Mixed-Frequency Deep Learning for GDP Forecasting: A Real-Time Framework*

Mohammad Soltaniehha† Lingyi Xu‡ Marialuz Moreno Badia§

March 26, 2026

Abstract

We develop a framework for macroeconomic time series forecasting using neural networks in a real-time setting with mixed-frequency data. Our approach processes quarterly, monthly, and weekly indicators through separate LSTM branches that preserve the natural sampling frequency of each data source, in contrast to conventional methods that require temporal aggregation or alignment. Applied to U.S. GDP growth, the framework produces forecasts that are statistically indistinguishable from the Survey of Professional Forecasters across all horizons. Three findings stand out. First, high-frequency data substantially reduce forecast errors at short horizons, particularly during economic turning points. Second, the informational value of data is dynamic: weekly indicators are most valuable for nowcasting, while quarterly variables dominate at longer horizons. Third, real-time vintage data are essential for honest forecast evaluation; models trained on revised data yield artificially inflated performance that does

*The views expressed are those of the authors and do not necessarily represent the views of the IMF, its Executive Board, or IMF management. We thank Paul Rudrendu Kumar for his support with data processing in the early stages of the project.

†Boston University, Questrom School of Business. E-mail: msoltani@bu.edu

‡Boston University, Faculty of Computing & Data Sciences. E-mail: lingyixu@bu.edu

§International Monetary Fund. E-mail: mmorenobadia@imf.org

not carry over to genuinely out-of-sample predictions. Model interpretability analysis using Shapley values reveals that the relative importance of variable groups shifts across time and forecast horizons, with the Aruoba-Diebold-Scotti Business Conditions Index playing a particularly prominent role during recessions.

Keywords: GDP forecasting, mixed-frequency data, neural networks, nowcasting, real-time data, LSTM, machine learning, SHAP

JEL Classification: C32, C45, C53, E37

1 Introduction

Forecasting GDP growth is a central task in macroeconomics, yet it remains inherently difficult. GDP is published at a quarterly frequency with a considerable lag, and the initial release is subject to multiple rounds of revision (Croushore, 2011). Forecasters must therefore rely on higher-frequency indicators—monthly employment figures, industrial production, financial variables, and increasingly, weekly or daily data—that become available at different times and are themselves revised. The resulting "ragged edge" of the data poses well-known challenges for real-time forecasting (Bańbura et al., 2013; Clark et al., 2022).

The standard approach to this problem employs dynamic factor models (DFMs), which extract a small number of common factors from a large panel of indicators and use these to produce forecasts at the desired frequency (Stock and Watson, 2002b; Giannone et al., 2008). An alternative class of methods, Mixed Data Sampling (MIDAS) regressions, directly accommodates regressors observed at different frequencies through parsimonious lag polynomials (Ghysels et al., 2004; Foroni et al., 2015). Both approaches have proven effective, but they rely on linear structures and typically require the researcher to impose a specific temporal aggregation scheme. A persistent finding in this literature is that the consensus forecast from the Survey of Professional Forecasters (SPF) is difficult to beat, reflecting both the judgment and the diverse information sets of professional forecasters (Liebermann, 2014; Stark, 2010; Diebold et al., 2025).

At the same time, the toolkit available for macroeconomic forecasting has expanded considerably. Machine learning methods, including random forests, gradient-boosted trees, and neural networks, have gained traction as forecasting tools (Coulombe et al., 2022; Richardson et al., 2021; Soybilgen and Yazgan, 2021). In a comprehensive benchmark of 17 nowcasting methodologies, Hopp (2024) finds that LSTM networks and Bayesian VARs consistently rank as the top-performing approaches for GDP nowcasting across both crisis and tranquil periods. A central advantage of these methods is their capacity to capture nonlinear relationships in high-dimensional settings without imposing a parametric structure, and their

use in macroeconomic forecasting has been explored in recent work (Sagheer and Kotb, 2019; Paranhos, 2023). Moreover, as Giannone et al. (2021) argue, the assumption of sparsity that underpins many traditional approaches may be an "illusion" when dealing with macroeconomic data; richer models that allow for dense interactions can yield meaningful gains even within a linear framework, and the potential benefits are amplified when combined with the nonlinear flexibility of machine learning methods. More recently, large language models and other foundation models have been evaluated for macroeconomic prediction, though their applicability to real-time forecasting with vintaged data remains limited (Carriero et al., 2025; Jha et al., 2025).

This paper makes four contributions. First, we develop a Long Short-Term Memory (LSTM) neural network architecture specifically designed to handle mixed-frequency data in a macroeconomic forecasting context. Unlike MIDAS or DFM-based approaches, our framework processes quarterly, monthly, and weekly indicators through separate branches that each pass through their own LSTM layer before merging. This preserves the natural sampling frequency of each data source and allows the model to learn frequency-specific temporal patterns. Second, we implement the entire framework in a real-time setting using vintage data. For each forecast date, the model receives only the information that would have been available to a forecaster at that time, including the specific data vintages with their associated publication lags and missing values. This stands in contrast to much of the machine learning forecasting literature, which often relies on the latest-available (revised) data, a practice that can introduce severe look-ahead bias (Kapoor and Narayanan, 2023; Croushore and Stark, 2006). Third, we provide a systematic analysis of how the informational value of data varies across frequencies, forecast horizons, and time periods. We document that high-frequency data are most valuable at short horizons and during economic turning points, while lower-frequency data dominate at longer horizons. Fourth, we use SHAP (Shapley Additive Explanations) values to provide transparent model interpretation, identifying which variable groups drive forecasts and how their importance evolves

over time.

Our empirical application focuses on forecasting U.S. GDP growth from 2001Q3 through 2024Q1 using an expanding window scheme. The dataset combines 232 quarterly features, 129 monthly features, and the weekly Aruoba-Diebold-Scotti (ADS) Business Conditions Index (Aruoba et al., 2009), drawn from the Federal Reserve Economic Data (FRED) system (McCracken and Ng, 2016). We compare our LSTM forecasts against the SPF median forecast across horizons ranging from nowcasting to four quarters ahead.

