$\boxed{\text{Milestone 4}}$

Model Training (ESN & Deep Baselines for Multi-Horizon Return Forecasting)

Group 3 – DS and AI Lab

October 19, 2025

1 Scope

This milestone executes the training plan from M3: we train the proposed Echo State Network (ESN) and three deep baselines (LSTM, Transformer, Temporal ConvNet/TCN), alongside a linear Ridge lower bound. All models run in our leakage-safe, walk-forward pipeline (Train $\approx$10y = 2520 days; Test $\approx$1y = 252 days; Step = 252 days). We report preliminary results (fold 0, h=1 day) and document the training protocol, hyperparameter grids, and early findings.

2 Training Protocol

Inputs. Per-fold standardized features $u_t \in \mathbb{R}^d$ (`ret_1, ret_2, ret_5, vol_20, ma_10, ma_20, ma_gap, rsi_14, vol_z, dow`); targets are forward log-returns y_{t+h} for $h \in \{1, 5, 20\}$.

Walk-forward & scaling. For each fold k: fit a *StandardScaler* on train features only, transform train/test, and save `train.csv`, `test.csv`, `scaler.json`. This prevents leakage of global moments.

Model-specific training.

- **Ridge (lower bound):** minimize MSE with L2 penalty λ.
- **ESN (state-space):** roll the leaky reservoir

$$x_t = (1 - a)x_{t-1} + a\ \tanh(W_{\text{in}}[1; u_t] + W x_{t-1}),$$

  discard a washout w, then solve ridge closed-form for the readout: $W_{\text{out}} = (H^\top H + \alpha I)^{-1} H^\top Y$, with $H = [\mathbf{1}\ \ X]$.
- **LSTM/Transformer/TCN (deep baselines):** sequence-to-one regressors on left-padded windows of length L; MSE loss, Adam optimizer. Validation = last 10% of the train window; best state by lowest val loss.

Metrics. On the test window: RMSE, MAE, R^2, direction accuracy (sign agreement). We also compute a *toy* sign-based backtest (1 bp per-trade cost):

$$\text{position}_t = \text{sgn}(\hat{y}_t), \qquad \text{pnl}_t = \text{position}_t \cdot y_t - \text{cost} \cdot |\Delta\text{position}_t|.$$

We summarize avg daily P&L, volatility, Sharpe (daily $\rightarrow$ annualized by $\sqrt{252}$), hit ratio, and turnover. (This is diagnostic, not a trading claim.)

3 Hyperparameter Grids

Small grids are used to keep runtime reasonable; they map one-to-one to `config/experiments.py`.

- **ESN:** $H \in \{400, 800\}$, spectral radius $\gamma \in \{0.85, 0.95\}$, leak $a \in \{0.3, 1.0\}$, ridge $\alpha \in \{0.3, 3.0\}$, density $= 0.1$, washout $= 100$, seed $= 0$.
- **LSTM:** $L \in \{32, 64\}$, hidden $\in \{128, 256\}$, layers $\in \{1, 2\}$, dropout $\in \{0.0, 0.1\}$, epochs $= 15$, batch $= 128$, lr $= 10^{-3}$.
- **Transformer:** $L \in \{32, 64\}$, d_{model}=128, heads $= 4$, layers $= 2$, feedforward $= 256$, dropout $= 0.1$, epochs $= 15$.
- **TCN:** $L \in \{32, 64\}$, channels $\in \{(64, 64), (64, 128)\}$, kernel $= 3$, dropout $\in \{0.0, 0.1\}$, epochs $= 15$.

