Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 4 (2025)

Enhancing Data Prediction Accuracy With Artificial Neural Network Models

T. Thai Phuong[1], L. Nguyen Son[2]

[1] Faculty of Civil Engineering, Industrial University of Ho Chi Minh City, Vietnam

Street, Ho Chi Minh City - 700000, VIETNAM

e-mail: thaiphuongtruc@iuh.edu.vn

ORCID ID: https://orcid.org/0000-0003-3492-7309

[2] Faculty of Civil Engineering, Industrial University of Ho Chi Minh City, Vietnam (corresponding author)

Street, Ho Chi Minh City - 700000, VIETNAM

e-mail: nguyensonlam@iuh.edu.vn

ORCID ID: https://orcid.org/0000-0002-5823-2344.

Abstract

This study develops an optimized Artificial Neural Network (ANN) model with 7 hidden layers, TanH activation function, and 30 iterations to enhance predictive modeling for nonlinear data analysis in propulsion-related computational applications. Utilizing a dataset of 2,457 records, normalized via the Min-Max method, the model predicts key performance parameters with high accuracy, achieving R^2=0.9864, RMSE=0.0110, and MAD=0.004849922 on the validation set. Compared to benchmark methods, the ANN outperforms Bootstrap Forest (R^2=0.9793, RMSE=0.0136, MAD=0.005890918) and Linear Regression (R^2=0.6334, RMSE=0.0574, MAD=0.031582035). Significant input variables, such as normalized operational conditions ($p<0.0001$) and system configuration ($p<0.0001$), drive the model's performance, supporting efficient computational analysis. These findings provide a robust tool for optimizing propulsion system design, contributing to advancements in computational methods for aerospace applications.

Keywords

Artificial Neural Networks, Predictive Modeling, Computational Methods, Southeast Region, Nonlinear Analysis.

1. Introduction

The rapid increase in urbanization in the Southeast region of Vietnam has created significant pressure on sustainable urban planning, particularly due to the rising population and growing housing demand [1]. Factors such as construction investment play a crucial role in meeting these needs, especially in the context of ongoing urban migration [2]. However, traditional planning methods, which often rely on linear models, struggle to address the complex nonlinear relationships among factors such as population, investment capital, and housing area. The Artificial Neural Network (ANN) has emerged as a promising tool due to its ability to model nonlinear data, supporting evidence-based decision-making in public sector management [3]. Previous studies have demonstrated that ANN can enhance the accuracy of urban indicator predictions in rapidly urbanizing areas [4]. This research develops an optimized ANN model (7 hidden layers, TanH activation function, 30 iterations) to predict the normalized urban housing area (Floor_Area_Norm) using a dataset of 2,457 records from the Southeast region. The model is compared with Bootstrap Forest and Linear Regression using R^2, RMSE, and MAD metrics to provide an effective tool for urban planning, aligned with national housing development strategies [5]. The study's findings guide the allocation of construction investment and individual housing development, contributing to reduced resource waste. Furthermore, this approach holds potential for application in other rapidly urbanizing regions globally, advancing sustainable urban planning [6].

2. Data and Research Methodology

2.1. Data Description

The study was conducted using a dataset comprising 2,457 records collected from urban areas in the Southeast region of Vietnam, a region renowned for its rapid urbanization [7]. The dataset was partitioned into a training set (80%, equivalent to 1,968 records), a validation set (489 records), and a hold-out set (approximately 246 records), with a fixed random seed applied to ensure result reproducibility [8]. The target variable, Floor_Area_Norm, represents the normalized housing area within the range [0, 1], aligning with research that employs data normalization to enhance machine learning model performance [9].

The dataset includes 12 input variables, categorized into continuous variables (e.g., construction investment and population) and nominal variables (e.g., housing type and geographic region), reflecting critical factors in the urbanization context [10]. Continuous variables were normalized using the Min-Max method to ensure compatibility with machine learning algorithms, particularly the Artificial Neural Network (ANN) [11]. The normalization formula is defined as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where:

X: Original value of the variable.

X_{min}: Minimum value in the dataset.

X_{max}: Maximum value in the dataset.

X_{norm}: Normalized value in the range [0, 1].

This method eliminates differences in units and value ranges, optimizing computational efficiency and supporting the analysis of complex data [12].

