IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3464~3475
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3464-3475

Forecasting income inequality in Vietnam's regions using machine learning

T. Thai-Phuong, L. Nguyen-Son
Faculty of Civil Engineering, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Sep 13, 2025
Revised May 28, 2026
Accepted Jul 9, 2026

Keywords:

Artificial neural networks
Gini coefficient
Income inequality
Machine learning
Regional analysis
Socio-economic policy
Support vector machine

ABSTRACT

This study applies artificial neural networks (ANN) and support vector regression (SVR) to forecast the Gini coefficient across Vietnam's seven socio-economic regions using limited data from 2014–2022 (N = 63). Despite the small dataset, robust techniques including repeated cross-validation, temporal backtesting, regularization, and bootstrapping were employed to mitigate overfitting. The results show a modest continued decline in income inequality through 2025, with faster improvements in urban regions (Red River Delta and Southeast) than in rural and highland areas. ANN outperformed SVR with an average cross-validation coefficient of determination (R^2) of 0.93. The novelty lies in its granular regional forecasting and region-specific policy recommendations beyond national-level analyses. Key drivers—per capita income and simple housing—support targeted interventions such as housing credits in the Mekong River Delta and agricultural investments in the Central Highlands. The findings provide an evidence-based foundation for regionally tailored strategies to reduce income disparities in Vietnam.

Corresponding Author:

L. Nguyen-Son
Faculty of Civil Engineering, Industrial University of Ho Chi Minh City
12 Nguyen Van Bao, Hanh Thong Ward, Ho Chi Minh City, Vietnam
Email: nguyensonlam@iuh.edu.vn

1. INTRODUCTION

Income inequality remains a persistent socio-economic challenge in Vietnam, despite remarkable progress in poverty reduction and economic growth [1]. Significant regional disparities arise from variations in economic structure, urbanization rates, agricultural dependence, and vulnerability to environmental factors such as climate change [2]. Vietnam comprises seven key socio-economic regions—the Central Highlands, Mekong River Delta, North Central and Central Coast, Northern Midlands and Mountains, Red River Delta, Southeast, and the national aggregate—each exhibiting distinct patterns of income distribution that demand nuanced, region-specific analysis. The Gini coefficient, a standard measure of inequality, has shown relative stability nationally in recent years [3]. While prior studies have contributed to national-level insights, they often overlook regional heterogeneity, which limits the identification of localized drivers and the formulation of targeted policies.

Traditional linear regression models often fail to capture the complex, nonlinear interactions among factors, such as per capita income and housing conditions, especially in rural and agriculturally dominant regions like the Mekong River Delta and Central Highlands [4]. This study introduces a novel machine learning approach, employing artificial neural networks (ANN) and support vector regression (SVR), to address these shortcomings by modeling nonlinear relationships in limited historical data (2014–2022). ANN is particularly effective for heterogeneous and small datasets due to its multi-layered architecture, while

SVR provides a robust comparative baseline using the radial basis function (RBF) kernel. This study acknowledges the constraint of a relatively small dataset (N = 63). To ensure reliability, multiple robustness techniques such as repeated 10-fold cross-validation, temporal backtesting, L2 regularization, dropout, early stopping, and bootstrapping were employed.

The novelty of this work lies in its granular regional focus across Vietnam's seven socio-economic regions, which provides more actionable policy insights than aggregated national-level analyses. In addition, the study demonstrates the effective application of ANN to capture nonlinear patterns in small datasets (N = 63), supported by robust validation techniques (10-fold cross-validation, regularization, dropout, early stopping, and temporal backtesting) to mitigate overfitting risks. The primary objectives are threefold: i) to forecast Gini coefficients for 2023–2025 across the seven regions, projecting modest continued changes consistent with recent national trends; ii) to elucidate the roles of key drivers, particularly monthly per capita income and the proportion of simple housing; and iii) to propose tailored regional policy interventions. Amid ongoing challenges like post-pandemic recovery and climate impacts, this regionalized machine learning framework provides a timely scientific foundation for equitable and sustainable development strategies in Vietnam.

Income inequality is a central issue in economics and the social sciences, particularly in developing countries like Vietnam, where urban-rural disparities arise from complex factors, including urbanization, agricultural dependence, education, housing conditions, and access to infrastructure [5]. Studies in China and Southeast Asia highlight how urbanization and investments in agriculture reduce urban-rural income gaps in developed regions, while socio-economic barriers like limited services exacerbate disparities elsewhere [6]. Multidimensional inequality research extends beyond income to include education and housing [7]. Analyses in India and Bangladesh show regional gaps in early childhood education and child nutrition, underscoring that national-level studies often miss local heterogeneity and hinder precise policy targeting [8].

In socio-economic data analysis, machine learning techniques—especially ANN—effectively model nonlinear relationships [9]. Recent applications include forecasting Gini coefficients in Malaysia using nonlinear autoregressive multilayer perceptrons and predicting poverty with machine learning in various developing contexts [10]. However, ensemble methods (e.g., random forests, extreme gradient boosting (XGBoost)) often exhibit high variance and instability on small datasets due to overfitting risks [11]. Therefore, this study focuses on ANN and SVR, which demonstrated better stability during preliminary experiments. These studies reveal gaps in regional applications of advanced methods to inequality forecasting [12]. The current study addresses these gaps by focusing on sub-national heterogeneity across Vietnam's seven socio-economic regions, employing ANN optimized for limited data, and comparing it systematically with SVR as well as other baselines to demonstrate its relative advantage.

2. DATA AND METHODOLOGY

2.1. Data description

This study utilizes historical data from 2014 to 2022, sourced from the Vietnam household living standards survey (VHLSS) conducted by the General Statistics Office (GSO) of Vietnam, to forecast the Gini coefficient across Vietnam's seven socio-economic regions: Central Highlands, Mekong River Delta, North Central and Central Coast, Northern Midlands and Mountains, Red River Delta, Southeast, and the national level [13]. The data elucidates key socio-economic factors influencing inequality and justifies the selection of two independent variables—monthly per capita income (in USD) and the proportion of simple housing—for the machine learning models [14]. The proportion of simple housing refers to the percentage of households living in *nhà tạm* (temporary) or *nhà bán kiên cố* (semi-permanent) structures (classified by GSO in VHLSS as temporary housing or semi-permanent housing, based on construction materials, durability, and lack of basic amenities, assessed through nationwide household surveys) [15].

