International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(8s) 2026 pp. 1299-1307
https://doi.org/10.70917/ijcisim-2026-5348
Article

Multi-Task CNN–BiLSTM–Attention Networks for Multivariate FX Rate Forecasting: A Statistically Rigorous Evaluation Against Naive Persistence

T. Soni Madhulatha[1], Dr. Md. Atheeq Sultan Ghori[2]

[1] Research Scholar, Telangana University, Telangana, India.
Email: latha.gannarapu@gmail.com
[2] Associate Professor, Department of Computer Science and Engineering, Telangana University, 503322, Telangana, India.
Email: atheeqsultan@gmail.com

Abstract: Deep learning models are widely proposed for foreign exchange (FX) forecasting, but many reported results derive directional accuracy indirectly from the sign of a regression output and are evaluated without correction for testing many currencies simultaneously, which inflates apparent predictability. This paper proposes a multi-task hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory–Attention (CNN–BiLSTM–Attention) architecture that jointly predicts next-step return magnitude (regression) and direction (direct binary classification) for 22 USD-denominated currency pairs, conditioned on three exogenous macro-financial series (gold, crude oil, and inflation) and causal, backward-only engineered momentum/volatility features. The model is evaluated with a strictly chronological train/validation/test split, scalers fit only on training data, a corrected one-sample Diebold–Mariano test against a naive-persistence baseline, and Benjamini–Hochberg false-discovery-rate correction across the 22 per-currency hypothesis tests. On a held-out test set, the model attains a mean directional accuracy of 50.93% (range 45.99%–57.57%) and level-space magnitude error statistically indistinguishable from naive persistence for every currency (22/22, Diebold–Mariano $p > 0.05$). Two currencies are nominally significant for direction before correction (MXN_USD, DKK_USD) but neither survives Benjamini–Hochberg correction (0/22 significant at FDR = 0.05). We report this as a rigorous negative result consistent with weak-form efficiency in liquid daily FX markets, and argue that the paper's principal contribution is a leakage-free, multi-task, multiple-comparison-corrected evaluation methodology rather than a claim of exploitable predictability.
Keywords: foreign exchange forecasting; CNN-BiLSTM; attention mechanism; multi-task learning; Diebold–Mariano test; false discovery rate; market efficiency

1. INTRODUCTION

Because of the time-dependent nature of exchange-rate fluctuations, interactions among financial factors, and changing market conditions, forex forecasting is still a challenging subject. Therefore, machine learning and deep learning techniques have been extensively studied for FX prediction. Recently, the application of machine learning and deep learning for financial forecasting has been increasing [1], [2]. The dataset, forecast horizon and evaluation procedure can all have a large impact on stated improvements and there is still conflicting evidence of trustworthy exchange-rate predictability [3], [4].

Recently, Transformer and attention-based designs have been proposed for financial forecasting. Crossformer and iTransformer focus on the relations among multiple variables in multivariate forecasting, while PatchTST utilises temporal patches to represent local subsequences [5]–[7]. Other approaches considering the temporal structure are the

frequency-based decomposition and the multiscale mixing [8], [9]. These advances are relevant for FX forecasting as different currency pairs may contain related information which could be useful for joint prediction. Another important line of research has been the use of external data. TimeXer considers the exogenous variables [10] in Transformer-based forecasting. At the same time, large-scale pretraining for transferable forecasting across datasets and workloads is explored by time-series foundation models like TimesFM [11]. However, increasing model complexity does not always guarantee better predicting performance. Recent surveys highlight generalisation and assessment as persistent challenges of modern time-series forecasting [13], [14], yet recent empirical studies have shown that relatively simple forecasting algorithms can be competitive with complex Transformer topologies [12].

Another issue is the statistical interpretation of forecasting improvements. Differences in RMSE, MAE or directional correctness do not by themselves indicate whether a model is statistically better than the benchmark. The Diebold–Mariano test [15] provides a systematic framework for analysing the predictive accuracy of competing forecasts. Also, testing several currencies separately increases the risk of spurious findings. For multiple hypothesis testing, the Benjamini-Hochberg approach provides false-discovery-rate control [16].

