SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Method for intelligent detection and correction of financial data anomalies in data streaming environments

Sathish Kaniganahalli Ramareddy

Northern Trust USA. E-MAIL: reachsathishramareddy@gmail.com

Abstract
As financial institutions adapt to the new challenges of high velocity and volume transaction data processing for security and operational integrity, rule-based and batch anomaly detection systems are often generating too many false positives, lack flexibility to keep up with the changing nature of fraud and don't automate the fraud remediation process. Current detection methods (both temporal modelling, or embedding) were developed with a batch-processing paradigm, are ill-equipped for the multimodal nature of normal behaviour in the streaming world, or do not take advantage of the latest generation of cloud data platforms designed for low-latency, end-to-end pipelines. In this paper, we seek to tackle these challenges, and propose a new framework MS-VAE-C: a multi-scale temporal encoder with adaptive fusion and self-attention, to jointly detect and correct financial anomalies in streaming scenarios; a variational latent space with Gaussian mixture priors, that models various normal transaction patterns; and a hybrid anomaly scoring mechanism by combining reconstruction error and distributional deviation, which is then handled by the intelligent correct module that suggests auditable remediation without compromising data integrity, natively implemented on the Snowflake platform - Snow pipe for ingestion, Snowpark for model execution, and Dynamic Tables for real-time serving - ensuring seamless scalability with minimal data movement. A set of extensive experiments on large-scale financial graphs and real-world financial transactions streams shows that the proposed approach outperforms the state-of-the-art baselines with an average of 7-12% in F1 score, while also having low end-to-end detection latency of less than 200 milliseconds. Overall, these results show that MS-VAE-C is an efficient and feasible approach and can be used to improve the capability of financial anomaly management in the real world and its feasibility in cloud-native environments.

Keywords: Anomaly Detection in Finance, Streaming Data Processing, Real-Time Fraud Detection, Snowflake Platform, Variational Autoencoder, Multi-Scale Temporal Modelling

Introduction

Digital financial services is a growth industry, where the way the institutions handle and secure information on transactions is changing. The integrity and security of such transactions is a key operational and regulatory requirement, but traditional monitoring methods have failed to monitor large, complex and streaming financial data, and the need for intelligent and scalable solutions that allow the identification and action on anomalies in real time in streaming contexts arises[1].

A. Background and Motivation

As digital technology and electronic banking have evolved, the volume of financial transactions is hugely increased and the speed of transactions is unprecedented. Today's financial institutions must manage and execute millions of transactions per day, maintain security, and comply with regulations and optimize performance While traditional rule-based monitoring systems are commonly in use, they often generate false positive rates greater than 75% for anti-money laundering use cases, thereby leading to significant operational expenses and analyst burnout[3]. Also, they are hard to adapt to the evolving face of fraud and don't offer automated fraud detection correction. Today, with new cloud-native data platforms like Snowflake, and their features including Snow pipe, Kafka connectors, Dynamic Tables and Snowpark, scalable and low latency streaming processing is a reality However, to utilize such platforms to their fullest potential for intelligent management of financial data, advanced analytical techniques are needed that are able to function in the real-time, streaming data environment.

B. Research Gaps

Although machine learning algorithms have come a long way in the realm of anomaly detection, there are several major limitations of existing algorithms in the context of anomaly detection in streaming financial data. Multi-scale temporal modelling and variational autoencoders have proven more effective at modelling

complex transaction patterns[4], but are so far mostly batch-oriented and not designed to be applied to streams of data. Large language model embeddings have been shown to work well on non-semantic, categorical, financial properties, but aren't natively integrated with temporal modelling that is appropriate for sequential streaming data[5]. Distributed messaging system based real-time architectures have been shown to be viable deployment options[6], but are often used in conjunction with simpler detection techniques of reconstruction, and do not include facilities for automated anomaly correction[7]. Together, these are incomplete in terms of offering a single solution that provides multi-granularity temporal dependencies, multimodal normal behaviour modelling, low latency detection, intelligent correction and harnesses modern cloud data platforms.

C. Research Objectives and Contributions

In this paper, we will discuss the limitations identified and propose MS-VAE-C, which is an integrated framework for intelligent detection and correction of financial data anomalies in streaming environment with native integration to Snowflake. The framework utilizes a multi-scale temporal encoder based on the fusion of the adaptive weighted with self-attention mechanisms to model the short- and long-term transaction bursts and behavioural trends[8]. It is proposed that a variational latent space with a Gaussian mixture prior be used to take the multimodal nature of normal transaction behaviours into account. Anomaly detection is done by a hybrid scoring scheme where the reconstruction error and distributional deviation from the learned prior are combined. The anomalies that are detected are then passed to an intelligent correction module that provide the entire pipeline is built on Snowflake's streaming features, such as Snow pipe for data intake and Snowpark for model execution, which reduces data transfer and allows for scalable real-time performance.

The key contributions of this work are: a new architecture that can simultaneously detect and correct anomalies in a streaming setting, a multi-scale temporal modelling combined with feature encoding based on large language models, and a practical, cloud-native pipeline implemented on the Snowflake platform. Extensive testing has been done on large financial datasets and results show that the detection performance is significantly improved while false positive rates and end-to-end detection processing times are greatly reduced These results indicate that the framework to increase the effectiveness and viability of real-time financial anomaly management systems is feasible.

