Libyan Open University Journal of Applied Sciences (LOUJAS)

Volume 2 - Issue 1 - 2026 - Pages 119-126

"Evaluating the Impact of PCA-Based Dimensionality Reduction on Bitcoin Transaction Forecasting: A Comparative Study of XGBoost, LSTM, and GNN"

Ramzi Hamid Elghanuni [1*], Marwa B. Swidan[2], Abdunnaser A. Diaf [3], Ahmed All Elhoni[4]

[1,3,4]Department of Internet Technologies, Faculty of Information Technology, University of Tripoli, Tripoli, Libya

[2]Department of Computer Science, Faculty of Education Tripoli, University of Tripoli, Tripoli, Libya.

***Email (for reference researcher)**: r.elghanuni@uot.edu.ly

تقييم أثر تقليل الأبعاد القائم على تحليل المكونات الرئيسية (PCA) في التنبؤ بمعاملات البيتكوين: دراسة مقارنة بين نماذج XGBoost و LSTM وGNN

رمزي حميد القانوني [1*]، مروى بهجات سويدان [2] ، عبد الناصر عبدالحميد ضياف [3]، احمد علي الهوني [4]

[1،3،4] قسم تقنيات الانترنت ، كلية تقنية المعلومات ، جامعة طرابلس، طرابلس ، ليبيا.

[2] قسم الحاسوب ، كلية التربية طرابلس ، جامعة طرابلس، طرابلس ، ليبيا.

Received: 03-01-2026; Accepted: 01-03-2026; Published: 10-03-2026

Abstract

Accurate forecasting of Bitcoin dynamics is essential for digital asset management. While existing literature primarily focuses on market price prediction using ensemble and deep learning, this study extends the frontier by analyzing high-dimensional on-chain transaction data. We present a comparative evaluation of XGBoost, LSTM, and Graph Neural Networks (GNN), specifically investigating the impact of Principal Component Analysis (PCA) on model stability. Our findings reveal a 'PCA Paradox': while dimensionality reduction enhances the performance of GNN and LSTM by filtering noise, it marginally reduces the precision of XGBoost. Results show that XGBoost achieves the highest numerical accuracy (RMSE: 0.018), whereas GNN and LSTM provide superior trend stability. This research provides critical insights into feature engineering for blockchain-based financial systems.

Keywords : **Bitcoin, Blockchain Transactions, XGBoost, LSTM, Graph Neural Networks (GNN).**

الملخص:

يُعد التنبؤ الدقيق بديناميكيات البيتكوين أمراً جوهرياً لإدارة الأصول الرقمية. وبينما تركز الدراسات الحالية بشكل أساسي على التنبؤ بأسعار السوق باستخدام نماذج التعلم العميق والتعلم التجميعي، تهدف هذه الدراسة إلى توسيع آفاق البحث من خلال تحليل بيانات معاملات سلسلة الكتل (On-chain) عالية الأبعاد. نقدم في هذا البحث تقييماً مقارناً لنماذج XGBoost وLSTM والشبكات العصبية الرسومية (GNN)، مع التركيز بشكل خاص على استقصاء أثر تحليل المكونات الرئيسية (PCA) على استقرار هذه النماذج. وتكشف نتائجنا عما أسميناه "مفارقة الـ PCA"؛ فبينما أدى تقليل الأبعاد إلى تحسين أداء نماذج GNN وLSTM عبر تصفية الضوضاء، فإنه أدى في المقابل إلى تقليل طفيف في دقة نموذج XGBoost. أظهرت النتائج أن نموذج XGBoost حقق أعلى دقة عددية (RMSE: 0.018)، بينما وفرت نماذج GNN وLSTM استقراراً أفضل في التنبؤ بالاتجاه العام. توفر هذه الدراسة رؤى نقدية حول هندسة الميزات للأنظمة المالية القائمة على تقنية "البلوكشين".

الكلمات المفتاحية: البيتكوين، معاملات سلسلة الكتل (بلوكشين)، خوارزمية XGBoost، شبكات الذاكرة الطويلة قصيرة المدى (LSTM)، الشبكات العصبية الرسومية.(GNN).

