Machine Learning Applications in Economic Growth Modeling: A Systematic Literature Review

Mutia Nur Estri[1], Lilis Siti Badriah[1], Supriyanto[2]
[1]Faculty of Economics and Business, Jenderal Soedirman University, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Jenderal Soedirman University, Indonesia
Corresponding email: nur.estri@mhs.unsoed.ac.id

Abstract

This systematic literature review evaluates the application of machine learning models in economic growth forecasting and analysis across the selected primary studies. This paper synthesises empirical consensus on study designs, target variables, predictive performance against traditional econometric baselines, and critical gaps in the existing literature. Methodologically, we analyse the extracted literature covering both single-country and cross-country settings, evaluating inputs ranging from standard macroeconomic series to alternative high-frequency datasets. The results indicate that machine learning algorithms consistently outperform or match standard linear models in short-term Gross Domestic Product (GDP) forecasting and nowcasting. However, the magnitude of these accuracy gains varies across forecasting horizons, geographic contexts, and underlying data regimes. Significant gaps remain due to an overemphasis on short-term headline GDP relative to broader metrics such as total factor productivity, limited model explainability, and continued reliance on final-vintage data. We conclude by recommending actionable avenues for future empirical research focused on expanded target metrics, emerging economy panels, and interpretable model designs.

Keywords: economic growth; forecast accuracy; machine learning; nowcasting; systematic literature review

I. INTRODUCTION

Gross domestic product remains a central measure of aggregate economic activity. Policymakers, firms, and researchers therefore need estimates of recent output before official statistics are released, as well as forecasts of future output. The two tasks differ. A nowcast estimates the current or most recent period using information arriving at different dates; a forecast predicts a later period from information already available. Both can use machine learning, but they require different information sets and should be judged at comparable horizons. Confusing these tasks can make a reported accuracy gain appear more general than the experiment permits (Richardson et al., 2021; Kant et al., 2025).

The growing supply of macroeconomic, financial, and online data has made flexible prediction methods attractive. Regularized regression, random forests, gradient boosting, support vector regression, and neural networks can incorporate many predictors or nonlinear associations. Yet macroeconomic observations are relatively few compared with many other ML applications, and the relationships between indicators and GDP can change after crises or changes in policy. Model complexity therefore creates opportunities to capture useful signals as well as risks of fitting historical noise. Strong claims about accuracy require out-of-sample comparison with plausible alternatives, not merely good fit to training data (Chu & Qureshi, 2023; Maccarrone et al., 2021).

An additional conceptual problem concerns the meaning of "economic growth modeling." Most reviewed papers predict quarterly GDP growth, monthly activity, or a forthcoming official release. They do not estimate the causal effects of education, technological progress, or institutions on growth. A predictor can improve a forecast because it moves with GDP, while its coefficient or variable-importance score need not identify an economic mechanism. Cross-country predictions over several years address a different

horizon, but predictive performance in that setting also does not establish causal determinants. Clear terminology helps keep the contribution of ML both useful and appropriately bounded (Mar Oo et al., 2025; Chen et al., 2025).

The purpose of this article is to evaluate how the selected empirical studies define their targets, organize the available predictors, compare algorithms, and assess performance. Its contribution is a cross-study framework organized around information timing, target and horizon, baseline quality, and validation design. This framework helps explain why apparently contradictory findings can coexist. It also identifies research designs that would make future findings more transferable across countries and periods. The article is a structured critical review of an identified corpus, rather than a systematic review claiming exhaustive retrieval of every eligible study (Richardson et al., 2021; Chu & Qureshi, 2023).

Three questions guide the discussion. First, which GDP-related outcomes and data environments receive the most attention? Second, when do flexible models improve upon parsimonious statistical benchmarks, and when are those benchmarks competitive? Third, how should future research test robustness and interpretability without equating prediction with causal explanation? These questions are particularly relevant where short samples, revised official data, and changing economic conditions can strongly influence the result of a model comparison (Muchisha et al., 2021; Kant et al., 2025; Qureshi et al., 2026).

