SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligen and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Self-Supervised Representation Learning in Sparse Data Regimes

Syed Rashid Anwar[1], Josephine R[2], Santosh Kumar Behera[3], Vinay Kumar Sadolalu Boregowda[4], Pushpa Nagini Sripada[5], Sivasankari V[6]

[1] Assistant Professor, Department of Computer Science & IT, Arka Jain University, Jamshedpur, Jharkhand, India. Email: syed.r@arkajainuniversity.ac.in, ORCID: 0000-0001-9810-8850
[2] Assistant Professor, Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. Email: josephine.r@presidencyuniversity.in, ORCID: 0009-0008-5982-1352
[3] Associate Professor, Centre for Artificial Intelligence and Machine Learning, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Email: santoshbehera@soa.ac.in, ORCID: 0009-0009-1797-9546
[4] Assistant Professor-I, Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, Karnataka, India. Email: sb.vinaykumar@jainuniversity.ac.in, ORCID: 0000-0001-7349-1697
[5] Professor, Department of English, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: sripadapn@maher.ac.in
[6] Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: sivasankariv@maher.ac.in

Abstract
Self-supervised representation learning enables scalable modeling by reducing reliance on labeled data. Yet fraud detection in financial systems remains challenged due to sparse, imbalanced observations and evolving transaction patterns, where existing methods fail to generalize and capture minority fraudulent behaviors effectively. This research aims to design a robust fraud detection framework for sparse and highly skewed financial datasets by leveraging self-supervised representations. Transactional data are aggregated from publicly available financial repositories and simulated streams reflecting realistic fraud scenarios, ensuring diversity in temporal and categorical attributes. Preprocessing incorporates normalization and missing value imputation. Feature extraction employs Time2Vec temporal encoding and rolling window statistical descriptors to capture evolving behavioral patterns. The proposed model integrates representation learning with Genetic Algorithm-tuned Dynamic Variational Autoencoders (GA-DVA), where data are first encoded into latent representations and subsequently refined for anomaly-aware discrimination. The Dynamic Variational Autoencoder models evolving transaction distributions, while the Genetic Algorithm optimizes latent space parameters and reconstruction constraints to enhance detection sensitivity under imbalanced datasets. This combination enables adaptive learning of rare fraud signatures. For robust fraud detection in sparse and imbalanced financial datasets, the framework prioritizes minority pattern amplification and distribution-aware learning using Python. Performance evaluation demonstrates improved precision (0.920), recall balance (0.912), F1-Score (0.916), AUC-ROC (0.950), Early Detection Rate (0.890), and reduced false alarms (0.050), and consistent adaptability to shifting data distributions. The approach delivers interpretable, scalable, and resilient fraud detection suitable for real-world financial environments.

Keywords: Self-Supervised Learning, Fraud Detection, Imbalanced Data, Anomaly Detection, Representation Learning.

1. Introduction

Financial fraud is a prevalent dynamic issue in the contemporary financial system, which has necessitated innovative methodologies and new technologies to effectively identify and prevent losses brought about by more sophisticated fraudsters [1]. The increased scale and complexity of financial fraud is becoming a major threat to the global financial system due to the rapid pace of digital transformation, and the use of Artificial Intelligence (AI)-based technologies, which offer new vulnerabilities exploited by criminals, especially in the regions that have inconsistent regulatory frameworks and varying degrees of technological readiness [2]. Bank account opening fraud is a serious issue in financial services and involves criminals using stolen or synthetic identities to open accounts, which they use to perpetrate money laundering, unauthorized credit access, and other illegal actions, causing significant operational and reputational risks to institutions [3]. Fraud detection is essentially a type of anomaly detection, where fraudulent behavior is very rare, far out of normal behavior, and often buried in highly unbalanced datasets with inconsistent or incomplete labels [4].

The high rate of development of electronic payment technologies based on online transactions in the mainstream are amplifying the incidence of fraud. The sheer amount of daily transactions however, makes it hard to effectively identify fraud manually by humans [5, 6]. Thus, the timely identification of fraudulent activities is vital and is conducted as the financial data streams are received. To overcome this problem, financial institutions put a lot of money into new methods that use AI to enhance the effectiveness of fraud detection systems. AI-based fraud detection algorithms are fast, efficient and flexible towards detecting fraudulent transactions [7, 8].

