Experimental Study of Deep Learning Models for Short-Term Crypto Price Forecasting

Waseem Mohamad
Student ID: g202401140
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Short term cryptocurrency forecasting remains a challenging problem due to the volatility, nonlinearity, and sentiment driven behavior of digital asset markets. Although recent studies highlight the potential of deep learning models for capturing complex temporal patterns, many rely on proprietary or restricted data sources, creating a significant reproducibility gap. This work addresses that limitation by developing a fully transparent forecasting framework using only publicly accessible datasets for Bitcoin and Ethereum. The objective is to evaluate how well a diverse set of neural architectures can predict next day price movements, local extrema, and closing prices under realistic academic constraints. The methodology integrates multimodal inputs including historical OHLCV data, technical indicators, Google Trends signals, macroeconomic variables, and news derived features into a unified preprocessing and modeling pipeline. Several architectures are tested under a walk forward expanding window evaluation, including MLP, LSTM, LSTM Deep, CNN-LSTM, and transformer based models for both regression and classification tasks. The empirical results show that simpler architectures, particularly MLP and CNN-LSTM, outperform deeper LSTM stacks and transformer models across all forecasting targets. The findings imply that, for short horizon crypto forecasting with limited data density, model stability and appropriate feature conditioning are more critical than architectural complexity. The study provides a reproducible benchmark and highlights directions for improving transformer performance through richer data and hybrid designs.**

***Index Terms*—Cryptocurrency, Forecasting, Deep Learning, Time Series Analysis, MLP, LSTM, Transformer Models, Financial Prediction, News Sentiment.**

I. Introduction

A. Background and Significance

Cryptocurrency markets have become a significant component of the global financial ecosystem, drawing considerable interest from researchers, traders, and institutional participants. Their defining characteristics persistent volatility, rapid information diffusion, and heightened sensitivity to investor sentiment distinguish them from traditional financial assets. For example, Bitcoin frequently exhibits far greater short-term volatility than benchmark indices such as the S&P 500 [1], underscoring the challenges involved in modeling cryptocurrency price dynamics. These markets operate continuously across decentralized exchanges, producing an uninterrupted stream of price, volume, and behavioral data that is well suited for data-driven forecasting approaches.

A key feature of cryptocurrencies is the diverse set of factors influencing price formation. Empirical studies report that classical macroeconomic indicators often exhibit only weak or inconsistent relationships with crypto asset valuations, whereas sentiment-driven and attention-based signals frequently demonstrate more immediate predictive relevance [1]. Public attention metrics, such as Google Trends search volume, and sentiment extracted from news headlines, online discussion forums, and micro-blogging platforms have been shown to lead price movements by several days, reflecting the behavioral and speculative nature of cryptocurrency markets. These observations position cryptocurrencies as an ideal testbed for machine learning models capable of integrating heterogeneous, non-stationary, and rapidly evolving data sources.

Traditional econometric approaches struggle with the nonlinear structure and regime shifts present in crypto time series [5]. In contrast, recent advances in machine learning and deep learning offer tools to capture complex temporal dependencies. Architectures such as Long Short-Term Memory (LSTM) networks, convolutional sequence models, and transformer-based frameworks have demonstrated promising performance in financial prediction tasks, particularly when combined with engineered technical indicators or alternative data modalities. In parallel, natural language processing (NLP) techniques, especially modern transformer models, have shown strong capabilities in extracting high-quality sentiment representations from unstructured text, enabling richer modeling of market psychology.

Within this evolving landscape, short-term forecasting of Bitcoin and Ethereum remains both practically relevant and methodologically challenging. The present study is situated at the intersection of these developments, focusing on evaluating deep learning models for daily-level crypto forecasting under realistic data constraints. By leveraging publicly accessible price histories, technical indicators, Google News headlines, and Google Trends data, the research aims to assess how far modern forecasting techniques can be pushed without reliance on paywalled APIs or proprietary sentiment datasets. This context highlights both the significance and the practical importance of developing reproducible, multimodal forecasting pipelines suitable for academic research and real-world analysis alike.

B. Challenges in Current Techniques

Although cryptocurrency forecasting has attracted extensive scholarly attention, significant methodological and practical obstacles remain. Classical econometric models struggle with the defining properties of cryptocurrency markets' extreme volatility, abrupt structural shifts, and strong nonlinear dependencies. Standard approaches such as ARIMA or GARCH cannot adequately accommodate regime changes or the complex interactions between behavioral, technical, and external drivers of crypto asset prices [5]. Even within machine learning research, many studies rely on limited feature sets centered around price and volume, lack rigorous temporal cross-validation, or overlook the importance of modeling long range dependencies.

Another persistent limitation concerns the treatment of sentiment. While investor sentiment plays a central role in cryptocurrency price dynamics, much of the literature continues to use dictionary-based sentiment analysis, which is known to miss contextual nuance, sarcasm, domain specific terminology, and subtle variations in tone (Liu, 2012). More advanced transformer-based methods offer richer linguistic representations, but these techniques are computationally demanding and often require large volumes of high-quality text data resources that are not always accessible.

Data accessibility presents an even more fundamental challenge. Many state-of-the-art studies rely on extensive social media and news datasets obtained through paid or deprecated APIs. Twitter's transition to a monetized API, Reddit's API pricing changes, and the restrictions imposed by CryptoCompare and CoinGecko on historical data have collectively made it difficult to reproduce earlier research pipelines. The motivating journal article for this project, for example, utilized large-scale sentiment streams from Twitter, Reddit, and specialized news outlets' data that are no longer freely obtainable. The broader implication is a reproducibility bottleneck; researchers without institutional access or paid APIs cannot replicate or extend existing methodologies, limiting the field's transparency and comparability.

Additional challenges arise on the modeling side. Deep neural networks can easily overfit noisy financial data without careful regularization and hyperparameter tuning. Training complex architectures, particularly transformer-based sequence models, requires significant computational resources and expertise. Furthermore, combining heterogeneous inputs such as price histories, technical indicators, Google Trends signals, and textual sentiment introduces additional complications in data preprocessing, temporal alignment, and feature integration.