II. METHOD

2.1 Scope and analytical method

The review corpus was restricted to empirical journal articles that report a GDP or GDP-growth prediction application, provide a resolvable DOI, and contain sufficient methodological information to discuss the target, data, model, or evaluation. The resulting corpus contains 14 journal articles. Conference proceedings, a bulletin chapter, and a journal item lacking a verified DOI were excluded to keep the bibliography consistent with the requested reference conditions. The final set is purposive: there is no claim that all relevant databases were searched or that article screening followed a registered protocol.

The 14 articles cover diverse settings rather than a statistically representative sample of countries. They include single-country and multi-country exercises, high-frequency nowcasting, conventional quarterly forecasting, per capita GDP prediction, and multi-year growth projection. This diversity is useful for comparing research choices but limits any pooled claim that one model class has a stable average advantage. Reported error metrics, evaluation windows, definitions of GDP growth, and information sets differ. Consequently, this article synthesizes patterns qualitatively and does not combine published error values in a meta-analysis (Chu & Qureshi, 2023; Mar Oo et al., 2025).

The extraction framework records six features for each study: geographic scope; target and forecast horizon; predictors and their publication timing; algorithm and tuning strategy; comparator model; and out-of-sample evaluation. When described by the article, particular attention is paid to whether the study uses revised or real-time data vintages and whether parameters and feature selection are re-estimated within each training window. These details matter because a model that uses later revisions or information unavailable at the forecast date can receive an unrealistically favorable score (Richardson et al., 2021; Qureshi et al., 2026).

The synthesis is deliberately cautious about the language of superiority. “Better” has meaning only with respect to a specified loss function, horizon, sample, and benchmark. For example, a model may lower root mean squared error while producing larger errors during recessions, or perform well at one quarter ahead but poorly at four quarters. A study-specific winner is therefore described as a result within its evaluation design. Unless a paper documents statistical uncertainty or repeated comparisons, modest numerical differences

should not be presented as a general ranking of algorithms (Kant et al., 2025; Chu & Qureshi, 2023).

This design has limitations. The corpus was assembled purposively from relevant studies and checked against publisher records; it was not retrieved through a documented, comprehensive database search with a dated query, duplicate screening, and independent reviewers. Publication and selection bias are therefore possible. In addition, abstracts and reported results cannot always establish the full reproducibility of code or data. The review offers a critical map of selected evidence and a practical standard for future experiments rather than a prevalence estimate for the entire forecasting literature.

2.1 Conceptual distinctions and evaluation principles

The first distinction is between the target and its measurement. Annual real GDP growth, quarterly growth at an annualized rate, year-on-year changes, monthly output indicators, and real GDP per capita are not interchangeable. A model predicting the revised quarterly level cannot automatically be compared with one predicting the first official estimate or a population-adjusted series. The target definition determines what information is admissible and which decisions the forecast might support. Reviews that group these outcomes under a single "growth" label risk obscuring meaningful differences in difficulty (Chen et al., 2025; Richardson et al., 2021).

The second distinction is between forecast origin and horizon. At origin t, a forecaster should only use variables that were available at t, even if the database later displays revised histories. This rule is especially consequential for nowcasting because monthly indicators are released on different schedules and the latest period contains missing values. A rolling-origin experiment refits or updates the model using successive historical cutoffs. An expanding window adds observations over time, whereas a fixed-length rolling window discards older observations. Neither scheme is intrinsically superior; each represents a different view of stability and adaptation (Richardson et al., 2021; Muchisha et al., 2021).

