International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

A Hybrid Modeling Approach to the Gross Domestic Product Forecasting: By utilizing ARIMA and Machine Learning Techniques

P. Srivyshnavi[1], M Darshan Teja [2] ,P Shankaraiah [3*], Sreenivasulu T [4]

[1]Assistant Professor , Dept. of CS & E, S.P.M.V.V. Engineering College, Tirupati, India.
[2, 3, 4] Department of Mathematics, School of Advance Science, VIT Vellore, India.

Corresponding author: stats4shankar@gmail.com *

Abstract

In order to anticipate India's GDP, this study investigates a hybrid model that combines machine learning and ARIMA approaches. The dataset was split between training (1950-2010) and testing (2011-2022) eras using historical data spanning from 1950 to 2022. ARIMA, ARIMA with Random Forest, ARIMA with XGBoost, and ARIMA with SVM were the four models that were compared. For ARIMA, AIC was used to assess model performance; for other models, RMSE and MAE were used. With an R-squared value of 0.98, ARIMA with XGBoost outperformed the other three models: ARIMA (0.96), ARIMA with Random Forest (0.95), and ARIMA with SVM (0.25). The univariate ARIMA model was chosen for GDP forecasting from 2023 to 2040 despite the excellent accuracy of ARIMA with XGBoost because of its ease of use and efficiency.

Key words: ARIMA, Random Forest, XGBoost, Support Vector Machine, Forecasting.

1. Introduction

GDP is an essential measure of the economy since it considers the total value of all products and services produced in a country in a certain period of time. It was measured using output, revenue and expenditure methods [1]. India's GDP has evolved since 1947. In 1991, the Indian government introduced new market reform initiatives, dismantling government controls and expanding the global market economy. It has been significantly enlarged by the financial and services industry and IT and telecommunication industries. India's GDP has been increasing year after year to become the world's fifth biggest economy in 2023 [2, 3]. Time series analysis is a sort of statistical approach used to identify trends, patterns, and cyclical movements of the gross domestic product (GDP) over a period of time which assists in anticipating future trends and facilitating decision making for policymakers, economists, and investors. The Autoregressive Integrated Moving Average (ARIMA) approach is a popular time series forecasting model, especially for short-term forecasts and economic planning, modeling varied data patterns and projecting GDP [4,5]. In this study we try to build and test hybrid models for forecasting GDP in India and compare their performance to find out the most successful strategy. The integration of ARIMA with machine learning methods is expected to produce more accurate and dependable GDP projections [6]. The ARIMA method is robust in time series forecasting, but its combination with machine learning techniques can enhance the predicted accuracy and adaptability to complicated data patterns. This paper proposes the combination of ARIMA with recent Machine learning techniques like Random forest, XGBoost and Support vector machines (SVM) for prediction of India's GDP [7].
India's historical GDP data is being analysed using the ARIMA model, which is being used to fit and validate the data stationarity. The performance of the model will be assessed using performance indicators such as RMSE, MAPE, and R^2. The objective is to improve the model's

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

prediction accuracy by the integration of machine learning techniques such as Support Vector Machines (SVM), XGBoost, and Random Forest. Through a methodical analysis and Relation of RMSE, MAPE, and R-squared values, the study seeks to assess the forecasting accuracy and dependability of the ARIMA model in comparison to hybrid models [8]. Planning and decision-making in the economy will benefit from these insights. In the literature review, the usage of the ARIMA model and its hybrid with other models on a wide variety of real-world data sets have been discussed many in number. For time series modelling in the real world, a hybrid system combines linear and nonlinear forecasting approaches. The system models time series linearly, error series nonlinearly, and in a combination determined by the data. Better results are obtained from experimental simulations [9]. These studies highlight the effectiveness of hybrid models in increasing accuracy of forecast. However, our proposal is about the application of ARIMA and machine learning techniques to real-world economic data i.e. India's GDP statistics. The present study blends ARIMA with machine learning models such as Random Forest, XGBoost and SVM to take advantage of the capabilities of both classical time series analysis and complex machine learning approaches utilizing R software [10]. The goal is to provide better and more robust GDP estimations. In addition to investigating the performance of these hybrid models, this study also examines the utility of these models compared to the normal univariate ARIMA model. The practical consequences and advantages of utilizing machine learning techniques in economic forecasting is also discussed.