The results indicate that the LSTM model performs on par with the SPF across all forecast horizons: the Diebold-Mariano test (Diebold and Mariano, 1995) reports no statistically significant differences in predictive accuracy over the full evaluation period. The model matches the SPF at short horizons and shows improvements at horizons of two quarters and beyond, consistent with the model's ability to exploit accumulated data dynamics for longer-range prediction. During the COVID-19 recession, the LSTM achieves a nowcasting RMSE roughly half that of the New York Fed's dynamic factor model (Almuzara et al., 2023), and it significantly outperforms the SPF at the two-quarter-ahead horizon.

The remainder of the paper is organized as follows. Section 2 describes the data and the construction of real-time vintages. Section 3 presents the model architecture and estimation procedure. Section 4 reports the forecasting results and comparisons with the SPF. Section 5 provides the model interpretation analysis. Section 6 concludes.

2 Data

2.1 Data Sources and Structure

Our dataset draws on the Federal Reserve Economic Data (FRED) system, which provides a comprehensive repository of macroeconomic indicators at various frequencies (McCracken and Ng, 2016, 2021). We organize the data into three frequency groups: quarterly, monthly, and weekly.

The quarterly component consists of 232 features spanning 12 variable groups: output and income, labor market, consumption and orders, housing, inventories, money and credit, interest and exchange rates, stock market prices, household balance sheets, nonhousehold balance sheets, prices, and surveys. The monthly component comprises 129 features drawn from the same variable groups where monthly counterparts are available, including industrial production, employment, retail sales, housing starts, and financial market indicators. The weekly component consists of a single variable, the ADS Business Conditions Index, which is a composite measure of real economic activity constructed from a dynamic factor model applied to six underlying indicators at mixed frequencies.

We use the median GDP growth forecast from the Survey of Professional Forecasters (SPF), maintained by the Federal Reserve Bank of Philadelphia, as our benchmark. The SPF provides quarterly forecasts at multiple horizons and is widely regarded as a strong benchmark that incorporates the collective judgment of professional forecasters (Stark, 2010; Diebold et al., 2025).

2.2 Real-Time Vintage Data

A distinguishing feature of our framework is the strict use of real-time vintage data for both model training and evaluation. Different economic indicators become available at different points in time, and most are subject to revision as more complete information arrives. For instance, variables such as net worth of households relative to disposable personal income typically report with a two-month lag, while stock price indices are updated with minimal delay. Using the latest-available data to train and evaluate forecasting models introduces look-ahead bias, a problem that Kapoor and Narayanan (2023) identify as a widespread source of reproducibility failures in machine-learning-based research. Croushore (2011) and Croushore and Stark (2006) provide detailed discussions of the challenges posed by data revisions for forecast evaluation.

To avoid these issues, we implement a real-time framework in which, for each forecast

date, the model receives only the data vintage that would have been available at that point. The earliest actual vintage data from the Federal Reserve are available from August 1999. For earlier periods, we construct artificial vintages by systematically dropping the most recent values that would not yet have been released, following the specific publication lag structure of each variable. While this procedure cannot replicate the exact data revisions that occurred in real time, these artificial vintages fall outside our out-of-sample evaluation period and therefore do not affect our forecast assessment. For the weekly ADS index, actual vintage data are available from December 2008, with earlier vintages constructed artificially using the same approach.

2.3 Data Preparation

The data undergo several preparation steps before entering the model. Variables with more than 30 percent of their observations missing are discarded to maintain data quality. Missing values at the interior and end of each vintage series are handled through forward-fill imputation, which preserves temporal continuity without introducing information from future observations. Remaining missing values at the beginning of a series are filled using the series median.

Each variable is then transformed to achieve approximate stationarity. Depending on the variable's characteristics, transformations include first or second differencing, logarithmic transformations, or combinations thereof, following the transformation codes established in the FRED-MD and FRED-QD databases (McCracken and Ng, 2016, 2021).

Finally, within each vintage, we standardize all features by subtracting the vintage-specific mean and dividing by the vintage-specific standard deviation. This ensures that no information from future vintages leaks into the scaling procedure.

3 Methodology

3.1 Model Architecture

The core challenge of incorporating mixed-frequency data into a neural network lies in managing the different temporal structures across data frequencies. Weekly data offer timely signals for short-horizon forecasting but provide limited information for predicting several quarters ahead. Quarterly data, by contrast, capture lower-frequency dynamics that are more informative for longer horizons. Conventional approaches to mixed-frequency forecasting, such as MIDAS regressions (Ghysels et al., 2004), address this by imposing parsimonious lag polynomial structures that map high-frequency observations to the low-frequency target. Factor-based approaches (Mariano and Murasawa, 2003; Bańbura et al., 2010) estimate latent factors jointly across frequencies using state-space models, though this comes with computational costs that grow with the number of frequencies modeled.

Our approach takes a different route. We process each frequency through a dedicated branch of a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997; Pascanu et al., 2013). Each branch receives a sequence of observations at its native frequency and passes them through an LSTM layer followed by a dense layer. The outputs of the three branches are then concatenated and fed into a final dense layer that produces the GDP growth forecast. This "multi-branch" architecture preserves the natural sampling frequency of each data source and allows the LSTM's gating mechanism—the forget gate, input gate, and output gate—to learn frequency-specific patterns of memory retention and updating.

Figure 1 illustrates this architecture. Specifically, the quarterly branch receives 8 lags (covering two years) of the quarterly principal components, the monthly branch receives 13 lags (one year plus the first month of the current quarter), and the weekly branch receives 13 lags (one quarter) of the ADS index.