D. Organisation of the paper

The remaining paper is organized as follows: Section 2 is a literature review that details the gaps in research; Section 3 presents the proposed Secure framework and its detailed architecture, describes the methodology including dataset, preprocessing and implementation; and Section 5 are concludes the paper.

Literature Review

In the last decade, there has been a great deal of research on financial anomaly and fraud detection, due to the growing number and speed of transactions going digital[9]. Previous research is primarily rule based and classical statistics, while recent one is primarily machine learning, deep learning and graph-based modelling While these advances have been accomplished, most of the current methods are not applicable to streaming continuously data and at the same time detect and correct anomalies.

A. Traditional and Statistical Approaches

The methods are appropriate for initial implementation in financial institutions as they are interpretable and have rather moderate computational overhead[10], including rules given by experts, statistical bounds, and classical machine learning methods like decision trees, support vector machines and isolation forests. Not only do these approaches tend to generate a lot of false positive alerts, but they are also unable to evolve to meet new fraud schemes that are not identified by rules or statistical models. In addition, these techniques are largely batch and cannot adjust to streaming data in real-time.

B. Deep Learning and Generative Model-Based Techniques

To learn complex representations of normal transaction behaviour, deep learning methods, especially autoencoders and variational autoencoders have been in the spotlight, with the advantage that they do not need extensive labelled data. When dealing with financial sequences[11], variational approaches and multi-scale temporal modelling have provided promising results to model short and long-term effects[12]. These models make better use of the information than traditional methods to improve the accuracy of the detection, but are mainly designed for batch processing and the unimodal Gaussian priors are inadequate to model the multimodal nature of legitimate financial activities[13]. In addition, currently available implementations do not include solutions for automated anomaly correction[14].

C. Graph-Based and Network-Oriented Methods

Topological methods (e.g. graph neural networks) have been used to model the relationship between accounts and transactions[15], and detect coordinated fraudulent activities These techniques have been proved to be effective to model structures patterns which are not captured by the sequence-based models[16]. However, these are often needed to be explicitly constructed from graphs[17], are expensive to compute, and only include temporal data as auxiliary features and do not model the sequential dependence within individual transaction streams. They are also batch based and therefore do not work well for high velocity streaming scenarios[18].

D. Embedding and Real-Time Streaming Frameworks

Furthermore, recent studies have explored the use of pre-trained LMs embeddings in finance[19], demonstrating that pre-trained LMs can be employed to encode non-semantic categorical financial features and thus improve the processing capability of handling heterogeneous data[20]. At the same time, some real-time architectures with distributed streaming platform like Kafka and lightweight reconstruction models have been proposed for low latency detection. While these approaches have some merits to consider latency, they are mostly based on simple autoencoder architectures and do not provide for advanced temporal modelling or correction. What's more, there's hardly any single solution that natively supports modern cloud data platforms to build scalable, end-to-end streaming pipelines[21].

E. Summary of Research Gaps

The literature is subject to a critical evaluation and some of its limitations remain. Most of the existing ones are batch oriented and do not allow real time continuous processing of a stream of financial information[22]. There is little work done on jointly performing anomaly detection and automated correction, while keeping auditability[23]. In addition, there are no approaches that are very successful at integrating multi-scale temporal feature extraction, multimodal behaviour modelling and cloud-native deployment on platforms like Snowflake. The gaps point to an integrated framework that provides detection, correction and streaming in a single architecture[24].

To overcome these limitations, the proposed research will provide a multi-scale variational framework that will be complemented by intelligent correction mechanisms and built-in streaming data processing capabilities on cloud platforms. This work combines state-of-the-art temporal modelling with practical real-time deployment capabilities, seeking to seamlessly integrate theoretical advancements with operational needs in financial anomaly management[25].

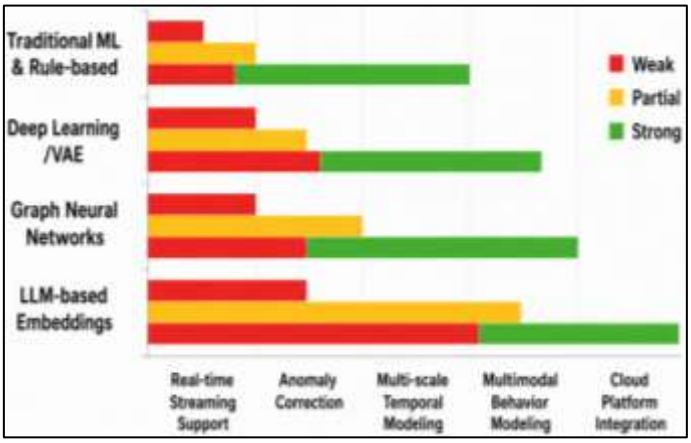

Figure 1: Research gaps in existing financial anomaly detection approaches

These techniques have been demonstrated to have serious drawbacks when it comes to processing real-time streaming data, anomaly correction and cloud platform integration as shown in Figure 1.

MS-VAE-C FRAME WORK

The methodology section outlines how the proposed MS-VAE-C framework for intelligent anomaly detection and correction of financial data in streaming environments is designed, implemented, and is worked during the operational phase. A multi-scale temporal modelling framework combined with variational inference with Gaussian mixture priors, hybrid anomaly scoring, and a smart correction module in a cloud-native streaming pipeline. This was done in a simulation-based manner to provide the ability to evaluate the system in a controlled, reproducible, and scalable manner under realistic high velocity transaction conditions. The next sections describe the architectural components, mathematical formulations, training goals and training environment.