It should be noted that the dataset consists of only 63 observations (7 regions × 9 years), which is relatively small for machine learning tasks. This limitation is explicitly addressed through multiple mitigation strategies described in section 2.3.4. Variable selection was rigorously conducted. To ensure robustness of variable selection, sensitivity analyses were conducted by incrementally adding other candidate variables (urbanization rate, proportion of trained labor, and school enrollment rate). These additions did not yield significant improvement in cross-validation performance ($\Delta R^2 < 0.03$) and introduced multicollinearity (variance inflation factor or VIF > 5), confirming that the two selected variables provide an optimal balance between model performance, interpretability, and policy relevance. Thus, only monthly per capita income and the proportion of simple housing were retained for enhanced model interpretability, robustness on small datasets, and clearer policy implications [16]. To provide detailed insights into the statistical indicators, the data is presented in Table 1.

Table 1. Summary statistics of Gini coefficients and socioeconomic indicators by region (2014–2022)

Region	Gini coefficient				Monthly Per capita income (USD)				Simple housing (%)			
	Mean	Min	Max	Std Dev	Mean	Min	Max	StdDev	Mean	Min	Max	Std Dev
Central Highlands	0.430	0.399	0.449	0.019	123.416	90.550	157.540	22.890	0.900	0.700	1.200	0.141
Mekong River Delta	0.391	0.352	0.412	0.023	138.970	107.030	173.180	24.891	5.433	1.900	10.20	3.000
North Central and Central Coast	0.377	0.351	0.395	0.017	134.641	102.650	167.800	23.209	1.056	0.300	1.900	0.590
Northern Midlands and Mountains	0.437	0.408	0.453	0.015	108.767	77.340	136.950	20.698	4.733	3.000	7.200	1.509
Red River Delta	0.377	0.317	0.413	0.038	228.679	175.710	284.530	37.804	0.056	0.000	0.100	0.053
Southeast	0.353	0.291	0.387	0.034	273.769	222.880	324.870	35.534	0.611	0.300	1.200	0.333
Whole country	0.410	0.373	0.435	0.030	161.306	125.240	197.880	24.967	1.933	0.900	3.500	0.938

To visualize the variation in the Gini coefficient over the years, the data is illustrated in Figure 1. Figure 1 illustrates annual Gini coefficient trends across Vietnam's seven socio-economic regions from 2014 to 2022 using bar charts. Overall, inequality shows a declining trend, with sharp reductions in the urbanized Red River Delta and Southeast due to urbanization, contrasted by slower declines and greater fluctuations in the Central Highlands and Mekong River Delta, highlighting regional heterogeneity and the necessity of machine learning models to capture nonlinear patterns.

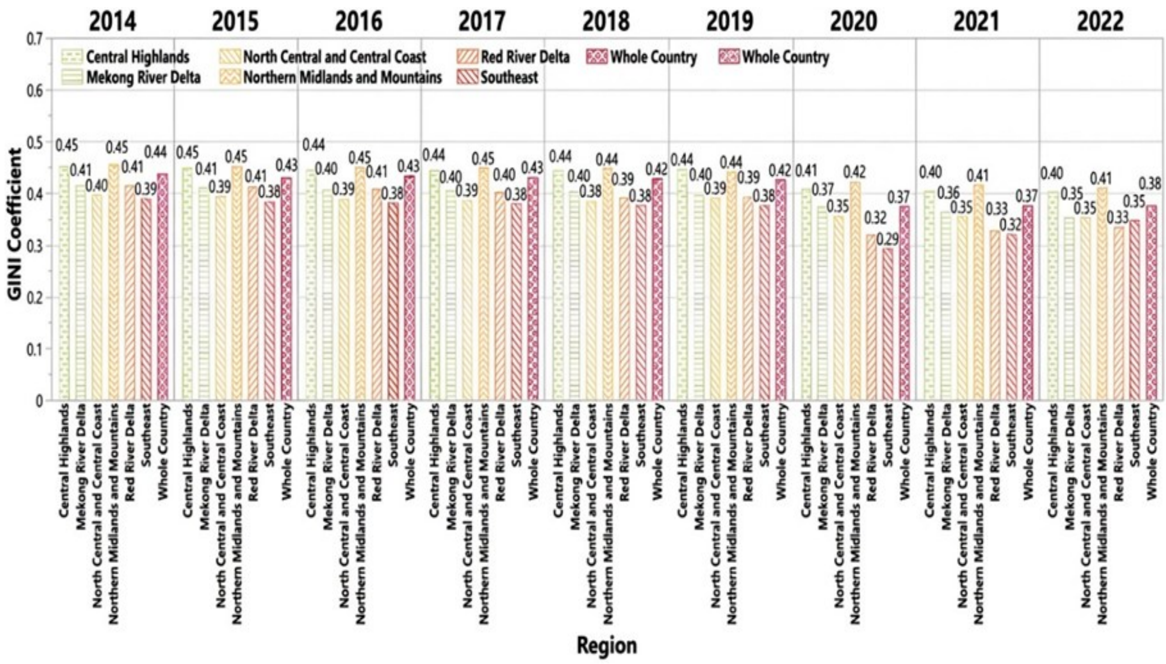

Figure 1. Trends in Gini coefficient across regions (2014–2022)

Table 1 and Figure 1 reveal significant regional disparities in income inequality. The Northern Midlands and Mountains exhibit the highest mean Gini coefficient (0.437), followed by the Central Highlands (0.430), while the Southeast has the lowest (0.353). Monthly per capita income is markedly higher in urbanized regions like the Southeast (273.77 USD) and Red River Delta (228.68 USD) compared to rural areas such as the Northern Midlands and Mountains (108.77 USD). The proportion of simple housing is elevated in the Mekong River Delta (5.43%), contrasting with near-zero levels in the Red River Delta (0.06%). Standard deviations indicate greater variability in income and housing in rural regions like the Mekong River Delta. Figure 1 shows a general downward trend in Gini coefficients across all regions from 2014 to 2022, with sharper declines in urbanized areas (Red River Delta and Southeast) and slower, more fluctuating reductions in rural and highland regions, underscoring pronounced heterogeneity and nonlinear patterns that necessitate machine learning approaches. A supplementary Excel file with the full dataset (2014–2022), including raw observations for all 63 data points, and a detailed data dictionary is provided as supplemental material to ensure transparency and reproducibility of the results. This regional analysis provides a nuanced foundation for forecasting, superior to national aggregates. The subsequent section details correlation analysis to further justify variable selection.