Aim of the research: The purpose of this study is to design a robust fraud detection model of sparse and unbalanced financial data. To identify changing fraudulent activities, it makes use of Genetic Algorithm-tuned Dynamic Variational Autoencoders (GA-DVA) and self-supervised representation learning. The approach enhances sensitivity to detection, precision-recall tradeoff, and flexibility to non-uniform data distributions by focusing on minority pattern amplification and distribution-sensitive learning.

Organization of the research: Section 1 gives the background of the research. Section 2 will be devoted to the discussion of the existing research. In Section 3, the data collection, preprocessing, and future extraction methodology and proposed model, GA-DVA, are discussed. Section 4 contains findings of the experimental study. The conclusion with the performance, limitations and future work is presented in section 5.

2. Related Work

A research, based on its purpose, procedures, obtained findings, and other limitations, a well-structured understanding of the latest developments and gaps in research in the field of detecting fraud in sparse data. To achieve higher rate of fraud detection, the problem involving extremely unbalanced data was explored [9], and the MoE algorithm which uses a DNN-SMOTE was used to synthesize data and improve the minority group. The technique has a high success rate and the use of synthetic samples minimizes the generalization and strength. Create proper financial statement fraud prediction model. An analytical model of data-based fraud prediction, which utilizes financial statements, was studied [10]. The sampling methods and BTC to detect fraud integrates under sampling and ensemble learning to detect fraud better. Findings indicate that there is an increased accuracy in the detection of fraud. Nonetheless, generalization of the model is constrained by the dependency and imbalance sensitivity of the dataset. A Fraud-Aware Fusion of detecting the fraud data problem was investigated [11], using a F-GNN to detect the fraud data. With high scores of AUC in the experiments, there are still problems concerning the high costs of spectra and issues of generalization. Design effective and interpretive contrastive learning to detect anomalies [12]. The contrastive objective is enhanced with a LFAM that co-optimizes the process of representation learning and explanation faithfulness. The datasets of experiments indicate a 47% AUC improvement over baselines. Some of the limitations are scalability limitations and domain transfer limitations. In [13], the problem of detecting fraud in the evolving heterogeneous dynamic graphs was examined. The proposed model is a Temporal Feedback Manager that uses sparse multi-scale LSTM networks with adaptive graph convolutions and attention fusion to capture the characteristic of time and generate grouping weights to learn the feature better. The experimental results indicate that the model is more effective than the current approaches in terms of the AUC and AP scores. In [14] an adaptive mechanism of reconstruction threshold was developed, and the AE, SOM, and RBM techniques were combined. The Kaggle dataset was used to test the model with the AE-ASOM model showing a 98.0% accuracy and 96.7% F1-score. The approach is, however, constrained by its reliance on particular datasets, and its susceptibility to overfitting when dealing with changing fraud patterns. GTCT which uses a combination of graph, temporal and contrastive modules was explored [15]. The result achieves (0.975) accuracy and (0.982) AUC, removing contrastive loss reduces recall/AUC to 0.805/0.948, removing the graph encoder lowers F1 to 0.786, and removing the temporal encoder reduces recall to 0.743 and AUC to 0.905, limited by imbalance and complexity. The detection of account opening fraud with highly unbalanced datasets was explored [16]. Compare CLR, RF to validate and Weighted Optimization, with class-weighted loss functions. The result indicates that classical algorithms achieve 1-6% recall, GRU achieves 78% recall with 5% precision (AUC 0.8800). The constraints include issues of data imbalances and metric bias in fraud detection systems. adaptive federated reinforcement learning with knowledge distillation to develop BSAFD [17] model integrated with PPO-based federated reinforcement learning. The proposed method brought about considerable improvements in AUC, F1-score, and average precision over the baseline methods. Nevertheless, the model has limitations based on the cost of computation, on the involvement of the client group, and on the quality of the synthetic data generation. In [18], the multilingual anomaly detection in low-resource settings with limited labels was investigated. It uses a LR-SSAD to detect anomalies in multilingual modeling

transactions. When evaluated on a financial dataset, it has a 93.2% accuracy, 91.4% precision, 89.1% recall and 94.8% AUC. Limitations consist of dependency and sensitivity. The fraud detection with the limited supervision with the help of unlabeled data was devised in [19]. With the class imbalance and heterophily improving the accuracy using the GFD it achieves the state-of-the-art results on three benchmarks with significant gains. One of the limitations is that they depend on pseudo-label quality bias. A self-supervised hierarchical learning architecture to detect accurately anomalies in accounting and to issue early risk warning [20]. The method uses HFSL for detecting fraud data. Results show precision 83.6%, recall 80.5%, F1 82.0%, early detection 72.6%, and false alarm 6.8%.

