IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Forecasting Income And Wealth Inequality In India Through Nonlinear Machine Learning Techniques

[1]Khelendra Kumar Yadav, [2]D P Singh
[1]Content Manager, [2]Professor
[1]Essentially Sports, New Delhi,India
[2]Amity University Uttar Pradesh, Greater Noida Campus, India

Abstract: Income and wealth inequality continue to pose deep-rooted structural challenges in India, carrying far-reaching consequences for economic stability, inclusive development, and effective policymaking. Consequently, accurate forecasting of inequality trends is crucial for evidence-based policy decisions. This study evaluates the suitability and comparative predictive performance of nonlinear machine learning and time-series techniques for forecasting income and wealth inequality in India. Departing from the constraints of traditional linear econometric models, the analysis employs Random Forest (RF), Autoregressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) models to better capture the nonlinear relationships and temporal dynamics that characterize inequality processes.

Drawing on historical inequality measures obtained from nationally representative secondary datasets, the forecasting models are developed and tested within a common analytical framework. Model performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The empirical evidence demonstrates that machine learning–driven approaches most notably XGBoost and LSTM outperform conventional ARIMA and other benchmark models in predictive accuracy. This advantage stems from their enhanced capacity to capture nonlinear dynamics and long-range temporal dependencies. The results highlight the growing importance of advanced machine learning methods as reliable and complementary instruments for forecasting inequality in emerging economies. Overall, the study makes a methodological contribution to the inequality forecasting literature and provides actionable insights for policymakers aiming to adopt data-driven strategies to track and mitigate income and wealth inequality in India.

Index Terms **-** Income Inequality; Wealth Inequality; India; Forecasting; Machine Learning; Time-Series Analysis; Random Forest; ARIMA; Support Vector Regression; XGBoost; LSTM.

I. Introduction

Income and wealth inequality have emerged as central issues in modern economic debates, particularly within emerging economies such as India. Although the country has experienced sustained economic growth over recent decades, gaps in income distribution and asset ownership have continued to widen. This growing disparity has intensified concerns regarding social cohesion, intergenerational mobility, and the long-term sustainability of development outcomes (Atkinson, [10] Piketty, [8]). In the Indian context, persistent inequality has been driven by structural factors such as segmented labor markets, unequal educational opportunities, regional disparities, and uneven returns on capital, with certain periods witnessing a pronounced expansion of these gaps (Banerjee & Piketty, [5]; Chancel & Piketty, [18]). Consequently, analyzing and forecasting trends in income and wealth inequality are essential for informed policy formulation and effective evaluation.

Conventional empirical studies of inequality in India have predominantly relied on descriptive measures and linear, stationarity-based econometric frameworks (Deaton & Dreze, [3]; Datt & Ravallion,[7]). Although these approaches have yielded important insights, their capacity to forecast inequality trends is limited when confronted with nonlinear dynamics, structural breaks, and complex interdependencies among macroeconomic and institutional factors. Inequality is inherently a dynamic phenomenon, shaped over time by globalization, technological change, fiscal policy interventions, and demographic transitions, rendering strictly linear modeling approaches increasingly inadequate (Milanovic,[13]).

In recent years, machine learning (ML) approaches have gained prominence in economic forecasting due to their flexibility, data-driven nature, and ability to capture nonlinear relationships without imposing restrictive parametric assumptions (Varian, [9]; Athey & Imbens,[17]). Techniques such as Random Forest (RF) and Support Vector Regression (SVR) can model complex functional relationships, while ensemble methods like Extreme Gradient Boosting (XGBoost) enhance predictive performance through the iterative correction of forecasting errors. Simultaneously, advances in deep learning—particularly Long Short-Term Memory (LSTM) networks—have demonstrated substantial success in time-series forecasting by effectively learning long-range dependencies and underlying temporal dynamics (Hochreiter & Schmidhuber, [2]; Zhang et al., [16]).

Despite the growing application of machine learning techniques in finance, macroeconomic analysis, and social data analytics, their use in forecasting income and wealth inequality in India remains relatively limited. Much of the existing literature focuses on cross-national inequality comparisons or static predictive analyses, with only a small number of studies adopting a structured, comparative forecasting framework within a single developing-country context. Moreover, inequality forecasting continues to rely heavily on traditional time-series models such as the Autoregressive Integrated Moving Average (ARIMA), despite their well-documented limitations in capturing nonlinear dynamics and structural regime shifts.

To address these gaps, the present study proposes an integrated forecasting framework for income and wealth inequality in India that combines classical econometric techniques with advanced nonlinear machine learning models. The analysis systematically evaluates and compares the forecasting performance of ARIMA, Random Forest, Support Vector Regression, XGBoost, and Long Short-Term Memory (LSTM) networks using historical inequality indicators derived from nationally representative secondary data sources. Forecast accuracy is assessed using standard error-based evaluation metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)—to ensure methodological rigor and comparability across models.

By integrating machine learning methods with inequality analysis, this research contributes to the literature in three important ways. First, it advances methodological research on inequality forecasting by applying nonlinear and deep learning models within the Indian context. Second, it provides comparative empirical evidence on the relative performance of alternative forecasting techniques under a unified evaluation framework. Third, it offers policy-relevant insights by demonstrating how advanced predictive models can facilitate early detection, continuous monitoring, and informed intervention to address income and wealth disparities in a rapidly evolving economy. Overall, the findings aim to support policymakers, researchers, and development practitioners in adopting evidence-based, data-driven strategies to reduce inequality in India.

II Literature Review

Income and wealth inequality has emerged as one of the most persistent socioeconomic challenges globally and in India. A substantial body of literature documents a sharp rise in inequality since the late twentieth century, driven by globalization, technological change, and institutional transformations (Piketty, [8]; Milanovic, [13]). Evidence from the World Inequality Lab shows that India exhibits one of the highest levels of income concentration among major economies, with the top 10% capturing nearly 58% of national income by the mid-2020s, while the bottom 50% receives a disproportionately small share (World Inequality Lab, 2025). These trends contradict the classical Kuznets [1], hypothesis, suggesting that economic growth alone does not guarantee declining inequality.

Long-run reconstructions of income distribution in India reveal a pronounced U-shaped trajectory. Inequality declined in the decades following Independence due to redistributive policies, progressive taxation, and public-sector expansion, but began rising sharply after economic liberalization in the 1990s (Chancel & Piketty, [18]). Recent gains among lower-income groups during the post-pandemic period appear temporary and policy-induced, underscoring the structural nature of inequality persistence (World Inequality Lab, 2026).

Beyond aggregate measures, the literature highlights deep intersectional inequalities across gender, caste, and region. Women's labor-income share in India remains among the lowest globally, while caste-based stratification continues to limit economic mobility despite affirmative action policies (Kabeer, [12]; Deshpande,[19]). Spatial disparities between rural and urban areas further reinforce unequal access to education, formal employment, and capital, indicating that inequality dynamics are multidimensional and nonlinear.