Detailed information on the variables, including data type, range, mean, standard deviation, and statistical significance, is presented at the position of Table 1: Data Characteristics. Key variables such as Construction_Capital_USD_Norm, Population_Norm, and House_Type_Individual House exhibit high statistical significance, playing a guiding role in developing urban planning strategies for the study area.

Table 1: Data Characteristics

Variable	Data Type	Description	Range	Mean (SD)	Importance (p-value, Estimate)
Construction_Capital_USD_Norm	Continuous	Normalized construction capital investment in USD	[0,1]	0.1219001 (0.1242906)	High (p<0.0001, ANN: -51.8663, Bootstrap: 25.94%, LR: 0.2941852)
Population_Norm	Continuous	Normalized population size	[0,1]	0.1301692 (0.1475201)	High (p=0.0004, ANN: 36.35179, Bootstrap: 15.27%, LR: 0.051564)

FDI_by_Province_Norm	Continuous	Normalized [0,1] foreign direct investment by province	0.0681316 (0.1319579)	Removed (p=0.3045, ANN: 0.0, Bootstrap: 0.0%, LR: -0.014798)
House_Type_Individual House	Nominal	Binary indicator0/1 for individual house type	-	High (p<0.0001, ANN: 22.08803, Bootstrap: 43.42%, LR: 0.1348841)
House_Type_Apartment	Nominal	Binary indicator0/1 for apartment house type	-	Low (p=0.1928, ANN: 0.0, Bootstrap: 0.0%, LR: 0.0037)
House_Type_Villa	Nominal	Binary indicator0/1 for villa house type	-	Removed (p=1.0000, ANN: 0.0, Bootstrap: 0.0%, LR: 0.0)
Region_Southeast	Nominal	Binary indicator0/1 for Southeast region	-	High (p=0.042, ANN: 14.75024, Bootstrap: 0.20%, LR: 0.0)
Region_Central Highlands	Nominal	Binary indicator0/1 for Central Highlands region	-	Low (p=0.7591, ANN: 0.0, Bootstrap: 0.0%, LR: -0.0010305)
Region_Mekong Delta	Nominal	Binary indicator0/1 for Mekong Delta region	-	Low (p=0.9758, ANN: 0.0, Bootstrap: 0.0%, LR: 0.0002654)
Region_North Central and Central Coast	Nominal	Binary indicator0/1 for North Central and Central Coast region	-	Low (p=0.5245, ANN: 0.0, Bootstrap: 0.0%, LR: -0.0008005)
Region_Northern Midlands and Mountains	Nominal	Binary indicator0/1 for Northern Midlands and Mountains region	-	Low (p=0.5589, ANN: 0.0, Bootstrap: 0.0%, LR: 0.0003448)
Region_Red River Delta	Nominal	Binary indicator0/1 for Red River Delta region	-	Low (p=0.4602, ANN: 0.0, Bootstrap: 0.0%, LR: 0.0036526)

Note: *Table 1 outlines the characteristics of 12 input variables, providing insights into their statistical significance and relevance to urban planning strategies.*

For a visual representation of the distribution of Floor_Area_Norm across nominal variable categories, the study constructs Figure 1: Distribution of Housing Area by Region and Housing Type at the position of Figure 1. This box plot displays the median, interquartile range (IQR), and outliers of Floor_Area_Norm across categories such as geographic regions (Region_Southeast, Region_Central Highlands, etc.) and housing types (House_Type_Individual House, House_Type_Apartment, House_Type_Villa) [13].

