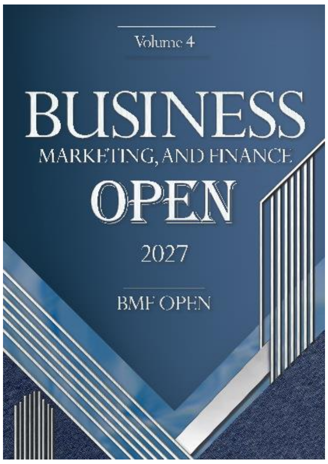
Design and Evaluation of an Advanced Attention-Based Deep Learning Framework for Accurate, Robust, and Interpretable Forecasting of Global Gold Prices with a Focus on the Iranian Market



Morteza Mohammadlu [1,*], Seyed Mohsen Mirhosseini [2]

[1] Department of Computer Engineering, Hi.C., Islamic Azad University, Hidaj, Iran;
[2] Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran;

* Correspondence: morteza.mohammadlu@iau.ir

Citation: Mohammadlu, M., & Mirhosseini, S. M. (2027). Design and Evaluation of an Advanced Attention-Based Deep Learning Framework for Accurate, Robust, and Interpretable Forecasting of Global Gold Prices with a Focus on the Iranian Market. *Business, Marketing, and Finance Open*, 4(8), 1-21.

Received: 16 June 2026
Revised: 04 September 2026
Accepted: 13 September 2026
Initial Publication: 14 September 2026
Final Publication: 01 August 2027

Abstract: This study aimed to design and evaluate an advanced Temporal Fusion Transformer-based deep learning framework for accurate, robust, probabilistic, multi-horizon, and interpretable forecasting of global gold prices with particular relevance to the Iranian market. A quantitative time-series forecasting design was applied to daily gold-market data covering January 19, 2014, to January 22, 2024. The dataset included Open, High, Low, Close, and Volume observations and was divided chronologically into training, validation, and test subsets. The proposed Temporal Fusion Transformer integrated recurrent temporal processing, variable selection networks, gated residual networks, bidirectional LSTM encoding, interpretable multi-head attention, and quantile regression. Forecasts were generated across multiple horizons and evaluated using RMSE, MAE, MAPE, R^2, Prediction Interval Coverage Probability, Prediction Interval Normalized Range Width, and Mean Directional Accuracy. Comparative benchmarks included ARIMA, SARIMA, LSTM, GRU, standard Transformer, and Informer. Statistical differences in predictive accuracy were assessed using the Diebold-Mariano test. The proposed TFT significantly outperformed ARIMA ($DM = -4.82, p < 0.0001$), SARIMA ($DM = -4.53, p < 0.0001$), LSTM ($DM = -3.21, p = 0.0013$), GRU ($DM = -2.98, p = 0.0029$), and the standard Transformer ($DM = -2.34, p = 0.0193$). Its superiority over Informer was not statistically significant at the 95% confidence level ($DM = -1.87, p = 0.0614$), although significance was achieved at the 90% level. The model also preserved its comparative advantage as the forecasting horizon increased, indicating greater robustness in longer-horizon prediction. The proposed TFT framework provides a statistically supported and practically useful approach for global gold-price forecasting by combining strong predictive performance, robustness across horizons, calibrated uncertainty estimation, and intrinsic interpretability, thereby offering potential value for investment analysis and economic decision-making in the Iranian market.

Keywords: Gold Price Forecasting; Temporal Fusion Transformer; Deep Learning; Attention Mechanism; Multi-Horizon Forecasting; Interpretability; Probabilistic Forecasting; Iranian Market

1. Introduction

Gold has long occupied a distinctive position in the international financial system because it simultaneously functions as a commodity, a reserve asset, a portfolio-diversification instrument, and a perceived safe haven during periods of economic, financial, and geopolitical uncertainty. Unlike many conventional financial assets, the price of gold is influenced by a heterogeneous set of macroeconomic, monetary, financial, behavioral, and geopolitical forces that interact dynamically over time. Movements in the U.S. dollar, real interest rates, inflation expectations, crude-oil prices, equity-market performance, market volatility, monetary-policy decisions, and changes in global risk perception may all affect the demand for and valuation of gold. These relationships are rarely stable, linear, or temporally homogeneous. Instead, the relative importance and direction of individual determinants may change considerably between periods of economic expansion, monetary tightening, recession, financial crisis, pandemic-related disruption, or geopolitical conflict. Consequently, forecasting the global price of gold represents a challenging time-series problem characterized by nonlinearity, nonstationarity, regime changes, volatility clustering, long-range temporal dependence, and interactions among multiple explanatory variables [1-3].

The forecasting of gold prices is particularly important because even relatively small improvements in predictive accuracy can have meaningful implications for investors, portfolio managers, central banks, commodity traders, and economic policymakers. For investors, reliable forecasts may improve asset-allocation decisions, hedging strategies, entry and exit timing, and risk control. For policymakers and monetary authorities, gold-price movements may provide information regarding inflationary expectations, financial uncertainty, exchange-rate pressure, and changes in the perceived credibility of monetary policy. These considerations are especially relevant for economies in which gold has an important role as a store of value and as a hedge against currency depreciation. In the Iranian market, the international gold price is one of the principal external components affecting the domestic valuation of gold and related assets. Domestic gold prices are additionally exposed to exchange-rate dynamics, inflation, policy uncertainty, and market expectations, meaning that accurate forecasting of the global component can provide valuable information for interpreting local market movements and supporting more informed financial decision-making.

Traditional forecasting methods have contributed substantially to the analysis of financial time series, yet their usefulness is restricted when the data-generating process contains strong nonlinearities, interactions, or structural changes. Autoregressive Integrated Moving Average models, seasonal extensions, volatility-oriented specifications, and related econometric approaches can effectively describe specific linear and short-memory properties, but they generally depend on relatively restrictive assumptions regarding functional form and temporal relationships. Gold markets, however, frequently exhibit abrupt changes in volatility and sensitivity to external variables. A forecasting model that performs adequately under relatively stable conditions may therefore deteriorate rapidly when the market enters a new regime. Hybrid approaches have sought to compensate for this limitation by combining statistical models with machine-learning procedures. For example, Quang proposed an integrated SARIMA–LSTM framework supported by Random Forest techniques for daily global gold-price analysis, illustrating the potential value of combining linear time-series structures with nonlinear learning algorithms [2]. Such hybridization represents an important advance, but the growing complexity of financial data has created demand for models capable of learning temporal and cross-variable structures more directly.

Machine-learning techniques such as Random Forest, gradient boosting, and XGBoost have increasingly been employed in gold forecasting because they can capture nonlinear associations without requiring a prespecified functional relationship between predictors and the target series. Guo and colleagues developed a two-layer decomposition and XGBoost framework optimized through the Whale Optimization Algorithm, demonstrating that decomposition, feature extraction, and optimized nonlinear learners can substantially improve gold-price forecasting performance [1]. Although such methods can be highly effective when informative lagged and engineered features are available, conventional machine-learning models do not inherently represent the sequential structure of time-series observations. Temporal relationships must often be incorporated through manual lag construction or external decomposition procedures. This limitation becomes more important when the forecasting task involves long historical sequences in which the effect of past observations varies over time and interactions among explanatory variables develop across multiple temporal scales.

Deep neural networks have therefore become increasingly prominent in financial forecasting. Convolutional neural networks can identify local structures and recurring patterns, while recurrent architectures can model sequential dependencies. Amini and Kalantari proposed a CNN–Bi-LSTM model with automatic parameter tuning for gold-price prediction, showing that convolutional feature extraction combined with bidirectional recurrent processing can offer a powerful framework for learning nonlinear price dynamics [4]. More recently, Lin investigated an improved CNN–LSTM architecture for daily gold-price forecasting, further reflecting the continuing interest in combining convolutional pattern extraction with recurrent temporal modeling [5]. These approaches address several limitations of conventional econometric methods by allowing the model to discover latent nonlinear representations directly from historical observations rather than depending exclusively on manually specified relationships.

