Science Journal of University of Zakho
DOI: https://doi.org/10.25271/sjuoz.2026.14.4.1887

Original Article

DETECTION OF BITCOIN FRAUD USING VARIATIONAL AUTO ENCODER AND LIGHT GRADIENT BOOSTING MACHINE

Muhammad Nazeer Musa[1], Abdulkadir Nafisat[2,*], Adebajo Olamilekan Adewale[3], Ramat Abba Imam[4], and Samaila Musa Abdullah[4]

[1]Department of Cyber Security, Nigerian Defence Academy, Kaduna State, Nigeria.
[2]Department of Computer Science, Confluence University of Science and Technology, Osara Kogi State, Nigeria.
[3]Division of Computing Science and Mathematics Faculty of Natural Sciences, University of Stirling, Stirling, Scotland, United Kingdom.
[4]Department of Cyber Security, Air Force Institute of Technology, Kaduna State, Nigeria.

[*]*Corresponding author, E-mail:*abdulkadirn@custech.edu.ng (Tel: +234-8165511221)

Received
January 25, 2026

Accepted
August 17, 2026

Published
October 20, 2026

ABSTRACT

The proliferation of fraudulent behaviour in cryptocurrency ecosystems poses a significant threat to financial security infrastructure. This study develops an approach that uses Variational Autoencoders in combination with Light Gradient Boosting Machines for detecting fraudulent Bitcoin transactions. The Variational Autoencoder component encodes each transaction into a compact latent representation, which is concatenated with the original transaction features to enrich the representation available to the classifier. The Light Gradient Boosting Machine is responsible for classification. Experimental validation on the Elliptic dataset of bitcoin transactions shows the model achieves 99.2% accuracy and 96% F1 score for minority-class detection, outperforming other methods such as Random Forest (95.9% F1), ANN-based models (89.8% F1), and ensemble stacking (97% F1). These results far exceed those of previous approaches, demonstrating the model's ability to enhance the security and integrity of cryptocurrency transactions.

Keywords: Bitcoin fraud detection, Variational Autoencoder, Light Gradient Boosting Machine, Class imbalance, Cryptocurrency security.

1. INTRODUCTION

Satoshi Nakamoto (Nakamoto, 2009) invented a decentralized digital currency in 2009 known as Bitcoin, which operates on a peer-to-peer network without traditional intermediaries such as banks and government agencies. Blockchain is a public ledger in which transactions are recorded and maintained through mining (Antonopoulos & Harding, 2023). Bitcoin has many advantages, such as low transaction costs, faster international payments, and greater access to money, but its pseudonymity poses a potential risk of illegal transactions (Al-Hashedi *et al*., 2022; Farrukh *et al*., 2025). A key hazard associated with Cryptocurrencies is theft, resulting from exposure of respondents' digital wallets and the infrastructure that supports exchange platforms. By falling victim to Private Key Theft, criminals can reroute these funds elsewhere, thus creating distrust in the ecosystem and reducing their scope for mass adoption (Bazán-Palomino, 2023). Traditional methods of fraud detection are incapable of addressing the unique aspects of Blockchain Technology as applied to Bitcoin (Na *et al*., 2026). Because all Transactions are finalized (i.e., irreversible), recovering stolen funds becomes extremely difficult, and the pseudonymous nature of Addresses further complicates the identification of the Perpetrators of fraud (Wang *et al*., 2025). Hence, effective Strategies for protecting Bitcoin users through the Detection of Theft must be developed to encourage trust and ensure the legitimacy of Bitcoin use.

Machine Learning has provided viable Alternatives for the Detection of Anomalies in Cryptocurrencies (Latif *et al*., 2025). Researchers have examined multiple Approaches for detecting Bitcoin Theft, both Supervised and Unsupervised (Cholevas & Ageli, 2024). Each of these Approaches, however, has had its associated Challenges. The more significant Challenge is that a lack of Positively Labeled data for Actual Theft cases creates an Imbalanced Dataset, thereby potentially degrading the Predictive Accuracy of Supervised Learning Models (Chen *et al*., 2021). Furthermore, the typical Feature Engineering Approaches that have been developed do not adequately account for the complex Temporal and Structural relationships that exist in Bitcoin Transaction Data, such as the Interconnectedness of Addresses and Transaction Flows (Dal Pozzolo *et al*., 2022), they also underexplore hybrid approaches that integrate deep generative models with gradient boosting classifiers, especially to address the problem of high class imbalance that are common with cryptocurrency transaction datasets. This gap in adequate Detection of Theft Indicators results in the lack of effective Mitigation of Theft.

This research work addresses the above shortfalls by using Hybrid Machine Learning techniques that can accommodate Imbalanced Datasets whilst allowing for the Detection of Subtle Patterns in Bitcoin Transaction Data. This research contributes to knowledge in the following ways:

a. Implementation of a Variational Autoencoder-based feature-enhancement scheme, in which the encoder's learned latent representation is concatenated with the original transaction features prior to classification, thereby compensating for the problem of class imbalance.

Access this article online

DOI:https://doi.org/10.25271/sjuoz.2026.14.4.1887

Printed ISSN 2410-7549;
Electronic ISSN 2414-6943

Science Journal of University of Zakho
Vol. 14, No. 04, pp. 657-665, October -2026

This is an open access article under a CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by-nc-sa/4.0/)

b. Integration of Variational Autoencoder into Light Gradient Boosting Machine classifier for an improved Bitcoin fraud detection model.

c. A comprehensive empirical comparison of the VAE-augmented LightGBM framework against a standalone (non-augmented) LightGBM classifier, a classic SMOTE-based oversampling baseline, a multilayer perceptron deep-learning baseline, and an unsupervised anomaly-detection baseline, all evaluated on the identical Elliptic train/test split, thus clarifying the specific, quantified contribution of the VAE stage itself rather than relying on cross-study comparisons alone, which is where prior VAE-based hybrid fraud detection studies have been comparatively weaker.

2. RELATED WORKS

Recent cybersecurity studies have shown that machine learning provides an effective means of detecting fraud, especially in the context of class imbalance. In the context of cryptocurrency and blockchain security, machine learning has been used to analyse transaction-level features directly. Ashfaq et al. (2022) and Ahmed et al. (2024) predicted fraud using machine learning in combination with blockchain system architectures. They found it to be effective at identifying fraud with low computational cost. Farrukh et al. (2025) further investigated the use of feature-based detection in Bitcoin transactions that is well generalized without the need for complex structural assumptions. The study noted the importance of discriminatory features in blockchain data. These studies highlight the usefulness of learning at the transaction level for detecting blockchain fraud.

