Course Project
Data Science and AI Lab (BSDA4001)

Group 3

November 25, 2025

Indian Institute of Technology Madras
Chennai-600036, Tamil Nadu, India

State-Space Echo State Networks for Multi-Horizon Financial Forecasting

Authors: Tarun Karmakar, Subhashree M, Ashish Mehta,
Haifa Abdul Sathar, Pradeep Singh

Code: `github.com/HaifaIITM/DSAI-PROJECT-GROUP-3`

Ethics & Disclaimer: This project is academic. Nothing herein constitutes financial advice. Any backtest is illustrative only and includes explicit modeling simplifications.

Team Roles & Contributions

- **Tarun Karmakar** — Data acquisition & archival (Yahoo), robust CSV parsing, RAG component.
- **Subhashree M** — Feature engineering (u_t), target construction ($y_{t,h}$), leakage-safe scaling, NLP component.
- **Ashish Mehta** — Training loops & hyperparameter sweeps, NLP component.
- **Haifa Abdul Sathar** — Evaluation metrics, result tables, RAG component.
- **Pradeep Singh** — ESN modeling (state-space framing), baselines.

All: literature review, ablations, code review, and slide preparation.

Why → How

- **Why**: Daily financial returns are noisy, weak-signal, nonstationary; evaluation is leakage-prone.
- **How**: Cast forecasting as a *state-space* problem; use ESN (fixed dynamics) vs. deep baselines under strict walk-forward.
- **Deliverables**: Clean data pipeline, provably causal preprocessing, multi-horizon targets, and reproducible backtesting.

Motivation & Impact

- **Fast iteration**: ESN trains only readouts $\Rightarrow$ broad sweeps, robust ablations.
- **Clarity**: State-space framing clarifies memory, stability, and inductive bias.
- **Reproducibility**: Leakage-safe folds, saved scalers, fully auditable artifacts.

Problem Definition

- Observed features $u_t \in \mathbb{R}^d$; targets are forward log-returns

$$y_{t,h} \;=\; \sum_{i=1}^{h} \big(\log P_{t+i} - \log P_{t+i-1} \big), \quad h \in \{1, 5, 20\}.$$

- Learn a *causal* fading-memory operator

$$F_h^\star \;:\; (u_1, \ldots, u_t) \;\mapsto\; y_{t,h}, \qquad \hat{y}_{t,h} = F_h(u_{1:t}),$$

  with diminishing sensitivity to remote past.
- **Goal**: minimize point loss (MSE) and improve directional metrics under walk-forward splits.

Data Acquisition & Archival

Symbols (diverse universe).

- **Indices/ETF:** GSPC, SPY
- **Crypto:** BTC-USD, ETH-USD
- **India (NSE):** NSEI, NSEBANK, RELIANCE.NS, TCS.NS
- **FX/Commodities/Vol:** EURUSD=X, USDINR=X, GC=F, CL=F, VIX

Horizon: ~20 years, daily bars ("1d") up to the run-time date.

Downloader: `yf.download(..., auto_adjust=False)` to preserve both `Close` and `Adj Close`; save canonical OHLCV to `data/raw/{symbol}_{start}_to_{end}_1d.csv`.

Causality & Leakage Safety

- Filtration $\mathcal{F}_t := \sigma(u_1, \dots, u_t)$. **Causality**: $\hat{y}_{t,h}$ must be $\mathcal{F}_t$-measurable.
- Train-only standardization (per fold), with train index set $\mathcal{T}$:

$$\mu_j = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} u_t^{(j)}, \quad \sigma_j = \left(\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (u_t^{(j)} - \mu_j)^2 \right)^{1/2}, \quad z_t^{(j)} = \frac{u_t^{(j)} - \mu_j}{\sigma_j}.$$

- No function of test data is used in preprocessing; labels use only future of the *same* point.

State-Space View of Forecasting

$$
\begin{aligned}
x_{t+1} &= f(x_t, z_t) + \xi_t, \qquad x_t \in \mathbb{R}^H, \;\; \xi_t \text{ process noise (optional)}, \\
\hat{y}_{t,h} &= g_h(x_t), \quad h \in \{1, 5, 20\}.
\end{aligned}
$$

- x_t compresses recent history (finite-memory proxy).
- **Bias–variance trade-off**: choose f simple (fixed ESN) and learn only g_h.

Pipeline Schematic

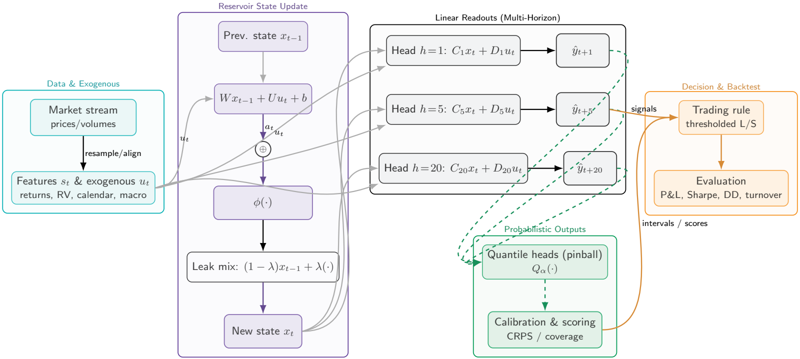


Figure: **State–space ESN pipeline.** Features $\rightarrow$ leaky reservoir (fixed $W, U, b, \phi, \lambda, \rho(W)<1$) $\rightarrow$ multi-horizon linear readouts $\hat{y}_{t+h} = C_h x_t + D_h u_t$; optional quantile heads for calibration; toy backtest for decision metrics. Only the readouts (and quantile heads) are trained.