3. Methodology

Figure 1 shows a powerful fraud detection model that is meant to overcome the issues of imbalanced, limited, and dynamically varying financial data. The proposed system incorporates Rolling Window Statistical Descriptors, feature extraction based on Time2Vec, data normalization techniques used for preprocessing, and hybrid GA-DVA model. The suggested approach would allow the adaptive, scalable, and the highest accuracy of detecting fraud.

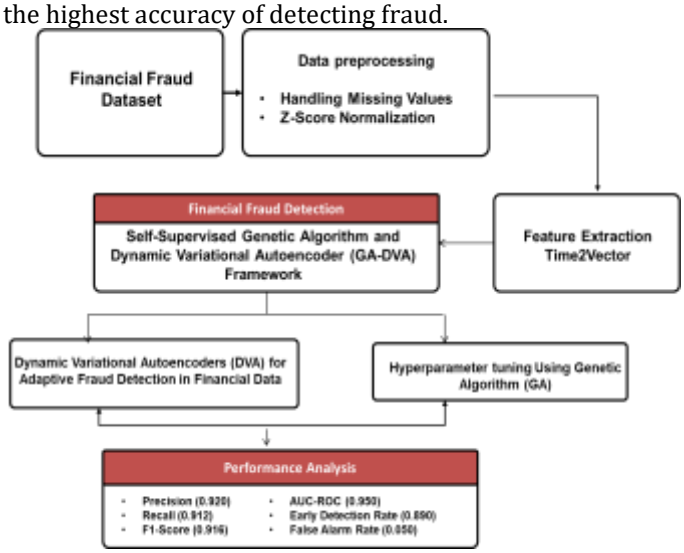

Figure 1: Methodology Workflow for Fraud Detection in Sparse and Imbalanced Datasets

3.1 Data Collection

The publicly available dataset utilized in the current study is the Fraud Detection Dataset in Kaggle (https://www.kaggle.com/datasets/goyaladi/fraud-detection-dataset?select=Data), which offers an in-depth collection of anonymous financial transactions to study the pattern of fraudulent behavior. The Kaggle dataset consists of a variety of heterogeneous elements, such as customer profiles, transactions, merchant information, fraud indicators, anomaly scores, transaction amounts, categorical labels, and behavioral characteristics. These various data sources can be used to model time and anomaly-based trends needed to support self-supervised learning, and to support rich and comprehensive feature representation in the context of fraud detection tasks. In order to make sure that the model evaluation is sound and in order to test the efficiency of the proposed fraud detection model based on the GA-DVA model, the dataset is split into 80% training data and 20% testing data.

3.2 Median Imputation Technique for skewed data in Financial value Data in Fraud Detection Systems

Missing data is a common issue in financial fraud detection that can adversely affect model performance, especially in unbalanced datasets where illegal transactions are rare. One approach that is popular in handling missing values is median imputation, particularly when the data is skewed or has outliers such is the case in financial information. This technique helps preserve the distribution of the data without being unduly impacted by extreme values or outliers by substituting missing values with the median of the observed values for a certain attribute. The given technique is represented in the following Equation (1).

$$\widehat{w}ji = \text{median}(wji : w_{ji} \in D_l) \quad (1)$$

The imputed value for the i^{th} transaction's missing j^{th} feature is represented as w_{ji}. Each wji refers to a single valid data point, and the phrase $w_{ji} \in D_l$ represents the observed values of the j^{th} feature across all accessible transactions in the labeled dataset D_l. In sparse and unbalanced financial fraud detection

datasets, missing items are filled in using the median of these values, guaranteeing robustness against skewed distributions.

3.3 Z-Score Normalization for Preprocessing Financial Data in Fraud Detection Systems

Once data has been collected, the data is preprocessed to detect financial fraud, especially when working with imbalanced datasets, which are characteristic of instantaneous situations. The preprocessing technique used to solve this problem is z-score normalization, which aims to normalize the distribution of the data points. It converts raw values to normalized scores that reflect the deviation of each data point to the mean with respect to the standard deviation. The approach can be especially useful in the case of financial data, where such characteristics as the number of transactions, the frequency of transactions, and time intervals may vary significantly in scale, as shown in Equation (2).

$$Y = \frac{W-\upsilon}{\sigma} \quad (2)$$

w is the initial transaction feature value of amount, υ is the dataset feature mean, and σ is the standard deviation. To ensure consistent feature representation across highly imbalanced and sparse fraud detection datasets, the normalized output y transforms raw financial data into a standardized scale.