A third distinction concerns the benchmark. An autoregressive model using lagged GDP, a dynamic factor model using many indicators, a regularized linear model, and an official forecast test different hypotheses about the benefit of ML. An improvement over a weak constant or naive benchmark is less informative than an improvement over a carefully tuned competitor with the same data at the same origin. Parsimonious models may be especially competitive when the sample is small or a few timely predictors are strong. The U.S. evidence of Chu and Qureshi illustrates that predictor quality can matter more than increasing model complexity (Chu & Qureshi, 2023).

Typical error measures answer different questions. Mean absolute error summarizes the average magnitude of misses, while root mean squared error gives more weight to large misses. Percentage errors may behave poorly when the growth rate is close to zero or negative, so the metric should be interpreted alongside the target scale. A useful evaluation reports more than one error measure, describes the forecast horizon precisely, and shows whether results persist across subperiods. Tests of predictive accuracy can supplement the comparison, but the test design must account for the relevant dependence and sample size (Maccarrone et al., 2021; Kant et al., 2025).

Finally, feature importance is not a substitute for causal identification. A model can use unemployment, consumption, searches, or sentiment to forecast GDP because those indicators respond to the same underlying conditions. SHAP values and permutation importance describe how a fitted prediction changes under specified perturbations; they do not show what would happen to GDP if a government intervened on a feature. Their most defensible contribution here is diagnostic: identifying unstable predictors, checking plausibility, and communicating which information drives predictions in different periods (Mar Oo et al., 2025; Qureshi et al., 2026).

III. RESULTS AND DISCUSSIONS

A. Results: patterns across the selected studies

1. Targets and geographic coverage

The reviewed studies concentrate on short-term GDP prediction in single-country settings. Applications include the United States, the Netherlands, Indonesia, New Zealand, Peru, South Korea, Japan, and Canada. This concentration reflects the policy demand for timely output estimates and the availability of frequent indicators. The target is often the current quarter or a small number of quarters ahead, rather than a long-term growth rate. This is an important characterization of the selected literature, not an estimate of how all global ML research is distributed (Yoon, 2021; Muchisha et al., 2021; Richardson et al., 2021; Kant et al., 2025).

Per capita and multi-year growth targets appear less frequently in this corpus. Chen and colleagues compare deep-learning architectures for per capita GDP prediction, while Mar Oo and colleagues examine cross-country predictions at horizons extending several years. Their targets are closer to broader development questions, yet differ from structural causal growth models. Longer horizons amplify uncertainty arising from policy changes, crises, and revisions in covariates. A model that explains useful short-term co-movement among indicators may not retain the same relationship over six to ten years (Chen et al., 2025; Mar Oo et al., 2025).

The contrast between settings is particularly instructive. A country with long, consistent, timely macroeconomic series can support high-dimensional methods that would be unstable in a country with a short annual record. Cross-country panels add observations but introduce comparability issues in measurement, institutions, and structural breaks. These challenges make external validity an empirical question: a model trained successfully in one setting cannot simply be exported to another without checking variable definitions, availability, and forecast performance (Chu & Qureshi, 2023; Mar Oo et al., 2025).

2. Predictor sets and timeliness

Official macroeconomic series remain the backbone of the studies. Monetary and financial variables, prices, labor-market measures, surveys, and activity indicators can provide signals ahead of the GDP release. Larger predictor sets may help flexible models discover interactions, but the increase in dimensions also raises the chance of unstable selection. A high-frequency indicator is valuable when it is timely and informative at the forecast origin, not merely because it expands the number of variables in a final dataset (Chu & Qureshi, 2023; Kant et al., 2025).

Alternative data have a more conditional role. Internet search intensity can arrive quickly, yet its relationship to aggregate activity can shift with media attention, platform use, or a shock. Qureshi and colleagues examine dynamic feature selection using Google Trends alongside official Canadian indicators, illustrating the importance of re-evaluating relevance within successive windows. The appropriate comparison is an official-data baseline under the same release calendar. Adding alternative variables without this comparison cannot isolate their incremental forecasting value (Qureshi et al., 2026).

