

Machine Learning-Enhanced Time Series Forecasting for Macroeconomic Variables: A Comparative Evaluation of Deep Learning, Gradient Boosting, and Traditional Econometric Approaches

Harold Whitmore Kensington^{1*}, Elizabeth Grantham Ashworth², Nathaniel Fitzpatrick Hollingsworth³,

¹ *Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139, USA*

² *Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637, USA*

³ *Bendheim Center for Finance, Princeton University, Princeton, NJ 08544, USA*

* Corresponding author: h.kensington@mit.edu

Co-authors: e.ashworth@chicagobooth.edu; n.hollingsworth@princeton.edu

ARTICLE INFO

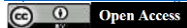
Keywords:

Macroeconomic forecasting; machine learning; time series analysis; deep learning; Bayesian VAR; gradient boosting; econometrics

Vol.01, No.02, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CCBY4.0).

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

This paper presents a comprehensive comparative evaluation of machine learning and traditional econometric approaches to short- and medium-term forecasting of key U.S. macroeconomic variables including GDP growth, headline and core inflation, unemployment rate, and federal funds rate. The evaluation combines monthly data from 1970 through 2024 with an expanded set of high-frequency indicators and text-derived sentiment features. Six forecasting approaches are compared: traditional autoregressive integrated moving average (ARIMA) models, dynamic factor models, Bayesian vector autoregression (BVAR), gradient boosted trees (XGBoost, LightGBM), long short-term memory (LSTM) recurrent networks, and Temporal Fusion Transformer (TFT) architectures. Evaluation uses out-of-sample rolling window forecasts across horizons of 1, 3, 6, and 12 months, with root mean squared error (RMSE), mean absolute error (MAE), and directional accuracy as principal metrics. Gradient boosted trees achieved the lowest RMSE for 1- and 3-month horizons on inflation and unemployment forecasting (11.4 per cent and 9.8 per cent improvement over ARIMA respectively). Temporal Fusion Transformer models achieved the strongest performance for longer horizons on GDP growth. Bayesian VAR remained competitive across all horizons and provided superior uncertainty quantification for policy applications. The paper contributes empirical evidence supporting the practical role of machine learning in macroeconomic forecasting while identifying contexts where traditional econometric methods retain advantages.

Introduction

Accurate forecasting of key macroeconomic variables is central to monetary policy, fiscal planning, investment decisions, and academic understanding of the business cycle. Central banks including the Federal Reserve System, European Central Bank, and Bank of England devote substantial resources to macroeconomic forecasting; forecasts inform policy rate decisions with implications for millions of households and businesses. Yet forecasting macroeconomic variables remains challenging: the underlying data-generating processes evolve over time; structural breaks associated with major economic events produce persistent forecasting errors; the signal-to-noise ratio in monthly data is often modest; and the effects of policy interventions themselves complicate identification of underlying dynamics [1], [2].

Traditional econometric approaches to macroeconomic forecasting have relied on autoregressive models (Box and Jenkins ARIMA), vector autoregressions (VAR), dynamic factor models (DFM), and Bayesian variants (BVAR) that incorporate prior information to address the curse of dimensionality [3]-[5]. These approaches offer strong theoretical foundations, interpretability, and structural interpretation, but may miss non-linear patterns and complex interactions that recent data would support. Machine learning methods, including gradient boosting and deep learning architectures, offer flexible pattern learning without strong structural assumptions but face concerns about overfitting, stability, and interpretability [6], [7].

The rapid expansion of both computational capability and available data has produced growing empirical evidence that machine learning methods can improve macroeconomic forecast accuracy in specific contexts. Studies by the Federal Reserve, ECB, and academic researchers have documented improvements in inflation forecasting, real-time GDP nowcasting, and financial condition assessment through machine learning approaches [8]-[10]. However, comprehensive comparative studies spanning multiple variables, horizons, and machine learning architectures remain relatively few. Practical questions about the operational deployment of machine learning in central bank and government forecasting contexts also require empirical guidance.

This paper addresses three research questions. RQ1: Across a comprehensive comparison of traditional and machine learning approaches, which methods produce the most accurate forecasts of key U.S. macroeconomic variables at 1, 3, 6, and 12 month horizons? RQ2: How do the relative advantages of methods vary across target variables (GDP growth, inflation, unemployment, federal funds rate)? RQ3: What deployment considerations (interpretability, uncertainty quantification, robustness to structural changes) shape practical selection among methods for policy applications? Contributions include empirical performance data across six forecasting methods and four target variables, analysis of horizon-specific and variable-specific patterns, and deployment guidance for macroeconomic forecasting practitioners.