3.4 Time2Vec-Based Adaptive Feature Extraction for Fraud Detection in Financial Transactions

After the median imputation missing value method of preprocessing unbalanced financial data in fraud detection systems, anomaly detection must be effective in capturing the temporal patterns. Time2Vec is an effective approach to model periodic and non-periodic temporal features by extracting adaptive features out of time-dependent transaction data. The approach allows detecting anomalous infrequent fraud that is not based on ordinary transaction patterns in unbalanced financial data. Moreover, Time2Vec not only increases the scalability and flexibility of fraud detection systems in the sense that it effectively captures time-varying patterns of fraud across different time intervals as mathematically represented in Equation (3).

$$T2V(T)[j] = \begin{cases} \omega_j T + \varphi_j, & \text{if } j = 0 \\ E(\omega_j T + \varphi_j), & \text{if } 1 \le j \le l \end{cases} \quad (3)$$

T2V(T)[j] is the j^{th} component of the Time2Vec embedding, which converts raw time into a learnt feature space for spotting fraud patterns in sparse and unbalanced datasets. T is the input transaction timestamp used to record temporal behavior in financial fraud detection. By modeling periodic and non-periodic transaction behaviors, the learnable frequency and phase variables ω_j and φ_j aid in the detection of infrequent and changing fraudulent activity. The embedding dimension is defined by the index $j = 0$ to l, where $j = 0$ captures linear time progression $(\omega_j T + \varphi_j)$ and $j > 1$ captures nonlinear temporal patterns using the function $E(\omega_j T + \varphi_j)$, enhancing fraud detection robustness under data imbalance.

3.5 Financial Fraud Detection using Self-Supervised Genetic Algorithm and Dynamic Variational Autoencoder (GA-DVA) Model

a self-supervised GA-DVA architecture to detect fraud in a small and unbalanced amount of financial data. The GA is used to optimize the latent space parameters, reconstruction constraints whereas the DVA is used to learn latent representations of transactions patterns. This integration yields a scalable and reliable fraud detection system by making it more sensitive to unusual fraudulent activity, more sensitive to anomalies and responsive to adaptive learning of evolving patterns. DVA to Adaptive Fraud Detection in Financial Data: This research applies DVA as an important component to adaptive fraud detection in unbalanced datasets to effectively detect the changing pattern of financial transactions. Using dynamic behavior modeling enables DVA to more effectively identify unusual and evolving fraudulent activities by continuously evolving to new transaction distributions. By mapping input data y into a latent variable w using an encoder $(w = Encoder(x))$ and reconstructing the input using a decoder$(y = Decoder(z))$, the AutoEncoder (AE) learns a compact representation. The optimization objective of Evidence Lower Bound (ELBO) is expressed in Equation (4).

$$\mathcal{L} = \mathbb{E}_{h_{\Phi(y|w)}} \log[p_\theta(y|w)] - C_{LK}(h_\Phi(y|w)||o(y)) \quad (4)$$

$\mathcal{L}$ represents the evidence of the lower bound, $\mathbb{E}_{h_{\Phi(y|w)}}$ denotes the distribution of the variational posterior, and $h_\Phi(y|w)$, where y is the input of financial detecting input transaction data, and w is the latent variable. $\log[p_\theta(y|w)]$ represents the reconstruction of fraudulent transaction data to detect. $C_{LK}(h_\Phi(y|w)$ is the Kullback–Leibler (KL) divergence, which regularizes the learned posterior distribution $h_\Phi(y|w)$ to remain latent variable for the financial transaction data o(y), preventing overfitting and improving generalization in sparse fraud datasets.

