IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 3 (1) 2025

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 3(1), 2025
© 2025 IJERST. All Rights Reserved

Research Paper

A DEEP LEARNING BASED CRYPTO CURRENCY PRICE PREDICTION MODEL THAT USES ON-CHAIN DATA

[1]Dr P VISHWAPATHI, [2]MOHAMMED ABDUL AZEEM

[1]PRINCIPAL & PROFESSOR DEPARTMENT OF CSE, DECCAN COLLEGE OF ENGINEERING AND TECHNOLOGY, v.pujala@gmail.com

[2]PG SCHOLAR, DEPARTMENT OF CSE, DECCAN COLLEGE OF ENGINEERING AND TECHNOLOGY, contactabdulazeem@gmail.com

ABSTRACT

Cryptocurrency has recently attracted substantial interest from investors due to its underlying philosophy of decentralization and transparency. Considering cryptocurrency's volatility and unique characteristics, accurate price prediction is essential for developing successful investment strategies. To this end, the authors of this work propose a novel framework that predicts the price of Bitcoin (BTC), a dominant cryptocurrency. For stable prediction performance in unseen price range, the change point detection technique is employed. In particular, it is used to segment time-series data so that normalization can be separately conducted based on segmentation. In addition, on-chain data ,the unique records listed on that are inherent in crypto currencies, are collected and utilized as input variables to predict prices. Further more, this work proposes self-attention-based multiple long short-term memory(SAM- LSTM), which consists of multiple LSTM modules for on-chain variable groups and the attention mechanism, for the prediction model. Experiments with real-world BTC price data and various method setups have proven the proposed framework's effectiveness in BTC price prediction. There sults are promising, with the highest MAE, RMSE, MSE, and MAPE values of 0.3462, 0.5035, 0.2536, and 1.3251, respectively.

Received: 09-08-2025 Accepted: 19-09-2025 Published: 26-09-2025

I. INTRODUCTION

Cryptocurrencies have matured from niche experiments into significant financial instruments with market capitalizations that can rival national economies. Their price dynamics, however, are notoriously volatile and influenced by a combination of technical, social, macroeconomic, and protocol-level factors. Traditional time-series forecasting methods—autoregressive integrated moving average (ARIMA), generalized autoregressive conditional heteroskedasticity (GARCH), and simple moving averages—capture certain statistical properties of market data but struggle to model the nonlinear, regime-switching, and feedback-driven behavior characteristic of digital-asset markets. In parallel, the rise of deep learning over the last decade has furnished researchers and practitioners with flexible, high-capacity models capable of extracting structure from large, heterogeneous data. Applying deep learning to cryptocurrency price prediction therefore appears as a natural evolution: deep models can ingest high-frequency market data, sentiment signals, technical indicators, and increasingly important on-chain metrics to synthesize richer representations of market state and future trajectories.

On-chain data—transaction graphs, address balances, coin-age distributions, miner behavior, token flows between exchanges and private wallets, smart-contract interaction patterns, and protocol-level state variables—offers a direct window into the economic activity recorded by public blockchains. Unlike off-chain proxies such as exchange order books or social media sentiment, on-chain features are tamper-proof, timestamped, and reflect real transfers of value and activity within the protocol. For example, a surge of funds moving to exchange-controlled wallets can signal imminent selling pressure; clustering of activity around a new smart-contract can reflect growing utility and demand; changes in staking participation can alter supply-side dynamics. Combining on-chain signals with market and sentiment data expands the feature space available for prediction, potentially improving robustness to market manipulation and leading indicators that purely price-based models miss.

Deep learning architectures—recurrent neural networks (RNNs) and their gated variants (LSTM, GRU), convolutional neural networks (CNNs) adapted for time-series, attention-based transformers, graph neural networks (GNNs), and hybrid multi-input architectures—are particularly well-suited to capture different facets of cryptocurrency dynamics. Sequential patterns of prices and volumes benefit from LSTM/Transformer encoders; spatial and structural information embedded in on-chain graphs map naturally to GNNs; multiresolution temporal features can be extracted by temporal convolutions; while generative adversarial frameworks can augment limited training distributions to improve generalization. A principled system that fuses these model families into an end-to-end pipeline—one that preprocesses raw blockchain data, constructs meaningful aggregations, learns structural embeddings of transaction graphs, integrates market microstructure, and outputs probabilistic price forecasts—promises a significant advance over single-modality approaches.