Echo State Network: Contractive Dynamics

$$x_t = (1-a)\,x_{t-1} \;+\; a\,\phi\big(Wx_{t-1} + W_{\text{in}}\,[1; z_t]\big), \quad \phi = \text{tanh},\; a \in (0,1], \tag{1}$$

$$\hat{y}_{t,h} = \begin{bmatrix} 1 \\ x_t \end{bmatrix}^{\top} w_{\text{out},h}. \tag{2}$$

Sufficient condition for ESP (echo-state property). If ϕ is 1-Lipschitz and

$$\kappa \;\equiv\; a\,\|W\|_2 \;<\; 1,$$

then the driven system (1) is a contraction in x and admits a unique input-caused state (fading memory).

ESP: Geometric Decay (Sketch)

Let x_t, x'_t be two trajectories under the same input. Using 1-Lipschitz ϕ,

$$\|x_t - x'_t\| \leq (1-a)\|x_{t-1} - x'_{t-1}\| + a\,\|W\|_2\,\|x_{t-1} - x'_{t-1}\| = \big(1 - a + a\|W\|_2\big)\,\|x_{t-1} - x'_{t-1}\|.$$

If $\kappa = a\|W\|_2 < 1$, then $\|x_t - x'_t\| \leq \kappa\,\|x_{t-1} - x'_{t-1}\| \Rightarrow \|x_t - x'_t\| \leq \kappa^t\|x_0 - x'_0\|$.

- Initial-condition influence decays geometrically $\Rightarrow$ fading memory.
- Design knobs: shrink $\|W\|_2$ (spectral radius control) and/or reduce a (leak).

Readout Learning: Closed-Form Ridge

- After a washout w, stack states and bias: $H = \begin{bmatrix} \mathbf{1} & X \end{bmatrix} \in \mathbb{R}^{n \times (H+1)}$.
- For horizon h, targets $Y_h \in \mathbb{R}^n$.

$$w^{\star}_{\text{out},h} \;=\; \arg\min_{w} \;\|Hw - Y_h\|_2^2 + \alpha\|w\|_2^2 \;=\; (H^\top H + \alpha I)^{-1} H^\top Y_h.$$

- **Per-horizon heads** $w_{\text{out},h}$; extremely fast to fit and cross-validate.
- Shapes: $H \in \mathbb{R}^{(T-w-h)\times(H+1)}$, $Y_h \in \mathbb{R}^{T-w-h}$.

Targets & Label Generation

$$r_t = \log P_t - \log P_{t-1}, \qquad y_{t,h} = \sum_{i=1}^{h} r_{t+i},$$

$$\mathcal{I}_{\text{train}}^{\text{eff}}(h) = \{t:\ t \geq t_0 + w,\ t \leq t_{T-1-h}\}, \quad \mathcal{I}_{\text{test}}^{\text{eff}}(h) = \{t:\ t \geq s_0,\ t \leq s_{S-1-h}\}.$$

- Drop edge rows to avoid peeking into the future.
- Multi-horizon ($h = 1, 5, 20$) handled with separate heads and effective indices.

Walk-Forward Splits

$$\mathcal{I} = \text{dates}(\text{SPY}),$$
$$\mathcal{T}_k = \mathcal{I}[k \cdot 252 : k \cdot 252 + 2520], \quad \mathcal{S}_k = \mathcal{I}[k \cdot 252 + 2520 : k \cdot 252 + 2520 + 252].$$

- Fit scaler on $\mathcal{T}_k$; apply to $(\mathcal{T}_k, \mathcal{S}_k)$.
- Train ESN/deep baselines on $\mathcal{T}_k$; evaluate on $\mathcal{S}_k$.
- Persist $\{\texttt{train.csv}, \texttt{test.csv}, \texttt{scaler.json}\}$ per fold.

Causal Windowing for Sequence Models

- For deep baselines (LSTM/Transformer/TCN), we use a **causal window** of length L:

$$X_t \;=\; \big[z_{t-L+1},\, z_{t-L+2},\, \ldots,\, z_t\big] \in \mathbb{R}^{L\times d}, \qquad \text{label: } y_{t,h}.$$

- **Train indices** (Fold k): $\mathcal{I}_k^{\mathrm{tr}}(L,h) = \{t \in \mathcal{T}_k :\; t \geq t_0 + L - 1,\; t \leq t_{T-1-h}\}$.
- **Test indices**: $\mathcal{I}_k^{\mathrm{te}}(L,h) = \{t \in \mathcal{S}_k :\; t \geq s_0 + L - 1,\; t \leq s_{S-1-h}\}$.
- Optional *bridge*: prepend the last $L-1$ train frames to the test stream (legal, causal).

