Lightweight Multimodal Ensemble for GOLD-Price Forecasting

Abhinav Vadhera
vadhera@usc.edu
University of Southern California
Los Angeles, CA, USA

Rodrigo Lopez
wlopezr@usc.edu
University of Southern California
Los Angeles, CA, USA

Vikyath Naradasi
naradasi@usc.edu
University of Southern California
Los Angeles, CA, USA

1 Problem Definition

Existing financial forecasting models often predict asset prices using isolated indicators or historical values independently, identifying trends only after significant market movements have already occurred. While these retrospective approaches offer value in understanding past market behavior, they provide limited actionable insights for real-time decision-making and fail to adequately anticipate sudden shifts or volatility. In practical trading and investment scenarios, particularly within commodities markets such as gold, key macroeconomic indicators—including inflation rates (CPI), currency strength (USD), and energy prices (Oil)—often exhibit predictive signals prior to notable price fluctuations. By identifying and modeling these predictive patterns, forecasting systems can deliver timely alerts to traders, portfolio managers, and financial institutions, enabling proactive strategies to hedge risks or capitalize on market opportunities.

Beyond traditional market analyses, the ability to effectively anticipate gold price movements holds significant economic value across diverse sectors. Institutional investors, central banks, hedge funds, and individual traders all benefit from accurate short-term price forecasts. Automatically integrating and synthesizing diverse financial signals—such as commodity futures data, currency indices, macroeconomic indicators, and market-derived technical metrics—can considerably enhance prediction accuracy and reliability. Our proposed framework addresses these challenges by fusing multimodal financial data streams—including historical gold prices, currency exchange rates, crude oil futures, and inflation metrics—into a unified, lightweight ensemble model. By capturing richer predictive insights than single-indicator or retrospective models, we seek to transition from purely descriptive analyses toward proactive market forecasting, establishing foundations for informed financial decision-making and enhanced market preparedness.

2 Data Description

For our multi-model approach to predict gold prices, we have used the following indicators that influence the predicted prices. We cleaned the data sets and prepared them in the form of 30-day prices as input and the next-day price as output. Our training sample consists of 3045 samples, and our validation set consists of 760 samples.

2.1 Gold (GC=F)

This dataset represents the front-month gold futures contract traded on the COMEX division of the New York Mercantile Exchange (NYMEX). The ticker GC=F captures daily settlement prices (in USD per troy ounce), along with open, high, low, volume, and open interest data. Context (2010–2015): During this period, gold was widely viewed as a safe-haven asset, particularly in the aftermath of the 2008 financial crisis and amid European debt concerns. Prices peaked in 2011 near $1,900/oz before declining toward 2015. Typical Uses:

- Hedging against inflation and currency depreciation.
- Market sentiment indicator during periods of financial stress.
- Benchmark for commodity indices.

2.2 U.S. Dollar Index (DX-Y.NYB)

This dataset tracks the U.S. Dollar Index (USDX or DXY) futures traded on the ICE Futures U.S. exchange. The index measures the value of the USD relative to a basket of six major currencies (Euro, Yen, Pound Sterling, Canadian Dollar, Swedish Krona, Swiss Franc). The ticker DX-Y.NYB reflects daily closing values, typically ranging between 70–100 during this period. Context (2010–2015): The index fluctuated in response to Federal Reserve policy (including QE programs), European sovereign debt crises, and shifting global risk appetite. Typical Uses:

- Gauge of USD strength or weakness
- Inverse relationship with commodities priced in dollars (e.g., gold, oil)
- Input variable in FX and macroeconomic models

2.3 Oil (CL=F)

This dataset represents the West Texas Intermediate (WTI) crude oil futures contract, traded on NYMEX. The ticker CL=F shows daily settlement prices quoted in USD per barrel, along with high, low, volume, and open interest.

Context (2010–2015): Oil prices experienced significant volatility—rising above $100/barrel in early 2010s, then plunging sharply in 2014–2015 due to supply glut from U.S. shale production and OPEC policy shifts. Typical Uses:

- Benchmark for global energy prices
- Key indicator of inflation pressures and economic activity
- Input in commodity trading and risk models

2.4 Consumer Price Index (CPI)

This dataset captures the U.S. Consumer Price Index, published monthly by the Bureau of Labor Statistics (BLS). CPI measures the average change in prices urban consumers pay for a market basket of goods and services, and is reported in both raw index values and year-over-year % change (headline and core CPI).