A conceptual gap also persists in the choice of predictive targets. Most studies emphasize regression-based forecasting of next-day returns, despite the well-documented noise present in daily price fluctuations. Emerging work suggests that discrete formulations, such as direction classification or local extrema detection, may produce more stable and interpretable signals. However, systematic comparisons across these target types remain scarce, particularly under controlled experimental settings.

Taken together, the shortcomings in data availability, sentiment extraction, modeling stability, and target formulation highlight why building robust and reproducible short-term cryptocurrency forecasting models remains an open challenge. These limitations motivate the present study's emphasis on publicly accessible datasets, careful model evaluation, and comparison across multiple prediction targets.

C. Problem Statement

Short term cryptocurrency forecasting remains a difficult task due to the nonlinear, highly volatile, and sentiment driven nature of crypto markets. While prior work demonstrates that deep learning can extract predictive patterns from financial and behavioral data, many existing studies rely on proprietary, paid, or recently restricted datasets, particularly from Twitter, Reddit, CryptoCompare, and CoinGecko. These data accessibility barriers undermine reproducibility and prevent broader use of advanced sentiment-based forecasting pipelines.

Given these constraints, the central problem addressed in this study is how can deep learning models be designed, trained, and evaluated to forecast short term cryptocurrency price movements using only publicly accessible datasets, while still capturing the predictive signals traditionally derived from sentiment, behavioral, and macro-financial indicators. Current literature presents several unresolved gaps that motivate this work; heavy dependence on inaccessible or paywalled sentiment sources, limited comparison across different prediction targets, and under exploration of diverse deep learning architectures in a unified framework.

Many influential studies extract sentiment directly from proprietary APIs of social networks. Replicating these pipelines is now infeasible due to API monetization, data-access restrictions, and the lack of open historical sentiment archives. Similarly, most research papers focus solely on continuous regression of next-day returns, despite evidence that classification-based targets such as direction forecasts or local extrema, may reduce noise and yield stronger signals. Systematic evaluation across these alternatives remains scarce. Furthermore, while LSTM-based methods are common, fewer studies conduct controlled comparisons between non-sequential models (MLPs), hybrid architectures (CNN-LSTM), and transformers adapted for time-series forecasting.

To address these limitations, the present study constructs a fully reproducible forecasting pipeline using publicly accessible data only, i.e. historical BTC and ETH prices from Kaggle, technical indicators derived from these prices, Google News headlines as a proxy for sentiment flow, Google Trends indices reflecting public interest, and macro-financial variables from Yahoo Finance (S&P 500, VIX, and gold). These inputs are integrated into a unified preprocessing framework and evaluated across multiple neural architectures, including multilayer perceptrons, deep LSTMs, CNN-LSTM hybrids, and a vanilla transformer model.

The research aims to determine which combinations of models, features, and target formulations price direction classification, regression, or local extrema detection offer the most reliable and accurate short-term forecasts for Bitcoin and Ethereum under realistic, non proprietary data constraints.

D. Objectives

The main goal of this project is to examine the capabilities and limitations of DL models for short-term cryptocurrency forecasting under realistic data availability constraints. The specific objectives are as follows:

- To construct a reproducible and publicly accessible multimodal dataset combining price history, technical indicators, Google News headlines and Google search trends.
- To implement and evaluate diverse DL architectures
- To compare forecasting performance across multiple target formulations including regression on next-day log-returns, binary direction classification and local extrema detection for 7, 14, and 21 day windows.
- To assess how model complexity and sequential structure affect predictive stability, especially during high volatility periods.
- To derive insights into which classes of models are most suitable for short-term forecasting tasks, given constraints on data quality and availability.

E. Scope of Study

This paper focuses exclusively on short-term forecasting, i.e. daily-level predictions for BTC and ETH, using publicly accessible data sources. The study does not directly incorporate paid APIs, proprietary sentiment feeds, or close-source blockchain analytics platforms. All models are implemented in Python using reproducible pipelines on Google Colab. The scope includes supervised DL architectures for time-series prediction, classical train and test protocols adapted for temporal data, and evaluation using regression and classification metrics. Furthermore, it includes the exploratory visualization of model predictions and residual dynamics along with the comparison across target formulations including local extrema. Generally, this paper draws conceptual inspiration from the work done by Gurgul et. al, while adapting the methodological emphasis to datasets that remain openly accessible.

II. Literature Review

A. Overview of Existing Techniques

Cryptocurrency forecasting has evolved through several methodological phases, reflecting the increasing complexity of market behavior and the increasing availability of diverse data sources. Early attempts relied on classical statistical models such as ARIMA, GARCH, and VAR. While useful for capturing linear relationships, these models perform poorly when confronted with the extreme volatility, non stationarity, and nonlinear dynamics characteristic of digital asset markets. The literature consistently shows that traditional approaches cannot adapt to sudden structural breaks or high-frequency fluctuations typical of BTC and ETH price movements.

As limitations of classical econometrics became evident, ML methods gained traction. Algorithms such as Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting (GB), and multilayer perceptrons (MLP) demonstrated improved performance by modeling nonlinear patterns without rigid distributional assumptions. However, these models still rely heavily on engineered features and struggle to represent long-term temporal dependencies in sequential financial data.

DL has become the dominant paradigm in financial time-series modeling due to its ability to learn hierarchical representations from raw sequences. Recurrent neural network (RNN) variants, particularly LSTM and GRU, have been widely adopted for forecasting cryptocurrency prices, motivated by their capacity to capture long range temporal structure. Numerous studies have shown these architectures outperform classical ML approaches, especially when datasets are noisy, stochastic, and exhibit complicated dynamics.

Hybrid architectures blending CNNs with LSTMs emerged to capture both local and long-term dependencies in temporal data. CNN components extract short term patterns while LSTM layers model broader structural evolution. More recently, transformer based architectures have gained visibility due to their parallelization advantages and ability to model distant temporal interactions via self attention. Several works, including those in [1], [2] , and [7], show that transformer based models outperform classical RNN based frameworks, especially in long-horizon forecasting tasks.