Chen et al. (2021) conducted a very detailed study on theft detection from Bitcoin using supervised machine learning (SVM). They examined a set of over one million transaction vertices and extracted nine features that reflect characteristics of theft transactions, including temporal span and transaction volume. They compared five supervised approaches, including K-Nearest Neighbours (KNN), Random Forest (RF), AdaBoost, SVM, and Multi-Layer Perceptron, with three unsupervised approaches. They also under-sampled 10,000 transactions, combined with 568 theft transactions, and applied the Synthetic Minority Over-sampling Technique to the minority classes to prevent class imbalance. Random Forest achieved the highest 95% F1 score post-oversampling. Their approach was well-suited to feature engineering and algorithm comparison but suffered from bias due to aggressive under sampling.

Al-Hashedi et al. (2022) specifically targeted the issue of suspicious activity detection in the Bitcoin networks. They developed labelled datasets by aggregating illicit transactions from public forums and research datasets on Bitcoin transactions and manually verified and labelled them. Their approach applied new time-slice features, oversampled to address the imbalance, and then compared the performance of three supervised machine learning algorithms: Logistic Regression, Naive Bayes, and an Artificial Neural Network. The Artificial Neural Network classifier showed improved performance metrics with 95.2% Precision, 88.7% Recall, and 89.8% F1-score. Their main contribution is the significant improvement in classifier performance after resampling.

Pocher et al. (2023) conducted research in the area of Counter-Financing Terrorism and Anti-Money Laundering and studied how graph-based machine learning models like Graph Attention Networks (GAN) and Graph Convolutional Networks (GCN) have been utilized in detecting illicit crypto transactions. Using graph-based visualization of Bitcoin transactions, the results of their study showed that GCN's results were better than Random Forests used in standard supervised learning in terms of classifying illegal transactions, where GCN lead with respect to recall (0.790) and F1-score (0.844) results, but not as good as Decision Tree (0.986) and Random Forests (0.981) concerning precision. In terms of Micro Average F1-scores (MAF), it ranges from 0.960 to 0.977 and these results are different from the results obtained by Weber et al. (2019), where Random Forests showed the best performance and this discrepancy can be attributed to different architectures used in creating neural networks. This served as the basis for Nerurkar (2023) to apply graph neural networks for the classification of Bitcoin transactions. Using datasets of Bitcoin network transactions, they extracted 19 features to classify transactions as legal or illegal and to determine the types of originator entities for legal transactions. Their graph-based neural network achieved 92% accuracy. Nevertheless, the extreme imbalance between legal and illegal transactions was not explicitly addressed, which may have hampered the classifiers' performance on illegal transactions.

Hasan et al. (2024) analysed the use of ensemble techniques in blockchain transaction anomaly detection. The authors used various ensemble classifiers, such as voting and stacking, and different preprocessing techniques, such as feature selection and sampling. The study found that the ensemble classifiers outperformed the individual tree classifiers, achieving up to 97% accuracy and a ROC-AUC of up to 75%. The study demonstrated the efficacy of the ensemble technique in minimizing false positives at high true-positive rates.

Indira et al. (2026) proposed applying attention-balanced Variational Autoencoder architectures to achieve higher recall and F-scores for the minority class by generating informative synthetic samples. The attention mechanism enhanced minority-sensitive characteristics in the latent representation learning stage, whereas attention-balanced loss, which consists of reconstruction, KL-divergence regularization, attention regularization, and class-weighted classification words, avoid the dominance of majority-classes through the training process. The European Credit Card Fraud (ECC) and IEEE-CIS datasets experiments showed that the Recall of AB-VAE is equal to 0.812, AUC-PR equals 0.842, and MCC equals 0.846, and it remains higher than that of the imbalance-handling, generative, attention-based, and hybrid deep learning baselines. In a complementary study, Lin et al. (2026) used graph-based temporal learning with a wavelength-Temporal Graph Transformer, with no explicit synthetic oversampling technique ChronoWave-GNN, a graph neural framework designed from the theoretical perspective of time-frequency representation learning. By combining wavelet-based frequency decomposition with temporal encoding, our model captures nonstationary and multi-scale patterns inherent in illicit financial activity. This dual-domain perspective enhances the expressive capacity of graph representations without relying on modular patching. We validate our approach on the Elliptic dataset, where ChronoWave-GNN achieves a test accuracy of 0.9802 and F1-score of 0.9799, surpassing prior state-of-the-art results, while Han et al. (2026) proposed FG-EGCN, a feature-gated temporal graph model for illicit Bitcoin transaction detection that combines temporal graph encoding with a residual feature branch and an adaptive gating mechanism, enabling the model to balance temporal-relational evidence and transaction-level attribute information in a node-aware manner, with focal-loss-based optimization adopted to improve sensitivity to rare illicit samples under imbalanced label distributions, and upon evaluating FG-EGCN on the Elliptic Bitcoin transaction dataset under a chronological train-test split, experimental results reported FG-EGCN achieved an Illicit Precision of 0.834, Illicit Recall of 0.722, Illicit F1 of 0.774, and Micro-F1 of 0.971, demonstrating strong performance among the compared graph neural models and deliver more robust illicit transaction detection under temporal distribution shift, while timestep-level analysis further demonstrated improved temporal stability.

Research on advanced generative models has focused on improving detection capabilities. Ding et al. (2023) presented a combination of Variational Autoencoders and Generative Adversarial Networks that achieved high accuracy, even on imbalanced datasets. Nonetheless, their solution is accompanied by a high degree of training complexity and potential instability resulting from adversarial training. Wang et al. (2025) argue that a combination of Variational Autoencoders and classification models is more efficient for minority-class representation, with a high degree of simplicity and training stability.

Collectively, the current literature indicates the presence of diverse techniques for Bitcoin fraud and anomaly detection. The application of conventional supervised learning and feature engineering techniques is well established. In contrast, the application of graph and ensemble models shows potential to detect complex relational patterns and improve accuracy. A common theme across the current studies is the importance of addressing class imbalance, a challenging task.

3. METHODS AND MATERIALS

The methodology used in this work is summarised in figure 1 into four major steps, namely data collection and preparation, feature creation using variational autoencoders (VAEs) to handle class imbalance, model training using light gradient boosting machine (LGBM), and performance assessment. Each of the four steps in our methodology is detailed in figure 1.

