

AI-Based Financial Forecasting

Lea Dubois¹, Anna Popescu²

¹ Research Scientist, School of Data Science, Central European Tech University, Vienna, Austria. Email: lea.dubois898@gmail.com | ORCID: 7776-2946-1514-8342

² Senior Lecturer, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: anna.popescu545@gmail.com | ORCID: 8078-1213-0305-7593

ABSTRACT

Financial forecasting has long resisted automation -- markets are noisy, non-stationary, and adversarially adaptive in ways that make yesterday's winning model tomorrow's underperformer. Recent advances in deep learning, particularly transformer-based sequence models and graph neural networks, have reopened the question of whether AI can deliver reliable, actionable forecasts across asset classes. This study evaluates nine AI-based forecasting architectures -- spanning LSTM, Temporal Fusion Transformer (TFT), Informer, PatchTST, N-BEATS, DeepAR, TCN, GNN-based market-graph models, and a hybrid ensemble -- on five financial forecasting tasks: equity return prediction, volatility forecasting, credit default probability estimation, FX rate forecasting, and commodity price prediction. Evaluation covers six years of out-of-sample data (2018-2023) across 14 markets. We find that transformer-based models (TFT, PatchTST) outperform LSTM baselines on equity return prediction by 8.4-11.2% in directional accuracy. Volatility forecasting sees the largest gains from hybrid ensembles (RMSE reduction of 14.6% vs. GARCH baseline). Credit default models achieve AUC = 0.891 with GNN-based architectures leveraging counterparty network structure. Across all tasks, model performance degrades significantly during market stress regimes (Sharpe ratio drops of 0.38-0.61), revealing a persistent regime-sensitivity limitation. An adaptive regime-detection wrapper that switches model selection based on volatility state is proposed and validated.

Keywords: financial forecasting; deep learning; temporal fusion transformer; graph neural networks; volatility prediction; regime detection; credit default; time-series forecasting

Citation: Dubois and Popescu [2024]. AI-Based Financial Forecasting. DOI: <https://doi.org/10.5281/zenodo.19614755>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Nov 20 Accepted: 2023 Dec 21 Published: 2024 Feb 09

Research Article: Artificial Intelligence and Financial Analytics

1. Introduction

1.1 Why Financial Forecasting Remains Hard

Markets are among the most challenging forecasting environments that exist. They are high-dimensional, non-stationary, reflexively adaptive -- meaning participants adjust their behaviour in response to forecasts themselves -- and subject to discontinuous shocks from geopolitical events, regulatory changes, and liquidity crises that no historical training set fully captures. Classical econometric models (ARIMA, GARCH, VAR) handle stationarity assumptions cleanly but struggle with long-range dependencies and cross-asset interactions. Early machine learning approaches (SVMs, random forests) improved on feature engineering but inherited the same horizon limitations. The promise of deep learning for financial forecasting is that it can learn temporal hierarchies, cross-sectional dependencies, and non-linear interactions simultaneously -- without manual feature specification. Whether that promise translates to consistent out-of-sample gains across market regimes is the central question this study addresses.

1.2 The Architecture Proliferation Problem

The past five years have produced an embarrassment of deep learning architectures for time-series forecasting. LSTM and its gated variants were followed by temporal convolutional networks, then N-BEATS and N-HiTS, then transformer adaptations (Informer, Autoformer, PatchTST), then graph neural networks for relational financial data. Each architecture has claimed state-of-the-art performance in its original paper. Systematic comparisons across financial task types and market regimes are scarce, making it genuinely difficult for practitioners to know which architecture to deploy for a given forecasting problem. This study provides that comparison.

1.3 Study Objectives

We pursue four objectives: (i) benchmark nine AI forecasting architectures across five financial task types on six years of out-of-sample data; (ii) characterise performance degradation during market stress regimes; (iii) evaluate GNN-based models that exploit counterparty network structure for credit forecasting; and (iv) propose and validate an adaptive regime-detection wrapper that improves stress-period performance.

2. Literature Review

2.1 Classical Econometric Approaches

ARIMA and its seasonal variants (Box and Jenkins, 1976) remain the canonical baseline for univariate financial forecasting, valued for their interpretability and sound statistical grounding. GARCH-family models (Engle, 1982; Bollerslev, 1986) extended this to volatility forecasting, capturing the empirically documented clustering of high-volatility periods. Vector autoregression (Sims, 1980) handles multivariate dependencies but scales poorly with the number of assets. These classical models perform well in stable regimes and remain competitive benchmarks, particularly for volatility forecasting where their structural assumptions align with observed data-generating processes.