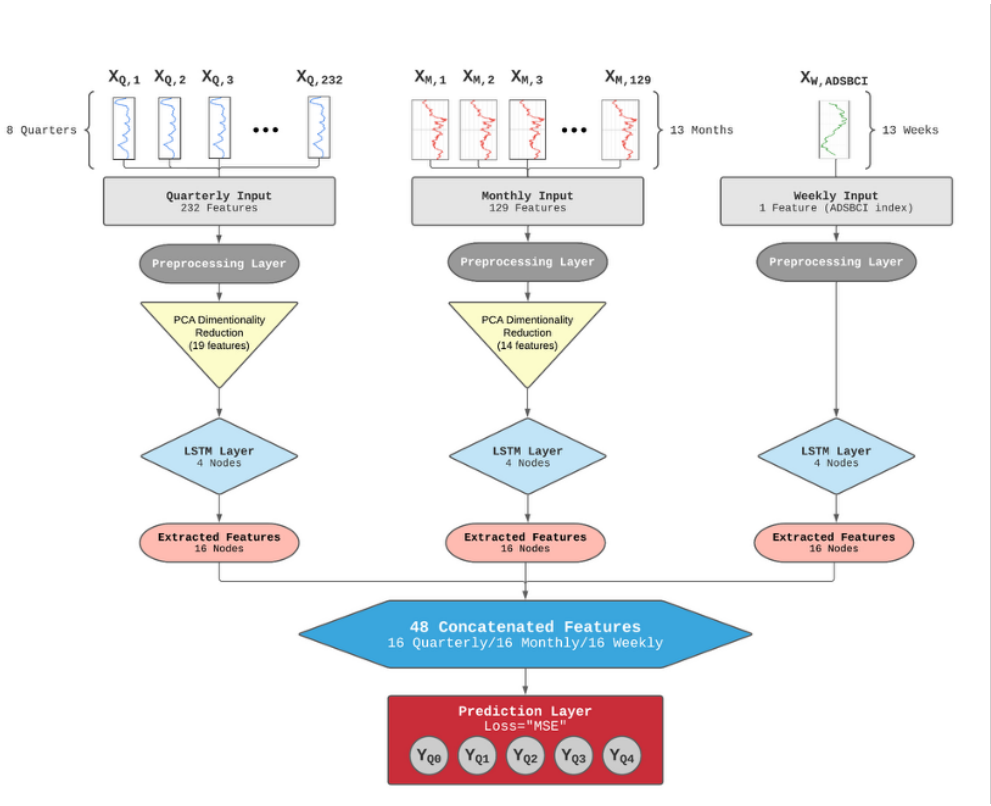


Figure 1: Multi-branch LSTM architecture. Quarterly and monthly features undergo within-group PCA for dimensionality reduction before entering their respective LSTM branches. The weekly branch receives the ADS index directly. Each branch processes its input sequence through an LSTM layer and a dense layer. The three branch outputs are concatenated and passed through a final dense layer to produce the GDP growth forecast.

These lag lengths were chosen to balance the trade-off between providing sufficient historical context and avoiding the inclusion of excessively distant observations that would dilute the signal. The model produces forecasts at five horizons: nowcasting (Q0) through four quarters ahead (Q4).

To explore the value of different frequency combinations, we estimate several model variants: a quarterly-only model (Q), a quarterly-monthly model (QM), and the full quarterly-monthly-weekly model (QMW). Additionally, for the QMW specification, we consider three sub-variants corresponding to different amounts of within-quarter information: month-one (Mo1), month-two (Mo2), and month-three (Mo3), where the monthly and weekly vintages reflect data available at the end of the first, second, or third month of the current quarter,

respectively. Comparing these variants allows us to quantify the marginal contribution of each frequency and of additional within-quarter data releases.

3.2 Dimensionality Reduction

The relatively small number of quarterly observations compared to the large number of features necessitates dimensionality reduction. We apply Principal Component Analysis (PCA) within each of the 12 variable groups separately, following the factor-based tradition in macroeconomic forecasting (Stock and Watson, 2002a; Bai and Ng, 2002). For each group, we retain one principal component if the first component explains at least 98 percent of the within-group variance, and two components otherwise. This procedure preserves over 80 percent of the variance for most groups while substantially reducing the input dimensionality.

A separate PCA model is estimated for each variable group at the quarterly and monthly frequencies. The PCA models are estimated using data through 2001Q2—the end of the validation period—and the resulting loadings are applied consistently across all subsequent vintages. This ensures that no information from the out-of-sample test period leaks into the dimensionality reduction step. PCA is not applied to the weekly frequency because only the ADS index is used, so no further reduction is needed. The resulting framework parallels the DFM literature, where information criteria guide the selection of the number of factors (Bai and Ng, 2002), while connecting naturally to recent work on PCA for potentially nonstationary series (Hamilton et al., 2024).

3.3 Training, Validation, and Testing

The dataset is divided into three segments: training (1959Q3 through 1990Q3), validation (1990Q4 through 2001Q2), and testing (2001Q3 through 2024Q1). Hyperparameter tuning, including the number of LSTM layers, the number of units per layer, the learning rate, and the dropout rates, is performed over the validation period using grid search. No information from the test period is used to select model hyperparameters.

During the test period, we employ an expanding-window scheme: before each forecast, the model is retrained using all available data up to that point. This allows the model to adapt to evolving economic conditions while maintaining the temporal integrity of the evaluation. The choice of 2001Q3 as the start of the test period ensures an adequate out-of-sample span that includes the 2001 recession, the Great Financial Crisis, and the COVID-19 recession, providing evaluation across diverse economic regimes. Figure 2 illustrates this expanding-window design, showing how each successive forecast date adds one quarter of observations to the training set while the forecast horizon shifts forward.

