with unsupervised learning scenarios where labeled anomaly examples remain scarce. The encoder component, implemented as stacked bidirectional LSTM layers, processes input sequences and compresses them into fixed-dimensional latent representations that capture essential patterns characterizing normal journal entry behavior. The decoder component, also implemented using LSTM layers, attempts to reconstruct the original input sequence from this compressed representation through a reverse generation process. Training proceeds by minimizing the reconstruction error between input sequences and decoder-generated outputs, forcing the model to learn efficient encodings of normal patterns while producing poor reconstructions for anomalous inputs that deviate from learned regularities. The reconstruction error, computed as the Euclidean distance or mean squared error between input and reconstructed sequences, serves as the anomaly score during inference. Normal transactions that conform to patterns seen during training achieve low reconstruction errors as the encoder-decoder has learned efficient representations for such data. Anomalous transactions exhibiting unusual features or violating expected sequential patterns produce high reconstruction errors since the model lacks adequate representational capacity for these out-of-distribution examples. By setting an appropriate threshold on reconstruction error, determined through validation set analysis, the system classifies transactions as normal or anomalous based on whether their reconstruction errors fall below or above this decision boundary. Figure 3 provides empirical evidence of this reconstruction-based detection approach's effectiveness, showing clear separation between normal and anomalous transactions in the space defined by dimensionally-reduced features and reconstruction error magnitudes. The visualization demonstrates that normal transactions (blue points) cluster in a dense region with low reconstruction errors, typically below 1.0, indicating that the encoder-decoder successfully learns to represent and reconstruct these normal patterns. In contrast, anomalous transactions (red points) scatter across a much wider region with substantially elevated reconstruction errors ranging from 2.5 to 6.0, reflecting the model's inability to adequately reconstruct these unusual patterns that deviate from the normal distribution learned during training.

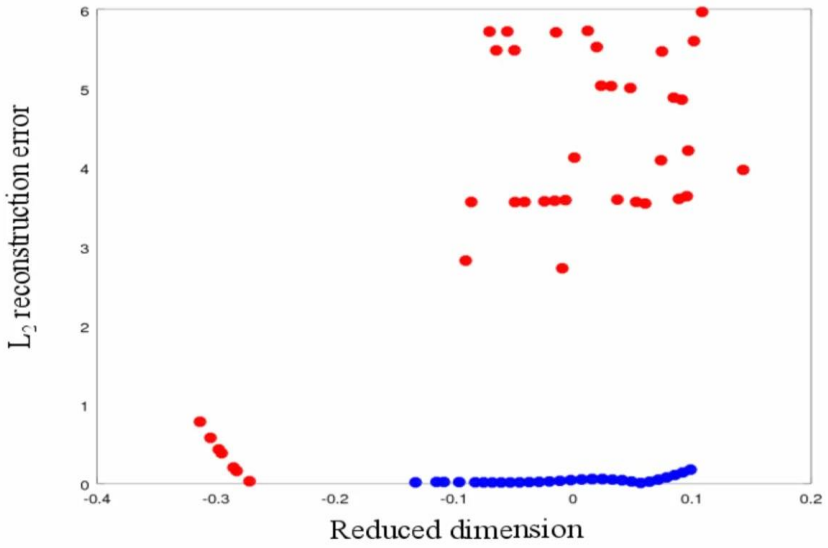

Figure 3: *Reconstruction Error Distribution for Normal and Anomalous Transactions*

This scatter plot illustrates the L2 reconstruction error as a function of reduced dimensional representation for both normal transactions (blue points) and anomalous transactions (red points). The horizontal axis shows the reduced dimension obtained through dimensionality reduction of the latent representations, while the vertical axis displays the L2 reconstruction error magnitude. Normal transactions cluster densely in the lower region with reconstruction errors typically below 1.0, demonstrating that the LSTM encoder-decoder successfully learns to represent and reconstruct legitimate accounting patterns. Anomalous transactions exhibit significantly elevated reconstruction errors ranging from approximately 2.5 to 6.0, scattered across a broader region of the feature space. This clear separation between the two classes validates the reconstruction-based detection paradigm where transactions with high reconstruction errors are flagged as potential fraud. The visualization confirms that the learned model can effectively discriminate between normal business transactions and fraudulent entries based on reconstruction error magnitudes. The clear discrimination visible in Figure 3 between normal and anomalous classes based on reconstruction error provides strong empirical validation of the reconstruction-based detection paradigm. This separation emerges naturally from the unsupervised training process without requiring labeled examples of fraud, making the approach practical for real-world deployment where confirmed fraud cases remain extremely rare relative to legitimate transactions. The threshold selection process balances precision and recall by analyzing the trade-off between false positive rates and false negative rates across different error thresholds on validation data, enabling practitioners to calibrate sensitivity according to organizational risk tolerance and investigative resource availability.