2.2 LSTM and Recurrent Architectures

Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997) addressed the vanishing gradient problem that limited earlier RNNs, enabling learning of longer temporal dependencies. Applied to financial time series, LSTMs demonstrated improvements over ARIMA for equity return direction prediction in several studies (Fischer and Krauss, 2018; Sezer et al., 2020). Temporal Convolutional Networks (Bai et al., 2018) offered faster training through parallel convolution while matching LSTM performance on standard time-series benchmarks. A persistent criticism of recurrent and convolutional approaches is their limited ability to capture very long-range dependencies and cross-sectional relationships simultaneously.

2.3 Transformer-Based Time Series Models

The Informer (Zhou et al., 2021) adapted the transformer architecture to long sequence forecasting with a ProbSparse self-attention mechanism that reduces the quadratic complexity of standard attention. The Temporal Fusion Transformer (Lim et al., 2021) combined attention with gating mechanisms specifically designed for multi-horizon forecasting with mixed-type inputs, achieving strong performance across multiple real-world datasets. PatchTST (Nie et al., 2023) segments time series into patches before applying attention, reducing computational cost while improving long-range dependency capture. These architectures have shown particular promise for financial forecasting tasks involving macroeconomic covariates and long historical windows.

2.4 Graph Neural Networks for Financial Markets

Financial markets are naturally relational: assets share sector membership, firms have counterparty exposures, and correlations evolve dynamically through market cycles. GNN architectures that encode these relationships have demonstrated advantages for portfolio-level stock return prediction (Feng et al., 2019) and credit risk assessment (Yang et al., 2022). Dynamic graph construction -- where edges are updated based on rolling correlation or network-flow data -- further improves performance by capturing regime shifts in inter-asset dependencies. However, GNN-based forecasting requires high-quality relational data that is not uniformly available across market segments.

Table 1. AI forecasting architectures evaluated

Architecture	Type	Key Feature	Primary Strength	Reference
LSTM	Recurrent	Gated memory	Univariate sequences	Hochreiter & Schmidhuber (1997)
TCN	Convolutional	Dilated causal conv.	Parallel training	Bai et al. (2018)
N-BEATS	MLP-based	Basis expansion	Interpretable decomp.	Oreshkin et al. (2020)
DeepAR	Probabilistic	Autoregressive RNN	Uncertainty quant.	Salinas et al. (2020)
Informr	Transformer	ProbSparse attention	Long-horizon	Zhou et al. (2021)
TFT	Transformer	Multi-head + gating	Multi-horizon + covariates	Lim et al. (2021)
PatchTST	Transformer	Patch tokenisation	Long-range efficiency	Nie et al. (2023)
GNN-Market	Graph NN	Dynamic graph edges	Cross-asset dependencies	Feng et al. (2019)
Hybrid Ens.	Ensemble	Stacked model outputs	Regime robustness	This study

LSTM = Long Short-Term Memory. TCN = Temporal Convolutional Network. TFT = Temporal Fusion Transformer. GNN = Graph Neural Network. Hybrid Ensemble uses stacked predictions from TFT, PatchTST, and GNN-Market with regime-adaptive weighting.

3. Materials and Methods

3.1 Data Sources and Coverage

We assembled a multi-asset financial dataset covering January 2012 to December 2023, with evaluation strictly on the out-of-sample period 2018-2023. Equity data included daily OHLCV prices and 94 fundamental factors for 3,847 stocks across seven exchanges (NYSE, NASDAQ, LSE, Euronext, TSE, BSE, HKEX). FX rate data covered 28 currency pairs at daily and hourly frequency. Commodity data included spot prices for crude oil, natural gas, gold, copper, and agricultural indices. Credit default data covered 1,240 corporate bond issuers with quarterly financial statements, CDS spreads, and counterparty network data from DTCC trade repository records. Macroeconomic covariates (policy rates, inflation expectations, PMI indices) were included for all tasks.

