Multi-Horizon Financial Time-Series Forecasting with Uncertainty Quantification

Track 2 (Sequences + Regularization + Transformers)

Zinetov Alikhan Yernur Bidollin IT-2301
Astana IT University

February 2026

Abstract

We build a predictive intelligence system for financial time-series forecasting under the Track 2 requirements (Sequences, Regularization, Transformers). Using daily market data from Yahoo Finance (SPY, 2010–2026), we predict multi-horizon future log-returns for $h \in \{1, 5, 20\}$ trading days. We implement a GRU baseline with attention, a Transformer encoder comparison, regularization controls (dropout and weight decay), and uncertainty estimation via deep ensembles. We report quantitative results, ablations, and naive baselines, and discuss calibration issues at longer horizons.

Keywords: time-series forecasting, GRU, Transformer, attention, regularization, deep ensembles, uncertainty

1 Introduction

Forecasting financial time-series is challenging due to low signal-to-noise ratio, non-stationarity, and regime shifts. Practical decision-making (risk management, position sizing) requires not only point forecasts but also uncertainty estimates. This project targets Track 2 (Predictive Intelligence for Time-Series) by integrating: (i) sequence modeling (GRU), (ii) Transformer comparison, (iii) attention for temporal dependency interpretation, (iv) regularization techniques, and (v) uncertainty quantification.

2 Related Work (brief)

Deep learning has been widely adopted for time-series forecasting, with RNN variants (LSTM/GRU) modeling sequential dependencies and attention improving interpretability. Transformer encoders apply self-attention to capture long-range dependencies. For uncertainty, deep ensembles are a strong practical baseline that often improves calibration and robustness in high-noise settings.

3 Data and Problem Formulation

3.1 Dataset

We use daily OHLCV data for SPY (S&P 500 ETF) from Yahoo Finance with the following configuration:

- Interval: 1 day
- Date range: 2010-01-01 to 2026-02-08 (about 4,200 trading days)
- Chronological splits: train $\leq$ 2022-12-31, validation 2023-01-01–2023-12-31, test 2024-01-01–2026-02-08

3.2 Targets: Multi-Horizon Log-Returns

Let C_t be the close price at time t. For each horizon h, the target is the future log-return:

$$y_{t,h} = \log\left(\frac{C_{t+h}}{C_t}\right), \quad h \in \{1, 5, 20\}. \tag{1}$$

3.3 Features and Windowing

We compute a feature vector $x_t \in \mathbb{R}^D$ including:

- $r_t = \log(C_t/C_{t-1})$ (1-day log return)
- High-low spread $(H_t - L_t)/(C_t + \epsilon)$ and open-close change $(C_t - O_t)/(O_t + \epsilon)$
- RSI(14)
- Rolling return mean/std over windows $\{5, 10, 20, 60\}$
- Volume z-score over the same windows
- Moving average and exponential moving average over $\{5, 10, 20, 60\}$

The model input is a lookback window of length $L = 60$: $X_t = [x_{t-L+1}, \ldots, x_t] \in \mathbb{R}^{L \times D}$, with output $\hat{y}_t \in \mathbb{R}^3$.

4 Models

4.1 GRU Baseline with Attention

We implement a GRU encoder producing hidden states h_ℓ across the lookback window. An attention layer scores each timestep and forms a context vector:

$$\alpha_\ell = \text{softmax}(w^\top h_\ell), \tag{2}$$

$$c = \sum_{\ell=1}^{L} \alpha_\ell h_\ell, \tag{3}$$

$$\hat{y} = Wc + b. \tag{4}$$

4.2 Transformer Encoder Comparison

We implement a Transformer encoder operating on the input window with positional encodings, followed by pooling and a linear head to predict $\hat{y} \in \mathbb{R}^3$.

4.3 Regularization

We evaluate dropout $p \in \{0, 0.2\}$ and weight decay (AdamW) $\lambda \in \{0, 10^{-4}\}$.

4.4 Uncertainty via Deep Ensembles

We train K independent GRU models with different seeds. For predictions $\{\hat{y}^{(k)}\}_{k=1}^{K}$:

$$\mu = \frac{1}{K}\sum_{k=1}^{K} \hat{y}^{(k)}, \quad \sigma^2 = \frac{1}{K}\sum_{k=1}^{K} \left(\hat{y}^{(k)} - \mu\right)^2. \tag{5}$$

A nominal 90% interval is approximated as $\mu \pm 1.645\,\sigma$. We report empirical coverage (Cov@90).

5 Training Setup and Metrics

We use MSE loss, AdamW (lr=10^{-3}), gradient clipping (1.0), ReduceLROnPlateau scheduling, and early stopping.

5.1 Metrics

For each horizon h, we report MAE, RMSE, and directional accuracy (DirAcc). We also report mean metrics averaged across horizons.

6 Results

6.1 Main Model Comparison

Table 1: Main model comparison on the test set (mean across horizons).