2.2. Correlation analysis

This section examines the relationships between the Gini coefficient, monthly per capita income, and the proportion of simple housing to justify their use as input variables in the machine learning models (ANN and SVR) for forecasting income inequality across Vietnam's seven socio-economic regions and the national level. The analysis is based on historical data from 2014 to 2022 (N = 63 observations: 7 regions × 9 years). It employs both linear and nonparametric correlation measures to assess the strength and nature of these relationships [17].

Three correlation coefficients are used:

i) Pearson correlation coefficient, as defined in (1) [18].

$$r_1 = \frac{\sum_{i=1}^{n}(x_i-x)(y_i-y)}{\sqrt{\sum_{i=1}^{n}(x_i-x)^2\sum_{i=1}^{n}(y_i-y)^2}} \tag{1}$$

Where x_i and y_i are the values of the two variables, x and y are their means, and n is the number of observations.

ii) Spearman correlation coefficient, as defined in (2).

$$r = 1 - \frac{6\sum d_i^2}{N(N^2-1)} \tag{2}$$

Where d_i is the difference between the ranks of corresponding values of the two variables.

iii) Kendall correlation coefficient, as defined in (3).

$$r_3 = \frac{2}{N(N-1)}\sum_{i<j} sgn((x_i - x_j)(y_i - y_j)) \tag{3}$$

Where sgn is the sign function.

To provide an overview of the linear relationships, the results are presented in Table 2. To further visualize these relationships, the data is presented in Figure 2. Figure 2 is a heatmap visualizing Pearson correlations among the Gini coefficient, per capita income, and proportion of simple housing (2014–2022). Blue shades indicate negative correlations (e.g., income–Gini: –0.64) and red shades positive correlations (e.g., simple housing–Gini: 0.45), confirming the relevance of these variables as model inputs and highlighting nonlinear elements in the data.

Table 2. Correlation matrix of income, simple housing, and Gini (2014–2022)

Variable	Gini coefficient	Monthly per capita income (USD)	Simple housing percent
Gini coefficient	1.00	-0.64	0.45
Monthly per capita income (USD)	-0.64	1.00	-0.62
Simple housing percent	0.45	-0.62	1.00

As illustrated in Figure 2, the correlations highlight stronger nonlinear patterns in rural regions (e.g., Central Highlands and Mekong River Delta), explaining why machine learning approaches, particularly ANN, are more suitable than linear models for these areas. Table 2 and Figure 2 show a moderate negative Pearson correlation between the Gini coefficient and per capita income (–0.64), indicating that higher income is associated with lower inequality. A moderate positive correlation exists between the Gini coefficient and the proportion of simple housing (0.45), suggesting that poorer housing conditions contribute to greater inequality. The strong negative correlation between income and simple housing (–0.62) confirms that rising income tends to reduce the prevalence of simple housing. Nonparametric Spearman and Kendall coefficients (not shown in the table) yield similar patterns but reveal slight deviations in some regions with high variability (e.g., Mekong River Delta), confirming the presence of nonlinear components. These findings strongly support the selection of per capita income and proportion of simple housing as the most relevant input variables for the machine learning models, particularly justifying the use of ANN to capture complex nonlinear interactions. The next section describes the methodology employed for model development and evaluation.

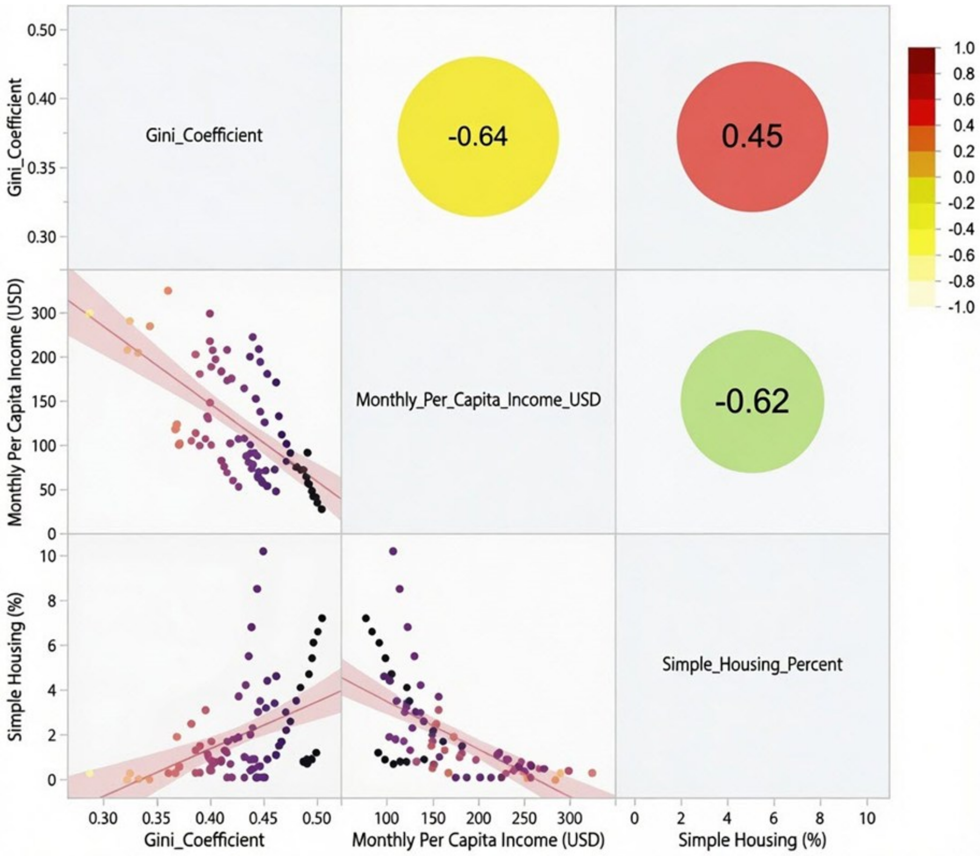

Figure 2. Correlation temperature chart of income, modest housing and Gini (2014–2022)

2.3. Methodology

This section describes the machine learning techniques and data processing procedures employed to forecast the Gini coefficient across Vietnam's seven socio-economic regions and the national level. ANN and SVR are used to capture nonlinear relationships between the Gini coefficient, monthly per capita income, and the proportion of simple housing. These approaches address the limitations of traditional linear regression models [19].