The research landscape has further expanded toward multimodal forecasting frameworks that incorporate price data, blockchain indicators, Google Trends, sentiment signals from news or social media, and even cryptographic enhancements for data privacy. For example, [3] introduces a forecasting pipeline that integrates deep learning with post quantum encryption, demonstrating the increasing practical scope of DL models.

Overall, existing techniques span a continuum from simple linear models to highly sophisticated hybrid networks integrating CNNs, RNNs, transformers, on-chain metrics, and external indicators. The trend clearly favours deep neural architectures capable of handling nonlinear, multiscale, and multimodal dynamics.

B. Related Work

The most relevant study motivating the present research is documented in the work done by Gurgul et al. [1], which introduces a hybrid attention–correlation transformer for cryptocurrency forecasting across multiple time scales. That study argues that transformers can better capture long-range dependencies while handling parallel computation efficiently. The paper also evaluates transformer variants such as Autoformer, Informer, and FEDformer, showing their competitive performance relative to RNN-based models. However, the methodology relies heavily on extensive historical datasets not publicly accessible today, including data from CryptoCompare and other market APIs.

Another substantial body of related work is represented by Almusfar et al. [4], which provides a comparative analysis of ensemble learning and deep learning models for cryptocurrency forecasting. The study finds that gated recurrent units (GRU), simple RNNs, and LightGBM outperform both statistical baselines and certain deep learning architectures. The results underscore the importance of algorithm selection and highlight that performance varies across different cryptocurrencies, reflecting heterogeneous market characteristics.

The broader landscape of deep learning in financial forecasting is systematically reviewed in [5], which covers 187 studies published between 2020–2024. This comprehensive survey identifies LSTM-based models and CNN–LSTM hybrids as dominant approaches while also noting the emergence of transformers, attention-augmented models, and multimodal systems. The review also outlines the methodological challenges, data scarcity, non stationarity, limited robustness under extreme conditions, several of which directly motivate the experimental design of the present study.

Deep learning architectures tailored to specific financial or blockchain-related applications are also relevant. For instance, [1] presents a hybrid attention-enhanced forecasting model and highlights the need to capture high-frequency patterns effectively. Similarly, [7] proposes a transformer-CNN hybrid for modeling Bitcoin energy consumption, demonstrating the importance of advanced feature engineering and long-range dependency modeling in complex blockchain-related time series.

In addition, several papers emphasize multimodal learning. Giantsidi et al. focus on integrating price data with sentiment or on-chain metrics using advanced attention mechanisms [5]. These studies highlight the increasing role of alternative data sources, news, Google Trends, social sentiment, and blockchain metrics, in improving prediction accuracy. Taken together, the related literature underscores three persistent themes; superiority of deep learning over classical statistical methods, growing relevance of transformer-based architectures, and importance of multimodal and hybrid feature integration.

C. Limitations in Existing Approaches

Despite the extensive research in cryptocurrency forecasting, several critical limitations remain unresolved, as evidenced across different papers discussed above. Many state-of-the-art forecasting pipelines depend on proprietary APIs such as Twitter's premium endpoints, Reddit API datasets, or commercial feeds from CryptoCompare and CoinGecko. As documented in paper [1], the original methods used large scale sentiment streams from social media platforms, data that is no longer easily accessible due to paywalls, API restrictions, or rate limitations. This creates a major reproducibility barrier in academic research.

Although transformers demonstrate exceptional ability to model long-range dependencies, several works including the base paper [1] acknowledge their difficulty in modeling high-frequency, short-term fluctuations. Few studies explicitly evaluate short horizon tasks such as next-day direction, local extrema detection, or regime classification. This leaves a gap in evaluating architectures under the precise conditions most relevant for real-world traders. Deep networks, especially LSTM based models, can easily overfit noisy financial data. Almusfar et al. highlight that even advanced architectures suffer from instability when exposed to sudden market shocks [4]. Many studies do not conduct rigorous temporal validation, nor do they address generalization across different market regimes.

Although sentiment and behavioral indicators are important drivers of cryptocurrency movements, their integration remains inconsistent. Some studies such as [7], incorporate news signals via transformer based NLP models, whereas others rely solely on price data. Moreover, timestamp alignment between sentiment and price is rarely handled systematically. Across papers particularly noted in [3] evaluation metrics, train test splits, and forecasting horizons vary significantly. This makes cross-paper comparisons difficult and obscures the true effectiveness of different model families.

Transformer-based models, especially hybrid ones, require substantial computational power. Papers such as [5] emphasize the significant resource demands of attention-based architectures, which limits their accessibility for smaller research groups and realtime trading systems. Given the diversity of preprocessing techniques (detrending, normalization, feature engineering), model choices (LSTM, CNN-LSTM, attention mechanisms), and external data inputs, many studies fail to provide replicable pipelines. This issue is explicitly noted in paper [4] and is one of the primary motivations behind your project's focus on publicly accessible data.

III. Proposed Methodology

A. Existing Model and Challenges

The original model that motivated this project is based on the transformer-driven hybrid architecture presented in [1], which combines an attention–correlation mechanism with temporal embeddings to improve long-range dependency modeling in cryptocurrency price forecasting. That framework integrates multiple feature families, price history, blockchain indicators, macroeconomic series, and sentiment signals mined from Reddit, Twitter, and CryptoCompare APIs, into a unified forecasting system. According to Gurgul et al. [1], this architecture demonstrated strong performance across multiple forecasting horizons, benefitting from transformers' ability to capture multiscale patterns and nonlinear temporal behavior.