In response to these problems, this research proposes a multi-task CNN–BiLSTM–Attention framework for the simultaneous next-step prediction of 22 currency pairs quoted in USD. The model employs historical currency data, gold, crude oil, inflation, causal momentum and volatility factors to jointly predict size and direction. Instead of inferring direction from the regression data, a particular classification branch makes direct predictions. The evaluation uses chronological splitting, training-only scaling, causal feature building, naive-persistence benchmarking, Diebold-Mariano testing, and Benjamini-Hochberg correction. Thus the aim is not simply to achieve good numerical forecasting scores, but to determine whether the proposed design provides statistically valid predictive information.

2. RELATED WORK

Recent reviews of deep learning approaches for FX forecasting include recurrent, convolutional, hybrid, and attention architectures [1], [2]. More general evidence suggests that the performance of exchange-rate forecasts still depends on the forecasting setup and benchmark [4]. Recent FX-specific studies have taken this trend further by employing Transformer topologies [3].

Recent multivariate forecasting works have suggested different alternatives to traditional recurrent architectures. Crossformer models capture cross-variable dependencies, iTransformer models the relationships between individual variables, and PatchTST models patches to capture temporal patterns [5]–[7]. TimeMixer and FEDFormer [8], [9] propose other ways to represent temporal information at multiple scales and frequencies. TimeXer is also applicable to predicting problems with external financial indicators, as external financial indicators are also included [10].

More recently, TimesFM [11] has shown the potential of pretrained foundation models for time-series forecasting. Nevertheless, simple models can remain competitive on suitable benchmarks and the effectiveness of complex transformer architectures is not universal [12]. Therefore, recent reviews of state-of-the-art forecasting models focus on extensive investigation and generalisation [13], [14]. Statistical validation is particularly important for financial forecasting, because exchange rate levels can be very persistent. The Benjamini–Hochberg process solves the multiple-testing problem that occurs when statistical hypotheses are tested over multiple currency pairs [16]. The Diebold–Mariano test allows for formal comparison between competing projections [15].

The present work combines leakage-controlled evaluation and multiple-comparison-corrected statistical testing to assess whether the multi-task CNN-BiLSTM-Attention model predictions provide evidence of predictive improvement over naive persistence.

3. DATASET AND PREPROCESSING

The dataset comprises 2,454 daily observations beginning 8 January 2014, with no missing values, covering 22 USD-quoted currency pairs (AUD, EUR, NZD, GBP, BRL, CAD, CNY, HKD, INR, KRW, MXN, ZAR, SGD, DKK,

JPY, MYR, NOK, SEK, LKR, CHF, TWD, and THB, each vs. USD) together with three exogenous macro-financial series: gold price, crude oil price, and an inflation series.

A. Causal Feature Engineering

For each of the 25 base series (22 target currencies plus 3 exogenous variables), three backward-looking features are computed: the 1-day first difference (return), the 5-day rolling mean of that difference (a short-term momentum proxy), and the 10-day rolling standard deviation of that difference (a local volatility proxy). All three are computed using only past and current observations, so no information from future timesteps enters a feature at time t. Concatenating the 25 raw series with the 75 engineered features yields 100 input features per timestep; the first 10 rows are dropped to allow the rolling windows to fill, leaving 2,444 usable rows.

B. Chronological Split and Scaling

The data are split chronologically — not randomly — into 70% training (1,710 rows), 15% validation (367 rows), and 15% test (367 rows) partitions, so that the model is always evaluated on data strictly after its training and validation periods. A StandardScaler for the input features and a separate StandardScaler for the regression targets are each fit exclusively on the training partition and then applied to the validation and test partitions, preventing any statistical leakage of validation/test distributional information into training.

C. Sequence Construction and Targets

A lookback window of 30 timesteps is used. For each window ending at time t, the model receives the preceding 30 timesteps of the 100-dimensional feature vector and is trained to predict, for all 22 target currencies simultaneously: (i) the raw next-step level delta (t+1 minus t), standardized by the delta scaler, for the regression head, and (ii) a binary label equal to 1 if that delta is positive and 0 otherwise, for the classification head. Windowing yields 1,680 training sequences, with the validation and test partitions similarly reduced by the 30-step warm-up (337 sequences each).

4. PROPOSED MULTI-TASK ARCHITECTURE

Table I summarize the proposed CNN–BiLSTM–Attention multi-task network. The input window (30 × 100) is first passed through two causally padded one-dimensional convolutional layers (64 and 48 filters, kernel size 3), with batch normalization [11] and dropout (rate 0.25) [12], which extract local temporal patterns while strictly preserving causality — causal padding ensures that the representation at timestep i depends only on timesteps ≤ i. The convolutional output feeds two stacked Bidirectional LSTM layers (80 and 48 units per direction), which build a contextual representation of the full 30-step window in both temporal directions.