Several structural drivers underpin rising inequality in India. Skill-biased technological change has widened wage differentials, benefiting highly educated and urban workers (Autor et al., [6]; Goldar & Aggarwal, [20]). Increased market concentration, weak labor institutions, and limited fiscal redistribution—reflected in India's relatively low tax-to-GDP ratio—have further amplified income and wealth concentration (Atkinson, [10]; Piketty & Saez,[8]).

Traditional econometric approaches, including time-series and panel models, have provided valuable insights into inequality dynamics but are constrained by assumptions of linearity and structural stability (Banerjee & Duflo, [4]; Ravallion, M.[14]). These limitations are particularly salient in forecasting contexts characterized by structural breaks, policy shocks, and heavy-tailed income distributions. Consequently, recent research increasingly advocates the use of machine learning (ML) techniques, which are better suited to capturing nonlinear relationships and complex interactions in economic data (Mullainathan & Spiess,[15] ; Varian, [21]).

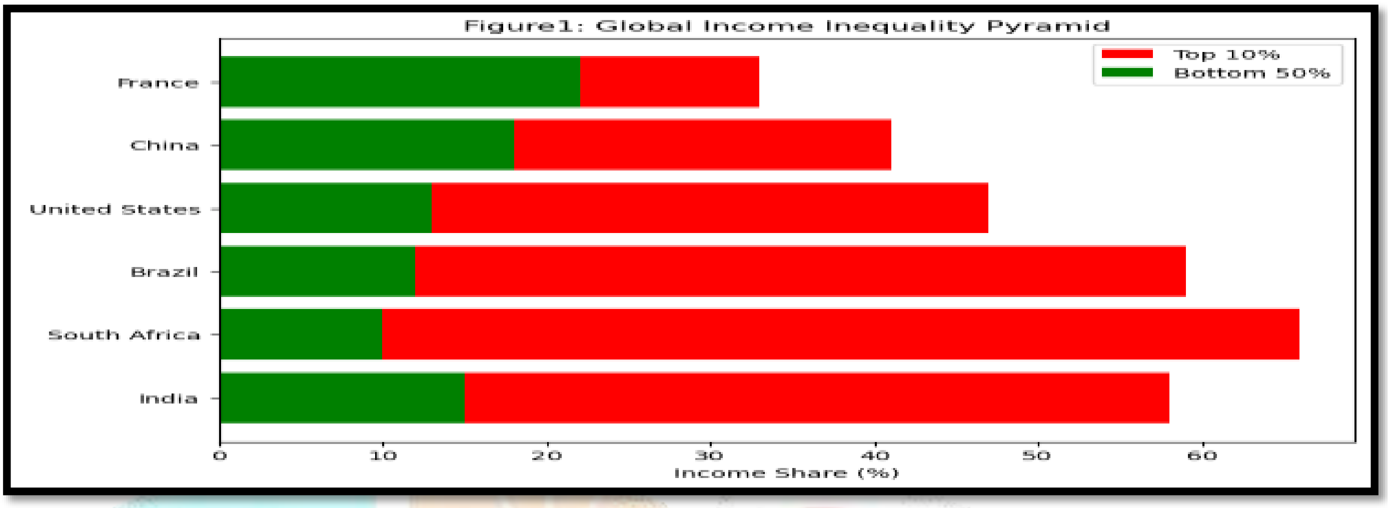
Machine learning models such as Random Forests, Gradient Boosting (XGBoost), Support Vector Regression, and Long Short-Term Memory networks have demonstrated superior predictive performance in macroeconomic and financial forecasting tasks, especially under conditions of instability and regime change (Athey & Imbens, [17]; Biau & Scornet,[11]). However, direct applications of these techniques to forecasting income and wealth inequality—particularly for India—remain limited, with most studies focusing on historical analysis rather than forward-looking projections.

Overall, the literature reveals a clear research gap: the absence of systematic, comparative forecasting studies that apply nonlinear machine learning models to India's income inequality dynamics. Addressing this gap requires integrating long-run distributional data with advanced predictive techniques and robust evaluation metrics. By doing so, future research can contribute both to the methodological literature on inequality forecasting and to evidence-based policy debates on inclusive growth in emerging economies.

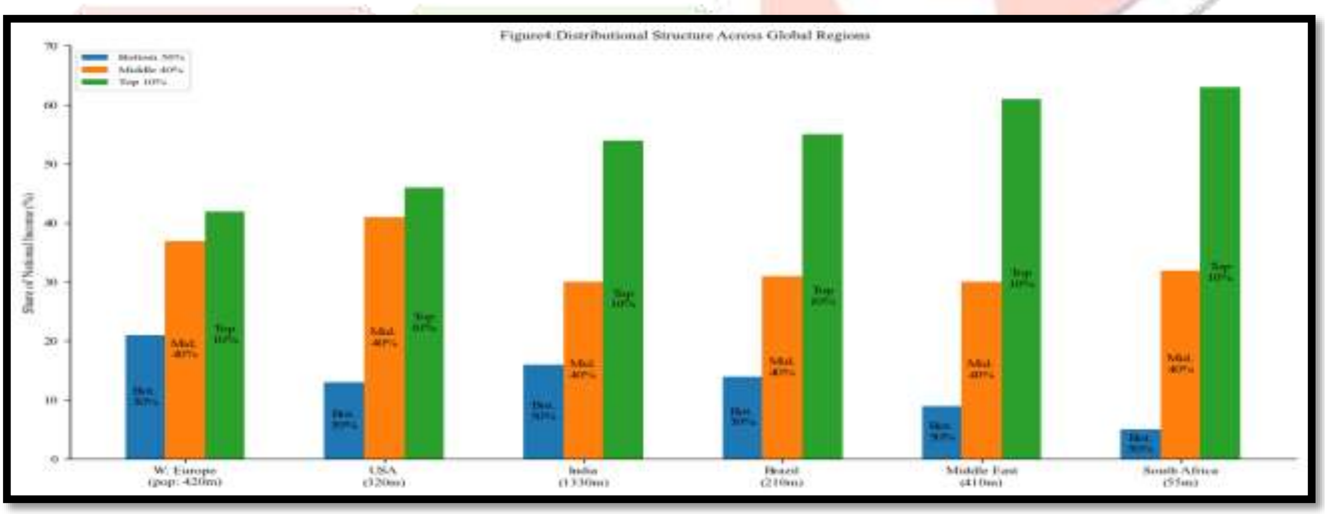
III Data and Cross Country Inequality Context

This study draws on a combination of long-run historical series and contemporary distributional data compiled by the World Inequality Lab2026 [29], supplemented by national-accounts-consistent estimates reported in successive World Inequality Reports (2025).