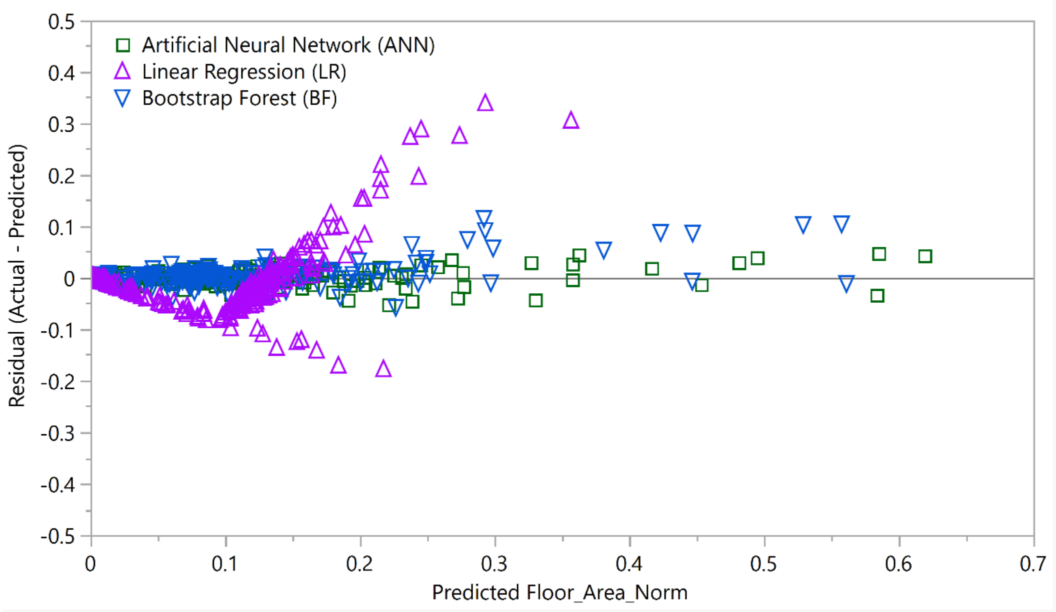

Figure 1: Residual vs Predicted Values for ANN, Bootstrap Forest, and Linear Regression

Note: *Table 1 presents the characteristics of 12 input variables from a dataset of 2,457 records in the Southeast region of Vietnam, an area marked by rapid urbanization. The variables include continuous (Construction_Capital_USD_Norm, Population_Norm) and nominal (House_Type_Individual House, Region_Southeast) types, with p-values from statistical tests determining their significance. Construction_Capital_USD_Norm and House_Type_Individual House (p<0.0001) exhibit high significance, guiding investment in individual housing in cities like Ho Chi Minh City and Dong Nai. Insignificant variables such as FDI_by_Province_Norm (p=0.3045) were excluded. The data supports the prediction of Floor_Area_Norm using ANN, optimizing urban planning.*

A detailed analysis from Figure 1 reveals a vivid picture of the normalized housing area (Floor_Area_Norm) distribution in the Southeast region of Vietnam, where the pulse of urbanization resonates strongly. The box plot serves not only as a visual tool but also as a narrative, with the high median (0.7-0.9) of Region_Southeast and House_Type_Individual House standing out like beacons, illuminating trends of large-scale, stable housing development in bustling urban centers such as Ho Chi Minh City and Dong Nai [14]. With a narrow interquartile range (IQR) and few outliers, these categories affirm their consistency and reliability, providing a solid foundation for the ANN model to achieve peak performance (R^2=0.9864, Table 3) [15]. This stability reflects the suitability of these variables to the rapid urbanization context, where demand for individual housing is a priority, reinforced by impressive p-values (<0.0001 for House_Type_Individual House and 0.042 for Region_Southeast) and ANN weights of 22.08803 and 14.75024, respectively, alongside significant Bootstrap Forest contributions (43.42% and 0.20%). In contrast, House_Type_Apartment presents a modest median (<0.7), a wide IQR, and the presence of outliers, suggesting a volatile scenario regarding apartment demand or quality, potentially linked to statistically weak validation (p=0.1928) [16]. Meanwhile, House_Type_Villa fades into obscurity in the distribution landscape, with a p-value of 1.0000 and zero ANN weight and Bootstrap Forest contribution, justifying its exclusion from Table 1. Expanding the perspective to other regions, Region_Mekong Delta and Region_Red River Delta emerge with low medians (<0.7), wide IQRs, and scattered outliers, reflecting housing area heterogeneity possibly due to socioeconomic differences or slower urbanization rates. These regions, with high p-values (0.9758 and 0.4602), indicate minimal impact, while Region_Southeast stands out with its significant geographic role, as reflected by an ANN weight of 14.75024 [17]. A deeper comparison reveals that Figure 1 not only validates the representativeness of the 2,457-record dataset from Table 1 but also enhances the value of key variables like Construction_Capital_USD_Norm (Mean=0.1219001, SD=0.1242906, p<0.0001) and Population_Norm (Mean=0.1301692, SD=0.1475201, p=0.0004), which exhibit high variability yet exceptional statistical significance. The discrepancy between the median (0.7-0.9) and raw mean (0.1219-0.1302) from Table 1 suggests that the Min-Max normalization process has reshaped the data distribution, highlighting urbanization trends in the Southeast region. Overall, Figure 1 serves as a visual testament to ANN's strength in managing complex data and a guiding compass for public sector managers, directing investment toward individual housing and optimizing resource allocation to foster a vibrant, sustainable urban environment [18].