Feature Engineering (u_t)

Let P_t be the adjusted (if available) or close price (PX); V_t volume.

$$\begin{aligned}
\text{Log return:}\quad & r_t = \log P_t - \log P_{t-1} \\
\text{Lagged cum. returns:}\quad & \texttt{ret_2}_t = r_{t-1} + r_{t-2}, \quad \texttt{ret_5}_t = \sum_{k=1}^{5} r_{t-k} \\
\text{Realized vol (ann.):}\quad & \texttt{vol_20}_t = \sqrt{252} \cdot \mathrm{std}(r_{t-19:t}) \\
\text{Moving avgs:}\quad & \texttt{ma_10}_t = \mathrm{mean}(P_{t-9:t}), \quad \texttt{ma_20}_t = \mathrm{mean}(P_{t-19:t}) \\
\text{MA gap:}\quad & \texttt{ma_gap}_t = \frac{P_t}{\texttt{ma_20}_t} - 1 \\
\text{RSI-14:}\quad & \Delta_t = P_t - P_{t-1}, \quad G_t = \mathrm{mean}(\max(\Delta, 0))_{14}, \\
& L_t = \mathrm{mean}(\max(-\Delta, 0))_{14}, \mathrm{RSI}_{14,t} = 100 - \frac{100}{1 + G_t/L_t} \\
\text{Volume z-score:}\quad & \texttt{vol_z}_t = \frac{V_t - \mu_{60}(V)}{\sigma_{60}(V)} \quad \text{(if } V \text{ exists; else NaN)} \\
\text{Calendar:}\quad & \texttt{dow}_t \in \{0, \dots, 6\} \text{ (day-of-week)}
\end{aligned}$$

Rationale: momentum (returns), trend (MAs/gap), risk/scale (vol/volume), seasonality (weekday); interpretable, low-variance inputs for ESN readouts.

Light Baseline Readout & Toy Backtest

Ridge readout on standardized features.

$$\hat{y}_{t+h} = w_h^\top z_t + b_h, \qquad (w_h, b_h) = \arg\min \sum_{t\in\mathcal{T}} \big(\hat{y}_{t+h} - y_{t+h}\big)^2 + \lambda \|w_h\|_2^2, \;\; h \in \{1, 5, 20\}.$$

Reported test metrics: RMSE, MAE, R^2, *Directional Accuracy* $\Pr[\mathrm{sign}(\hat{y}) = \mathrm{sign}(y)]$.

Toy sign P&L diagnostic (not a trading claim).

$$\mathrm{position}_t = \mathrm{sign}(\hat{y}_t), \qquad \mathrm{pnl}_t = \mathrm{position}_t \cdot y_t - \mathrm{cost} \cdot |\Delta \mathrm{position}_t| \quad (\mathrm{cost} = 1\ \mathrm{bp}).$$

Purpose: sanity-check if forecasts carry usable direction after costs (summarize Avg-PnL, Vol, Sharpe, Hit Ratio, Turnover).

LSTM Regressor

- One-layer LSTM (hidden H_{lstm}), final state $\tilde{x}_t \to$ linear head:

$$\hat{y}_{t,h} = v_h^\top \tilde{x}_t + c_h.$$

- LSTM cell (per time τ):

$$\begin{aligned} i_\tau &= \sigma(W_i x_{\tau-1} + U_i z_\tau + b_i), \quad f_\tau = \sigma(W_f x_{\tau-1} + U_f z_\tau + b_f), \\ o_\tau &= \sigma(W_o x_{\tau-1} + U_o z_\tau + b_o), \quad \tilde{c}_\tau = \tanh(W_c x_{\tau-1} + U_c z_\tau + b_c), \\ c_\tau &= f_\tau \odot c_{\tau-1} + i_\tau \odot \tilde{c}_\tau, \quad x_\tau = o_\tau \odot \tanh(c_\tau). \end{aligned}$$

- Loss: MSE; optimizer: Adam; validation: last 10% of train (chronological).

Transformer Encoder (Causal Mask)

- Input $X_t \in \mathbb{R}^{L \times d}$ is linearly projected to d_{model} and added sinusoidal positions.
- Self-attention with **causal mask** M (upper-triangular $-\infty$):

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}} + M\Big)V.$$

- Final token representation $\tilde{x}_t \to$ linear head: $\hat{y}_{t,h} = v_h^\top \tilde{x}_t + c_h$.
- Loss: MSE; regularization: dropout, weight decay; small L preferred on daily data.

Temporal ConvNet (TCN)

- Causal, dilated 1-D convolutions with residual blocks:

$$y_\tau = \sum_{m=0}^{k-1} W_m \, x_{\tau - d\cdot m} \quad (\text{dilation } d, \ \text{kernel } k).$$

- Receptive field grows exponentially with layers, efficient for local motifs.
- Last-time embedding $\tilde{x}_t \to$ linear head: $\hat{y}_{t,h} = v_h^\top \tilde{x}_t + c_h$.

