Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33623

# *Can Deep Learning Beat a Naive Baseline for Short-Horizon Stock Price Prediction? An Empirical Study Across 20 Large-Cap US Stocks*

**Hanwen Zhao**

*Faculty of Spanish and Portuguese Studies, Beijing Foreign Studies University, Beijing, China*
*edwardzzz42@gmail.com*

***Abstract.*** This paper asks a simple question: can LSTM, GRU, and Transformer models outperform a zero-return rule at predicting next-hour stock prices, when the only inputs are hourly OHLCV data? We test this across 20 large-cap US stocks on two separate datasets — one drawn live from Yahoo Finance covering the past 730 days, the other a static archive spanning 1998 to 2024 with around 46,000 records per stock. Training and evaluation follow strict chronological order throughout to avoid any forward-looking bias. None of the three models reliably beats the Naive baseline on either dataset. Win-rates sit between 10% and 35%, and Wilcoxon signed-rank tests put model performance consistently below the baseline. The pattern holds across both datasets and both data volumes, which points to a fairly clear conclusion: OHLCV data alone, however much of it, does not provide enough signal for reliable short-horizon price prediction — consistent with what the Efficient Market Hypothesis would lead one to expect.

***Keywords:*** stock price prediction, LSTM, GRU, Transformer, Naive baseline, Efficient Market Hypothesis, time series forecasting

## 1. Introduction

Stock price prediction has kept both finance researchers and machine learning practitioners busy for a long time, and for obvious reasons. When deep learning took off, LSTM and GRU found their way into financial time series modeling fairly naturally — the sequential structure of price data seemed like a good fit, and a number of studies came back with encouraging results. Transformer models followed, swapping out the sequential processing for self-attention and offering more flexibility in capturing dependencies across longer time windows.

One thing that tends not to get much direct attention in this literature, though, is the baseline problem. A lot of studies compare one deep learning model against another, or pit them against classical statistical methods, but stop short of asking the more fundamental question: can any of these models actually beat simply predicting no change? Without that reference point, it is genuinely hard to know what the reported performance gains mean.

This paper sets up that comparison directly. We train LSTM, GRU, and Transformer models on hourly OHLCV data and test them against a Naive zero-return baseline across 20 large-cap US

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33623

stocks, keeping the chronological order of the data intact throughout and using both price-space metrics and statistical tests to assess the results. None of the three models consistently beats the baseline — a finding that sits naturally alongside the weak-form Efficient Market Hypothesis, which holds that historical price data alone is unlikely to support reliable excess returns in liquid markets.

## 2. Related work

The theoretical anchor for this study is Fama's Efficient Market Hypothesis [1]. The weak form makes a specific claim: current prices have already absorbed all available historical price information, which leaves no room for past prices alone to predict future returns. In practice, this means short-horizon price movements should behave much like a random walk.

LSTM came out of work by Hochreiter and Schmidhuber [2], who addressed a long-standing problem with standard RNNs — the tendency for gradients to vanish during backpropagation — by building in gating mechanisms. Cho et al. [3] then introduced GRU as a stripped-down version, fewer parameters, similar performance, and both architectures were picked up by the stock forecasting community fairly early on. The Transformer [4] came later, and its self-attention mechanism has since been pulled into financial time series work as well.

There is, however, a thread of research that has started to push back on the results coming out of this space. Kadak et al. [5] showed that once temporal structure is properly accounted for, several prominent LSTM-based forecasting models turn out to be producing false positives. A large cross-market study covering 55 markets [6] found something similarly deflating — model accuracy tends to go down as market efficiency goes up, and simpler approaches often perform just as well as the more elaborate deep learning ones. Xiong et al. [7] further demonstrated that stock market forecasts are systematically overestimated when denoising procedures are applied only once, highlighting the fragility of reported gains. This paper follows that same line of questioning, bringing a multi-asset setup and an explicit baseline into the picture.

## 3. Methodology

### 3.1. Data

We use two independent datasets to evaluate robustness. Dataset 1 retrieves hourly OHLCV data for 20 large-cap US stocks from Yahoo Finance covering approximately 730 days, yielding roughly 3,000 rows per stock. Dataset 2 is a static historical dataset from NASDAQ and NYSE covering January 1998 to March 2024, providing approximately 46,000 rows per stock. Both datasets cover the same 20 stocks spanning Technology (AAPL, MSFT, AMZN, GOOGL, META, NVDA, TSLA, AVGO, ORCL, CRM), Financials (JPM, BAC, GS, V, MA), Energy (XOM, CVX), and Consumer Staples (KO, PG, WMT). Non-consecutive hours are removed to ensure temporal continuity.