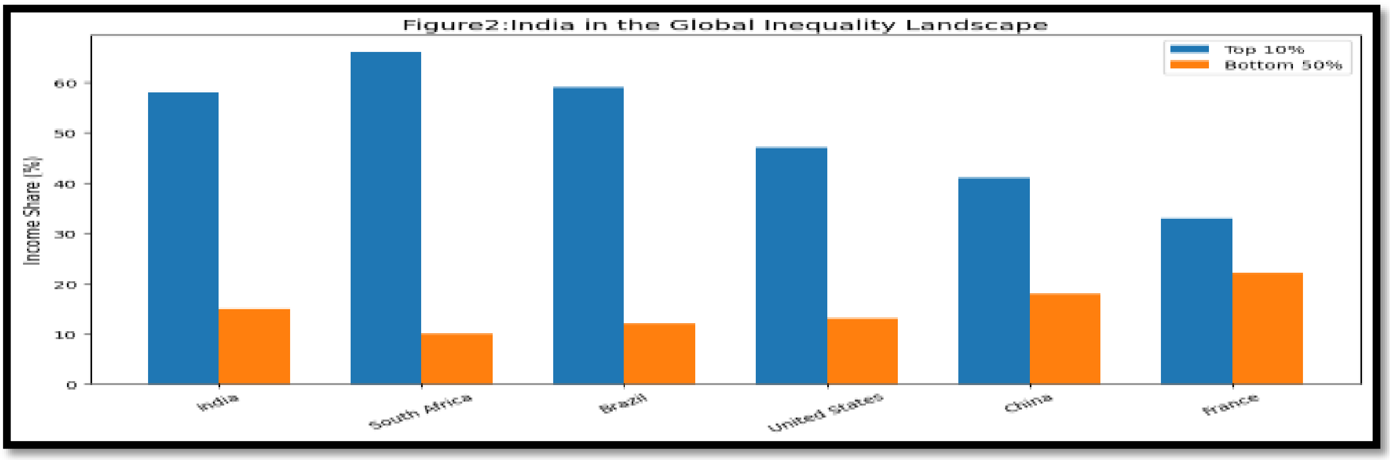
3.1 Data Sources, Construction, and Coverage: The analysis examines how income and wealth are distributed among the top 1 percent, top 10 percent, and bottom 50 percent of the population, measured as shares of total national income or wealth. It relies on harmonized methods combining tax data, national accounts, household surveys, and wealth capitalization to ensure consistency across countries and over time. For India, the dataset covers the early 1920s to 2025, with projections to 2030 using machine-learning models, capturing key phases from colonial inequality to post-liberalization concentration and recent post-pandemic shifts. Cross-country comparisons with economies such as South Africa, Brazil, the United States, China, and France provide a global context for assessing India's inequality trends.



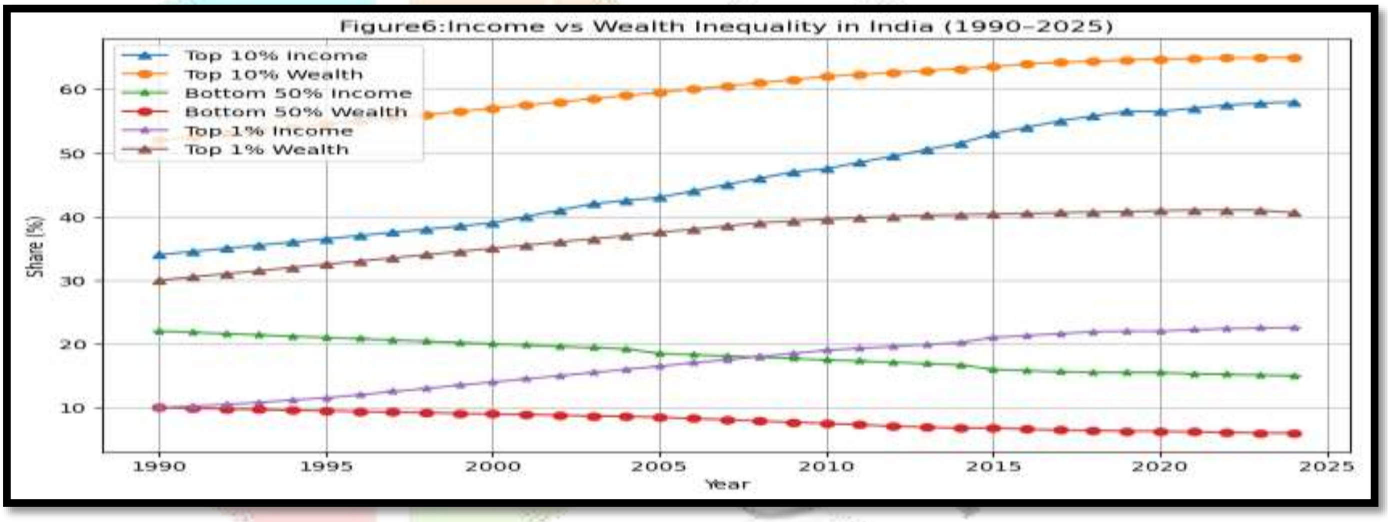
3.2 Distributional Structure across Global Regions: Income distribution varies sharply by region: Western Europe shows a strong middle-income base, while countries like South Africa and India have a compressed middle and a highly concentrated top. In India, the weak middle 40 percent and continued disadvantage of the bottom half highlight deep structural inequality with significant implications for growth, consumption, and savings.



3.3 India in the Global Income Inequality Landscape: India is among the world's most unequal large economies, with the top 10% earning about 58% of national income while the bottom half receives just 15%. Inequality levels are comparable to Brazil and nearing South Africa, and significantly higher than in China, France, and the US. The data spans the 1920s to 2025, covering colonial, post-Independence, and post-liberalization phases, with machine-learning–based projections to 2030. Recent estimates account for post-pandemic changes and harmonized national accounts, allowing reliable cross-country comparisons.