The real-time problem extends beyond the timing of a feature's publication. GDP itself is revised, and related indicators are sometimes backfilled or reclassified. A dataset downloaded today can contain values that forecasters could not see at a historical forecast origin. Richardson and colleagues explicitly use real-time vintages for New Zealand output and

predictors, giving their exercise a stronger basis for claims about operational nowcasting. Pseudo-real-time studies remain informative, but their design and interpretation should be labeled accurately (Richardson et al., 2021).

3. Algorithm comparisons

Tree ensembles repeatedly perform well in selected nowcasting exercises, but their advantages are conditional. Kant and colleagues report strong random-forest forecasts and nowcasts for the Netherlands while a dynamic factor model is more accurate for backcasts, especially in recessions. The same study therefore resists a single all-purpose winner. Results in Japan and Indonesia also support the usefulness of tree-based approaches, but their evaluation samples and benchmarks differ. Cross-study ranking from these separate exercises would be unjustified (Kant et al., 2025; Yoon, 2021; Muchisha et al., 2021).

The U.S. study by Chu and Qureshi provides a direct warning against equating flexibility with superiority. Density-based approaches can perform well with large predictor sets for short horizons, yet compact models using strong timely indicators can compete at longer horizons. Forecast combinations can also reduce sensitivity to the failure of one method. The underlying lesson is that signal quality, horizon, and evaluation design jointly determine performance. Algorithm labels alone convey little about the practical value of a forecast (Chu & Qureshi, 2023; Muchisha et al., 2021).

Deep-learning results are also mixed by task. Pyo compares recurrent and hybrid architectures for South Korean GDP growth. Chen and colleagues examine deep-learning methods for per capita GDP, and Mar Oo and colleagues find an artificial neural network performs strongly in their longer-horizon cross-country comparison. These findings demonstrate possible value in modeling nonlinear and temporal patterns; they do not prove that deep learning dominates simple models whenever macroeconomic data are scarce. Comparability requires attention to sample size, architecture tuning, features, and held-out periods (Pyo, 2025; Chen et al., 2025; Mar Oo et al., 2025).

Maccarrone and colleagues compare machine-learning, linear, and autoregressive approaches to U.S. GDP forecasting, illustrating why the choice of information structure matters. Statistical models can exploit persistence and economic indicators with fewer estimated parameters. If an ML model uses a richer predictor set, the observed gain may reflect better information rather than the learning algorithm itself. Factorial comparisons using the same predictors across model classes are especially valuable for separating the effects of data from modeling flexibility (Maccarrone et al., 2021; Chu & Qureshi, 2023).

Recent additions to the selected corpus widen the range of national contexts and methods. A Somalia application examines macroeconomic and environmental indicators in an economy where annual data and structural volatility complicate model training. It is useful as a reminder that predictive models designed for long, stable quarterly datasets cannot be assumed to retain their advantage with a smaller and differently measured series. The interpretation of such a comparison must emphasize the precise holdout design and uncertainty associated with the limited number of observations, rather than extrapolating a model ranking from one country to all developing economies (Ibrahim et al., 2026).

A study of quarterly GDP growth rates by Jönsson similarly highlights that forecasting performance should be assessed against plausible competing methods at the forecast horizon of interest. Comparisons based on quarterly data are sensitive to the number of observations,

the treatment of lags, and whether evaluation periods include economic disruptions. This type of evidence contributes to the current review by strengthening the case for clearly specified benchmark and horizon choices, but the mere availability of another national experiment does not establish a universal winner among tree ensembles, neural networks, or linear models (Jönsson, 2026; Chu & Qureshi, 2023).

Peru's monthly GDP nowcasting application examines unstructured information alongside more conventional leading indicators. Search data may deliver a timely signal, particularly when official information arrives with delay, but its incremental value should be reported relative to forecasts using the same official predictors at the same date. Platform behavior and language can change over time, making a historical association difficult to transport without fresh testing. The Peruvian case therefore contributes evidence about the design of an alternative-data experiment rather than proof that online searches will improve every macroeconomic forecast (Tenorio & Pérez, 2025; Qureshi et al., 2026).