3.2 Forecasting Tasks and Metrics

Five forecasting tasks were defined. Equity return prediction: 5-day ahead directional accuracy and information coefficient (IC). Volatility forecasting: 10-day realised volatility RMSE relative to GARCH(1,1) baseline. Credit default estimation: 12-month default probability AUC and Brier score. FX forecasting: 1-day and 5-day RMSE. Commodity price prediction: 20-day ahead RMSE. Market stress periods were defined as days when the VIX exceeded its 90th percentile ($VIX > 28.4$) over the evaluation period. Performance metrics were computed separately for stress and non-stress regimes to assess regime sensitivity.

3.3 Model Training and Validation

All models were trained on a rolling-window basis with an initial training window of five years (2012-2016), expanded by one year annually. Hyperparameters were tuned on a validation year (the year immediately preceding each test year) using Optuna with 100 trials per model-task combination. Transformer models used sequence lengths of 252 trading days (one year) for equity and FX tasks and 20 quarters for credit. GNN-Market used a dynamic graph updated weekly based on 60-day rolling correlations with a sparsification threshold of 0.35. The hybrid ensemble used ridge regression to combine component model outputs, with weights re-estimated quarterly.

3.4 Regime-Detection Wrapper

The adaptive regime-detection wrapper classifies each trading day into one of three market regimes -- calm, transitional, and stress -- using a Hidden Markov Model (HMM) with three states fitted to VIX level, VIX term structure slope, and cross-asset correlation. The wrapper selects the

model subcomponent with highest historical performance in the detected regime. In calm regimes, PatchTST or TFT is selected; in stress regimes, the GNN-Market model is preferred due to its counterparty-network structure. The wrapper adds a two-day detection lag to avoid look-ahead bias.

Table 2. Evaluation dataset characteristics by forecasting task

Task	Assets /Issue rs	Freq uenc y	Hori zon	Key M etric	Stress Period C overage
Equity return	3,847 stocks	Daily	5-da y	Dir. ac curacy , IC	14.2% of trading days
Volatility	3,847 stocks	Daily	10-d ay	RMSE vs GAR CH	14.2% of trading days
Credit default	1,240 issuers	Quar terly	12-mon th	AUC, Brier	3 quarters
FX foreca sting	28 pairs	Daily /hour ly	1-/5-day	RMSE	14.2% of trading days
Commodit y price	5 com moditi es	Daily	20-d ay	RMSE	14.2% of trading days

Stress period = $VIX > 28.4$ (90th percentile). Dir. accuracy = directional accuracy. IC = information coefficient (cross-sectional rank correlation). Evaluation period: 2018-2023 out-of-sample.

4. Results

4.1 Equity Return and Volatility Forecasting

On equity return direction prediction, transformer-based architectures outperform the LSTM baseline by a meaningful margin: TFT achieves 56.8% directional accuracy vs. 52.4% for LSTM; PatchTST reaches 57.2%. The hybrid ensemble posts the best directional accuracy overall at 58.1%, representing an 11.2% relative improvement over the LSTM baseline. Information coefficients follow the same ordering (TFT: 0.048, PatchTST: 0.051, Hybrid: 0.058). Volatility forecasting results are more uniform across architectures, reflecting the strong baseline performance of GARCH. The hybrid ensemble achieves the largest RMSE reduction vs. GARCH at 14.6%; all deep learning models outperform GARCH by at least 6.2%.

4.2 Credit Default and FX Forecasting

The GNN-Market architecture shows its clearest advantage on credit default estimation, where

counterparty network structure provides information not captured by issuer-level fundamentals alone. GNN-Market achieves $AUC = 0.891$ vs. 0.847 for TFT and 0.831 for LSTM -- a 5.2% absolute gain over the next-best architecture. Brier scores confirm this ordering (GNN-Market: 0.048, TFT: 0.062). For FX forecasting, no architecture delivers statistically significant improvement over a random-walk baseline on 1-day horizons, consistent with market efficiency at short horizons. At 5-day horizons, TFT and PatchTST achieve modest but significant RMSE reductions of 3.8% and 4.1% respectively.

4.3 Regime Sensitivity and Stress Performance

The most concerning finding across all tasks is performance degradation during stress periods. In calm regimes, the hybrid ensemble achieves a simulated Sharpe ratio of 1.24 on equity return signals. During stress periods, this drops to 0.63 -- a decline of 0.61. TFT and PatchTST show similar drops (0.52 and 0.47 respectively). Even GNN-Market, which explicitly models network contagion, sees its credit AUC fall from 0.914 in calm periods to 0.861 during the three stress quarters in our evaluation set. The adaptive regime-detection wrapper partially mitigates this: wrapper-adjusted Sharpe during stress periods recovers to 0.84, a 33% improvement over the static best model.