However, reproducing this pipeline poses several practical challenges. First, the original dataset is not publicly accessible; social media text streams, historical sentiment scores, and paid APIs used in [1] require premium endpoints or institutional access. Twitter's API pricing changes, Reddit's rate limitations, and CryptoCompare's historical data restrictions make these data sources infeasible for academic replication. Second, the original architecture is computationally heavy. Its multi stage attention correlation blocks require substantial GPU resources and extensive hyperparameter tuning, making full reproduction

difficult within the constraints of a course project. Third, the model relies on multimodal signals not easily available today, limiting its reproducibility and undermining empirical comparability.

These constraints necessitated a shift toward a fully open source, fully reproducible alternative pipeline. The objective became not to replicate the original model verbatim but instead to construct a methodologically comparable deep learning framework using public data sources while preserving the core conceptual motivations of the transformer based approach.

B. Proposed Enhancements

In response to the reproducibility issues described above, this study proposes a redesigned methodology that preserves the conceptual strengths of the original model but adapts it to a publicly accessible environment. Several enhancements distinguish the proposed framework from the original pipeline. The first of them was to fully open source data ecosystem. All inputs are obtained from publicly available sources, ensuring complete reproducibility. Data include; daily OHLCV and market cap for BTC or ETH from Kaggle, Google News headlines scraped via a custom RSS pipeline in our functions related to data acquisition, Google Trends indices collected through a robust multi window query strategy, and macroeconomic variables (S&P 500, VIX, Gold) fetched using Yahoo Finance. This eliminates reliance on proprietary APIs and enables transparent, academic grade evaluations.

Second enhancement is to expand the feature engineering. The proposed system introduces a substantially richer feature set than the original baseline. The system pipeline generates; multi horizon log returns, volatility estimators, ATR, OBV, and Bollinger band based indicators, multiple moving-average ratios, sentiment proxies from Google News (bullish and bearish keyword scoring), Google Trends velocity and z-score transformations, macroeconomic z-scores and return features. This expanded feature space enables learning of short term dynamics without reliance on unavailable social media sentiment.

Next enhancement and one of the most critical one is to use a framework that predicts multi-targets. Unlike Gurgul et al. work [1], which focuses primarily on continuous value forecasting, the proposed pipeline evaluates three distinct learning objectives. Firstly, next-day direction classification (binary i.e. up and down). Secondly, local extrema detection (7-day rolling maxima and minima), and the third one is next-day close-price regression. This extension enables more comprehensive evaluation of how architectures respond to various market representations.

Moving forward from the data perspective, diversification of models was required to provide a controlled comparison of sequential and non-sequential signal extraction. Therefore, multiple architectures were implemented such as Multilayer Perceptron (MLP), standard LSTM, deep two-layer LSTM, CNN-LSTM hybrid, and custom implementations for transformer classifier and transformer regressor. This broad spectrum of models far more diverse than in paper [1] allows us to quantify the incremental value of sequence depth, convolutional filtering, and self attention.

Lastly, The pipeline adopts a multi split walk forward evaluation with non overlapping validation windows, producing realistic, time consistent performance estimates. This addresses an important limitation identified in [4] regarding the lack of rigorous temporal validation in existing studies. Collectively, these enhancements produce a reproducible, multimodal forecasting framework aligned with the conceptual motivations of the original model but adapted to real world data constraints.

C. Algorithm and Implementation

The system is implemented in Python using TensorFlow (Keras), Pandas, NumPy, and scikit-learn. The implementation consists of three main stages namely; data acquisition, feature engineering, and model training.

1) Data Acquisition Pipeline
    a) Kaggle OHLCV datasets for BTC and ETH are loaded and transformed into consistent daily time series.
    b) Google News scraping: A custom RSS scraping engine (GoogleNewsRSS class) retrieves daily news headlines for BTC and ETH. The script handles request errors, time-zone normalization, duplicate removal, and XML parsing.
    c) Google Trends collection: A segmented 250-day windowing strategy is used to bypass API rate limits and retrieve long-term daily Trends indices for multiple keywords (“bitcoin”, “ethereum”, “cryptocurrency”, “blockchain”, “investing”).
    d) Yahoo Finance macro series: Daily S&P 500, VIX, and gold prices are retrieved and aligned to the crypto dataset.
2) Feature Engineering and Dataset Construction
    a) Price derived features: log returns, volatility (7-day, 30-day), range-based measures, moving averages, ATR, RSI, MACD, stochastic oscillator, OBV, Williams %R, KDJ, PSAR.
    b) Local-extrema targets: computed via centered rolling maxima and minima (add_local_extrema_targets).
    c) Direction target: next-day binary up/down indicator (add_nextday_target).
    d) News sentiment proxies: simple keyword based scoring and headline statistics (count, length, sentiment).
    e) Google Trends features: 7-day differences and 90-day z-score transformations for each keyword.
    f) Macroeconomic features: z-score standardized series for S&P 500, VIX, and gold; log returns where relevant.
3) Data Splitting and Standardization: Temporal splits are generated using a function with 3 folds, 180-day validation windows and expansion based training sets. All features are standardized, i.e. fit on training only, using the StandardScalar.

4) Model Implementations: The following architectures are implemented
   a) MLP: batch norm with dense layers (128→64→32→16) togther with dropout and binary cross entropy.
   b) LSTM: single layer 64 unit LSTM with dropout.
   c) Deep LSTM: two stacked LSTM layers (128→64).
   d) CNN-LSTM: causal Conv1D (3 and 5 filters) → max pooling → LSTM → dense layers.
   e) Transformer classifier: custom encoder block with multi head attention, residual connections, layer normalization, feed forward projection, temporal pooling, and sigmoid head.
   f) Transformer regressor: same structure as classifier but with linear output and MAE/RMSE metrics.

   All sequence models use a function to convert tabular data into sliding windows of length 30–60 days.
5) Evaluation and Backtesting
   Classification outputs are evaluated using AUC and accuracy. Regression models use MAE and RMSE. A simple long only backtesting engine evaluates economic utility of next-day direction outputs by comparing the model driven equity curve, buy-and-hold baseline, and annualized Sharpe ratio.

D. Loss Function and Optimization

Different architectures require different optimisation strategies, selected to ensure stable convergence given the noisy and non-stationary nature of crypto time series.