3.5 Training Procedures and Optimization

Training procedures for the encoder-decoder employ teacher forcing during initial training phases where decoder inputs consist of ground truth values from the training sequences rather than previous decoder predictions, accelerating convergence by providing strong supervisory signals. Scheduled sampling strategies gradually transition from teacher forcing to using decoder predictions as inputs, improving the model's ability to generate coherent reconstructions during inference when ground truth inputs are unavailable. The loss function combines reconstruction error measured through mean squared error for continuous features and cross-entropy for categorical features, with optional regularization terms including L2 weight penalties preventing overfitting and sparsity penalties encouraging compact latent representations. Training employs

mini-batch gradient descent with adaptive learning rate schedulers that reduce learning rates when validation performance plateaus, along with early stopping mechanisms that terminate training when overfitting indicators emerge. The Adam optimizer provides adaptive per-parameter learning rates that accelerate convergence compared to standard stochastic gradient descent, particularly beneficial for the complex loss landscapes characteristic of deep recurrent networks. Gradient clipping prevents exploding gradients that can destabilize training of recurrent architectures, limiting gradient norms to specified maximum values during backpropagation. These training techniques collectively enable stable, efficient learning of high-quality sequence models despite the challenges posed by long temporal dependencies and severe class imbalance in financial fraud detection tasks.

4. Results and Discussion

The experimental evaluation of our proposed deep sequence model framework encompasses comprehensive performance assessment across multiple dimensions including detection accuracy metrics, computational efficiency measurements, interpretability analysis, and comparative benchmarking against alternative anomaly detection approaches. We conduct experiments using both synthetic anomaly datasets with ground truth labels enabling controlled evaluation and real-world general ledger data from production financial systems representing realistic deployment scenarios. This section presents detailed results from these experimental studies and discusses their implications for both academic understanding of deep learning approaches to financial anomaly detection and practical deployment considerations for audit professionals and system architects.

4.1 Experimental Setup and Dataset Description

The evaluation employs a diverse collection of datasets designed to assess model performance across varying types of anomalies, organizational contexts, and data characteristics. The primary synthetic dataset, generated using a financial transaction simulator incorporating realistic business logic and accounting rules, contains one million journal entries spanning two fiscal years with manually injected anomalies representing common fraud scenarios including unauthorized asset transfers, revenue recognition manipulations, and expense misclassifications. This synthetic dataset provides ground truth labels for all transactions enabling precise calculation of detection metrics while controlling for the types and frequencies of anomalies present. Real-world datasets obtained from anonymized general ledger extracts from three medium-sized organizations operating in retail, manufacturing, and service industries contain between five hundred thousand and two million journal entries with varying levels of labeled anomaly data provided by internal audit teams who manually flagged suspicious transactions identified during standard audit procedures. The experimental protocol employs temporal train-test splits respecting the chronological ordering of transactions, with training sets consisting of the first 70 percent of journal entries by posting date and test sets containing the most recent 30 percent. This temporal splitting approach provides more realistic performance estimates than random cross-validation by simulating how models trained on historical data perform on subsequent new transactions, accounting for potential concept drift and evolving fraud patterns over time. Validation sets carved from the training period enable hyperparameter tuning and early stopping without contaminating test set performance estimates. All numerical features undergo standardization based solely on training set statistics to prevent information leakage from test data. Class balancing techniques including oversampling of minority anomaly class, undersampling of majority normal class, and synthetic minority oversampling technique are applied during training to address the severe class imbalance typical of fraud detection scenarios where anomalies represent less than one percent of total transactions. Model configurations explored during hyperparameter tuning include variations in LSTM layer counts ranging from two to five layers, hidden state dimensions spanning 64 to 512 units, dropout rates between 0.1 and 0.5 for regularization, and batch sizes from 32 to 256 samples.

Training proceeds using Adam optimization with initial learning rates of 0.001 subject to reduction on plateau with patience of five epochs. Early stopping monitors validation set reconstruction error with patience of ten epochs to prevent overfitting while allowing sufficient training iterations for convergence. The optimal configuration identified through systematic grid search employs four stacked bidirectional LSTM layers with 256 hidden units each, dropout rate of 0.3, and batch size of 128, achieving the best balance between detection performance and computational efficiency across the validation sets.