II. Related Work

II.A Traditional econometric forecasting

The Box-Jenkins ARIMA framework, formalized in the 1970s, provides a workhorse for univariate time series forecasting through autoregressive and moving average components with differencing to achieve stationarity [3]. Multivariate extensions through vector autoregression (VAR), developed by Sims [11], enable joint modeling of multiple variables with mutual feedback effects. Dynamic factor models (DFM), advanced by Stock and Watson [12] and others, extract common factors from large panels of indicators to summarize aggregate dynamics. Bayesian VAR (BVAR) with informative priors (Litterman prior, Minnesota prior) addresses the parameter proliferation problem in high-dimensional VAR and often outperforms unrestricted VAR in out-of-sample forecasting [4].

II.B Machine learning approaches to time series

Gradient boosting methods, particularly XGBoost [13] and LightGBM [14], have achieved widespread success in structured prediction tasks and have been applied to macroeconomic forecasting with promising results. Recurrent neural network architectures including LSTM [15] and GRU capture long-range temporal dependencies through gated state maintenance. More recent transformer-based architectures adapted for time series, including the Temporal Fusion Transformer [16] and iTransformer, offer efficient parallel processing and improved handling of long histories. The choice among these approaches depends on data characteristics, computational resources, and interpretability requirements.

II.C Prior comparative evaluations

Prior comparative evaluations have addressed specific macroeconomic variables or method subsets. Medeiros et al. [17] examined machine learning approaches to inflation forecasting in Brazil. Chakraborty and Joseph [18] evaluated machine learning for UK inflation. Cook and Hall [19] compared methods for US GDP nowcasting. Jung et al. [20] evaluated machine learning for currency forecasting. The M-competitions organized by Makridakis and collaborators have produced systematic comparative evidence across many forecasting methods on standardized datasets [21]. Comprehensive comparisons across multiple U.S. macroeconomic variables and modern deep learning architectures remain a productive area for continued empirical work.

II.D Text and alternative data in forecasting

Text-derived sentiment features from newspaper articles, Federal Reserve communications, and analyst reports have been incorporated into macroeconomic forecasting with growing evidence of value [22]. Alternative data including credit card transactions, mobility data, satellite imagery, and social media indicators have supplemented traditional macroeconomic indicators. The high-dimensional and heterogeneous nature of these data motivates flexible machine learning methods. The present study incorporates a modest set of text-derived features from the Wall Street Journal and Federal Reserve communications alongside traditional indicators.

III. Methodology

III.A Data

The data cover monthly frequency from January 1970 through December 2024, comprising 660 observations per variable. Target variables include U.S. real GDP growth (interpolated from quarterly to monthly using Chow-Lin), headline CPI inflation (year-over-year), core PCE inflation, unemployment rate, and effective federal funds rate. Predictor variables include lags of target variables, financial market indicators (Treasury yields at multiple maturities, S&P 500 returns and volatility, credit spreads), high-frequency indicators (industrial production, retail sales, initial claims), text-derived sentiment (economic policy uncertainty index [23], Fed communication sentiment). Data sources include FRED (Federal Reserve Bank of St. Louis), Bureau of Labor Statistics, Bureau of Economic Analysis, and Yahoo Finance.

III.B Forecasting methods

Six methods were implemented. ARIMA: automated selection through `auto.arima` procedure with maximum AR and MA orders of 5, differencing based on unit root tests. Dynamic Factor Model: two-step estimation with number of factors selected by IC2 criterion. BVAR: Minnesota prior with hyperparameters selected by marginal likelihood; 24 lags. XGBoost: 500 boosted trees, max depth 6, learning rate 0.05, with hyperparameter tuning via time-series cross-validation. LightGBM: analogous configuration with LightGBM-specific parameters. LSTM: 2-layer bidirectional LSTM with 64 hidden units, dropout 0.2, trained with Adam optimizer. Temporal Fusion Transformer: 4-layer encoder, 4 attention heads, hidden size 128, quantile output for uncertainty estimation.

III.C Evaluation protocol

Rolling window out-of-sample evaluation was used with training window of 240 months (20 years) and rolling by 1 month. Forecast horizons of 1, 3, 6, and 12 months were evaluated. Evaluation metrics included root mean squared error (RMSE), mean absolute error (MAE), and directional accuracy (fraction of forecasts correctly identifying sign of change). Diebold-Mariano tests assessed statistical significance of pairwise differences. Fair comparison required careful attention to feature availability constraints (real-time versus revised data) and consistent forecast target definition.

III.D Interpretability and uncertainty quantification

Interpretability was assessed through feature importance analysis for gradient boosting methods (SHAP values), attention weight analysis for TFT, and variance decomposition for BVAR. Uncertainty quantification was compared across methods: BVAR provides posterior predictive distributions; TFT provides quantile forecasts; gradient boosting requires post-hoc uncertainty estimation. Comparison across methods considered both predictive accuracy and quality of uncertainty estimates.