### 3.2. Task formulation

The prediction target is the one-step-ahead simple return $r_{t+1} = (C_{t+1} - C_t) / C_t$. Each model receives a 48-hour window of raw OHLCV features as input and outputs a predicted return. Predicted prices are reconstructed as $\hat{C}_{t+1} = C_t \times (1 + \hat{r})$. The Naive persistence baseline predicts $\hat{r} = 0$ at every step, equivalent to forecasting no price change.

DOI: 10.54254/2755-2721/2026.GU33623

### 3.3. Data splitting and normalization

The full sequence is divided into three non-overlapping segments following the original time order: the first 70% for training, the next 15% for validation, and the remaining 15% for testing. No shuffling is applied at any stage. Each input feature is scaled by subtracting the training-set mean and dividing by its standard deviation. Return targets undergo the same standardization procedure and are converted back to the original scale before computing evaluation metrics.

### 3.4. Model architectures

LSTM [2]: Single-layer LSTM (hidden=64) with a two-layer MLP head (64-64-1, ReLU).

GRU [3]: Identical to LSTM but using a GRU layer.

Transformer [4]: Linear input projection to d_model=64, 2-layer TransformerEncoder with 4 attention heads, feedforward dim 128, dropout 0.1, linear output layer on the last time step.

All three architectures share identical training configuration: the Adam optimizer with a learning rate of $10^{-3}$, mean squared error as the loss function, and mini-batches of 64 samples. Training runs for a maximum of 20 epochs; a patience-based early stopping criterion halts training if the validation loss fails to improve for 5 consecutive epochs, and gradient norms are clipped at 1.0 to stabilize optimization.

### 3.5. Evaluation

Primary metrics are RMSE and MAE in price space (USD). Statistical significance is assessed using paired t-tests and Wilcoxon signed-rank tests across the 20 stocks, with Cohen's d as effect size measure.

## 4. Results and discussion

### 4.1. Overall performance

Tables 1 and 2 present mean RMSE and MAE across all 20 stocks for Dataset 1 and Dataset 2 respectively. On both datasets, the Naive baseline achieves the lowest RMSE. All deep learning models perform worse on average, with win-rates ranging from 10% to 35% across the two experiments.

Table 1. Mean RMSE and MAE (price, USD) across 20 stocks — Dataset 1 (yfinance, ~3,000 rows/stock)

| Model | RMSE | MAE | Win-rate |
|---|---|---|---|
| Naive-last | 2.3937 | 1.4164 | — |
| LSTM | 2.3997 | 1.4300 | 25% |
| GRU | 2.4023 | 1.4326 | 25% |
| Transformer | 2.4018 | 1.4301 | 30% |

DOI: 10.54254/2755-2721/2026.GU33623

Table 2. Mean RMSE and MAE (price, USD) across 20 stocks — Dataset 2 (1998–2024, ~46,000 rows/stock)

| Model | RMSE | MAE | Win-rate |
|---|---|---|---|
| Naive-last | 1.2263 | 0.8201 | — |
| LSTM | 1.2301 | 0.8282 | 25% |
| GRU | 1.2329 | 0.8318 | 10% |
| Transformer | 1.2269 | 0.8217 | 35% |

## 4.2. Statistical analysis

Table 3 reports statistical comparisons against the Naive baseline using Dataset 2. All mean ΔRMSE values are positive, confirming DL models are on average worse. Wilcoxon signed-rank tests confirm statistical significance for all three models ($p < 0.05$). The same directional pattern holds on Dataset 1, where GRU and Transformer also show significant results.

