Milestone 5

Model Evaluation

Group 3 – DS and AI Lab

October 19, 2025

1 Overview

This milestone evaluates all models on three *walk-forward* folds (Fold 0–2) and three horizons $h \in \{1, 5, 20\}$ using the leakage-safe pipeline from earlier milestones. For each (fold, horizon), we report test metrics—RMSE, MAE, R^2, directional accuracy—and a simple sign-based backtest (avg. daily P&L, volatility, Sharpe, hit ratio, turnover). No averaging across folds is performed; results are preserved *per configuration.*

Artifacts produced.

- Tabular results saved to `data/experiments/m5_plots/results_table.csv` and `.json`.
- Metric-only bar plots (300 dpi) per (fold, horizon) in `data/experiments/m5_plots/`.

2 Experimental Protocol (recap)

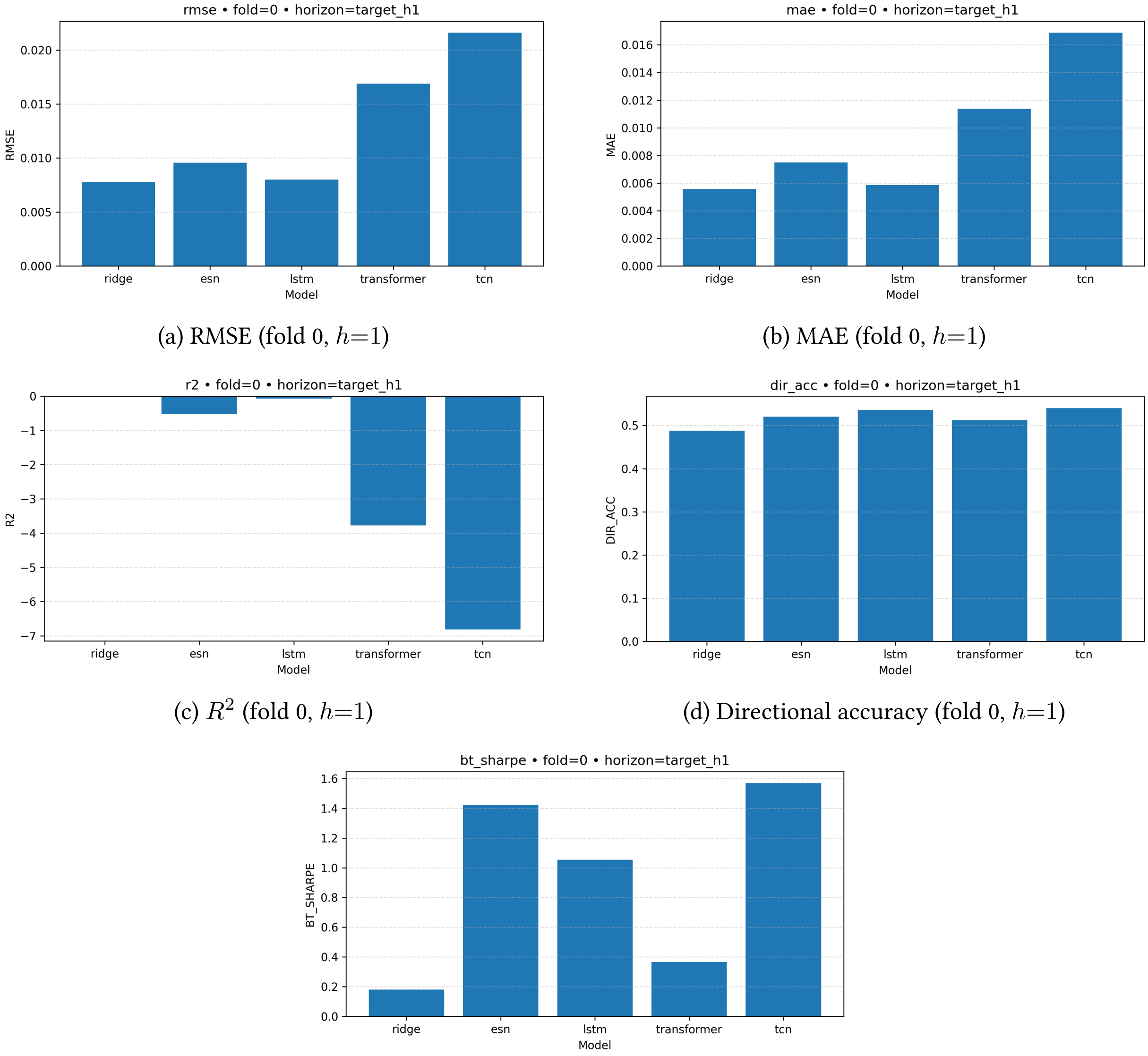
- **Folds.** Train=2520 trading days (≈10y), Test=252 days (≈1y), Step=252 days; Fold 0–2.
- **Targets.** Forward log-returns $y_{t,h}$ at $h \in \{1, 5, 20\}$.
- **Inputs.** Standardized features z_t built from returns/volatility/trend/RSI/volume/weekday. Scalers fit on *train only.*
- **Models.** Ridge (linear), ESN (leaky reservoir + ridge readout), LSTM, Transformer, TCN. Identical loss (MSE), identical folds, per-horizon training.
- **Metrics.** RMSE, MAE, R^2, directional accuracy; and a toy sign-based backtest with 1 bp per-trade cost: avg. daily P&L, volatility, Sharpe (annualized by $\sqrt{252}$), hit ratio, turnover.

Table 1: Fold 0, horizon `target_h1`: test metrics and backtest statistics (values truncated to 4 decimals).

Model	Fold	Horizon	RMSE	MAE	R^2	Dir. Acc.	Avg. Daily PnL	Vol	Sharpe	Hit Ratio	Turnover
ridge	0	target_h1	0.0077	0.0055	-0.0139	0.4880	0.0000	0.0077	0.1810	0.4880	0.7260
esn	0	target_h1	0.0095	0.0074	-0.5287	0.5198	0.0006	0.0077	1.4230	0.5200	0.8130
lstm	0	target_h1	0.0079	0.0058	-0.0723	0.5357	0.0005	0.0077	1.0540	0.5360	0.7100
transformer	0	target_h1	0.0168	0.0113	-3.7752	0.5119	0.0001	0.0077	0.3660	0.5120	0.4480
tcn	0	target_h1	0.0215	0.0168	-6.8132	0.5396	0.0007	0.0077	1.5680	0.5400	0.7100

3 Brief Observations

- **Error metrics.** RMSE/MAE vary substantially across folds/horizons; negative R^2 values indicate the difficulty of daily-return prediction and frequent regime shifts.
- **Decision proxy.** Despite weak R^2, some configurations show positive Sharpe in the toy backtest (e.g., certain LSTM/TCN/ESN cases), illustrating that directional consistency can diverge from squared-error fit.
- **Fold sensitivity.** Performance dispersion across folds underscores the need for robust cross-fold reporting without pooling.



(a) RMSE (fold 0, h=1)

(b) MAE (fold 0, h=1)

(c) R^2 (fold 0, h=1)

(d) Directional accuracy (fold 0, h=1)

(e) Sharpe (toy backtest; fold 0, h=1)

Figure 1: Headline metrics for Fold 0 at horizon h=1.