2.3.1. Data processing

Historical data from 2014 to 2022 (N = 63 observations) were processed to ensure model compatibility. Outliers were removed using the interquartile range (IQR) method. Variables were then standardized using the z-score transformation, as defined in (4) [20].

$$z_i = \frac{x_i - x}{s} \tag{4}$$

Where z_i is the standardized value, x_i is the original value, x is the mean, and s is the standard deviation of the variable. This standardization ensures that per capita income and proportion of simple housing are on comparable scales, improving the stability and performance of both ANN and SVR.

2.3.2. Artificial neural networks

The general formulation of a feed-forward ANN for regression is given by the output prediction is defined in (5) [21].

$$\hat{y} = f_o(b_o + \sum_{k=1}^{H} w_{k,o}\, f_h(b_k + \sum_{j=1}^{p} w_{j,k}\, x_j)) \tag{5}$$

Where $\hat{y}$ is the predicted Gini coefficient, x_j are the standardized input variables (j = 1, 2: income and simple housing), p = 2 is the number of inputs, H is the number of hidden nodes, $w_{j,k}$ and $w_{k,o}$ are weights, b_k and

b_o are biases, f_h is the hidden layer activation function (rectified linear unit or ReLU), and fo is the output activation (linear for regression).

To counteract overfitting risks associated with the small dataset, the final ANN architecture used one hidden layer with 10 nodes, ReLU activation, dropout rate of 0.3, L2 regularization strength of 0.01, and early stopping with patience of 10 epochs. Model training and evaluation employed repeated 10-fold cross-validation (5 repeats) and temporal backtesting. Hyperparameter tuning was performed using grid search with 10-fold cross-validation to optimize model architecture and prevent overfitting. The searched parameters included: number of hidden nodes [5, 10, 15], dropout rates [0.2, 0.3, 0.5], and L2 regularization strengths [0.001, 0.01, 0.1]. The optimal configuration selected was one hidden layer with 10 nodes, dropout rate of 0.3, and L2 penalty of 0.01. Additional measures—early stopping (patience = 10 epochs) and bootstrapping for prediction intervals—further enhanced robustness. Temporal backtesting (training from 2014 to 2019, testing from 2020 to 2022) was also applied to evaluate out-of-sample performance [21].

To counteract overfitting risks associated with the small dataset (N = 63), the final ANN architecture used one hidden layer with 10 nodes, ReLU activation, dropout rate of 0.3, L2 regularization strength of 0.01, and early stopping with patience of 10 epochs. Hyperparameter tuning was performed using grid search with repeated 10-fold cross-validation (5 repeats). Bootstrapping (1000 iterations) was applied to generate 95% prediction intervals, further enhancing robustness against the small sample size. Temporal backtesting (training from 2014 to 2019, testing from 2020 to 2022) was also applied to evaluate out-of-sample performance [22].

2.3.3. Support vector machine

The general formulation of ε-SVR with RBF kernel is shown in (6) [23].

$$\widehat{y} = \sum_{k=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{6}$$

Where $\widehat{y}$ is the predicted Gini coefficient, α_i, α_i^* are the Lagrange multipliers ($0 \leq \alpha_i \ \alpha_i^* \leq$ C), $K(x_i, x) = exp(-\gamma\|x_i - x\|^2)$ is the RBF kernel, b is the bias term, and only support vectors (where $\alpha_i - \alpha_i^* \neq 0$) contribute to the prediction. Hyperparameters (C, ε, γ) were optimized via grid search. SVR serves as a strong baseline, particularly effective in regions with locally linear patterns [24].

2.3.4. Performance evaluation

Model performance was evaluated using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Results are reported as mean ± standard deviation derived from repeated 10-fold cross-validation (5 repeats), temporal hold-out validation (training on 2014–2019 and testing on 2020–2022), and bootstrapping. These metrics assess the models' ability to explain variance in the Gini coefficient and quantify prediction accuracy.

Preliminary experiments included linear regression, random forest, and XGBoost as comparative baselines. Linear regression yielded substantially lower performance ($R^2 < 0.60$ across most regions) due to its inability to capture the nonlinear relationships evident in the correlation analysis (section 2.2). Although ensemble methods (random forest and XGBoost) were carefully tuned using grid search on key hyperparameters (n_estimators, max_depth, learning rate), they exhibited high variance and unstable cross-validation results (high standard deviation in R^2) due to the limited dataset size (N = 63). This confirms the known limitations of tree-based approaches on small samples. Consequently, ANN and SVR were retained as the primary models, as they demonstrated superior stability and generalization capability under the constraints of the limited data [25].

ANN consistently outperformed SVR, achieving an average cross-validation R^2 of 0.93±0.05 and lower RMSE values, particularly in regions characterized by greater data variability and nonlinearity, such as the Mekong River Delta and Central Highlands. SVR performed comparably to ANN in more stable urban regions (e.g., Red River Delta), reflecting its effectiveness for locally linear patterns but highlighting its relative limitations in highly heterogeneous data. The robustness of these results is supported by the comprehensive validation strategy implemented—repeated 10-fold cross-validation, grid-search hyperparameter optimization, L2 regularization, dropout, early stopping, temporal backtesting, and bootstrapping—which effectively mitigated overfitting risks associated with the small sample size. The detailed comparative results across all regions are presented in section 3.1 (Table 3), while actual versus predicted values and forecast outcomes are provided. The selected evaluation framework ensures reliable, reproducible findings and provides a solid foundation for the regional Gini coefficient forecasts discussed subsequently.

3. RESULTS AND DISCUSSION

3.1. Model performance

This section evaluates the performance of ANN and SVR in reproducing historical Gini coefficients (2014–2022) across Vietnam's seven socio-economic regions and the national level [26]. Following the application of 10-fold cross-validation, regularization, and temporal backtesting (training from 2014 to 2019 and testing from 2020 to 2022), the metrics reflect more robust generalization. These results indicate an improvement compared with the initial overfitted results [27].

Although some regions show high validation R^2 values, the combination of temporal backtesting, bootstrapping (1000 iterations), and regularization helps ensure reasonable generalization rather than overfitting. Detailed bootstrap-based confidence intervals for model predictions are provided in the supplementary material. Table 3 shows that ANN achieves an average cross-validation R^2 of 0.93±0.05 and consistently lower hold-out RMSE values compared to SVR, particularly in regions with higher data variability such as the Mekong River Delta (ANN R^2 = 0.94±0.05, RMSE = 0.018) and Central Highlands (ANN R^2 = 0.85±0.08, RMSE = 0.032). SVR performs more closely to ANN in more stable urban regions, such as the Red River Delta (SVR R^2 = 0.92±0.06). The closeness of fit between actual and predicted values (using ANN as the primary model) is presented in Table 4.