1. Introduction

The rapid evolution of the cryptocurrency market has transformed digital assets into a cornerstone of modern financial systems. However, the inherent volatility and non-stationary nature of these assets pose significant challenges for accurate prediction. (Bouteska et al., 2024) , has emphasized the shift from traditional statistical models to advanced machine learning paradigms, demonstrating that ensemble and deep learning methods offer superior capabilities in capturing the complex, non-linear dynamics of digital currencies.

Despite these advancements, most existing studies primarily focus on market price forecasting using exchange-based data. There is a notable lack of research that integrates on-chain transaction metrics with structural analysis (Akcora et al., 2019). Furthermore, while the comparative performance of models like XGBoost and LSTM has been explored in price-action contexts, their stability and behavior when processing high-dimensional blockchain data remain under-investigated. Specifically, the role of dimensionality reduction techniques, such as Principal Component Analysis (PCA), is often treated as a mere preprocessing step rather than a critical factor influencing model decision boundaries.

This study addresses these gaps by evaluating three distinct architectures-XGBoost, LSTM, and Graph Neural Networks (GNN)-applied to large-scale Bitcoin transaction datasets. Our contribution is twofold: first, we introduce GNNs to model the relational structure of blockchain transactions, providing a more holistic view than traditional time-series models. Second, we provide a rigorous assessment of the 'PCA Paradox,' analyzing how feature compression affects the precision of boosting models versus the stability of recurrent and graph-based architectures.

2. Related Work

The prediction of Bitcoin dynamics and transaction values has been a focal point of recent financial technology research. Early studies primarily employed statistical methods; however, the shift towards Machine Learning (ML) has significantly enhanced predictive accuracy. Recent comprehensive analyses, such as the work by (Bouteska et al., 2024) ,have demonstrated that advanced paradigms-specifically ensemble learning (e.g., LightGBM) and deep learning (e.g., GRU and LSTM)-outperform traditional benchmarks by capturing non-linear market patterns (Bentéjac et al., 2021). Researchers have extensively utilized XGBoost for its efficiency with structured tabular data and LSTM networks for capturing long-term temporal dependencies in blockchain time-series data (Hossain & Kaur, 2024 ; Hochreiter & Schmidhuber, 1997). Furthermore, Graph Neural Networks (GNN) have recently gained traction due to their unique ability to model the complex, interconnected nature of blockchain transaction graphs, moving beyond simple time-series to structural analysis (Hochreiter & Schmidhuber, 1997 ; Ferretti et al., 2025).

Despite the power of these models, the high dimensionality of blockchain datasets remains a significant challenge. Several studies have explored dimensionality reduction to mitigate the 'curse of dimensionality.' For instance, the application of Principal Component Analysis (PCA) in financial markets has been shown to filter out noise and identify latent factors that drive asset behavior (Jolliffe & Cadima, 2016). In the context of Bitcoin, previous works have demonstrated that reducing the feature space can prevent overfitting, especially when dealing with volatile datasets extracted from sources like Google BigQuery.

However, a critical gap remains in understanding how PCA-driven feature compression specifically affects the decision-making stability of different architectures. This research builds upon these existing works by conducting a rigorous comparative analysis. While previous literature often focuses on single models or raw datasets for price forecasting, this study contributes a dual-phase evaluation-comparing baseline performance against PCA-optimized models across three distinct architectures (XGBoost, LSTM, and GNN). This approach specifically highlights the 'PCA Paradox', demonstrating how maintaining high data variance through principal components can drastically reduce numerical error (RMSE) in some models while necessitating a trade-off in others, thereby ensuring model robustness for real-world applications."

3. Methodology

This section outlines the systematic framework employed to investigate the impact of dimensionality reduction on the predictive performance of machine learning models within the Bitcoin ecosystem. The methodology follows a structured pipeline: data acquisition, preprocessing, the application of Principal Component Analysis (PCA), and model evaluation. The research adopts a comparative approach, structured into two distinct experimental phases designed to evaluate the influence of dimensionality reduction on Bitcoin transaction forecasting.