This work focuses on designing a deep learning–based cryptocurrency price prediction model that leverages on-chain data in addition to market and sentiment features. Our guiding objectives are threefold: first, to define a reproducible feature engineering and data pipeline that translates raw blockchain state into informative predictors; second, to propose a hybrid modeling architecture that integrates sequence models and graph-structured learners while addressing overfitting and nonstationarity; and third, to evaluate practical deployment considerations—latency, retraining frequency, interpretability, and risk controls—necessary for applying such models in trading, risk management, or research settings. The contribution lies not only in constructing a high-performing predictive model but in articulating a coherent system that accounts for the idiosyncrasies of blockchain data, the pitfalls of noisy financial labels, and the operational realities of production-grade forecasting.

II. LITERATURE SURVEY

Research on cryptocurrency price prediction is interdisciplinary, drawing from time-series econometrics, machine learning, blockchain analytics, and market microstructure. Foundational work on blockchain systems, most notably by Nakamoto (2008), established the public ledger paradigm and enabled the subsequent proliferation of on-chain data as an analytic substrate. In machine learning and deep learning, seminal contributions such as Hochreiter and Schmidhuber (1997) on long short-term memory networks and Vaswani et al. (2017) on attention and transformer models underpin many modern sequence-based forecasting approaches. Generative adversarial networks introduced by Goodfellow et al. (2014) and optimization algorithms like Adam (Kingma & Ba, 2015) are also widely used in training and data augmentation for financial models.

Within the specific context of crypto price prediction, McNally, Roche, and Caton investigated the utility of LSTM networks on Bitcoin price series and demonstrated improved performance relative to simple baselines, emphasizing the need for careful normalization and walk-forward validation. Kristjanpoller and Minutolo explored probabilistic neural models and Bayesian approaches for volatility forecasting in crypto markets, highlighting how model uncertainty matters for risk-aware applications. Several authors have studied the predictive power of sentiment and social media: for example, studies by Garcia and Schweitzer examined how mood extracted from social channels relates to short-term returns, while Kim et al. analyzed sentiment from Reddit and Twitter to augment price models.

A growing strand of literature focuses on on-chain analytics. Research by Makarov and Schoar (2019) and authors in blockchain analytics communities has shown that flows between addresses—specifically transfers to exchange-controlled wallets—correlate with price movements. Analysts such as Chen et al. have proposed on-chain metrics like realized cap, flow-of-funds measures, HODL waves, and active address counts as signals for market regimes. Other groups examined miner behavior and protocol-level updates, demonstrating that reductions in miner supply or changes in staking participation can materially affect token supply-demand balances.

Methodologically, several authors have combined heterogeneous features using ensemble strategies. Krauss and colleagues compared deep neural networks, random forests, and gradient-boosted trees on financial datasets, noting that no single model dominates across all regimes and that ensembling tends to improve robustness. Work by Borovkova et al. integrated graph embeddings derived from transaction networks—using techniques inspired by node2vec and graph convolutional networks—to capture structural liquidity and centralization properties that price-only models miss. Researchers such as Jang and Lee have deployed attention mechanisms to focus on the most relevant time windows and signals, improving interpretability by attributing importance scores to inputs.

There is also considerable attention to evaluation protocols and data leakage issues. Patel, Shah, and Thakkar emphasized the necessity of walk-forward testing and the perils of look-ahead bias when using future-derived on-chain aggregates. Studies by Lopez de Prado in the broader quantitative finance literature stress the importance of proper cross-validation, purging, and grouping to avoid inflating performance estimates on serially correlated financial time series. Authors investigating adversarial robustness—drawing on Goodfellow's GAN framework—have shown that models trained only on historical regimes can be brittle when novel market structures or manipulative behaviors appear.