ESN Design: Spectral Radius & Leak

- Reservoir W initialized sparse; scale to target γ:

$$\rho(W) \leftarrow \text{power\_iter}(W), \qquad W \leftarrow \frac{\gamma}{\rho(W)}\,W.$$

- **Contraction factor**: $\kappa = a\,\|W\|_2 \approx a\,\rho(W) = a\gamma$; require $\kappa < 1$.
- **Leak** a: low-pass memory; smaller $a \Rightarrow$ smoother state, longer effective memory.
- **Washout** w: discard initial transients $\Rightarrow$ stable state–label alignment.

Objectives & Regularization (Per Horizon h)

- **Point objective (common)**:

$$\min_{\theta_h} \frac{1}{|\mathcal{T}_k^{\text{eff}}(h)|} \sum_{t \in \mathcal{T}_k^{\text{eff}}(h)} \left(\hat{y}_{t,h}(\theta_h) - y_{t,h}\right)^2.$$

- **ESN readout**: closed-form ridge $\Rightarrow$ fast CV over α.
- **Deep baselines**: early stopping on chronological val; dropout / weight decay.
- **Ensembling (optional)**: ESN seed-averaging over reservoirs to reduce variance.

Metrics & Backtest (Decision Proxy)

- $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(\hat{y} - y)^2}$, $\quad \mathrm{MAE} = \frac{1}{n}\sum|\hat{y} - y|$, $\quad R^2 = 1 - \frac{\sum(\hat{y}-y)^2}{\sum(y-\bar{y})^2}$.
- Directional accuracy: $\mathrm{DA} = \frac{1}{n}\sum \mathbf{1}\{\mathrm{sign}(\hat{y}) = \mathrm{sign}(y)\}$.
- Sign strategy (*toy*): $p_t = \mathrm{sign}(\hat{y}_t)$,

$$\mathrm{PnL}_t = p_t\, y_t - c\,|p_t - p_{t-1}|, \quad \mathrm{Sharpe} = \frac{\overline{\mathrm{PnL}}}{\mathrm{Std}(\mathrm{PnL})}\sqrt{252}.$$

- $\mathrm{Turnover} = \frac{1}{n}\sum|p_t - p_{t-1}|$, Hit ratio $= \frac{1}{n}\sum \mathbf{1}\{p_t\, y_t > 0\}$.

Tensors & Shapes (Example)

Assume $T{=}2520$, $S{=}252$, $d{=}10$, $L{=}32$, $H{=}600$, $w{=}100$.

- **Ridge**: $Z_{\text{tr}} \in \mathbb{R}^{(T-h)\times d}$, $\quad Z_{\text{te}} \in \mathbb{R}^{(S-h)\times d}$.
- **ESN**: $H \in \mathbb{R}^{(T-w-h)\times(H+1)}$, $\quad Y_h \in \mathbb{R}^{T-w-h}$, $\quad$ test preds $\in \mathbb{R}^{S-h}$.
- **LSTM/TF/TCN**: $X_{\text{tr}} \in \mathbb{R}^{(T-(L-1)-h)\times L\times d}$, $\quad$ same for test with S.

$$\text{E.g., } h{=}1: \quad H \in \mathbb{R}^{2419\times 601}, \quad X_{\text{tr}} \in \mathbb{R}^{(2520-31-1)\times 32\times 10}.$$

Compute Complexity & Practical Notes

- **ESN (training)**: state roll $O(T \cdot \mathrm{nnz}(W))$; ridge solve $(H^\top H + \alpha I)^{-1}$ in $O((H+1)^3)$ once per h (small H feasible, or use Cholesky).
- **LSTM**: $O(T \cdot L \cdot d \cdot H_{\mathrm{lstm}})$ per epoch.
- **Transformer**: $O(T \cdot L^2 \cdot d_{\mathrm{model}})$ per layer per epoch (attention is L^2).
- **TCN**: $O(T \cdot L \cdot C \cdot k)$ per epoch (causal dilated convs).
- **Practice**: prefer smaller L and moderate hidden sizes on daily returns; ESN enables broad hyperparam sweeps; deep models need careful regularization.

End-to-End Pipeline & Artifacts

- **Acquisition** → **Parsing** → **Feature Engineering** (u_t) → **Targets** (y_{t+h})
- **Walk-Forward Splits** (Train=2520d, Test=252d, Step=252d) with **train-only** scaling.
- **Models**: Ridge (lower bound), ESN (state-space), LSTM, Transformer, TCN.
- **Outputs per fold**: `train.csv`, `test.csv`, `scaler.json`, `preds_h*.csv`, `metrics_h*.json`.
- **Reproducibility**: config slugs → `data/experiments/{model}/{exp_id}/fold_k/`.

All transforms are causal: features at time t depend only on $\{1{:}t\}$.

Leakage-Safe Preprocessing

Standardization (per feature j) on train-only window $\mathcal{T}_k$:

$$\mu_j^{(k)} = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} u_t^{(j)}, \quad \sigma_j^{(k)} = \sqrt{\frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \left(u_t^{(j)} - \mu_j^{(k)}\right)^2},$$

$$z_t^{(j)} = \frac{u_t^{(j)} - \mu_j^{(k)}}{\sigma_j^{(k)}} \ \ \forall t \in \mathcal{T}_k \cup \mathcal{S}_k.$$

Causality claim. If u_t is constructed via backward-looking windows, i.e. $u_t = \mathcal{F}(P_{1:t})$, then z_t uses only $\{u_\tau\}_{\tau \le t}$ and (μ, σ) fitted on $\mathcal{T}_k$.