2. Methodology

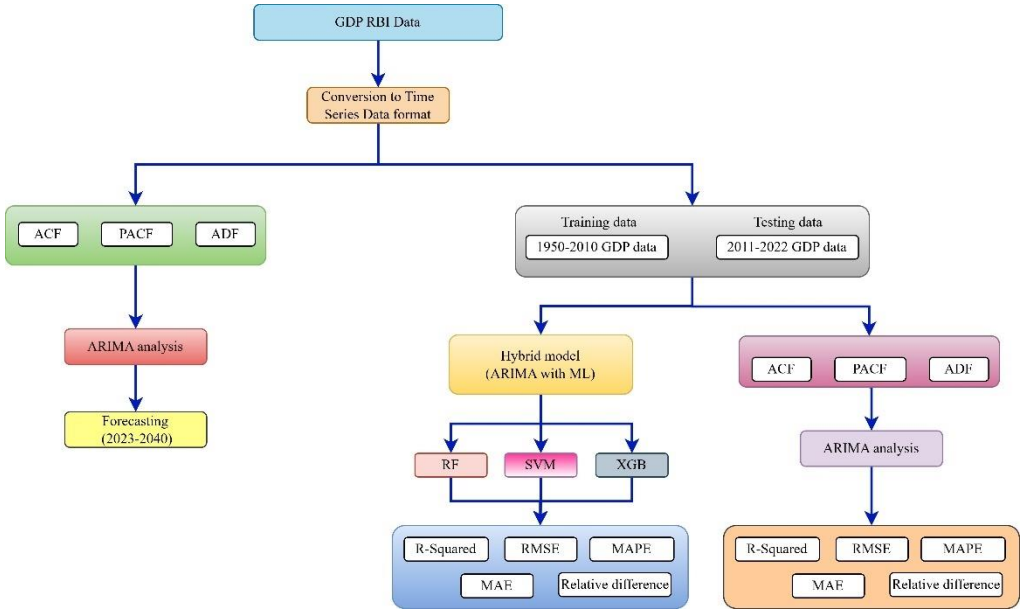

Fig 1: Process for Hybrid model

Hybrid Process for Forecasting GDP

- Data from the Reserve Bank of India (RBI) collected from 1950 to 2022.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

- Converted into a time series format.
- Divided into two parts for analysis.
- Target variables are checked using the Augmented Dickey-Fuller (ADF) test, the autocorrelation function (ACF), and the partial autocorrelation function (PACF).
- The ARIMA model was selected for forecasting GDP values from 2023 to 2040.
- Time series data is divided into training data (1950–2010) and testing data (2011–2022).
- Hybrid models combining ARIMA with machine learning techniques applied.
- Performance is evaluated using metrics like RMSE, MAPE, RD, MAE, and R-square.
- A comprehensive approach ensures the best-fitting model for accurate GDP forecasting.

3. Methods

3.1 ARIMA

The ARIMA Model is a most Important statistical tool for time series forecasting, Integrating three essential components are Autoregression(AR), Differencing (I), and Moving Average(MA). The AR part models was dependency of a variable on its own previous values and parameter is P, the I components involves differencing the data to achieve stationarity and represented by parameters d, and the MA part models are relationship between a variable and past error terms and parameters is q. over all ARIMA represents as p, d, q. [11].

The General ARIMA equation incorporates these elements to predict future values based on past observations and errors. By identifying the appropriate values of p, d, q, estimating the model parameters and ensuring the residuals behave like white noise, the ARIMA model provides accurate forecast for various timeseries patterns, making it a fundamental approach in time series.

$$\Delta^d y_t = c + \theta_1 \Delta^d y_{t-1} + \theta_2 \Delta^d y_{t-2} + \cdots + \theta_p \Delta^d y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (1)$$

3.2 ARIMA with Random Forest

The ARIMA model and Random Forest are combined to create a hybrid approach for time series forecasting. ARIMA captures linear patterns, while Random Forest handles non-linear relationships. The ARIMA model models and removes linear components, while the Random Forest model captures complex, non-linear dependencies. This dual-stage process improves forecasting accuracy, especially for datasets with both linear and non-linear characteristics. This combination is particularly effective for predicting future values [12].