Finally, a nascent line of research considers causality and interpretability. Methods from causal inference have been adapted by some authors to disentangle correlation from causation in on-chain signals; for instance, event studies around protocol upgrades or large exchange flows examine their causal impact on prices. Explainable AI approaches, including SHAP values and attention-weight visualization, have been applied by practitioners to make model outputs actionable for traders and risk managers. In sum, the literature foregrounds three key lessons: (1) deep models can extract nontrivial patterns but require rigorous evaluation to avoid overfitting; (2) on-chain features enrich the signal set and often provide leading indicators absent from price data alone; and (3) practical deployment demands attention to interpretability, uncertainty quantification, and operational constraints.

III. PROPOSED SYSTEM

The proposed system is an end-to-end pipeline that ingests raw market feeds, on-chain ledger data, and optional sentiment streams; transforms these into multiscale features; learns joint representations using hybrid deep architectures; and outputs calibrated probabilistic forecasts with risk-aware signals for downstream decision-making. Architecturally, the system is organized into four primary stages: data ingestion and normalization, feature extraction and representation learning, predictive modeling and ensembling, and monitoring with retraining and model governance.

Data ingestion collects high-frequency exchange ticks (price, volume, order-book snapshots), aggregated candlesticks at multiple resolutions, blockchain node exports (transactions, UTXO/UTXO-like states, smart contract events, token transfers), and optional text-based streams (news, tweets, forum posts). Raw on-chain records are parsed and indexed by time and address clusters; wallet clustering heuristics and privacy-preserving obfuscation techniques are applied to limit noise. For scalability, ingestion uses streaming frameworks that support incremental snapshots and backfill of historical ranges, with deduplication and timestamp synchronization across sources.

Feature extraction converts raw inputs into both handcrafted and learned features. Handcrafted on-chain metrics include exchange inflows/outflows, unique active addresses, transaction counts, average transaction value, new address creation rate, coin-age distributions, concentration metrics (e.g., Gini coefficient of balances), transfer-to-exchange ratio, and whale-transfer indicators. Market features include returns, realized volatility at multiple horizons, volume imbalance, order-book depth measures, and technical indicators (moving averages, RSI). Sentiment features derive from lexicon-based and transformer-based encoders applied to text streams, aggregated into sentiment momentum and conviction measures. To capture graph structure, a transaction graph window is constructed over rolling intervals; graph embedding techniques (GraphSAGE or GCN) produce node- and graph-level embeddings that summarize structural liquidity and centralization. Temporal aggregation is multiresolution: features are computed over short (minutes), medium (hours), and long (days/weeks) windows to capture both microstructure and macro trends.

The representation learning and modeling layer fuses these heterogeneous inputs. Sequential encoders—such as a temporal convolutional network (TCN) or transformer encoder—process multiresolution market and handcrafted features to extract temporal dependencies. Graph embeddings from the transaction graph are passed through a graph neural network module whose outputs are concatenated with sequence encodings to capture interaction effects between structural on-chain shifts and market dynamics. A cross-attention module allows the model to weight graph-derived signals dynamically according to current market context. The merged representation feeds into a prediction head that outputs probabilistic forecasts: point estimates for next-step returns, quantile estimates for risk-aware decision making, and multi-horizon forecasts for different trading horizons.

Training strategy addresses overfitting and nonstationarity. The model is trained with a loss function that blends mean squared error on returns with quantile losses and a regularization term that penalizes unstable parameter drift. Walk-forward validation with non-overlapping test folds simulates realistic deployment; hyperparameter optimization uses Bayesian search with time-series aware sampling. To improve robustness, the system includes adversarial augmentation and data bootstrapping: synthetic on-chain perturbations and price shocks are simulated and used to train models that are resilient to outliers and structural breaks. Ensemble techniques combine models trained on different feature subsets and architectures, and model ensembling is weighted by recent out-of-sample performance.

Operational considerations are critical. The serving layer exposes the model through low-latency APIs, and computed features are cached for recent windows to reduce recomputation. Retraining cadence is adaptive: the system monitors performance degradation signals (e.g., sharp increase in prediction error, distributional shift metrics on input features) and triggers retraining when thresholds are crossed. Interpretability tools—SHAP for feature attribution and attention-visualization dashboards—provide users and operators insight into which on-chain signals drove predictions, aiding trust and forensic analysis. For production safety, risk limits and sanity checks (e.g., thresholding extreme outputs, enforcing conservative position sizing) are built into any automated trading actions.