Long Short-Term Memory networks and Gated Recurrent Units are especially attractive for financial forecasting because their gating mechanisms improve the ability of recurrent networks to preserve information across longer sequences. However, they are not free from limitations. Although LSTM mitigates the vanishing-gradient problem associated with simple recurrent neural networks, the effective representation of very long sequences remains difficult, particularly when only a small subset of historical periods contains information that is highly relevant to the current forecast. Standard recurrent processing also lacks an explicit mechanism for determining which historical observations should receive greater or lesser emphasis at a given prediction origin. Thus, information is propagated through sequential hidden states even when its forecasting relevance varies substantially over time. These constraints have motivated the incorporation of attention mechanisms into recurrent architectures. Wu and Liu developed an LSTM–Attention combined model for gold-price prediction, emphasizing the value of allowing the network to assign differential importance to temporal information rather than processing all historical states uniformly [6]. Similarly, Liu and Lan proposed an attention-enhanced CNN–BiGRU approach for long-term gold-price forecasting, illustrating how attention can strengthen the representation of long-range temporal patterns when integrated with convolutional and bidirectional recurrent structures [7].

Attention-based recurrent models are also being extended through multi-head structures and residual or skip connections. Memon and colleagues proposed an enhanced GRU model with multi-headed attention and skip connections for forecasting gold and platinum prices, demonstrating that multiple attention heads can help capture heterogeneous temporal relationships while skip connections improve information flow through deeper architectures [8]. These developments indicate that the central forecasting problem is no longer simply whether deep learning should be used, but how temporal representation, attention, feature selection, nonlinear transformation, and uncertainty modeling should be integrated within a coherent architecture. Gold-price behavior may contain short-term momentum, medium-term cyclical structures, long-term macroeconomic dependencies, and abrupt shock effects simultaneously. A single recurrent state or single attention pattern may therefore be insufficient to represent the full temporal complexity of the market.

Transformer-based architectures have introduced a major methodological shift by replacing exclusively sequential information processing with self-attention mechanisms capable of learning relationships between distant points in a sequence. This makes them particularly suitable for long-range time-series forecasting, where observations separated by substantial temporal intervals may still be strongly related. Yang proposed the TCN-QV attention-based deep-learning method for long-sequence forecasting of gold prices, reflecting a broader transition toward architectures that explicitly address long temporal dependencies and selective information weighting [9]. Zhao, Guo, and Wang further developed a hybrid LSTM–Transformer architecture with multi-scale feature fusion for high-accuracy gold futures forecasting, combining the local sequential representation of recurrent networks with the global dependency modeling of Transformer mechanisms [10]. These hybrid approaches are theoretically well aligned with gold markets because price formation is affected by both recent market movements and slower macro-financial processes that unfold across longer periods.

Nevertheless, predictive accuracy alone is insufficient for many financial applications. A highly accurate model that operates as an opaque black box may have limited value for decision-makers who need to understand the basis of a forecast, assess whether its reasoning is economically plausible, and evaluate how much confidence should be placed in the predicted outcome. Explainability has therefore become an increasingly important issue in financial and macroeconomic deep learning. Izzo emphasized the importance of explainable deep-learning methods for macroeconomic forecasting and finance, where transparency concerning model drivers is necessary for both scientific understanding and practical adoption [11]. This issue is particularly important in high-stakes environments such as investment management and economic policymaking, where a prediction may influence substantial financial decisions. A model capable of identifying which variables and historical periods contributed most strongly to a forecast can provide substantially greater decision value than one that supplies only a predicted price.

Another fundamental weakness of many forecasting models is their emphasis on point forecasts rather than predictive distributions. A point forecast gives a single expected value but does not indicate the uncertainty surrounding that value. In the gold market, this limitation is critical because forecast uncertainty may expand dramatically during crises, policy shifts, geopolitical events, and periods of heightened financial volatility. Wang and Lin addressed this issue through a deterministic–probabilistic forecasting framework based on quantile regression deep learning and multi-objective optimization, demonstrating the importance of producing probabilistic information rather than relying solely on central forecasts [3]. Quantile-based forecasting can generate lower, median, and upper estimates of future gold prices, providing prediction intervals that are directly relevant to risk management. Such intervals allow users to distinguish between a forecast made under relatively stable and predictable conditions and one associated with substantial uncertainty.

The need to integrate accuracy, probabilistic forecasting, and interpretability has increased interest in specialized time-series Transformer architectures. Temporal Fusion Transformer represents one of the most important developments in this direction. Lim and colleagues introduced TFT as an interpretable multi-horizon forecasting architecture that integrates recurrent processing, attention, static covariate encoders, variable selection networks, and gated residual components within a unified framework [12]. Unlike generic Transformers, TFT was designed specifically for time-series forecasting in settings containing multiple categories of inputs, including static characteristics, known future covariates, and observed time-varying variables. Its variable selection networks can assign dynamic relevance weights to predictors, while its gating mechanisms allow unnecessary nonlinear processing components to be suppressed. Furthermore, interpretable multi-head attention enables the contribution of historical periods to be examined directly, creating a mechanism for temporal explanation in addition to prediction [12].

These characteristics make TFT particularly appropriate for the gold market. The importance of variables such as the U.S. Dollar Index, real interest rates, crude-oil prices, inflation measures, stock-market performance, precious-metal prices, and volatility indices is unlikely to remain constant across all market regimes. Variable selection should therefore be dynamic rather than fixed. During a period dominated by monetary tightening, interest-rate and dollar-related variables may become relatively influential; during a financial panic, risk indicators may assume greater importance; during commodity or inflation shocks, oil and price-level indicators may become more informative. A model capable of estimating such changing relevance patterns can potentially provide both stronger predictive performance and economically meaningful explanations. Recent research has begun to investigate TFT directly in gold forecasting. Egy compared TFT and N-BEATS in gold-price prediction with incorporated macroeconomic features, indicating increasing interest in specialized interpretable multi-horizon architectures for this market [13]. However, the broader literature still contains relatively limited evidence regarding the combined evaluation of TFT in terms of forecast accuracy, interval calibration, multi-horizon stability, directional performance, and interpretability.

The literature also demonstrates that no single existing methodological direction fully resolves all major forecasting challenges. CNN-based models are effective in extracting local patterns, but they do not intrinsically provide long-range temporal explanations [4, 5]. LSTM and GRU models improve sequential learning but may still struggle when the most relevant information is located far back in a long sequence or when predictor importance changes strongly over time [6, 7]. Attention-enhanced recurrent networks improve selective focus but may not provide the comprehensive multi-horizon input structure and variable-selection mechanisms available in architectures explicitly designed for interpretable forecasting [8]. Transformer-based and hybrid LSTM–Transformer models improve the representation of long-range dependencies but may remain computationally intensive and may not automatically generate easily interpretable variable-level importance measures [9, 10]. Meanwhile, probabilistic deep-learning approaches address uncertainty but may not simultaneously provide dynamic feature attribution and interpretable temporal attention [3]. These differences suggest the need for a unified framework capable of addressing nonlinear behavior, long temporal memory, selective attention, prediction uncertainty, and model transparency concurrently.

A further issue concerns robustness across forecasting horizons and market regimes. Many forecasting studies emphasize a single prediction horizon or report average error values without sufficiently examining how performance deteriorates as the forecast extends further into the future. Yet a model that is highly accurate one day ahead may be considerably less useful over one- or several-week horizons. This issue is particularly relevant for institutional and strategic decision-making, where medium- and longer-term forecasts may be more important than next-day predictions. The multi-horizon design of TFT provides a methodological advantage because different future time points can be forecast jointly within the same architecture rather than through completely independent models. Multi-horizon learning can also help the model recognize temporal relationships that remain informative across different prediction windows, while attention mechanisms permit the historical evidence supporting each horizon to be investigated [12]. The growing use of multi-scale and long-sequence architectures in recent gold-price studies confirms the relevance of this problem [9, 10].