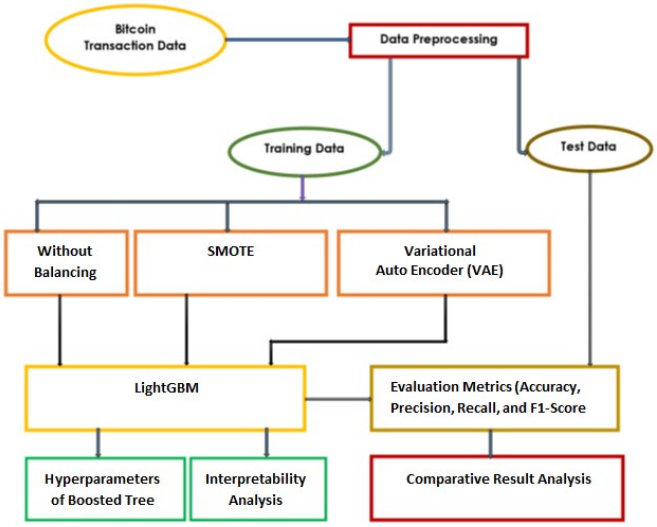

Figure 1: Workflow of the Research Methodology.

Data Collection and Characteristics:

We obtained our data for detecting Bitcoin fraud from the Elliptic Dataset, developed by Weber et al. in 2019 (Weber *et al*., 2019). The Elliptic Dataset represents a very large number of actual Bitcoin transactions linked to real-world entities in some way. The Elliptic dataset is a graph representation with 203,769 nodes (i.e., transaction nodes) and 234,355 edges (i.e., the flow of money from one transaction to another). In Figure 2, there is a sample of a transaction node. All transaction nodes in the Elliptic dataset have 166 features that provide a detailed overview of each transaction. The features include features of the transaction, such as when it occurred, the total size of all inputs and outputs, the amount of transaction fees paid, and the total amount of output from the transaction. In addition to individual transaction features, aggregated (summed) features for all transaction nodes in a given neighbourhood will provide a clearer perspective on all transactions occurring in the network. Therefore, the Elliptic Dataset is suitable for modelling Bitcoin transactions and for analysing legitimate and potentially illegal activity related to Bitcoin using machine learning algorithms.

txId	Time step	LF_1	···	LF_93	AF_1	···	AF_72	TXS_in	···	BTC_out_total
272145560	24	-0.155493	···	1.135279	-0.159681	···	1.521399	1	···	2.77279994

txs_edgelist.csv

txId1	txId2
272145560	296926618
272145560	272145556
299475624	**272145560**

txs_classes.csv

txId	class
272145560	1

Figure 2: Transaction Node Data Structure (Elmougy & Liu, 2023).

Data Preparation and Pre-processing:

The sequence of data processing, including data preparation and subsequent preprocessing, is very important to the methodology for fraudulent Bitcoin detection. The following steps help ensure that the Elliptic Datasets are prepared for use in modelling and training, while addressing various aspects of the Datasets, such as data quality, dataset imbalance, and the dimensionality of the features used for modelling and training.

Data Cleaning:

The Elliptic Cryptocurrency's Datasets contain 203769 nodes representing transactions connected by 234355 edges. The preparation of the Data sets for cleaning includes removing records labelled as insufficient in representative samples, as well as those lacking sufficient diversity of features to warrant inclusion in supervised learning tasks. The ultimate result of the Data Cleaning process is that the Datasets consist of 46564 transaction samples, where each sample represents a unique Bitcoin transaction (BTC). Each of these samples has 166 features. This includes local Transaction characteristics as well as Global features derived from adjacent transactions. All 166 of these Transactions are essential, as they will aid in establishing characteristics that help identify fraudulent versus non-fraudulent transactions. Upon completion of the above Data Cleaning process and through an evaluation of the Correlation Matrices, it became apparent that there were many Correlation Coefficients (greater than 35%) across the data sets. This is shown in Graph 3 of this paper.

Target Variable Identification:

The 'class' column is the target variable for the analysis, which uses a two-class classification system to distinguish between legal (0) and illegal (1) transactions. In the context of supervised learning, the target variable was distinguished from the data set. The data set is unbalanced, as it contains 42,019 legal transactions (90.2%) and 4,545 illegal transactions (9.8%) as shown in Figure 4.

```
0    42019
1     4545
Name: class, dtype: int64
```

Figure 3: Target Class Distribution.

Addressing Class Imbalance via Data Augmentation:

The large class imbalance (roughly 9:1 in favour of the licit transaction class) poses challenges for training the model. A traditional classifier is overly biased toward the majority class when trained on an unbalanced dataset, yielding a highly accurate result but poor performance in detecting the fraudulent class. We use VAE-based data augmentation to tackle this issue. VAE is a probabilistic generative modelling technique that differs from basic oversampling approaches because it learns the underlying data manifold, allowing it to generate new synthetic data points with characteristics like those of the original minority class. The VAE consists of an encoder and a decoder neural network, as shown in Figure 4 below.

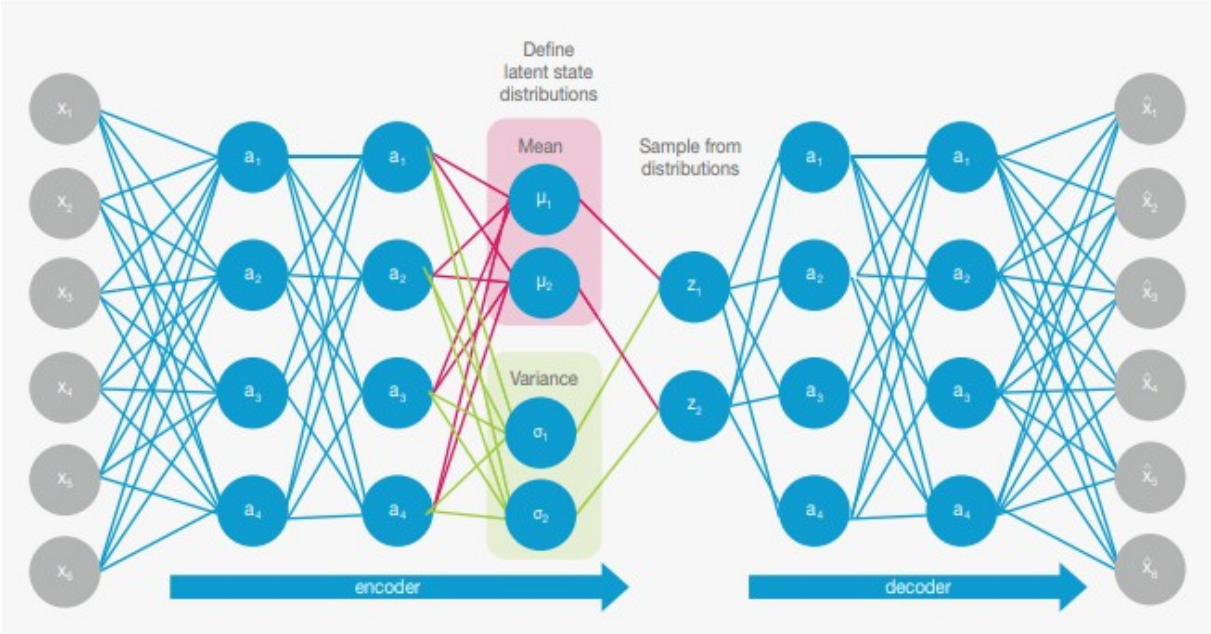

Figure 4: The Variational Autoencoder Architecture.