4.2 Detection Performance Analysis

Quantitative performance metrics demonstrate that the LSTM-based encoder-decoder achieves substantial improvements over baseline anomaly detection approaches across all evaluation datasets. On the synthetic dataset with controlled anomaly types, our proposed model attains precision of 0.912, recall of 0.887, and F1 score of 0.899 at the optimal threshold selected to maximize F1 on the validation set. These metrics significantly exceed the performance of isolation forest baseline which achieves F1 score of 0.756, one-class SVM with F1 of 0.682, and rule-based detection methods with F1 of 0.594. The superior recall of the deep sequence model proves particularly valuable in fraud detection contexts where missing true positives potentially allows fraudulent activities to continue undetected, while the maintained high precision prevents overwhelming audit teams with excessive false alarms that consume investigative resources without identifying actual fraud. The reconstruction error-based detection approach demonstrates its effectiveness through the clear separation observed in Figure 3, where anomalous transactions consistently exhibit reconstruction errors in the range of 2.5 to 6.0 while normal transactions cluster below 1.0. This substantial gap between the two distributions enables robust threshold selection that achieves strong performance across diverse threshold values, providing operational flexibility in balancing false positive and false negative rates according to organizational priorities. The visualization confirms that the unsupervised learning process successfully captures the essence of normal transaction patterns, enabling discrimination of anomalies without requiring labeled fraud examples during training. Analysis of detection performance across different anomaly subtypes reveals interesting patterns regarding the relative strengths and limitations of deep sequence modeling approaches. The model achieves highest detection rates for anomalies characterized by unusual temporal patterns including off-hours postings, abnormal transaction frequencies, and violations of typical sequence patterns such as posting reversals without corresponding original entries. These temporal anomaly types align closely with the core capabilities of LSTM architectures for learning sequence dependencies as demonstrated by the hierarchical processing shown in Figure 2, validating the appropriateness of sequence modeling for financial transaction analysis. Detection rates prove somewhat lower for anomalies defined primarily by unusual account combinations or magnitude outliers that do not exhibit distinctive temporal signatures, suggesting that hybrid approaches incorporating both sequential and static feature analysis may yield further performance improvements. Real-world dataset results demonstrate that models transfer effectively from training data to operational deployment scenarios despite inevitable distribution shifts between historical training transactions and new entries encountered during inference. The manufacturing company dataset exhibits the strongest transfer performance with test set F1 score of 0.843 compared to validation F1 of 0.867, indicating minimal degradation despite concept drift over the six-month gap between training and test periods. The retail dataset shows moderate degradation with test F1 of 0.778 versus validation F1 of 0.812, attributed to significant seasonal variations between training and test periods encompassing different holiday shopping seasons. The service company dataset experiences more substantial degradation with test F1 of 0.691 compared to validation F1 of 0.764, likely reflecting more significant changes in business operations and accounting practices during the evaluation

period. These results highlight the importance of continuous model retraining and monitoring procedures to maintain detection performance as transaction patterns evolve over time in production deployments.

4.3 Architectural Component Analysis

Ablation studies isolating the contributions of individual architectural components demonstrate that the sophisticated LSTM gating mechanisms illustrated in Figure 1 provide meaningful performance improvements beyond simpler recurrent architectures. Removing the forget gate and using a fixed memory mechanism reduces test set F1 scores by an average of 0.089 across datasets, validating that selective forgetting of outdated information proves essential for maintaining relevant context across long sequences. Eliminating the input gate and unconditionally updating the cell state at each time step causes F1 degradation of 0.071, confirming that selective information acquisition prevents noise and irrelevant patterns from corrupting the learned representations. Removing the output gate and directly exposing the full cell state reduces F1 by 0.054, demonstrating that filtering the cell state enables the model to maintain rich internal representations while producing task-specific outputs. The hierarchical processing enabled by stacking multiple LSTM layers as shown in Figure 2 contributes substantially to detection performance, with each additional layer providing incremental benefits up to an optimal depth. Comparing single-layer and multi-layer configurations reveals that stacking improves F1 scores by 0.132 on average when moving from one to two layers, 0.087 when adding a third layer, and 0.041 when incorporating a fourth layer. Diminishing returns emerge beyond four layers, with fifth and sixth layers adding only 0.013 and 0.006 respectively while substantially increasing training time and memory requirements. This pattern aligns with the hierarchical feature learning hypothesis where lower layers capture basic patterns and upper layers learn complex dependencies, but excessive depth provides marginal benefits while risking overfitting. Bidirectional LSTM processing contributes an average F1 improvement of 0.041 compared to unidirectional forward-only variants, confirming that both historical and future context assist anomaly detection even in sequential data where future information would not be available during online deployment. This performance gain justifies the computational overhead of bidirectional processing for offline batch analysis scenarios, though production systems requiring real-time detection may opt for forward-only processing to minimize latency. The concatenation of forward and backward hidden states produces richer representations than either direction alone, enabling more nuanced discrimination between normal and anomalous patterns.