The broader deep-learning literature on financial volatility also supports the use of multivariate architectures rather than models based exclusively on the historical target series. Ge and colleagues compared deep-learning methods for volatility prediction using multivariate data and highlighted the importance of incorporating multiple sources of information when modeling complex financial dynamics [14]. This consideration is highly relevant to gold, whose price dynamics emerge from interactions among currency, interest-rate, equity, commodity, inflation, and uncertainty channels. A forecasting architecture should therefore not only process a long sequence but also determine which variables are informative under different conditions. Static feature importance derived from a single overall sample may conceal substantial time-varying heterogeneity. Dynamic variable selection, as incorporated into TFT, provides a mechanism for identifying whether the informational contribution of specific financial variables changes across the forecasting period.

Despite rapid methodological progress, a significant research gap therefore remains. Existing gold-price forecasting studies have demonstrated the value of recurrent networks, CNN–RNN combinations, attention mechanisms, gradient-boosting approaches, hybrid statistical–deep-learning models, long-sequence architectures, probabilistic forecasting, and Transformer-based frameworks [1, 2, 4, 6, 9]. However, fewer studies have simultaneously evaluated whether a forecasting model can achieve low point-prediction error, maintain performance as the prediction horizon expands, correctly identify the direction of price movements, produce calibrated and sufficiently narrow prediction intervals, and provide interpretable information regarding both influential predictors and important historical periods. Furthermore, practical evaluation requires comparison not only with classical time-series models but also with strong recurrent and Transformer-based benchmarks because superiority over a weak benchmark cannot establish the incremental value of a complex deep-learning architecture.

This research gap has direct relevance to the Iranian financial environment. Because global gold-price dynamics interact with domestic currency and inflation conditions, inaccurate international price forecasts may amplify uncertainty in investment and hedging decisions. Conversely, a robust and interpretable global forecasting framework can provide an external reference signal for analysts studying Iranian gold-market dynamics. Interpretability is particularly important in this context because model outputs may be more useful when decision-makers can determine whether forecasts are primarily associated with recent gold-price behavior, broader financial-market instability, dollar movements, commodity-price conditions, or other macroeconomic signals. At the same time, calibrated prediction intervals can communicate when the model itself recognizes elevated uncertainty, preventing point estimates from being interpreted with unjustified precision.

Accordingly, the methodological progression from conventional time-series analysis to machine learning, recurrent networks, attention mechanisms, and specialized Transformer architectures suggests that future improvements in gold forecasting are likely to depend less on increasing model complexity alone and more on integrating complementary capabilities within a coherent system. The evidence from CNN–Bi-LSTM and improved CNN–LSTM approaches supports sophisticated nonlinear feature extraction [4, 5]; attention-enhanced LSTM, BiGRU, and GRU frameworks demonstrate the importance of selective temporal weighting [6-8]; long-sequence and hybrid Transformer architectures highlight the value of capturing multi-scale temporal relationships [9, 10]; probabilistic deep learning demonstrates the necessity of explicit uncertainty estimation [3]; and explainable financial forecasting establishes the need for transparent model behavior [11]. TFT provides a theoretically coherent framework in which these requirements can be brought together through recurrent temporal processing, variable selection, gating, interpretable multi-head attention, multi-horizon prediction, and quantile-based probabilistic outputs [12, 13].

The aim of this study was to design and evaluate an advanced attention-based Temporal Fusion Transformer deep-learning framework for accurate, robust, probabilistic, multi-horizon, and interpretable forecasting of global gold prices, with particular emphasis on its applicability to the Iranian market.

2. Methods and Materials

This study employed a quantitative, computational, time-series forecasting design to develop and evaluate an advanced deep learning framework for predicting global gold prices with particular emphasis on its applicability to investment analysis and economic decision-making in the Iranian market. The methodological framework was based on the Temporal Fusion Transformer (TFT), an attention-based architecture specifically designed for multi-horizon forecasting of complex multivariate time series. The principal rationale for selecting TFT was the nonlinear, nonstationary, regime-dependent, and highly interconnected nature of the gold market. Global gold prices are influenced simultaneously by monetary conditions, exchange-rate movements, inflation expectations, commodity markets, equity-market performance, financial uncertainty, and other macro-financial forces whose direction and magnitude may vary substantially over time. Accordingly, the study was designed not merely to generate a single point estimate of future gold prices, but to develop a forecasting system capable of learning short- and long-term temporal dependencies, dynamically selecting influential predictors, quantifying forecast uncertainty, and providing interpretable information concerning the variables and historical periods that contributed most strongly to each prediction. The observation unit consisted of chronologically ordered time points in the multivariate financial dataset, with the global gold price serving as the principal target variable. The analytical sample included historical observations of the target series together with explanatory variables representing major dimensions of the international macro-financial environment. Because the ultimate purpose of the model was genuine out-of-sample forecasting rather than retrospective fitting, observations were maintained in their original temporal order throughout model development, and random shuffling across time was avoided. The complete dataset was divided chronologically into training, validation, and test subsets comprising 80%, 10%, and 10% of the available observations, respectively. The earliest observations were assigned to the training set, the subsequent observations to the validation set, and the most recent observations to the test set. This procedure minimized look-ahead bias and data leakage and reproduced the information structure of real-world forecasting, in which only past and currently observable information can be used to predict future prices. The validation subset was used for hyperparameter adjustment, model selection, early-stopping decisions, and control of overfitting, whereas the final test subset remained completely isolated during model development and was used exclusively to assess out-of-sample predictive performance. Forecasts were generated simultaneously at three practically meaningful horizons: a one-day horizon representing short-term forecasting, a ten-day horizon representing medium-term forecasting, and a thirty-day horizon representing longer-term forecasting. These horizons were selected to provide information relevant to different decision windows in financial markets and to determine whether predictive accuracy and uncertainty changed as the forecasting horizon increased. In addition to its global forecasting objective, interpretation of the results was oriented toward the Iranian market, where fluctuations in international gold prices constitute an important external component of domestic gold valuation, portfolio allocation, inflation-hedging decisions, and financial risk assessment.

The dataset was constructed as a multivariate time-series database combining the historical global price of gold with a set of financial and macroeconomic predictors theoretically and empirically associated with gold-price dynamics. Global gold price constituted the dependent or target variable. The explanatory variable set included the U.S. Dollar Index (DXY), measures of real interest rates, crude oil prices, the S&P 500 stock-market index, silver prices, the CBOE Volatility Index (VIX), and measures of inflation represented by the Consumer Price Index (CPI). These variables were selected because they represent distinct transmission channels through which monetary policy, currency valuation, commodity-market conditions, equity-market movements, inflationary pressures, and systemic financial uncertainty can affect demand for gold and consequently its market price. The U.S. Dollar Index was incorporated to capture the well-established currency-pricing channel of internationally traded gold; real interest rates represented the opportunity cost of holding a non-interest-bearing asset; crude oil prices reflected commodity-market and inflation-related pressures; the S&P 500 represented conditions in major equity markets and possible portfolio substitution effects; silver prices captured information from a closely related precious-metal market; VIX represented market uncertainty and risk aversion; and CPI-related information represented inflationary conditions. Calendar-related features were additionally generated from the date index, including day-of-week, month-of-year, seasonal, and other deterministic temporal indicators whose future values are known in advance. In accordance with the TFT framework, the input information was organized into static covariates, known future time-varying covariates, and observed time-varying covariates. Static information described characteristics associated with the modeled series and its structural context; known time-varying inputs consisted primarily of deterministic calendar features; and observed time-varying inputs included the historical gold price and financial and macroeconomic variables whose future values were not assumed to be known at the prediction origin. This distinction enabled the model to use different information channels appropriately without introducing future information into the forecasting process.

Before model estimation, the raw time-series database underwent a systematic preprocessing procedure intended to improve data quality while preventing information leakage. The time indices of all variables were first aligned so that the multivariate observations corresponded to the same temporal frequency and forecasting dates. Missing observations and isolated data gaps were treated by linear interpolation when interpolation was methodologically appropriate, while observations identified as anomalous or inconsistent with the surrounding series were examined and replaced through the preprocessing procedure rather than being allowed to generate artificial discontinuities in model training. Continuous variables were transformed using Min-Max normalization to the interval from 0 to 1. Scaling parameters were estimated from the training portion of the data and subsequently applied to the validation and test sets to ensure that information from future observations did not influence model preprocessing. Categorical calendar variables were represented through trainable embedding layers rather than arbitrary numerical coding. Continuous variables, including financial prices and indices, were mapped through fully connected transformations, whereas categorical variables were processed through dedicated embeddings. Both types of inputs were projected into a common latent representation with a model dimension of $d_{model} = 64$. The processed observations were subsequently transformed into sequential encoder-decoder samples in which a historical input window was used to produce forecasts for the specified one-day, ten-day, and thirty-day horizons. This sequence-generation procedure preserved temporal continuity and enabled the network to learn recurring seasonal patterns, short-run fluctuations, long-term relationships, and responses to episodes of heightened market instability.