1) Classification Tasks: For next-day direction and local-extrema detection tasks, models are trained using Binary cross entropy (BCE) loss, and Adam optimizer with learning rates. This loss is appropriate for probability based binary classification and aligns directly with evaluation metrics such as AUC and accuracy. The learning rate was set to 3e-4 for MLP, standard LSTM, and transformer classifier. Whereas, for the deep LSTM, and CNN-LSTM, learning rate of 1e-4 was set. Generally, class imbalance in extrema tasks is mitigated via dynamic class weighting, computed based on positive class frequency.
2) Regression Tasks: Next-day close price regression uses Mean Squared Error (MSE) loss for MLP, Mean Absolute Error (MAE) for transformer regression and Adam optimizer with learning rate 3e-4 for both. Both approaches standardize the target (z-score) before training to stabilize gradients.
3) Regularization and Stabilization: To prevent overfitting, dropout layers are applied throughout all deep architectures, and batch normalization stabilizes MLP feature scaling. Early stopping (patience=10) monitors validation AUC for classification models, and validation MAE for regression models. Transformer models additionally benefit from residual connections, layer normalization, feed-forward dimensionality reduction, and reducing vanishing/exploding gradients and improving training stability.

IV. Experimental Design and Evaluation

A. Datasets and Preprocessing

The experiments rely on a multi modal dataset combining cryptocurrency market data, online news statistics, Google Trends indicators, and macroeconomic variables. BTC price data span from 2010–2024, while ETH data begin in 2015. The dataset includes open–high–low–close (OHLC) values, trading volumes, market capitalization, and derived technical indicators (e.g., RSI, Stochastic, Bollinger Bands, moving averages, and ATR).

Daily news features include article counts, mean title length, sentiment scores, and sentiment variance. Google Trends queries track interest in "bitcoin," "ethereum," "blockchain," "cryptocurrency," and related terms. Macroeconomic signals include S&P 500 prices and volume, VIX, gold prices, and their respective returns and z-scores.

All features undergo forward–backward fill for missing values, z-normalisation, and rolling-window feature construction. After feature selection, the final dataset contains 62 predictive variables, producing 8,342 samples for the combined BTC–ETH time span.

B. Performance Metrics

Evaluation uses metrics aligned with the task type:

1) Regression (Next-Day Close Prediction)
   - MAE (Mean Absolute Error): Measures average absolute deviation from the true closing price.
   - RMSE (Root Mean Squared Error): Heavily penalizes larger errors; critical in volatile markets.
2) Classification (Next-Day Direction; Min/Max Local Predictions)
   - Accuracy: Correct directional predictions relative to total predictions.
   - AUC: Robust to class imbalance; measures ranking ability.
   - Sharpe Ratio (Backtesting): Evaluates risk-adjusted profitability of model-generated trading signals.
   - Equity Curve vs. Buy-and-Hold: Measures cumulative return generated by model predictions.

C. Experiment Setup

A walk forward expanding window validation strategy is employed, aligning with real world trading conditions. The first main split used for training was based on data available until the early 2023 and at that time the validation split ranged from early 2023 to the mid of 2023. The second train split then included until the end of the first validation and the second validation ranged until the end of the year 2023. Similarly the third train and validation split was done as follows:

1) Train $\leq$ 2023-01-05, Val = 2023-01-06 $\rightarrow$ 2023-07-04
2) Train $\leq$ 2023-07-04, Val = 2023-07-05 $\rightarrow$ 2023-12-31
3) Train $\leq$ 2023-12-31, Val = 2024-01-01 $\rightarrow$ 2024-06-28

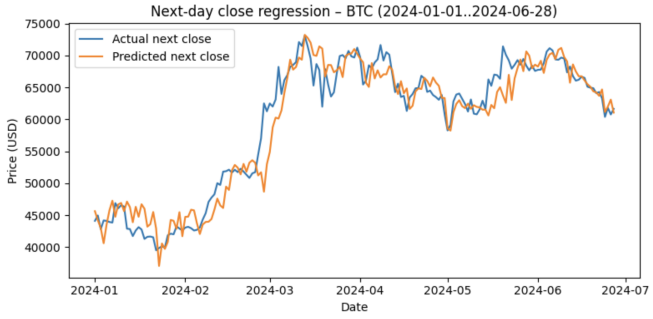

Fig. 1. Comparison of actual and predicted next-day Bitcoin closing prices for the period January 2024 to June 2024.

The models that were evaluated composed of MLP, LSTM, LSTM-DEEP, CNN-LSTM in order to compare them to the baseline. Then the transformer models were also added for both regression and classification. All models trained separately for both BTC and ETH to respect their different dynamics.

D. Results Comparative Analysis

1) Regression Results (Next-Day Close): Across all three time splits, the baseline MLP (per coin) demonstrates stable and comparatively strong performance. For BTC, MAE declines from 3693 to 2045, and for ETH, RMSE remains consistently below 360. The walk forward MLP in the baseline framework achieves even stronger results (such as RMSE 1100–3700), indicating that classical neural networks especially those leveraging historical technical indicators remain well suited for short horizon crypto regression tasks.

By contrast, the Transformer regression model performs significantly worse. For example, MAE ranges between 12,468 and 28,668, and RMSE ranges between 14,851 and 35,577, which is an order of magnitude higher than baseline models. This confirms that, in this setting, Transformers struggle with high variance numerical regression targets, particularly when price levels differ by several orders of magnitude (BTC vs ETH). According to the results found, MLP and LSTM based models clearly outperform Transformers for next-day close prediction. The MLP results for the next-day close regression, train and validation loss, and train and validation MAE can be found as shown in figures 1, 2 and 3 respectively. The overall results for the MLP regressor are shown in Figure 4.

2) Classification Results (Next-Day Direction): Baseline models again show strong performance. MLP achieves an AUC of approximately 0.64 to 0.66 with an Accuracy of 0.58 to 0.62 which is quite good for a model forecasting financial time series. The Sharpe ratios exceeded 3.0 while the Equity curves outperformed buy-and-hold in all three splits. LSTM, LSTM-DEEP, and CNN-LSTM achieve moderate but consistent performance, with AUC between 0.53 and 0.59.