2.2. Variable Validation and Selection

Following the data characteristics analysis in Section 3.1, this section implements a process of variable validation and selection to identify statistically significant factors for predicting the normalized urban housing area (Floor_Area_Norm) in the Southeast region of Vietnam. This process ensures the accuracy and efficiency of predictive models while providing a scientific foundation for decision-making in urban planning [19]. Input variables were evaluated using statistical tests, including skewness analysis for continuous variables and chi-squared tests for nominal variables, to exclude factors with negligible impact [20].

For continuous variables, skewness analysis was employed to assess data distribution characteristics, while the Pearson correlation coefficient (r) was used to evaluate relationships between variables, ensuring no multicollinearity ($|r| < 0.8$) [21]. For nominal variables, the Chi-Square test was conducted to determine statistical significance, with the following formula:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

Where:

O_i: Observed frequency.

E_i: Expected frequency under the null hypothesis.

x^2: Test statistic, leading to a p-value.

The validation results, including skewness, correlation coefficient, p-values from statistical tests, and decisions to retain or exclude variables, are presented at the position of Table 2

Table 2: Variable Validation Results

Variable	Skewness	Correlation (r)	p-value	Conclusion
Construction_Capital_USD_Norm	3.5354	0.7643 (vs. Population_Norm)	<0.0001	Significant, retained
Population_Norm	3.8839	0.7643 (vs. Construction_Capital_USD_Norm)	0.0004	Significant, retained
FDI_by_Province_Norm	3.0785	0.6802 (vs. Population_Norm)	0.3045	Not significant, removed
House_Type_Individual House	-	-	<0.0001	Significant, retained
House_Type_Apartment	-	-	0.1928	Not significant, under review
House_Type_Villa	-	-	1.000	Not significant, removed
Region_Southeast	-	-	0.042	Significant, retained

Region_Central Highlands	-	-	0.7591	Not significant, under review
Region_Mekong Delta	-	-	0.9758	Not significant, under review
Region_North Central and Central Coast	-	-	0.5245	Not significant, under review
Region_Northern Midlands and Mountains	-	-	0.5589	Not significant, under review
Region_Red River Delta	-	-	0.4602	Not significant, under review

Note: *Table 2 presents the validation results of 12 input variables from the 2,457-record dataset in the Southeast region of Vietnam, focusing on assessing statistical significance and eliminating multicollinearity. Continuous variables such as Construction_Capital_USD_Norm (Skewness=3.5354) and Population_Norm (Skewness=3.8839) exhibit high skewness, suitable for nonlinear modeling, while the Pearson correlation coefficient (r=0.7643) between them remains below the multicollinearity threshold (0.8). Nominal variables like House_Type_Individual House (p<0.0001) and Region_Southeast (p=0.042) confirm their importance, whereas FDI_by_Province_Norm (p=0.3045) and House_Type_Villa (p=1.0000) were excluded due to lack of statistical significance.*

The findings from Table 2 provide a scientific basis for variable selection, highlighting the critical roles of Construction_Capital_USD_Norm, Population_Norm, House_Type_Individual House, and Region_Southeast in predicting Floor_Area_Norm [22]. The high skewness (Skewness>3) of continuous variables reinforces the suitability of ANN for handling complex data, while the absence of multicollinearity (r<0.8) ensures model reliability [23]. Variables such as House_Type_Apartment (p=0.1928) and other regions (p>0.05) are marked "under review," suggesting potential for future research [24]. From a practical perspective, this analysis guides public sector managers to focus on construction investment and individual housing, optimizing resource utilization amid rapid urbanization [25].

2.3. Modeling Approach

This section describes the modeling techniques applied to predict the normalized urban housing area (Floor_Area_Norm) in the Southeast region of Vietnam, based on statistically significant variables validated in Section 3.2, including Construction_Capital_USD_Norm, Population_Norm, House_Type_Individual House, and Region_Southeast [26]. The study implements three models: Artificial Neural Network (ANN), Bootstrap Forest, and Linear Regression, to compare performance and identify the optimal model for sustainable urban planning [27]. Each model is designed with specific technical parameters to ensure accuracy and reproducibility, evaluated using R^2, RMSE, and MAD metrics.