Hyperparameter tuning using Genetic Algorithm (GA): This research uses a GA to optimize model parameters, specifically in the suggested learning model, to improve fraud detection performance in sparse and unbalanced financial datasets. Because the optimization problem requires tuning parameters like latent space configurations and reconstruction constraints, it is intrinsically non-convex and discrete. GA is used as a heuristic, evolutionary method to effectively search the solution space and obtain near-optimal configurations because traditional optimization techniques are computationally costly. GA starts with a random population and uses crossover, mutation, and selection to iteratively evolve solutions. By enhancing model flexibility, this procedure makes it possible to detect uncommon fraudulent patterns effectively without the need for gradient-based optimization, which makes it appropriate for complex financial data distributions, as expressed in Equation (5).

$$\underset{z}{\text{Minimize}}\ o(z) \quad (5)$$

For sparse and unbalanced financial datasets, z is the decision vector that represents all tunable model parameters (latent dimensions, reconstruction weights) utilized in fraud detection. To increase the detection accuracy of uncommon fraudulent patterns, the Genetic Algorithm seeks to minimize the objective or fitness function (such as reconstruction loss or detection error) represented by the function o(z). Consistency of fraud data is expressed in Equation (6).

$$h(z) = 0 \quad (6)$$

Having equality constraints on the parameter vector z, $h(z)$ represents this constraint. These constraints assure that the optimized solution satisfies the structural or model-specific conditions, which ensure the stability and consistency of detection of fraud under highly imbalanced distributions of financial data, and the optimization feasibility procedure expressed in Equation (7).

$$g(z) = 0 \quad (7)$$

$g(z)$ are optional constraint functions which are used to define the feasibility requirements of the optimization process. It removes over model configurations of sparse datasets of financial fraud detection model. Equation (8) represents mathematically the tendency of fraud.

$$z^k \leq z_j \leq z^v : integer, j = 1, \dots, m \quad (8)$$

The z_j is a parameter of the j th parameter and z^k and z^v are the lower and upper limits of the parameter. This enhances the strength of the Genetic Algorithm in detecting minority fraud tendencies in unbalanced financial data by regulated exploration of search space. The overall number of adjustable parameters is specified by the index integer values $j = 1, \dots, m$. The disparity in the financial dataset parameters is in the form of Equation (9).

$$z = [z_1, z_2, \dots, z_m]^T \quad (9)$$

z is the whole solution vector (z_1 *to* z_{m}) that includes all model parameters. The GA-optimized fraud detection model uses each element to represent a particular hyperparameter. T represents transpose, which converts row representation into optimization processing for extremely imbalanced and sparse financial datasets. The data on financial transactions is preprocessed by Algorithm 1 in the form of median imputation, Z-score normalization, and Time2Vec feature extraction. The GA optimizes hyperparameters of the DVA, which in turn recreates the pattern of transactions. The fraudulent transactions are detected on the basis of reconstruction error thresholding.

Algorithm: GA-DVA for Financial Fraud Detection

```
Input dataset D={x1,x2,...,xn}, threshold θ
Split D into training set Dt and testing set Ds
Initialize latent vector z, fitness function F(z), population P
For each transaction xi ∈ Dt do
  If xi contains missing values then
    xi ← MedianImputation(xi)
  End If
  yi ← ZScoreNormalization(xi)
  ti ← Time2Vec(yi)
End For
While generation g < Gmax do
  z ← Encoder(ti), x̂i ← Decoder(z)
  If F(z) is optimal then update best parameters
  P ← Selection + Crossover + Mutation(P)
If ||xi − x̂i|| > θ then Fraud else Normal
```

Hyperparameter of Proposed GA-DVA: The proposed self-supervised representation learning model on sparse and skewed financial data employs well-tuned hyperparameters to achieve consistent learning,

useful latent representation and efficient fraud detection. Although latent dimension (32256) that allows compact representation learning, training parameters such as batch size (32256), learning rate (1e-41e-3), and Adam / AdamW optimizer provide efficient convergence. To balance between VAE and contrastive losses, α (0.5-1.0), 0 (0.001-0.1), and temperature 0 (0.05-0.5) are applied. Scalability, stability and generalization are boosted by the architecture and regularization parameters.

4. Result

The suggested model is much better at detecting fraud in thin and imbalanced financial data through utilising GA-DVA and Self-Supervised Representation Learning. The model is better than the other models in improving metrics, Early Detection Rate, and False Alarm rate by adapting to the dynamics of transaction distributions. It ensures that detection can be reliable and scalable in instantaneous settings by enhancing minority fraud tendencies, and by optimizing latent representations.

4.1. Experimental Setup

The experimental system is compatible with scalable training on sparse and imbalanced financial data with a high-performance computer, including Intel i7/i9 or Ryzen 7/9 central processing unit, NVIDIA RTX graphics card (≥8GB VRAM), 16-64 GB RAM, and SSD memory. The software environment is Python (3.83.11), PyTorch/TensorFlow, and CUDA 11.x, NumPy, and Pandas to effectively implement it.