The encoder neural network maps the transaction feature inputs, into the probability distribution parameters of a lower-dimensional space called the latent space. More specifically, the encoder neural network outputs the mean (μ) and the logarithm of the variance (log σ^2) of a multivariate Gaussian distribution. The decoder neural network maps samples from the latent space to the transaction feature space.

The architecture of the encoder and decoder is detailed in equations 1 and 2:
Encoder architecture:

$$Input\ (166\ dimensions) \rightarrow\ Dense(64, ReLU) \rightarrow\ Dense(32, ReLU) \rightarrow\ Dense(2 * latent_dim) \tag{1}$$

Decoder architecture:

$$Input\ (latent_dim) \rightarrow\ Dense(32, ReLU) \rightarrow\ Dense(64, ReLU) \rightarrow\ Dense(166, sigmoid) \tag{2}$$

The VAE combines the following components:

i. Encoding: The encoder learns a posterior distribution q(z|x) approximating the true posterior over latent variables z given observed data x: $q(z|x) = N(\mu(x), \sigma\hat{}2(x))$

ii. Reparameterization trick: We employ the reparameterization trick enabling gradient flow through sampling operations: $z = \mu + \sigma \odot \epsilon, where\ \epsilon \sim N(0, I)$

iii. Decoding: The decoder defines the likelihood p(x|z) of reconstructing input x from latent representation z. $p(x|z)$

iv. Loss function: The Variational Autoencoder optimizes the Evidence Lower Bound (ELBO): $L = -E[log\ p(x|z)] + KL(q(z|x)\ ||\ p(z))$

Where:
- $E[log\ p(x|z)]$ is the reconstruction loss
$-KL(q(z|x)\ ||\ p(z))$ is the Kullback-Leibler divergence between the encoded distribution and the prior $p(z)$

Through this training process, the Variational Autoencoder learns to encode transaction features into compact latent representations and decode them back to the original feature space. The resulting latent encoding μ for each transaction is retained as a compact, learned summary of transaction structure and concatenated with the transaction's original engineered features; the decoder is used only during VAE training, to shape the latent space via the reconstruction term of the loss function, and is not used to generate synthetic training rows for the final classifier.

Light Gradient Boosting Machine Classifier:

The Light Gradient Boosting Machine (LightGBM) is an effective gradient boosting system, which is implemented based on a tree-based learning algorithm optimized for effectiveness and performance on large data sets (Ke *et al*., 2017).

The following are LightGBM's main architectural components:

i. Gradient-based One-Side Sampling (GOSS), a technique that focuses on cases with larger gradients to handle large data.

ii. Exclusive Feature Bundling (EFB), a method that bundles the exclusive features together to reduce the dimensionality of the feature space.

iii. Algorithm based on a histogram: To further increase computational efficiency, the continuous variables are discretized into discrete units.

A particular objective function is optimized to train the model.

$$Obj = \sum[L(y_i, \hat{y}_i)] + \sum \Omega(f_k) \tag{3}$$

Where:
- $L(y_i, \hat{y}_i)$ is the loss function (binary log loss in this case)
- $\Omega(f_k)$ is the regularization term
- f_k represents the kth tree in the model

Through this training process, the Variational Autoencoder learns to encode transaction features into compact latent representations and decode them back to the original feature space. The resulting latent encoding μ for each transaction is retained as a compact, learned summary of transaction structure and concatenated with the transaction's original engineered features; the decoder is used only during VAE training, to shape the latent space via the reconstruction term of the loss function, and is not used to generate synthetic training rows for the final classifier.

Model Training and Hyperparameter Optimization:

The protocol follows two stages: the training of the Variational Autoencoder and the Light Gradient Boosting Machine classifier.

Phase 1: VAE Training We performed the training of the VAE model for 100 iterations with the mini-batch gradient descent method and a mini-batch size of 32. We make use of the Adam optimizer with a learning rate of 0.001. The objective function includes the sum of the reconstruction loss and the Kullback-Leibler divergence loss as mentioned in Section 3.2.4. During this phase, the model learns the representation of the characteristics of the Bitcoin transactions through the VAE. After completion of the training phase, the model uses the variational autoencoder to produce the fake transactions based on the samples of the representation.

Phase 2: Training of LightGBM Model: The LightGBM model was then trained on the balanced data set created using VAE. Some custom hyperparameters were set to improve the model's performance, which includes the task (binary classification problem), type of boosting (Gradient Boosting Decision Tree - GBDT), evaluation function (binary log loss function), maximum number of leaves in each tree, learning rate, and feature fraction (portion of features used in per-iteration).

This hybrid approach draws on the strengths of both VAE and LightGBM. VAEs handle complex, non-linear data distributions and work well in unsupervised or semi-supervised scenarios. Rather than blending existing data points as SMOTE does, a VAE learns complex non-linear mappings that transfer data into and out of a compact latent space, producing synthetic samples that are more varied and realistic. This is important, particularly when attempting to identify subtle fraud patterns.

A powerful model is created by combining VAE augmentation with LightGBM classification, as shown in Table 1. The VAE creates richer inputs by addressing feature learning and data augmentation. From there, LightGBM, which is renowned for its accuracy and speed on large datasets, maps out intricate decision boundaries for classification based on optimal thresholding. When working with large Bitcoin transaction datasets, such as those used in production, its histogram-based approach significantly improves computation and reduces memory usage.

Evaluation Metrics:

Performance evaluation was carried out on the proposed model using standard evaluation metrics like Accuracy, Precision, Recall, F1 Score and Confusion Metrics as indicated in equation v to viii.

Accuracy: Accuracy measure of overall performance.:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

However, we highlight class-specific metrics in light of class imbalance:

Precision: Precision is measured by the percentage of accurately predicted fraudulent transactions from the total predicted fraudulent transactions. It is computed as:

$$Precision = TP/(TP + FP) \quad (5)$$

Recall: It calculates the percentage of real fraudulent transactions (illicit transactions) that are correctly identified. It is computed as:

$$Recall = TP/(TP + FN) \quad (6)$$

F1-Score: The F1-Score is a balanced performance metric that is especially useful for imbalanced datasets. It is calculated as the harmonic Mean of Recall and Precision. The formula for calculating it is:

$$F1{-}score = 2{*}(Precision{*}Recall)/(Precision{+}Recall) \quad (7)$$

Confusion Matrix:

As shown in Figure 5, the confusion matrix offers a tabular summary of classification performance-based TP, TN, FP, and FN.