$\Rightarrow$ **No look-ahead leakage.**

Walk-Forward Indices & Effective Sets

Let common date index $\mathcal{I}$ and fold k:

$$\mathcal{T}_k = \mathcal{I}[k \cdot S : k \cdot S + T], \qquad \mathcal{S}_k = \mathcal{I}[k \cdot S + T : k \cdot S + T + S],$$

with $T{=}2520$, $S{=}252$.

$$\text{Targets: } y_{t,h} = \sum_{i=1}^{h} r_{t+i} \;\Rightarrow\; \mathcal{T}_k^{\text{eff}}(h) = \{t \in \mathcal{T}_k \mid t \leq t_{T-1-h}\},$$

$$\mathcal{S}_k^{\text{eff}}(h) = \{t \in \mathcal{S}_k \mid t \leq s_{S-1-h}\}.$$

Deep models require windows:

$$X_t = [z_{t-L+1}, \ldots, z_t] \Rightarrow \; t \geq t_0 + L - 1.$$

Cross-Asset Exogenous Features (Causal)

- Augment u_t with lagged exogenous signals $\tilde{z}_t$ from other tickers:

$$\tilde{z}_t = \Big[\ \underbrace{r^{(\mathrm{VIX})}_{t-1:t-5}}_{\text{vol-of-vol}},\ \underbrace{\Delta \log(\mathrm{USDINR})_{t-1:t-5}}_{\text{FX}},\ \underbrace{r^{(\mathrm{Crude})}_{t-1:t-5}}_{\text{commodities}}\ \Big].$$

- All exogenous features are aligned by date and **lagged** to avoid contemporaneous leak.
- Final feature vector: $u_t = [\text{core features}_t,\ \tilde{z}_t]$.

News Features: Construction (Lookback L days)

Aggregates for a symbol s:

$$\texttt{news\_count\_}L(t) = \sum_{\tau=t-L+1}^{t} \mathbf{1}\{\text{headline}_\tau \text{ for } s\},$$

$$\texttt{news\_sent\_mean\_}L(t) = \frac{1}{\texttt{news\_count}} \sum_{\tau} \hat{m}_\tau,$$

$$\texttt{news\_sent\_std\_}L(t) = \sqrt{\frac{1}{\texttt{news\_count}} \sum_{\tau} (\hat{m}_\tau - \overline{m})^2},$$

where $\hat{m}_\tau$ is headline sentiment. **Text factor** via TF–IDF PCA:

$$X_t = \text{tfidf}(\text{bag-of-words in window}), \quad \texttt{news\_tfidf\_pc1\_}L(t) = \mathrm{PC}_1^\top X_t.$$

All computed **before** or at t (causal).

ESN Echo-State via Contraction (Sufficient Condition)

State update with 1-Lipschitz ϕ (e.g., tanh):

$$x_t = (1-a)x_{t-1} + a\,\phi(Wx_{t-1} + W_{\text{in}}\tilde{z}_t).$$

Two trajectories (x_t) and (x'_t) under same inputs satisfy

$$\|x_t - x'_t\| \le \big((1-a) + a\|W\|_2\big)\,\|x_{t-1} - x'_{t-1}\|.$$

If $L_\phi{=}1$ and $\|W\|_2 \approx \rho(W) = \gamma$, a sufficient contraction is

$$\kappa \equiv (1-a) + a\gamma < 1 \quad \Rightarrow \quad \|x_t - x'_t\| \le \kappa^t \|x_0 - x'_0\| \to 0.$$

Hence: unique input-driven state (ESP) and fading memory.

Threats to Validity & Mitigations

- **Nonstationarity/regime shifts**: use walk-forward; consider rolling re-fit, regime features.
- **Multiple testing/overfitting**: restrict grids; report all runs; use DM tests; consider deflated Sharpe.
- **Data issues/holidays**: strict intersection of trading calendars; forward-fill only safe fields; no future peeks.
- **Scale drift**: volatility-normalized targets; calibration layers (isotonic) on validation.
- **Variance of reservoirs**: seed-ensembles; average readouts; report dispersion.

Toy Backtest Design (Decision Proxy)

Position rule (per horizon h):

$$\pi_t \;=\; \operatorname{sgn}\big(\hat{y}_{t,h}\big) \in \{-1, 0, 1\}, \quad \Delta\pi_t = \pi_t - \pi_{t-1}.$$

PnL with linear frictions (c bps/trade):

$$\mathrm{pnl}_t \;=\; \pi_t\, y_{t,h} \;-\; c\,|\Delta\pi_t|, \qquad \text{Sharpe} \;=\; \frac{\overline{\mathrm{pnl}}}{\widehat{\sigma}(\mathrm{pnl})} \times \sqrt{252}.$$

Diagnostics: hit ratio $\Pr[\operatorname{sgn}(\hat{y}) = \operatorname{sgn}(y)]$, turnover $\sum_t |\Delta\pi_t|/T$.