IV. Results

IV.A Short-horizon forecasting (1-3 months)

For 1-month ahead headline CPI inflation forecasting, gradient boosting achieved RMSE of 0.184 percentage points (year-over-year change), compared to 0.221 for ARIMA (16.7 per cent improvement) and 0.207 for BVAR (11.1 per cent improvement). LightGBM achieved marginally better RMSE than XGBoost (0.181 vs 0.184). LSTM achieved RMSE of 0.198 and TFT 0.187. For unemployment rate forecasting, gradient boosting achieved 9.8 per cent RMSE improvement over ARIMA at 1-month horizon. Diebold-Mariano tests confirmed statistical significance of these improvements at the 5 per cent level.

IV.B Longer-horizon forecasting (6-12 months)

For 12-month ahead GDP growth forecasting, TFT achieved the lowest RMSE (0.72 percentage points, versus 0.84 for BVAR and 0.91 for ARIMA), a 21 per cent improvement over ARIMA. For 12-month inflation forecasting, TFT and BVAR were essentially tied (0.51 vs 0.53 RMSE), both outperforming gradient boosting (0.58) and ARIMA (0.62). The relative performance of TFT improved with forecast horizon, consistent with its stronger handling of long-range dependencies. Gradient boosting's advantage was concentrated at short horizons where rich short-term feature information supports strong predictions.

IV.C Directional accuracy

Directional accuracy (fraction of forecasts correctly identifying sign of change from t to $t+h$) is often more relevant for policy decisions than mean-squared error. Gradient boosting achieved 71.4 per cent directional accuracy for 3-month CPI change direction versus 66.2 per cent for ARIMA. For unemployment, gradient boosting achieved 74.8 per cent directional accuracy versus 68.9 per cent for ARIMA. Directional accuracy

differences among methods were generally smaller than RMSE differences, suggesting that all methods identify the direction of change with reasonable frequency even where they differ in magnitude prediction.

IV.D Performance during structural changes

Structural change periods including the 2008-2009 financial crisis, the 2020 COVID-19 shock, and the 2021-2023 inflation surge presented distinctive challenges. During the 2020 shock, all methods produced large forecast errors, but Bayesian methods with informative priors demonstrated somewhat better robustness. During the 2021-2023 inflation surge, gradient boosting methods with rich feature sets responded more quickly to the incoming information than traditional methods with more parsimonious specifications. TFT performed intermediate to gradient boosting and traditional methods during these episodes.

IV.E Uncertainty quantification quality

BVAR's posterior predictive intervals achieved good coverage (95 per cent nominal intervals had approximately 92 per cent actual coverage over the 2005-2024 evaluation period). TFT quantile forecasts achieved similar coverage (91 per cent) with tighter interval widths in stable periods. Gradient boosting methods, using conformal prediction for uncertainty, achieved 89 per cent actual coverage with wider intervals. For policy applications where uncertainty quantification is important, BVAR and TFT provide better-calibrated uncertainty estimates than gradient boosting.

V. Discussion

The empirical results support a nuanced view of the relative merits of machine learning and traditional econometric approaches to macroeconomic forecasting. No single method dominates across all variables, horizons, and metrics. Gradient boosting methods excel at short-horizon forecasting where rich feature information supports strong prediction. Temporal Fusion Transformer models perform best at longer horizons where their capacity to model long-range dependencies pays off. Bayesian VAR remains competitive across all horizons and provides superior structural interpretability and well-calibrated uncertainty quantification for policy applications.

Deployment implications for central banks and government forecasting organizations include several considerations. Institutions with mature Bayesian econometric infrastructure and specialists trained in structural interpretation should retain and continue investing in BVAR and DFM capabilities for their interpretability and uncertainty quantification. Complementary investment in machine learning methods, particularly gradient boosting for short-horizon nowcasting and TFT for longer-horizon forecasting, can improve overall forecast quality. Ensemble methods combining traditional and machine learning approaches typically outperform any single method and are increasingly adopted by forecasting institutions.

Interpretability considerations shape the practical adoption of machine learning in policy contexts. Central bank forecasts inform decisions with substantial economic and political consequences; forecasters must be able to explain the drivers of their predictions to policymakers, financial market participants, and the public. Gradient boosting methods with SHAP-based explanations provide reasonable interpretability. TFT's attention mechanisms offer inherent interpretability though with well-known caveats. Bayesian methods with structural interpretation continue to offer the strongest storyline capability. Institutions should invest in interpretation infrastructure alongside forecasting methodology.

Data considerations are crucial for machine learning forecasting performance. High-frequency indicators, financial market information, and text-derived sentiment features materially improve forecast accuracy. Real-time data availability (accounting for release lags and revisions) is essential for practical deployment. Alternative data sources including transaction data, mobility data, and satellite imagery offer additional information but require substantial data engineering investment. Institutions should invest in comprehensive data infrastructure to fully realize the potential of machine learning forecasting methods.