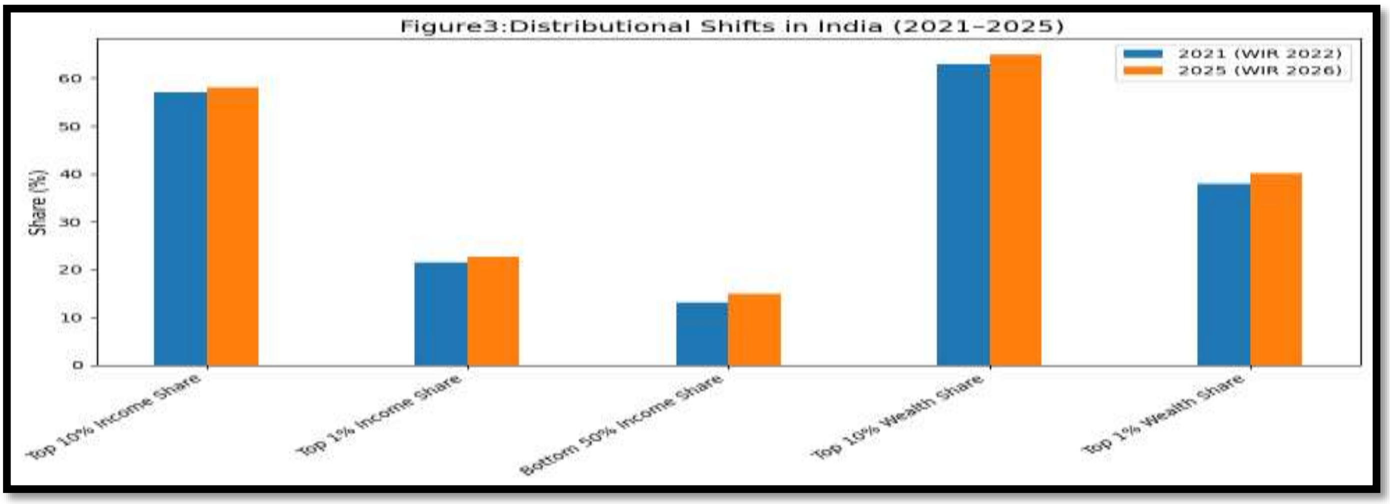


3.4 Income versus Wealth Inequality in India (1990–2025): Figure 6 highlights that income and wealth inequality in India follow nonlinear, regime-dependent patterns. Income inequality declined after Independence but has risen sharply since liberalization, reaching or surpassing colonial-era levels by the 2010s. Wealth inequality remains higher and more persistent than income inequality, accelerating after 1990 due to asset concentration at the top. Since liberalization, the top 1% and 10% have steadily increased their income and wealth shares, while the bottom 50% especially in wealth has seen sustained losses. Overall, the trends indicate growing polarization and entrenched wealth concentration, underscoring the limits of linear models and the need for nonlinear, machine-learning-based forecasting approaches.



3.5 Income and Wealth inequality Distributional in India (2021–2025): Figure 3 shows that income and wealth inequality in India intensified between 2021 and 2025. The income shares of the top 10 percent and top 1 percent increased steadily, indicating that post-pandemic income gains were concentrated among the richest groups. In contrast, the bottom 50 percent experienced only a modest rise in income share, suggesting limited redistribution.

Wealth inequality remains even more pronounced. The wealth shares of both the top 10 percent and top 1 percent rose further over the period, reinforcing persistent asset concentration at the top. Overall, the figure highlights that recent improvements for the lower half of the population have been marginal, while income and wealth gains have been heavily skewed toward the top, underscoring the structural nature of inequality in India and motivating the use of advanced forecasting approaches.



IV Methodology and Models

Five regression models, are evaluated for pattern recognition. Prior studies by Singh, D. P.[22,23,24,25,26,27,28] highlight their importance in numerical prediction, statistical analysis, and decision-making by integrating statistical and machine learning techniques.

4.1. Random Forest (RF): Random Forest is an ensemble learning method that combines many decision trees to improve prediction accuracy and reduce overfitting.

Mathematical Model-Let: Training data: $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots \dots \dots (x_n, y_n)\}$

Random Forest builds T decision trees using:

Bootstrap sampling (random samples with replacement)

Random feature selection at each split

Each tree produces a prediction: $\hat{y}^{(t)}(x), t = 1,2,3 \dots \dots . T$

Final Prediction

Regression: $\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)}(x)$

4.2. ARIMA (Autoregressive Integrated Moving Average): ARIMA is a time-series forecasting model that explains future values using: Past values, Past errors, Differencing for stationarity.

Mathematical Model-ARIMA(p, d, q): p = autoregressive order, d = differencing order, q = moving average order.

AR(p) $y_t = c + \sum_{i=1}^{p} \emptyset_i y_{t-i} + \varepsilon_t$

MA(q) $y_t = c + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$

ARIMA(p, d, q): After differencing d times: $\Delta^d = c + \sum_{i=1}^{p} \emptyset_i \Delta^d y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$

4.3. Support Vector Regression (SVR):SVR finds a function that fits data within an ε-tube, while keeping the model as flat as possible.

Mathematical Model- Function:$f(x) = \omega^T x + b$

Objective:$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$

Subject to: $y_i - (\omega^T x_i + b) \le \varepsilon + \xi_i$

$(\omega^T x_i + b) - y_i \le \varepsilon + \xi_i^*$

$\xi_i, \xi_i^* \ge 0$

Where: = error tolerance, = penalty parameter.

Kernel Trick: $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$

Allows non-linear regression.

4.4. XGBoost (Extreme Gradient Boosting): XGBoost builds models sequentially, where each new tree corrects previous errors using gradient descent.

Mathematical Model-Prediction:$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$, where f_k ∈decision trees.

Objective Function: $\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$

Regularization term:$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda\|\omega\|^2$.

4.5. LSTM (Long Short-Term Memory): LSTM is a recurrent neural network designed to handle long-term dependencies in sequence data.

Mathematical Structure- Let: x_t = input, h_t = hidden state, c_t = cell state

Gates-

Forget Gate: $f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$

Input Gate: $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$

Candidate Memory: $\check{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c)$

Cell State Update: $c_t = f_t \odot c_{t-1} + i_t \odot \check{c}_t$

Output Gate: $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$

Hidden State: $h_t = o_t \odot \tanh(c_t)$.

V Model Evaluation and Performance Metrics

Singh, D. P.[28] highlighted the benefits of Mean Absolute Error (MAE) compared to Root Mean Square Error (RMSE) for assessing the overall performance of models. To assess model accuracy, we use the following metrics:

5.1Mean Squared Error: Mathematical Model of MSE is defined as: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Where: n is Total number of observations. y_i is the actual value. $\hat{y}_i$ is the predicted value. $(y_i - \hat{y}_i)^2$is Squared error for each observation

5.2Mean Absolute Error: The Mean Absolute Error (MAE) is a widely used metric for assessing the performance of regression models. It quantifies the average size of the errors between predicted and actual values, ignoring whether the errors are positive or negative. The mathematical model for MAE is given by: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ where: n is the total count of data points, y_i denotes the actual value of the i-th data point, $\hat{y}_i$ represents the predicted value of the i-th data point, $|y_i - \hat{y}_i|$ signifies the absolute error for each observation.

5.3 R² Score: The R² Score (Coefficient of Determination) is mathematically defined as: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where: SS_{res} (Residual Sum of Squares) is given by: $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

SS_{tot} (Total Sum of Squares) is given by: $SS_{tot} = \sum_{i=1}^{n}(y_i - \overline{y_i})^2$ where: y_i = Actual values of the dependent variable. $\hat{y}_i$ = Predicted values from the regression model. $\overline{y_i}$ = Mean of the actual values. n = Number of data points.

5.4 Root Mean Squared Error (RMSE) $RMSE = \sqrt{MSE}$

VI Results and Discussion

This section reports and interprets the empirical findings from forecasting income and wealth inequality in India using five competing models ARIMA, Random Forest, Support Vector Regression (SVR), XGBoost, and Long Short-Term Memory (LSTM). Model performance is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), with lower values indicating superior forecasting accuracy.

6.1 Model Performance in Forecasting Income Inequality: The forecasting results reveal substantial variation in model performance across income inequality measures, reflecting the heterogeneous dynamics of income distribution segments.

6.1.1 Income Share of the Top 1 Percent: For the income share of the top 1 percent, the LSTM model consistently outperforms all competing approaches, achieving the lowest RMSE (0.0165) and MAE (0.0107). This superior performance indicates that income concentration at the extreme upper tail follows strongly nonlinear and path-dependent dynamics, which are more effectively captured by deep learning architectures than by traditional econometric or kernel-based models. The relatively weak performance of ARIMA and SVR underscores their limited capacity to model such complex temporal dependencies.