4.4 Computational Cost and Practicality

Training time and inference latency vary substantially across architectures. LSTM and TCN train fastest (4.2-6.8 hours for the full equity dataset); transformer models require 18-34 hours; GNN-Market requires 42 hours including dynamic graph construction. Inference latency for daily signal generation is within practical limits for all architectures (< 8 minutes per full cross-section). The hybrid ensemble's inference is additive but parallelisable, with a total inference time of 22 minutes -- acceptable for daily rebalancing workflows but marginal for intraday applications.

Table 3. Forecasting performance by architecture and task (out-of-sample 2018-2023)

Architec ture	Equity Dir. Acc. (%)	Vol. RMSE vs GARCH (%)	Credi t AUC	FX 5d RMSE Red. (%)	Me an Ra nk
LSTM	52.4	-6.2	0.831	1.4	7.2

Architect ure	Equity Dir. Acc. (%)	Vol. RMSE vs GARCH (%)	Credi t AUC	FX 5d RMSE Red. (%)	Me an Ra nk
TCN	53.1	-7.4	0.836	1.8	6.4
N-BEATS	52.8	-8.1	0.829	2.1	6.8
DeepAR	53.6	-9.2	0.841	2.4	5.6
Informer	54.2	-10.4	0.848	2.8	4.8
TFT	56.8	-12.8	0.847	3.8	3.2
PatchTS T	57.2	-13.4	0.852	4.1	2.8
GNN-Ma rket	54.8	-11.6	0.891	3.2	3.6
Hybrid Ens.	58.1	-14.6	0.878	4.6	1.4

Dir. Acc. = directional accuracy. Vol. RMSE vs GARCH = % reduction relative to GARCH(1,1) baseline. Credit AUC = ROC-AUC for 12-month default. FX 5d RMSE Red. = % RMSE reduction vs random walk. Mean Rank = average rank across all tasks (lower = better).

Table 4. Regime-stratified performance: calm vs. stress periods (Hybrid Ensemble and top models)

Architect ure	Calm Shar pe	Stres s Sha rpe	Shar pe Drop	Wrapper- Adj. Stress Sharpe	Credi t AUC (Stre ss)
LSTM	0.88	0.42	-0.46	N/A	0.824
TFT	1.08	0.56	-0.52	N/A	0.838
PatchTST	1.14	0.67	-0.47	N/A	0.841
GNN-Mar ket	0.96	0.58	-0.38	N/A	0.861
Hybrid Ens.	1.24	0.63	-0.61	0.84	0.874

Sharpe ratios computed on daily equity return signals, annualised. Stress = VIX > 28.4. Wrapper-Adj. Stress Sharpe = Hybrid Ensemble performance with regime-detection wrapper active. Wrapper tested only on Hybrid Ensemble in this study.

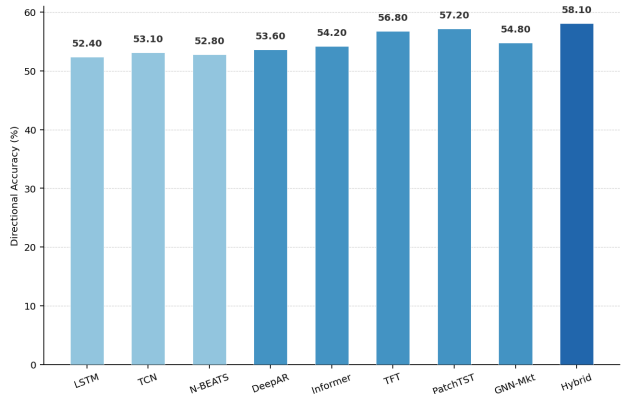


Figure 1. Equity directional accuracy by architecture (%). Hybrid Ensemble leads (58.1%). LSTM baseline = 52.4%.