4.2. Exploratory Analysis

Figure 2(a) displays a concentration of transaction amounts between 20 and 100, and it presents the sparsity and skewness of the Imbalanced Financial data. Figure 2(b) shows both classes, the median transaction amount is close to sixty, and also illustrates the imbalance of Fraud Data and the overlap of classes. Figure 2(c) visualizes the 98% noise, and fraudulent points (yellow) within dense legitimate transactions are barely noticeable, highlighting the "needle in a haystack" problem. Figure 2(d) presents the Time2Vec temporal encoding, which maps cyclical transaction behaviors into a geometric space to improve the GA-DVA's capacity to identify changing fraud signatures.

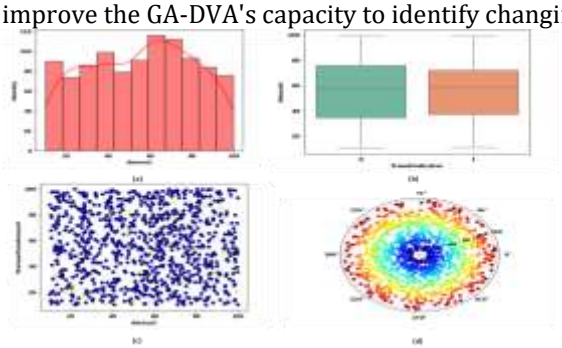

Figure 2: Exploratory Data Analysis of (a) Transaction Amount Density, (b) Transaction Amounts for Non-Fraudulent and Fraudulent Indicators, (c) Signs of Minority Fraud in Transaction Trends, and (d) Behavioral Patterns for Temporal Encoding

4.3. Evaluation Metrics

The model suggested measures the performance of fraud detection in sparse and skewed financial information using various measures. Precision is a measure of how accurately the model can identify all instances of fraud, especially those in the minority category. Recall is a measure of how well the model can identify all instances of fraud, especially those who belong to the minority category. F1-score is a balance of precision and recall to achieve strong performance. AUC-ROC measures the model to discriminate between cases of fraud and non-fraud. Also, the Early Detection Rate is the measure of timely detecting the fraud, and the False Alarm Rate is the measurement of the timely false detection of the fraud.

4.4. Performance Evaluation

Table 1 and Figure 3 are comparative studies of the performance of different existing methods of fraud detection using six major metrics. Uniclass Support Vector Machines (SVM) [20] presents the least measures. The One-Class Support Vector Machines (SVM) [20] show the lowest metrics. The Isolation Forest [20] method shows the results better to compare with the One-Class SVM, and the LSTM-AE [20] shows the lowest metrics. The VAE [20] method performs the fewest results compared to the LSTM-AE. The HFSL [20] method shows the results better for comparison with the previous four methods. The proposed GA-DVA method is the most efficient method overall and performs better than any other. Its False Alarm Rate (FAR) is much lower at 5.0%, its Early Detection Rate (EDR) is 89.0%, and its AUC-ROC is 95.0%.

Table 1: Comparison Analysis of Existing and Proposed Methods

Methods	Precision	Recall	F1-Score	AUC-ROC	EDR	FAR
One-Class SVM [20]	0.722	0.663	0.691	0.724	0.504	0.107
Isolation Forest [20]	0.736	0.682	0.708	0.753	0.522	0.103
LSTM-AE [20]	0.795	0.734	0.763	0.832	0.631	0.080
VAE [20]	0.772	0.727	0.749	0.821	0.617	0.089
HFSL [20]	0.836	0.805	0.820	0.883	0.726	0.068
GA-DVA [Proposed]	0.920	0.912	0.916	0.950	0.890	0.050

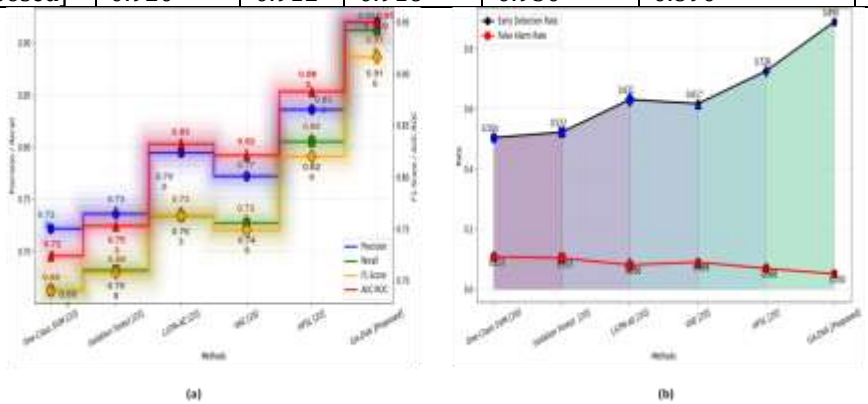

Figure 3: Fraud Detection in Sparse and Imbalanced Financial Datasets ,(a) Comparative Metric Analysis and (b) Operational Efficiency Analysis

