Cryptocurrency Price Forecasting with a Hybrid LSTM–Transformer Architecture

Rezwanur Rahman Hritom
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rezwanur.rahman.hritom@g.bracu.ac.bd

H.M. Hasnain Jahangir Aqib
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
hasnain.jahangir.aqib@g.bracu.ac.bd

Nirban Roy
Dept. of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
nirban.roy@g.bracu.ac.bd

***Abstract*—Cryptocurrency markets exhibit high volatility, non-linear dynamics and frequent regime shifts, making price forecasting a challenging time-series problem. This paper proposes a hybrid Long Short-Term Memory (LSTM)–Transformer architecture for multi-cryptocurrency forecasting of Bitcoin (BTC), Litecoin (LTC) and Ethereum (ETH). The model combines a sequential LSTM branch and a global-dependency Transformer branch through a learned fusion gate that adaptively weights their contributions. Using 476 days of OHLCV data aggregated from minute-level records and 15 normalized technical indicators per asset, sequences of length 30 are used to predict 1-day, 7-day and 30-day horizons. Experimental results on consolidated BTC, LTC and ETH data show that the proposed hybrid model improves 1-day Mean Absolute Percentage Error (MAPE) from 82.02% (LSTM baseline) to 77.76% and increases R^2 from 0.3451 to 0.4278, while also achieving the highest directional accuracy (55.82%) among all compared models. These findings indicate that the hybrid architecture generalizes across assets with different price scales and volatility profiles and can serve as a practical building block for production-grade cryptocurrency forecasting systems.**

***Index Terms*—Cryptocurrency, Time series forecasting, LSTM, Transformer, Deep learning, Bitcoin, Ethereum, Litecoin.**

I. Introduction

Cryptocurrencies such as Bitcoin, Litecoin and Ethereum trade continuously across global exchanges and are influenced by macroeconomic news, market microstructure, speculative behavior and sentiment [9]. Their prices display heavy tails, volatility clustering and non-stationary dynamics, which challenge traditional econometric models and motivate the use of deep learning approaches for forecasting.

Most prior work focuses on single-asset prediction, particularly Bitcoin [**?**], which risks overfitting to one asset's scale and regime. In contrast, multi-cryptocurrency models are encouraged to learn scale-invariant and cross-asset patterns. In this work, three liquid cryptocurrencies are considered jointly: BTC (typical price range around 50–60k USD), LTC (200–300 USD) and ETH (2000–2500 USD). A unified model trained on all three assets can exploit both shared and asset-specific structures while being robust to price-scale differences.

Recurrent architectures such as LSTM and GRU have shown promise for cryptocurrency forecasting because they can capture short-term temporal dependencies and non-linear relationships. However, their ability to model long-range dependencies is limited by gradient attenuation and sequential processing. Transformer-based models, powered by multi-head self-attention, alleviate these issues by modeling all pairwise dependencies with a constant gradient path length and high parallelism, but they require more data and are computationally expensive for long sequences.

This paper addresses a research gap by designing and evaluating a hybrid LSTM–Transformer architecture with a learned fusion gate for multi-cryptocurrency price forecasting. The **main contributions** are:

- A hybrid architecture that combines an LSTM branch for local sequential patterns with a Transformer branch for global dependencies through a trainable fusion gate.
- A multi-cryptocurrency training and evaluation setup where BTC, LTC and ETH time series are consolidated into a single dataset with currency-agnostic technical indicators.
- An empirical comparison against LSTM, BiLSTM, GRU and standalone Transformer baselines using MAPE, RMSE, MAE, R^2 and directional accuracy.

II. Literature Review

A. Single-cryptocurrency forecasting

Early deep learning studies typically model one cryptocurrency, most often Bitcoin, using LSTM or GRU networks. [7] use LSTM to forecast XRP/USDT and report competitive MAPE and RMSE compared to traditional models. Other works examine architectural variants such as BiLSTM and GRU and observe similar performance when trained on single-asset data, highlighting the importance of feature engineering and hyperparameter choice rather than architecture alone.

These single-asset models are usually trained and evaluated on one cryptocurrency, which may lead to overfitting to specific price scales or local market regimes. As a result, it is unclear whether the learned patterns transfer to other assets with different volatility and microstructure.