Notes. (i) *Diagnostic only*, not a strategy. (ii) Exact same cost applied across models for fairness. (iii) No leverage, no slippage modeling.

Empirical Snapshot (Fold 0, h=1 day)

- **Magnitude (RMSE/MAE):** Ridge lowest error $\Rightarrow$ high-bias/low-variance wins at daily noise scale.
- **Direction/Sharpe (toy):** ESN achieves strongest risk-adjusted signal despite larger RMSE.
- **Dispersion:** Transformer/TCN sensitive to hyperparams; calibration impacts R^2 vs. hit ratio.

Ablation Highlights: ESN Controls

Contraction factor $\kappa = (1-a) + a\gamma$ with $\gamma \approx \rho(W)$.

$$\text{Memory depth } \uparrow \text{ as } \kappa \uparrow \quad \text{but} \quad \text{stability margin } \downarrow .$$

Observed trends (qual.):

1. Moderate $\gamma \in [0.85, 0.95]$ and $a \in [0.3, 0.6] \Rightarrow$ best hit ratio.
2. Larger H improves Sharpe up to variance limits; ridge α regularizes over-active states.
3. Washout $w \approx 100$ stabilizes readout; too small w harms ESP.

$$\text{Bias–variance via } \alpha : \quad W_{\text{out}} = (H^\top H + \alpha I)^{-1} H^\top Y.$$

Ablation Highlights: Deep Baselines

Window length L (context) vs. data scale:

$X_t \in^{L\times d} \Rightarrow$ LSTM/TCN stable for $L \leq 64$; Transformer needs stronger regularization.

Regularization: dropout/weight decay reduce overfit but may attenuate sign.
Calibration: post-hoc scaling improves R^2 without degrading hit ratio.

$$\hat{y}^{\text{cal}} = a^\star \hat{y} + b^\star, \quad (a^\star, b^\star) = \arg\min_{a,b} \sum_{t\in\text{val}} (y_t - a\hat{y}_t - b)^2.$$

Default Hyperparameters

Model	**Key Params (defaults)**	**File**
Ridge	α=1.0	`src/models/ridge_readout.py`
ESN	H=500, γ=0.9, a=1.0, density=0.1, w=100, α=1.0	`src/models/esn.py`
LSTM	L=32, hidden=128, layers=1, epochs=10	`src/models/lstm.py`
Transformer	L=32, d_{model}=128, heads=4, layers=2, epochs=10	`src/models/transformer.py`
TCN	L=32, channels=(64, 64), k=3, epochs=10	`src/models/tcn.py`

Tensor Shapes (Summary)

Let $(T, S) = (2520, 252)$, horizon h, window L, features d, ESN size H, and washout w.

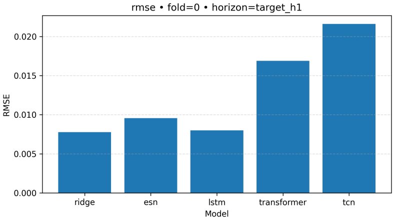
Model	Train tensors (Fold–1)	Test tensors (Fold–1)
Ridge	$Z_{\text{tr}} \in \mathbb{R}^{(T-h)\times d}, \ y_{\text{tr}} \in \mathbb{R}^{T-h}$	$Z_{\text{te}} \in \mathbb{R}^{(S-h)\times d}, \ y_{\text{te}} \in \mathbb{R}^{S-h}$
ESN	$H \in \mathbb{R}^{(T-w-h)\times(H+1)}, \ Y_h \in \mathbb{R}^{T-w-h}$	states $\{x_{s_i}\}$, preds $\hat{Y}_h \in \mathbb{R}^{S-h}$
LSTM	$X_{\text{tr}} \in \mathbb{R}^{(T-(L-1)-h)\times L\times d}$	$X_{\text{te}} \in \mathbb{R}^{(S-(L-1)-h)\times L\times d}$
Transformer	same as LSTM (internal d_{model})	same as LSTM
TCN	same as LSTM (internal C_{out})	same as LSTM

Results — Fold 0 (h=1 day)

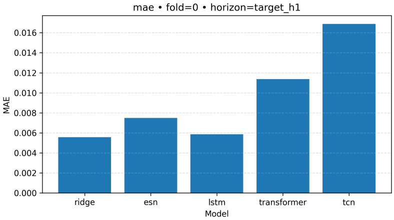
model	fold	horizon	RMSE	MAE	R^2	DirAcc	AvgPnL	Vol	Sharpe	Turnover
ridge	0	h1	0.008335	0.005989	-0.005	0.492	0.000077	0.008323	0.147	0.687
lstm	0	h1	0.009392	0.007025	-0.276	0.508	0.000208	0.008325	0.396	0.337
esn	0	h1	0.010098	0.007852	-0.475	0.528	0.000841	0.008278	**1.612**	0.853
transformer	0	h1	0.014366	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

Notes: All models use identical features, scaling, and test window. Sharpe from toy sign backtest with 1 bp cost.