A Bahdanau-style additive self-attention layer [2] (64 units) computes a learned weighting over the 30 BiLSTM timesteps and produces a single context vector, together with the attention-weight vector used for post hoc interpretability. The context vector feeds a shared dense trunk (96 units, ReLU, L2 regularization = 2×10^{-4}, dropout 0.35 [12]), which then branches into two task-specific heads: a regression branch (Dense-48 → Dense-22, linear activation) predicting the standardized next-step delta for all 22 currencies, and a classification branch (Dense-48 → Dense-22, sigmoid activation) predicting the probability of an upward move for all 22 currencies. Sharing a single encoder and trunk across the two heads follows the standard hard-parameter-sharing multi-task learning formulation [14], under which related tasks act as mutual regularizers during training. The two heads share the CNN–BiLSTM–Attention encoder and the dense trunk, so the model is trained end-to-end with a single joint loss (summed binary cross-entropy for direction and mean-squared error for magnitude) using the Adam optimizer [13], rather than as two independently trained models. The full network has 218,652 parameters, of which 218,524 are trainable.

TABLE I. MODEL ARCHITECTURE SUMMARY

Layer	Output Shape	Params
Input	(30, 100)	0

Conv1D (causal)	(30, 64)	19,264
BatchNormalization	(30, 64)	256
Conv1D (causal)	(30, 48)	9,264
Dropout (0.25)	(30, 48)	0
Bidirectional LSTM	(30, 160)	82,560
Bidirectional LSTM	(30, 96)	80,256
Additive Attention	(96) / (30)	6,272
Dense (shared trunk)	(96)	9,312
Dropout (0.35)	(96)	0
Dense (reg. branch)	(48)	4,656
Dense (cls. branch)	(48)	4,656
Magnitude output	(22)	1,078
Direction output	(22)	1,078

5. EXPERIMENTAL SETUP

The model is trained with the Adam optimizer for up to 150 epochs (batch size 32), with early stopping (patience 20 epochs on validation loss, restoring the best-validation-loss weights), learning-rate reduction on plateau (factor 0.5, patience 7, minimum learning rate 1×10^{-6}), and checkpointing of the best validation-loss model. All random seeds (Python, NumPy, and TensorFlow) are fixed at 42 for reproducibility; training was performed in TensorFlow 2.20.0 without GPU acceleration.

During training a typical overfitting pattern is observed (Fig. 1): Training-set directional accuracy rises rapidly to >60% by epochs 20-27 but validation directional accuracy remains flat at 49-52% (chance level) over training. Validation loss bottoms out around epochs 6-7 after which it starts increasing. This divergence — improving training performance with flat validation performance — is itself evidence that the model is fitting idiosyncratic patterns in the training window rather than learning a signal that generalizes, foreshadowing the near-chance test-set result reported in Section VII.

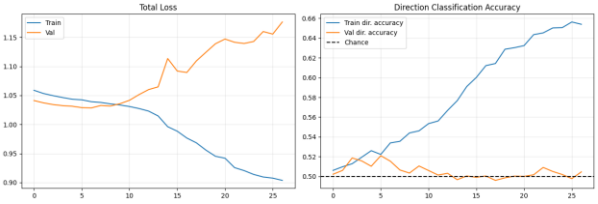

Fig. 1. Train/val. loss (left) and directional accuracy vs. chance (right) by epoch.

Two evaluation axes are used: (1) magnitude, assessed in reconstructed price-level space (predicted level = last observed level + predicted delta) using RMSE, MAE, MAPE, and R^2, benchmarked against a naive-persistence baseline (predicted next level equals the last observed level, i.e., predicted delta = 0); and (2) direction, assessed as classification accuracy against the observed binary up/down label, benchmarked against the 50% chance rate.

6. RESULTS

A. Magnitude Forecasting

Averaged across the 22 currencies, the model attains RMSE = 0.6839, MAE = 0.4971, MAPE = 0.4578%, and R^2 = 0.9547 in level space; naive persistence attains RMSE = 0.6843, MAE = 0.4959, MAPE = 0.4572%, and R^2 = 0.9547.