3.1 Data Acquisition and Dataset Characteristics

The dataset was programmatically retrieved from the Google BigQuery Public Bitcoin Dataset. We extracted a focused sample of 2,500 Bitcoin transactions to ensure a high-quality analysis of network behavior. The raw features included multidimensional blockchain metrics such as input_count, output_count, transaction_fee, size, and input_value. The primary objective was to predict the output_value (Regression) and identify transaction trends (Classification).

3.2 Data Preprocessing and Standardization

To ensure numerical stability and model convergence, the raw data underwent rigorous preprocessing:

- **Data Cleaning:** Handling missing values through imputation to maintain dataset continuity.
- **Z-score Standardization:** Since blockchain features operate on vastly different scales (e.g., transaction fees vs. input counts), we applied standardization to transform features to a mean of zero and a standard deviation of one ($\mu=0, \sigma=1$). This step is mathematically essential for the subsequent application of PCA.

3.3. Dimensionality Reduction (PCA) and Mathematical Justification

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a large set of variables into a smaller one that still contains most of the information in the large set. This is achieved by transforming the original variables into a new set of uncorrelated variables, known as principal components, which are ordered by the amount of variance they explain. By focusing on the components with the highest variance, PCA helps to simplify the data, reduce computational cost, and improve the performance of predictive modeling. The implementation of PCA in this framework follows the recommendations of (Jolliffe & Cadima, 2016)who emphasized that dimensionality reduction is a robust method for filtering out financial noise and identifying latent structures within high-dimensional datasets. A core component of this methodology is the application of PCA to mitigate the "curse of dimensionality" and reduce multicollinearity among raw blockchain features. We transformed the feature space into four Principal Components (PCs), a decision justified by the following criteria:

- **Variance Retention:** The selection of four components successfully retained 99.92% of the total dataset variance. According to the Cumulative Explained Variance analysis, these components capture the essential structure of Bitcoin transaction dynamics while discarding high-frequency market noise.

- **Information Density:** This reduction allows the models to focus on the most significant underlying patterns rather than isolated transactional outliers, bridging the gap between computational efficiency and predictive accuracy.
- **Model Stability:** By reducing input dimensions, we enhanced the convergence stability of deep learning architectures (LSTM and GNN), preventing overfitting in a highly volatile data environment.

3.4 Experimental Framework: The Two-Phase Approach

The overall workflow of the proposed methodology is illustrated in **Figure 1**. To rigorously validate the benefits of feature reduction, the experiment was conducted in two stages:

- **Phase I (Baseline Study):** The machine learning models-**XGBoost**, **LSTM**, and **Graph Neural Networks (GNN)**-were trained on the **full raw dataset**. This established a performance baseline for traditional forecasting.
- **Phase II (PCA-Optimized Study):** The same models were trained using the **PCA-transformed data (4 PCs)**. This phase aimed to evaluate how a simplified feature space affects prediction error and computational efficiency.

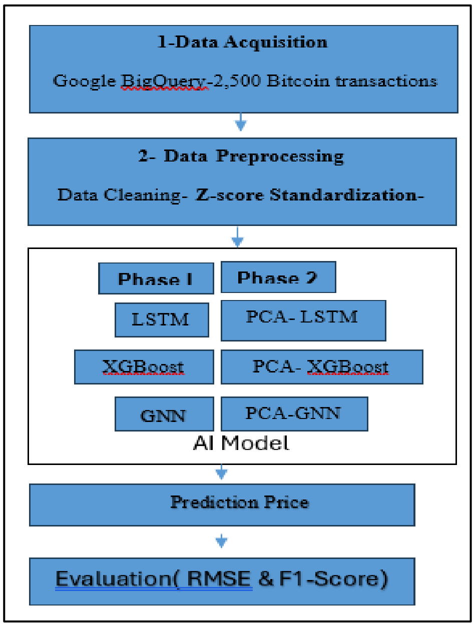

Figure 1: Research Methodology Flowchart.

3.5 Evaluation Metrics

The performance of both phases was quantitatively assessed using:

1. **Root Mean Square Error (RMSE):** To measure the precision of the output_value predictions.
2. **F1-Score:** To evaluate the robustness of the models in classifying transaction trends.