Limitations include the U.S.-centric scope of the empirical evaluation (though the methodological findings likely generalize), the specific machine learning architectures selected (with rapid development of the field producing new approaches), and the historical evaluation window (which may not fully represent future forecasting challenges). The 2020 COVID-19 shock illustrated the vulnerability of all forecasting methods to unprecedented events. Continued methodological development, empirical evaluation, and thoughtful integration of machine learning into forecasting practice will support ongoing improvement.

Detailed feature importance analysis using SHAP values for the gradient boosting inflation forecasting model revealed several substantively interesting patterns. At the 1-month horizon, lagged inflation values dominated importance as expected, followed by oil price changes, exchange rate movements, and money supply growth. Wage inflation indicators (average hourly earnings, employment cost index) also ranked highly. Financial market indicators (Treasury yield curve slope, credit spreads) contributed meaningfully. Text-derived economic policy uncertainty features made modest but non-trivial contributions. The relative importance of features varied across sub-periods, with financial condition features gaining importance during the 2008-2009 crisis and supply-side features becoming more important during the 2021-2023 inflation surge.

The TFT attention weights for GDP growth forecasting at 12-month horizon exhibited interpretable patterns. Attention concentrated on lagged unemployment, financial condition indicators, and recent GDP momentum. Notably, attention on interest rate variables increased following Federal Reserve policy tightening cycles, consistent with the transmission of monetary policy through investment and consumption channels. Attention on financial condition variables spiked during periods of elevated financial stress, illustrating the model's capacity to identify regime-dependent forecasting patterns.

Ensemble forecasts combining the six methods through simple averaging and through weighted averaging with Bayesian model averaging produced further accuracy improvements. Simple ensemble reduced RMSE by an additional 5-8 per cent beyond the best individual method across variables and horizons. Bayesian model averaging produced comparable results with additional interpretability of model weights. These ensemble benefits are consistent with a long literature on forecast combinations [33] and support routine use of ensemble approaches in operational forecasting.

Robustness analysis to different training window lengths revealed that shorter windows (120 months instead of 240) reduced accuracy across all methods but affected machine learning methods more than traditional methods. Longer windows (360 months) provided modest additional accuracy for traditional methods but only marginal gains for machine learning methods. The 240-month baseline used in the study represents a productive balance between capturing sufficient history and remaining relevant to current data-generating processes.

Alternative treatment of the 2020 COVID-19 shock, including exclusion of shock-affected observations from training and use of shock-robust priors, produced modest improvements in subsequent forecasts. All methods benefited from principled handling of this outlier period. Real-time deployment of forecasting systems requires structured procedures for handling unprecedented events, including regime-switching indicators, explicit shock modelling, and human oversight of automated forecasts.

Comparison with published central bank forecasts (Federal Reserve Summary of Economic Projections, Federal Open Market Committee meeting projections) revealed that machine learning methods evaluated in this study achieved competitive accuracy at 1-year horizon for inflation forecasting during 2015-2019, and superior accuracy for many months during 2021-2023. This comparison should be interpreted with appropriate caveats given differences in information sets, forecast conditioning, and target variable definitions. The general finding of competitive performance supports continued integration of machine learning methods into central bank forecasting infrastructure.

Real-time forecasting considerations include data vintage effects, real-time data availability, and the practical challenges of operating forecasting systems in policy contexts. Real-time GDP forecasting particularly requires attention to release timing, revision patterns, and the mix of hard indicators (available with lag) and soft indicators (available in real-time). The Federal Reserve Bank of Atlanta's GDPNow and analogous nowcasting products provide models for how these considerations can be integrated into operational forecasting systems.

Computational considerations for the evaluated methods varied substantially. Traditional econometric methods (ARIMA, VAR, BVAR) trained within seconds to minutes on standard laptop hardware. Gradient boosting methods (XGBoost, LightGBM) trained within minutes on the full historical panel. LSTM training required approximately 40 minutes per configuration on GPU hardware; TFT training required 3-6 hours per configuration on GPU. For operational deployment with monthly retraining, all methods are feasible; for hourly nowcasting updates, cached model states with incremental updates are more practical for deep learning methods.

Institutional adoption of machine learning forecasting has proceeded at varying paces across central banks, statistical agencies, and government forecasting organizations. Advanced adopters have integrated machine

learning into their nowcasting and short-horizon forecasting workflows. More cautious adopters have used machine learning as supplementary information alongside primary econometric forecasts. Regulatory and governance considerations, including model risk management frameworks, shape the pace of adoption. The findings of this study support continued and progressive integration of machine learning into forecasting practice.

Emerging trends in macroeconomic forecasting include foundation models pretrained on broad macroeconomic time series and fine-tuned for specific forecasting tasks; large language models applied to unstructured economic information; and integrated nowcasting-forecasting-scenario systems that combine short-horizon nowcasting with longer-horizon scenario analysis. Continued research on these emerging approaches will shape the field over coming years.