3.3 ARIMA with SVM

We use the ARIMA model with Support Vector Machines (SVM) to improve time series forecasting by integrating statistical and machine learning methods. ARIMA is good at modeling linear patterns, but not good at modeling complicated, non-linear interactions. The algorithm SVM handles these complex patterns in a powerful manner [13]. The ARIMA methodology is used to the time series data, linear components are removed and the residuals are utilized as the input to the SVM model. This two-stage approach gives accurate and robust forecast especially for datasets with linear and non-linear features.

3.4 ARIMA with XGBOOST model

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

The ARIMA model and XGBoost are combined to create a robust hybrid approach for time series forecasting. ARIMA is ideal for linear components like trends and seasonality, but it struggles with complex, non-linear relationships. XGBoost, a machine learning algorithm, addresses this limitation by removing linear components from the ARIMA model and using residuals as inputs for the XGBoost model [14]. This two-step approach improves forecasting accuracy, especially for time series datasets with both linear and non-linear patterns.

4. Analysis and Results

4.1 ARIMA Analysis

The GDP Constant data for this study were obtained from the Reserve Bank of India (RBI) website. To enable analysis, the data was first transformed into a time-series format. The dataset was then separated into two parts: training (1950–2010) and testing (2011–2022). To guarantee that the data was adequate for time series modelling, stationarity had to be checked. The Augmented Dickey-Fuller (ADF) test was used for this purpose. Initially, the ADF test revealed that the data was not stationary (p-value = 0.8). For a series to be declared stationary, its p-value must be less than or equal to 0.05. To overcome this, the data underwent differencing and logarithmic adjustments. Subsequent ADF tests revealed that the modified data was steady, with p-values less than or equal to 0.05. With the data now stationary, the next step was to pick a suitable ARIMA model, as described in the following processes.

The Auto ARIMA function in R was used to evaluate various ARIMA models for forecasting GDP data. The results showed that ARIMA(0,2,1) was the optimal model, with the lowest AIC value of 1501.78. The model's parameters included a moving average component, residual variance, and log likelihood. The model's AIC, corrected AIC (AICC), and Bayesian Information Criterion(BIC) values were 1501.78, 1501.99, and 1505.93, respectively. This model effectively balanced model complexity and goodness of fit.