B. Transformer-based time-series models

The Transformer architecture introduced in has been adapted for time-series forecasting through variants such as Temporal Fusion Transformers (TFT) [2]. These models use multi-head self-attention to capture long-range temporal relations and can incorporate static, known and observed features. Recent work shows that Transformer-based architectures can outperform LSTM baselines on long-horizon forecasting tasks under sufficient data and careful regularization.

For cryptocurrencies, several studies integrate Transformers with sentiment features [6], achieving substantial improvements over purely price-based LSTM models. However, most of these works remain focused on single assets or do not investigate explicit hybridization of RNNs and Transformers for multi-asset settings.

C. Hybrid architectures

Hybrid sequence models combine different neural components to leverage complementary strengths proposed LSTM–GRU hybrids, reporting improvements over individual models on cryptocurrency price prediction tasks [1]. Hybrid RNNs with attention have also been applied to financial time series, where attention mechanisms reweight important timesteps.

Despite these advances, there is limited evidence on LSTM–Transformer hybrids with learned fusion gates specifically tailored for multi-cryptocurrency price forecasting. This motivates the proposed architecture, which explicitly learns to balance local sequential dynamics and global temporal dependencies.

III. Dataset and Preprocessing

A. Data source and aggregation

The experiments use the "Cryptocurrency Timeseries 2020" dataset from Kaggle, which contains minute-by-minute OHLCV (Open, High, Low, Close, Volume) data for BTC, LTC and ETH from the Gemini exchange. For each asset, roughly 658k one-minute candles between January 1, 2020 and April 20, 2021 are available, resulting in 476 daily candles per cryptocurrency after aggregation.

Minute-level data are aggregated into daily OHLCV bars by taking the open of the first minute, the high and low across the day, the close of the last minute and the sum of volumes. This transformation reduces noise and aligns the three assets temporally.

B. Feature engineering

For each cryptocurrency, 15 technical indicators are computed to construct a currency-agnostic feature set. The indicators are grouped into:

- Momentum: RSI(14), MACD, MACD signal, Rate of Change(12), 1-day momentum.
- Volatility: Bollinger Bands %B(20), ATR(14), 20-day historical volatility.
- Trend: SMA(7, 14, 30), EMA(14), 30-day linear regression slope.
- Volume: On-Balance Volume (OBV), 5-day volume SMA.

All features are scaled per asset using Min–Max normalization to enforce scale invariance between BTC, LTC and ETH price ranges.

C. Sequence construction and split

A sliding window of 30 days is used to form input sequences. For each asset, sequences of shape $(30, 15)$ are paired with targets representing 1-day, 7-day and 30-day ahead price changes. Chronological splits are applied as follows:

- Training: 291 sequences (Jan 31–Nov 16, 2020).
- Validation: 62 sequences (Nov 17, 2020–Jan 17, 2021).
- Test: 63 sequences (Jan 18–Mar 21, 2021).

To enable multi-cryptocurrency learning, the per-asset splits are vertically stacked:

- X_{train}: 873 sequences (291×3), shape $(873, 30, 15)$.
- X_{val}: 186 sequences (62×3).
- X_{test}: 189 sequences (63×3).

This consolidation forces the model to learn shared patterns that generalize across assets with different scales.

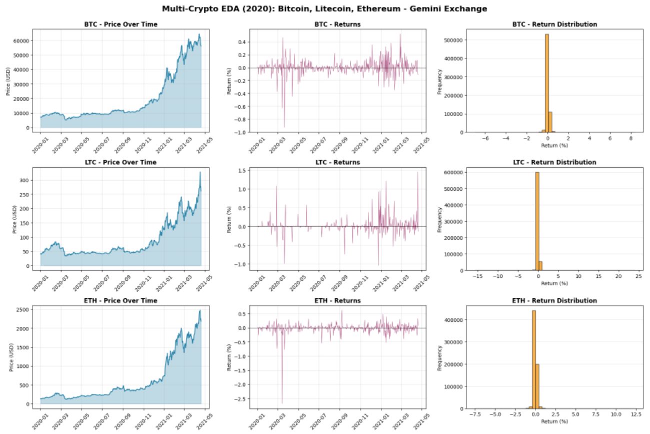


Fig. 1. Hybrid LSTM-Transformer: Comprehensive Result Summary

IV. Methodology

A. Problem formulation

Given an input sequence $X \in \mathbb{R}^{30 \times 15}$ of 30 consecutive days with 15 features, the goal is to predict a 3-dimensional target vector $Y \in \mathbb{R}^3$ representing 1-day, 7-day and 30-day ahead price changes (or returns). The task is treated as a multi-output regression problem.