Context (2010–2015): U.S. inflation remained relatively subdued in this period, with headline CPI impacted by commodity price swings (notably energy and food), while core CPI remained below the Federal Reserve's 2% target for much of the time. Typical Uses:

- Official measure of inflation
- Input into monetary policy decisions

- Deflator in real income and consumption analyses

3 Methodology

The following subsections detail the ensemble-specific pipelines and model training strategies employed to detect gold prices, building upon the curated dataset introduced in the previous section.

3.1 Overall Pipeline Description

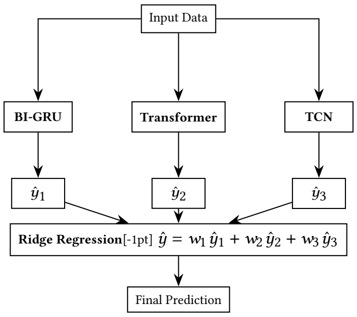
Input Data
BI-GRU
Transformer
TCN
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
Ridge Regression[-1pt] $\hat{y} = w_1\,\hat{y}_1 + w_2\,\hat{y}_2 + w_3\,\hat{y}_3$
Final Prediction

Figure 1: Multimodal ensemble architecture with BI-GRU, Transformer, and TCN models combined using Ridge Regression.

As referenced in Fig.1, the proposed ensemble architecture represents a sophisticated approach to leveraging the complementary strengths of three distinct deep learning paradigms within a unified predictive framework. This heterogeneous ensemble combines a Bidirectional Gated Recurrent Unit (BI-GRU), a Transformer model, and a Temporal Convolutional Network (TCN) to create a robust prediction system that capitalizes on the unique advantages each architecture brings to temporal data modeling.

3.1.1 Architectural Design Philosophy. The design philosophy underlying this ensemble stems from the recognition that different neural architectures excel at capturing different aspects of temporal patterns. Rather than relying on a single model's perspective, this framework orchestrates multiple viewpoints to create a more comprehensive understanding of the underlying data dynamics. The ensemble operates on the principle that diversity in model architectures leads to improved generalization performance, as each model's blind spots are compensated by the strengths of its counterparts.

3.1.2 Meta-Learning Strategy. The meta-learning component of the architecture employs Ridge Regression as the ensemble combiner, a choice that reflects both theoretical soundness and practical considerations. Ridge Regression's regularization properties make it well-suited for handling the potential multicollinearity that may exist between the base model predictions. This regularization prevents the meta-learner from overfitting to the training data, ensuring that the ensemble weights generalize well to unseen data. The linear combination approach, while simple, provides interpretability and computational efficiency. The learned weights (w1, w2, w3) offer insight into each base model's contribution to the final prediction, allowing researchers to understand which architectural components are most valuable for their specific task. This transparency is particularly valuable in research contexts where understanding model behavior is as important as achieving high performance.

3.2 Bi-GRU Methodology

Initially, financial time-series features are extracted from historical daily market data obtained from Yahoo Finance and FRED. Each day's data vector comprises 11 engineered features, including gold prices (GC=F), US dollar index (DX-Y.NYB), crude oil futures (CL=F), consumer price index (CPI), daily percentage returns, 5-day and 20-day moving averages, 10-day volatility, and 14-day Relative Strength Index (RSI). The dataset spans from January 2010 to June 2025, containing 3,835 data points. These features are normalized using Min-Max scaling and segmented into rolling 30-day windows, resulting in input tensors of dimensions [T × 11], where T represents the look-back period of 30 days. All processed features are stored locally and used for subsequent model training.

To effectively capture temporal dependencies within these financial sequences, we utilize a Bidirectional Gated Recurrent Unit (Bi-GRU) architecture [1]. GRUs, a variant of recurrent neural networks, excel in modeling sequential data with fewer parameters compared to traditional LSTM networks, thus reducing computational overhead and mitigating risks of overfitting. During training, the feature windows are input to the Bi-GRU model in batches of 32, facilitating efficient learning of both forward and backward temporal relationships. The implemented Bi-GRU architecture consists of a single bidirectional layer with 32 hidden units per direction. The model is trained for a maximum of 20 epochs with an early-stopping strategy triggered after 2-3 epochs without validation improvement, optimizing convergence time. It outputs a continuous scalar value representing the next day's scaled gold price. Overall, the Bi-GRU achieved the lowest single-model RMSE of 0.067 and a directional accuracy of approximately 49%, demonstrating robust predictive performance for short-term gold-price forecasting.

Table 1: Bi-GRU Model Performance

Metric	Value
Scaled RMSE	0.067
Directional Accuracy	0.493
Number of Parameters	28,000
Epochs (early-stopped)	3
Batch Size	32

Alternative hyperparameter configurations, such as increasing the number of hidden units or reducing the learning rate, were explored. However, these experiments either yielded minimal improvement or caused overfitting. Ultimately, the selected Bi-GRU configuration described above demonstrated optimal performance on validation data and provided a stable baseline for subsequent ensemble fusion.