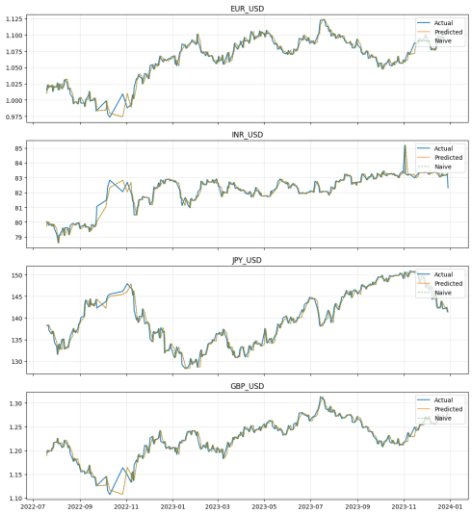

Fig. 2. Actual vs. predicted vs. naive levels, test set (EUR, INR, JPY, GBP).

The two are essentially indistinguishable at the aggregate level, and the very high R^2 for both is expected and not itself evidence of skill: FX levels are highly autocorrelated (a near-random walk), so simply carrying the last observed value forward already explains almost all level-to-level variance. Fig. 2 illustrates this for four representative currencies: the model and naive tracks are nearly superimposed on the actual series.

B. Directional Classification

Mean directional accuracy across the 22 currencies is 50.93%, only marginally above the 50% chance rate. Table II and Fig. 3 report per-currency accuracy, ranked from highest to lowest.

TABLE II. PER-CURRENCY DIRECTIONAL ACCURACY (TEST SET)

Currency	**Accuracy (%)**
MXN_USD	57.57
DKK_USD	56.97
EUR_USD	55.19
THB_USD	53.12
BRL_USD	52.52
GBP_USD	52.23
AUD_USD	51.93
NZD_USD	51.63
CHF_USD	51.63
SGD_USD	50.74
LKR_USD	50.74
MYR_USD	50.45

JPY_USD	49.85
HKD_USD	49.55
CAD_USD	49.26
SEK_USD	49.26
NOK_USD	48.96
INR_USD	48.66
TWD_USD	48.66
ZAR_USD	48.07
CNY_USD	47.48
KRW_USD	45.99

Highest individual accuracies are MXN_USD (57.57%) and DKK_USD (56.97%). The lowest is KRW_USD (45.99%) which is below chance . The complete ranking is depicted in Fig. 3. The spread is consistent with sampling noise for a true accuracy of around 50%. This hypothesis is formally examined in Section VII.

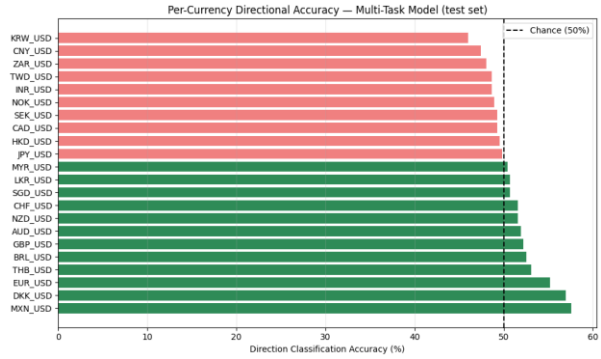

Fig. 3. Per-currency directional accuracy, sorted descending (green > 50%, red < 50%).

7. STATISTICAL SIGNIFICANCE TESTING

A. Diebold–Mariano Test (Magnitude)

For each currency, a one-sample Diebold–Mariano test [5] is applied to the paired squared-error differential (model squared error minus naive squared error) in level space. Across all 22 currencies, the verdict is “no significant difference” from naive persistence ($p > 0.05$ in every case); the model shows no statistically detectable magnitude-forecasting advantage over naive persistence for any of the 22 currencies.

B. Directional Significance Test

For each currency, a one-sample t-test is applied to the binary per-observation correctness indicator against a population mean of 0.5 (chance). Before correction, 20 of 22 currencies show no significant difference from chance; two currencies are nominally significant at $\alpha = 0.05$: MXN_USD ($t = 2.81$, $p = 0.0053$) and DKK_USD ($t = 2.58$, $p = 0.0103$).