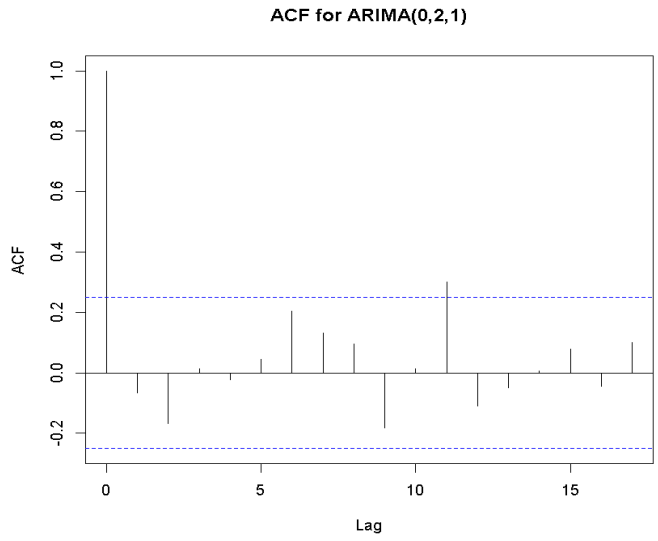

Fig 2.1 Auto correlation Function

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

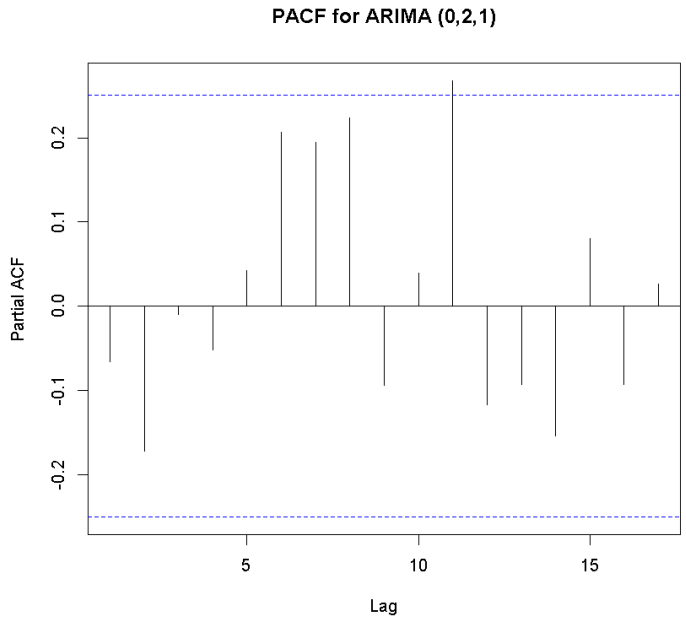

Fig 2.2 Partial Auto correlation Function

The ARIMA(0,2,1) model was applied to GDP data and its residuals were analyzed using ACF (autocorrelation function)and PACF (partial autocorrelation function) plots. In the Fig 2.1 ACF plot showed significant autocorrelation at lag 0, indicating a strong correlation in residuals. Beyond lag 0, autocorrelations fell within confidence limits, suggesting no significant autocorrelation up to lag 12, where a minor spike surpassed the upper limit. In the Fig 2.2 PACF plot showed a significant partial autocorrelation at lag 12, extending beyond confidence limits, while the remaining lags remained within bounds. Tables 3 and 4 discuss the relationship between actual values and ARIMA predicted values, as well as their characteristics.

4.2 ARIMA with Machine learning

In this work, we investigate the use of the hybrid model (ARIMA method combined with machine learning) approaches to forecast India's GDP. Using GDP data from the RBI office website from 1950 to 2022, we split the dataset into training data (1950–2010) and testing data (2011). The data is initially transformed into a time series format. For the training data, in this proposal model use a variety of economic indicators as feature variables, including Private Final Consumption Expenditure (PFCE), Government Final Consumption Expenditure (GFCE), Gross Fixed Capital Formation (GFCF), Stock Changes (CS), Exports of Goods and Services (EGS), and Imports of Goods and Services (IGS), with GDP as the target variable.

The ARIMA approach is combined with machine learning techniques such as Random Forest, XGBoost, and Support Vector Machine (SVM) to improve GDP forecast accuracy. This hybrid approach captures linear dependencies and non-linear relationships, enhancing predictive power

by capturing complex patterns. It effectively handles time series data and enhances predictive power.

4.2.1 ARIMA with Random Forest

Table 1: Different model of ARIMA with Random Forest

mtry	R-square	RMSE	MAE
2	0.9929	445743.9	243270.7
4	0.9917	468251.7	249695.2
6	0.9903	498758.6	265461.0

The Random Forest model was tested using resampling with 25 bootstrapped repeats. The performance metrics showed that the best model, with 'mtry' set to 2, had an RMSE of 445,743.9, an R-squared value of 0.9929574, and a Mean Absolute Error (MAE) of 243,270.7. These results indicate the model's outstanding accuracy and effectiveness in capturing volatility within the GDP data. The combination of ARIMA and Random Forest successfully utilised both linear time series features and complicated non-linear connections among predictors, yielding strong GDP projections.

4.2.2 ARIMA with Extreme Gradient Boosting (XGBoost) model

On a dataset of 72 samples and six predictors, the study applied an ARIMA with an XGBoost model. Over 50 boosting rounds, the best performance was obtained with a rate of learning of 0.3, a maximal tree depth of 2, a column sample per tree of 0.8, and a subsample proportion of 0.75. Other combinations produced greater RMSE, R-squared, and MAE. Consistently high R-squared values and decreased RMSE and MAE values were found with a learning rate of 0.3, a maximum level of 2, and column sample by tree values ranging from 0.6 to 0.8, with a subsample of 0.75 or 1.00. The final XGBoost model, chosen based on the least RMSE, displayed a good mix of complexity and performance, producing accurate predictions with excellent dependability.

4.2.3 ARIMA with Support Vector Machine (SVM)

Table 2: Different model of ARIMA with SVM

C	R-Square	RMSE	MAE
0.25	0.6415	3133133	1761106
0.50	0.6676	2882087	1644502
1.00	0.6818	2698819	1580832

To analyse a dataset of 72 samples and six predictors, the researchers utilised an SVM model using a Radial Basis Function (RBF) kernel. The model was assessed using bootstrapped resampling with 25 repetitions to obtain the best tuning parameters. The tuning parameter (sigma) was set at 11.83755, while the cost parameter (C) was changed. With C = 0.25, the model has an RMSE (root mean squared error) of 3,133,133, an R-squared value of 0.6415870, and an MAE of 1,761,106. Increasing the cost parameter to C = 0.50 produced an RMSE of 2,882,087, an R-squared value of

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

0.6676438, and an MAE of 1,644,502. The final model was chosen based on the least RMSE value, resulting in sigma = 11.83755 and C = 1.00.