The Transformer classifier underperforms compared to all the baselines with an AUC around 0.48 to 0.53 with an Accuracy of 0.44-0.54. Sharpe ratios appear high due to noise in the strategy but the equity curves are notably weaker relative to

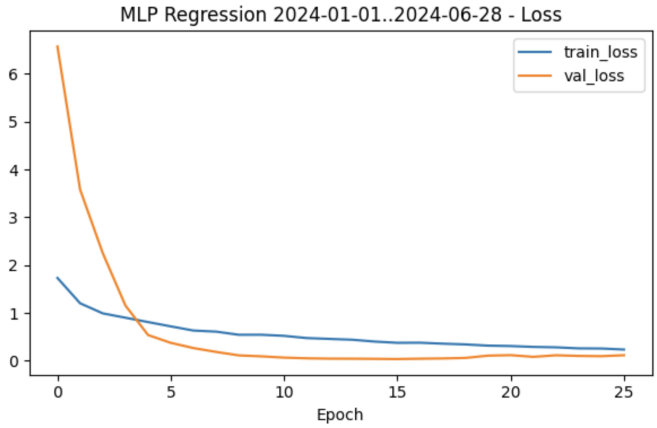

Fig. 2. Training and validation loss curves for the MLP regression model over the January 2024 to June 2024 period.

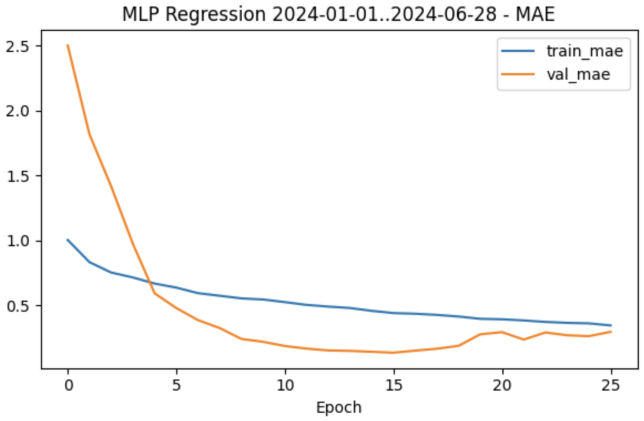

Fig. 3. Training and validation MAE curves for the MLP regression model from January 2024 to June 2024.

	coin	fold	mae	rmse
0	BTC	2023-01-06..2023-07-04	3693.157715	4497.223810
1	ETH	2023-01-06..2023-07-04	157.296707	202.967006
2	BTC	2023-07-05..2023-12-31	2663.279053	3356.958147
3	ETH	2023-07-05..2023-12-31	144.334122	173.368826
4	BTC	2024-01-01..2024-06-28	2045.326416	2809.220532
5	ETH	2024-01-01..2024-06-28	281.657715	355.625231

Fig. 4. Cross-validation performance summary for BTC and ETH across three temporal folds.

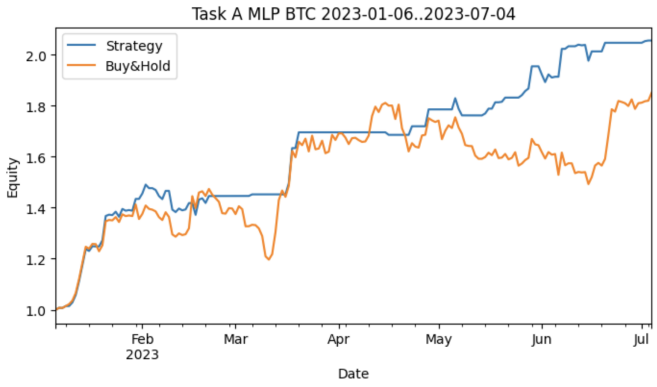

Fig. 5. Equity curve comparison for the MLP-based BTC trading strategy versus a buy and hold baseline over the period January 2023 to July 2023.

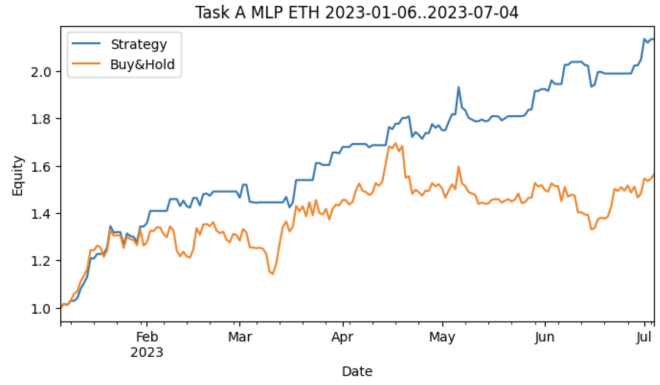

Fig. 6. Equity curves for the MLP driven ETH trading strategy compared with a buy and hold baseline from January 2023 to July 2023.

MLP and CNN-LSTM baselines. In the final split, Transformer equity barely grows, whereas MLP achieves approximately 1.75 to 2.62 growth. Under the light of the current results for the Classification, Transformers do not provide a performance gain and trail behind MLP and LSTM family models. Figures 5 and 6 show the best results from MLP classifier model for the equity and buy-and-hold plot. Whereas, figure 7 shows the train and validation loss converging in the end for the same block.