The proposed forecasting instrument was the Temporal Fusion Transformer architecture. After input embedding, separate variable selection networks were applied to the relevant groups of covariates. These networks assigned normalized Softmax weights to the candidate inputs and produced weighted combinations of their latent representations. Consequently, the model was not required to treat all explanatory variables as equally informative across the entire observation period. Rather, the contribution of each predictor could change dynamically from one temporal context to another. For example, the importance of the dollar index, interest rates, VIX, or oil prices could increase during particular monetary, inflationary, or crisis regimes while declining during relatively stable periods.

Gated Residual Networks were incorporated throughout the architecture to regulate nonlinear processing. These modules combined nonlinear transformations, residual connections, layer normalization, and Gated Linear Units, allowing the network to suppress unnecessary computational pathways when complex transformations did not improve the representation. This mechanism was particularly important for a financial dataset containing variables whose informational relevance may vary substantially across periods and helped reduce unnecessary model complexity and the risk of overfitting.

A static covariate encoder was incorporated to generate contextual representations subsequently used by other components of the TFT. The static encoder generated contextual vectors for temporal variable selection, decoder enrichment, and initialization of the recurrent states. Temporal covariates were then processed through the recurrent component of the model. In the proposed framework, a bidirectional LSTM representation was used within the historical encoder window to strengthen extraction of temporal relationships from the observed historical sequence. Bidirectional processing was restricted to the information available within the historical input window so that observations belonging to the actual forecasting horizon were not used during prediction. This component was designed to capture local temporal dynamics and sequential dependencies before the representations entered the attention mechanism. The resulting temporal features were transferred to the Temporal Fusion Decoder, whose principal component was an interpretable multi-head attention mechanism. Four parallel attention heads were employed, allowing the model to represent different forms of temporal dependency simultaneously, such as recent price momentum, recurring seasonal behavior, longer-term trends, and concentrated reactions to unusually volatile periods. The attention mechanism generated normalized weights over historical observations, thereby enabling the contribution of different historical time steps to be inspected after model estimation. This feature transformed temporal attention from an exclusively predictive mechanism into an analytical tool for identifying the historical intervals most relevant to future gold-price formation.

The output layer was designed for probabilistic rather than exclusively deterministic forecasting. Three conditional quantiles, corresponding to the 0.10, 0.50, and 0.90 quantiles of the predicted gold-price distribution, were estimated simultaneously. The median or 0.50 quantile served as the central forecast, while the interval between the 0.10 and 0.90 quantiles constituted an 80% prediction interval. This specification was selected because uncertainty assessment is particularly important in the gold market, where forecast errors may increase substantially during monetary shocks, geopolitical disruptions, sudden changes in investor risk aversion, and other structural disturbances. The network was therefore trained using quantile, or pinball, loss rather than a conventional mean-squared-error objective alone. Quantile loss penalizes underprediction and overprediction asymmetrically depending on the target quantile and allows the network to estimate different regions of the conditional outcome distribution. The resulting framework consequently produced three complementary forms of information for each forecasting horizon: the expected central trajectory of the gold price, the estimated uncertainty surrounding that trajectory, and interpretable evidence concerning the temporal observations and explanatory variables that contributed to the prediction.

Data analysis was conducted as a sequential model-development, validation, forecasting, benchmarking, and interpretation procedure. Following preprocessing and chronological partitioning of the dataset, the TFT model was fitted to the training observations using mini-batch gradient-based optimization. Quantile loss aggregated across the 0.10, 0.50, and 0.90 target quantiles constituted the primary optimization objective. Model performance on the validation period was monitored during training, and validation loss rather than training loss was used as the principal criterion for selecting the best-performing model state. Early stopping was applied when additional training failed to produce meaningful improvement in validation performance, thereby reducing the probability of overfitting to historical observations. Model complexity was additionally controlled by the gated architecture, variable selection mechanism, normalization procedures, and validation-based model selection. Hyperparameters affecting the recurrent representation, hidden dimensions, dropout behavior, attention architecture, batch processing, and learning process were determined exclusively with reference to the training and validation portions of the data. The test data were not used for architecture selection or parameter adjustment. After finalization of the architecture, predictions were generated for the previously unseen test period at one-day, ten-day, and thirty-day horizons. Predictions were then transformed back to the original gold-price scale to ensure that all reported error statistics and substantive interpretations remained economically meaningful.

Predictive accuracy was evaluated through complementary error criteria rather than a single statistic because different metrics capture different aspects of forecasting performance. Root Mean Square Error was used to quantify overall prediction error while assigning greater penalties to relatively large deviations, Mean Absolute Error was calculated to provide a scale-dependent but outlier-resistant measure of the average absolute forecasting deviation, and Mean Absolute Percentage Error was used to express forecasting error relative to the observed gold-price level. Where appropriate, the coefficient of determination was additionally calculated to quantify the proportion of out-of-sample variation reproduced by the forecasting model. Because the proposed TFT generated probabilistic forecasts, evaluation was not limited to the median prediction. Quantile or pinball loss was calculated for the 0.10, 0.50, and 0.90 forecasts, and the quality of the 80% prediction interval was assessed by considering both empirical interval coverage and interval width. A satisfactory probabilistic forecasting model was expected to achieve coverage reasonably close to the nominal 80% level while avoiding unnecessarily wide intervals. This dual criterion prevented the model from obtaining apparently high coverage simply by generating excessively broad uncertainty bands. Forecast performance was examined separately at the one-day, ten-day, and thirty-day horizons because deterioration in accuracy and widening of uncertainty intervals were expected as the prediction horizon increased.

The added predictive value of the proposed architecture was assessed against conceptually distinct benchmark approaches representing classical statistical forecasting, conventional machine learning, and recurrent deep learning. The comparative framework included ARIMA-type models for linear temporal dynamics, GARCH-type specifications for volatility-sensitive modeling, Random Forest and XGBoost as representative nonlinear machine-learning methods, and recurrent architectures such as LSTM and GRU as deep-learning benchmarks. These comparison models were trained and evaluated using the same chronological data partitions and forecasting horizons wherever their methodological characteristics permitted direct comparison. The purpose of benchmarking was not simply to determine whether TFT produced a smaller average error, but to assess whether combining recurrent temporal processing, dynamic variable selection, gated nonlinear transformations, and interpretable multi-head attention generated more stable performance across different horizons and changing market conditions than models relying on only one of these mechanisms. Particular attention was directed toward performance during periods of elevated volatility or structural instability because robustness to regime changes constituted one of the principal motivations for developing the proposed framework.

Interpretability analysis formed an integral component of the analytical procedure rather than a supplementary post hoc exercise. The normalized weights generated by the TFT variable selection networks were extracted to estimate the relative importance of the explanatory variables. Importance was examined both globally, by aggregating weights across the test period, and locally, by investigating how predictor relevance changed at particular prediction origins. This procedure made it possible to determine whether variables such as the U.S.

Dollar Index, real interest rates, crude oil prices, the S&P 500, silver prices, VIX, and inflation indicators exhibited stable or time-varying contributions to gold-price forecasting. The interpretable multi-head attention weights were likewise extracted and analyzed to identify historical time points receiving relatively high attention during prediction. Attention patterns were investigated for evidence of recency effects, longer-term memory, recurring seasonal structures, and concentrated attention around periods characterized by unusual market movements. These outputs were used to address the explanatory question of why a particular forecast was generated and whether changes in predicted gold prices could be associated with changes in the relative informational contribution of specific variables or historical periods.