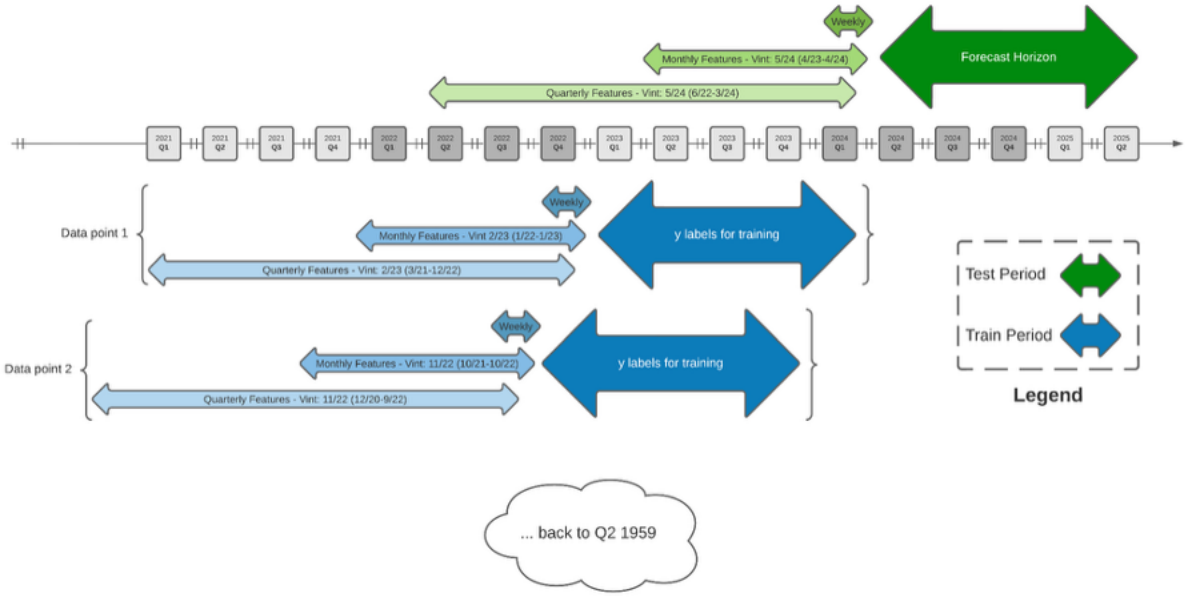


Figure 2: Expanding-window data structure. Each data point pairs a set of quarterly features (8 lags) and monthly features (13 lags) with the corresponding GDP growth target. The training window expands from 1959Q3 forward as each new forecast date is reached, while the test period (green) covers 2001Q3–2024Q1. For each forecast, the model uses only information available at that point, preserving real-time validity.

3.4 Regularization and Ensemble Forecasting

Overfitting is a well-known concern when applying deep learning methods to the relatively small samples typical of macroeconomic data. We address this through two forms of regularization. First, following Zaremba et al. (2014), we apply dropout to the non-recurrent

connections of the LSTM, which preserves the model's ability to maintain long-term memory while preventing co-adaptation of neurons. Second, we apply dropout to the dense layers that follow each LSTM branch. In addition, batch normalization (Ioffe and Szegedy, 2015) is applied within the model layers to stabilize training.

Forecasts are constructed as the average of 20 model runs, each initialized with a different random seed but sharing the same architecture and training data. This ensemble approach reduces the variance of individual predictions and provides a natural measure of forecast uncertainty through the dispersion across runs (Dietterich, 2000). The choice of 20 ensemble members was found to provide stable results in practice, though this number can be increased at the cost of additional computation.

3.5 Forecast Evaluation

We evaluate forecasts using root mean squared error (RMSE), with the ground truth defined as the most recently available GDP release from FRED. To test whether differences in forecast accuracy between the LSTM and the SPF are statistically significant, we apply the Diebold-Mariano test (Diebold and Mariano, 1995), which assesses the null hypothesis of equal predictive ability based on the loss differential between two competing forecasts.

4 Results

4.1 Overall Performance

Table 1 reports the RMSE of the LSTM model (using the QMW Mo2 specification) and the SPF median forecast across five horizons, along with the p-values of the Diebold-Mariano test for the full evaluation period (2001Q3–2024Q1), the pre- and post-COVID period (2001Q3–2019Q4 and 2021Q1–2024Q1), and the COVID period (2020Q1–2020Q4).

Table 1: Forecast Performance: LSTM vs. SPF (RMSE)

	Q0 (Nowcast)	Q1	Q2	Q3	Q4
Full Sample (2001Q3–2024Q1)					
LSTM	4.131	5.423	5.186	5.206	5.279
SPF	2.558	4.655	5.257	5.309	5.286
DM test (p-value)	0.142	0.319	0.245	0.327	0.941
Pre- and Post-COVID (2001Q3–2019Q4 & 2021Q1–2024Q1)					
LSTM	1.900	2.240	2.351	2.281	2.312
SPF	1.797	2.206	2.384	2.412	2.437
DM test (p-value)	0.447	0.784	0.785	0.365	0.401
COVID Period (2020Q1–2020Q4)					
LSTM	17.597	23.663	22.172	22.436	22.756
SPF	8.869	19.675	22.476	22.685	22.507
DM test (p-value)	0.178	0.399	0.047*	0.651	0.577

* denotes significance at the 5% level.

Over the full evaluation period, the Diebold-Mariano test reports no statistically significant difference between the LSTM and the SPF at any horizon. Looking at the point estimates, the LSTM trails the SPF at the nowcasting and one-quarter-ahead horizons, but produces lower RMSEs from the two-quarter-ahead horizon onward. The same pattern holds in the pre- and post-COVID subsample. This pattern likely reflects the different information advantages of each approach: the SPF incorporates the real-time judgment and diverse information sets of professional forecasters, which is especially valuable at very short horizons, while the LSTM benefits from its ability to extract systematic patterns from accumulated data that become more informative as the forecast horizon lengthens.

The COVID period warrants separate discussion. Quarter-over-quarter GDP growth fell by 28 percent (annualized) in 2020Q2 before rebounding by nearly 35 percent in 2020Q3.

Variations of this magnitude are notoriously difficult for any model to capture. For context, the New York Fed's dynamic factor model (Almuzara et al., 2023) recorded a nowcasting RMSE of 38.73 during this period. The LSTM's nowcasting RMSE of 17.60 is less than half that figure. Moreover, at the two-quarter-ahead horizon, the LSTM significantly outperforms the SPF ($p = 0.047$), with an RMSE improvement of 0.30 percentage points. This result provides evidence that the LSTM architecture can process rapidly changing information effectively, even during unprecedented economic dislocations.

4.2 The Value of Mixed-Frequency Data

A central question motivating our framework is whether richer information structures improve forecasting performance. We address this by comparing models that use different frequency combinations: quarterly only (Q), quarterly-monthly (QM), and the full quarterly-monthly-weekly (QMW) specification (Table 2).

Table 2: Diebold-Mariano Test p-values Across Model Pairs

	Q0 (Nowcast)	Q1	Q2	Q3	Q4
Across Frequency Model Pairs (2001Q3–2024Q1)					
Q-QM	0.059*	0.731	0.903	0.530	0.974
Q-QMW	0.071*	0.693	0.376	0.438	0.773
QM-QMW	0.096*	0.747	0.374	0.399	0.401
Within QMW Model Variants (2001Q3–2024Q1)					
QMW Mo1–QMW Mo2	0.212	0.475	0.932	0.204	0.279
QMW Mo1–QMW Mo3	0.197	0.075*	0.053*	0.649	0.153
QMW Mo2–QMW Mo3	0.213	0.163	0.084*	0.936	0.473

* denotes $p < 0.1$.