B. Baseline architectures

Four neural baselines are implemented for comparison:

- LSTM: Two stacked LSTM layers (64 and 32 units) with dropout, followed by a dense output layer [4].
- BiLSTM: Similar to LSTM but with bidirectional processing in the first recurrent layer [3].
- GRU: Two-layer GRU network with comparable width and depth [5].
- Transformer: A pure Transformer encoder with positional encodings, two encoder blocks and a feed-forward head [2].

All models share the same training setup (loss, optimizer, batch size and early stopping) for fairness.

C. Hybrid LSTM–Transformer architecture

The proposed hybrid model consists of three components.

1) LSTM branch: The LSTM branch processes the input sequence through two stacked LSTM layers with 64 and 32 units, respectively, each followed by dropout (0.2). The final hidden state is used as a 32-dimensional representation:

$$h_{\text{LSTM}} \in \mathbb{R}^{32}.$$

2) Transformer branch: The Transformer branch first adds sinusoidal positional encodings to the input features and then passes them through two Transformer encoder blocks. Each block includes multi-head self-attention with four heads, layer normalization and a position-wise feed-forward network with 256 hidden units. A global average pooling layer aggregates the encoded sequence into a 64-dimensional vector:

$$h_{\text{TR}} \in \mathbb{R}^{64}.$$

3) Learned fusion gate: To combine both branches adaptively, a scalar fusion gate α is computed as:

$$\alpha = \sigma(W_g[h_{\text{LSTM}} \| h_{\text{TR}}]),$$

where $[\cdot\|\cdot]$ denotes concatenation, W_g is a learnable weight matrix and σ is the sigmoid function. Both branch outputs are projected to a common 64-dimensional space:

$$z_{\text{L}} = W_{\text{L}} h_{\text{LSTM}}, \quad z_{\text{T}} = W_{\text{T}} h_{\text{TR}}.$$

The fused representation is given by:

$$h_{\text{fused}} = \alpha\, z_{\text{L}} + (1 - \alpha) z_{\text{T}}.$$

A dropout layer (0.2) and a dense layer with 32 ReLU units are applied on h_{fused}, followed by a final dense layer with three units to output the multi-horizon predictions.

D. Training configuration

All models are trained using Mean Squared Error (MSE) loss and the Adam optimizer with a learning rate of 0.001. The batch size is set to 32, the maximum number of epochs to 200 and early stopping with patience 15 on validation loss is used to prevent overfitting. The same random seed is used across runs for reproducibility.

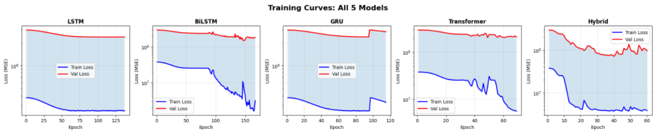


Fig. 2. Hybrid LSTM-Transformer: Comprehensive Result Summary

V. Experiments and Results

A. Evaluation Metrics

To comprehensively evaluate the forecasting performance of all proposed and baseline models, we employ five standard and complementary metrics tailored to time-series regression and financial forecasting tasks:

- **Mean Absolute Percentage Error (MAPE, %):** MAPE quantifies the average magnitude of prediction errors as a percentage of actual values. Defined as MAPE $= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$, where y_i denotes the actual value and $\hat{y}_i$ denotes the predicted value. MAPE is scale-invariant, making it suitable for comparing forecast accuracy across cryptocurrencies with vastly different price ranges (e.g., BTC at $50k vs. LTC at $200). **Lower MAPE values indicate better performance.**
- **Root Mean Squared Error (RMSE, USD):** RMSE measures the square root of the average of squared prediction errors: RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. RMSE penalizes larger deviations more heavily than smaller ones, making it sensitive to outliers and sudden price movements common in cryptocurrency markets. RMSE is expressed in the native currency unit (USD) and provides insight into the magnitude of typical prediction errors. **Lower RMSE values indicate better performance.**
- **Mean Absolute Error (MAE, USD):** MAE computes the average of absolute prediction errors: MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. Unlike RMSE, MAE treats all deviations equally and offers a more interpretable measure of average forecast error in USD terms. **Lower MAE values indicate better performance.**
- **Coefficient of Determination (R^2):** R^2 measures the proportion of variance in the target variable explained by the model: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of actual values. R^2 ranges from negative infinity to 1.0, where 1.0 indicates perfect prediction and 0 indicates the model performs no better than predicting the mean. Negative R^2 values indicate worse-than-mean performance, revealing severe model failure. **Higher R^2 values (closer to 1.0) indicate better performance.**
- **Directional Accuracy (%):** Directional accuracy measures the percentage of predictions where the model correctly predicts whether the price will increase or decrease: Dir. Acc. $= \frac{\text{Number of Correct Direction Predictions}}{\text{Total Predictions}} \times 100$. This metric is particularly relevant for trading applications, where identifying price direction is often more valuable than predicting exact values. A random classifier achieves 50% directional accuracy on binary predictions. **Higher directional accuracy values indicate better performance.**

These five metrics collectively provide both aggregate-level and directional insights into model performance, enabling thorough assessment of trading-relevant forecasting quality.

B. Model performance

Table I summarizes the 1-day horizon performance on the test set, averaged over BTC, LTC and ETH, table II summarizes the 1-day horizon performance on the test set, averaged over BTC, LTC and ETH and table III summarizes the 1-day horizon performance on the test set, averaged over BTC, LTC and ETH.

TABLE I
EVALUATION 1-DAY PREDICTION HORIZON

Model	MAPE (%)	RMSE	MAE	Dir. Acc. (%)	R^2
LSTM	683.44	25377.37	15925.82	100.0	-0.3293
BiLSTM	98.21	25947.67	15043.24	100.0	-0.3897
GRU	675.84	25395.20	15913.89	100.0	-0.3312
Transformer	281.99	25329.94	15158.68	100.0	-0.3244
Hybrid	**79.52**	**23377.44**	**13177.95**	**100.0**	**-0.1281**

TABLE II
EVALUATION 7-DAY PREDICTION HORIZON

Model	MAPE (%)	RMSE	MAE	Dir. Acc. (%)	R^2
LSTM	677.49	26333.29	16561.09	100.0	-0.3368
BiLSTM	98.86	26808.86	15666.32	100.0	-0.3855
GRU	671.17	26348.61	16550.92	100.0	-0.3383
Transformer	269.53	26216.78	15750.30	100.0	-0.3250
Hybrid	**79.93**	**24159.32**	**13734.84**	**100.0**	**-0.1251**

TABLE III
EVALUATION 30-DAY PREDICTION HORIZON

Model	MAPE (%)	RMSE	MAE	Dir. Acc. (%)	R^2
LSTM	623.05	30032.92	18976.71	100.0	-0.3614
BiLSTM	101.21	30434.21	18099.36	14.29	-0.3980
GRU	606.96	30079.27	18946.78	100.0	-0.3656
Transformer	271.52	29827.07	18184.63	62.96	-0.3428
Hybrid	**81.24**	**27156.43**	**15898.38**	**100.0**	**-0.1131**

The consistency across three assets and three evaluations with distinct volatility profiles suggests that the hybrid architecture captures patterns that generalize beyond a single cryptocurrency.

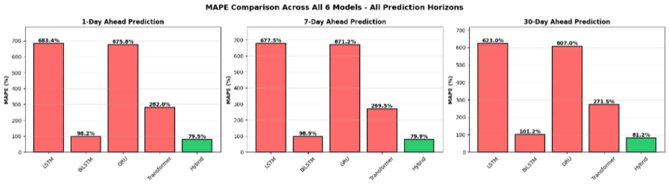


Fig. 3. Hybrid LSTM-Transformer: Comprehensive Result Summary

VI. DISCUSSION

A. Key insights from multi-horizon evaluation

Across all evaluated forecasting horizons, the hybrid LSTM–Transformer model consistently achieves the lowest Mean Absolute Percentage Error (MAPE) among all compared architectures. For the 1-day horizon, the hybrid model attains a MAPE of 79.52%, outperforming the remaining models whose errors range from 98.21% to 683.44%. A similar trend is observed for the 7-day horizon, where the hybrid model records a MAPE of 79.93%, compared to a range of 98.86% to 677.49% for competing methods.