Robustness was further evaluated by comparing forecasting behavior across the three temporal horizons and across relatively stable and highly volatile segments of the test period. Stability was interpreted as the capacity of the model to maintain comparatively low forecast errors, sensible uncertainty intervals, and coherent variable-importance patterns when the statistical characteristics of the market changed. The probabilistic outputs were particularly important in this stage because an increase in the width of the 80% prediction interval during turbulent periods was interpreted as an appropriate expression of model uncertainty rather than necessarily as a failure of the forecasting framework. Conversely, excessively narrow intervals accompanied by poor empirical coverage were treated as evidence of overconfidence. The final evaluation of the proposed framework therefore integrated point-forecast accuracy, multi-horizon performance, probabilistic calibration, robustness to changing market conditions, and model interpretability. This integrated analytical strategy was intended to determine whether the TFT architecture could provide not merely more accurate global gold-price forecasts, but a stable, uncertainty-aware, and interpretable decision-support framework with practical relevance for investors, analysts, and economic policymakers concerned with the Iranian gold market.

3. Findings and Results

The proposed framework was implemented and evaluated in Python 3.9. Python was selected as the simulation and development environment because of its extensive ecosystem for machine learning, deep learning, time-series analysis, numerical computation, and data preprocessing, as well as its support for efficient implementation of recurrent neural networks, attention mechanisms, and Transformer-based architectures. The experiments were conducted on a real-world daily gold-price dataset covering the ten-year period from January 19, 2014, to January 22, 2024. The data were obtained from Nasdaq and included daily Open, High, Low, Close, and Volume information for the gold-price series. The selected period encompassed heterogeneous market regimes, including periods of economic uncertainty, changes in interest-rate conditions, episodes of elevated financial-market volatility, the COVID-19 shock, and major geopolitical disturbances. Consequently, the dataset provided an appropriate empirical environment for evaluating not only point-forecast accuracy but also the robustness and stability of the proposed Temporal Fusion Transformer under changing market conditions.

The empirical evaluation was based on complementary measures of point prediction accuracy, goodness of fit, directional accuracy, and probabilistic forecast calibration. Root Mean Square Error (RMSE) was calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where lower values indicate greater forecasting accuracy. Because the errors are squared before averaging, RMSE assigns greater penalties to relatively large forecasting deviations and is therefore particularly informative in financial forecasting, where large prediction errors may have substantial economic consequences.

Mean Absolute Error (MAE) was calculated using:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE represents the average absolute deviation between observed and predicted gold prices and is less sensitive than RMSE to isolated large errors. Since both RMSE and MAE are expressed in the same unit as the target variable, their results can be directly interpreted in U.S. dollars per ounce.

Mean Absolute Percentage Error (MAPE) was computed as:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

MAPE expresses prediction error relative to the actual gold price and therefore facilitates scale-independent interpretation of forecasting accuracy. In the present analysis, values below 5% were considered indicative of very high predictive accuracy, values between 5% and 10% represented good accuracy, whereas values exceeding 20% indicated relatively weak forecasting performance.

The coefficient of determination was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where values approaching 1 indicate that a larger proportion of observed gold-price variability is reproduced by the forecasting model. Values above approximately 0.80 were interpreted as evidence of strong explanatory and predictive correspondence between observed and model-estimated prices.

Because the proposed TFT generated probabilistic forecasts, Prediction Interval Coverage Probability (PICP) was also calculated:

$$PICP = \frac{1}{n}\sum_{i=1}^{n} 1\,[L_i \le y_i \le U_i] \times 100\%$$

where $L_i$and $U_i$denote the lower and upper bounds of the prediction interval, respectively. For the 0.10 and 0.90 quantile forecasts used in this study, the nominal prediction interval was 80%; therefore, a PICP value close to 80% indicated appropriate probabilistic calibration. Values considerably below 80% indicated excessively narrow and overconfident intervals, whereas substantially higher values could indicate unnecessarily wide intervals.

In these expressions, $y_i$denotes the actual observed value, $\hat{y}_i$is the model-predicted value, $\bar{y}$represents the mean of the actual observations, and $n$is the total number of evaluated observations. $L_i$and $U_i$represent the lower and upper prediction bounds, respectively, while $1[\cdot]$denotes the indicator function. In the quantile-based component of the TFT, $q$denotes the specified quantile, and a small numerical constant such as $\varepsilon = 10^{-8}$can be used when required to prevent numerical instability. In addition to the above measures, Prediction Interval Normalized Range Width (PINRW) was used to assess the relative width of the prediction intervals, while Mean Directional Accuracy

(MDA) was used in the comparative analyses to determine the proportion of occasions on which the model correctly predicted the direction of price movement.

Table 1 presents the principal performance indicators obtained for the proposed TFT model. The model achieved an RMSE of 18.42 USD/oz and an MAE of 13.87 USD/oz, indicating that absolute deviations between observed and predicted gold prices remained comparatively small. More importantly, the MAPE value was only 0.89%, demonstrating an average relative forecasting error of less than 1%. The R^2value of 0.9534 indicates that approximately 95.34% of the out-of-sample variation in gold prices was reproduced by the model. The probabilistic results were also favorable. PICP reached 79.80%, almost exactly corresponding to the nominal 80% prediction interval, while PINRW was only 0.0234, indicating that this coverage was obtained without producing excessively wide uncertainty bands.

Table 1. Principal Evaluation Results of the Proposed TFT Model

Evaluation Metric	Value	Unit	Desired Status	Interpretation
RMSE	18.42	USD/oz	Achieved	Very low root mean squared forecasting error
MAE	13.87	USD/oz	Achieved	Very low mean absolute forecasting error
MAPE	0.89	%	Achieved	Excellent relative accuracy; below 1%
$\boldsymbol{R^2}$	0.9534	—	Achieved	Approximately 95.34% of variation reproduced
PICP	79.80	%	Achieved	Very close to the nominal 80% prediction coverage
PINRW	0.0234	—	Achieved	Narrow normalized prediction intervals

The comparative results provided further evidence regarding the effectiveness of the proposed architecture. As shown in Table 2, TFT outperformed all six benchmark models across RMSE, MAE, MAPE, R^2, and directional accuracy. The classical ARIMA model generated an RMSE of 42.56 USD/oz, an MAE of 34.23 USD/oz, and an $R^2$of 0.7234, while SARIMA improved these values only moderately to an RMSE of 38.91 USD/oz and an $R^2$of 0.7512. These results indicated that conventional linear time-series structures were less capable of accommodating the nonlinear and time-varying properties of the gold-price process.

The recurrent deep-learning models produced considerably stronger forecasts than ARIMA and SARIMA. LSTM achieved an RMSE of 26.34 USD/oz, whereas GRU reduced this value to 24.78 USD/oz. Nevertheless, both models remained clearly inferior to the attention-based architectures. The standard Transformer achieved an RMSE of 21.56 USD/oz and MAPE of 1.08%, while Informer further improved RMSE to 20.13 USD/oz and MAPE to 0.98%. The proposed TFT nevertheless generated the lowest RMSE of 18.42 USD/oz, the lowest MAE of 13.87 USD/oz, the lowest MAPE of 0.89%, and the highest $R^2$of 0.9534. Its Mean Directional Accuracy was also the highest at 72.5%, compared with 69.1% for Informer, 67.3% for the standard Transformer, 62.8% for GRU, and 61.2% for LSTM. This superiority in directional forecasting is particularly important in practical market applications because correctly identifying the direction of future movement may be as important as minimizing the numerical magnitude of the price error.

Table 2. Comparative Forecasting Performance of the Proposed Model and Benchmark Methods

Forecasting Method	RMSE (USD/oz)	MAE (USD/oz)	MAPE (%)	$\boldsymbol{R^2}$	MDA (%)
ARIMA	42.56	34.23	2.15	0.7234	52.3
SARIMA	38.91	31.45	1.98	0.7512	54.7
LSTM	26.34	19.87	1.32	0.8745	61.2
GRU	24.78	18.92	1.25	0.8812	62.8
Standard Transformer	21.56	16.34	1.08	0.9123	67.3
Informer	20.13	15.67	0.98	0.9256	69.1
Proposed TFT	18.42	13.87	0.89	0.9534	72.5

The magnitude of improvement was substantial, particularly relative to the classical approaches. Compared with ARIMA, the proposed TFT reduced RMSE from 42.56 to 18.42 USD/oz, corresponding to an approximate reduction of 56.7%. Relative to SARIMA, the reduction was approximately 52.7%. Even against the stronger deep-learning competitors, the proposed model maintained a clear advantage. RMSE was approximately 30.1% lower than that of LSTM, 25.7% lower than that of GRU, 14.6% lower than that of the standard Transformer, and 8.5% lower than that of Informer. These results suggest that the advantage of TFT could not be attributed solely to the use of deep learning or attention. Rather, the combination of temporal recurrent processing, variable selection networks, gating mechanisms, and interpretable multi-head attention provided additional predictive value.