The results indicate that the marginal value of adding higher-frequency data depends on the forecast horizon. For nowcasting and one-quarter-ahead forecasts, incorporating monthly

and especially weekly data yields meaningful improvements, consistent with the idea that higher-frequency indicators capture timely signals about the current state of the economy. For longer horizons—three and four quarters ahead—the gains from adding monthly or weekly data are modest, and the quarterly-only model performs comparably. This finding is consistent with the time-series data's time-sensitive nature: the informational content of high-frequency data depreciates as the forecast horizon lengthens, while the lower-frequency dynamics embedded in quarterly series remain informative.

Within the QMW specification, the comparison across monthly vintages (Mo1, Mo2, and Mo3) reveals that the first within-quarter data release (Mo1 to Mo2) provides a statistically significant improvement for nowcasting and one-quarter-ahead forecasts in the pre- and post-COVID period. The second update (Mo2 to Mo3) offers no additional significant improvement at any horizon. This suggests that the bulk of the within-quarter information gain occurs with the initial data release, and subsequent updates contribute primarily to precision rather than to qualitative changes in the forecast.

4.3 Real-Time vs. Latest-Available Data

To quantify the importance of using real-time vintage data, we compare the QMW model estimated with vintage data against a "snapshot" model that uses the latest-available (fully revised) data at each forecast date (Table 3). The snapshot model consistently produces lower RMSEs for short time horizons, reflecting its access to information that would not have been available in real time. This gap demonstrates the severity of look-ahead bias when revised data are used for model evaluation. The real-time model produces the forecasts that a practitioner would actually have obtained, and it is these forecasts that should be used for fair comparison with benchmarks such as the SPF. This underscores the importance of vintage-based evaluation, a point emphasized by Croushore and Stark (2006) and Stark (2010) in the context of survey forecasts.

Table 3: Forecast Performance: LSTM QMW Model using Real-Time vs. Latest-Available Data (RMSE)

	Q0 (Nowcast)	Q1	Q2	Q3	Q4
Full Sample (2001Q3–2024Q1)					
LSTM (QMW; real-time)	4.131	5.423	5.186	5.206	5.279
LSTM (QMW; latest-available)	3.986	4.849	5.264	5.263	5.217

5 Model Interpretation

Neural network models are sometimes criticized for their lack of interpretability (Karpathy et al., 2015; Guidotti et al., 2018; Samek et al., 2021). We address this by applying SHAP (Shapley Additive Explanations) values (Lundberg and Lee, 2017; Shapley, 1953) to decompose the LSTM's predictions into contributions from individual features and feature groups (Lundberg et al., 2020). Specifically, we use the gradient explainer (Erion et al., 2021; Sundararajan et al., 2017; Smilkov et al., 2017), which combines integrated gradients with gradient smoothing and is well-suited to models with correlated features—a characteristic of macroeconomic data.

For interpretation, we aggregate absolute SHAP values at the variable-group level, summing over the principal components within each group to obtain a group-level importance measure. Using absolute values ensures that positive and negative contributions do not offset one another, and it focuses the analysis on the magnitude of each group's influence on the forecast.

5.1 Variable Importance Across Frequencies, Lags, Horizons, and Time

Across specifications, we find a consistent pattern in how predictive information is distributed over data frequencies, lags, and forecast horizons. High-frequency (weekly) indicators are most informative for nowcasting and short-horizon forecasts, particularly around economic turning points, while lower-frequency (quarterly) variables become increasingly important as the forecast horizon extends. In addition, recent observations (short lags) contribute most to predictive power, especially at longer horizons. These patterns are stable over time and become even more pronounced during periods of rapid economic change, such as the COVID-19 recession.

Figure 3 summarizes variable importance across quarterly, monthly, and weekly frequencies, aggregated over all lags from 2001Q3 to 2024Q1. Weekly data are mostly informative for nowcasting (Q0), whereas quarterly data dominate at longer forecast horizons (Q1-Q4). In addition, recent lags (1-3) account for the largest share of predictive power, especially at longer horizons. These patterns reflect the time-sensitive nature of macroeconomic data: higher-frequency indicators are most valuable for near-term predictions, while lower-frequency information becomes more relevant as the forecast horizon extends.

Complementing this, Figure 4 shows how feature importance by frequency evolves over time. Weekly data become especially influential around turning points for nowcasting (Q0), notably during the Global Financial Crisis and the COVID-19 recession in 2020. For longer-horizon forecasts, quarterly data play a more prominent role, with their importance increasing with the forecast horizon, while the contribution of weekly data declines sharply. This again highlights the differential roles of data frequencies: high-frequency indicators are more responsive to contemporaneous economic conditions, whereas lower-frequency data provide more stable signals for longer-term forecasts.