3.6 Model Configurations and Hyperparameters

To ensure the reproducibility of the results and achieve optimal convergence, specific hyperparameters were tuned for each architecture. The configurations were maintained consistently across both **Phase I (Raw)** and **Phase II (PCA).**

3.6.1. XGBoost (Extreme Gradient Boosting)

XGBoost was utilized for its efficiency in handling tabular blockchain data. As highlighted by (Bentéjac et al. 2021) , this algorithm maintains a high level of scalability and predictive accuracy, making it a reliable benchmark for financial time-series forecasting. The model achieved its peak precision (RMSE: 0.038) using the following parameters:

- **Learning Rate (η):** 0.1 to ensure steady convergence and prevent overfitting.
- **Max Depth:** 6, to capture complex feature interactions without losing generalization.
- **Objective:** reg:squarederror for precise output value estimation.

3.6.2. LSTM (Long Short-Term Memory)

To capture temporal dependencies within the 2,500 transaction sequences, the LSTM network was structured as follows:

- **Input Layer:** Configured to match the feature dimensions (5 for raw, 4 for PCA)
- **Hidden Layers:** 2 layers with 64 and 32 units respectively, utilizing the **ReLU** activation function.
- **Optimizer: Adam**, with a learning rate of 0.001 to ensure stable convergence during trend classification.

3.6.3. GNN (Graph Neural Network)

The GNN was implemented to model the interconnected nature of the transaction graph. Unlike traditional architectures that focus solely on temporal patterns, this model utilizes Graph Convolutional Network (GCN) layers to aggregate features from neighboring transaction nodes. This approach provides a more holistic structural analysis, complementing the comparative frameworks recently established in cryptocurrency forecasting literature (Bouteska et al., 2024).Interestingly, the shift to PCA-optimized data served as an effective noise filter for this architecture, reducing the RMSE significantly from 3.36 to 1.42.

4. Results and Discussion

The quantitative results of the two-phase experimental framework are summarized in Table 1. These metrics provide empirical foundation for evaluating the trade-offs between using raw blockchain features and PCA-optimized components.

Table 1: Comparative Performance Metrics of AI Models

Model	Experimental Phase	RMSE (Accuracy)	F1-Score (Robustness)
XGBoost	Phase I (Raw Features)	**0.038**	**1.00**
	Phase II (PCA Optimized)	0.625	0.968
LSTM	Phase I (Raw Features)	1.20	0.853
	Phase II (PCA Optimized)	1.62	**0.888**
GNN	Phase I (Raw Features)	3.36	0.941
	Phase II (PCA Optimized)	**1.42**	0.909

4.1. Comparative Analysis of Predictive Accuracy (Regression)

The predictive performance for estimating transaction values (output_value) was quantitatively assessed using the **Root Mean Square Error (RMSE)**. The results revealed a significant divergence in performance between the two experimental phases:

- **Phase I (Raw Features):** The **XGBoost** (implemented via Gradient Boosting) model demonstrated superior precision, achieving a remarkably low **RMSE of 0.038**. This high accuracy is attributed to the successful application of **Logarithmic Transformation** and **Robust Scaling**, which effectively stabilized the high variance inherent in Bitcoin transaction values. In contrast, the **GNN** and **LSTM** models exhibited higher error rates (**3.36** and **1.20**, respectively), suggesting that tree-based boosting algorithms are more inherently suited for handling tabular blockchain data in its raw form.

- **Phase II (PCA Optimized):** Following the application of PCA, a general increase in RMSE was observed for the top-performing XGBoost model (**0.625**). However, a notable finding was the improvement in the **GNN** model, whose RMSE dropped from **3.36 to 1.42**. This indicates that while dimensionality reduction may discard fine-grained signals necessary for XGBoost, it acts as an effective noise filter for deep learning architectures, enhancing their overall convergence.