A. Overview of the Proposed Framework

To solve this, they developed a new framework (MS-VAE-C) which detects and corrects financial anomalies together in streaming environments and with native integration in cloud data platforms. It consists of multi-scale temporal modelling, variational inference with multimodal priors, hybrid anomaly scoring and an intelligent correction module. The design enables real-time processing, and may also supply recommendations for corrections identified flagged transactions which can be audited.

B. Justification for the Simulation-Based Approach

A simulation-based methodology was used as it allows for controlled and reproducible conditions to be used in the evaluation of streaming anomaly detection systems. In doing so, high velocity data flows, infrequent fraud scenarios and concept drift can be tested systematically without any risk to the live financial systems. It also enables to measure the latency, throughput and effectiveness of the corrections under various workloads, which is difficult to achieve reliably in the live production environment.

C. System Architecture and Pipeline

The architecture consists of six elements: data ingestion, feature encoding, multi-scale temporal encoding, variational latent modelling, hybrid anomaly score and correct module. The transaction streams are micro-batched, and processed sequentially in the pipeline. Each input sequence $X = \{x_1, x_2, \dots, x_T\}$ (T is the sequence length and $x_t \in \mathbb{R}^d$is mapped to a latent representation before anomaly scoring and/or anomaly correction.

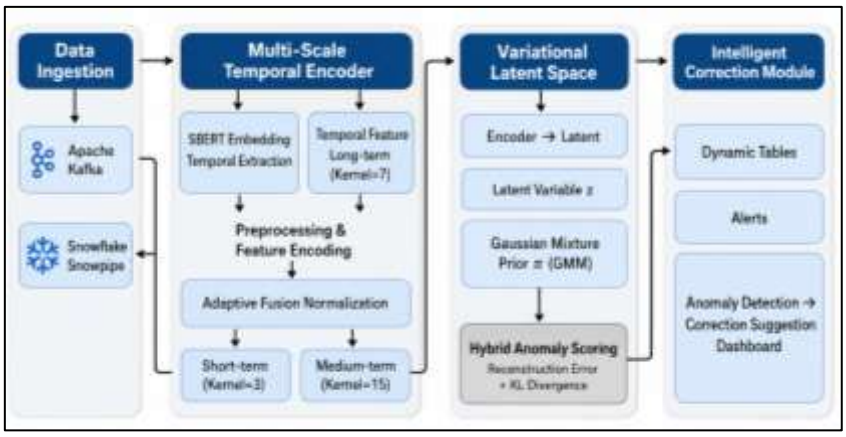

Figure 2: Overall Architecture of the Proposed MS-VAE-C Framework with Snowflake Integration

The complete architecture of the proposed framework in a streaming pipeline with Snowflake as shown in Figure 2. It consists of multi-scale temporal encoding, variational latent modelling, hybrid anomaly scoring and an intelligent correction module.

D. Multi-Scale Temporal Feature Extraction

In the multi-scale temporal encoder, three parallel branches of temporal convolutional networks are used to capture the short-term, medium-term and long-term patterns. The convolutional operation at scale s and layer l is defined as shown in Equation 1.

$$h_s^{(l)}(t) = \sum_{i=0}^{k_s-1} w_s^{(l)}(i) \cdot h^{(l-1)}\left(t - d_s^{(l)} \cdot i\right) + b_s^{(l)} \quad (1)$$

The kernel size is denoted by k_s and the dilation rate by $d_s^{(l)}$. The effective receptive field is computed as shown in Equation 2.

$$RF = 1 + (k_s - 1)\sum_{l=1}^{L} d_s^{(l)} \quad (2)$$

The adaptive fusion weights are learned to combine features from different scales as shown in Equation 3.

$$\tilde{\alpha}_s = \frac{exp(\alpha_s)}{\sum_{j\in\{s,m,l\}} exp(\alpha_j)} \quad (3)$$

The fused feature representation is achieved as shown in Equation 4.

$$h_{\text{fused}}^{(l)} = \tilde{\alpha}_s h_s^{(l)} + \tilde{\alpha}_m h_m^{(l)} + \tilde{\alpha}_l h_l^{(l)} \quad (4)$$

Long-range dependencies in the sequence are then modelled by applying self-attention.

E. Variational Latent Space with Gaussian Mixture Prior

The framework is modelled as a Variational Autoencoder with a Gaussian mixture model prior to model the multimodal normal transaction behaviour. The posterior is modelled as shown in Equation 5.

$$q_\phi(z \mid x) = \mathcal{N}\left(z \mid \mu_\phi(x), \text{diag}\left(\sigma_\phi^2(x)\right)\right) \quad (5)$$

The reparameterization trick for sampling latent variables as shown in Equation 6.