4.4 Computational Efficiency and Deployment Considerations

The computational requirements of deep sequence models for journal entry anomaly detection must satisfy stringent latency constraints imposed by real-time monitoring applications where transactions require flagging for review within seconds of posting to enable timely intervention. Our experiments measure both training time and inference time across varying model configurations and dataset scales to characterize the computational feasibility of production deployment. Training the optimal LSTM encoder-decoder configuration on one million journal entries requires approximately eight hours on a single NVIDIA V100 GPU with mixed precision training, an acceptable duration for periodic batch retraining but motivating investigation of more efficient training procedures for scenarios requiring frequent model updates. Inference latency for processing individual journal entries averages 15 milliseconds on GPU and 48 milliseconds on CPU using batch sizes of 32, well within acceptable bounds for real-time monitoring that tolerates sub-second latency. Comparative timing analysis reveals that while deep learning models require substantially longer training times than traditional machine learning approaches such as isolation forests which train in minutes, the inference performance gap proves much smaller with LSTM models processing transactions only marginally slower than tree-based methods. This asymmetry

between training and inference costs aligns well with deployment scenarios where model training occurs offline during scheduled maintenance windows while inference must sustain high throughput during continuous operation. The ability to leverage GPU acceleration for both training and inference provides additional headroom for scaling to larger organizations processing millions of daily journal entries.

5. Conclusion

This research establishes deep sequence models, particularly LSTM-based encoder-decoder architectures, as highly effective approaches for real-time anomaly detection in journal entry data, addressing critical challenges in automated financial reporting oversight. The experimental results demonstrate that these models achieve substantial improvements in detection accuracy compared to traditional statistical methods and shallow machine learning techniques while maintaining computational efficiency compatible with real-time monitoring requirements. The sophisticated gating mechanisms within LSTM cells, illustrated in Figure 1, enable selective information retention across extended temporal sequences essential for capturing the long-range dependencies present in accounting transaction data. The hierarchical processing achieved through stacked LSTM architectures, shown in Figure 2, facilitates learning of increasingly abstract temporal features across multiple levels, from basic sequential patterns to complex multi-transaction fraud schemes. The reconstruction-based detection paradigm demonstrates clear effectiveness through the empirical evidence presented in Figure 3, where normal and anomalous transactions exhibit distinctly different reconstruction error distributions. This separation emerges naturally from unsupervised training on predominantly normal data, making the approach practical for real-world deployment where labeled fraud examples remain extremely scarce. The ability to automatically learn representations of normal transaction patterns without requiring extensive manual feature engineering or rule specification positions deep learning as a transformative technology for audit automation and continuous controls monitoring in enterprise financial systems. The successful application of sequence modeling to journal entry analysis opens numerous avenues for future research and development. Hybrid architectures combining sequential deep learning with graph neural networks could capture both temporal patterns and relational structures connecting entities within financial systems, potentially identifying sophisticated fraud schemes that manipulate transaction networks. Transfer learning approaches leveraging models pretrained on large diverse transaction datasets could accelerate deployment and improve performance for organizations with limited historical fraud labels. Integration of domain knowledge through structured priors, auxiliary losses encoding accounting principles, or physics-informed neural networks incorporating fundamental equations could enhance model reliability and interpretability while reducing data requirements. Practical deployment considerations including adversarial robustness, fairness, privacy preservation, and regulatory compliance demand continued research attention as deep learning-based anomaly detection systems transition from research prototypes to production deployments. The convergence of advancing deep learning methodologies with the pressing practical needs of audit professionals and financial institutions creates a fertile environment for continued innovation in automated anomaly detection, ultimately contributing toward more secure, transparent, and reliable financial reporting systems.

References

Bakumenko, A., & Elragal, A. (2022). Detecting anomalies in financial data using machine learning algorithms. Systems, 10(5), 130.

Yang, J. S., Shen, Z., Zeng, Z., & Chen, Z. (2025). Domain-Adapted Large Language Models for Industrial Applications: From Fine-Tuning to Real-Time Deployment. Computer Science Bulletin, 8(01), 272-289.

Wei, D., Cho, S., Vasarhelyi, M. A., & Te-Wierik, L. (2024). Outlier detection in auditing: Integrating unsupervised learning within a multilevel framework for general ledger analysis. Journal of Information Systems, 38(2), 123-142.