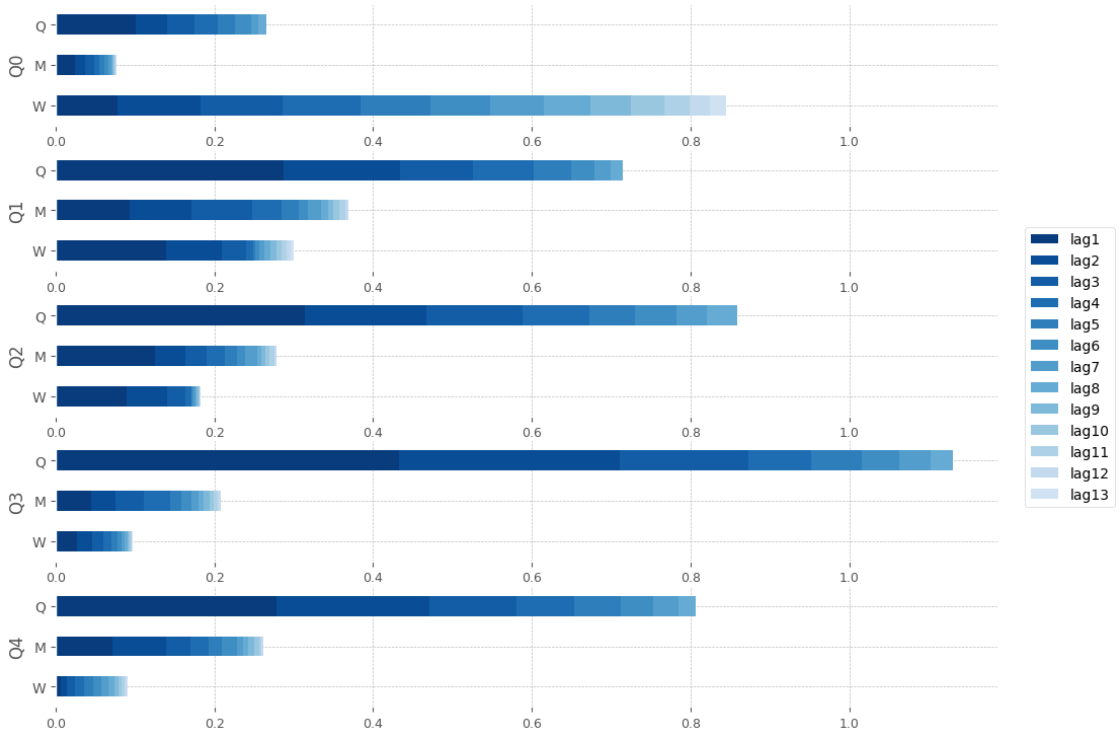


Figure 3: SHAP value by forecast horizon and data frequency, full out-of-sample period (2001Q3 - 2024Q1).

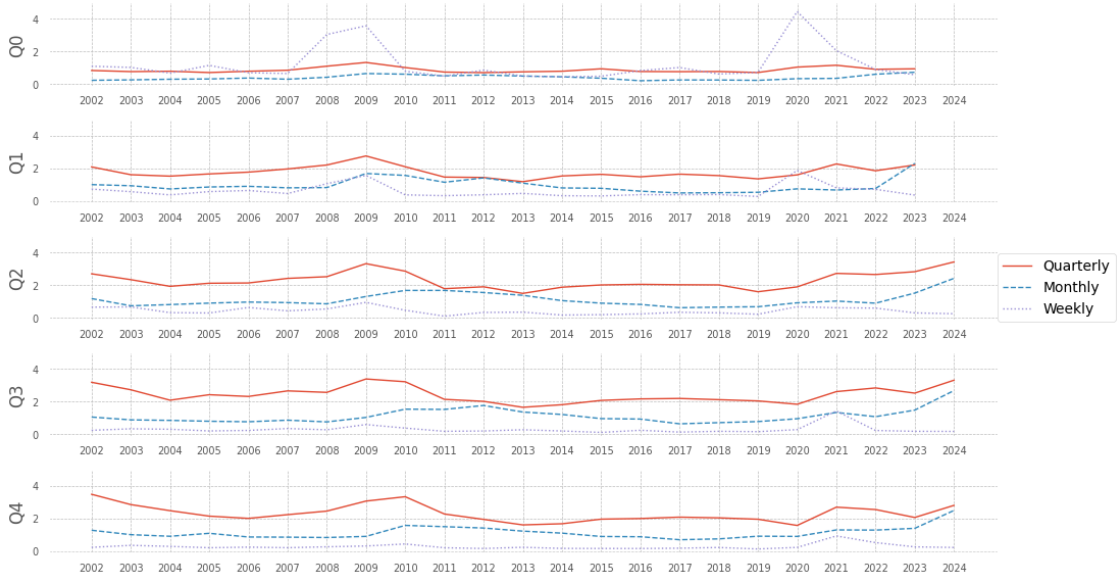


Figure 4: SHAP value over time by forecast horizon and data frequency, aggregated by fiscal year. Note that nowcasting (Q0) and one quarter ahead forecasting (Q1) lack data for 2024, due to the setting that predictions are up to 2024Q2 and 2024Q3, respectively, therefore no fiscal aggregation could be generated.

These patterns are consistent in both the pre- and post-COVID periods (Figure 5). During the COVID period (Figure 6), the importance of weekly data becomes even more pronounced for nowcasting (Q0) and short-horizon forecasting (Q1). This provides further evidence that higher-frequency data offer more timely information during periods of rapid economic change, when lower-frequency indicators may lag behind underlying dynamics.

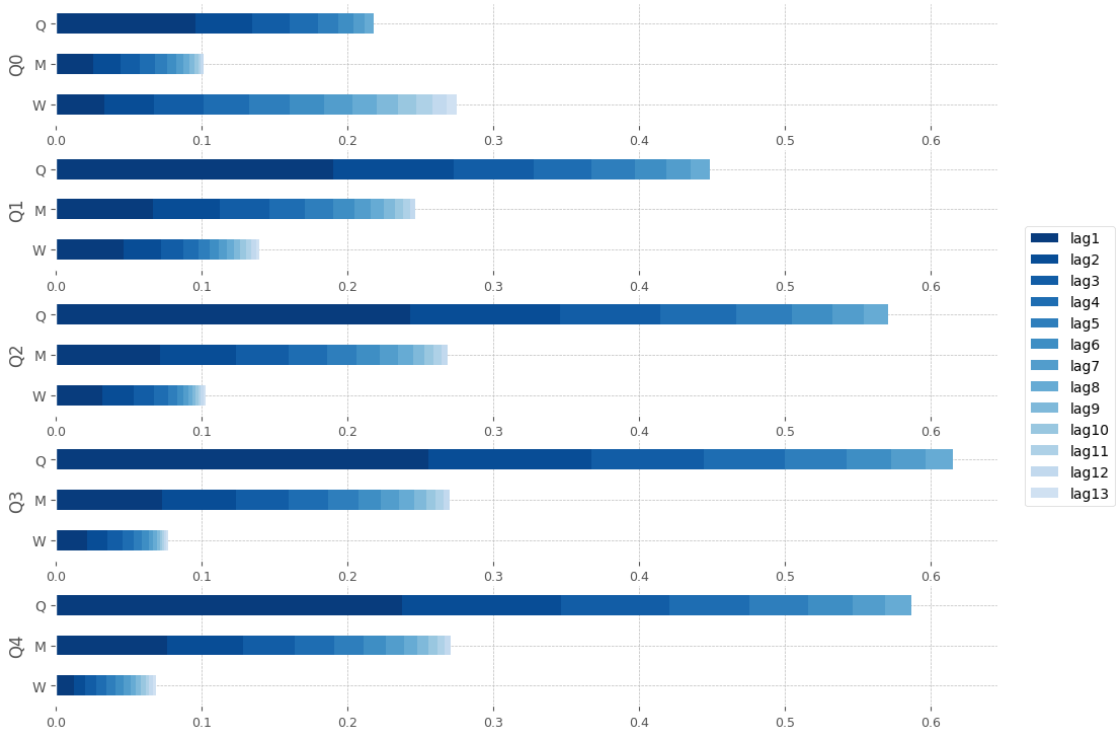


Figure 5: SHAP value by forecast horizon and data frequency, pre- and post- COVID period (2001Q3 - 2019Q4 and 2021Q1 - 2024Q1).