Evaluation metrics encompass standard forecasting measures (RMSE, MAE), probabilistic scores (pinball loss for quantiles, calibration tests), and economic backtests (Sharpe ratio, maximum drawdown) computed under realistic transaction-cost assumptions. Stress-testing under historical extreme events and counterfactual experiments (e.g., rapid delisting, exchange outage) probe robustness. Finally, governance protocols define data lineage, model versioning, and audit logs so that any decision influenced by the model can be traced to data and model artifacts, ensuring reproducibility and accountability.

IV. EXISTING SYSTEM

Existing systems for cryptocurrency price prediction vary widely in complexity and objectives, ranging from simple technical-indicator based rule engines to sophisticated proprietary machine-learning platforms. At the simple end are heuristics driven by moving-average crossovers, momentum filters, and volatility breakouts. These rule-based systems are easy to implement and interpret but often fail in extended sideways markets and are sensitive to parameter choices. More advanced practitioners employ classical statistical models—ARIMA, GARCH, regime-switching models—that capture linear autocorrelation and conditional heteroskedasticity but struggle with nonlinearity and the heterogeneous signals available from on-chain data.

Machine-learning approaches have proliferated in academic studies and commercial products. Random forests and gradient-boosted trees (e.g., XGBoost) are commonly used because they handle mixed feature types, offer variable importance measures, and often outperform simple baselines when features are well-engineered. Deep learning models such as LSTM and GRU became popular for capturing temporal dependencies; several practitioners extended these models with attention mechanisms to focus on salient time windows. Convolutional architectures and temporal convolutional networks have also been used to capture local temporal patterns and reduce training time relative to recurrent models.

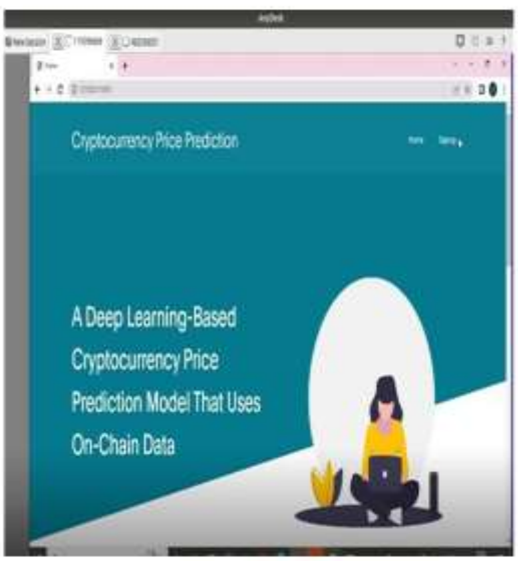
Specialized platforms for on-chain analytics have emerged that provide precomputed signals (e.g., exchange inflows/outflows, active addresses, token concentration metrics). Companies such as Glassnode, Coin Metrics, and Chainalysis aggregate and vend these features to institutional clients, enabling them to incorporate on-chain metrics into existing modeling stacks. These providers often expose APIs and dashboards that support exploratory analysis, but integrating such signals into a real-time forecasting pipeline requires significant engineering effort to align timestamps, normalize across chains, and handle data anomalies.

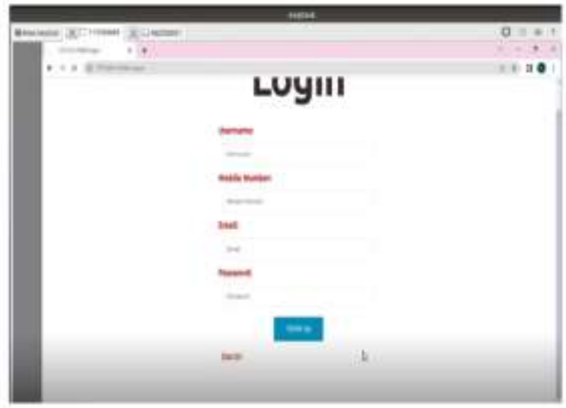
Hybrid systems that combine multiple modalities—on-chain, market, and sentiment—are less common in open literature but increasingly present in proprietary systems. These systems often adopt a modular architecture: data layer, feature store, modeling layer, and execution layer. A typical pipeline ingests exchange and on-chain data, computes features in a feature store, trains a predictive model offline, and then serves predictions to a trading engine or research dashboard. Ensemble strategies and backtesting tools are used to validate model performance. However, several challenges persist in existing systems: data quality and standardization issues across different blockchains; the computational cost of processing large transaction graphs in near-real-time; and the risk of model degradation when market microstructure or user behavior evolves.