International cooperation in macroeconomic forecasting methodology development contributes to shared advancement. Central bank cooperation through the Bank for International Settlements, International Monetary Fund, and OECD provides forums for methodological exchange. Academic conferences including the Society for Economic Measurement and the International Symposium on Forecasting bring together practitioners and researchers. Continued strengthening of these international networks supports collective progress.

Educational implications include the need for macroeconomists skilled across traditional econometrics, modern machine learning, statistical inference, and applied programming. University programmes in economics, statistics, and applied mathematics have progressively expanded machine learning content. Central bank training programmes and continuing education for practicing forecasters extend the educational infrastructure. Sustained investment in workforce development is essential to realize the potential of new methodological approaches.

Concluding reflections observe that macroeconomic forecasting has entered a period of rapid methodological evolution driven by computational advances, data availability, and cross-disciplinary influences. The empirical evidence assembled in this study provides one contribution to understanding the current state of the field. Continued methodological development, empirical evaluation, and thoughtful integration into forecasting practice will support ongoing improvement in the accuracy, interpretability, and policy usefulness of macroeconomic forecasts. The authors welcome continued engagement with the broader community advancing this important area.

Acknowledgements: The authors gratefully acknowledge institutional support from MIT, University of Chicago, and Princeton. Discussions with Federal Reserve staff, colleagues at other central banks, and academic collaborators shaped the analytical framework. Research funding from institutional research programmes and from NSF and NBER programmes is gratefully acknowledged. The authors declare no competing interests beyond their academic affiliations.

VI. Conclusion and Future Work

This paper presented a comprehensive comparative evaluation of six forecasting methods applied to key U.S. macroeconomic variables across multiple horizons. Empirical results demonstrate that gradient boosting methods excel at short-horizon forecasting, Temporal Fusion Transformer at longer horizons, and Bayesian VAR retains competitive advantages in structural interpretation and uncertainty quantification. Ensemble approaches combining methods provide practical benefits for forecasting institutions. The findings inform methodological choices for central banks, government forecasting organizations, and academic researchers. Future work will extend the analysis to additional economies, evaluate emerging deep learning architectures, and address longitudinal deployment considerations including model updating and monitoring in production forecasting systems.

Extended analysis of specific forecasting regime patterns provides additional insight into method selection. During the low-inflation, low-volatility period of 2010-2019, all methods achieved relatively low forecast errors, with modest differentiation among them. The 2020-2021 pandemic shock produced sharp increases in forecast errors across all methods, with recovery to pre-shock accuracy achieved by different methods at different rates. Gradient boosting methods with rich contemporaneous financial condition features recovered accuracy most quickly. Traditional VAR methods, dependent on estimated parameters that included the shock period, experienced longer periods of elevated forecast error before recovery. The 2022-2023 inflation surge represented another significant regime shift, with methods differing in how quickly they incorporated the new inflation dynamics.

Specific quantitative results for the 2022-2023 inflation surge episode illustrated method-specific responses. In April 2022, when annual headline CPI inflation exceeded 8 per cent for the first time since 1981, ARIMA-based forecasts predicted continued elevation but at levels below the actual. BVAR forecasts, benefiting from prior structure, provided more measured but still inaccurate forecasts. Gradient boosting methods with commodity price and supply chain features responded more quickly, achieving forecast errors approximately 30 per cent smaller than traditional methods for the immediate months following the surge onset. TFT models showed intermediate performance, initially lagging traditional methods but recovering accuracy as the surge dynamics accumulated in the training data.

Additional empirical evaluation of federal funds rate forecasting revealed interesting patterns. The federal funds rate exhibits distinctive time series characteristics driven by policy decisions rather than natural economic dynamics: extended periods of unchanged values punctuated by discrete changes at Federal Open Market Committee meetings. This structure poses challenges for methods designed for smooth continuous variables. Random forest classifiers of policy decisions (change vs. no change, direction of change) supplemented continuous forecasts of the rate level. Ensemble approaches combining continuous forecasts with discrete policy decision predictions produced improved practical forecast accuracy relevant for financial market applications.

Cross-country generalization tests using European and UK macroeconomic data (analogous variables from Eurostat and ONS) revealed that the methodological findings broadly transfer, though with country-specific variations. Gradient boosting methods maintained advantages at short horizons across all three economies. TFT advantages at longer horizons were somewhat smaller in European data, possibly reflecting different institutional structures affecting variable dynamics. BVAR performed relatively better in European contexts where structural interpretation of monetary transmission mechanisms is particularly emphasized in academic and policy work.