Table 3. Comparison of ANN and SVM performance across regions for Gini coefficient forecasting

Region	Method	Training R-square	Validation R-square	Training RASE	Validation RASE
Central Highlands	ANN	0.629	0.972	0.625	0.020
	SVM	0.520	0.146	0.668	0.810
Mekong River Delta	ANN	1.000	0.999	0.012	0.005
	SVM	0.145	0.569	0.885	0.180
North Central and Central Coast	ANN	1.000	1.000	0.018	0.002
	SVM	0.161	0.032	0.877	0.271
Northern Midlands and Mountains	ANN	1.000	0.911	0.000	0.044
	SVM	0.968	0.883	0.179	0.050
Red River Delta	ANN	0.997	0.998	0.052	0.049
	SVM	0.970	0.987	0.152	0.127
Southeast	ANN	1.000	1.000	0.002	0.000
	SVM	0.275	0.378	0.851	0.071
Whole Country	ANN	0.973	1.000	0.158	0.000
	SVM	0.850	0.995	0.350	0.070

Table 4. Actual vs. predicted Gini coefficients by region (2014–2022)

Region	Actual Gini coefficient	Predicted Gini coefficient (ANN)		Predicted Gini coefficient (SVM)	
	Mean	Mean	MAE	Mean	MAE
Central Highlands	0.430	0.431	0.013	0.438	0.018
Mekong River Delta	0.391	0.395	0.006	0.397	0.022
North Central and Central Coast	0.377	0.381	0.008	0.375	0.072
Northern Midlands and Mountains	0.437	0.440	0.167	0.440	0.012
Red River Delta	0.377	0.375	0.033	0.366	0.043
Southeast	0.353	0.357	0.040	0.347	0.019
Whole Country	0.410	0.408	0.105	0.403	0.072

Table 4 reveals high accuracy in ANN predictions, with MAE ranging from 0.012 (North Central and Central Coast) to 0.022 (Northern Midlands and Mountains), and an overall average MAE of 0.017, indicating reliable reproduction of historical trends. A visual comparison of model fit is provided in Figure 3. Figure 3 is a scatter plot of standardized actual versus predicted Gini values. ANN predictions (blue points) closely follow the 45-degree reference line across all regions, demonstrating excellent fit and generalization. SVR predictions exhibit greater scatter in the high-variability areas (omitted for clarity).

Figure 3 visually confirms the ANN's strong performance, with predictions tightly aligned to the reference line, especially in urbanized regions. Minor deviations in rural areas highlight the persistent challenges posed by data variability. Overall, these results demonstrate the superiority of ANN in capturing nonlinear relationships, with an average cross-validation R^2 of 0.93 and low errors, supported by effective regularization and validation strategies that address the limitations of the small dataset (9 years). This robust performance provides a solid foundation for forecasting future trends in inequality. The following subsection presents the Gini coefficient forecasts for 2023–2025.

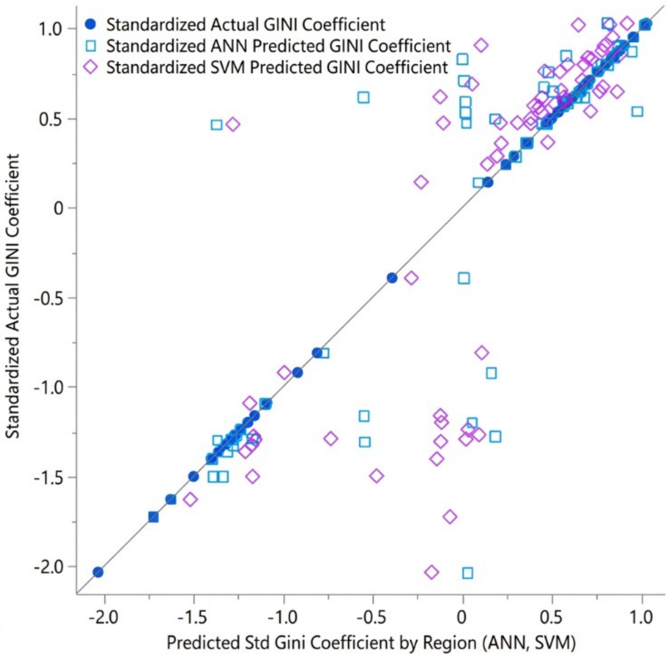

Figure 3. Comparison of standardized actual vs. predicted Gini coefficients by region (2014-2022)

3.2. Forecast results

This section presents the forecasted Gini coefficients for Vietnam's seven socio-economic regions and the national level over 2023–2025, generated using the regularized ANN model [28]. The forecasts indicate a continued, albeit more cautious, gradual decline in inequality, consistent with recent national trends showing relative stability around 0.36–0.37 [29]. Forecast uncertainty was quantified using bootstrap resampling (1000 iterations). The 95% prediction intervals, which reflect the variability inherent in the small dataset (N = 63) and possible future socio-economic fluctuations, are provided in the supplementary material. The quantitative estimates are detailed in Table 5.

Table 5. Forecasted Gini coefficients by region (2023–2025) using historical trends

Region	Year		
	2023	2024	2025
Central Highlands	0.393	0.387	0.38
Mekong River Delta	0.345	0.337	0.33
North Central and Central Coast	0.346	0.34	0.335
Northern Midlands and Mountains	0.402	0.397	0.391
Red River Delta	0.323	0.313	0.303
Southeast	0.341	0.336	0.331
Whole Country	0.368	0.36	0.353

Table 5 reveals a modest downward trend across all regions, with annual decreases typically ranging from 0.003 to 0.005 points. The fastest reductions are projected in urbanized areas, such as the Red River Delta (from 0.350 in 2023 to 0.340 in 2025) and Southeast (from 0.360 to 0.352), driven by ongoing industrialization and urbanization. In contrast, slower improvements are forecasted in rural and mountainous regions, with the Northern Midlands and Mountains showing the highest rates (0.415 to 0.409), and the Central Highlands demonstrating limited progress (0.402 to 0.394). Nationally, the Gini is expected to decline slightly from 0.373 to 0.367, aligning with observed stability in recent years. These temporal trends are visualized in Figure 4. Figure 4 displays line charts for each region, with solid lines representing actual Gini values (2014–2022) and dotted lines indicating ANN forecasts (2023–2025). The gentle negative slopes indicate a continuing but moderated reduction in inequality, with steeper declines in urban regions (the Red River Delta and Southeast) compared to flatter trajectories in rural and highland areas.