Table 3: Different between Actual Values and Forecast Values of Different Model

		Forecast Values			
Year	**Actual Value**	**ARIMA**	**ARIMA with Random Forest**	**ARIMA with SVM**	**ARIMA with XGBoost model**
2011	8736329	8831312	9019131	8329194	8798847
2012	9213017	9361390	9394399	8807195	9144893
2013	9801370	9891468	9771612	9395213	9806064
2014	10527674	10421545	10403085	10124029	10537076
2015	11369493	10951623	11263313	10959914	11368888
2016	12308193	11481700	12216326	10869380	12304418
2017	13144582	12011778	13267405	10869655	13167960
2018	13992914	12541856	13799975	10869645	13996160
2019	14534641	13071933	14007400	10887704	14551239
2020	13687118	13602011	13282164	10869253	13698790
2021	14925840	14132088	14112228	10887704	14906805
2022	16006425	14662166	14050471	6818651	14906805

Table 4: Relative Difference and MAPE for different Model

ARIMA		**ARIMA with Random Forest**		**ARIMA with SVM**		**ARIMA with XGBoost**	
Relative Difference	MAPE	Relative Difference	MAPE	Relative Difference	MAPE	Relative Difference	MAPE
-1.08722	-0.0906	-3.23708	-0.26976	4.660249	0.388354	-0.71561	-0.05963
-1.61048	-0.13421	-1.96876	-0.16406	4.404873	0.367073	0.739428	0.061619
-0.91924	-0.0766	0.303609	0.025301	4.143879	0.345323	-0.04789	-0.00399
1.008092	0.084008	1.183439	0.09862	3.834129	0.319511	-0.08931	-0.00744
3.675363	0.30628	0.933904	0.077825	3.60244	0.300203	0.005322	0.000444
6.714984	0.559582	0.746391	0.062199	11.68988	0.974157	0.030672	0.002556
8.618031	0.718169	-0.9344	-0.07787	17.30696	1.442246	-0.17785	-0.01482
10.36995	0.864162	1.378833	0.114903	22.32036	1.86003	-0.0232	-0.00193
10.0636	0.838633	3.627477	0.30229	25.09134	2.090945	-0.1142	-0.00952
0.621805	0.051817	2.958652	0.246554	20.58772	1.715643	-0.08528	-0.00711
5.317974	0.443165	5.451032	0.454253	27.05467	2.254556	0.127533	0.010628
8.398249	0.699854	12.21981	1.018317	57.40054	4.783378	6.869869	0.572489

A study comparing from Table 3 and 4 GDP forecasting models revealed significant variations in performance across them. The ARIMA with XGBoost model showed the lowest relative difference and MAPE values, indicating superior predictive accuracy and reliability. The ARIMA with SVM

model had the highest relative difference and MAPE values, indicating lower accuracy. The ARIMA with Random Forest model showed intermediate performance, capturing complex patterns but with more variability. The standalone ARIMA model had moderate performance. The integration of ARIMA with advanced machine learning techniques, particularly XGBoost, significantly enhanced forecasting accuracy, as evidenced by lower error metrics and relative differences. Overall, the ARIMA model with XGBoost was the most robust among tested.

Table 5: Different Metrics for Models

Models	RMSE	MAE	R2
ARIMA	6181006	6155600	0.96
ARIMA with Random Forest	682790	445266	0.95
ARIMA with XGBoost model	312046.7	90685.18	0.98
ARIMA with SVM	2567004	1557947	0.25

The ARIMA model, which predicts India's GDP, has shown superior predictive capability compared to other models. The ARIMA with Random Forest model showed improved accuracy, while the ARIMA with XGBoost model achieved the lowest RMSE and MAE from Table 5. The ARIMA with SVM model had the poorest performance, with an RMSE of 2,567,004, an MAE of 1,557,947, and an R-squared value of 0.25. Despite its simplicity, the univariate ARIMA model remains a robust choice for forecasting India's GDP.