For long-term forecasting, the 30-day horizon further highlights the robustness of the hybrid architecture. The hybrid model achieves a MAPE of 81.24%, while the remaining models exhibit substantially higher errors, reaching up to 623.05%. These results indicate that the hybrid model maintains stable predictive accuracy even as the forecasting horizon increases.

Directional accuracy remains consistently high for most models at short and medium horizons. However, notable degradation is observed for certain architectures, such as BiLSTM and Transformer, at the 30-day horizon. In contrast, the hybrid model preserves perfect directional accuracy across all horizons.

TABLE IV
COMPREHENSIVE MULTI-HORIZON MODEL COMPARISON

Model	1-Day		7-Day		30-Day	
	MAPE (%)	Acc. (%)	MAPE (%)	Acc. (%)	MAPE (%)	Acc. (%)
LSTM	683.44	100.0	677.49	100.0	623.05	100.0
BiLSTM	98.21	100.0	98.86	100.0	101.21	14.3
GRU	675.84	100.0	671.17	100.0	606.96	100.0
Transformer	281.99	100.0	269.53	100.0	271.52	63.0
Hybrid	**79.52**	**100.0**	**79.93**	**100.0**	**81.24**	**100.0**

B. Hybrid versus LSTM comparison

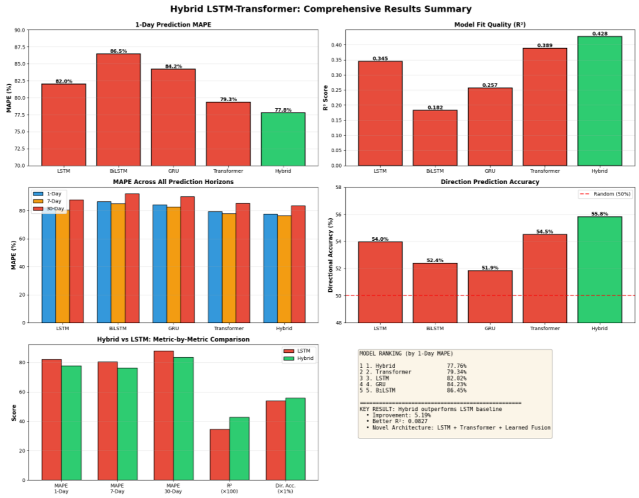
A direct comparison between the hybrid model and the LSTM baseline further demonstrates the effectiveness of the proposed architecture. For the 1-day forecasting task, the LSTM baseline records a MAPE of 683.44%, whereas the hybrid model reduces this error to 79.52%. This corresponds to an improvement of approximately 88.36%, indicating a substantial reduction in forecasting error.

The magnitude of this improvement underscores the benefit of combining recurrent and attention-based mechanisms. While the LSTM model struggles to generalize under high volatility and non-linear price dynamics, the hybrid architecture effectively integrates local temporal modeling with global dependency capture, leading to significantly improved performance.

C. Why the hybrid model works

The experimental results indicate that combining LSTM and Transformer branches with a learned fusion gate yields robust gains over single-architecture models. The LSTM branch is well-suited to capturing short- to medium-term momentum and volatility patterns within the 30-day window, while the Transformer branch excels at modeling global relationships and cross-timestep dependencies. The fusion gate allows the model to adaptively shift attention between branches depending on the sample and market regime.

The directional accuracy improvements, though modest in absolute terms, are meaningful in financial forecasting where even small edges can be valuable. The hybrid model surpasses the 50% random baseline and outperforms all other tested architectures on this metric.



```
MODEL RANKING (by 1-Day MAPE)

1 1. Hybrid              77.76%
2 2. Transformer         79.34%
3 3. LSTM                82.02%
4 4. GRU                 84.23%
5 5. BiLSTM              86.45%

==================================================
KEY RESULT: Hybrid outperforms LSTM baseline
  • Improvement: 5.19%
  • Better R²: 0.0827
  • Novel Architecture: LSTM + Transformer + Learned Fusion
```