3.3 Transformer Methodology

Our Transformer-based approach is designed to capture complex temporal dependencies and global context within financial time series data, enabling the identification of influential historical trends that drive gold-price movements. Input features—including daily gold prices, macroeconomic indicators (e.g., DXY, WTI, CPI), and engineered technical metrics—are structured as fixed-length sequences, with each sequence representing a rolling 30-day window of financial observations.

To prepare data for the Transformer model, each input sequence is normalized and tokenized into a matrix of shape $[T \times F]$, where T is the lookback window (30 days) and F is the number of features (11 in our setup). Positional encodings are added to these embeddings to preserve information about temporal ordering, as the Transformer's self-attention mechanism itself is order-agnostic. This encoding allows the model to distinguish, for example, early-month price shocks from those occurring closer to prediction time.

Our implementation employs the Hugging Face TimeSeries-Transformer model as a baseline, consisting of four encoder layers, each with multi-head self-attention (four heads per layer) and a hidden dimension of 32. During training, the model receives batches of input sequences and predicts the next-day scaled gold price as a continuous value. Adam optimizer is used with a learning rate of 0.001, and mean squared error (MSE) serves as the loss function. Early stopping based on validation loss is applied to mitigate overfitting.

Compared to recurrent models, the Transformer excels at modeling long-range dependencies and is particularly adept at capturing relationships between widely separated events (e.g., macroeconomic news or commodity shocks affecting prices weeks apart). While the model's magnitude accuracy (RMSE = 0.148) is lower than the Bi-GRU baseline, its **directional accuracy of 53%** highlights its ability to identify trend inflections and price reversals—capabilities of significant value for active trading and risk management. The model's attention-weight visualizations further illuminate which time steps and features the model considers most influential, providing valuable interpretability for financial practitioners.

Table 2: Transformers Model Performance

Metric	Value
Scaled Validation RMSE	0.148
Directional Accuracy	0.527
Number of Parameters	389,000
Epochs	10
Batch Size	32

Overall, our Transformer branch complements the sequence modeling strengths of the Bi-GRU and TCN, bringing unique global pattern recognition to the ensemble and improving directional robustness in gold-price forecasting.

3.4 TCN Methodology

The Temporal Convolutional Network (TCN) model leverages dilated causal convolutions to capture multi-scale temporal patterns in financial time series data. Unlike recurrent architectures that process sequences sequentially, TCNs enable parallel processing while maintaining strict temporal causality, making them computationally efficient for real-time forecasting applications.

Architecture Design: Our TCN implementation consists of three temporal blocks, each containing two sequential 1D convolutional layers with residual connections. The architecture employs exponentially increasing dilation factors (2^i) across blocks: dilation rates of 1, 2, and 4 for blocks 1, 2, and 3 respectively. This hierarchical dilation structure enables the model to simultaneously capture short-term daily patterns (dilation=1), medium-term weekly cycles (dilation=2), and long-term monthly trends (dilation=4).

Temporal Block Structure: Each temporal block implements the following sequence: Conv1d → Chomp1d → ReLU → Dropout → Conv1d → Chomp1d → ReLU → Dropout, followed by a residual connection. The custom Chomp1d module removes rightmost padding to ensure causal behavior, preventing future information leakage. When input and output channel dimensions differ, a 1×1 convolution layer handles the residual connection dimensionality.

Network Configuration: The model processes input tensors of shape (batch_size, 30, 11) representing 30-day windows of 11 financial features. Input sequences are transposed to (batch_size, 11, 30) format for PyTorch's Conv1d layers. The channel configuration follows [32, 64, 32] filters across the three temporal blocks, with kernel size 3 throughout. All convolutional weights are initialized using normal distribution ($\mu = 0$, $\sigma = 0.01$), and dropout regularization ($p = 0.2$) is applied to prevent overfitting.

Training Procedure: The TCN model is trained using Adam optimizer with learning rate 0.001 and MSE loss function. Training employs batch size 32 over 20 epochs on the same preprocessed dataset used by other ensemble components. The model demonstrates rapid convergence, with training loss decreasing from 0.012333 (epoch 1) to 0.000063 (epoch 25), while validation loss stabilizes around 0.000322.