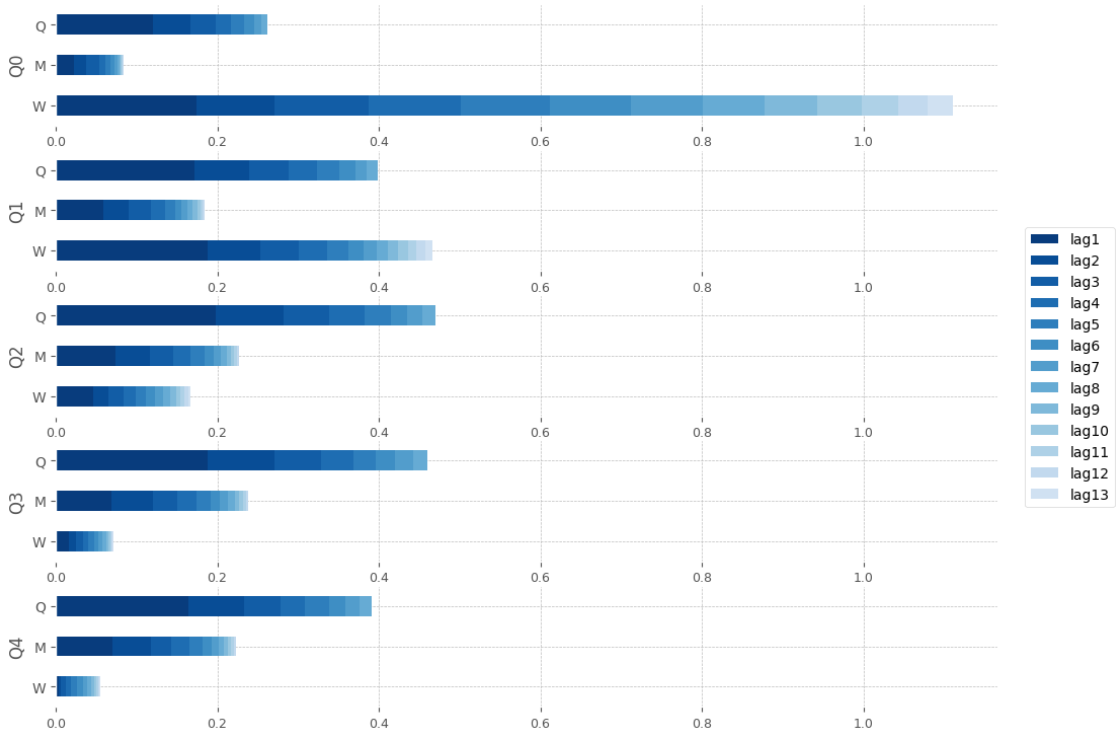


Figure 6: SHAP value by forecast horizon and data frequency, COVID period (2020Q1 - 2020Q4).

5.2 Specific Variable Groups that Dominate Predictions

At the group level, predictive importance varies systematically across data frequencies and forecast horizons. Quarterly groups are the primary contributors at longer horizons, monthly groups play an intermediate role, and weekly indicators are most influential for nowcasting and short-horizon forecasts.

Figure 7 presents group SHAP values by data frequency over the full out-of-sample period. Among quarterly groups (left panel), household and non-household balance sheets are the most important contributors. At the monthly frequency (middle panel), housing and output and income groups dominate, although their quarterly counterparts contribute less. At the weekly frequency (right panel), the ADS index is the primary driver of nowcasting with a SHAP value of 0.312, before declining to 0.154 at Q1, 0.105 at Q2, 0.077 at Q3, and 0.068 at Q4 as the forecast horizon extends. These patterns are consistent with the frequency-horizon

relationships documented in Figures 3 and 4.

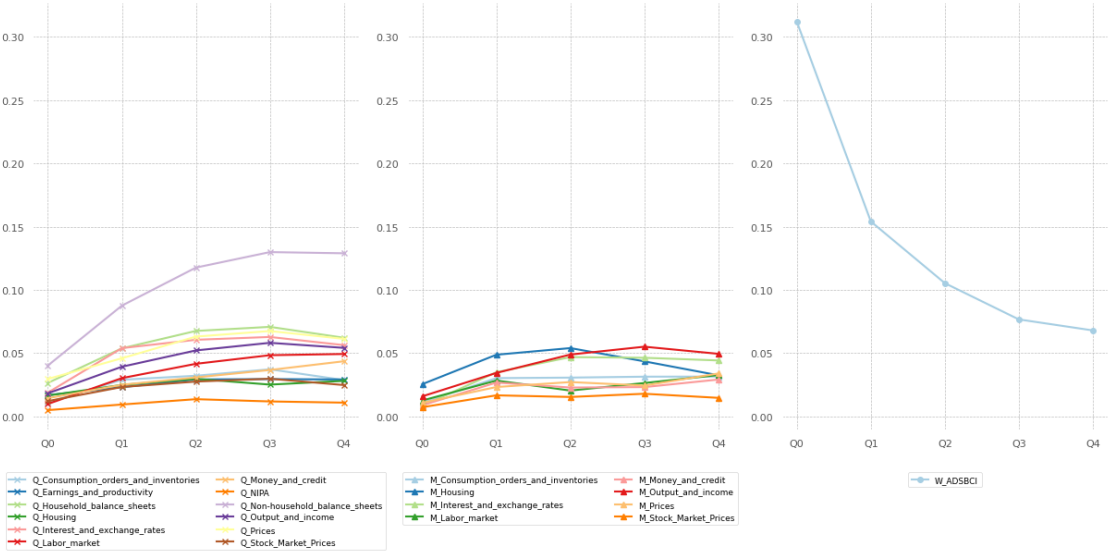


Figure 7: SHAP value by group, forecast horizon, and data frequency (left: quarterly, middle: monthly, right: weekly), full out-of-sample period (2001Q3 - 2024Q1).