Model	MAE_{mean}	$\text{RMSE}_{\text{mean}}$	$\text{DirAcc}_{\text{mean}}$	$\text{Cov@90}_{\text{mean}}$
GRU + Attention (single)	0.0231	0.0302	0.6029	–
Transformer Encoder	0.0466	0.0570	0.4425	–
Deep Ensemble (GRU, K=5)	0.0192	0.0252	0.4970	0.7324

6.2 Per-Horizon Performance

Table 2: Per-horizon performance for core models.

Model	Horizon	MAE	RMSE	DirAcc
GRU + Attention	1	0.0253	0.0288	0.5799
	5	0.0159	0.0222	0.5897
	20	0.0282	0.0395	0.6391
Transformer	1	0.0242	0.0306	0.5010
	5	0.0189	0.0257	0.5661
	20	0.0967	0.1146	0.2604
Ensemble (K=5)	1	0.0068	0.0104	0.5503
	5	0.0187	0.0237	0.3846
	20	0.0322	0.0415	0.5562

6.3 Ablation Studies

Table 3: Ablations (GRU family). Mean metrics across horizons.

Variant	MAE_{mean}	$\text{RMSE}_{\text{mean}}$	$\text{DirAcc}_{\text{mean}}$
GRU + Attention (dropout=0.2, wd=1e-4)	0.0231	0.0302	0.6029
GRU (no attention)	0.0231	0.0302	0.6029
Dropout=0 (GRU + attention)	0.0216	0.0285	0.5641
Weight decay=0	0.0231	0.0302	0.6029
Ensemble K=1 (no ensemble)	0.0231	0.0302	0.6029
Ensemble K=5	0.0192	0.0252	0.4970

6.4 Naive Baselines

Table 4: Naive baselines (mean across horizons).

Baseline	MAE_{mean}	$\text{RMSE}_{\text{mean}}$	$\text{DirAcc}_{\text{mean}}$
Zero return	0.0175	0.0230	0.0000
Last return	0.0191	0.0263	0.5155
Mean return (lookback mean)	0.0173	0.0230	0.6410

6.5 Visualizations

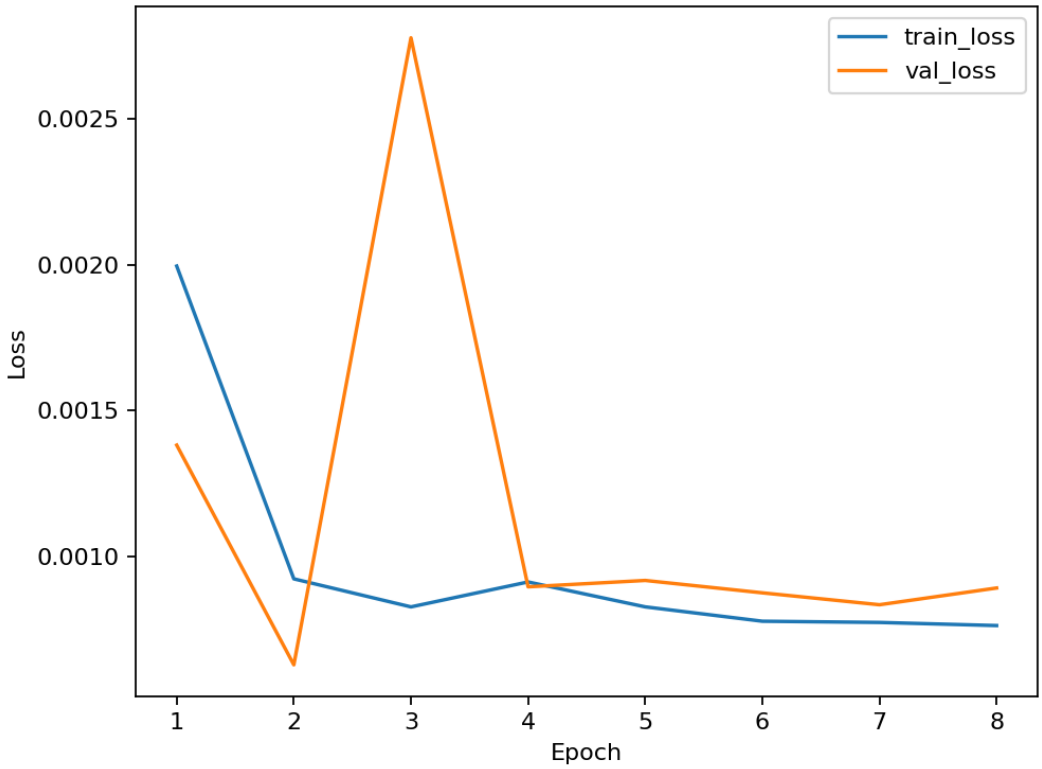


Figure 1: Training/validation learning curves (example run).