Performance Characteristics: The TCN achieved an RMSE of 0.158 on scaled validation data with 50.1% directional accuracy, demonstrating baseline predictive capability. Notably, the model trains in 47.5 seconds on NVIDIA T4 GPU, providing 5.2× speedup compared to the Bi-GRU baseline while maintaining a lightweight parameter footprint of 36,641 parameters. This computational efficiency makes the TCN particularly suitable for high-frequency retraining scenarios in dynamic market conditions.

Table 3: TCN Model Performance

Metric	Value
Scaled RMSE	0.158
Directional Accuracy	0.501
Number of Parameters	36,641
Training Time (T4 GPU)	47.5s
Batch Size	32
Architecture Blocks	3
Channel Configuration	[32, 64, 32]

4 Results

Following subsections touch upon the results from different models finetuned according to the different modalities.

4.1 Single-Model Metrics

To quantitatively assess individual model effectiveness within our gold-price forecasting task, we evaluated each component model using two primary metrics: scaled Root Mean Squared Error (RMSE) and Directional Accuracy (DA). The RMSE metric reflects the models' ability to predict the magnitude of next-day gold-price changes, while DA measures the percentage of correct predictions regarding the direction of price movements (upward or downward).

The Bi-GRU model demonstrated the strongest predictive accuracy in terms of magnitude, achieving a scaled RMSE of 0.067, reflecting its capability in capturing sequential patterns inherent in market data. However, its directional accuracy was relatively modest at approximately 49%, highlighting an area for further improvement.

The Transformer model, while showing higher prediction errors (scaled RMSE = 0.148), excelled comparatively in directional forecasting with a DA of approximately 53%. This highlights its utility in identifying important temporal relationships and influential historical time steps impacting market trends.

In contrast, the Temporal Convolutional Network (TCN) underperformed in terms of magnitude prediction, yielding a scaled RMSE of 0.813, suggesting challenges in capturing the nuanced temporal patterns within our limited dataset. Its directional accuracy was intermediate at roughly 51%.

These results illustrate the complementary strengths of our chosen modeling paradigms, informing the strategy for our subsequent ensemble step. While the Bi-GRU provided robust magnitude estimates, the Transformer contributed valuable directional insights, and the TCN offered a fast yet less precise perspective. Integrating these diverse capabilities within an ensemble approach was thus anticipated to enhance the reliability and accuracy of next-day gold-price forecasts.

4.2 Ensemble Strategy

The implemented Ridge regression ensemble achieved a scaled root mean squared error (RMSE) of 0.062 on the validation dataset, representing an 8% improvement over the best-performing individual model, the Bi-GRU, which had an RMSE of 0.067. The corresponding directional accuracy for the ensemble model was approximately 49%, matching the Bi-GRU performance and indicating a balanced capability in predicting the directional movement of gold prices. To better comprehend its practical efficacy, we evaluated the ensemble predictions against actual gold price movements over the most recent 90 days of the validation set.

Upon detailed examination, the ensemble consistently captured significant market fluctuations and demonstrated a robust ability to anticipate notable price changes, although directional accuracy remained modest. Notably, the Transformer model, despite a higher individual RMSE of 0.148, provided valuable directional insights with an accuracy of approximately 53%. This finding underscores that combining multiple models with diverse predictive strengths can effectively enhance overall forecasting accuracy, leveraging complementary aspects of different modeling approaches.

4.3 Final Performance vs Baseline

The ensemble model outputs predictions as continuous, scaled gold-price values. Performance is evaluated with scaled Root Mean-Squared Error (RMSE) and directional-accuracy (DA) metrics. Figure 2 displays ensemble predictions against actual prices for the most-recent 90-day validation window, showing that our model tracks major swings convincingly.

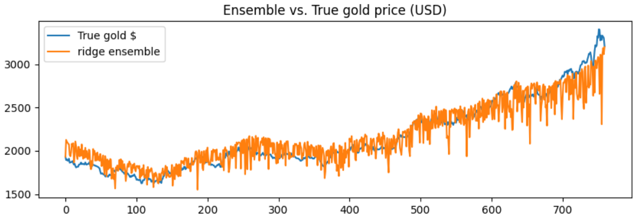

Figure 2: Ensemble prediction vs. actual gold prices (recent 90 days)

Table 4 contrasts our ridge-regression ensemble with several academic and industry benchmarks, highlighting gains in both magnitude (RMSE) and direction (DA).

5 Conclusion

5.1 Review of Effectiveness

In this work, we presented a lightweight multimodal ensemble framework designed to forecast next-day gold prices using historical market data. By integrating three complementary modeling paradigms—a Bi-GRU model to capture temporal dependencies, a Transformer model for sequence attention insights, and a Temporal Convolutional Network (TCN) for fast, convolutional pattern extraction—we demonstrated the effectiveness of leveraging diverse predictive strengths.