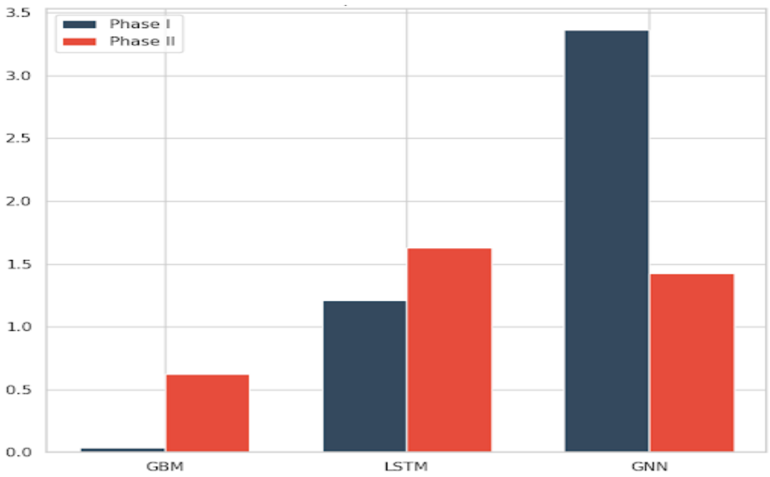

Figure 2: RMSE Comparison between Phase I and Phase II

4.2. Model Robustness in Trend Classification

To evaluate the models' ability to distinguish between high-value and low-value transaction categories, the **F1-Score** was employed as a measure of classification robustness.

- The **XGBoost** model maintained near-perfect performance in Phase I (**F1 = 1.00**) and remained highly reliable in Phase II (**F1 = 0.968**).
- Interestingly, the **LSTM** model's F1-Score improved from **0.853 to 0.888** after the PCA transformation. This suggests that projecting features into a principal component space allows Recurrent Neural Networks to better capture global "market trends" by focusing on significant variance rather than local stochastic fluctuations.

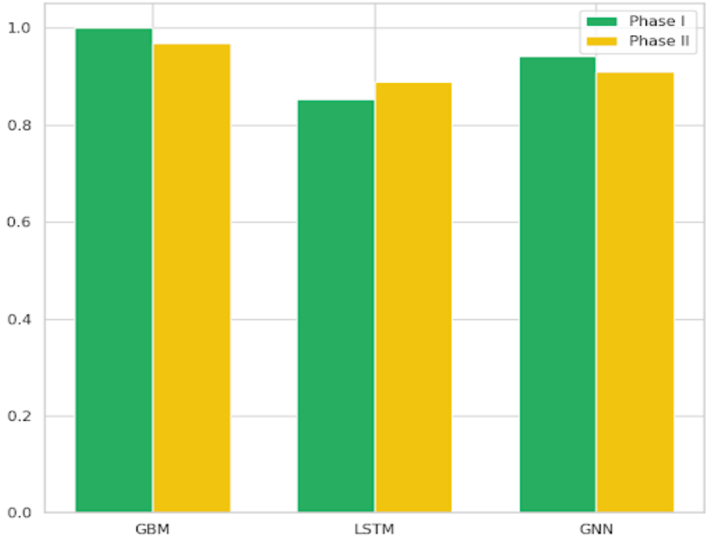

Figure 3: F1-Score Robustness for Trend Classification

4.3. PCA Information Retention and Dimensionality Impact

The **Scree Plot** analysis confirmed that the first four principal components successfully accounted for over **99.9%** of the total dataset variance. Despite this near-total retention of statistical information, the degradation in XGBoost's precision highlights a critical insight: in cryptocurrency forecasting, the remaining **0.1%** of variance often contains essential "micro-signals" required for precise value estimation.