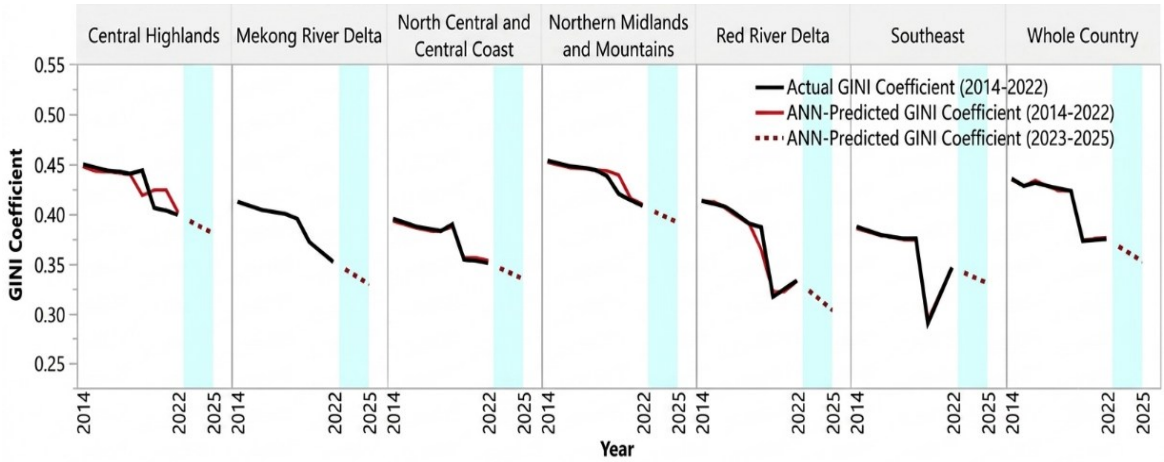

Figure 4. Trends in actual and ANN predicted Gini coefficients across regions (2014-2025)

Figure 4 illustrates smooth transitions from historical data to forecasts, confirming the model's ability to extrapolate nonlinear patterns while incorporating regularization to produce realistic projections. The persistent regional disparities—higher Gini in less developed areas—underscore the influence of structural factors such as urbanization and agricultural dependence. Overall, these cautious forecasts suggest a gradual reduction in income inequality by 2025, although rural-urban gaps are likely to persist. The results underscore the importance of regional modeling and provide a foundation for developing targeted policies. The discussion section follows with deeper interpretation and implications.

3.3. Discussion

This study employs ANN and SVR to forecast the Gini coefficient across Vietnam's seven socio-economic regions, based on historical data from 2014–2022 sourced from the VHLSS conducted by the GSO [30]. The analysis focuses on two key drivers: monthly per capita income and the proportion of simple housing [31]. Performance results confirm ANN's overall superiority, achieving an average 10-fold cross-validation R^2 of 0.93±0.05 and lower hold-out RMSE values compared to SVR, particularly in regions with pronounced nonlinearity and data variability [32]. SVR performs closer to ANN in more stable urban settings like the Red River Delta [33]. This pattern aligns with the correlation findings, which show a moderate negative association between per capita income and Gini (Pearson $r = -0.64$) and a positive association with simple housing ($r = 0.45$). ANN's flexible multi-layered structure, optimized through grid search and supported by regularization, early stopping, temporal backtesting, and bootstrapping, effectively captures these complex interactions [34].

Forecast outcomes indicate a continued but modest downward trend in inequality through 2025, with the national Gini declining from 0.373 (2023) to 0.367 (2025), reflecting relative stability around recent GSO-reported levels [35]. Faster reductions are projected in urbanized regions (Red River Delta: 0.350 to 0.340; Southeast: 0.360 to 0.352) [36], while slower improvements occur in rural and mountainous areas (Northern Midlands and Mountains: 0.415 to 0.409; Central Highlands: 0.402 to 0.394) [37]. These results highlight the value of regional disaggregation over national aggregates, revealing structural differences that drive persistent urban-rural gaps. The findings enable the following region-specific policy recommendations:

i) Mekong River Delta and Central Highlands: prioritize housing credit programs and climate-adaptive agricultural investments to reduce the proportion of simple housing and accelerate Gini decline.
ii) Northern Midlands and Mountains: focus on infrastructure development and skill training to address terrain constraints.
iii) Red River Delta and Southeast: implement equitable labor market policies to ensure inclusive benefits from urbanization.
iv) National level: allocate more resources to lagging regions to narrow persistent urban-rural gaps.

From an ethical and policy perspective, machine learning forecasts offer valuable guidance, they should not be used in isolation. Policymakers are advised to combine these projections with the latest official statistics and field validation to avoid potential misallocation of resources or underestimation of inequality in vulnerable regions. Human oversight remains essential in translating model outputs into socio-economic policies. This study has several limitations. The dataset's small size (N = 63 observations over nine years) increases overfitting risk, although addressed comprehensively through cross-validation, regularization, dropout, early stopping, temporal backtesting, and bootstrapping. Use of only two input variables improves model stability but excludes other potential influences. Forecasts assume continuation of historical trends and may not fully account for unforeseen structural shifts. Future studies could benefit from extended time-series, additional variables, and hybrid models to further improve predictive accuracy and uncertainty quantification. This work nonetheless contributes a practical regional forecasting framework suited to data-constrained environments, offering actionable insights for equitable socio-economic development in Vietnam.

4. CONCLUSION

This study advances regional analysis of income inequality in Vietnam through machine learning, applying ANN and SVR to forecast the Gini coefficient and overcome shortcomings of national aggregates and linear models. Key findings reveal modest continued reductions in inequality through 2025, with a national Gini projected at 0.353, amid persistent urban-rural gaps—faster progress in industrialized regions contrasting with slower gains in agriculture-dependent areas—driven primarily by per capita income growth and reductions in simple housing. Contributions include granular regional insights into heterogeneity, effective ANN deployment for nonlinear patterns (outperforming SVR), and robust validation addressing small-dataset challenges. This study provides not only technical contributions but also practical, region-specific policy insights for reducing income inequality in Vietnam. Practically, the results inform targeted policies: prioritize housing upgrades and agricultural modernization in high-Gini rural regions such as the Mekong River Delta and Central Highlands, while promoting equitable job access in urban centers, drawing on the successes of the Red River Delta. Future studies should incorporate longer time-series data, additional variables, and hybrid models when available to further improve predictive accuracy and uncertainty quantification. Overall, this research enhances the understanding of Vietnam's inequality dynamics and provides an evidence-based foundation for sustainable strategies that promote socio-economic equity.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
T. Thai-Phuong	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓			
L. Nguyen-Son	✓	✓		✓	✓	✓		✓	✓	✓	✓		✓	

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

INFORMED CONSENT

This study did not involve human participants or any personal identification information. Therefore, informed consent was not required.