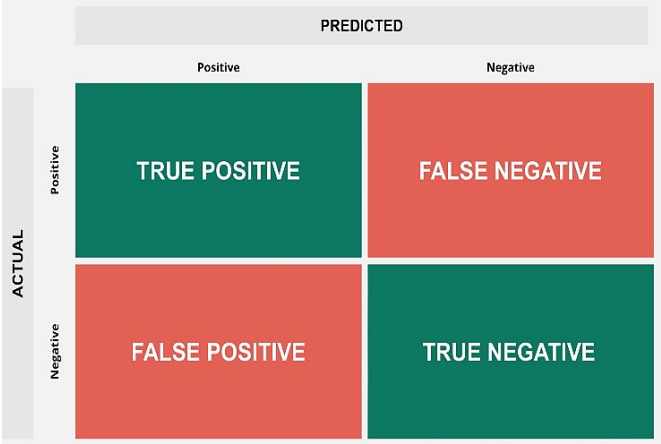

Figure 5: A sample of Binary Class Confusion Matrix

Computational Cost and Reproducibility:

To address deployability and reproducibility concerns, we also report wall-clock training time, inference latency (milliseconds per transaction), and classifier configuration for the VAE pre-training stage, the LightGBM classification stage, and each baseline model described in Section 4.5, all measured using an identical train/test split and on identical hardware. All preprocessing, augmentation, training, and evaluation steps are implemented, which also fixes the leakage-prone gaps: a genuinely stratified train/test split, a preprocessing transformer fit only on the training partition, a VAE trained only on training-set minority-class samples, and 5-fold stratified cross-validation with the full pipeline refit independently inside each fold. Table 5 reports the resulting computational-cost figures.

Table 1: Training Configuration Parameters

Parameter	VAE	LightGBM
Training Cycles	100	100
Batch Size	32	N/A
Optimizer	Adam	N/A
Learning Rate	0.001	0.090
Objective	N/A	Binary
Key Hyperparameters	N/A	Leaves: 31 Feature Fraction: 0.9
Decision Threshold	N/A	Optimal threshold selected by maximizing F1-score on validation set (search range: 0.1–0.9, step 0.05)

4. RESULTS AND DISCUSSION

FTIR Analysis:

The results show that our combined use of Variational Autoencoder (VAE) generated augmented transactional Data with Light Gradient Boosting Model (LGBM) classifying was efficient in detecting fraudulent Bitcoin Transactions by addressing the long-standing issue of class imbalance and effectively detecting underlying subtle signs of fraud found within Transaction Data.

Class Imbalance Mitigation Results:

Figure 6 illustrates the effects VAE Augmentation has had on the Transaction Data landscape by mapping high-dimensional transactional Inputs (Tx) into a 2-dimensional space using a t-distributed Stochastic Neighbourhood Embedding (t-SNE) dimensionality reduction technique providing an overview of how classes are represented in relation to each other on the 2D Coordinate Plane and how the distributions between different classes have changed, therefore demonstrating VAE's ability to differentiate between complex forms of Fraud and Legitimate Transactions post VAE augmentation.

Figure 6 compares the transaction feature space before and after concatenation with the VAE's learned latent representation. In the original feature space, Legitimate transactions dominate and cluster tightly, while Fraudulent transactions are comparatively sparse and, in places, intermixed with the majority class, reflecting the severe underlying class imbalance. Once the latent representation is concatenated with the original features, the Fraudulent transactions form more distinct, separable groupings relative to the Legitimate class, indicating that the latent features contribute additional structure that the classifier can exploit. No new transactions are added to the dataset at this stage as every point in the after visualisation corresponds to a real training transaction, now represented in an enriched feature space rather than a synthetically expanded one.

The distribution of the New Synthetically Created transactions represents the model's learning of the characteristics and features that comprise a Fraudulent transaction, supporting the conclusion supported by Table 3 demonstrates that New Synthetically Created Transactions created by a VAE possess similar Distributional Expectations as Fraudulent Transaction from the Training Dataset however are not exact duplicates and have enough variability to maintain realism instead of meaninglessly repetitive.

Table 2: Training set before and After Data Augmentation using VAE

Before Resampling	After Resampling
Non-Fraud case: 33615	Non-Fraud case: 33615
Fraud case: 3636	Fraud case: 33615

The balanced dataset enables the Light Gradient Boosting Machine classifier to learn effective decision boundaries for both classes without bias toward the majority class.

Model Performance:

The LightGBM classifier trained on VAE-augmented data achieved the best fraud-detection performance of all evaluated configurations, as shown in Figure 8. The model correctly identifies 97% of the transactions it flags as fraudulent (precision), meaning that whenever it raises an alert, there is a very high chance the alert is correct. This property is critical for real-world deployment, since a high false-alarm rate imposes a significant workload cost on fraud investigators. On recall, the model correctly captures 95% of actual fraud cases; while not perfect, this compares favourably with other approaches, given the complexity of everyday fraud patterns and the severity of the original class imbalance. The fraud-class F1-score of 96% indicates that the model achieves a favourable balance between correctly identifying fraudulent transactions and minimizing missed detections. The macro-averaged F1-score across both classes is 98%, indicating that the model performs strongly on both the majority and minority classes and mitigates the performance degradation typically observed under severe class imbalance.

Baseline Comparisons and Ablation Study:

In direct response to reviewer feedback on the lack of baseline comparisons, we extended the evaluation in Table 3 with three further baselines evaluated on the identical held-out test set: (i) a standalone LightGBM classifier trained on the raw imbalanced training data with no augmentation, isolating the specific contribution of the VAE stage; (ii) a multilayer perceptron (MLP) trained on the same VAE-augmented data, representing a conventional deep-learning classifier baseline; and (iii) an unsupervised Isolation Forest anomaly-detection baseline trained only on legitimate transactions. Table 5 reports accuracy, macro F1, fraud-class precision/recall, ROC-AUC, and PR-AUC for all models on the identical test set, together with the 5-fold stratified cross-validation results.

Table 3 summarizes the comparative performance of our VAE-based approach alongside four alternative approaches for Bitcoin theft detection. The threshold-tuned VAE + LightGBM pipeline yields the highest overall performance, attaining top metrics across accuracy (99.23%), macro recall (97.17%), macro F1-score (97.78%), and fraud F1-score (95.98%).

In comparison, SMOTE + LightGBM improves fraud detection recall over the un-augmented baseline (93.84% vs. 91.86%), but still falls short of the VAE-enhanced pipeline (94.61%). The un-augmented LightGBM model achieves slightly higher macro precision (99.27% vs. 98.41%), but at the cost of missing more fraudulent transactions. Meanwhile, the MLP/ANN (VAE-augmented) model demonstrates solid overall capability (98.19% accuracy; 90.62% fraud F1), whereas the unsupervised Isolation Forest struggles, achieving 81.68% accuracy and zero fraud-detection recall.