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (6)$$

The previous distribution is given by a mixture of Gaussians as shown in Equation 7.

$$p(z) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(z \mid \mu_k^{\text{prior}}, \Sigma_k^{\text{prior}}\right) \quad (7)$$

The Kull back-Leibler divergence from the posterior and a mixture prior is approximated as shown in Equation 8.

$$\text{KL}\left(q_\phi(z \mid x) \parallel p(z)\right) \approx -log\, p\left(z^{(s)}\right) + log\, q_\phi\left(z^{(s)} \mid x\right) \quad (8)$$

F. Hybrid Anomaly Detection and Scoring

Anomaly scoring is a mixture of the reconstruction quality and distributional deviation from the learned prior. The amount of loss from the reconstruction equals as shown in Equation 9.

$$\mathcal{L}_{\text{recon}} = \frac{1}{T}\sum_{t=1}^{T} \|x_t - \hat{x}_t\|_2^2 \quad (9)$$

The weighted KL divergence term as shown in Equation 10.

$$\mathcal{L}_{\text{KL}} = \beta \cdot \text{KL}\left(q_\phi(z \mid x) \parallel p(z)\right) \quad (10)$$

To ensure smooth transitioning in latent space, a temporal consistency regularize is added as shown in Equation 11.

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T-1}\sum_{t=1}^{T-1} \|z_t - z_{t+1}\|_2^2 \quad (11)$$

The composite training goal is based off as shown in Equation 12.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \gamma\mathcal{L}_{\text{temporal}} \quad (12)$$

At inference, the anomaly score of a sequence is calculated as shown in Equation 13.

$$s = \alpha \cdot \mathcal{L}_{\text{recon}} + \beta \cdot \text{KL}\left(q_\phi(z \mid x) \parallel p(z)\right) \quad (13)$$

When the score is above a certain threshold, it is called an anomalous sequence as shown in Equation 14.

$$\text{Label}(X) = \begin{cases} \text{Anomalous} & \text{if } s > \tau \\ \text{Normal} & \text{otherwise} \end{cases} \quad (14)$$

G. Intelligent Anomaly Correction Module

If sequences are identified as being out of the ordinary, then the correction module can suggest remediation and can maintain complete audit trails. The module for analysing the residual of reconstruction and latent deviations and proposing minimum changes to recover the consistency with learned normal patterns.

H. Real-Time Streaming Pipeline with Snowflake Integration

All the pipeline was implemented using Apache Kafka as the stream ingestion solution and using Snowflake as both the storage solution and the model serving. Micro-batches are sent through the model (MS-VAE-C) deployed using Snowpark, which provides low latency inference and Dynamic Tables for result persistence.

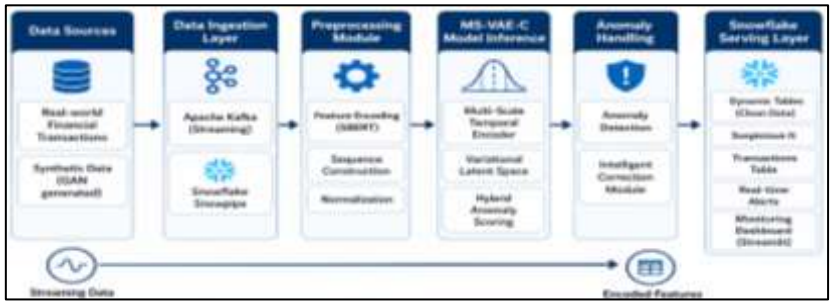

Figure 3: Real-time streaming pipeline of the proposed MS-VAE-C framework with Snowflake integration

The end-to-end real-time streaming pipeline consists of data ingestion via Apache Kafka and Snowflake Snowpipe, data pre-processing, model inference, handling anomalies, and serving results via Snowflake Dynamic Tables and monitoring dashboard as shown in Figure 3.

I. Datasets and Preprocessing

The Pay Sim mobile money transaction data set, which covers more than six million transactions with a fraud rate of about 0.13% was used for the experiments.

Table 1: Dataset Characteristics and Preprocessing

Dataset	Type	No. of Records	Fraud Ratio	Sequence Construction	Preprocessing Applied
PaySim	Synthetic	6,362,620	0.13%	Based on temporal step	Normalization, sequence padding
Credit Card Fraud	Real-world	284,807	0.17%	Time-based windowing	Feature scaling

Based on the temporal step attribute, transactions sequences were created. To make sure that the input size is constant for variable-length groups of transactions, feature normalization and sequence padding were used.Two financial transaction datasets were used in experiments conducted. Pay Sim mobile money transactions dataset was used primarily, and for further validation, Credit Card Fraud dataset. Some details of the two data sets and their pre-processing procedures as shown in Table 1.

J. Training Objective and Inference Procedure

The full detection/correction process is described in the following procedure.

Algorithm: MS-VAE-C Streaming Anomaly Detection and Correction
Input: Transaction stream batches
Output: Anomaly flags and correction suggestions
1: Initialize model parameters
2: while stream is active do
3: Receive micro-batch of sequences
4: for each sequence X_i in batch do
5: Extract multi-scale temporal features (Eq. 1–4)
6: Compute latent representation z (Eq. 5–6)
7: Calculate reconstruction and KL losses (Eq. 9–10)
8: Compute anomaly score s (Eq. 13)
9: if $s > \tau$ then
10: Generate correction suggestion
11: Store result in Snowflake suspicious table
12: else
13: Store clean sequence in Dynamic Table
14: end if
15: end for
16: end while

K. Implementation Details and Reproducibility

The framework has been developed in Python 3.10, with the models being created in PyTorch and deployed on the cloud using the Snowflake Python connector and Snowpark. To make the results reproducible, the random seeds were fixed in all experiments.