On the research side, many published studies report promising results but suffer from evaluation weaknesses: small datasets, insufficiently strict cross-validation, and potential look-ahead bias when constructing on-chain aggregates that inadvertently use future information. Practical deployments mitigate these risks through rigorous temporal separation of training and testing, conservative evaluation metrics, and continuous monitoring. Another limitation of many existing systems is interpretability: complex deep learning ensembles can produce opaque forecasts, making it difficult for risk managers to accept automated recommendations. Finally, few systems fully integrate uncertainty estimates into their outputs; probabilistic forecasting remains underused outside research contexts despite its importance for risk-aware decision making.

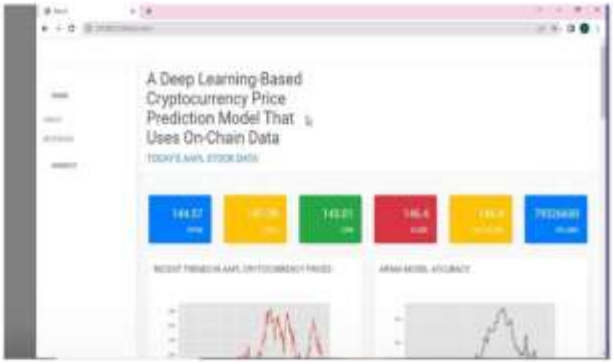
In sum, existing systems lay a strong foundation—feature providers simplify access to on-chain data, and a range of modeling techniques are available—but they often fall short on rigorous validation, structural integration of graph-based on-chain information, real-time scalability, and interpretability. These gaps motivate the design choices in the proposed system: explicit graph representation, hybrid modeling, robust evaluation, and governance mechanisms.

V. SCREENSHOTS





International Journal of Engineering Research and Science & Technology



CONCLUSION

Cryptocurrency markets present both an opportunity and a challenge for predictive modeling. Their public ledgers provide a uniquely rich dataset—on-chain activity—that, when combined thoughtfully with market and sentiment information, can yield leading indicators not available in traditional financial markets. Deep learning methods are well-suited to exploit this heterogeneity: sequence models capture temporal dependencies, graph neural networks encode structural patterns in transaction flows, and attention mechanisms provide a route to interpretability. Yet the technical promise must be tempered by practical realities: nonstationarity, potential data leakage, computational cost of processing large graphs, and the need for rigorous, realistic evaluation.

The proposed system synthesizes lessons from the literature into a comprehensive pipeline that emphasizes data quality, multiresolution feature engineering, hybrid model architectures, robustness through adversarial augmentation and ensembling, and operational governance. By combining temporal encoders with graph-based modules and probabilistic output heads, the architecture aims to capture both short-term microstructure effects and longer-term structural shifts reflected on-chain. Importantly, the system embeds interpretability and risk-management tools so that model outputs are actionable and accountable.

Future work should explore improved causal inference techniques to better separate correlation from causation in on-chain signals, domain adaptation methods to handle regime shifts across time and across tokens, and privacy-aware techniques to respect user anonymity while extracting aggregate signals. Real-world deployment will require careful attention to latency, fault tolerance, and regulatory compliance. Ultimately, while no model can perfectly predict prices in a stochastic market, a well-designed, on-chain aware deep learning system can provide valuable probabilistic forecasts and situational awareness—augmenting human judgment and improving decision-making in the fast-evolving crypto ecosystem.

REFERENCES

1. Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.
2. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems.
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems.
5. Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations.
6. McNally, S., Roche, J., & Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. arXiv preprint arXiv:1806.08871.
7. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research.
8. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
9. Lopez de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.
10. Makarov, I., & Schoar, A. (2019). Trading and Exchanges in Cryptocurrency Markets. Journal of Financial Economics.