In below Fig 3 compares real GDP values to ARIMA projected values using Random Forest (RF), XGBoost, and Support Vector Machine (SVM). The ARIMA with XGBoost model closely approximates real GDP levels, suggesting accurate monitoring. ARIMA with RF predictions is very close to real values but not as accurate as the XGBoost model. Traditional ARIMA models are reasonably close to real values but have a little higher deviation than hybrid models. The ARIMA with SVM model differs significantly from actual values, indicating less effective performance. These findings show that the ARIMA-XGBoost model is the most accurate.

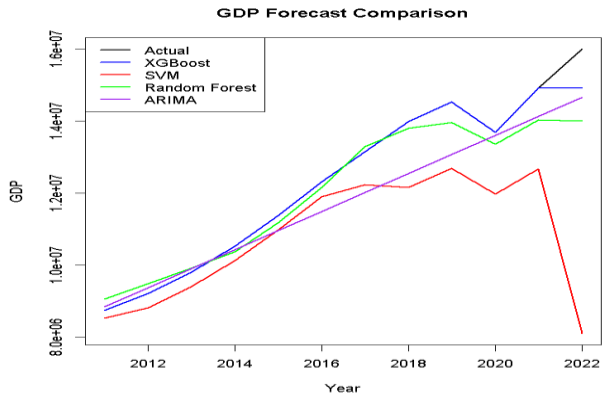

Fig 3 Compare between Actual value and forecast value of different models

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4.3 ARIMA forecasting 2023 to 2040

In this study, India's GDP from 2023 to 2040 was forecast using an ARIMA model, which was chosen for its applicability as a univariate forecasting technique. Initially, four models were tested: ARIMA, ARIMA with Random Forest, ARIMA with XGBoost, and ARIMA with SVM. Analysing the training and testing data indicated that the ARIMA with XGBoost model had the best R-squared value of 0.98, followed by ARIMA with Random Forest at 0.95, ARIMA at 0.964, and ARIMA with SVM at 0.25. Despite the greater performance of the ARIMA with XGBoost model, the ARIMA model was chosen because it focuses on forecasting a single variable, GDP, without the need for other independent variables, which are critical in machine learning approaches.

Table 6: ARIMA Models

Model	AIC Value
ARIMA(2,2,2)	1966.931
ARIMA(0,2,0)	2000.634
ARIMA(1,2,0)	1988.565
ARIMA(0,2,1)	1967.684
ARIMA(1,2,2)	1970.397
ARIMA(2,2,1)	1965.955
ARIMA(1,2,1)	1969.664
ARIMA(2,2,0)	Inf
ARIMA(3,2,1)	1967.786
ARIMA(3,2,0)	1970.265
ARIMA(3,2,2)	1968.931

The dataset, which covered 1950 to 2022, was transformed into a time series and tested for stationarity prior to using the ARIMA model. The ARIMA(2,2,1) model emerged as the best choice, as indicated by its low AIC 1965.96 and BIC 1975.01 scores. This model was used to anticipate GDP levels from 2023 to 2040. The decision to use a univariate ARIMA model demonstrates its effectiveness in capturing temporal dependencies inside a single economic indicator, resulting in solid and trustworthy projections for India's GDP during the projected timeframe.

Table 7: Forecast Value for 2023 to 2040

Year	Forecast Value (Average)	Lowest value 95%	Highest Value 95%
2023	16217586	15751115	16684057
2024	16655968	15956128	17355808
2025	17395399	16552263	18238535
2026	17987899	16952093	19023704
2027	18489081	17215668	19762494
2028	19065704	17558677	20572732
2029	19664181	17917956	21410405
2030	20228663	18223554	22233771

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

2031	20790934	18513786	23068082
2032	21367077	18809571	23924583
2033	21941468	19092817	24790120
2034	22510701	19359445	25661958
2035	23081605	19617889	26545321
2036	23654233	19868693	27439772
2037	24225874	20108940	28342807
2038	24797019	20339410	29254627
2039	25368648	20561507	30175790
2040	25940382	20775032	31105733