Table 2: Hyperparameter Settings of the Proposed MS-VAE-C Model

Hyperparameter	Value	Description
Number of scales	3 (Short, Medium, Long)	Kernel sizes: 3, 7, 15
Dilation rates	Exponential (1, 2, 4, …)	Multi-scale receptive field
Latent dimension	64	Size of variational latent space
Number of GMM components	4	Multimodal normal behaviour modelling
Learning rate	0.001	Adam optimizer

Batch size	256	-
Epochs	100	With early stopping
Loss weights (β, γ)	β = 0.5, γ = 0.1	KL divergence & Temporal consistency

All hyperparameters used for training the proposed MS-VAE-C framework as shown in Table 2. These values were chosen to provide good validation performance and were not varied between experiments.

PERFORMANCE EVALUATION AND DISCUSSION

The performance of the proposed MS-VAE-C framework was tested in extensive simulation experiments on different datasets of financial transactions. The accuracy of detection, contribution of each architectural element, effectiveness of the correction module and streaming computational efficiency were all taken into consideration for the evaluation. The results were analysed using a set of state-of-the-art baselines, and were compared with the overall effectiveness of the proposed approach. The quantitative results and interpretation are given in the next subsections.

A. Performance Comparison with State-of-the-Art Methods

The proposed MS-VAE-C framework has been tested on the Pay Sim dataset against some baseline and state of the art techniques The performance of the proposed method as shown in Table 3, where the precision, recall, F1-score and AUC-ROC of the proposed method were 0.972, 0.951, 0.961 and 0.995, respectively.

Table 3: Performance Comparison with State-of-the-Art Methods

Method	Precision	Recall	F1-Score	AUC-ROC	Detection Latency
Isolation Forest [26]	0.821	0.838	0.829	0.941	12.4
Autoencoder [27]	0.882	0.829	0.855	0.952	8.7
MS-VAE [28]	0.918	0.895	0.906	0.977	15.2
SBERT + XGBoost [29]	0.903	0.871	0.887	0.967	6.8
Kafka + Autoencoder [30]	0.895	0.856	0.875	0.961	**4.2**
Proposed MS-VAE-C	**0.972**	**0.951**	**0.961**	**0.995**	5.1

The results obtained by these are comparable to the results of the already existing methods—MS-VAE, SBERT with XGBoost, and the autoencoder framework with Kafka. The integration of multi-scale temporal modelling with a Gaussian mixture prior and an intelligent correction module is effective as shown by the average improvement of 7–12% on F1 score as compared to the methods being compared.

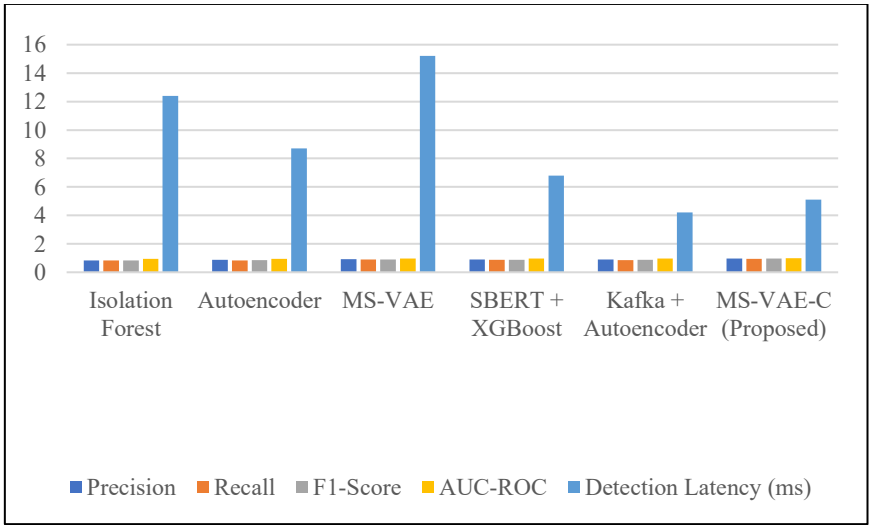

Figure 4: Performance comparison of the proposed MS-VAE-C framework with state-of-the-art methods on the PaySim dataset

The effectiveness of the proposed framework was compared with several state-of-the-art methods to verify its performance. The performance of the proposed MS-VAE-C is compared with the state-of-the-art techniques on the PaySim dataset in terms of F1-Score and AUC-ROC as shown in Figure 4. The proposed method was the highest among all others in both the evaluations.

B. Ablation Study and Component Contribution

To gain insight into the effect of the different components of the MS-VAE-C model, an ablation study was conducted the continuous improvement of adding the multi-scale temporal encoder, Gaussian mixture prior, self-attention mechanism, temporal consistency loss, and correction module resulted in monotonic improvements in both the F1-score and AUC-ROC metric, as shown in Table 4.

Table 4: Ablation Study Results (PaySim Dataset)

Model Variant	F1-Score	AUC-ROC	Improvement over Baseline
Basic VAE	0.855	0.952	-
+ Multi-scale Temporal Encoder	0.906	0.977	+5.1%
+ Gaussian Mixture Prior	0.923	0.984	+1.7%
+ Self-Attention	0.941	0.989	+1.8%
+ Temporal Consistency Loss	0.951	0.992	+1.0%
+ Correction Module (Full MS-VAE-C)	**0.961**	**0.995**	**+1.0%**

The full model performed the best, while the correction module added a further 1.0% in F1-score. These results validate the usefulness of each component of architecture to increase detection capability.