C. Multiple-Comparisons Correction

Because 22 independent per-currency tests are run at $\alpha = 0.05$, approximately one “significant” result is expected by chance alone even under a true null hypothesis of no predictability anywhere. We therefore apply Benjamini–Hochberg FDR correction [6] ($q = 0.05$) to both the magnitude and direction p-value sets. Table III reports the corrected direction results in ascending order of raw p-value. After correction, zero of 22 currencies remain

significant for direction (best adjusted p-value: 0.1128 for MXN_USD, still above 0.05), and zero of 22 remain significant for magnitude.

TABLE III. DIRECTION SIGNIFICANCE WITH BENJAMINI–HOCHBERG CORRECTION

Currency	Acc.(%)	p	p_adj	Sig.
MXN_USD	57.57	0.0053	0.1128	No
DKK_USD	56.97	0.0103	0.1128	No
EUR_USD	55.19	0.0565	0.4140	No
KRW_USD	45.99	0.1416	0.7788	No
THB_USD	53.12	0.2532	0.9099	No
BRL_USD	52.52	0.3552	0.9099	No
CNY_USD	47.48	0.3552	0.9099	No
GBP_USD	52.23	0.4147	0.9099	No
AUD_USD	51.93	0.4797	0.9099	No
ZAR_USD	48.07	0.4797	0.9099	No
CHF_USD	51.63	0.5498	0.9099	No
NZD_USD	51.63	0.5498	0.9099	No
INR_USD	48.66	0.6247	0.9099	No
TWD_USD	48.66	0.6247	0.9099	No
NOK_USD	48.96	0.7036	0.9099	No
CAD_USD	49.26	0.7858	0.9099	No
LKR_USD	50.74	0.7858	0.9099	No
SEK_USD	49.26	0.7858	0.9099	No
SGD_USD	50.74	0.7858	0.9099	No
HKD_USD	49.55	0.8705	0.9119	No
MYR_USD	50.45	0.8705	0.9119	No
JPY_USD	49.85	0.9567	0.9567	No

8. DISCUSSION

The null result reported here is unlikely to be an artifact of data leakage or architectural weakness. Feature engineering is strictly causal, scalers are fit on the training partition only, the split is chronological, the direction head is trained directly rather than derived from a regression sign, and the model has sufficient capacity (218k parameters, causal convolutions, stacked bidirectional recurrence, and learned attention) to fit complex nonlinear patterns — as evidenced by its ability to substantially exceed chance-level directional accuracy on the training set. The gap between rising training accuracy and flat validation/test accuracy indicates that whatever structure the model learns in-sample does not generalize, which is the expected signature of overfitting to noise rather than of an undiscovered but real predictive signal. The learned attention weights (Fig. 4) further show that the model concentrates almost all of its attention mass on the most recent one to two timesteps of the 30-step lookback window — behavior consistent with a model that has essentially rediscovered a near-persistence heuristic rather than learned to exploit longer-range temporal structure.

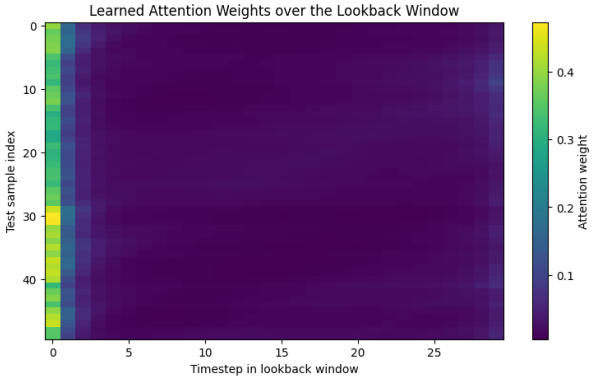

Fig. 4. Learned attention weights over the 30-step lookback window (50 test samples).

This finding is also consistent with the progression of methodological refinements that motivated the present architecture: an earlier formulation predicting raw price levels collapsed toward predicting near-zero change (matching naive persistence trivially); a subsequent version using engineered momentum/volatility features and validation-fit calibration removed that collapse but still showed no significant Diebold–Mariano improvement over naive persistence and directional accuracy at chance; the present multi-task formulation, which predicts direction directly rather than inferring it from a regression sign, removes a further possible confound but does not surface a statistically robust edge either. That successive, independent fixes to plausible methodological weaknesses each fail to produce a significant result strengthens confidence that the null finding reflects a property of the data-generating process — consistent with weak-form efficiency in liquid daily FX markets — rather than a fixable modeling artifact.