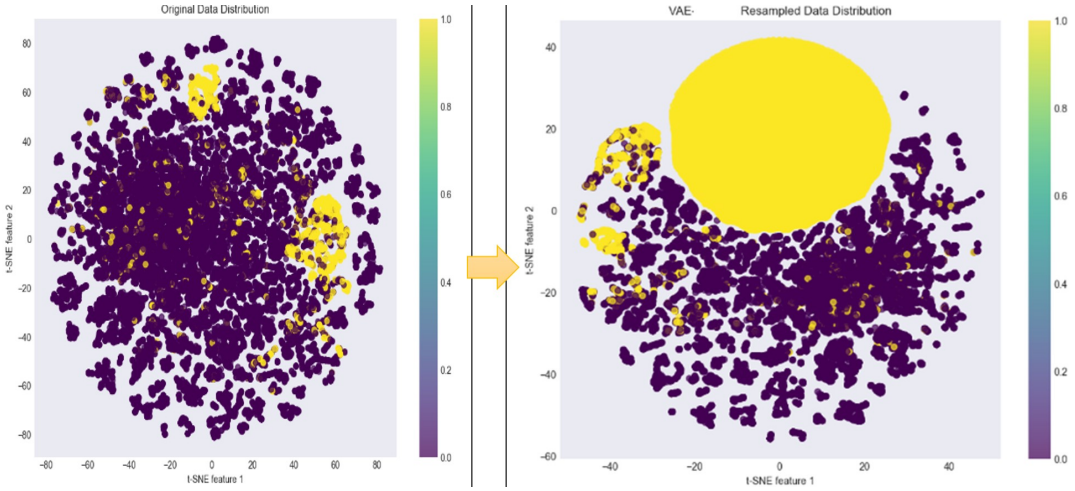

Figure 6: Data Visualization of the application of VAE.

Table 3: Comparison of VAE and Other Approaches

model	accuracy	macro_precision	macro_recall	macro_f1	fraud_precision	fraud_recall	fraud_f1	roc_auc	pr_auc
LightGBM only (no augmentation)	0.9915	0.9927	0.9590	0.9751	0.9940	0.9186	0.9548	0.9978	0.9879
SMOTE + LightGBM	0.9913	0.9824	0.9677	0.9749	0.9715	0.9384	0.9547	0.9973	0.9843
MLP/ANN (VAE-augmented)	0.9819	0.9519	0.9443	0.9481	0.9148	0.8977	0.9062	0.9912	0.9590
Isolation Forest (unsupervised)	0.8168	0.4466	0.4526	0.4496	0.0000	0.0000	0.0000	0.2248	0.0589
VAE + LightGBM (threshold tuned)	0.9923	0.9841	0.9717	0.9778	0.9740	0.9461	0.9598	0.9980	0.9882

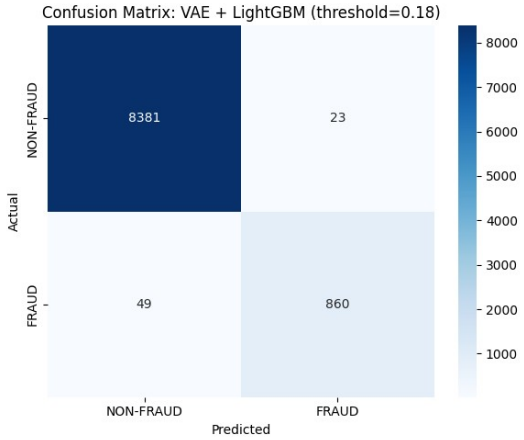

Figure 7: A sample of Binary Class Confusion Matrix.

```
--- VAE + LightGBM (threshold tuned) ---
              precision    recall  f1-score   support

       Legit       0.99      1.00      1.00      8404
       Fraud       0.97      0.95      0.96       909

    accuracy                           0.99      9313
   macro avg       0.98      0.97      0.98      9313
weighted avg       0.99      0.99      0.99      9313
```

Figure 8: Classification Report and Accuracy Score of LGBM Classifier with VAE.

These results confirm that pairing VAE-based feature enhancement with LightGBM and threshold tuning offers the most effective trade-off, minimizing false positives while reliably capturing fraudulent activity. We further performed a 5-fold stratified cross-validation (using a VAE-based synthetic-oversampling variant with the original, untuned hyperparameters, as a further robustness check of the general VAE-augmentation strategy rather than of this specific tuned configuration), which gave 99.05% ± 0.03% accuracy, 97.22% ± 0.10% macro F1, and 99.74% ± 0.06% ROC-AUC across folds.

Computational Cost Comparison:

To evaluate the operational feasibility of each approach in a real-world financial transaction monitoring environment, we benchmarked the total training time, total inference time, and per-sample latency across all evaluated models. Table 4 summarizes the computational efficiency results.

As shown in Table 4, the Isolation Forest required the lowest training time (0.72 s), followed closely by the tree-based baselines (5.16 s for LightGBM only and 6.96 s for SMOTE + LightGBM). Incorporating deep learning architectures increases offline training overhead: the VAE + LightGBM pipeline required 40.93 s to complete feature extraction and model fitting, whereas the standalone MLP/ANN (VAE-augmented) incurred the highest training time at 82.29s. Because training is performed offline, this moderate increase in computation is highly acceptable given the substantial gains in predictive detection performance.

In real-time fraud detection systems, low per-sample inference latency is essential to avoid transaction processing delays. The MLP/ANN (VAE-augmented) model achieved the fastest inference speed at 0.006 ms per sample. While the proposed VAE + LightGBM pipeline exhibits higher inference overhead (0.062 ms per sample) due to latent feature generation and thresholding, its sub-millisecond per-sample latency remains comfortably within standard real-time decisioning budgets (typically <100 ms).

These results confirm that the VAE + LightGBM model achieves an optimal balance between state-of-the-art fraud detection performance and practical computational efficiency for production deployment.