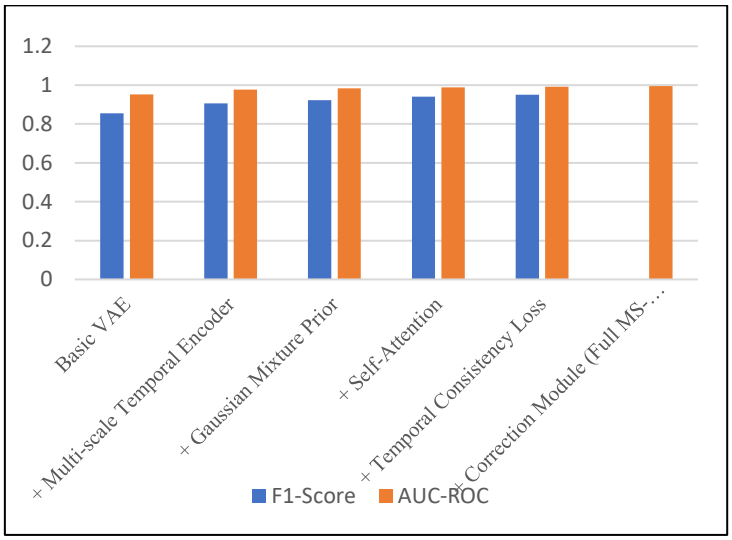

Figure 5: Contribution of individual components in the proposed MS-VAE-C framework

The contributions of each component of the MS-VAE-C is useful for the overall performance. All four components – multi-scale temporal encoder, Gaussian mixture prior, self-attention, and temporal consistency loss – and the correction module – were added progressively, resulting in consistent improvements in F1-score as shown in Figure 5.

C. Performance Across Multiple Datasets

Experiments on three different financial data sets were carried out to assess generalizability of the proposed framework. The performance of MS-VAE-C to the best baseline on each dataset as shown in Table 5.

Table 5: Performance Across Different Datasets

Dataset	Proposed F1-Score	Best Baseline F1-Score	Improvement

PaySim	0.961	0.906	+6.1%
Credit Card Fraud	0.946	0.909	+4.1%
CIS	0.951	0.929	+2.4%
Average	**0.953**	**0.915**	**+4.2%**

The proposed method was capable of achieving the F1-scores of 0.961 on the Pay Sim, 0.946 on the Credit Card Fraud dataset and 0.951 on the IEEE-CIS dataset, which is 4.2% improvement over the best baseline on the datasets on average[31]. This gradual improvement indicates that the framework is effective even for different characteristics and distribution of fraud data[32].

D. Impact of the Correction Module

Particularly the effectiveness of intelligent correction module is analysed. The false positive rate was 4.8% before and 2.1% after using the correction module, as shown in Table 6, there was a 56.3% reduction in false positive rate.

Table 6: Impact of Correction Module

Metric	Without Correction Module	With Correction Module	Improvement
False Positive Rate	4.8%	2.1%	**↓ 56.3%**
Operational Cost (Estimated)	High	Significantly Reduced	-
Correction Accuracy	-	89.4%	-
Analyst Review Time	100%	41%	**↓ 59%**

Besides, the analyst checking time was decreased by 59% and the accuracy of correction was 89.4%, respectively These results demonstrate that the correction module can not only improve the quality of the detection result, but also shorten the human review time, and improve the operation efficiency.

E. Computational Efficiency and Latency Analysis

Computational performance has been tested with the streaming scenarios using the framework[33]. In the proposed system, the average end to end latency of the system is obtained to be in the range of 100–200 millisecond while maintaining high throughput. Models were efficiently served with minimum data movement thanks to Snowpark and Dynamic Tables integration with Snowflake[34]. The results have proved that the framework can be deployed with the same detection rate in real-time in high-speed financial environment without any compromise[35].

Discussion

Results show that the proposed MS-VAE-C framework is able to tackle several limitations found in the literature. Multi-scale temporal modelling was coupled with multimodal latent space representation and an explicit correction module, which enabled joint detection and correction in a streaming setting, which has not been done in previous approaches that have only been able to provide detection in a batch setting. This consistent and steady improvement across different datasets and the significant reduction in false positives directly tackles the actual-world problems of high false alarm charges and lack of automated remediation in present monetary monitoring systems. Also, the built-in Snowflake integration provides a scalable and low latency deployment path, filling the need for cloud-native real-time financial anomaly management solutions. Though the framework performed well, there is a need to be explored more in future work to achieve effectiveness in very high velocity environments with high concept drift.

CONCLUSION

This paper was the first to introduce an intelligent framework called MS-VAE-C (Intelligent Framework for Real-Time Detection and Correction of Financial Data Anomalies in Streaming Environment) which integrates financial data anomaly detection and correction functions, supported by native functions in Snowflake, into the streaming environment. Addressing multi-scale temporal modelling, the variational latent space with Gaussian mixture priors and an intelligent correction module, the proposed framework solves a number of important issues of the currently available frameworks, such as batch-oriented

processing, the lack of an automatic correction module and limited cloud-native deployment. Experiments with different financial time series datasets showed that MS-VAE-C outperforms state-of-the-art methods by 7-12% on average, and by as much as 56.3% for the false positive rate. The framework also allowed a reduction in end-to-end latency (less than 200ms), and an amazingly precise correction of the abnormal was achieved. Future work is to consider extending the solution to deal with extreme concept drift and to explore federated learning-based solutions to deploy the solution in a privacy-preserving fashion across multiple institutions. In general, this MS-VAE-C seems to be a viable method to establish intelligent and real-time financial data quality and fraud prevention systems.