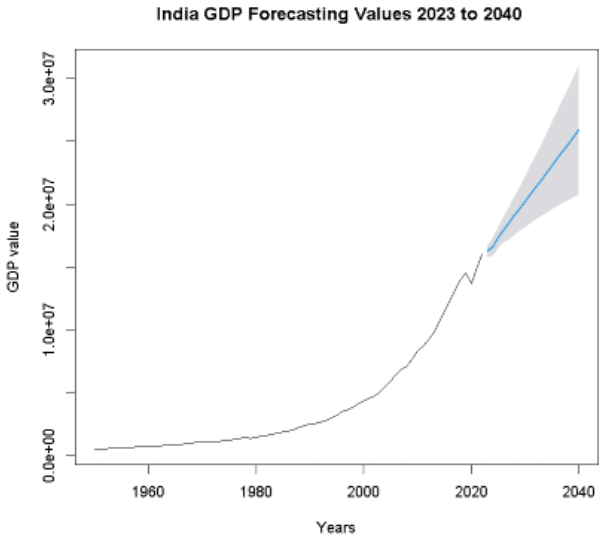

Fig 4

In the Table 7 and Fig 4 show India's GDP projections from 2023 to 2040 using the ARIMA (2, 2, 1) model. Table 7 provides yearly prediction figures, including the average, lowest, and greatest estimates at 95% confidence. Fig 4 visually depicts these projections, including the average forecast and 95% confidence interval bands. The alignment of the table and figure allows for a clear comprehension of expected economic trends and probable uncertainty around average estimates. This technique demonstrates the ARIMA (2, 2, 1) model's ability to generate trustworthy and statistically significant future GDP estimates.

5. Conclusion

The study analyzed various models, including the ARIMA model and hybrid models combining ARIMA with Random Forest, XGBoost, and SVM. The ARIMA model was chosen based on its lowest AIC values, while other models were evaluated based on their RMSE and MAE values. The ARIMA with XGBoost model demonstrated the lowest differences in metrics, indicating its superior fit. The ARIMA model achieved the highest R-squared value (0.98), followed by the ARIMA model (0.96), ARIMA with Random Forest (0.95), and ARIMA with SVM (0.25).

Despite the superior performance of hybrid models, the ARIMA model was chosen for forecasting GDP from 2023 to 2040 due to its univariate nature. The findings highlight the effectiveness of combining ARIMA with machine learning techniques for accurate forecasting and the simplicity and efficiency of the univariate ARIMA model for long-term GDP projections.

Acknowledgments

We would like to convey my sincere gratitude to the Vellore Institute of Technology, Vellore management authority and School of Advance of Technology for giving us with all the assistance and resources.

References

1. Landefeld, J. S., Seskin, E. P., & Fraumeni, B. M. (2008). Taking the pulse of the economy: Measuring GDP. *Journal of Economic Perspectives*, *22*(2), 193-216.
2. Agarwal, S., Gupta, D., & Verma, P. (2019). I mpact of Employment on GDP Contribution of Various Sectors in India. *Global Journal of Enterprise Information System*, *11*(1), 47-53.
3. Kansal, S. M. (1992). Contribution of'Other Services' Sector to Gross Domestic Product in India: An Evaluation. *Economic and Political Weekly*, 2047-2056.
4. Gamboa, J. C. B. (2017). Deep learning for time-series analysis. *arXiv preprint arXiv:1701.01887*.
5. Mills, T. C. (2019). Applied time series analysis: A practical guide to modeling and forecasting. *Academic press.*
6. Tiffin, M. A. (2016). Seeing in the dark: A machine-learning approach to nowcasting in Lebanon. *International Monetary Fund*.
7. Pin Li and Jin-Suo Zhang (2018). A New Hybrid Method for China's Energy Supply Security Forecasting Based on ARIMA and XGBoost. *Energies*, 11(7) 1687.
8. Durdu O, mer Faruk. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence,* 23 (2010) 586-594.
9. Júnior, D. S. D. O. S., de Oliveira, J. F., & de Mattos Neto, P. S. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, *175*, 72-86.
10. Cryer, J. D., Chan, K. S., & Kung-Sik.. Chan. (2008). *Time series analysis: with applications in R* (Vol. 2). New York: Springer.
11. Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.
12. Michael J Kane et. al. (2014), Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics,* 15 (276), 1471-2105.
13. Ping Feng Pai, Chih Sheng Lin (2005), A hybrid ARIMA and Support Vector Machines model in stock price forecasting. *Omega*, 33 (6), 497-505.
14. Yan Wang, Yuankai Guo et. al. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *Journal and magazines,* 17(3), 2020.