Table1: Forecasting Income_Top1_pct

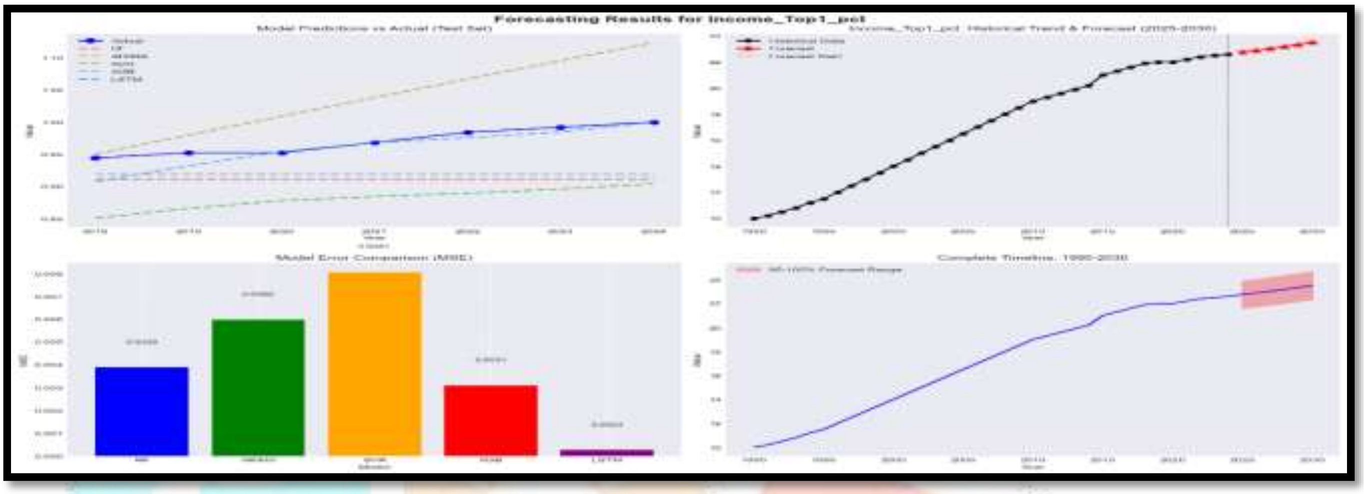
Evaluation Results for Income_Top1_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.003891	0.062381	0.059014
1	ARIMA	0.005975	0.077300	0.067196
2	SVR	0.008064	0.089802	0.089485
3	XGBoost	0.003097	0.055650	0.051847
4	LSTM	0.000271	0.016455	0.010748

Best model for Income_Top1_pct: LSTM (Lowest RMSE)

Forecast for Income_Top1_pct (2025-2030):

Year	Income_Top1_pct_Forecast
2025	22.762041
2026	22.883528
2027	23.022299
2028	23.181797
2029	23.332251
2030	23.499598



6.1.2 Income Share of the Top 10 Percent: For the income share of the top 10 percent, XGBoost delivers the highest forecasting accuracy, with an RMSE of 0.0917. The comparatively poorer performance of LSTM (RMSE = 0.2025) suggests that deep neural networks may be less effective when the underlying series exhibits moderate volatility and limited nonlinear structure. This finding highlights the advantage of tree-based ensemble methods in balancing flexibility with generalizability.

Table2: Forecasting Income_Top10_pct

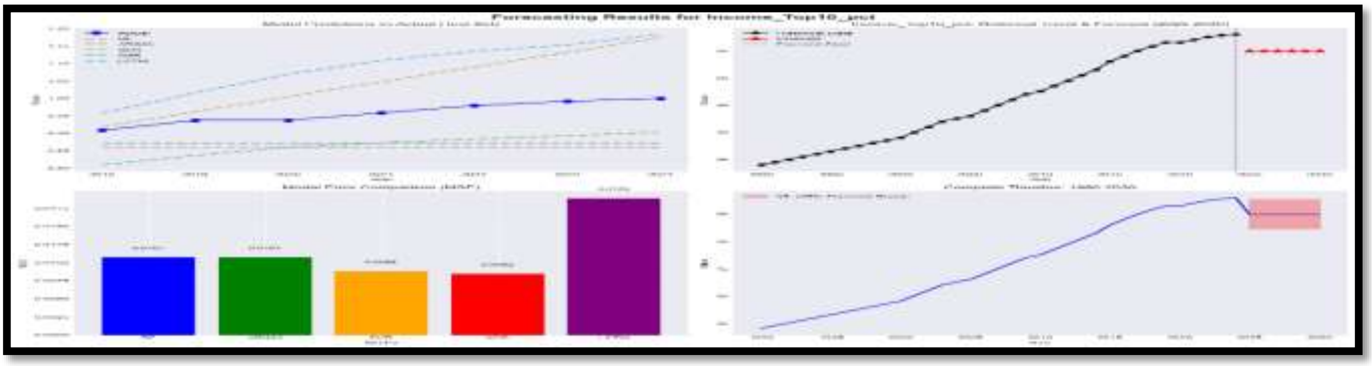
Evaluation Results for Income_Top10_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.010736	0.103617	0.098929
1	ARIMA	0.010724	0.103555	0.087831
2	SVR	0.008757	0.093580	0.093243
3	XGBoost	0.008401	0.091657	0.086322
4	LSTM	0.041025	0.202547	0.195703

Best model for Income_Top10_pct: XGBoost (Lowest RMSE)

Forecast for Income_Top10_pct (2025-2030):

Year	Income_Top10_pct_Forecast
2025	54.942558
2026	54.942558
2027	54.942558
2028	54.942558
2029	54.942558
2030	54.942558



6.1.3 Income Share of the Bottom 50 Percent: For the income share of the bottom 50 percent, XGBoost again emerges as the best-performing model (RMSE = 0.0527). This result reflects the model's ability to capture subtle nonlinearities while remaining robust to noise and structural stability at the lower end of the income distribution.