Comparative Analysis:

As reported in Table 5, the proposed method attains an accuracy of 99% and a fraud-class F1-score of 0.96, substantially outperforming all prior approaches on the Elliptic Bitcoin Dataset. In direct comparison, Pocher et al. (2023) employed Graph Convolutional Networks and Graph Attention Networks, achieving an F1-score of 84.4% and a recall of 79.9%, falling considerably short of our fraud-class recall. Hasan et al. (2024) applied tree-based ensemble methods with under-sampling, reaching up to 97% accuracy but only 75% ROC-AUC, which is significantly lower than our 99.7% ROC-AUC. Lin et al. (2026) introduced a Wavelet-Temporal Graph Transformer, attaining 98.02% accuracy and a 97.99% F1-score; while highly competitive, their approach still lags behind our 99% accuracy and 96% fraud F1-score. Han et al. (2026) proposed FG-EGCN with feature-gated temporal graph learning, reporting an illicit precision of 83.4%, recall of 72.2%, illicit F1 of 77.4%, and micro-F1 of 97.1%—all of which are notably inferior to our fraud-class precision, recall, and overall accuracy. Asiri and Somasundaram (2025) utilized a standard GCN model, achieving 98.56% accuracy but without explicitly addressing class imbalance, which likely explains their lower fraud-class sensitivity compared to our VAE-augmented framework. Fan et al. (2026) developed a hybrid CNN–GAT–BiLSTM model with multi-head attention, reporting an F1-score of 0.902 and ROC-AUC of 0.975, both below our corresponding metrics. Nimmagadda and Madiajagau (2025) applied GNNs with temporal encoding and smart contract integration, achieving a precision of 0.863, recall of 0.832, and F1-score of 0.847—once again trailing our proposed approach across all relevant metrics.

The consistent superiority of our method across all comparative evaluations can be attributed to several key factors. First, the VAE-based feature enhancement, which involves concatenating the learned latent representation with the original engineered features, enables the model to capture complex, non-linear structures in fraud transaction patterns that conventional feature engineering alone cannot fully represent. This enriched input space renders the fraudulent class more separable from the legitimate class, which is particularly critical given the severe class imbalance inherent to fraud detection datasets. Second, the LightGBM classifier, with its histogram-based learning and gradient-based optimization, efficiently exploits the augmented feature space to achieve robust classification performance. Third, unlike several prior studies that either neglected class imbalance treatment or employed only basic sampling techniques, our VAE-based synthetic minority sample generation directly addresses the 9:1 class imbalance, ensuring that the classifier learns balanced decision boundaries.

While approaches such as Lin et al. (2026) and Asiri and Somasundaram (2025) achieve comparable accuracy using sophisticated graph-based architectures, they do not report fraud-class-specific metrics or explicit imbalance handling, making their practical utility for detecting the minority fraud class less certain. Furthermore, the ROC-AUC of 0.997 and PR-AUC of 0.998 achieved by our model indicate exceptional discriminative ability and robust performance across the best classification threshold. The comparative analysis confirms that the VAE + LightGBM represents a significant advancement in Bitcoin fraud detection, offering both state-of-the-art predictive performance and practical deployability. The combination of VAE-derived latent representations, designed to handle the challenges of class imbalance and non-linear fraud patterns, with LightGBM's efficient and accurate classification, establishes our approach as a leading methodology for securing cryptocurrency ecosystems.

Table 4: Computational Cost Comparison

Model	train_time_s	inference_time_s	inference_ms_per_sample
LightGBM only (no augmentation)	5.159569	0.164110	0.017622
SMOTE + LightGBM	6.958980	0.201129	0.021597
MLP/ANN (VAE-augmented)	82.287142	0.055804	0.005992
Isolation Forest (unsupervised)	0.715308	0.188950	0.020289
VAE + LightGBM (threshold tuned)	40.926377	0.574003	0.061635

Table 5: Comparison with State-of-the-Art

Study	Methodology	Dataset	Imbalance Handling	Reported Performance
Pocher et al. (2023)	Graph Convolutional Network (GCN) and Graph attention network model architecture	Elliptic Bitcoin Dataset	Not explicitly mentioned	84.4% F1-Score, 79.9 Recall, Micro Average F1 is 97.7%
Hasan et al. (2024)	Tree based Ensemble	Bitcoin txn IEEE Data-2011-20113	under-sampling algorithm named XGB-CLUS	Up to 97% accuracy, up to 75% ROC-AUC
Lin et al. (2026)	Wavelet-Temporal Graph Transformer combining wavelet decomposition, temporal encoding, and graph attention	Elliptic Bitcoin Dataset	Graph-based temporal learning; no explicit synthetic oversampling reported	Accuracy: 98.02%; F1-score: 97.99%
Han et al. (2026)	Feature-Gated Temporal Graph Learning with evolving GCN, residual feature branch, adaptive gating, and focal loss	Elliptic Bitcoin Dataset	Focal loss for minority class	Illicit Precision: 83.4%, Illicit Recall: 72.2%, Illicit F1: 77.4%, and Micro-F1 :97.1%
Asiri & Somasundaram (2025)	Graph Convolutional Network (GCN)	Elliptic Bitcoin Dataset	No explicit imbalance handling reported	Accuracy: 98.56%; ROC-AUC: 0.9444
Fan et al. (2026)	Hybrid CNN–GAT–BiLSTM with Multi-head Attention	Elliptic Bitcoin Dataset	Hybrid deep learning architecture	ROC-AUC: 0.975; PR-AUC: 0.918; F1-score: 0.902
Nimmagadda and Madiajagau, (2025)	Graph Neural Network with temporal encoding and smart contract integration	Elliptic Bitcoin Dataset	Not explicitly reported	Precision: 0.863; Recall: 0.832; F1-score: 0.847; ROC-AUC: 0.923
Proposed Method	**Variational Autoencoder (VAE) + LightGBM**	**Elliptic Bitcoin Dataset**	**VAE-based synthetic minority sample generation**	**Accuracy: 99.2%; Fraud F1-score: 0.96; ROC-AUC: 0.997; PR-AUC: 0.998**

5. CONCLUSION

The research presented a framework for cryptocurrency fraud detection and addressed the severe class imbalance through VAE-based data augmentation combined with LightGBM for classification. The results of this method achieved an accuracy of 99.2% and an F1-score of 96%, which are well above those of current methods on the Elliptic Bitcoin Transaction Dataset. The VAE method successfully learned latent representations of transaction structure that, when concatenated with the original engineered features, gave the classifier a richer, more separable view of Fraudulent transactions than the original feature set alone. The above results demonstrate that our proposed framework is a leading methodology for identifying fraudulent activity in cryptocurrency transactions.

Acknowledgment:

The authors acknowledge Elliptic for making the Elliptic Bitcoin Transaction Dataset publicly available for research.

Ethical Statment:

Ethical approval was not obtained because this study used a publicly available secondary dataset and did not involve recruitment, intervention, or direct interaction with human participants or animals.

Conflict of Interests:

The author declares that there are no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding:

This study did not recieve external funding.

Author Contributions:

M.N.M., A.N., A.O.A., R.A.I., and S.M.A. contributed to the conception and design of the study, methodology, data analysis and interpretation, and preparation and revision of the manuscript. All authors read and approved of the final manuscript.