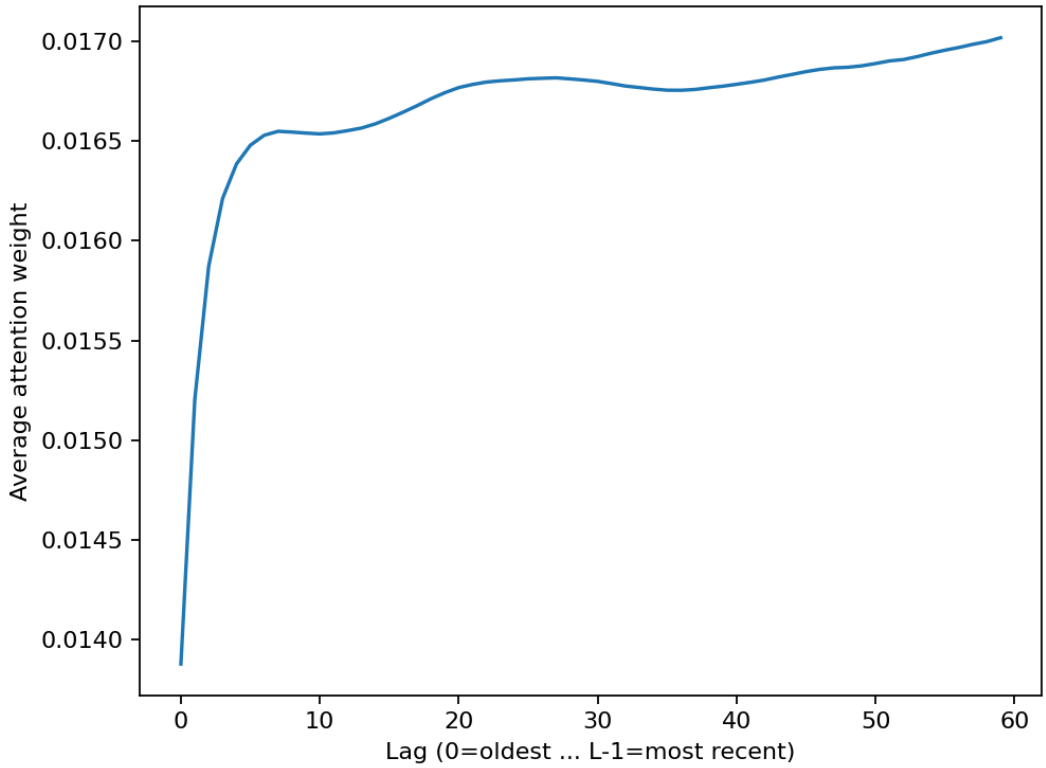
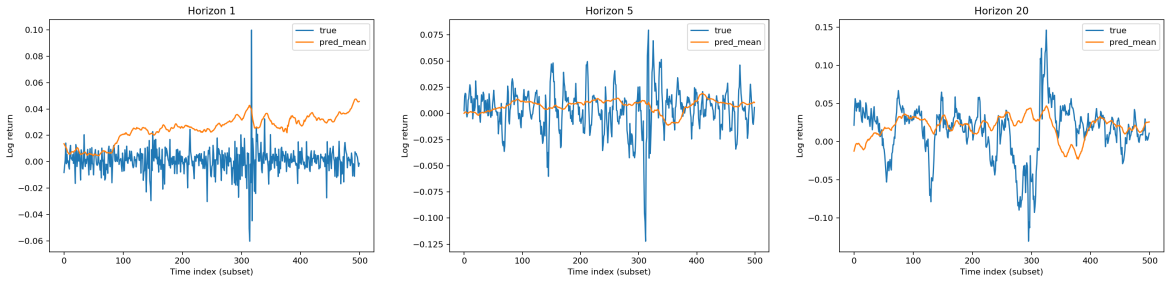


Figure 2: Average attention weights across the 60-day lookback window (GRU+Attention).



(a) $h = 1$ (b) $h = 5$ (c) $h = 20$

Figure 3: Predicted vs. true log-returns on the test set for each horizon.

7 Discussion

Transformer performance degraded at $h = 20$ suggesting sensitivity to configuration or limited inductive bias for this feature-driven setup. Dropout produced a trade-off between point error and directional accuracy. Ensemble improved MAE/RMSE and gave uncertainty intervals; coverage is high at $h = 1$ (0.943) but under-covers at $h = 20$ (0.465). Naive baselines are very competitive, reflecting low predictability of returns.

8 Limitations and Future Work

Single asset (SPY), no transaction-cost-aware backtesting, limited Transformer tuning, and imperfect long-horizon calibration. Future work: multi-asset setting, probabilistic heads (quantile regression), calibration, and economic utility backtests.

9 Conclusion

We implemented a Track 2 time-series system combining GRU, Transformer, regularization, and uncertainty via ensembles. Ensembles improve point metrics and provide useful intervals at short horizons; longer horizons require calibration.

References

[1] A. Vaswani et al., "Attention is All You Need," *NeurIPS*, 2017.

[2] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," *EMNLP*, 2014.

[3] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," *NeurIPS*, 2017.

[4] R. Ranaroussi, *yfinance: Yahoo Finance market data downloader*. `https://github.com/ranaroussi/yfinance`