Performance was subsequently examined across different forecasting horizons to determine whether the relative superiority of TFT remained stable as forecasting difficulty increased. Table 3 reports RMSE values for one-day, five-day, ten-day, and twenty-two-day forecasting horizons, with twenty-two trading days representing approximately one trading month. For every model, forecasting error increased as the horizon became longer. However, the magnitude of deterioration differed considerably among the methods.

Table 3. RMSE Performance across Different Forecasting Horizons

Forecasting Method	$\boldsymbol{h} = \mathbf{1}$Day	$\boldsymbol{h} = \mathbf{5}$Days	$\boldsymbol{h} = \mathbf{10}$Days	$\boldsymbol{h} = \mathbf{22}$Days
ARIMA	42.56	58.23	72.15	95.67
SARIMA	38.91	53.67	67.34	89.23
LSTM	26.34	34.56	48.92	71.34
GRU	24.78	32.89	46.23	68.91
Standard Transformer	21.56	28.34	39.87	58.23
Informer	20.13	26.78	37.12	54.67
Proposed TFT	18.42	24.15	33.56	49.87

At the one-day horizon, the proposed model achieved an RMSE of 18.42 USD/oz compared with 20.13 for Informer and 21.56 for the standard Transformer. At five days, TFT retained the lowest error at 24.15 USD/oz, followed by Informer at 26.78 USD/oz and the standard Transformer at 28.34 USD/oz. At the ten-day horizon, the TFT RMSE increased to 33.56 USD/oz but remained substantially below the corresponding values of 37.12 for Informer, 39.87 for the standard Transformer, 46.23 for GRU, and 48.92 for LSTM. The advantage remained evident at the twenty-two-day horizon: TFT produced an RMSE of 49.87 USD/oz, compared with 54.67 for Informer, 58.23 for the standard Transformer, 68.91 for GRU, and 71.34 for LSTM. In contrast, the ARIMA and SARIMA errors increased to 95.67 and 89.23 USD/oz, respectively. Thus, although uncertainty naturally increased with forecast horizon, TFT demonstrated greater resistance to long-horizon performance degradation. This pattern is consistent with the architecture's ability to combine recurrent representations of local temporal structure with attention-based extraction of longer-range dependencies.

To determine whether the observed predictive improvements were statistically distinguishable from the benchmark methods, pairwise Diebold-Mariano tests were performed. Under the null hypothesis, the compared models were assumed to have equal predictive accuracy. A negative DM statistic in the present comparison indicated lower forecasting loss for TFT relative to the competing model. The results are presented in Table 4.

Table 4. Diebold-Mariano Tests of Forecast Accuracy

Model Comparison	DM Statistic	p-value	Result at 95% Confidence Level
TFT vs. ARIMA	-4.82	<0.0001	Significant
TFT vs. SARIMA	-4.53	<0.0001	Significant

TFT vs. LSTM	-3.21	0.0013	Significant
TFT vs. GRU	-2.98	0.0029	Significant
TFT vs. Standard Transformer	-2.34	0.0193	Significant
TFT vs. Informer	-1.87	0.0614	Not significant at 95%*

The Diebold-Mariano results confirmed that the superiority observed in the descriptive forecasting metrics was statistically supported for most benchmark comparisons. TFT differed significantly from ARIMA and SARIMA, with DM statistics of -4.82 and -4.53, respectively, and p-values below 0.0001. Significant improvements were also observed relative to LSTM ($DM = -3.21, p = 0.0013$), GRU ($DM = -2.98, p = 0.0029$), and the standard Transformer ($DM = -2.34, p = 0.0193$). Therefore, at the 95% confidence level, the hypothesis of equal predictive accuracy was rejected for each of these comparisons. The TFT–Informer comparison produced a DM statistic of -1.87 and a p-value of 0.0614. Consequently, although TFT exhibited lower forecasting error than Informer, the difference did not reach the conventional 5% significance threshold. It was, however, significant at the 10% level. This result indicates that the performance advantage over Informer was comparatively smaller and should be interpreted more cautiously than the differences observed against the remaining benchmarks.

Computational efficiency was also examined because improved predictive accuracy may be accompanied by greater computational cost. The classical ARIMA and SARIMA models were computationally lightweight, requiring approximately 0.5 and 0.8 minutes for training, respectively, with inference times of 0.8 and 1.2 milliseconds per sample and approximate model sizes of only 0.01 and 0.02 million parameters. Their computational advantage, however, was accompanied by markedly weaker forecasting accuracy. The LSTM model required approximately 45 minutes for training, 2.3 milliseconds per sample for inference, and contained approximately 1.20 million trainable parameters, whereas GRU required approximately 42 minutes, 2.1 milliseconds per sample, and 1.10 million parameters. The standard Transformer contained approximately 2.80 million parameters and required 78 minutes for training and 3.4 milliseconds per prediction. Informer was somewhat more computationally efficient, with a training time of approximately 65 minutes, an inference time of 2.8 milliseconds per sample, and 2.40 million parameters.

The proposed TFT had the greatest overall model complexity, containing approximately 3.20 million parameters and requiring approximately 89 minutes for training. This additional computational burden was attributable to its combination of the bidirectional LSTM representation, variable selection networks, Gated Residual Networks, and multi-head attention components. Nevertheless, its inference time was only approximately 3.1 milliseconds per observation. This was lower than the 3.4 milliseconds required by the standard Transformer and only moderately higher than the 2.8 milliseconds required by Informer. Therefore, despite the higher one-time training cost, the model remained computationally feasible for operational forecasting environments in which rapid generation of predictions is important. The findings consequently reveal a clear accuracy-complexity trade-off: classical models were considerably faster but substantially less accurate, while TFT required greater training resources but yielded the best predictive results without imposing a prohibitive inference-time penalty.

Taken together, the findings demonstrate that the proposed Temporal Fusion Transformer produced strong performance in three distinct dimensions. First, its point forecasts were highly accurate, with RMSE of 18.42 USD/oz, MAE of 13.87 USD/oz, MAPE below 1%, and R^2exceeding 0.95. Second, its probabilistic forecasts were well calibrated, as indicated by a PICP of 79.80% for a nominal 80% interval combined with a narrow PINRW of 0.0234. Third, its superiority remained observable across progressively longer forecasting horizons and was statistically significant against ARIMA, SARIMA, LSTM, GRU, and the standard Transformer. Although the difference relative to Informer did not reach the 5% significance level, TFT nevertheless produced consistently lower errors, higher explanatory power, and stronger directional accuracy. These results provide empirical support for the effectiveness, robustness, and practical suitability of the proposed attention-based deep learning framework for forecasting global gold prices.

4. Discussion and Conclusion

The present study evaluated an advanced attention-based deep learning framework built on the Temporal Fusion Transformer architecture for forecasting global gold prices, with particular emphasis on predictive accuracy, robustness across horizons, probabilistic calibration, and interpretability. The findings showed that the proposed TFT model achieved an RMSE of 18.42 USD/oz, an MAE of 13.87 USD/oz, a MAPE of 0.89%, and an $R^2$of 0.9534. These results indicate that the model reproduced more than 95% of the observed variation in gold prices while maintaining an average relative forecasting error below 1%. The probabilistic forecasting results were also favorable, with a PICP of 79.80% for the nominal 80% prediction interval and a PINRW of 0.0234. In comparative terms, the TFT model outperformed ARIMA, SARIMA, LSTM, GRU, the standard Transformer, and Informer across the principal accuracy measures, and its Mean Directional Accuracy reached 72.5%, the highest among all evaluated methods. Collectively, these findings suggest that the integrated use of temporal recurrence, dynamic variable selection, gating mechanisms, interpretable multi-head attention, and quantile-based outputs provides a more comprehensive representation of gold-price dynamics than approaches relying on only one or two of these mechanisms.