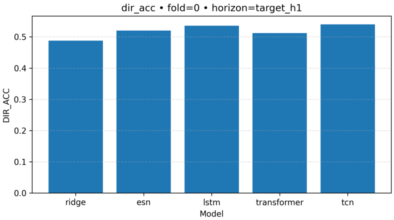
Headline Metrics — Fold 0 (h=1)



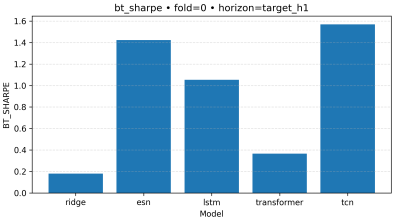
RMSE (fold 0, h=1)



MAE (fold 0, h=1)



Directional Accuracy (fold 0, h=1)



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

Compute & Complexity

ESN: roll T steps with sparse W (density p):

$$\mathcal{O}(T \cdot pH^2) \text{ (state roll)} \quad + \quad \mathcal{O}(H^3 + TH^2) \text{ (closed-form ridge).}$$

LSTM/TCN:

$$\mathcal{O}(T \cdot L \cdot d \cdot H_{\text{hid}}) \text{ (per epoch).}$$

Transformer:

$$\mathcal{O}(T \cdot (L\, d_{\text{model}}^2 + L^2 d_{\text{model}})) \text{ (per epoch).}$$

Implication: ESN excels in sweep-heavy, fold-rich evaluation; Transformers need careful sizing.

Pipeline Schematic

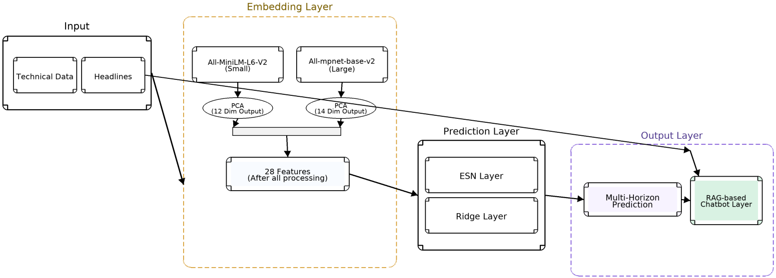


Figure: **Overall pipeline.** Technical bars and headlines → SBERT embeddings (MiniLM, MPNet) + PCA → 28-D features; leaky ESN with ridge readouts produces multi-horizon return forecasts under walk-forward splits; optional RAG chatbot explains/answers from headlines.

NLP Stack: Inputs → Embedding Layer

- **Inputs:** Technical data (numerical) & **Headlines** (text).
- **Two-encoder setup (diagram):**
  - `all-MiniLM-L6-v2` (*small* sentence encoder).
  - `all-mpnet-base-v2` (*large* sentence encoder).
- Each headline → dense sentence vector; downstream PCA compresses and stabilizes.
- Output of the "Embedding Layer" feeds a compact **28-d feature block**.

Sentence Embeddings (Dual Encoders)

Given headline h_t:

$$\mathbf{e}_t^{(\mathrm{mini})} = \mathrm{Enc}_{\mathrm{MiniLM}}(h_t), \qquad \mathbf{e}_t^{(\mathrm{mpnet})} = \mathrm{Enc}_{\mathrm{MPNet}}(h_t),$$

$$\tilde{\mathbf{e}}_t^{(\cdot)} \leftarrow \frac{\mathbf{e}_t^{(\cdot)}}{\|\mathbf{e}_t^{(\cdot)}\|_2} \quad \text{(L2-normalize for cosine geometry).}$$

- Two complementary encoders improve semantic coverage & robustness.
- Normalization reduces scale mismatch before PCA.

Dimensionality Reduction via PCA (Per Encoder)

For encoder matrix $X \in \mathbb{R}^{n\times d}$ (rows = headlines around t):

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i, \quad \hat{X} = X - \mathbf{1}\bar{\mathbf{x}}^\top, \quad \hat{X} = U\Sigma V^\top \text{ (SVD).}$$

Keep top-k components (V_k):

$$\mathbf{p}_t^{(\mathrm{mini})} = \hat{\mathbf{e}}_t^{(\mathrm{mini})} V_k^{(\mathrm{mini})}, \quad \mathbf{p}_t^{(\mathrm{mpnet})} = \hat{\mathbf{e}}_t^{(\mathrm{mpnet})} V_\ell^{(\mathrm{mpnet})}.$$

- **Diagram constraints:** e.g., $k{=}12$ (MiniLM), $\ell{=}14$ (MPNet).
- PCA filters noise & yields stable low-d embeddings for the model.

News Feature Fusion (28-d Block)

$$\mathbf{f}_t^{(\text{news})} = \left[\, \mathbf{p}_t^{(\text{mini})};\ \mathbf{p}_t^{(\text{mpnet})} \,\right] \in \mathbb{R}^{12+14=\mathbf{28}}.$$

- Concatenation preserves complementary semantics of both encoders.
- Optional post-PCA standardization per fold (train-only stats).
- This **28-d news block** is what the diagram denotes as the Embedding Layer output.