E. Ablation Study

The ablation study evaluates the incremental benefit of introducing the Transformer architecture compared to the baseline designs described in the paper by Gurgul et al. [1]. The findings indicate the following:

1) Effect of Model Architecture: MLP consistently achieves the strongest performance across regression and classification. LSTM and LSTM-DEEP provide stability but not superior accuracy. CNN-LSTM improves directional prediction in volatile regimes.
2) Impact of Feature Set: Since all models share the same 62 feature engineered set, the performance gap is at-

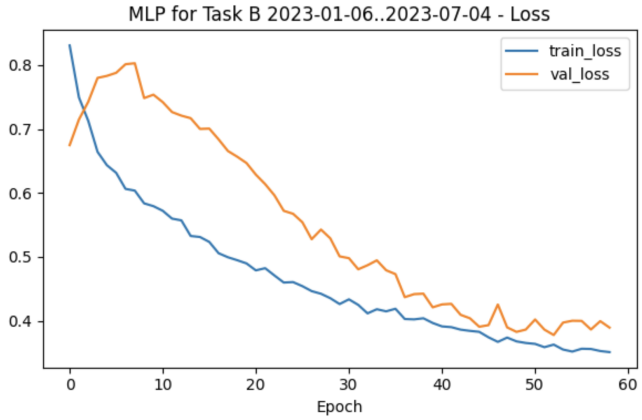

Fig. 7. Training and validation loss trajectories for the MLP model on Task B over the period January 2023 to July 2023.

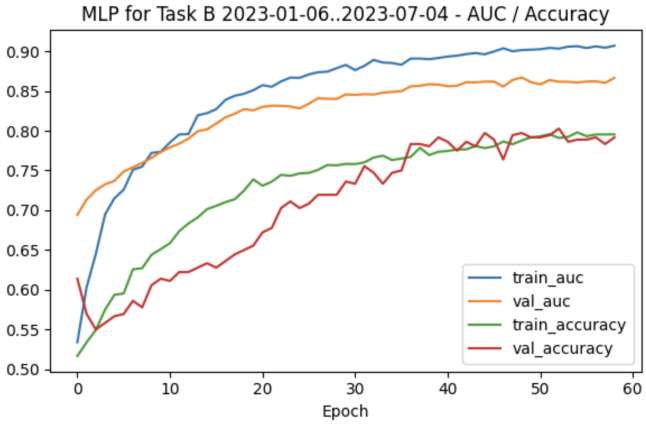

Fig. 8. Evolution of AUC and accuracy for the MLP model on Task B from January 2023 to July 2023.

#	task	model	coin	fold	auc	acc	sharpe	equity_end	bh_end
0	A	MLP	BTC	2023-01-06..2023-07-04	0.6449	0.62222	4.110076	2.055667	1.851228
1	A	MLP	ETH	2023-01-06..2023-07-04	0.6449	0.62222	3.80658	2.13447	1.563673
2	A	LSTM	BTC	2023-01-06..2023-07-04	0.47777	0.48667	0.269086	1.017839	1.336177
3	A	LSTM	ETH	2023-01-06..2023-07-04	0.46688	0.48667	-1.117027	0.922688	1.168392
4	B_MIN	MLP	POOLED	2023-01-06..2023-07-04	0.86763	0.79722	NaN	NaN	NaN
5	B_MAX	MLP	POOLED	2023-01-06..2023-07-04	0.90029	0.7	NaN	NaN	NaN
6	A	LSTM_DEEP	BTC	2023-01-06..2023-07-04	0.49147	0.46667	-0.364964	0.908995	1.336177
7	A	LSTM_DEEP	ETH	2023-01-06..2023-07-04	0.49751	0.51333	0.10316	0.988722	1.168392
8	A	CNN_LSTM	BTC	2023-01-06..2023-07-04	0.47849	0.47333	1.218754	1.280771	1.336177
9	A	CNN_LSTM	ETH	2023-01-06..2023-07-04	0.48754	0.46	0.551901	1.088456	1.168392
10	A	MLP	BTC	2023-07-05..2023-12-31	0.57211	0.56667	2.850123	1.346014	1.371341
11	A	MLP	ETH	2023-07-05..2023-12-31	0.57211	0.56667	1.850428	1.401919	1.190882
12	A	LSTM	BTC	2023-07-05..2023-12-31	0.58375	0.53333	1.696883	1.332114	1.445249
13	A	LSTM	ETH	2023-07-05..2023-12-31	0.58393	0.53333	1.657052	1.322044	1.257041
14	B_MIN	MLP	POOLED	2023-07-05..2023-12-31	0.8949	0.72222	NaN	NaN	NaN
15	B_MAX	MLP	POOLED	2023-07-05..2023-12-31	0.87161	0.75556	NaN	NaN	NaN
16	A	LSTM_DEEP	BTC	2023-07-05..2023-12-31	0.58929	0.56	2.755544	1.550055	1.445249
17	A	LSTM_DEEP	ETH	2023-07-05..2023-12-31	0.58893	0.54667	1.745805	1.323734	1.257041
18	A	CNN_LSTM	BTC	2023-07-05..2023-12-31	0.56214	0.53333	2.029463	1.445249	1.445249
19	A	CNN_LSTM	ETH	2023-07-05..2023-12-31	0.55661	0.53333	1.245719	1.257041	1.257041
20	A	MLP	BTC	2024-01-01..2024-06-28	0.65534	0.58611	2.057179	1.75234	1.459973
21	A	MLP	ETH	2024-01-01..2024-06-28	0.65534	0.58611	3.180924	2.621051	1.5099
22	A	LSTM	BTC	2024-01-01..2024-06-28	0.50544	0.48	0.852679	1.139912	1.43645
23	A	LSTM	ETH	2024-01-01..2024-06-28	0.528	0.48667	1.319379	1.210722	1.470005
24	B_MIN	MLP	POOLED	2024-01-01..2024-06-28	0.86365	0.75	NaN	NaN	NaN
25	B_MAX	MLP	POOLED	2024-01-01..2024-06-28	0.90563	0.65556	NaN	NaN	NaN
26	A	LSTM_DEEP	BTC	2024-01-01..2024-06-28	0.54377	0.52667	1.574382	1.43645	1.43645
27	A	LSTM_DEEP	ETH	2024-01-01..2024-06-28	0.50868	0.54	1.502626	1.470005	1.470005
28	A	CNN_LSTM	BTC	2024-01-01..2024-06-28	0.54662	0.55333	3.009055	1.421085	1.43645
29	A	CNN_LSTM	ETH	2024-01-01..2024-06-28	0.55144	0.5	2.02438	1.282045	1.470005

Fig. 9. Summary of performance metrics for all models across Tasks A and B, assets, and temporal folds.

tributable to architectural suitability rather than feature selection.