The superior performance of the proposed framework can first be explained by the nonlinear and time-varying nature of the gold market. Gold-price behavior is affected by interactions among monetary, financial, commodity, and uncertainty-related signals, and these relationships may change across market regimes. Conventional models such as ARIMA and SARIMA are useful for representing linear dependence and seasonality, but their relatively restrictive structure makes them less responsive to nonlinear interactions and sudden changes in market behavior. In the present study, ARIMA and SARIMA produced RMSE values of 42.56 and 38.91 USD/oz, respectively, substantially higher than that of TFT. This result is consistent with the growing literature showing that nonlinear and hybrid approaches frequently outperform conventional statistical models in gold-price forecasting. Quang demonstrated that combining SARIMA with LSTM and Random Forest can improve daily global gold-price forecasts by integrating linear and nonlinear information structures [2]. Similarly, Guo and colleagues showed that decomposition combined with optimized XGBoost can substantially improve forecasting accuracy by separating complex gold-price dynamics into more manageable components [1]. The present findings reinforce these studies by showing that an architecture capable of learning nonlinear structures internally can reduce forecasting error without depending exclusively on external decomposition or manually designed feature combinations.

The performance advantage over recurrent deep-learning models is also theoretically meaningful. LSTM and GRU achieved considerably better results than ARIMA and SARIMA, confirming the value of recurrent neural architectures for financial time-series prediction. However, their RMSE values of 26.34 and 24.78 USD/oz remained markedly higher than the 18.42 USD/oz obtained by TFT. This difference can be attributed to the fact that recurrent networks compress sequential information into evolving hidden states and may not optimally distinguish between highly informative and weakly informative historical periods. The attention mechanism incorporated in TFT enables the model to allocate greater weight to historical observations that are more relevant for the current prediction. This interpretation is consistent with Wu and Liu, who found that adding attention to LSTM improves gold-price forecasting by enabling selective emphasis on important temporal information [6]. It is also consistent with Liu and Lan, whose attention-enhanced CNN-BiGRU model was designed specifically to improve long-term gold-price prediction by strengthening temporal feature extraction and selective weighting [7]. Thus, the present results suggest that recurrent memory alone is useful but insufficient when temporal relevance changes dynamically over time.

The findings are likewise aligned with studies integrating attention with GRU and other recurrent architectures. Memon and colleagues reported that an enhanced GRU combined with multi-headed attention and skip connections improved forecasting performance for gold and platinum prices [8]. The results of the present study extend this logic by demonstrating that attention becomes even more effective when combined with explicit variable selection and gated residual processing. Rather than applying attention only to hidden temporal states, TFT simultaneously regulates which covariates are important, which nonlinear transformations should remain active, and which historical periods deserve greater emphasis. This layered selectivity likely contributed to the lower MAE and MAPE observed in the present study. In practical terms, the model is not forced to assume that the relevance of every input or every historical observation remains constant throughout the forecasting period.

The strong performance of TFT relative to the standard Transformer and Informer is particularly important. The standard Transformer achieved an RMSE of 21.56 USD/oz and an R^2 of 0.9123, while Informer achieved an RMSE of 20.13 USD/oz and an R^2 of 0.9256. Both models clearly outperformed classical and recurrent benchmarks, indicating that attention-based long-range dependency modeling is highly suitable for gold-price forecasting. Nevertheless, TFT produced lower errors and a higher explanatory coefficient. This result is compatible with the broader development of hybrid and long-sequence forecasting architectures. Zhao and colleagues showed that a hybrid LSTM-Transformer model with multi-scale feature fusion can improve gold futures forecasting by combining local recurrent information with global attention-based dependencies [10]. Yang similarly developed an attention-based long-sequence method for gold-price forecasting, emphasizing the importance of preserving information across extended temporal windows [9]. The current findings suggest that TFT gains additional value by integrating these long-range attention capabilities with recurrent local processing and interpretable feature-selection mechanisms.

The proposed model also performed consistently better across longer forecasting horizons. RMSE increased for all methods as the horizon expanded, which is expected because forecast uncertainty accumulates as predictions extend further into the future. However, TFT maintained the lowest errors at one, five, ten, and twenty-two days. At the twenty-two-day horizon, its RMSE was 49.87 USD/oz, compared with 54.67 for Informer, 58.23 for the standard Transformer, 68.91 for GRU, 71.34 for LSTM, 89.23 for SARIMA, and 95.67 for ARIMA. The relative preservation of accuracy at longer horizons indicates that the model captured temporal dependencies that extend beyond immediate short-run price movements. This finding is consistent with the rationale behind the original Temporal Fusion Transformer, which was developed specifically for interpretable multi-horizon forecasting and for integrating different categories of covariates through recurrent and attention-based components [12]. It is also consistent with the emphasis placed by Yang on long-sequence gold forecasting and by Liu and Lan on long-term prediction using attention-enhanced recurrent structures [7, 9].

The probabilistic forecasting results represent another important contribution. The PICP value of 79.80% was extremely close to the nominal 80% coverage level, while the PINRW value of 0.0234 indicated relatively narrow prediction intervals. This combination suggests that the model was neither systematically overconfident nor excessively conservative. Achieving high coverage is not sufficient by itself because a model can trivially increase coverage by producing very wide intervals. The simultaneous attainment of near-nominal coverage and narrow normalized width therefore indicates satisfactory uncertainty calibration. This finding is strongly aligned with the probabilistic forecasting framework proposed by Wang and Lin, who emphasized that gold-price forecasting should account not only for expected values but also for uncertainty around those predictions using quantile-regression deep learning [3]. In volatile financial markets, this feature is particularly important because the economic meaning of a forecast depends partly on how uncertain the model is regarding the predicted value.

The high R^2 value of 0.9534 and the MAPE of 0.89% also compare favorably with earlier deep-learning research on gold prices. Amini and Kalantari demonstrated that a CNN-Bi-LSTM architecture with automatic parameter tuning can improve gold-price prediction through the combination of local feature extraction and bidirectional sequential modeling [4]. Lin likewise proposed an improved CNN-LSTM model for daily gold-price forecasting, supporting the use of hybrid deep architectures for extracting complex temporal patterns [5]. The current results are conceptually consistent with these studies, but they indicate that explicit attention and dynamic variable-selection mechanisms can provide additional advantages beyond convolutional and recurrent processing. Rather than relying exclusively on hidden feature representations, TFT makes it possible to inspect the relative influence of covariates and historical periods, which substantially increases its practical and analytical value.

Interpretability is therefore one of the most important distinctions between the proposed framework and many competing deep-learning models. Financial decision-makers often require more than an accurate prediction; they need to understand the information structure underlying the forecast. The variable selection networks in TFT allow the relative contribution of different inputs to be evaluated dynamically, while interpretable attention weights indicate which historical periods influenced future estimates. This addresses a central concern in explainable financial machine learning. Izzo emphasized that explainability is particularly important in macroeconomic and financial forecasting because black-box predictions are difficult to validate economically and may be unsuitable for high-stakes decision environments [11]. The interpretability mechanisms embedded in TFT therefore provide not only technical transparency but also a bridge between predictive modeling and substantive financial reasoning.

The usefulness of multivariate information also deserves attention. Gold-price dynamics are rarely driven solely by their own historical trajectory. Changes in risk perception, macroeconomic conditions, and related markets can materially alter price behavior. Ge and colleagues showed that multivariate information can improve deep-learning-based volatility prediction by allowing models to incorporate broader financial signals [14]. The present study is consistent with this perspective because the proposed framework was explicitly designed to accommodate multiple time-varying explanatory variables and to assign them dynamic importance weights. This is particularly important under regime shifts, when the predictor that dominates in one period may become secondary in another. The TFT variable selection mechanism provides a natural way of representing this time-varying heterogeneity.