4.5 Discussion

A Self-Supervised Representation Learning to develop a strong fraud detection model specifically for sparse and unbalanced financial information. The One-Class SVM [20] method faces problems related to scalability and effectiveness due to high imbalance and dimensionality of financial data. IF [20] demonstrates low sensitivity to changes in patterns of fraud in complex transaction scenarios. LSTM-AE [20] is limited by overfitting issues, which prevent the generalization process under conditions of sparse labeled financial datasets. VAE [20] can result in blurred features in the latent space, leading to decreased precision in distinguishing fraud. HFSL [20] is limited by scalability issues on very large-scale datasets, while it provides a low false alarm rate (6.8%) and a high early detection rate (72.6%). GA-DVA predicts changing transaction distributions and improves latent representations for better anomaly identification, the suggested method tackles these issues.

5. Conclusion

The research develops a robust fraud detection model for sparse and imbalanced financial datasets using self-supervised representation learning. Transactional data is gathered from publicly accessible repositories, and preprocessing methods like Normalization, Missing Value Imputation, and Time2Vec encoding are used for data preprocessing and feature extraction purpose. The GA-DVA model demonstrated notable improvements, achieving a precision of 92.0%, a recall of 91.2%, and an F1-score of 91.6%, an AUC-ROC of 0.950, an EAD of 0.890, and a FAR of 0.050. While the proposed model demonstrates robust performance, its scalability to extremely large financial datasets remain a challenge. Future research includes improving model efficiency, transfer learning of rapidly changing and dynamic data distributions, and much more sophisticated approaches to handling highly volatile and changing data distributions.

REFERENCES

1. Breskuvienė, D. and Dzemyda, G., 2024. Enhancing credit card fraud detection: highly imbalanced data case. Journal of Big Data, 11(1), p.182. https://doi.org/10.1186/s40537-024-01059-5
2. Al-Daoud, K.I. and Abu-AlSondos, I.A., 2025. Robust AI for financial fraud detection in the GCC: A hybrid framework for imbalance, drift, and adversarial threats. Journal of Theoretical and Applied Electronic Commerce Research, 20(2), p.121. https://doi.org/10.3390/jtaer20020121
3. Maashi, M., Alabduallah, B. and Kouki, F., 2023. Sustainable financial fraud detection using Garra Rufa fish optimization algorithm with ensemble deep learning. Sustainability, 15(18), p.13301. https://doi.org/10.3390/su151813301