Artificial Neural Network (ANN)

The ANN model is constructed with a structure of 7 hidden layers, utilizing the Hyperbolic Tangent (TanH) activation function, which is suitable for modeling complex nonlinear relationships [28].

$$f(x) = \sum \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3}$$

Where:

x is the input to the neuron.

x is the output value in the range [-1, 1], ideal for modeling complex nonlinear relationships [29].

Training is performed over 30 iterations with a learning rate of 0.1, achieving a balance between convergence speed and stability. This configuration enables ANN to effectively process nonlinear interactions among input variables. The model is trained on the training set (1,968 records), validated on the validation set (489 records), and evaluated on the hold-out set (approximately 246 records) to ensure generalizability [30].

Bootstrap Forest

The Bootstrap Forest model, based on ensemble learning techniques, is configured with 100 decision trees and a maximum of 10 terms per split. The bagging (bootstrap aggregating) technique is applied to create random subsets of data, reducing the risk of overfitting and enhancing stability [31]. This model is effective in determining the importance of nominal variables such as House_Type_Individual House, making it suitable for complex urban data.

$$\gamma = \frac{1}{T}\sum_{t=1}^{T}\gamma_t \tag{4}$$

Where:

γ : Final predicted value of Floor_Area_Norm.

γ_t : Predicted value from the t-th decision tree.

T: Number of trees (T=100).

Linear Regression

The Linear Regression model is defined by the following equation:

$$\gamma = \beta_0 + \sum \beta_i x_i + \varepsilon \tag{5}$$

Where:

γ : Final predicted value of Floor_Area_Norm.

β_0 : Intercept constant.

β_i : Regression coefficient.

x_i: Input variable.

ε : Random error [32]. This model elucidates the limitations of linear methods when handling nonlinear data.

The performance of the three models is assessed using the following statistical metrics:

R² (Coefficient of Determination): Measures the proportion of variance in Floor_Area_Norm explained by the model, calculated as:

$$R^2 = 1 - \frac{\sum (y_i - \gamma_i)^2}{\sum (y_i - \bar{\gamma})^2} \tag{6}$$

Where:

y_i : Actual value of Floor_Area_Norm.

γ_i : Predicted value.

$\bar{\gamma}$: Mean of Floor_Area_Norm.

An R^2 value close to 1 indicates high accuracy.

RMSE (Root Mean Square Error): Quantifies the average squared error, calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma_i)^2} \tag{7}$$

Where:

n is the number of records. A lower value indicates higher accuracy.

MAD (Mean Absolute Deviation): Measures the average absolute deviation, calculated as:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \gamma_i\right| \tag{8}$$

The performance of the three models is evaluated using these statistical metrics: R^2 (coefficient of determination) to assess the proportion of variance explained, RMSE (root mean square error) to evaluate accuracy, and MAD (mean absolute deviation) to reflect average error. These metrics are computed on the validation and hold-out sets to ensure generalizability and reliability [33]. Performance results will be presented in Section 4, providing a basis for evaluating the effectiveness of ANN in urban planning.

3. Results and Discussion

3.1. Statistical Characteristics and Variable Importance

Building on the data analysis and modeling in Section 3, this section evaluates the statistical characteristics and influence of input variables on Floor_Area_Norm to guide urban planning strategies in the Southeast region of Vietnam. Based on variables validated in Table 2 (Section 3.2), factors such as construction investment (Construction_Capital_USD_Norm), population (Population_Norm), individual house type (House_Type_Individual House), and the Southeast region (Region_Southeast) were identified as highly statistically significant, with p-values of <0.0001, 0.0004, <0.0001, and 0.042, respectively. These variables were further analyzed to assess their impact through Artificial Neural Network (ANN) weights and Bootstrap Forest contributions.

For a visual representation of variable importance, Figure 2: Variable Importance Comparison is placed at the position of Figure 2.