A hybrid recurrent architecture examined by Shams and colleagues adds to the diversity of algorithmic designs in the literature. Its application differs from a conventional national quarterly nowcast in data context and model structure, which makes direct error comparisons inappropriate. Hybrid models can combine representation steps with sequence learning, yet each additional component increases tuning decisions and potential overfitting if the test period influences construction. The appropriate contribution of this paper to the synthesis is to illustrate the need for transparent preprocessing, a strong baseline, and evaluation on held-out periods when interpreting a reported improvement (Shams et al., 2024).

4. Study comparison

Table 1 summarizes representative contrasts rather than presenting a numerical league table. Its findings reflect the authors' reported evaluation designs. The columns emphasize the target and comparator so readers can see why outcomes cannot be pooled without a harmonized experiment (Yoon, 2021; Muchisha et al., 2021; Richardson et al., 2021; Kant et al., 2025; Mar Oo et al., 2025).

Tabel 1. Selected Studies Comparing Machine Learning Approaches to GDP Growth Forecasting and Nowcasting

Study	Target and setting	Comparison	Reported implication
Yoon (2021)	Japan, real GDP growth	Boosting vs forest and benchmarks	Gradient boosting performs well in its tested setup.
Muchisha et al. (2021)	Indonesia, quarterly GDP	ML models vs AR(1)	Forecast combinations and ensembles are competitive.
Chu and Qureshi (2023)	United States, GDP growth	Many vs few predictors	Strong timely inputs can beat complex models at longer horizons.
Richardson et al. (2021)	New Zealand, GDP nowcast	ML with real-time vintages	Real-time release information matters for evaluation.
Kant et al. (2025)	Netherlands, GDP nowcast	Forest vs dynamic factor model	The preferred method depends on forecast task and horizon.
Mar Oo et al. (2025)	109 countries, multi-year growth	Five ML model classes	Neural network strong at longer horizons in this dataset.

B. Discussion

1. What the evidence supports

Across this corpus, ML is a credible addition to the forecasting toolkit. Flexible models can exploit large indicator panels, nonlinearities, and interactions, and several selected applications report lower errors than particular statistical baselines. Yet the evidence supports a conditional proposition rather than "ML consistently outperforms econometrics." A dynamic factor model wins some backcasting comparisons, compact predictors can outperform more elaborate approaches at longer horizons, and results are sensitive to evaluation periods. A balanced conclusion treats performance as specific to the target, forecast origin, horizon, data, and benchmark (Kant et al., 2025; Chu & Qureshi, 2023).

The literature also supports a practical distinction between operational and retrospective success. Retrospective studies can uncover promising signals and methods, but policymakers require forecasts made with the data actually available at the time of decision. Real-time vintages, the publication calendar, and the treatment of missing latest-period observations can materially change the ranking of methods. Replication should therefore report both the final-data experiment and, where feasible, a genuinely real-time or carefully reconstructed pseudo-real-time evaluation (Richardson et al., 2021; Qureshi et al., 2026).

The reviewed evidence does not establish which inputs cause economic growth. Even when a model predicts growth from unemployment, consumer spending, or online attention, these variables can be consequences, correlates, or proxies of the underlying cycle. Causal policy claims require separate identification assumptions and designs. Predictive tools can still be useful for monitoring and scenario preparation, but their use should not imply that changing a highly ranked feature will change GDP in the direction suggested by its learned association (Mar Oo et al., 2025; Maccarrone et al., 2021).

2. Why some findings differ

Differences in target construction can change apparent performance. Nowcasting a quarterly figure after several monthly indicators are published is a different task from predicting the same figure before the quarter begins. A GDP revision can change the dependent variable used to score historical forecasts. Countries may also differ in release schedules and indicator reliability. These design choices must be recorded before concluding that an algorithm's advantage in one paper contradicts a different paper's result (Richardson et al., 2021; Kant et al., 2025).