Table3: Forecasting Income_Bottom50_pct

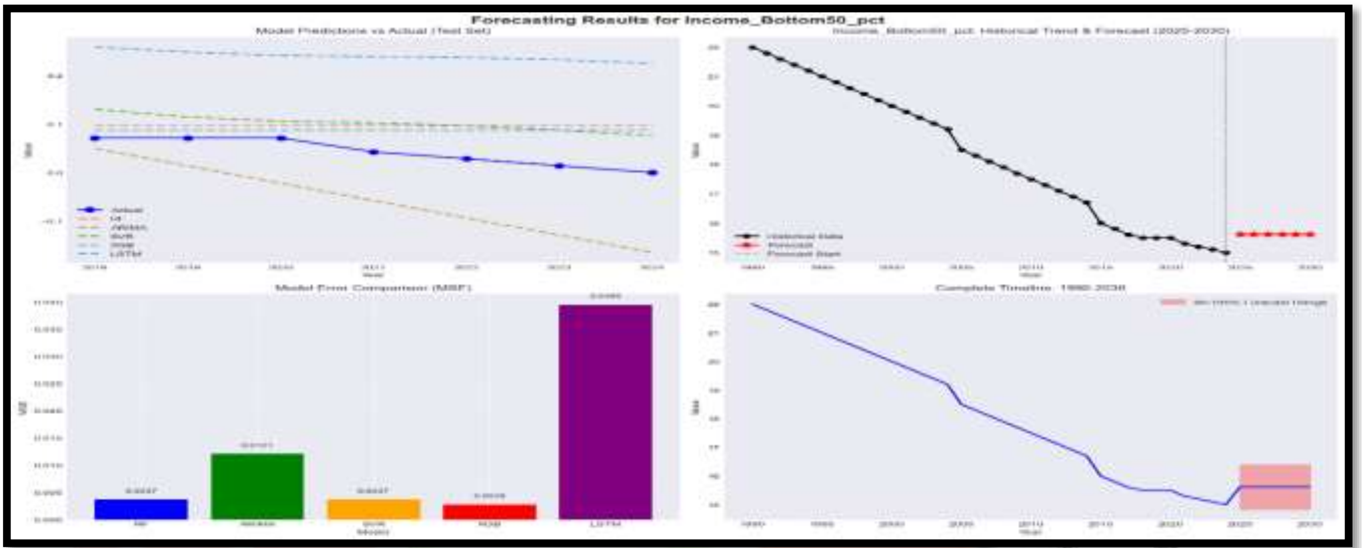
Evaluation Results for Income_Bottom50_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.003705	0.060868	0.054286
1	ARIMA	0.012132	0.110145	0.100314
2	SVR	0.003705	0.060869	0.059167
3	XGBoost	0.002782	0.052749	0.044994
4	LSTM	0.039470	0.198671	0.197729

Best model for Income_Bottom50_pct: XGBoost (Lowest RMSE)

Forecast for Income_Bottom50_pct (2025-2030):

Year	Income_Bottom50_pct_Forecast
2025	15.614959
2026	15.614959
2027	15.614959
2028	15.614959
2029	15.614959
2030	15.614959



Taken together, these results indicate that nonlinear machine learning models dominate income inequality forecasting, with LSTM being most effective for highly concentrated income segments and XGBoost performing best for broader population groups.

6.2 Model Performance in Forecasting Wealth Inequality: Wealth inequality exhibits markedly different statistical properties from income inequality, leading to a distinct pattern of model performance.

6.2.1 Wealth Share of the Top 1 Percent: For the wealth share of the top 1 percent, ARIMA outperforms all machine learning models, achieving the lowest RMSE (0.0240). This finding suggests that wealth concentration at the very top evolves through relatively stable and linear autoregressive processes, where traditional time-series models retain a strong comparative advantage.

Table4: Forecasting Wealth_Top1_pct

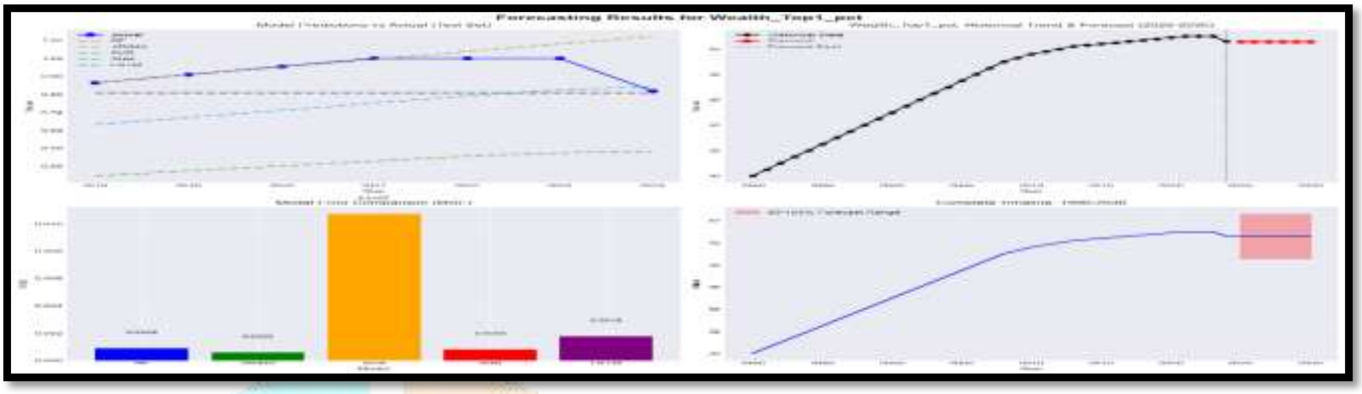
Evaluation Results for Wealth_Top1_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.000896	0.029929	0.026649
1	ARIMA	0.000575	0.023988	0.012375
2	SVR	0.010712	0.103500	0.102437
3	XGBoost	0.000800	0.028284	0.024789
4	LSTM	0.001757	0.041912	0.039237

Best model for Wealth_Top1_pct: ARIMA (Lowest RMSE)

Table: Forecast for Wealth_Top1_pct (2025-2030):

Year	Wealth_Top1_pct_Forecast
2025	40.584469
2026	40.584469
2027	40.584469
2028	40.584469
2029	40.584469
2030	40.584469



6.2.2 Wealth Share of the Top 10 Percent: For the wealth share of the top 10 percent, however, LSTM provides the most accurate forecasts (RMSE = 0.0404). This result indicates the presence of nonlinear accumulation effects and long-memory dynamics in upper-wealth trajectories that are better captured by recurrent neural networks.

Table5: Forecasting Wealth_Top10_pct

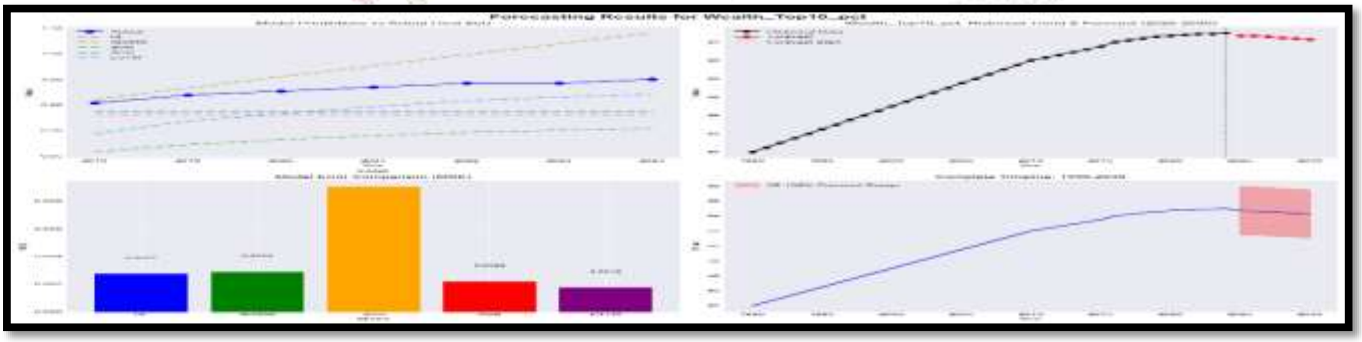
Evaluation Results for Wealth_Top10_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.002742	0.052360	0.050242
1	ARIMA	0.002872	0.053594	0.044935
2	SVR	0.009020	0.094974	0.094961
3	XGBoost	0.002202	0.046931	0.044555
4	LSTM	0.001636	0.040445	0.039023

Best model for Wealth_Top10_pct: LSTM (Lowest RMSE)

Forecast for Wealth_Top10_pct (2025-2030):

Year	Wealth_Top10_pct_Forecast
2025	64.702194
2026	64.687042
2027	64.635315
2028	64.481695
2029	64.433914
2030	64.353081



6.2.3 Wealth Share of the Bottom 50 Percent: In the case of the wealth share of the bottom 50 percent, ARIMA again delivers the best performance (RMSE = 0.0410), implying that wealth changes among lower-wealth households are slow-moving, persistent, and dominated by autoregressive behavior.