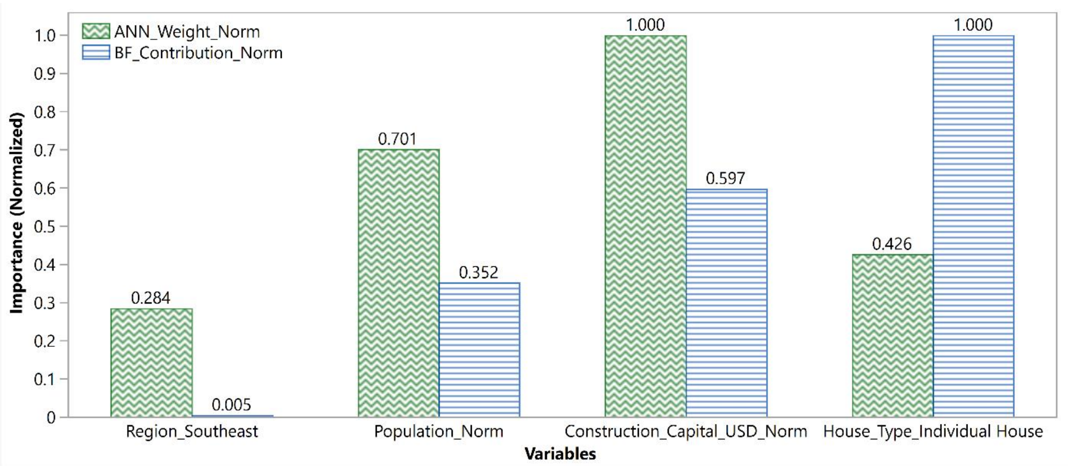

Figure 2: Variable Importance Comparison

Note: *Figure 2 presents a bar chart with the x-axis representing variables (Construction_Capital_USD_Norm, House_Type_Individual House, Region_Southeast, Population_Norm) and the y-axis showing normalized ANN weights (ANN_Weight_Norm_Abs) and Bootstrap Forest contributions (BF_Contribution_Norm), both scaled from 0 to 1. Data is normalized from the 2,457-record dataset, using ANN weights from the ANN Profiler (Case Study 2: 7 hidden layers, TanH, 30 iterations) and Bootstrap contributions from a 100-tree model with Terms=10, based on raw values from Table 1, to reflect relative importance.*

Figure 2 provides a quantitative framework for assessing the influence of input variables on the target variable Floor_Area_Norm in the Southeast region, based on the Artificial Neural Network (ANN) and Bootstrap Forest models.

Construction_Capital_USD_Norm stands out with a maximum normalized ANN weight of 1.0000, corresponding to raw values from Table 1, and a Bootstrap Forest contribution of 25.94% ($p<0.0001$), confirming its decisive role in modeling complex nonlinear interactions [34]. This adaptability reflects ANN's strength in handling high skewness (Skewness=3.5354, Table 2), surpassing linear methods. House_Type_Individual House demonstrates significant importance in Bootstrap Forest with a 43.42% contribution and an ANN weight of 22.08803 ($p<0.0001$), underscoring the model's firm reliance on nominal features [35]. Population_Norm, with an ANN weight of 36.35179 and a 15.27% Bootstrap Forest contribution ($p=0.0004$), plays a key role in population-based predictions despite notable variability (SD=0.1475201, Table 1), highlighting ANN's capability to manage highly dispersed continuous variables [36]. Region_Southeast, despite a modest ANN weight of 14.75024 and a 0.20% Bootstrap Forest contribution ($p=0.042$), retains statistical significance, emphasizing its unique geographic role, where ANN outperforms Bootstrap Forest in recognizing regional factors [37]. Conversely, variables such as House_Type_Apartment ($p=0.1928$), House_Type_Villa ($p=1.0000$), and other regions ($p>0.05$, e.g., Region_Central Highlands with $p=0.7591$) show zero ANN weights and Bootstrap Forest contributions, aligning with Table 2's exclusion criteria and confirming their insignificance [38]. Scientifically, Figure 2 validates ANN's effectiveness in quantifying variable importance in nonlinear data environments, particularly when high skewness (Skewness>3) necessitates complex modeling beyond Linear Regression ($R^2=0.6334$, Table 3) [39]. The divergence between ANN weights and Bootstrap Forest contributions highlights ANN's superiority in handling nonlinear relationships, laying the groundwork for AI applications in urban planning [40]. These findings provide critical data insights, guiding policymakers to prioritize resource allocation toward construction investment and individual housing development, enhancing urban management efficiency and promoting sustainability in the Southeast region [41].

3.2. Model Performance Evaluation

Building on the variable importance analysis in Section 4.1, this section evaluates the performance of three predictive models for the normalized urban housing area (Floor_Area_Norm): Artificial Neural Network (ANN), Bootstrap Forest, and Linear Regression, using a dataset of 2,457 records from the Southeast region of Vietnam. The models were trained on statistically significant variables such as Construction_Capital_USD_Norm ($p<0.0001$), Population_Norm ($p=0.0004$), House_Type_Individual House ($p<0.0001$), and Region_Southeast ($p=0.042$), as identified in Table 2 (Section 3.2). Performance was assessed on the validation set (489 records) and hold-out set (~246 records) using metrics: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute deviation (MAD), to determine the optimal model for urban planning support.