Evaluation-window choice matters for macroeconomic data because recessions, pandemics, commodity shocks, and recoveries have different statistical patterns. An average error over a quiet period can hide poor performance when an early warning is most valuable. Reporting errors by subperiod and alongside a measure of large misses can expose such weaknesses. Similarly, a gain from Google Trends during a volatile episode need not persist during stable expansion. Stability is a separate dimension of performance, not merely a footnote to the average score (Chu & Qureshi, 2023; Qureshi et al., 2026).

Tuning and feature selection can introduce leakage when performed on the whole dataset before the holdout period is defined. Proper rolling evaluation repeats preprocessing, scaling, variable filtering, and parameter selection using training observations available at each

origin. That rule applies to neural networks as well as comparatively simple regularized models. Without it, the reported out-of-sample result can indirectly use future information. Published comparisons should specify which steps were nested inside each training window so other researchers can reproduce the claimed improvement (Muchisha et al., 2021; Qureshi et al., 2026).

The distinction between benchmark strength and empirical relevance is also central to synthesis. A comparison can be methodologically clean and still address a narrow decision if the target arrives too late for a policymaker's use. Conversely, an operationally timely model can have only a small average advantage over a simpler alternative. A review should therefore avoid translating statistical ranking into practical value without asking who uses the forecast, when the number is needed, and what errors are especially costly. The literature's contribution is strongest when the authors align those decisions explicitly with the reported evaluation design (Richardson et al., 2021; Kant et al., 2025).

Another useful distinction is between model uncertainty and data uncertainty. Hyperparameter tuning, random initialization, and sample selection can change the fitted model, while revisions in the official GDP series can change the target used to score it. Repeating model fits across seeds may assess one kind of variability, but it cannot replace an analysis of real-time vintages or release lags. Researchers should report both where appropriate, since a model that is stable under random initialization may still be fragile to revised national-account observations (Richardson et al., 2021; Chu & Qureshi, 2023).

Researchers should also avoid treating an explanation as stable merely because the same broad variable category appears in multiple papers. Financial indicators, consumption measures, and search information can overlap in the signals they convey. Their importance can shift depending on the chosen model, correlated covariates, or the part of the forecast window being studied. A transparent analysis would show how explanations change when groups of predictors are removed or the evaluation period changes, then state the remaining uncertainty. This is an informative diagnostic of model behavior, not an estimate of structural economic effects (Mar Oo et al., 2025; Qureshi et al., 2026).

For studies involving annual outcomes, the effective sample size deserves explicit discussion. Several decades of annual observations may still provide fewer independent forecast origins than a quarterly panel, especially after holding out a meaningful test period and allowing for lagged predictors. A flexible architecture with many tuning options can thus appear impressive in training data while providing unstable external forecasts. Pre-specifying a small candidate set, reporting uncertainty, and testing on genuinely unseen years can make findings more credible than selecting the best result from a large unreported search (Ibrahim et al., 2026; Mar Oo et al., 2025).

3. Gaps and research agenda

The concentration on headline GDP nowcasts leaves room to examine other well-defined targets, such as real GDP per capita, sectoral output, or productivity. These extensions should begin with the economic question and data availability rather than a preference for a more complex algorithm. Sectoral output may respond differently to external shocks and policy, whereas per capita GDP incorporates population dynamics. Measures of total factor productivity introduce additional measurement assumptions and should not be treated as

directly observed truth. Carefully specifying the outcome is therefore the first step toward useful research (Chen et al., 2025; Mar Oo et al., 2025).