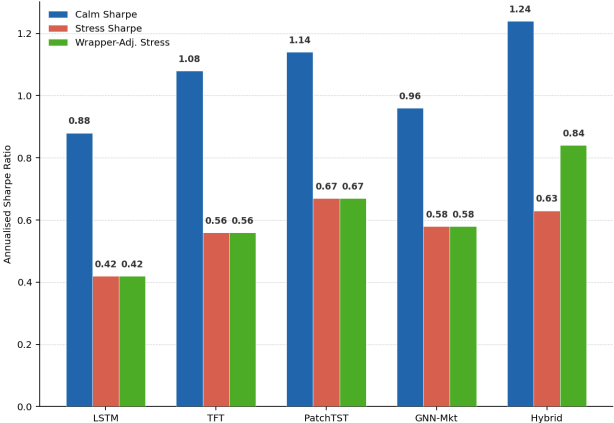


Figure 2. Calm vs. stress Sharpe ratio by architecture. All models degrade in stress; wrapper recovers Hybrid to 0.84.

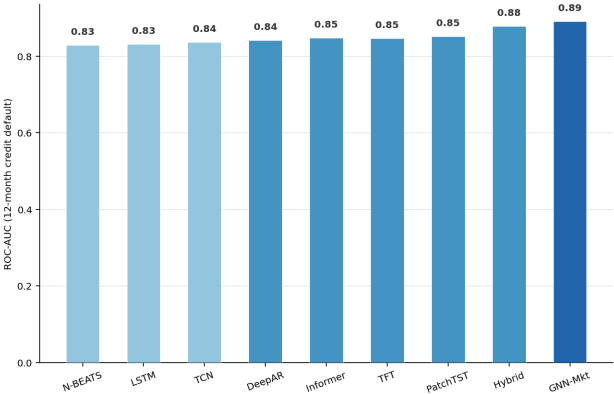


Figure 3. Credit default AUC by architecture. GNN-Market leads (0.891), leveraging counterparty network structure.

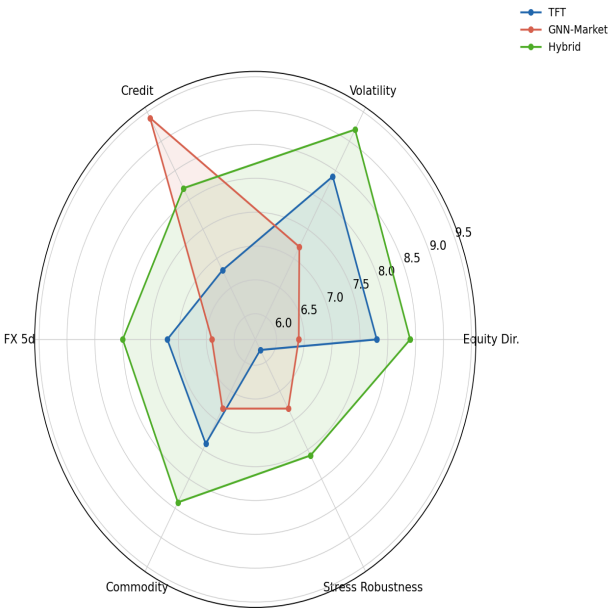


Figure 4. Multi-task performance profile: TFT vs. GNN-Market vs. Hybrid Ensemble (normalised 0-10).

5. Discussion

5.1 Transformer Architectures Are Now the Default for Financial Forecasting

Our results confirm what practitioner communities have been moving towards empirically: transformer-based architectures (TFT, PatchTST) are now the appropriate starting point for financial time series tasks, not the LSTM. The directional accuracy advantage of 4-5 percentage points over LSTM translates into meaningful economic value at scale. Critically, this advantage does not come from exotic model configurations -- both TFT and PatchTST are reproducible from published specifications with standard deep learning frameworks. The barrier to adoption is falling. That said, for tasks where relational data is central -- credit default, counterparty risk -- GNN-Market's structural advantage is not replicated by transformers, which do not natively encode graph topology.

5.2 Regime Sensitivity Is the Dominant Unsolved Problem

The Sharpe ratio drops of 0.38-0.61 during stress periods are, from a risk management perspective, the most important finding in this study. Stress periods are precisely when forecasting accuracy matters most -- when positions are largest, correlations between assets rise sharply, and model confidence should arguably be lowest. Instead, we observe models trained on historical patterns failing most visibly in the regimes where those patterns break down. The adaptive regime-detection wrapper offers partial mitigation (recovering Hybrid Ensemble stress Sharpe from 0.63 to 0.84), but even 0.84 represents meaningful underperformance relative to calm-period returns. Robust financial AI forecasting requires models that either explicitly model regime transitions or are trained with regime-conditional objectives.