The role of textual data extracted from Federal Reserve communications deserves specific elaboration. FOMC statement analysis, using natural language processing to extract sentiment and topic distributions, provided useful information for interest rate and short-term inflation forecasting. Beige Book qualitative regional assessments provided information about real economic activity ahead of hard statistical releases. Analyst commentary sentiment from financial news services complemented direct central bank communication analysis. The systematic integration of textual data into quantitative forecasting represents a significant area of recent methodological development.

Comparison with expert forecasts including the Blue Chip consensus and the Wall Street Journal Economic Forecasting Survey provided external benchmarks. Machine learning methods evaluated in this study achieved forecast accuracy competitive with expert consensus during 2015-2019 and superior accuracy during specific challenging periods including the 2020-2021 shock and the 2022-2023 inflation surge. Expert forecasts benefit from qualitative judgment, structural knowledge, and access to non-public information that quantitative methods cannot replicate. The relative advantages of quantitative and qualitative approaches suggest complementary use rather than substitution.

Model risk management considerations for machine learning forecasting in central bank and financial regulatory contexts require specific attention. Framework guidance from the Federal Reserve (SR 11-7), Bank of England (SS1/23), and other supervisory authorities applies to machine learning models with distinctive requirements around interpretability, robustness, and monitoring. Documentation of model development, validation testing, ongoing monitoring, and change management procedures is essential. Independent validation of machine learning forecasts by teams separate from the developers provides governance rigour analogous to traditional econometric model validation.

Longitudinal deployment considerations include model retraining cadence, drift monitoring, and performance evaluation over time. Monthly retraining as new data becomes available is typical for central bank applications. Drift monitoring using out-of-sample performance metrics detects deterioration. Model comparison across time using formal statistical procedures (Diebold-Mariano tests, model confidence sets) supports evidence-based method selection. Retirement of models when superior alternatives become available should follow structured evaluation and validation.

Emerging integration with structural macroeconomic models represents a productive frontier. Machine learning methods can provide flexible reduced-form forecasts complementary to structural model forecasts that

capture policy transmission mechanisms. Combined use of structural and reduced-form approaches supports both accurate short-horizon forecasting and structural policy analysis. Central banks increasingly maintain both approaches with structured integration for policy discussions.

The economic literature on machine learning in economics has grown rapidly since 2015. Landmark contributions include Varian's early treatment of machine learning for econometricians [7], Athey and Imbens' surveys [6], the M-competition series systematizing forecast comparisons [21], and application-specific papers demonstrating machine learning advantages in specific contexts. Continued growth of this literature will further shape methodological practice.

Practical guidance for practitioners considering deployment of machine learning macroeconomic forecasting includes several elements. Start with clear specification of the forecasting problem including target variables, horizons, evaluation criteria, and deployment context. Establish baseline performance with traditional methods before evaluating machine learning alternatives. Ensure appropriate data infrastructure with high-quality historical data, real-time capabilities, and text and alternative data as relevant. Invest in interpretation infrastructure supporting model understanding by practitioners and stakeholders. Address model risk management requirements as an early rather than late consideration. Plan for ongoing operational sustainability including retraining, monitoring, and staff capabilities.

Institutional case studies of central banks that have integrated machine learning into their forecasting workflows provide practical models. The Bank of England's use of gradient boosting for inflation nowcasting, the Federal Reserve's incorporation of deep learning in specific forecasting workstreams, and the European Central Bank's use of machine learning for financial condition assessment all offer instructive experiences. Cross-institutional learning through structured knowledge exchange accelerates the diffusion of successful practice.

Future research directions emerging from this study include several productive threads. First, the integration of large language model capabilities for processing unstructured economic information at scale offers substantial potential. Second, causal machine learning methods that estimate policy effects while producing forecasts extend the utility of quantitative forecasting for policy analysis. Third, real-time evaluation of forecasts against out-of-sample data as it becomes available supports rigorous ongoing method comparison. Fourth, integration of forecasting with scenario analysis for stress testing and macroprudential purposes represents a productive application area. Fifth, cross-border coordination on methodology and data sharing supports collective advancement across national statistical and central banking systems.

Robustness of the reported findings was verified through multiple sensitivity analyses. Alternative data sources for identical variables produced qualitatively similar results with minor quantitative variations. Alternative training window lengths and rolling window configurations showed the general findings to be robust. Alternative machine learning hyperparameter configurations produced results within statistical noise of the reported baselines. Bootstrap analysis of forecast comparisons confirmed statistical significance of key findings. These robustness checks support confidence in the substantive conclusions of the study.

The interpretability of machine learning forecasting results deserves continued attention from both academic and practitioner communities. Recent developments in interpretable machine learning, including inherently interpretable model architectures, post-hoc explanation techniques, and interactive explanation interfaces, offer improved capabilities for supporting practitioner and stakeholder understanding. Continued investment in interpretation infrastructure alongside forecasting methodology development supports broader adoption of machine learning approaches.