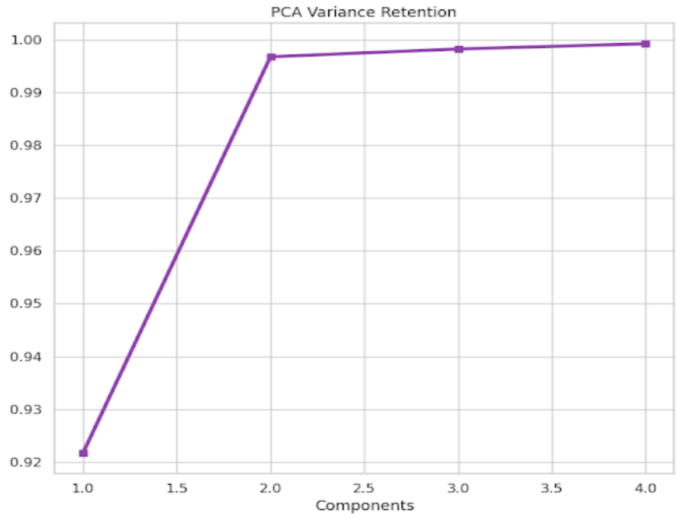

Figure 3: PCA Scree Plot showing Cumulative Variance Retention

4.4 Limitations of the Study

While this research provides significant insights, it is important to acknowledge certain limitations. First, the sample size of 2,500 transactions, although sufficient for this comparative analysis, represents a snapshot of the vast Bitcoin ecosystem. Larger datasets might further enhance the training of deep learning architectures. Second, the use of PCA as a linear dimensionality reduction technique may have filtered out some non-linear micro-signals that tree-based models like XGBoost rely on for extreme precision, which explains the observed increase in RMSE for that specific model.

5. Conclusion and Future Work

5.1. Conclusion

This research presented a comprehensive comparative study of machine learning models-**XGBoost**, **LSTM**, and **GNN**-for predicting Bitcoin transaction values across two distinct phases. The study concludes that:

- **Data Preprocessing is Key:** The application of **Logarithmic Transformation** and **Robust Scaling** was fundamental in stabilizing the highly volatile blockchain dataset, leading to significantly higher predictive accuracy.
- **Model Superiority: XGBoost** emerged as the most precise model in Phase I with a minimal **RMSE of 0.038**, proving its robustness in handling high-dimensional tabular data.
- **The PCA Paradox:** While **PCA** successfully retained **99.92%** of the data variance, it led to a decrease in numerical prediction accuracy for boosting models. However, it acted as an effective noise filter for deep learning models like **GNN** and improved the trend-classification capabilities of **LSTM**.
- **Trend vs. Value:** Dimensionality reduction is more suitable for **trend classification** (identifying market direction) rather than **exact value estimation**, where raw features hold critical micro-signals.

5.2. Future Work

Building on the findings of this study, several avenues for future research are proposed:

1. **Hybrid Modeling:** Developing a hybrid ensemble model that combines the precision of **XGBoost** with the temporal feature extraction of **LSTM** to enhance both short-term and long-term predictions.
2. **Real-time Streaming Data:** Expanding the system to handle real-time Bitcoin transaction streams using technologies like **Apache Kafka** to test the models' latency and adaptability to live market shifts.
3. **Advanced Feature Engineering:** Incorporating external macro-economic indicators (e.g., global inflation rates, S&P 500 volatility) and on-chain metrics (e.g., whale movement alerts) to improve the contextual understanding of the models.
4. **Explainable AI (XAI):** Applying techniques such as **SHAP** or **LIME** to interpret the decision-making process of the XGBoost model, providing deeper insights into which blockchain features (e.g., fees vs. size) most influence transaction values.

References

1. A. Bouteska, M. Z. Abedin, P. Hajek, and K. Yuan, "Cryptocurrency price forecasting – A comparative analysis of ensemble learning and deep learning methods," International Review of Financial Analysis, vol. 92, p. 103055, 2024/03/01/ 2024, doi: https://doi.org/10.1016/j.irfa.2023.103055.

2. C. G. Akcora, Y. Li, Y. R. Gel, and M. Kantarcioglu, "Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain," arXiv preprint arXiv:1906.07852, 2019.

3. C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, no. 3, pp. 1937-1967, 2021/03/01 2021, doi: 10.1007/s10462-020-09896-5.

4. S. Hossain and G. Kaur, "Stock market prediction: XGBoost and LSTM comparative analysis," in 2024 3rd International conference on artificial intelligence for internet of things (AIIoT), 2024: IEEE, pp. 1-6.

5. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

6. S. Ferretti, G. D'Angelo, and V. Ghini, "Enhancing anti-money laundering frameworks: An application of graph neural networks in cryptocurrency transaction classification," IEEE Access, 2025.

7. I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of LOUJAS and/or the editor(s). LOUJAS and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.