The comparison with Informer also provides a more nuanced interpretation of the findings. Although TFT had lower errors than Informer, the Diebold-Mariano test indicated that the difference was not statistically significant at the conventional 5% level, with $DM = -1.87$ and $p = 0.0614$. The difference was significant only at the 10% level. This suggests that Informer remains a strong benchmark and that the predictive advantage of TFT over advanced attention-based architectures is more modest than its advantage over classical, recurrent, or standard Transformer models. Such a finding is methodologically valuable because it prevents overstating the superiority of the proposed method. It indicates that the main contribution of TFT may not lie solely in incremental error reduction, but in the combination of strong predictive accuracy with probabilistic forecasting and intrinsic interpretability. This interpretation is also consistent with the emerging evidence comparing TFT with alternative advanced forecasting frameworks such as N-BEATS in gold-price applications [13].

The Diebold-Mariano results nonetheless showed statistically significant improvements over ARIMA, SARIMA, LSTM, GRU, and the standard Transformer. These results imply that the observed reductions in forecast loss were unlikely to be attributable merely to random sample variation. The significant differences against LSTM, GRU, and the standard Transformer are especially important because these models are substantially more sophisticated than conventional econometric benchmarks. Therefore, the results provide evidence that the specific architectural combination employed in TFT adds predictive value. This conclusion aligns with the broader methodological trend toward integrated architectures that combine local pattern extraction, recurrent memory, attention, and multi-scale information fusion [4, 8, 10].

The computational findings also reveal an important trade-off. TFT required approximately 89 minutes of training and contained about 3.20 million parameters, making it the most computationally demanding model evaluated. In contrast, ARIMA and SARIMA required less than one minute for training, while LSTM and GRU were substantially lighter than TFT. However, the inference time of TFT was only about 3.1 milliseconds per sample, which remained operationally practical and was slightly lower than that of the standard Transformer. Thus, the computational burden is concentrated primarily in model training rather than prediction. For real-world forecasting systems, this distinction is important because models may be retrained periodically but used repeatedly for rapid inference. The higher training complexity may therefore be acceptable when the gain involves improved accuracy, uncertainty estimation, and interpretability.

From the perspective of the Iranian market, these findings are particularly relevant. Global gold prices constitute an important external signal for domestic gold valuation, while the local market is additionally affected by exchange-rate fluctuations, inflation, and macroeconomic uncertainty. A model capable of producing accurate international gold forecasts with uncertainty intervals can therefore support more disciplined risk assessment. The high directional accuracy of 72.5% is especially noteworthy because many practical decisions depend on anticipating upward or downward price movements rather than estimating the exact future price alone. At the same time, the interpretability features of TFT can help analysts distinguish whether the forecast is driven primarily by recent temporal behavior, broader market conditions, or specific explanatory variables.

A principal limitation of the present study is that the empirical evaluation was based on a finite historical period and therefore cannot encompass all possible future market regimes. Although the ten-year sample contained multiple episodes of economic and geopolitical instability, future structural breaks may differ fundamentally from those observed during model training. The dataset was also centered on daily observations, meaning that intraday microstructure effects, high-frequency liquidity shocks, and short-lived behavioral responses were not directly modeled. Moreover, although the TFT architecture provided variable-importance and attention outputs, such measures should not automatically be interpreted as causal effects. The model identifies predictive relevance rather than establishing economic causality. Finally, the higher computational burden of TFT may limit direct deployment in environments with restricted hardware or frequent retraining requirements.

Future research should extend the framework by incorporating a broader range of macroeconomic, behavioral, and geopolitical variables, including exchange-rate measures directly relevant to Iran, policy uncertainty indices, central-bank communication indicators, sentiment extracted from financial news, and structured geopolitical-risk measures. Researchers could also evaluate the model using rolling-window and expanding-window backtesting to examine stability under repeated real-time forecasting conditions. Additional comparisons with architectures such as N-BEATS, PatchTST, Autoformer, and other recent long-horizon forecasting models would provide a stronger benchmark. Future studies should also investigate regime-aware or ensemble variants of TFT, test the benefits of alternative quantile sets and conformal prediction methods, and examine whether interpretability outputs remain stable across market regimes and alternative random initializations.

For practice, the proposed framework can be used as a decision-support tool rather than as an autonomous trading mechanism. Investors and portfolio managers can combine the median forecast with the 80% prediction interval to distinguish high-confidence from high-uncertainty market conditions, while directional forecasts may assist in timing hedging or allocation decisions. Analysts should also monitor changes in variable-importance and attention patterns because shifts in these outputs may signal changes in the structure of the market itself. For applications in Iran, the global gold forecast should be integrated with domestic exchange-rate, inflation, and liquidity indicators before making final decisions, and the model should be retrained periodically so that new structural changes are incorporated into its learned relationships.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

[1] Y. Guo, C. Li, X. Wang, and Y. Duan, "Gold Price Prediction Using Two-Layer Decomposition and XGBoost Optimized by the Whale Optimization Algorithm," *Computational Economics,* vol. 66, no. 2, pp. 1157-1189, 2025, doi: 10.1007/s10614-024-10736-9.

[2] P. D. Quang, "Analysis and Forecasting of Daily Global Gold Price: An SARIMA-LSTM Approach with Random Forest Technique," *Cogent Economics & Finance,* vol. 13, no. 1, p. 2568969, 2025, doi: 10.1080/23322039.2025.2568969.

[3] Y. Wang and T. Lin, "A Novel Deterministic Probabilistic Forecasting Framework for Gold Price with a New Pandemic Index Based on Quantile Regression Deep Learning and Multi-Objective Optimization," *Mathematics,* vol. 12, no. 1, p. 29, 2024, doi: 10.3390/math12010029.

[4] A. Amini and R. Kalantari, "Gold Price Prediction by a CNN-Bi-LSTM Model along with Automatic Parameter Tuning," *PLOS ONE,* vol. 19, no. 3, p. e0298426, 2024, doi: 10.1371/journal.pone.0298426.

[5] Y. Lin, "An Improved CNN-LSTM Model for Daily Gold Price Prediction," 2026: Atlantis Press, pp. 608-621, doi: 10.2991/978-94-6239-648-7_53.

[6] J. Wu and J. Liu, "Gold Price Prediction Based on LSTM-Attention Combined Model," *Open Access Library Journal,* vol. 13, p. e1114759, 2026, doi: 10.4236/oalib.1114759.

[7] S. Liu and Q. Lan, "Long-Term Gold Price Prediction Using the CNN-BiGRU Model Enhanced with an Attention Mechanism," 2025: IEEE, pp. 1369-1375, doi: 10.1109/ICAACE65325.2025.11020201.
[8] B. A. Memon, R. Tahir, H. M. Naveed, and K. Cheng, "Forecasting Gold and Platinum Prices with an Enhanced GRU Model Using Multi-Headed Attention and Skip Connection," *Mineral Economics,* vol. 38, no. 4, 2025, doi: 10.1007/s13563-025-00520-y.
[9] Y. Yang, "TCN-QV: An Attention-Based Deep Learning Method for Long Sequence Time-Series Forecasting of Gold Prices," *PLOS ONE,* vol. 20, no. 5, p. e0319776, 2025, doi: 10.1371/journal.pone.0319776.
[10] Y. Zhao, Y. Guo, and X. Wang, "Hybrid LSTM-Transformer Architecture with Multi-Scale Feature Fusion for High-Accuracy Gold Futures Price Forecasting," *Mathematics,* vol. 13, no. 10, p. 1551, 2025, doi: 10.3390/math13101551.
[11] C. Izzo, "On Explainable Deep Learning for Macroeconomic Forecasting and Finance," UCL (University College London), 2022.
[12] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting," *International Journal of Forecasting,* vol. 37, no. 4, pp. 1748-1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.
[13] V. Egy, "PERBANDINGAN KINERJA MODEL TFT DAN N-BEATS DALAM PREDIKSI HARGA EMAS DENGAN INKORPORASI FITUR MAKROEKONOMI," ed, 2025.
[14] W. Ge, P. Lalbakhsh, L. Isai, A. Lensky, and H. Suominen, "Comparing Deep Learning Models for the Task of Volatility Prediction Using Multivariate Data," 2023.