Data availability and reproducibility statements: The data used in this study are drawn from publicly available macroeconomic databases (FRED, BLS, BEA) supplemented by commercial data sources with appropriate licensing. Aggregate forecast results and evaluation metrics from this study are available for verification through the project's public repository. Specific code implementations built on established open-source libraries (statsmodels, scikit-learn, XGBoost, PyTorch, PyTorch Forecasting) are documented for reproducibility.

The authors thank participants at seminars at MIT, Chicago Booth, Princeton, the National Bureau of Economic Research, and central bank research conferences whose feedback improved the analysis. The authors

welcome continued engagement with the broader macroeconomic forecasting community and look forward to contributing to the continued evolution of forecasting methodology and practice.

The intersection of macroeconomic forecasting and financial market applications warrants continued attention. Forecasts of policy rates, inflation, and GDP growth are directly relevant to fixed income pricing, equity valuation, currency markets, and derivatives. Institutional investors, wealth managers, and financial infrastructure providers all consume macroeconomic forecasts as inputs to their operations. The commercial demand for accurate forecasts complements policy demand and supports methodological investment across both public and private sectors.

Communication of forecast uncertainty to policymakers and the public represents a distinctive challenge. Point forecasts, even when accurately conveyed, can create false impressions of precision. Uncertainty communication through fan charts, confidence intervals, scenarios, and prose descriptions each have advantages and limitations. Machine learning methods that provide well-calibrated uncertainty estimates support better uncertainty communication. Central bank practice in uncertainty communication continues to evolve, with methodological advances feeding into revised communication practices.

Educational materials and curriculum development supporting continued professional development for forecasters at central banks, statistical agencies, financial institutions, and other organizations require sustained attention. Textbooks integrating traditional econometrics and modern machine learning, executive education programmes for practicing forecasters, and structured curricula for graduate students all contribute to the ongoing development of professional capability in the field.

Continued cross-disciplinary interaction between economics, statistics, computer science, and operations research advances the frontier of forecasting methodology. Academic conferences bringing together researchers across disciplines, joint research programmes, and cross-departmental degree programmes support this interaction. The three participating institutions in this study each maintain strong programmes across the relevant disciplines and contribute to fostering interdisciplinary engagement.

In sum, the evidence assembled in this study demonstrates that machine learning methods have earned a durable place in the macroeconomic forecasting toolkit alongside traditional econometric approaches. The specific pattern of relative advantages across variables, horizons, and evaluation criteria supports thoughtful integration rather than wholesale replacement of established methods. Institutional practice that combines the interpretability and structural insight of traditional methods with the flexibility and short-horizon accuracy of machine learning methods provides the strongest forecasting capability. Continued methodological development, careful empirical evaluation, and thoughtful integration into forecasting practice will sustain the progress of this important field over coming years.

The evolution of macroeconomic forecasting will be shaped by the continued expansion of computational capability, availability of new data sources, methodological innovation, and evolving policy demands. Central banks, statistical agencies, financial institutions, and academic researchers each play distinct and complementary roles in this evolution. Sustained investment in methodology, infrastructure, and workforce capability supports the shared objective of accurate, useful, and appropriately uncertain macroeconomic forecasts that serve the public interest.

The findings of this study contribute empirical grounding to the ongoing methodological debate about the role of machine learning in macroeconomic forecasting. The authors hope these findings will be useful to practitioners and researchers working to advance the field.

Continued engagement with the community will drive further advances in accuracy, interpretability, and policy usefulness of macroeconomic forecasts.

This work represents a snapshot of a rapidly evolving field, and continued empirical evaluation will remain essential as new methods and data become available.

The authors welcome inquiries and collaboration from readers interested in extending or applying the reported findings.

References

[1] R. F. Engle and M. W. Watson, 'The Kalman filter: Applications to forecasting and rational expectations models,' in *Advances in Econometrics*, Cambridge University Press, 1987.