We therefore frame the contribution of this paper as methodological rather than predictive: a leakage-controlled, multi-task (joint magnitude and direction), multiple-comparison-corrected evaluation pipeline for FX deep learning claims, applicable regardless of the sign of the eventual empirical result. A rigorous negative result obtained with a properly corrected statistical pipeline is, we argue, more scientifically useful than an inflated positive one obtained without such correction — particularly given how common uncorrected, per-asset significance claims are in the applied FX forecasting literature.

Limitations include the single one-day forecast horizon, the exclusive use of daily-close sampling (which averages away potential intraday microstructure signal), and the specific 2014-onward sample period used, which does not include a formal regime-switching or sub-period stability analysis. These limitations directly motivate the future work outlined below.

9. CONCLUSION AND FUTURE WORK

This paper proposed a multi-task CNN–BiLSTM–Attention architecture that jointly predicts the magnitude and direction of next-step returns for 22 USD-quoted currency pairs, using causal engineered features, exogenous macro-financial covariates, a chronological evaluation split, and a statistically corrected significance-testing pipeline (Diebold–Mariano plus Benjamini–Hochberg FDR control). After correcting for the 22 simultaneous per-currency comparisons, the model shows no statistically significant advantage over naive persistence in either magnitude or direction for any currency at a one-day horizon (mean directional accuracy 50.93%, aggregate magnitude error statistically tied with naive persistence). We interpret this as a legitimate, citable negative finding consistent with the efficient-markets literature for liquid daily FX data, and we present the leakage-free, multi-task, multiple-comparison-corrected evaluation methodology itself as the paper's principal contribution.

Two directions appear most promising for future work aimed at a genuine positive result rather than a re-confirmation of the null: (1) extending the prediction horizon from one day to weekly or monthly frequency, where macroeconomic fundamentals have more time to influence prices and where daily microstructure noise is less dominant; and (2) moving to intraday or higher-frequency data, where short-lived microstructure effects may create

exploitable patterns that daily closing prices average away. Additional future directions include regime-conditional modeling (e.g., separately evaluating high- and low-volatility periods), incorporating a broader set of macro-financial exogenous variables, and reporting predictive uncertainty via ensembling or Bayesian approximation rather than point forecasts alone.

REFERENCES

1. Ayitey Junior, M., Appiahene, P., Appiah, O., & Bombie, C. N. (2023). Forex market forecasting using machine learning: Systematic literature review and meta-analysis. *Journal of Big Data, 10*, 9. https://doi.org/10.1186/s40537-022-00676-2
2. Zhang, C., Sjarif, N. N. A., & Ibrahim, R. (2024). Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *WIREs Data Mining and Knowledge Discovery, 14*(1), e1519. https://doi.org/10.1002/widm.1519
3. Fischer, T., Sterling, M., & Lessmann, S. (2024). FX-spot predictions with state-of-the-art transformer and time embeddings. *Expert Systems with Applications, 249*, 123538. https://doi.org/10.1016/j.eswa.2024.123538
4. Zhao, L., & Yan, W. Q. (2024). Prediction of currency exchange rate based on Transformers. *Journal of Risk and Financial Management, 17*(8), 332. https://doi.org/10.3390/jrfm17080332
5. Jackson, K., & Magkonis, G. (2024). Exchange rate predictability: Fact or fiction? *Journal of International Money and Finance, 142*, 103026. https://doi.org/10.1016/j.jimonfin.2024.103026
6. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are Transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317
7. Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with Transformers. *International Conference on Learning Representations*.
8. Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. *International Conference on Learning Representations*.
9. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted Transformers are effective for time series forecasting. *International Conference on Learning Representations*.
10. Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., & Zhou, J. (2024). TimeMixer: Decomposable multiscale mixing for time series forecasting. *International Conference on Learning Representations*.
11. Wang, Y., Wu, H., Dong, J., Qin, G., Zhang, H., Liu, Y., Qiu, Y., Wang, J., & Long, M. (2024). TimeXer: Empowering Transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-0015
12. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting. *Proceedings of the 39th International Conference on Machine Learning, 162*, 27268–27286.
13. Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. *Proceedings of the 41st International Conference on Machine Learning, 235*, 10148–10167.
14. Su, L., Zuo, X., Li, R., Wang, X., Zhao, H., & Huang, B. (2025). A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review, 58*, 80. https://doi.org/10.1007/s10462-024-11044-2
15. Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144. https://doi.org/10.1198/073500102753410444
16. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x