Emerging-economy comparisons would benefit from evaluation protocols that respect short samples and structural instability. Researchers could use multiple countries while testing performance separately by country and regime, and prevent information from the test period from influencing feature engineering. A panel model that scores well on average may obscure poor performance in a particular country. External validation on countries or years excluded from training can provide more informative evidence about transferability than a random split of pooled observations (Mar Oo et al., 2025; Muchisha et al., 2021).

Interpretation should be a planned component of evaluation. Feature-importance profiles can be recalculated across forecast origins to show whether a model relies on stable signals or shifts sharply after shocks. Researchers can compare the explanation of a flexible model with the coefficients or factors of a benchmark, check whether influential variables were available on time, and examine large-error episodes. These checks support model governance and economic plausibility. They remain predictive diagnostics and cannot alone resolve causal questions (Mar Oo et al., 2025; Qureshi et al., 2026).

A stronger reporting standard would present the exact target transformation, forecast dates, information-release calendar, training window, preprocessing pipeline, hyperparameter search, data vintage, comparator, and loss function. It would also report uncertainty around relative errors and make code or an executable specification available when possible. Such transparency allows future reviewers to distinguish genuine advances from changes in the experiment. A small, well-described improvement that persists in real-time evaluation is more useful than a large but irreproducible gap against a weak baseline (Richardson et al., 2021; Chu & Qureshi, 2023).

4. Implications for empirical design

A proposed follow-up study could compare regularized regression, tree ensembles, and a deliberately small neural network against autoregressive and dynamic factor baselines on the same GDP target. Its first stage should freeze the target definition and forecast origins, then reconstruct the predictor set as it was available at each origin. Every model should receive the same underlying information when the research question concerns algorithm choice. A second comparison could add alternative data to each model class to estimate the incremental value of that information (Richardson et al., 2021; Qureshi et al., 2026).

Model selection should occur within the training window. An outer rolling evaluation would produce the forecasts used for the final comparison; an inner time-ordered procedure would choose features and parameters without looking at the outer holdout. This nested arrangement is computationally costly, but it prevents selection from benefiting from outcomes that were not yet known. The researcher should record the number of forecasts at each horizon and avoid comparing scores derived from different dates as though they came from the same task (Chu & Qureshi, 2023; Muchisha et al., 2021).

The reporting design should include mean absolute and root mean squared errors, errors during high-volatility periods, and a comparison against each baseline on matched forecast dates. A formal predictive-accuracy test can be added if its assumptions fit the overlapping horizons and sample size. With short macroeconomic series, a narrow error gap

should be treated cautiously even when one model ranks first. Error distributions and individual forecast trajectories can reveal episodes that average scores conceal (Kant et al., 2025; Richardson et al., 2021).

Finally, policy usefulness should be evaluated separately from pure error reduction. Decision-makers may value timely detection of a contraction, calibrated uncertainty, or stable revisions across successive nowcasts. A model that slightly improves average RMSE while producing unreliable signals around turning points could be less useful in practice. Reporting these additional outcomes would connect technical performance to the decisions for which the GDP forecast is intended, while preserving the distinction between monitoring and causal policy evaluation (Richardson et al., 2021; Maccarrone et al., 2021).

IV. CONCLUSION

The selected journal literature shows that machine learning can improve GDP prediction in well-defined settings, particularly when timely indicators and nonlinear patterns provide useful information, but no single model consistently outperforms all econometric baselines. Reported gains depend on the target, forecast horizon, predictors, validation design, and historical period; the quality and timing of information can matter as much as algorithmic flexibility (Chu & Qureshi, 2023; Kant et al., 2025; Richardson et al., 2021). Future research should compare models using identical information sets, reconstruct which data were available at each forecast origin, assess performance across economic conditions, and examine longer-horizon outcomes with appropriate attention to measurement and validation. Model explanation can help assess predictions, while causal claims require separate research designs (Mar Oo et al., 2025; Qureshi et al., 2026). Because this review examines a purposively selected set of DOI-identified journal articles rather than an exhaustive, pooled synthesis, its contribution is to clarify what the available comparisons support and how future GDP forecasting studies can produce more credible and useful evidence (Richardson et al., 2021; Chu & Qureshi, 2023).