Merging with Technical Features

$$\mathbf{u}_t = \Big[\underbrace{\mathbf{u}_t^{(\text{tech})}}_{\text{returns/vol/MA/RSI/vol\_z/dow}}, \underbrace{\mathbf{f}_t^{(\text{news})}}_{\text{28-d PCA news block}} \Big],$$

- Train-only `StandardScaler` on $\mathbf{u}_t$ (per fold) prevents leakage.
- Unified feature vector feeds the ESN/Ridge (and deep) predictors.

RAG Layer: Purpose & Scope

- **Output layer:** *Multi-Horizon Prediction* $\rightarrow$ **RAG-based Chatbot Layer**.
- Goal: **grounded Q&A** on *recent headlines & model outputs.*
- Users can ask: "What drove today's signal?"; "Summarize last week's news relevant to SPY."

Retrieval: TF–IDF Index Over Headlines

Indexing (rolling window, e.g., ± 30 days):

$$\mathrm{tfidf}(t,d) = \mathrm{tf}(t,d) \cdot \log \frac{N+1}{\mathrm{df}(t)+1}, \qquad \mathrm{score}(q,d) = \cos\big(\mathbf{v}(q), \mathbf{v}(d)\big).$$

- Tokenize & build TF–IDF matrix; keep top-k docs by cosine similarity.
- Optionally tag documents with dates/tickers for temporal filtering.

Grounded Generation: Prompting & Fusion

- **Context pack:** top-k headlines (title, snippet, date, source) + current model outputs (e.g., $\hat{y}_{t+h}$ summaries).
- **Prompt skeleton:**

```
You are a financial explainer. Use ONLY the provided headlines and
model summaries to answer. Cite date/source snippets inline.
Question: ”user query”
Context:
[1] YYYY-MM-DD — Source title + snippet
...
Model: horizon-wise metrics or sign summaries
```

- **Guardrails:** refuse advice; note uncertainty; surface conflicting headlines.

RAG Quality: What We Measure

- **Retrieval:** Recall@k, MRR; freshness filter (recentness bias).
- **Answer quality:** Faithfulness (citation coverage), conciseness, hallucination rate.
- **Latency:** indexing time, query time, generation time; enable caching for hot windows.

Failure Modes & Lessons

- **High RMSE, decent Sharpe:** sign captured, magnitude miscalibrated $\Rightarrow$ add calibration layer.
- **Transformer collapse:** small data vs. big model $\Rightarrow$ reduce depth/heads, increase weight decay.
- **ESN variance across seeds:** ensemble K reservoirs $\Rightarrow$ average states or stack $[X^{(1)} \cdots X^{(K)}]$.
- **Feature drift:** RSI/vol windows stale in new regime $\Rightarrow$ retune window sizes or add regime flags.

Roadmap & Future Work

1. **Vol-normalized targets** and heteroscedastic readouts.
2. **Multi-asset ESN**: shared reservoir, asset-specific readouts $W_{\text{out}}^{(s)}$.
3. **Regime-aware** leak/spectral scheduling a_t, γ_t (safe projection).
4. **Probabilistic ESN**: direct quantile/state mixtures; CRPS training.
5. **Richer exogenous**: macro calendars, lagged cross-asset graphs, verified *causal* news signals.

Conclusions

- **SSM framing of ESN** yields controllable memory and efficient training; strong directional utility.
- **Ridge** provides a robust magnitude floor; deep baselines need careful regularization/calibration.
- **Protocol-first**: leakage-safe splits, train-only scalers, fold-wise reporting, DM tests.

Code & Artifacts: reproducible pipeline with per-fold CSV/JSON outputs.

Acknowledgments: Course Team (Data Science and AI Lab), IITM.

Questions?

References I

H. Jaeger, *The "Echo State" Approach to Analysing and Training Recurrent Neural Networks*, GMD Report 148, German National Research Center for Information Technology, 2001.

M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, 3(3):127–149, 2009.

J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press, 2012.

H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert, "Optimization and applications of echo state networks with leaky-integrator neurons," *Neural Networks*, 20(3):335–352, 2007.

R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, 2021. Available online: `https://otexts.com/fpp3/`.

R. Koenker and G. Bassett Jr., "Regression Quantiles," *Econometrica*, 46(1):33–50, 1978.

T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477):359–378, 2007.

D. H. Bailey, J. Borwein, M. Lopez de Prado, and Q. J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the AMS*, 61(5):458–471, 2014.

F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

References II

- J. Lin and G. Michailidis, "Deep Learning-based Approaches for State Space Models: A Selective Review," *arXiv preprint* arXiv:2412.11211v1 [stat.ML], Dec. 2024. Available at: `https://arxiv.org/abs/2412.11211`.
- Yahoo Finance Historical Data. Accessed programmatically via community Python package `yfinance`. URL: `https://finance.yahoo.com/`.
- Stooq: Free historical market data (equities, FX, indices). URL: `https://stooq.com/`.
- Alpha Vantage: Free API for financial time series. URL: `https://www.alphavantage.co/`.
- Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. URL: `https://fred.stlouisfed.org/`.

Thanks!