4. Kennedy, R.K., Salekshahrezaee, Z., Villanustre, F. and Khoshgoftaar, T.M., 2023. Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. Journal of Big Data, 10(1), p.106. https://doi.org/10.1186/s40537-023-00750-3
5. Cheah, P.C.Y., Yang, Y. and Lee, B.G., 2023. Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. International Journal of Financial Studies, 11(3), p.110. https://doi.org/10.3390/ijfs11030110
6. KK, G., BV, A., Reddy, H.P., Mayur, N.R. and Bhowmik, B., 2026. Mitigating Class Imbalance in Banking Transactions: A Graph-Based GAN Solution for Fraud Detection. Computational Economics, pp.1-53. https://doi.org/10.1007/s10614-025-11219-1
7. Goswami, S. and Singh, A.K., 2026. An adaptive oversampling method WRM-SMOTE for credit card fraud imbalanced datasets. Intelligent Data Analysis, p.1088467X261422918. https://doi.org/10.1177/1088467X261422918
8. Tayebi, M. and El Kafhali, S., 2025. Generative modeling for imbalanced credit card fraud transaction detection. Journal of Cybersecurity and Privacy, 5(1), p.9. https://doi.org/10.3390/jcp5010009
9. Yang, Z., Wang, Y., Shi, H. and Qiu, Q., 2024. Leveraging mixture of experts and deep learning-based data rebalancing to improve credit fraud detection. Big Data and Cognitive Computing, 8(11), p.151. https://doi.org/10.3390/bdcc8110151
10. Zhou, Y., Xiao, Z., Gao, R. and Wang, C., 2024. Using data-driven methods to detect financial statement fraud in the real scenario. International Journal of Accounting Information Systems, 54, p.100693. https://doi.org/10.1016/j.accinf.2024.100693
11. Zhao, W. and Chen, H., 2026. Learning Frequency-aware Graph Fraud Detection. Neural Networks, p.108600.
12. Zheng, J., Qiu, W. and Huang, S., 2026. Interpretable Contrastive Learning for Robust and Explainable Anomaly Detection in Financial and Organizational Data. International Journal of Pattern Recognition and Artificial Intelligence, 40(01), p.2551029. https://doi.org/10.1142/S0218001425510292
13. Zhen, W., Dai, Q., Chen, L. and Liu, D., 2026. TF-GNN: A temporal feedback graph neural network with discriminative regularization for heterogeneous fraud detection. Neurocomputing, p.133694. https://doi.org/10.1016/j.neucom.2026.133694
14. Adejoh, J., Owoh, N., Ashawa, M., Hosseinzadeh, S., Shahrabi, A. and Mohamed, S., 2025. An Adaptive Unsupervised Learning Approach for Credit Card Fraud Detection. Big Data and Cognitive Computing, 9(9), p.217. https://doi.org/10.3390/bdcc9090217
15. Olaniyan, J., Olaniyan, D., Obagbuwa, I.C. and Ngafeeson, M., 2025. Graph-Temporal Contrastive Transformer for Financial Fraud Detection Using Transaction Behavior Modeling. Algorithms, 18(12), p.770. https://doi.org/10.3390/a18120770
16. Sun, W., Shen, Q., Gao, Y., Mao, Q., Qi, T. and Xu, S., 2025. Objective over architecture: Fraud detection under extreme imbalance in bank account opening. Computation, 13(12), p.290. https://doi.org/10.3390/computation13120290
17. Zhang, N., Li, Q., Zhu, K. and Zhu, D., 2025. Beyond siloed aggregation: An adaptive federated reinforcement learning model with multi-level knowledge distillation against evolving financial fraud. Displays, p.103205. https://doi.org/10.1016/j.displa.2025.103205
18. Wang, M., Wang, N., Mei, L., Li, Y., Liu, X., Hua, S. and Li, M., 2026. Data-Driven Cross-Lingual Anomaly Detection via Self-Supervised Representation Learning. Electronics, 15(1), p.212. https://doi.org/10.3390/electronics15010212
19. Yu, C., He, X., Wang, J. and Zhang, B., 2025. GSLA: Graph fraud detection with structure enhancement and label augmentation under limited supervision. Neurocomputing, p.132291. https://doi.org/10.1016/j.neucom.2025.132291
20. Zhang, Y. and Duan, B., 2025. Accounting data anomaly detection and prediction based on self-supervised learning. Frontiers in Applied Mathematics and Statistics, 11, p.1628652. https://doi.org/10.3389/fams.2025.1628652

Variable / Term	Explanation
MoE	Mixture of Experts model for handling imbalanced fraud data
DNN-SMOTE	Deep Neural Network-based Synthetic Minority Over-sampling Technique
BTC	Boosting Tree Classifiers for ensemble-based fraud detection
F-GNN	Frequency-aware Gated Neural Network for fraud pattern learning
LFAM	Learnable Feature Attribution Mask for interpretable anomaly detection
LSTM	Long Short-Term Memory network for temporal feature learning
AE	Autoencoder for unsupervised fraud pattern reconstruction
SOM	Self-Organizing Map for clustering and anomaly detection

RBM	Restricted Boltzmann Machine for probabilistic representation learning
AE-ASOM	Autoencoder with Adaptive Self-Organizing Map framework
GTCT	Graph-Temporal Contrastive Transformer for financial fraud detection
CLR	Classical Logistic Regression for fraud classification
RF	Random Forest classifier for validation and fraud prediction
GRU	Gated Recurrent Unit for sequential fraud detection
BSAFD	Beyond Siloed Aggregation Fraud Detection framework
PPO	Proximal Policy Optimization for federated reinforcement learning
LR-SSAD	Low-Resource Self-Supervised Anomaly Detector
GFD	Graph-based Fraud Detection framework
HFSL	Hierarchical Fusion Self-Supervised Learning framework
AUC	Area Under the Curve for classification performance evaluation