ACKNOWLEDGEMENT

The author acknowledges the use of ChatGPT to assist with language editing and drafting portions of this review. The author reviewed and revised the text, verified the citations and references, and takes full responsibility for the final manuscript.

REFERENCES

Journal Articles

Chen, X.-S., Kim, M. G., Lin, C.-H., & Na, H. (2025). Development of per capita GDP forecasting model using deep learning: Including consumer goods index and unemployment rate. *Sustainability, 17*(3), 843. https://doi.org/10.3390/su17030843

Chu, B. M., & Qureshi, S. (2023). Comparing out-of-sample performance of machine learning methods to forecast U.S. GDP growth. *Computational Economics, 62*, 1567–1609. https://doi.org/10.1007/s10614-022-10312-z

Ibrahim, A. D., Mohamed, A. M., & Abdulle, B. M. (2026). Comparative performance of machine learning models in forecasting Somalia's GDP growth using macroeconomic and

environmental indicators. *Discover Sustainability, 7*, Article 1541. https://doi.org/10.1007/s43621-026-03895-2

Jönsson, K. (2026). Forecasting quarterly GDP growth rates using machine learning methods. *Journal of Forecasting*. Advance online publication. https://doi.org/10.1002/for.70184

Kant, D., Pick, A., & de Winter, J. (2025). Nowcasting GDP using machine learning methods. *AStA Advances in Statistical Analysis, 109*, 1–24. https://doi.org/10.1007/s10182-024-00515-0

Maccarrone, G., Morelli, G., & Spadaccini, S. (2021). GDP forecasting: Machine learning, linear or autoregression? *Frontiers in Artificial Intelligence, 4*, Article 757864. https://doi.org/10.3389/frai.2021.757864

Mar Oo, Z., Lin, C.-Y., & Kakinaka, M. (2025). Deciphering long-term economic growth: An exploration with leading machine learning techniques. *Journal of Forecasting, 44*(4), 1531–1562. https://doi.org/10.1002/for.3254

Muchisha, N. D., Tamara, N., Andriansyah, A., & Soleh, A. M. (2021). Nowcasting Indonesia's GDP growth using machine learning algorithms. *Indonesian Journal of Statistics and Its Applications, 5*(2), 355–368. https://doi.org/10.29244/ijsa.v5i2p355-368

Pyo, D.-J. (2025). Enhancing GDP growth forecasting with LSTM, GRU, and hybrid model: Evidence from South Korea. *SAGE Open, 15*(1), Article 21582440251359828. https://doi.org/10.1177/21582440251359828

Qureshi, S., Chu, B. M., Demers, F. S., Khan, N., & Irshad, A.-U.-R. (2026). Dynamic feature selection for Canadian GDP forecasting: Machine learning with Google Trends and official data. *Machine Learning and Knowledge Extraction, 8*(3), Article 66. https://doi.org/10.3390/make8030066

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting, 37*(2), 941–948. https://doi.org/10.1016/j.ijforecast.2020.10.005

Shams, M. Y., Tarek, Z., El-Kenawy, E.-S. M., Eid, M. M., & Elshewey, A. M. (2024). Predicting gross domestic product (GDP) using a PC-LSTM-RNN model in urban profiling areas. *Computational Urban Science, 4*, Article 3. https://doi.org/10.1007/s43762-024-00116-2

Tenorio, J., & Pérez, W. (2025). Nowcasting del PBI mensual peruano con machine learning y datos no estructurados [Nowcasting Peru's monthly GDP with machine learning and unstructured data]. *Apuntes, 52*(99), 5–42. https://doi.org/10.21678/apuntes.99.2189

Yoon, J. (2020). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics, 57*, 247–265. https://doi.org/10.1007/s10614-020-10054-w