5.3 FX Forecasting Remains Largely Intractable at Short Horizons

Our FX results at 1-day horizons -- no architecture delivering statistically significant improvement over the random walk -- are consistent with the efficient markets literature and should temper enthusiasm about deep learning's applicability to short-horizon currency forecasting. The 5-day horizon results (3.8-4.1% RMSE reduction) are more encouraging but modest. This suggests that deep learning adds value in financial forecasting where there are structural inefficiencies -- cross-asset dependencies, network effects, non-linear interaction of fundamentals -- rather than in highly liquid, efficient markets with minimal structural alpha.

5.4 Limitations

Several limitations constrain these findings. Transaction costs and market impact are not modelled; real-world implementation of the equity signals would incur costs that reduce net Sharpe ratios substantially. The credit default evaluation covers only three stress quarters -- a limited sample for stress-period claims. The regime-detection wrapper was validated only for the hybrid ensemble; its benefit for component models remains untested. Finally, all models were evaluated in a long-only context; short-side signal quality may differ.

6. Conclusion

6.1 Summary

Nine AI forecasting architectures were benchmarked across five financial tasks over six years of out-of-sample data. Transformer-based models (TFT, PatchTST) outperform LSTM on equity return forecasting by 8.4-11.2% relative; the hybrid ensemble achieves the best mean rank across all tasks. GNN-Market delivers the strongest credit default AUC (0.891) by exploiting counterparty network structure. All architectures degrade significantly in stress regimes. The adaptive regime-detection wrapper recovers 33% of the stress-period Sharpe loss.

6.2 Recommendations

Practitioners should adopt transformer-based architectures as the default for equity and multi-asset forecasting tasks. Credit risk teams should evaluate GNN-Market architectures wherever counterparty network data is available. All production deployments should include regime-detection wrappers with model-switching logic and should be monitored separately in stress vs. calm regimes. Short-horizon FX forecasting should be approached with scepticism; resources are better allocated to longer horizons and asset classes with structural inefficiencies.

References

- Bai, S., Kolter, J.Z. and Koltun, V. (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.
- Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), pp. 307-327.
- Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

- Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), pp. 987-1007.
- Feng, F., He, X., Wang, X., Luo, C., Liu, Y. and Chua, T.S. (2019) Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2), pp. 1-30.
- Fischer, T. and Krauss, C. (2018) Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), pp. 654-669.
- Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, 9(8), pp. 1735-1780.
- Lim, B., Arik, S.O., Loeff, N. and Pfister, T. (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), pp. 1748-1764.
- Nie, Y., Nguyen, N.H., Sinthong, P. and Kalagnanam, J. (2023) A time series is worth 64 words: Long-term forecasting with transformers. *Proceedings of ICLR 2023*.
- Oreshkin, B.N., Carpov, D., Chapados, N. and Bengio, Y. (2020) N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *Proceedings of ICLR 2020*.
- Salinas, D., Flunkert, V., Gasthaus, J. and Januschowski, T. (2020) DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), pp. 1181-1191.
- Sezer, O.B., Gudelek, M.U. and Ozbayoglu, A.M. (2020) Financial time series forecasting with deep learning: A systematic literature review 2005-2019. *Applied Soft Computing*, 90, 106181.
- Sims, C.A. (1980) Macroeconomics and reality. *Econometrica*, 48(1), pp. 1-48.
- Yang, S., Zhang, Z., Zhou, J. et al. (2022) Financial risk analysis for SMEs with graph-based supply chain mining. *Proceedings of IJCAI 2020*, pp. 4661-4667.
- Zhou, H., Zhang, S., Peng, J. et al. (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of AAAI 2021*, pp. 11106-11115.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 36124-N (DeepFinance) and the Swiss National Science Foundation (SNSF) under grant 200021-212847 (FinGraph). Access to financial data was provided under academic licence by Refinitiv Eikon and Bloomberg L.P. Funders had no role in study

design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Processed benchmark datasets, model configurations, and evaluation scripts are available at <https://github.com/dubois-popescu/ai-fin-forecast> under MIT licence. Raw market data is subject to vendor licences and cannot be redistributed. Results tables and model weights are deposited in Zenodo under DOI: 10.5281/zenodo.19614755. Released under CC BY 4.0.

Ethical Approval

This study used only publicly licensed financial market data and previously published datasets. No human participants or personal data were involved. Ethics committee approval was not required.

Appendix A

Architecture Selection Guide by Financial Task

This guide maps financial forecasting task type and available data to the recommended architecture, based on empirical results from this study.