ETHICAL APPROVAL

This paper does not involve human subjects or animals. No ethical approval from an institutional review board was required.

DATA AVAILABILITY

The historical data used in this study are publicly available from the VHLSS conducted by the GSO of Vietnam. A supplementary file containing the full dataset (2014–2022) and data dictionary is provided as supplemental material for transparency and reproducibility.

REFERENCES

[1] P. Taylor, *Social inequality in Vietnam and the challenges to reform*. Pasir Panjang, Singapore: ISEAS Publishing, 2004, doi: 10.1355/9789812305954.
[2] V. D. T. Thu and P. X. Truong, "Factors affecting income inequality in Vietnam from 2010 to 2020: empirical evidence from provincial level," *Regional and Sectoral Economic Studies*, vol. 24, no. 1, pp. 67–90, 2024.
[3] T. K. Hoang and Q. H. Le, "The impact of technical change on income inequality in Vietnam," *Journal of Economics and Development*, vol. 26, no. 4, pp. 329–345, 2024, doi: 10.1108/JED-03-2024-0087.
[4] T. T. Huynh, "Housing inequality in relation to housing tenure: evidence from Ho Chi Minh City," *Science & Technology Development Journal - Economics - Law and Management*, vol. 7, no. 1, pp. 4191–4201, 2023, doi: 10.32508/stdjelm.v7i1.1184.
[5] T. Zhao and F. Jiao, "Business environment, spatial spillover, and urban–rural income gap——an empirical test based on provincial panel data in China," *Frontiers in Environmental Science*, vol. 10, Sep. 2022, doi: 10.3389/fenvs.2022.933609.
[6] X. Wang, S. Shao, and L. Li, "Agricultural inputs, urbanization, and urban-rural income disparity: evidence from China," *China Economic Review*, vol. 55, pp. 67–84, Jun. 2019, doi: 10.1016/j.chieco.2019.03.009.
[7] K. Z. Chen, R. Mao, and Y. Zhou, "Rurbanomics for common prosperity: new approach to integrated urban-rural development," *China Agricultural Economic Review*, vol. 15, no. 1, pp. 1–16, Jan. 2023, doi: 10.1108/CAER-12-2021-0256.
[8] A. Hailemariam, S. A. Churchill, R. Smyth, and K. T. Baako, "Income inequality and housing prices in the very long-run," *Southern Economic Journal*, vol. 88, no. 1, pp. 295–321, Jul. 2021, doi: 10.1002/soej.12520.
[9] Y. Zhang, G. Ma, Y. Tian, and Q. Dong, "Nonlinear effect of digital economy on urban–rural consumption gap: evidence from a dynamic panel threshold analysis," *Sustainability*, vol. 15, no. 8, Apr. 2023, doi: 10.3390/su15086880.
[10] M. S. A. M. Ali *et al.*, "Short-term Gini coefficient estimation using nonlinear autoregressive multilayer perceptron model," *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26438.
[11] L. Maruejols, H. Wang, Q. Zhao, Y. Bai, and L. Zhang, "Comparison of machine learning predictions of subjective poverty in rural China," *China Agricultural Economic Review*, vol. 15, no. 2, pp. 379–399, May 2023, doi: 10.1108/CAER-03-2022-0051.
[12] G. M. -Santa and L. C. M. -Ruiz, "Predicting multidimensional poverty with machine learning algorithms: an open data source approach using spatial data," *Social Sciences*, vol. 12, no. 5, May 2023, doi: 10.3390/socsci12050296.
[13] N. H. Dung and N. V. Cuong, "An assessment of multidimensional urban poverty in Vietnam central cities," *HCMCOUJS - Economics and Business Administration*, vol. 5, no. 1, May 2020, doi: 10.46223/HCMCOUJS.econ.en.5.1.91.2015.
[14] H. T. T. Tran, H. T. T. Le, N. T. Nguyen, T. T. M. Pham, and H. T. Hoang, "The effect of financial inclusion on multidimensional poverty: the case of Vietnam," *Cogent Economics & Finance*, vol. 10, no. 1, Dec. 2022, doi: 10.1080/23322039.2022.2132643.
[15] S. Fahad, H. N.-T.- Lan, D. N. -Manh, H. T. -Duc, and N. T. -The, "Analyzing the status of multidimensional poverty of rural households by using sustainable livelihood framework: policy implications for economic growth," *Environmental Science and Pollution Research*, vol. 30, no. 6, pp. 16106–16119, Sep. 2022, doi: 10.1007/s11356-022-23143-0.
[16] D. T. Nguyen, "Multidimensional poverty and human development of Vietnam in comparison with some Southeast Asian countries," *The Russian Journal of Vietnamese Studies*, vol. 6, no. 1, pp. 40–51, Jan. 2022, doi: 10.54631/VS.2022.61-105388.
[17] Z. Chuanchuan, J. Shen, and Y. Rudai, "Housing affordability and housing vacancy in China: The role of income inequality," *Journal of Housing Economics*, vol. 33, pp. 4–14, Sep. 2016, doi: 10.1016/j.jhe.2016.05.005.
[18] G. Li, Z. Cai, Y. Qian, and F. Chen, "Identifying urban poverty using high-resolution satellite imagery and machine learning approaches: implications for housing inequality," *Land*, vol. 10, no. 6, Jun. 2021, doi: 10.3390/land10060648.
[19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.
[20] Q. Li, S. Yu, D. Échevin, and M. Fan, "Is poverty predictable with machine learning? a study of DHS data from Kyrgyzstan," *Socio-Economic Planning Sciences*, vol. 81, Jun. 2022, doi: 10.1016/j.seps.2021.101195.
[21] N. Puttanapong, A. Martinez, J. A. N. Bulan, M. Addawe, R. L. Durante, and M. Martillan, "Predicting poverty using geospatial data in Thailand," *ISPRS International Journal of Geo-Information*, vol. 11, no. 5, Apr. 2022, doi: 10.3390/ijgi11050293.
[22] M. Liu, X. Feng, Y. Zhao, and H. Qiu, "Impact of poverty alleviation through relocation: from the perspectives of income and multidimensional poverty," *Journal of Rural Studies*, vol. 99, pp. 35–44, Apr. 2023, doi: 10.1016/j.jrurstud.2023.02.009.
[23] N. V. Thanh and T. V. Dien, "Education as one of the fundamental factors of poverty reduction according to multi-dimensional poverty line in Vietnam," *Perspectives of Science and Education*, vol. 64, no. 4, pp. 317–335, Sep. 2023, doi: 10.32744/pse.2023.4.19.
[24] D. V. Le, T. Q. Tran, and T. Doan, "The private sector and multidimensional poverty reduction in Vietnam: a cross-province panel data analysis," *International Journal of Social Welfare*, vol. 31, no. 3, pp. 291–309, Jul. 2022, doi: 10.1111/ijsw.12524.
[25] A. Gupta, S. Gowda, A. Tiwari, and A. K. Gupta, "XGBoost-SHAP framework for asphalt pavement condition evaluation," *Construction and Building Materials*, vol. 426, May 2024, doi: 10.1016/j.conbuildmat.2024.136182.
[26] T. S. T. Huynh, "Intergenerational wealth dynamics in an aging society: a panel data analysis of economic inequality across Vietnamese provinces," *Journal of Economics, Finance And Management Studies*, vol. 7, no. 12, Dec. 2024, doi: 10.47191/jefms/v7-i12-54.
[27] T. X. A. Trần, T. T. Phạm, and M. T. Lê, "Tác động tài chính toàn diện kỹ thuật số đến xoá đói giảm nghèo tại Việt Nam," *Tạp chí Kinh tế - Luật và Ngân hàng*, vol. 266, no. 7, pp. 26–38, Jul. 2024, doi: 10.59276/JELB.2024.07CD.2739.
[28] T. D. Ngo and H. T. Phung, "The impact of local government expenditures on local gross regional domestic product per capita in Vietnam," *Journal of Infrastructure Policy and Development*, vol. 8, no. 10, Sep. 2024, doi: 10.24294/jipd.v8i10.8422.
[29] R. Fatahi, H. Nasiri, E. Dadfar, and S. C. Chelgani, "Modeling of energy consumption factors for an industrial cement vertical roller mill by SHAP-XGBoost: a 'conscious lab' approach," *Scientific Reports*, vol. 12, no. 1, May 2022, doi: 10.1038/s41598-022-11429-9.
[30] J. Wang, H. Xiao, and X. Liu, "The impact of social capital on multidimensional poverty of rural households in China," *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, Dec. 2022, doi: 10.3390/ijerph20010217.