These relationships remain broadly stable outside the COVID period, with the ranking of key groups largely unchanged: the ADS index continues to dominate nowcasting, while quarterly balance sheet variables remain the most important contributors at longer horizons. For example, the ADS index has a SHAP value of 0.275 for nowcasting in the pre- and post-COVID subsample, comparable to its magnitude in the full sample.

During the COVID period, however, the role of high-frequency data becomes substantially more pronounced. The ADS index reaches a SHAP value of 1.111 for nowcasting and remains dominant at 0.466 (Q1) and 0.167 (Q2) for short-horizon forecasts before declining at longer horizons. These values are notably larger than those observed outside the COVID period (Q0: 0.275, Q1: 0.140, Q2: 0.102), highlighting the importance of high-frequency indicators during periods of rapid economic change. Updated weekly, the ADS index captures abrupt shifts in economic activity more quickly than lower-frequency measures.

Figure 8 highlights differences in how similar economic concepts behave across frequencies. For nowcasting, the monthly housing group exhibits sustained variation over time, rising during the Global Financial Crisis, fluctuating afterward, and remaining elevated through

the mid-2010s. In contrast, the quarterly housing group shows a sharp increase during the crisis followed by a rapid decline. This contrast illustrates how lower-frequency measures smooth short-term fluctuations, making them less responsive to rapid economic changes and therefore less informative for near-term predictions.



Figure 8: Quarterly and monthly housing group SHAP value for nowcasting.

6 Conclusion

This paper develops a mixed-frequency LSTM framework for forecasting U.S. GDP growth in real time. The approach makes three design choices that distinguish it from much of the existing literature on machine learning for macroeconomic forecasting. First, it processes quarterly, monthly, and weekly data through separate network branches, preserving the native sampling frequency of each source rather than imposing temporal aggregation. Second, it operates entirely on real-time vintage data, ensuring that performance metrics reflect what a practitioner would actually have achieved. Third, it provides transparent model interpretation through SHAP values, addressing the "black box" criticism commonly leveled at neural network models.

The results demonstrate that the LSTM model produces forecasts competitive with the Survey of Professional Forecasters across all horizons from nowcasting through four quarters

ahead. The model is particularly effective during periods of economic turbulence, where its capacity to process rapidly changing high-frequency information translates into meaningful forecast improvements. The SHAP analysis reveals that the informational value of data is dynamic and frequency-dependent: weekly data are most valuable at short horizons and during turning points, while quarterly data dominate for longer-range prediction.

Several directions for future research emerge from this work. The framework could be extended to incorporate additional high-frequency data sources, including daily financial variables, text-based indicators, or alternative data such as satellite imagery and mobility data, which have proven informative during the COVID-19 period (Woloszko, 2020; Lewis et al., 2022). Replacing the LSTM with attention-based architectures such as Temporal Fusion Transformers (Lim et al., 2021) could improve the model's ability to handle longer sequences and provide a built-in mechanism for variable selection. Finally, developing density forecasts from the ensemble of model runs would allow for a richer characterization of forecast uncertainty, enabling direct comparison with the probabilistic forecasts produced by the SPF and the methods developed in Clark et al. (2022).

References

Almuzara, M., Baker, K., O'Keeffe, H., and Sbordone, A. M. (2023). The New York Fed staff nowcast 2.0. *Federal Reserve Bank of New York Staff Technical Paper*.

Aruoba, S. B., Diebold, F. X., and Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of Economic Forecasting*, 2:195–237.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Nowcasting. *ECB Working Paper No. 1275*.

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2025). Macroeconomic forecasting with large language models. *arXiv preprint arXiv:2407.00890*.

Clark, T. E., Ganics, G., and Mertens, E. (2022). Constructing fan charts from the ragged edge of SPF forecasts. *Federal Reserve Bank of Cleveland Working Paper No. 22-36*.

Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964.

Croushore, D. (2011). Frontiers of real-time data analysis. *Journal of Economic Literature*, 49(1):72–100.

Croushore, D. and Stark, T. (2006). Forecasting with real-time macroeconomic data. *Handbook of Economic Forecasting*, 1:961–982.

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

Diebold, F. X., Mora, J., and Shin, M. (2025). The wisdom of crowds: Cross-sectional variation in the informational content of the SPF. *arXiv preprint arXiv:2503.09287*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, pages 1–15.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3:620–631.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted mixed data sampling (UMIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A*, 178(1):57–82.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *CIRANO Working Paper*.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Hamilton, J. D., Ma, X., and Xi, J. (2024). Principal component analysis for a mix of stationary and nonstationary variables. *NBER Working Paper No. 32068*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hopp, D. (2024). Benchmarking econometric and machine learning methodologies in nowcasting GDP. *Empirical Economics*, 66:2191–2247.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456.

Jha, M., Qian, J., Weber, M., and Yang, B. (2025). Harnessing generative AI for economic insights. *arXiv preprint arXiv:2410.03897*.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Lewis, D. J., Mertens, K., and Stock, J. H. (2022). Measuring real activity using a weekly economic index. *Journal of Applied Econometrics*, 37(4):667–687.

Liebermann, J. (2014). Real-time nowcasting of GDP: A factor model vs. professional forecasters. *Oxford Bulletin of Economics and Statistics*, 76(6):783–811.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, 18(4):427–443.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

McCracken, M. W. and Ng, S. (2021). FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review*, 103(1):1–44.

Paranhos, L. (2023). Predicting inflation with recurrent neural networks. Version 2.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2013). How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.

Richardson, A., van Florenstein Mulder, T., and Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2):941–948.

Sagheer, A. and Kotb, M. (2019). Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems. *Scientific reports*, 9(1):19038.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2:307–317.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Soybilgen, B. and Yazgan, E. (2021). Nowcasting US GDP using tree-based ensemble models and dynamic factors. *Computational Economics*, 57:387–417.

Stark, T. (2010). Realistic evaluation of real-time forecasts in the Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Research Paper*.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328.

Woloszko, N. (2020). Tracking activity in real time with Google trends. *OECD Economics Department Working Paper No. 1634*.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.