4 Preliminary Results (Fold 0, h=1 day)

Table 1 reports the first trained runs per model on fold 0. All models use the exact same features, scaling, and test window.

model	fold	horizon	RMSE	MAE	R^2	DirAcc	AvgPnL	Vol	Sharpe	Turnover
ridge	0	h1	0.008335	0.005989	-0.005	0.492	0.000077	0.008323	0.147	0.687
lstm	0	h1	0.009392	0.007025	-0.276	0.508	0.000208	0.008325	0.396	0.337
esn	0	h1	0.010098	0.007852	-0.475	0.528	0.000841	0.008278	**1.612**	0.853
transformer	0	h1	0.014366	0.011092	-1.986	0.480	-0.000127	0.008332	-0.242	0.456
tcn	0	h1	0.028566	0.019333	-10.807	**0.552**	0.000433	0.008313	0.826	0.631

Table 1: Preliminary test metrics (fold 0, h=1). Sharpe computed from the toy sign backtest (1 bp cost).

Key findings (so far).

- **Best forecaster by magnitude:** *Ridge* has the lowest RMSE/MAE (others yield $R^2 < 0$, i.e., not beating the mean on scale).
- **Best directional/trading signal:** *ESN* delivers the highest Sharpe (1.612) and the highest AvgPnL, with DirAcc > 0.52.
- **TCN** achieves the highest DirAcc (≈ 0.552) but is poorly calibrated (very large RMSE/MAE); sign still produces Sharpe $= 0.826$.
- **LSTM** is mid-pack: DirAcc > 0.50, moderate Sharpe.
- **Transformer** underperforms on this setup (likely undertrained/overparameterized for daily-noise scale).

5 Discussion

Why ESN helps on direction. The nonlinear reservoir with controlled memory (γ, a) can capture regime-dependent sign cues even when absolute magnitudes are noisy. Although ESN's RMSE trails Ridge, its *sign* accuracy and turnover profile yield the strongest toy Sharpe.

Why Ridge wins on RMSE. A linear readout on carefully standardized, low-variance features is hard to beat for scale in noisy daily returns. Linear bias often reduces variance and overfit.

On TCN/Transformer. TCN's causal dilations seem to learn directional motifs but over/under-scale predictions; calibration (or training on vol-normalized targets) should help. Transformers may require more data/regularization or shorter windows L to stabilize.

6 Error Analysis (h=1, fold 0)

Residuals for the Ridge baseline are roughly centered and thin-tailed; ESN residuals show heavier tails yet improved sign. Autocorrelation of ESN residuals beyond small lags is weak, suggesting limited temporal structure remains once the sign is extracted. (Figures produced by `src/viz/plots.py`: residual histograms, ACF, true-vs-pred scatter, and last-N time series.)

7 Ablations & Next Steps

Planned sweeps. Expand grids per model, and evaluate across *all* folds and horizons:

- **ESN:** sweep H, γ, a, α, washout; try seed-averaged ensembles of reservoirs.
- **Targets:** train on volatility-normalized returns $y/\widehat{\sigma}$ to improve magnitude calibration; compare to direction-only training.
- **Features:** add cross-asset exogenous inputs (e.g., VIX for SPY) and market-regime indicators; test robustness.
- **Calibration:** post-hoc scaling of predictions (isotonic/Platt on val) to address TCN/Transformer magnitude errors without harming sign.
- **Statistics:** aggregate metrics over folds and conduct Diebold–Mariano tests vs. Ridge for RMSE and sign loss.

8 Reproducibility & Artifacts

- **Code.** Modularized under `src/models/*`, `src/train/runner.py`, `src/viz/plots.py`; experiments use unique IDs via parameter slugs.
- **Data.** `data/raw/` (Yahoo), `data/processed/` (features), `data/splits/` (per-fold scaled CSVs + scalers).
- **Experiments.** `data/experiments/¡model¿/¡exp_id¿/fold_¡k¿/{preds,metrics}`; summaries stored per experiment.

9 Conclusions

On the first trained fold, **Ridge** remains the *magnitude* benchmark, while the proposed **ESN** is the strongest *directional* model by risk-adjusted returns. These patterns are consistent with expectations for noisy daily horizons. Milestone 5 will expand to all folds/horizons, refine grids, and include statistical significance and calibration analyses.

Ethics/Disclaimer. Academic project; not financial advice. Backtests here are simplified diagnostics and do not constitute tradable strategies.