Table6: Forecasting Wealth_Bottom50_pct

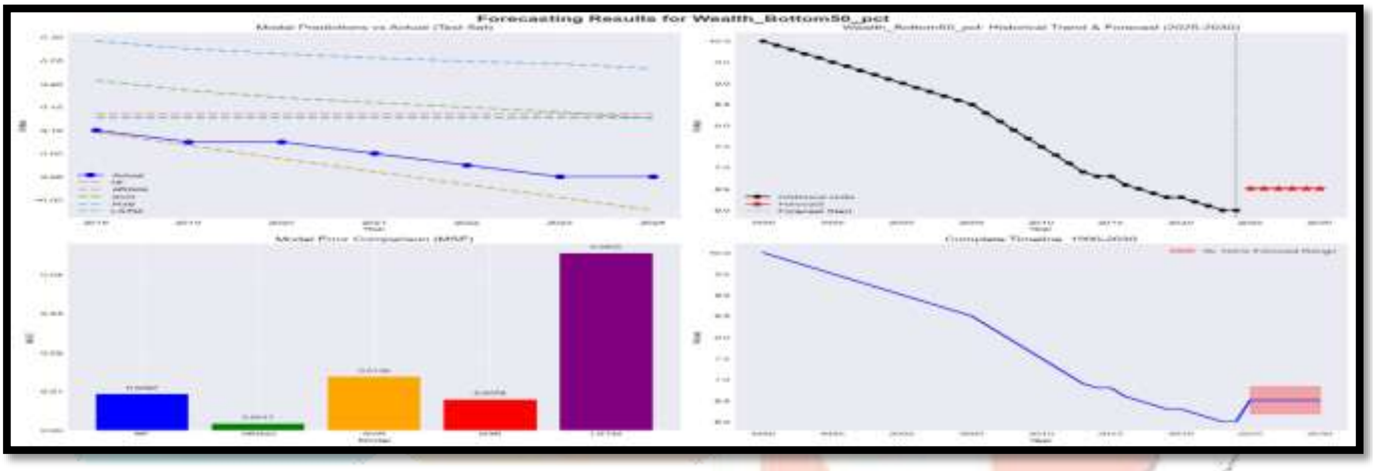
Evaluation Results for Wealth_Bottom50_pct:

	Model	MSE	RMSE	MAE
0	Random Forest	0.009216	0.095999	0.088821
1	ARIMA	0.001680	0.040983	0.034991
2	SVR	0.013633	0.116762	0.115988
3	XGBoost	0.007819	0.088423	0.080574
4	LSTM	0.045514	0.213341	0.212468

Best model for Wealth_Bottom50_pct: ARIMA (Lowest RMSE)

Table: Forecast for Wealth_Bottom50_pct (2025-2030):

Year	Wealth_Bottom50_pct_Forecast
2025	6.50801
2026	6.50801
2027	6.50801
2028	6.50801
2029	6.50801
2030	6.50801



Overall, the wealth inequality results highlight a greater degree of structural rigidity and persistence, reducing the relative gains from complex nonlinear models in several cases.

6.3 Overall Model Comparison and Robustness: To assess overall robustness, Table 7(a) reports average forecasting performance across all income and wealth indicators. Based on mean RMSE, MAE, and MSE, XGBoost emerges as the best overall-performing model, with the lowest average RMSE (0.0606). Random Forest and ARIMA follow closely, while SVR and LSTM exhibit weaker average performance due to inconsistent accuracy across indicators.

These findings imply that, although no single model dominates every inequality measure, XGBoost provides the most reliable and stable forecasting performance across heterogeneous distributional outcomes, making it particularly suitable for applied inequality forecasting.

Table7(a): OVERALL MODEL PERFORMANCE COMPARISON

Average Performance Across All Variables:

Model	RMSE	MAE	MSE
XGBoost	0.0606	0.0555	0.0042
Random Forest	0.0675	0.0630	0.0052
ARIMA	0.0683	0.0579	0.0057
SVR	0.0932	0.0925	0.0090
LSTM	0.1097	0.1071	0.0180

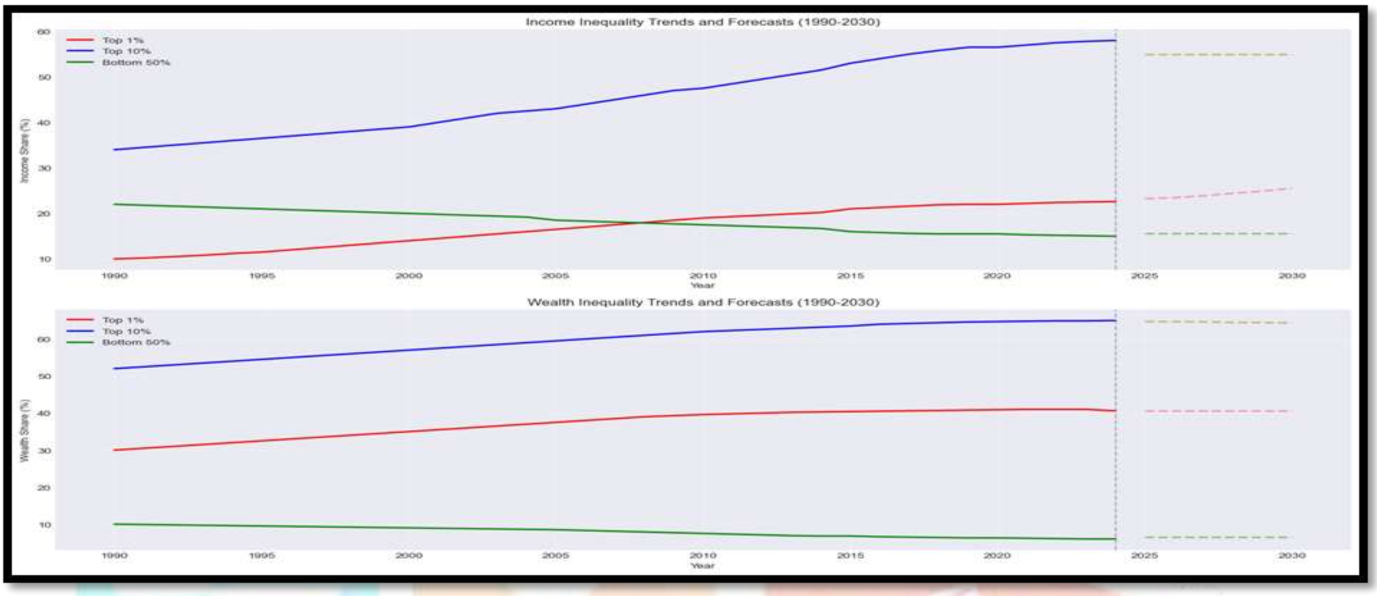
Overall Best Model: XGBoost
Average RMSE: 0.0606