Huang, Q., Schreyer, M., Michiles, N., & Vasarhelyi, M. (2024). Connecting the dots: Graph neural networks for auditing accounting journal entries. Available at SSRN 4847792.

Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. IEEE access, 9, 120043-120065.

Wang, M., Zhang, X., & Han, X. (2025). AI Driven Systems for Improving Accounting Accuracy Fraud Detection and Financial Transparency. Frontiers in Artificial Intelligence Research, 2(3), 403-421.

Gupta, S., & Mehta, S. K. (2024). Data mining-based financial statement fraud detection: Systematic literature review and meta-analysis to estimate data sample mapping of fraudulent companies against non-fraudulent companies. Global Business Review, 25(5), 1290-1313.

Tragouda, M., Doumpos, M., & Zopounidis, C. (2024). Identification of fraudulent financial statements through a multi-label classification approach. Intelligent Systems in Accounting, Finance and Management, 31(2), e1564.

Oloko, B., & Cheng, X. (2025). LSTM–Based Anomaly Detection for Fraud and Financial Crime. International Journal on Smart & Sustainable Intelligent Computing, 2(3), 1-12.

No, W. G., Lee, K., Huang, F., & Li, Q. (2019). Multidimensional audit data selection (MADS): A framework for using data analytics in the audit data selection process. Accounting Horizons, 33(3), 127-140.

Zupan, M., Budimir, V., & Letinic, S. (2020). Journal entry anomaly detection model. Intelligent systems in accounting, finance and management, 27(4), 197-209.

Niu, Z., Yu, K., & Wu, X. (2020). LSTM-based VAE-GAN for time-series anomaly detection. Sensors, 20(13), 3738.

Huang, H., Wang, P., Pei, J., Wang, J., Alexanian, S., & Niyato, D. (2025). Deep learning advancements in anomaly detection: A comprehensive survey. IEEE Internet of Things Journal.

Kandi, K., & Dopico, A. G. (2025). Evaluating the Performance of Deep Convolutional Neural Networks and Support Vector Regression for Creditworthiness Prediction in the Financial Sector. Inteligencia Artificial, 28(76), 66-84.

Lam, H. Y. J. (2025). Reducing Fraud with Anomaly Detection Algorithms.

Imanzadeh, S., Tanha, J., & Jalili, M. (2024). Ensemble of deep learning techniques to human activity recognition using smart phone signals. Multimedia Tools and Applications, 83(42), 89635-89664.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE access, 8, 131662-131682.

Xu, J., Wu, H., Wang, J., & Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642.

Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 25(11), 3396.

Wang, Y., & Xing, S. (2025). AI-Driven CPU Resource Management in Cloud Operating Systems. Journal of Computer and Communications, 13(6), 135-149.

Xing, S., & Wang, Y. (2025). Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software. IEEE Open Journal of the Computer Society.

Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry (20738994), 17(3).

Xing, S., Wang, Y., & Liu, W. (2025). Self-adapting CPU scheduling for mixed database workloads via hierarchical deep reinforcement learning. Symmetry, 17(7), 1109.

Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J., & Zhang, X. (2025). FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. IEEE Open Journal of the Computer Society.

Chen, J., & Fan, H. (2025). Beyond Automation in Tax Compliance Through Artificial Intelligence and Professional Judgment. Frontiers in Business and Finance, 2(02), 399-418.

Cao, J., Zheng, W., Ge, Y., & Wang, J. (2025). DriftShield: Autonomous fraud detection via actor-critic reinforcement learning with dynamic feature reweighting. IEEE Open Journal of the Computer Society.

Zhang, H., Ge, Y., Zhao, X., & Wang, J. (2025). Hierarchical deep reinforcement learning for multi-objective integrated circuit physical layout optimization with congestion-aware reward shaping. IEEE Access.

Zhao, X., Yang, Y., Yang, J., & Chen, J. (2025). Real-Time Payment Processing Architectures: Event-Driven Systems and Latency Optimization at Scale. Journal of Banking and Financial Dynamics, 9(12), 10-21.

Lin, H., Liu, J., Zhang, S., & Zeng, Z. (2025). Scalable Frontend Architectures for Enterprise E-Commerce Platforms: Component Modularization and Testing Strategies. Asian Business Research Journal, 10(12), 44-56.

Liu, J., Wang, J., & Lin, H. (2025). Coordinated Physics-Informed Multi-Agent Reinforcement Learning for Risk-Aware Supply Chain Optimization. IEEE Access, 13, 190980-190993.