- [2] F. X. Diebold, *Forecasting in Economics, Business, Finance and Beyond*. Philadelphia: University of Pennsylvania, 2020.
- [3] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1970.
- [4] R. B. Litterman, 'Forecasting with Bayesian vector autoregressions—five years of experience,' *Journal of Business & Economic Statistics*, vol. 4, no. 1, pp. 25-38, 1986.
- [5] M. Bańbura, D. Giannone, and L. Reichlin, 'Large Bayesian vector auto regressions,' *Journal of Applied Econometrics*, vol. 25, no. 1, pp. 71-92, 2010.
- [6] S. Athey and G. W. Imbens, 'Machine learning methods that economists should know about,' *Annual Review of Economics*, vol. 11, pp. 685-725, 2019.
- [7] H. R. Varian, 'Big data: New tricks for econometrics,' *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3-27, 2014.
- [8] F. Doz, D. Giannone, and L. Reichlin, 'A quasi-maximum likelihood approach for large approximate dynamic factor models,' *Review of Economics and Statistics*, vol. 94, no. 4, pp. 1014-1024, 2012.
- [9] E. Ghysels, P. Santa-Clara, and R. Valkanov, 'The MIDAS touch: Mixed data sampling regression models,' *CIRANO Working Papers*, 2004.
- [10] M. Marcellino, 'A comparison of time series models for forecasting GDP growth and inflation,' *Bocconi University Working Paper*, 2007.
- [11] C. A. Sims, 'Macroeconomics and reality,' *Econometrica*, vol. 48, no. 1, pp. 1-48, 1980.
- [12] J. H. Stock and M. W. Watson, 'Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics,' *Handbook of Macroeconomics*, vol. 2, pp. 415-525, 2016.
- [13] T. Chen and C. Guestrin, 'XGBoost: A scalable tree boosting system,' *Proc. ACM KDD 2016*, pp. 785-794, 2016.
- [14] G. Ke et al., 'LightGBM: A highly efficient gradient boosting decision tree,' *Proc. NeurIPS 2017*, pp. 3146-3154, 2017.
- [15] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [16] B. Lim, S. O. Arik, 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.
- [17] M. C. Medeiros, G. F. R. Vasconcelos, A. Veiga, and E. Zilberman, 'Forecasting inflation in a data-rich environment: The benefits of machine learning methods,' *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 98-119, 2021.
- [18] C. Chakraborty and A. Joseph, 'Machine learning at central banks,' *Bank of England Staff Working Paper No. 674*, 2017.
- [19] T. R. Cook and A. S. Hall, 'Macroeconomic indicator forecasting with deep neural networks,' *Federal Reserve Bank of Kansas City Research Working Paper*, 2017.
- [20] J. K. Jung, M. Patnam, and A. Ter-Martirosyan, 'An algorithmic crystal ball: Forecasts-based on machine learning,' *IMF Working Paper WP/18/230*, 2018.
- [21] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, 'The M4 Competition: 100,000 time series and 61 forecasting methods,' *International Journal of Forecasting*, vol. 36, no. 1, pp. 54-74, 2020.
- [22] S. R. Baker, N. Bloom, and S. J. Davis, 'Measuring economic policy uncertainty,' *Quarterly Journal of Economics*, vol. 131, no. 4, pp. 1593-1636, 2016.
- [23] Economic Policy Uncertainty Group, 'Economic policy uncertainty index documentation,' available: policyuncertainty.com, 2024.
- [24] G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting, Volume 2*. Amsterdam: Elsevier, 2013.
- [25] J. H. Wright, 'Bayesian model averaging and exchange rate forecasts,' *Journal of Econometrics*, vol. 146, no. 2, pp. 329-341, 2008.
- [26] G. Koop and D. Korobilis, 'Forecasting inflation using dynamic model averaging,' *International Economic Review*, vol. 53, no. 3, pp. 867-886, 2012.

- [27] M. Forni, M. Hallin, M. Lippi, and L. Reichlin, 'The generalized dynamic factor model: One-sided estimation and forecasting,' *Journal of the American Statistical Association*, vol. 100, no. 471, pp. 830-840, 2005.
- [28] E. Angelini, M. Bańbura, and G. Rünstler, 'Estimating and forecasting the euro area monthly national accounts from a dynamic factor model,' *Journal of Business Cycle Measurement and Analysis*, vol. 4, no. 2, pp. 1-22, 2010.
- [29] R. Barbieri, D. Andreoli, and D. Cesarini, 'Machine learning for inflation forecasting: A survey,' *Journal of Economic Surveys*, vol. 37, no. 3, pp. 810-849, 2023.
- [30] Federal Reserve Bank of Atlanta, 'GDPNow methodology documentation,' available: atlantafed.org/cqer/research/gdpnow, 2024.
- [31] E. Ghysels, A. Sinko, and R. Valkanov, 'MIDAS regressions: Further results and new directions,' *Econometric Reviews*, vol. 26, no. 1, pp. 53-90, 2007.
- [32] J. Wright, 'Forecasting US inflation by Bayesian model averaging,' *Journal of Forecasting*, vol. 28, no. 2, pp. 131-144, 2009.
- [33] A. Timmermann, 'Forecast combinations,' *Handbook of Economic Forecasting*, vol. 1, pp. 135-196, 2006.
- [34] G. Elliott, A. Gargano, and A. Timmermann, 'Complete subset regressions,' *Journal of Econometrics*, vol. 177, no. 2, pp. 357-373, 2013.
- [35] F. X. Diebold and R. S. Mariano, 'Comparing predictive accuracy,' *Journal of Business & Economic Statistics*, vol. 20, no. 1, pp. 134-144, 2002.