Table7(b):SUMMARY FORECAST FOR 2030

Indicator	2024 Value	2030 Forecast	Change (%)
Income_Top1_pct	22.60	25.53	12.95
Income_Top10_pct	58.00	54.94	-5.27
Income_Bottom50_pct	15.00	15.61	4.10
Wealth_Top1_pct	40.60	40.58	-0.04
Wealth_Top10_pct	65.00	64.35	-1.00
Wealth_Bottom50_pct	6.00	6.51	8.47

6.4 Forecasted Income and Wealth Inequality Trends: The projections for 2025–2030 show rising income concentration at the very top, with the top 1 percent increasing its income share significantly, while changes elsewhere are modest. The top 10 percent sees a slight decline, reflecting redistribution within the elite rather than toward lower-income groups, and the bottom 50 percent gains only marginally. Overall, income inequality becomes more polarized.

Wealth inequality remains even more rigid. The top 1 percent's wealth share is essentially unchanged, the top 10 percent declines slightly, and the bottom 50 percent sees a relative increase. However, because the bottom half still holds very little wealth in absolute terms, overall wealth inequality remains largely unchanged, highlighting its strong structural persistence.



6.5 Discussion and Policy Implications: From a methodological perspective, the results demonstrate that model suitability depends critically on the underlying distributional dynamics. Nonlinear machine learning models significantly improve forecasting accuracy for income inequality, while traditional econometric approaches remain highly effective for stable wealth series. Among all models, XGBoost offers the best balance between flexibility, robustness, and generalizability.

From an economic perspective, the forecasts raise important concerns. Income inequality in India is projected to worsen at the extreme upper tail, while wealth inequality remains deeply entrenched, with only marginal improvements for the bottom half of the population. These findings underscore the need for targeted redistributive policies and long-term wealth-building strategies, particularly for lower-income and lower-wealth households.

VII Conclusion

This study demonstrates the value of integrating nonlinear machine learning techniques with traditional econometric models for forecasting income and wealth inequality. While no single model dominates across all indicators, XGBoost emerges as the most reliable overall forecasting tool, with LSTM and ARIMA serving as complementary models for specific segments of the distribution. The projected trends point to persistent and in some cases worsening inequality in India, highlighting the urgency of evidence-based policy interventions aimed at addressing deep-rooted income and wealth disparities.

References

[1].Kuznets, S. (1955). Economic growth and income inequality. American Economic Review, 45(1), 1–28.

[2].Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.https://doi.org/10.1162/neco.1997.9.8.1735

[3].Deaton, A., & Dreze, J. (2002). Poverty and inequality in India: A re-examination. Economic and Political Weekly, 37(36), 3729–3748.

[4].Banerjee, A., & Duflo, E. (2003). Inequality and growth: What can the data say? Journal of Economic Growth, 8(3), 267–299.

[5].Banerjee, A., & Piketty, T. (2005). Top Indian incomes, 1922–2000. The World Bank Economic Review, 19(1), 1–20.https://doi.org/10.1093/wber/lhi001

[6].Autor, D., Katz, L., & Kearney, M. (2008). Trends in U.S. wage inequality. Review of Economics and Statistics, 90(2), 300–323.

[7].Datt, G., & Ravallion, M. (2011). Has India's economic growth become more pro-poor in the wake of economic reforms? The World Bank Economic Review, 25(2), 157–189.

[8].Piketty, T. (2014). Capital in the twenty-first century. Harvard University Press.

[9].Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3–28.https://doi.org/10.1257/jep.28.2.3

[10].Atkinson, A. B. (2015). Inequality: What can be done? Harvard University Press.

[11].Biau, G., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227.

[12].Kabeer, N. (2016). Gender equality, economic growth, and women's agency. Feminist Economics, 22(1), 295–321.

[13].Milanovic, B. (2016). Global inequality: A new approach for the age of globalization. Harvard University Press.

[14] Ravallion, M. (2016). The economics of poverty: History, measurement, and policy. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190212766.001.0001

[15].Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106.

[16].Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (2018). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62.

[17].Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

[18].Chancel, L., & Piketty, T. (2019). Indian income inequality, 1922–2015: From British Raj to Billionaire Raj? Review of Income and Wealth, 65(S1), S33–S62. https://doi.org/10.1111/roiw.12439

[19].Deshpande, A. (2019). The grammar of caste. Oxford University Press.

[20].Goldar, B., & Aggarwal, S. C. (2019). Trade liberalisation, labour regulations and unionisation in Indian manufacturing: Implications for labour share, wage gap and emolument inequality. The Indian Journal of Labour Economics, 62(3), 441–461. https://doi.org/10.1007/s41027-019-00179-4

[21].Varian, Hal R. 2019. "Artificial Intelligence, Economics, and Industrial Organization." In The Economics of Artificial Intelligence: An Agenda, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press.

[22].Singh, D P Jassi J S, Sunaina(2023). Exploring the Significance of Statistics in the Research: A Comprehensive Overview, Eur. Chem. Bull. (ISSN 2063-5346), 2023; 12(2): 2089-2102, doi: 10.31838/ECB/2023.12.s2.262.

[23].Singh, D. P. (2024). An extensive analysis of machine learning models to predict breast cancer recurrence. Tuijin Jishu/Journal of Propulsion Technology, 45(2).

[24].Singh, D. P. (2024). An extensive analysis of machine learning techniques for predicting the onset of lung cancer. Tuijin Jishu/Journal of Propulsion Technology, 45(4).

[25].Singh, D P(2024). An Extensive Examination of Machine Learning Methods for Identifying Diabetes, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055, 2024; 45: 2.

[26].Singh, D. P. (2024). An extensive analysis of machine learning techniques for predicting the onset of lung cancer. Tuijin Jishu/Journal of Propulsion Technology, 45(4).

[27].Singh, D. P. (2024).Comprehensive analysis of machine learning models for cardiovascular disease detection and diagnosis. Tuijin Jishu/Journal of Propulsion Technology, 45(4).

[28].Singh, D P (2025). Exploring Machine Learning-Driven Advanced Regression Models For Predicting Prime Number Distributions, IJCRT | Volume 13, Issue 4, | ISSN: 2320-2882,b229-b239

[29].World Inequality Lab. (2026). World inequality report 2026. https://wid.world