Table 3. Statistical comparison: DL models vs. naive-last baseline — Dataset 2 (n=20 stocks)

| Model | Mean ΔRMSE | t-test p | Wilcoxon p | Cohen's d |
|---|---|---|---|---|
| LSTM | +0.0038 | 0.0700 | 0.0018 | 0.43 (small) |
| GRU | +0.0066 | 0.1919 | 0.0002 | 0.30 (small) |
| Transformer | +0.0006 | 0.1357 | 0.0291 | 0.35 (small) |

## 4.3. Discussion

Across both datasets, none of the three models managed to beat the Naive baseline. This points to a fairly straightforward conclusion: OHLCV data alone is not enough to predict short-term price movements, which lines up with what weak-form EMH would predict. What is perhaps more telling is the data volume experiment — going from roughly 4,000 to over 46,000 rows per stock, more than ten times the data, made essentially no difference. The models did not get better. That rules out data scarcity as the explanation and points instead to the signal itself: there simply is not enough predictable structure in hourly price movements for these models to learn from. This sits comfortably alongside the critical findings in recent literature [5-7], which have raised similar doubts about the robustness of positive results in this area.

These findings stand in contrast to a number of prior studies that report strong predictive performance from LSTM and Transformer architectures on individual stocks. For instance, studies applying LSTM to single equities such as Apple or Tesla frequently report accuracy figures above 90%. However, such results are typically obtained on a single stock, without a Naive baseline for comparison, and without statistical validation across multiple assets. When those controls are applied — as in the present study — the apparent advantage of complex models disappears. The gap between reported results and what survives rigorous multi-asset testing is consistent with the methodological concerns raised by [5] and [7], and underscores the importance of baseline comparison as a minimum standard in this line of research.

DOI: 10.54254/2755-2721/2026.GU33623

## 5. Limitations

Several limitations of this study deserve acknowledgment. First, only OHLCV features are used as inputs; no sentiment data, earnings announcements, or macroeconomic indicators are incorporated. Second, the prediction horizon is fixed at one hour, and the random walk property may not hold equally at longer time scales. Third, no traditional ML baseline such as linear regression or SVR is included between the Naive rule and the DL models. Fourth, hyperparameters were not extensively tuned, which was intentional for fair comparison but may understate the potential of individual models.

A more significant concern relates to distributional shift across the train, validation, and test splits. When checked empirically, the standardized variance of the training set is 1.0 by definition, but the validation and test sets show a variance of only 0.034 — roughly 30 times lower. This reflects the fact that the training data covers 1998 to approximately 2017, a period that includes high-volatility episodes such as the dot-com crash and the 2008 financial crisis, while validation and test fall in the comparatively calmer period that follows. This distributional gap may affect the reliability of early stopping, which is based on validation loss. Future work should consider rolling-window evaluation or volatility-stratified splitting to address this issue.

## 6. Conclusion

This paper set out to test whether LSTM, GRU, and Transformer models can beat a zero-return rule at predicting next-hour stock prices using only OHLCV data, across 20 large-cap US stocks and two independent datasets. None of them did. The result held whether we used 730 days of recent data or a 26-year historical archive — win-rates stayed between 10% and 35%, and Wilcoxon signed-rank tests confirmed that all three models performed significantly worse than the Naive baseline. More data did not help. More model complexity did not help either.

The core contribution of this work is methodological: by introducing a Naive zero-return baseline, enforcing strict chronological splitting, and applying statistical hypothesis testing across 20 stocks and two datasets, this study provides a more rigorous standard of evaluation than is typical in the literature. The finding that no DL model clears even this minimal bar supports the weak-form EMH and adds to a growing body of evidence [5, 6] that many reported gains in this space do not survive proper experimental controls. Future directions include incorporating alternative data sources such as news sentiment, extending the prediction horizon to daily or weekly scales, and applying the same evaluation framework to markets with lower efficiency where predictable structure may be more accessible.

## Authorship contribution

Hanwen Zhao: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, Writing – Original Draft, Writing – Review & Editing.

## References

[1] Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. J. Finance 25(2): 383–417.
[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9(8): 1735–1780.
[3] Cho, K., van Merrienböer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv: 1406.1078.

DOI: 10.54254/2755-2721/2026.GU33623

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst., 30.
[5] Kadak, A. et al. (2025). Stock market trend prediction using deep neural network via chart analysis: a practical method or a myth? Humanit. Soc. Sci. Commun. (Nature).
[6] Bedi, P. et al. (2025). Machine learning, stock market forecasting, and market efficiency: a comparative study. Int. J. Data Sci. Anal., Springer.
[7] Xiong, M., Xu, K., Yin, J., & Wang, W. (2025). Overestimated stock market forecasts resulting from the one-time denoising. Appl. Intell. 55: 112.