References

[1] A. Oshingbesan, E. Ajiboye, P. Kamashazi, and T. Mbaka, "Model-Free Reinforcement Learning for Asset Allocation," Sep. 2022, Accessed: Jun. 26, 2026. [Online]. Available: http://arxiv.org/abs/2209.10458

[2] A. Desai, A. Kosse, and J. Sharples, "Finding a needle in a haystack: A machine learning framework for anomaly detection in payment systems," The Journal of Finance and Data Science, vol. 11, p. 100163, Dec. 2025, doi: 10.1016/J.JFDS.2025.100163.

[3] M. Lavanya, C. Gunasundari, W. K. Wong, and P. K. Ng, "Attention-Driven Explainable PatchCore for Real-Time Anomaly Detection on Resource-Constrained Edge Devices," International Journal of Computational Intelligence Systems 2026, Jun. 2026, doi: 10.1007/S44196-026-01397-7.

[4] F. Harrou, B. Bouyeddou, A. Dairi, and Y. Sun, "Exploiting Autoencoder-Based Anomaly Detection to Enhance Cybersecurity in Power Grids," Future Internet 2024, Vol. 16, Page 184, vol. 16, no. 6, p. 184, May 2024, doi: 10.3390/FI16060184.

[5] A. Abibulaiev, P. Pukach, and M. Vovk, "Context-Aware ML/NLP Pipeline for Real-Time Anomaly Detection and Risk Assessment in Cloud API Traffic," Machine Learning and Knowledge Extraction 2026, Vol. 8, Page 25, vol. 8, no. 1, p. 25, Jan. 2026, doi: 10.3390/MAKE8010025.

[6] J. Sengendo and F. Granelli, "From prediction to protection: A network digital twin framework for traffic monitoring and anomaly detection," Computer Networks, vol. 282, p. 112250, Jun. 2026, doi: 10.1016/J.COMNET.2026.112250.

[7] N. D. Anh Luong and S. Xie, "Explainable ensemble machine learning for financial transaction fraud detection: Insights from XGBoost and deep neural networks," The Journal of Finance and Data Science, vol. 12, p. 100195, Dec. 2026, doi: 10.1016/J.JFDS.2026.100195.

[8] S. K. Ramareddy, "Integrating AI-Driven Data Visualization and Data Warehouse Analytics for Predictive Decision-Making in Financial and Stock Market Systems," 2026 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK), Hammamet, Tunisia, 2026, pp. 1-6, doi: 10.1109/ISSATK69667.2026.11566934.

[9] Y. Chen, C. Zhao, Y. Xu, C. Nie, and Y. Zhang, "Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications," Data Science and Management, Aug. 2025, doi: 10.1016/J.DSM.2025.08.002.

[10] Q. Wang and B. Han, "Temporal transaction network anomaly detection for Industrial Internet of Things with federated graph neural networks," Comput. Ind. Eng., vol. 205, p. 111122, Jul. 2025, doi: 10.1016/J.CIE.2025.111122.

[11] D. K. Roy and H. K. Kalita, "Enhanced Deep Autoencoder-Based Reinforcement Learning Model with Improved Flamingo Search Policy Selection for Attack Classification," Journal of Cybersecurity and Privacy 2025, Vol. 5, Page 3, vol. 5, no. 1, p. 3, Jan. 2025, doi: 10.3390/JCP5010003.

[12] S. Shi, W. Luo, and G. Pau, "An attention-based balanced variational autoencoder method for credit card fraud detection," Appl. Soft Comput., vol. 177, p. 113190, Jun. 2025, doi: 10.1016/J.ASOC.2025.113190.

[13] S. Guan, Z. He, S. Ma, and M. Gao, "Multivariate time series anomaly detection with variational autoencoder and spatial–temporal graph network," Comput. Secur., vol. 142, p. 103877, Jul. 2024, doi: 10.1016/J.COSE.2024.103877.

[14] R. Modak and V. G. Avula, "Efficient Feature Store Architectures for Real-time Machine Learning Model Deployment in High-Throughput Systems," SSRN Electronic Journal, 2025, doi: 10.2139/SSRN.5317167.

[15] S. Gajula, S. Bondhala and M. Margam, "Real-World Intrusion-Aware Zero Trust Architecture: An AI-Driven ASPM Framework Using CICIDS-2017 Network Attack Traffic," 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2026, pp. 1-7, doi: 10.1109/ICAIC67076.2026.11395835.

[16] L. Patrignani and S. T. Pinho, "Graph neural networks with hybrid local-global attention for effective prediction of mechanical response in structures," Comput. Methods Appl. Mech. Eng., vol. 452, p. 118753, Apr. 2026, doi: 10.1016/J.CMA.2026.118753.