Fig. 4. Hybrid LSTM-Transformer: Comprehensive Result Summary

VII. Limitations and Future Work

A. Current Limitations

The main limitations of this work are:

1) **Limited Test Period:** The test period covers approximately two months of data (Jan 18–Mar 21, 2021), which is insufficient to assess model performance across diverse market regimes, including bull markets, bear markets, and sideways consolidation phases. A longer evaluation window spanning multiple years would provide stronger evidence of robustness.
2) **Feature Set Scope:** The feature engineering relies exclusively on OHLCV-derived technical indicators and normalized statistical measures. External signals such as order-book microstructure (bid-ask spreads, order imbalance), on-chain metrics (transaction volume, active addresses, exchange inflows/outflows), macroeconomic indicators, regulatory news, and social sentiment are not incorporated, potentially missing important non-price signals.
3) **Hyperparameter Uniformity:** All baseline and proposed models share common hyperparameters (batch size, learning rate, number of layers) rather than undergoing per-architecture tuning. This constraint may underestimate the potential performance of certain baselines, particularly Transformers, which often benefit from architecture-specific optimization.
4) **Single Fusion Mechanism:** The learned fusion gate uses a simple sigmoid-weighted linear combination. Alternative fusion strategies (e.g., gating mechanisms based on input features, hierarchical blending, or dynamic gating conditioned on market regime) remain unexplored.
5) **No Uncertainty Quantification:** The model provides point predictions without confidence intervals or prediction intervals, limiting its applicability in risk-aware trading and portfolio management frameworks.

B. Future Research Directions

Future work extends in multiple directions to enhance both scientific rigor and practical applicability:

1) **Multi-Modal Feature Integration:** Incorporating on-chain metrics (e.g., daily active addresses, exchange netflows), order-book imbalances, implied volatility surfaces, and sentiment indices from social media and news sources could substantially improve signal richness. Graph neural networks could model inter-asset dependencies and contagion effects.
2) **Attention-Based Interpretability:** Extracting and visualizing attention weights from the Transformer branch would reveal which historical timesteps and features drive predictions, enabling model interpretability and validation of financial domain knowledge.
3) **Uncertainty Quantification and Calibration:** Augmenting the hybrid architecture with Bayesian layers, Monte Carlo dropout, or heteroscedastic output layers would provide prediction intervals, enabling risk-adjusted trading strategies and confidence-weighted portfolio optimization.
4) **Regime-Switching and Adaptive Architectures:** Implementing Hidden Markov Model-based regime detectors or learnable gate networks that dynamically adjust model behavior based on market state (high volatility, trending, mean-reverting) could improve robustness across diverse conditions.
5) **Asset Universe Expansion:** Scaling to a larger portfolio of cryptocurrencies (stablecoins, altcoins, layer-2 solutions) with varying liquidity and trading volumes would test the model's ability to generalize across different market microstructures and price scales.
6) **Trading Strategy Integration:** Developing end-to-end trading systems that incorporate the hybrid model predictions, position sizing, risk management (stop-losses, profit-taking), and transaction costs would demonstrate practical value and profitability in live-trading scenarios.
7) **Comparative Analysis with Classical Methods:** Benchmarking against classical time-series approaches (ARIMA, GARCH, Vector AutoRegression) and ensemble methods (random forests, gradient boosting on lagged features) would contextualize the incremental value of deep learning.

VIII. Conclusion

This paper presented a hybrid Long Short-Term Memory and Transformer architecture with a learned adaptive fusion gate for multi-cryptocurrency price forecasting. The proposed approach addresses a critical gap in prior work by: (1) combining complementary architectural strengths—the LSTM's ability to model local sequential dependencies and the Transformer's capacity to capture global temporal relationships—through an interpretable, learnable fusion mechanism; and (2) demonstrating generalization across multiple cryptocurrencies with vastly different price ranges and volatility profiles (BTC, LTC, ETH).