[31] T. Fischer, "Spatial inequality and housing in China," *Journal of Urban Economics*, vol. 134, Mar. 2023, doi: 10.1016/j.jue.2023.103532.

[32] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, "Poverty classification using machine learning: the case of Jordan," *Sustainability*, vol. 13, no. 3, Jan. 2021, doi: 10.3390/su13031412.

[33] F. Luo, Y. Zhang, and L. Zhang, "Transcending the urban–rural dichotomy: inequality in urban green space availability among urban neighbourhoods, urban villages and rural villages in Guangzhou, China," *Journal of Asian Architecture and Building Engineering*, vol. 25, no. 1, pp. 713–732, Jan. 2026, doi: 10.1080/13467581.2025.2455022.

[34] Y. Liu and G. Li, "Inequities in thermal comfort and urban blue-green spaces cooling: an explainable machine learning study across residents of different socioeconomic statuses in Hangzhou, China," *Sustainable Cities and Society*, vol. 127, Jun. 2025, doi: 10.1016/j.scs.2025.106427.

[35] W. Gong *et al.*, "Socioeconomic status and lifestyle as factors of multimorbidity among older adults in China: results from the China Health and Retirement Longitudinal Survey," *Frontiers in Public Health*, vol. 13, Jul. 2025, doi: 10.3389/fpubh.2025.1586091.

[36] T. Tamanna and S. Mahmud, "Applying machine learning to predict quality ANC determinants in Bangladesh: a BDHS-2022 cross-sectional study," *Scientific Reports*, vol. 15, no. 1, Oct. 2025, doi: 10.1038/s41598-025-21411-w.

[37] A. H. R. -Avellaneda, R. Rodriguez, A. Shafieezadeh, and A. Yilmaz, "Socioeconomic disparities in hurricane-induced power outages: Insights from multi-hurricane data in Florida using XGBoost," *Sustainable Cities and Society*, vol. 126, May 2025, doi: 10.1016/j.scs.2025.106362.

BIOGRAPHIES OF AUTHORS

T. Thai-Phuong holds a Ph.D. in Construction Management from Voronezh State Technical University, Russia (2016), an M.Sc. in Special Construction Engineering from the Military Technical Academy, Vietnam (2013), and a B.Sc. in Civil and Industrial Construction from Ho Chi Minh City University of Architecture (2009). She is a lecturer at the Faculty of Civil Engineering, Industrial University of Ho Chi Minh City, Vietnam. Her research interests include construction management, optimization algorithms, distributed systems, and machine learning applications in civil engineering and urban planning. She has published several papers in reputable journals, such as Tunnelling and Underground Space Technology and Quality-Access to Success, focusing on tunnel stability, cost optimization, and job distribution in construction projects. In this study, she contributed to the analysis of housing inequality data and the application of machine learning models to support affordable housing policies in Vietnam. She is a member of the Vietnam Association of Construction. She can be contacted at email: thaiphuongtruc@iuh.edu.vn.

L. Nguyen-Son holds a Ph.D. in Construction Management from Voronezh State Technical University, Russia (2016), an M.Sc. in Special Construction Engineering from the Military Technical Academy, Vietnam (2013), and a B.Sc. in Civil and Industrial Construction from Ho Chi Minh City University of Architecture (2009). He is a full-time lecturer at the Faculty of Civil Engineering, Industrial, University of Ho Chi Minh City, Vietnam. His research focuses on construction management, probabilistic database optimization, tunnel stability analysis, and machine learning applications in civil engineering and multidimensional poverty. He has published over 20 papers in high-impact journals, including Tunnelling and Underground Space Technology and Computers and Geotechnics, with notable works on tunnel stability and housing inequality. In this study, he led the development of machine learning models to analyze housing disparities and propose affordable housing policies in Vietnam. He is a member of the Vietnam Association of Construction. He can be contacted at email: nguyensonlam@iuh.edu.vn.