6. REFERENCES

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. DOI:https://doi.org/10.1007/978-3-319-14142-8

Al-Hashedi, K. G., Magalingam, P., Maarop, N., Samy, G. N., Rahim, F. B. A., Shanmugam, M., & Hasan, M. K. (2022). A supervised model to detect suspicious activities in the bitcoin network. In *International Conference on Emerging Technologies and Intelligent Systems* (pp. 606–615). Springer International Publishing. DOI:https://doi.org/10.1007/978-3-031-25274-7_53

Antonopoulos, A. M., & Harding, D. A. (2023). *Mastering Bitcoin: Programming the open blockchain*. O'Reilly Media, Inc.

Asiri, A., & Somasundaram, K. (2025). Graph convolution network for fraud detection in bitcoin transactions. *Scientific Reports*, 15(1), 11076. DOI:https://doi.org/10.1038/s41598-025-95672-w

Bazán-Palomino, W. (2023). The increased interest in Bitcoin and the immediate and long-term impact of Bitcoin volatility on global stock markets. *Economic Analysis and Policy*, 80, 1080–1095. DOI:https://doi.org/10.1016/j.eap.2023.08.016

Böhme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2), 213–238. DOI:https://doi.org/10.1257/jep.29.2.213

Bolton, R. J., & Hand, D. J. (2004). Statistical fraud detection: A review. *Quality Control and Applied Statistics*, 49(3), 313–314. DOI:https://doi.org/10.1214/ss/1042727940

Chen, B., Wei, F., & Gu, C. (2021). Bitcoin theft detection based on supervised machine learning algorithms. *Security and Communication Networks*. DOI:https://doi.org/10.1155/2021/6643763

Chen, W., Wu, J., Zheng, Z., Chen, C., & Zhou, Y. (2019). Market manipulation of Bitcoin: Evidence from mining the Mt. Gox transaction network. In *Proceedings of IEEE INFOCOM 2019* (pp. 964–972). DOI:https://doi.org/10.1109/INFOCOM.2019.8737416

Cholevas, C., Angeli, E., Sereti, Z., Mavrikos, E., & Tsekouras, G. E. (2024). Anomaly detection in blockchain networks using unsupervised learning: A survey. *Algorithms*, 17(5), 201. DOI:https://doi.org/10.3390/a17050201

Dai, Z., & Huang, W. (2025). Improving energy management practices through accurate building energy consumption prediction: analyzing the performance of LightGBM, RF, and XGBoost models with advanced optimization strategies. *Electrical Engineering*, 107(9), 12583–12605.

Ding, H., Sun, Y., Huang, N., Shen, Z., & Cui, X. (2023). TMG-GAN: Generative adversarial networks-based imbalanced learning for network intrusion detection. *IEEE Transactions on Information Forensics and Security*, 19, 1156–1167. DOI:https://doi.org/10.1109/TIFS.2023.3331240

Elliptic. (2019). Elliptic dataset. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/ellipticco/elliptic-data-set

Elmougy, Y., & Liu, L. (2023). Demystifying fraudulent transactions and illicit nodes in the bitcoin network for financial forensics. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3979–3990). DOI:https://doi.org/10.1145/3580305.3599803

Fan, J., Mao, Y., Zhang, S., & Tan, Y. (2026). Transformer-Aware Graph Convolutional Network for Anomalous Bitcoin Transaction Detection. In *2026 IEEE International Conference on Power, Electronics and Green Energy (ICPEGE)* (pp. 1675–1679). IEEE.

Farrukh, H., Zafar, S., Rehman, Z. U., Shah, A. A., & Alshammry, N. (2025). Blockchain-Based Fraud Detection: A Comparative Systematic Literature Review of Federated Learning and Machine Learning Approaches. *Electronics*, 14(24), 4952. DOI:https://doi.org/10.3390/electronics14244952

Han, N., Zhang, R., Liu, X., & Zhang, H. (2026). Illicit Bitcoin transaction detection via feature-gated temporal graph learning.

Hasan, M., Rahman, M. S., Janicke, H., & Sarker, I. H. (2024). Detecting anomalies in blockchain transactions using machine learning classifiers. *Blockchain: Research and Applications*, 100207. DOI:https://doi.org/10.1016/j.bcra.2024.100207

Indira, M., Basha, R. M., Sivamma, P., Babu, M. S., & Ketan, S. H. N. (2026). Attention-Balanced Variational Autoencoder Framework for Credit Card Fraud Detection under Extreme Class Imbalance. In *2026 7th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1060–1067). IEEE.

Jordan, J. (2018). *Variational autoencoders*. Retrieved from https://www.jeremyjordan.me/variational-autoencoders/

Latif, M. N., Kaplan, M., & Khan, A. U. I. (2025). Unsupervised machine learning based anomaly detection in high frequency data: Evidence from Cryptocurrency Market. *Pakistan Journal of Commerce and Social Sciences*, 19(3), 407–440.

Lin, Z., Luo, Q., Wu, D., Shen, J., Li, L., Nong, X., & Qin, Z. (2026). Detecting illicit transactions in bitcoin: a wavelet-temporal graph transformer approach for anti-money laundering. *Scientific Reports*, 16(1), 1548. DOI:https://doi.org/10.1038/s41598-026-01548-x

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Na, G., Park, M., Kim, S., Shin, J., & Chai, S. (2026). Knowledge-Integrated Representation Learning for Crypto Anomaly Detection under Extreme Label Scarcity; Relational Domain-Logic Integration with Retrieval-Grounded Context and Path-Level Explanations. *arXiv preprint arXiv:2601.12839*. DOI:https://doi.org/10.48550/arXiv.2601.12839

Pocher, N., Zichichi, M., Merizzi, F., Shafiq, M. Z., & Ferretti, S. (2023). Detecting anomalous cryptocurrency transactions: An AML/CFT application of machine learning-based forensics. *Electronic Markets*, 33(1), 37. DOI:https://doi.org/10.1007/s12525-023-00650-z

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. Retrieved from https://bitcoin.org/bitcoin.pdf

Nerurkar, P. (2023). Illegal activity detection on Bitcoin transaction using deep learning. *Soft Computing*, 27(9), 5503–5520. DOI:https://doi.org/10.1007/s00500-022-07779-1

Nimmagadda, N., & Madiajagan, M. (2025). Graph Neural Networks for Illicit Transaction Detection in the Elliptic Bitcoin Dataset. In *2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)* (pp. 1121–1127). IEEE.

Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A., & Lima Netto, S. (2021). Variational autoencoder. In *Variational methods for machine learning with applications to deep networks* (pp. 111–149). Springer. DOI:https://doi.org/10.1007/978-3-030-70604-6_6

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*. DOI:https://doi.org/10.48550/arXiv.1908.02591