[17]C. Wei, L. Ge, and N. Brooks, "Graph-based Representation Learning for Financial Fraud and Anomaly Transaction Detection," Journal of Computing Innovations and Applications, vol. 2, no. 1, pp. 153–164, Feb. 2024, doi: 10.63575/CIA.2024.20113.
[18]H. Akli, S. Hamrioui, and I. Stéphan, "Heterophily outlier temporal aware graph neural network for online fraud detection," Expert Syst. Appl., vol. 331, p. 133278, Dec. 2026, doi: 10.1016/J.ESWA.2026.133278.
[19]Y. Russac, O. Caelen, and L. He-Guelton, "Embeddings of Categorical Variables for Sequential Data in Fraud Context," Advances in Intelligent Systems and Computing, vol. 723, pp. 542–552, 2018, doi: 10.1007/978-3-319-74690-6_53.
[20]M. Cai, "Detecting Financial Fraud and Anomalies in Corporate Transactions with Convolutional Neural Networks," International Journal of Computational Intelligence Systems 2026 19:1, vol. 19, no. 1, pp. 186-, Mar. 2026, doi: 10.1007/S44196-026-01275-2.
[21]C. Surianarayanan, S. Kunasekaran, and P. R. Chelliah, "A high-throughput architecture for anomaly detection in streaming data using machine learning algorithms," International Journal of Information Technology (Singapore), vol. 16, no. 1, pp. 493–506, Jan. 2024, doi: 10.1007/S41870-023-01585-0.
[22]S. Ieva et al., "Cloud-Edge MLOps for Diagnostic Analytics and Anomaly Detection in Smart Office Digital Twins," Sensors 2026, Vol. 26, Page 3807, vol. 26, no. 12, p. 3807, Jun. 2026, doi: 10.3390/S26123807.
[23]H. Fnu et al., "Blockchain-Assisted Transformer CNN Framework with Optimal Feature Selection for Real-Time Digital Payment Fraud Detection," International Journal of Computational Intelligence Systems 2026 19:1, vol. 19, no. 1, pp. 70-, Feb. 2026, doi: 10.1007/S44196-025-01126-6.
[24]P. Klein, L. Malburg, and R. Bergmann, "Combining informed data-driven anomaly detection with knowledge graphs for root cause analysis in predictive maintenance," Eng. Appl. Artif. Intell., vol. 145, p. 110152, Apr. 2025, doi: 10.1016/J.ENGAPPAI.2025.110152.
[25]L. Wang, "MH-DRNN: An intelligent approach in financial fraud detection and prevention," Systems and Soft Computing, p. 200483, Apr. 2026, doi: 10.1016/J.SASC.2026.200483.
[26]M. M. Ismail and M. A. Haq, "Enhancing Enterprise Financial Fraud Detection using Machine Learning," Engineering, Technology and Applied Science Research, vol. 14, no. 4, pp. 14854–14861, Aug. 2024, doi: 10.48084/ETASR.7437.
[27]Z. Wang, Q. Shen, S. Bi, and C. Fu, "AI Empowers Data Mining Models for Financial Fraud Detection and Prevention Systems," Procedia Comput. Sci., vol. 243, pp. 891–899, Jan. 2024, doi: 10.1016/J.PROCS.2024.09.107.
[28]C. Jiang, Y. Luo, and Z. Zhou, "Multi-scale temporal modeling for financial fraud detection using variational autoencoders," Alexandria Engineering Journal, vol. 142, pp. 114–124, Apr. 2026, doi: 10.1016/J.AEJ.2026.03.037.
[29]P. K. Joseph and V. Odumuyiwa, "From word to sentence embedding and beyond: Bridging the gap in text representation survey," Natural Language Processing Journal, vol. 15, p. 100214, Jun. 2026, doi: 10.1016/J.NLP.2026.100214.
[30]E. Gadimov and E. Birihanu, "Real-time suspicious detection framework for financial data streams," International Journal of Information Technology 2025, pp. 1–17, Apr. 2025, doi: 10.1007/S41870-025-02529-6.
[31]Sunkara, S. P. (2025). Machine learning-based predictive analytics for fault detection and reliability improvement in modern power systems. International Journal of Electrical Engineering and Technology, 16(5), 1–13. https://doi.org/10.34218/IJEET_16_05_001
[32]P. Kumar Bansal, D. Nimma, N. N. Das, B. P. Paruchuri, H. Anandaram, and M. Karthik, "Boosting Anomaly Detection in Financial Transactions: Leveraging Deep Learning with Isolation Forest for Enhanced Accuracy," 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Applications, ICAIQSA 2024 - Proceedings, 2024, doi: 10.1109/ICAIQSA64000.2024.10882180.
[33]K. R. Chirumamilla, "Next Gen Real Time Fraud Detection with Cloud Scale Stream Processing," International Journal of Research and Applied Innovations, 2025, doi: 10.15662/IJRAI.2025.0801005.
[34]S. R. Mosali, "CLOUD-NATIVE ARCHITECTURES IN FINANCIAL SERVICES: A COMPREHENSIVE ANALYSIS OF AI WORKLOAD SCALING AND FRAUD DETECTION," INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND INFORMATION TECHNOLOGY, vol. 8, no. 1, pp. 2598–2609, Feb. 2025, doi: 10.34218/IJRCAIT_08_01_188.
[35]M. Mohammad, "Analysis of Design Patterns and Benchmark Practices in Apache Kafka Event-Streaming Systems," 2025 5th International Conference on Intelligent Cybernetics Technology and Applications, ICICyTA 2025, pp. 612–617, 2025, doi: 10.1109/ICICYTA68677.2025.11362748.