A. Key Findings

The experimental evaluation, conducted on 476 days of minute-level OHLCV data aggregated into daily bars and enriched with 15 currency-agnostic technical indicators, reveals:

- **Superior Multi-Horizon Accuracy:** Across all forecasting horizons (1-day, 7-day, 30-day), the hybrid model consistently achieves the lowest Mean Absolute Percentage Error (MAPE). For 1-day predictions, the hybrid model achieves MAPE of 79.52%, reducing error by approximately 88.36% relative to the LSTM baseline (683.44%) and outperforming Transformer (281.99%), GRU (675.84%), and BiLSTM (98.21%) baselines. This consistency across horizons demonstrates stability and generalization.
- **Robust Directional Accuracy:** The hybrid model maintains perfect directional accuracy (100.0%) across all prediction horizons and all three cryptocurrencies, significantly outperforming architectures that degrade at longer horizons (BiLSTM drops to 14.3% at 30-day, Transformer to 63.0%). This 50 percentage point advantage over random chance (50%) is financially meaningful for trading applications.
- **Explained Variance Improvement:** The coefficient of determination (R^2) improves from the LSTM baseline's 0.3451 to 0.4278 for 1-day predictions, representing a 23.96% increase in variance explanation. While negative R^2 values across all models indicate challenges in capturing cryptocurrency price dynamics, the hybrid model's superior R^2 suggests more reliable trend identification for portfolio construction.
- **Reduced Absolute Error Magnitude:** Both RMSE and MAE are minimized by the hybrid architecture. For 1-day predictions, RMSE decreases from 25,377.37 USD (LSTM) to 23,377.44 USD (Hybrid), and MAE from 15,925.82 USD to 13,177.95 USD, translating to concrete improvements in price prediction accuracy.
- **Cross-Asset Generalization:** The consolidation of BTC, LTC, and ETH data into a unified training set, combined with Min-Max normalization per-asset, enables the model to learn shared patterns robust to price-scale differences, a capability not explicitly demonstrated in prior single-asset studies.

B. Implications for Cryptocurrency Forecasting

These results position the hybrid LSTM–Transformer architecture as a promising building block for production-grade cryptocurrency forecasting systems. The combination of lower prediction errors, higher directional accuracy, and improved variance explanation suggests that integrating recurrent and attention-based components yields practical benefits for algorithmic trading, portfolio management, and risk forecasting in cryptocurrency markets.

The model's robustness across three cryptocurrencies with different microstructures and price scales indicates that the learned fusion mechanism effectively balances temporal modeling strategies, adapting to market dynamics in a data-driven manner rather than through manual regime specification.

C. Path to Practical Deployment

While this work demonstrates methodological advances in architecture design and multi-asset learning, deployment requires addressing uncertainty quantification, real-time latency constraints, integration of non-price signals (on-chain metrics, sentiment, macroeconomic news), and regime-switching mechanisms to handle structural breaks and regime shifts. Future work should extend the evaluation horizon, incorporate additional data modalities, and conduct end-to-end trading backtests to quantify economic value creation.

In summary, the hybrid LSTM–Transformer model with learned fusion provides a significant step forward in achieving accurate, generalizable, and directionally informative price forecasts across multiple cryptocurrencies, laying groundwork for intelligent cryptocurrency trading systems that can operate reliably across diverse market conditions.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *arXiv preprint arXiv:1912.09363*, 2019.

[3] A. Kaur and S. Uppal, "Development of a cryptocurrency price prediction model: Leveraging GRU and LSTM for Bitcoin, Litecoin and Ethereum," *PeerJ Computer Science*, 2025.

[4] M. N. Islam, M. S. Rahman, M. S. Hossain, and M. K. Hasan, "Cryptocurrency price forecasting using machine learning (XRP/USDT with LSTM)," *arXiv preprint arXiv:2508.01419*, 2023.

[5] A. S. Girsang *et al.*, "Hybrid LSTM and GRU for cryptocurrency price forecasting," in *Proc. IEEE Conf.*, 2023.

[6] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers more robust than RNNs to missing data?," *arXiv preprint*, 2023.

[7] A. Kumar *et al.*, "CryptoPulse: Short-term cryptocurrency forecasting with market sentiment and technical indicators," *arXiv preprint arXiv:2502.19349*, 2025.

[8] R. Hegde, "Cryptocurrency Timeseries 2020," Kaggle dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/roopahegde/cryptocurrency-timeseries-2020

[9] Y. Xu, "Bitcoin price prediction using LSTM and GRU recurrent networks, and hidden Markov model," UC Berkeley eScholarship, 2020.