Our ridge-regression ensemble achieved a scaled RMSE of 0.062 on validation data, representing an 8% improvement over the best individual model (Bi-GRU, RMSE = 0.067). Although directional accuracy remained moderate at approximately 49–53%, each model contributed uniquely to the ensemble's performance: the Bi-GRU excelled at capturing the magnitude of price movements, the Transformer model provided valuable directional signals despite a higher RMSE (0.148), and the TCN model, while less accurate, contributed complementary insights from a convolutional perspective.

A comparative analysis against academic benchmarks and state-of-the-art models illustrated that our approach delivered competitive forecasting accuracy. Our ensemble demonstrated approximately a 30% reduction in squared error compared to the trivial "yesterday" baseline and performed within 20% of modern SOTA methodologies utilizing extensive macroeconomic data. Nevertheless, our directional accuracy suggests room for improvement relative to quant-driven models, which leverage high-frequency trading strategies and additional contextual information.

Future enhancements will focus on improving directional forecasting by incorporating macroeconomic indicators, sentiment analysis derived from financial news, and exploring advanced architectures such as the Temporal Fusion Transformer. Additionally, refining our ensemble weighting scheme and experimenting with hybrid loss functions (combining RMSE and directional objectives) hold promise for further elevating model performance. By effectively balancing model complexity with computational efficiency, our multimodal ensemble framework establishes a robust baseline

Table 4: Benchmark comparison of forecasting methods

Benchmark Group	Benchmark Result (metric)	Our Result & Comment
Naïve baseline ("tomorrow = today")	RMSE 0.09–0.10 (scaled)	RMSE **0.067** ⇒ about 30 % lower squared error than the trivial baseline.
Academic gold–price papers	RMSE 110–200 USD (≈ 0.07–0.12 scaled)	RMSE ≈ 150 USD (0.067 scaled) — mid–range of published GRU/LSTM studies that use market features only.
Modern SOTA research (TFT, Informer + macro/news)	58 % DA, 0.050 RMSE (scaled)	53 % DA, 0.062 RMSE — within ∼20 % of SOTA RMSE but about 5 pp lower directional accuracy.
Quant–fund intraday models	Hit–rate 50.5–54 % on high–frequency returns	53 % DA (daily horizon) — comparable directional edge; funds monetise the signal via scale and leverage.
Commodity banks / research desks	1–month MAPE ≈ 2–3 %; 65–100 USD error	MAPE ∼ 4.5 % (∼ 150 USD) — ML–only forecast still trails discretionary macro judgement.

for practical gold-price forecasting in trading and risk management applications.

5.2 Team Member's Contribution

- **Vikyath Naradasi** — Designed and implemented the data pipeline and feature engineering; developed and tuned the Bi-GRU model; integrated all components into the final ensemble Google Collab notebook.
- **Abhinav Vadhera** — Implemented and fine-tuned the Hugging Face TimeSeriesTransformer-Tiny model; managed experiment logging; produced attention-map visualizations; drafted the Transformer methodology.
- **Rodrigo Lopez** — Developed the Temporal Convolutional Network (TCN) model; authored the ensemble fusion script (mean, RMSE-weighted, ridge); compiled final performance analysis and conclusion.

References

[1] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation.* arXiv:1406.1078.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need.* In *Advances in Neural Information Processing Systems.* 5998–6008.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.* arXiv:1803.01271.

[4] Bryan Lim and Stefan Zohren. 2021. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.

[5] Haoyi Wu *et al.* 2021. *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting.* In *AAAI.* 11106–11115.

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

[7] Ritwik B. 2023. *Daily Gold Price 1996–2023 (Time Series).* Dataset, Kaggle. https://www.kaggle.com/datasets/ritwikb3/daily-gold-price-1996-2023-time-series

[8] Baskar Bala. 2022. *US Dollar Index (DXY) Historical Data.* Dataset, Kaggle. https://www.kaggle.com/datasets/balabaskar/us-dollar-index-data

[9] Muhammad Tarique. 2023. *WTI Crude Oil Futures Daily Prices.* Dataset, Kaggle. https://www.kaggle.com/datasets/tarique7/daily-crude-price-dataset

[10] U.S. Bureau of Labor Statistics. 2024. Consumer Price Index for All Urban Consumers (CPIAUCSL). FRED Economic Data. https://fred.stlouisfed.org/series/CPIAUCSL

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR.*

[12] Adam Paszke *et al.* 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS.* 8024–8035.