3) Impact of Transformer: Adding transformer model and its impact was nonexistent since it does not improve either regression or classification performance.

The proposed Transformer based extensions do not contribute positively to predictive performance. Traditional models, especially MLP and CNN-LSTM, remain the most effective architectures in this framework.

V. Extended Contributions

This study offers several meaningful contributions that extend beyond reproducing the reference architecture from the main paper. First, the project provides a fully operational end to end pipeline that consolidates heterogeneous data sources, including market statistics, Google Trends indicators, macroeconomic variables, and news derived features. The merged dataset contains eighty four hundred observations and seventy nine raw attributes before feature selection, giving a significantly broader basis for short horizon forecasting than the datasets provided in existing literature. The inclusion of news and search interest derived variables creates a richer representation of market attention which is often incomplete or omitted in similar academic studies.

Second, the project provides an extensive comparison across several neural architectures under a unified walk forward evaluation regime. The detailed logs and plots in the such figures 5 - 9, illustrate the stability and generalization behavior of classifying architectures across three distinct temporal regimes. This creates a consistent evaluation standard that improves reproducibility and strengthens the reliability of the reported findings.

Third, the results clarify where classic architectures outperform recent transformer based approaches when the forecasting horizon, sample size, and feature scale create mismatches between model structure and signal properties. The equity curves and summaries reported clearly illustrate that transformer based strategies are more volatile and less aligned with true market direction than MLP and CNN-LSTM baselines . The regression outputs show a similar behavior where transformer regression predictions systematically fail to track BTC and ETH price levels, often deviating by an order of magnitude from the true series .

Fourth, the study documents how model architectures differ in practical trading impact by pairing classification outputs with realistic backtesting. The results tables in the figure 9 show that the MLP baseline achieves stable Sharpe ratios of two to four and consistent equity growth across all validation splits, while transformer models do not provide improvement under identical market conditions . This demonstrates the importance of model choice in financial forecasting where predictive performance must translate to execution level outcomes.

Together, these contributions provide a broader and more transparent perspective on short term crypto forecasting. They highlight when increased architectural complexity is justified and when simpler, well conditioned models offer stronger predictive utility.

VI. Conclusion and Future Work

This research evaluated several deep learning architectures for short horizon forecasting of BTC and ETH using an extended multimodal dataset. Across classification and regression tasks, the results show that simple architectures such as MLP and CNN LSTM consistently outperform deeper or more complex configurations.The figure 9 confirm that MLP models achieve higher AUC, higher accuracy, and significantly better trading performance than LSTM DEEP or transformer based models . The regression summaries in the figure 4 also show that MLP models provide stable MAE and RMSE values across all walk forward splits, while transformer regression results remain an order of magnitude worse.

These findings support the conclusion that short horizon crypto price dynamics require architectures capable of learning smooth temporal relationships without over fitting sparse or noisy components of the feature space. The results also show that high dimensional attention based models require either larger training corpora or different feature engineering strategies to capture useful structure. In practice, this means that more complex architectures are not automatically superior when applied to financial time series forecasting.

Future work should address several limitations encountered in this study. One direction is to incorporate regime detection or market volatility features to dynamically select or weight models based on observed conditions. Another direction is to explore hybrid transformer designs that combine convolutional encoders with reduced attention layers to stabilize training. Extending the dataset with intraday data could also improve transformer performance due to increased temporal density. A more advanced backtesting pipeline that includes slippage, transaction costs, and position sizing logic would allow strategies to be evaluated under more realistic constraints. A final direction is to investigate representation learning methods that compress the feature space before forecasting, which may reduce noise sensitivity and improve generalization for all architectures.

Overall, the study provides a clear understanding of when specific deep learning architectures can contribute meaningful predictive power in short term crypto forecasting and identifies several promising avenues for targeted methodological improvements.

VII. References

References

[1] V. Gurgul, S. Lessmann, and W. K. Härdle, "Deep learning and NLP in cryptocurrency forecasting: Integrating financial, blockchain, and social media data," *International Journal of Forecasting*, vol. 41, no. 4, Mar. 2025, doi: https://doi.org/10.1016/j.ijforecast.2025.02.007.

[2] N. N. AlMadany, O. Hujran, G. Al Naymat, and A. Maghyereh, "Forecasting cryptocurrency returns using classical statistical and deep learning techniques," *International journal of information management data insights*, vol. 4, no. 2, pp. 100251–100251, Nov. 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100251.

[3] R. Younas, H. M. Raza Ur Rehman, and G. S. Choi, "Crypto foretell: a novel hybrid attention correlation based forecasting approach for cryptocurrency," *Journal of Big Data*, vol. 12, no. 1, Oct. 2025, doi: https://doi.org/10.1186/s40537-025-01291-7.

[4] L. A. Almusfar and F. Mallouli, "Deep Learning-Based Financial Forecasting with Post-Quantum Cryptographic Integration: A CNN-LSTM and F2N-ECC Hybrid Framework," *Procedia Computer Science*, vol. 270, no. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025), pp. 4573–4584, 2025, doi: https://doi.org/10.1016/j.procs.2025.09.583.

[5] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," *International Review of Economics & Finance*, vol. 104, p. 104719, Nov. 2025, doi: https://doi.org/10.1016/j.iref.2025.104719.

[6] 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, Mar. 2024, doi: https://doi.org/10.1016/j.irfa.2023.103055.

[7] H. Kırkgöz and O. Kurt, "Modeling bitcoin network energy demand: Price-adjusted hybrid deep learning approach to complex time series forecasting," *Chaos Solitons & Fractals*, vol. 200, pp. 117075–117075, Aug. 2025, doi: https://doi.org/10.1016/j.chaos.2025.117075.