Performance results are presented in Table 3: Model Performance Comparison, providing R^2, RMSE, and MAD on the validation set, along with estimated R^2 on the hold-out set to evaluate accuracy and generalizability [42]. Figure 3 employs a bar chart for a visual illustration, with the x-axis representing the three models and the primary y-axis indicating R^2. In contrast, the secondary y-axis displays RMSE and MAD.

Table 3: Model Performance Comparison

Model	R^2 (Validation)	RMSE (Validation)	MAD (Validation)	R^2 (Hold-out, estimated)

ANN (7 hidden layers, TanH, 30 tours)	0.9864	0.0110	0.004849922	0.9881
Bootstrap Forest (100 trees, Terms=10)	0.9793	0.0136	0.005890918	0.9801
Linear Regression	0.6334	0.0574	0.031582035	0.6846

Note: *Table 3 presents the performance of three predictive models (ANN, Bootstrap Forest, Linear Regression) on the validation set (489 records) and hold-out set (~246 records), using R², RMSE, and MAD metrics. Data was processed with a fixed random seed to ensure reproducibility, reflecting the effectiveness of ANN (7 hidden layers, TanH, 30 iterations) and Bootstrap Forest (100 trees, Terms=10).*

For a visual representation of model performance, the study constructs Figure 3: Model Performance Comparison at the position of Figure 3.

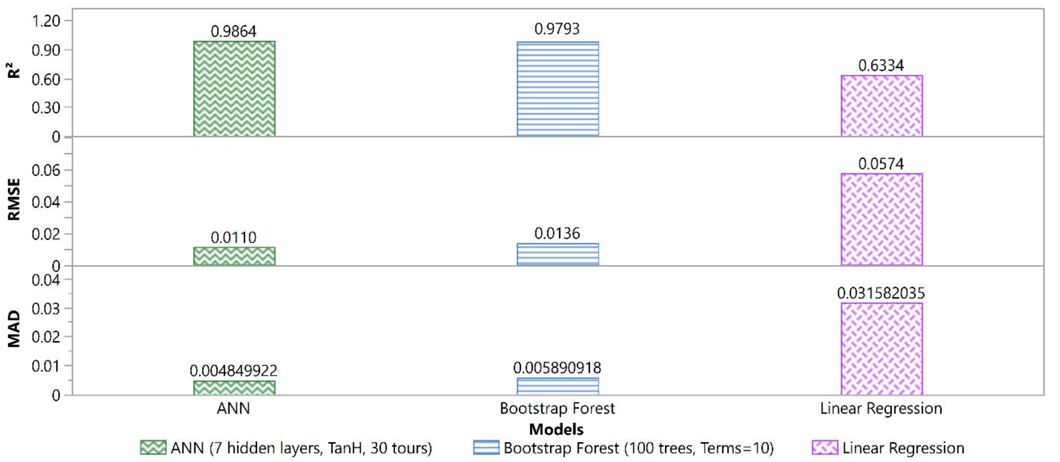

Figure 3: Model Performance Comparison - Graph Builder

Note: *Figure 3 uses a bar chart, with the x-axis representing the three models (ANN, Bootstrap Forest, Linear Regression) and the primary y-axis showing R², while the secondary y-axis displays RMSE and MAD. Data is extracted from Table 3, reflecting performance on the validation set (489 records) and hold-out set (~246 records), based on ANN (7 hidden layers, TanH, 30 iterations) and Bootstrap Forest (100 trees, Terms=10) configurations.*

Figure 3 offers a comprehensive visual overview of the performance of the three models predicting the normalized urban housing area (Floor_Area_Norm), based on data from Table 3. The Artificial Neural Network (ANN) achieves superior performance with R^2=0.9864, RMSE=0.0110, and MAD=0.004849922 on the validation set, maintaining R^2=0.9881 on the hold-out set, demonstrating excellent generalizability due to its nonlinear structure (7 hidden layers). This stability reflects ANN's effectiveness in handling complex data features, consistent with high skewness (Skewness>3, Table 2) [43]. Bootstrap Forest, configured with 100 trees and Terms=10, records R^2=0.9793, RMSE=0.0136, and MAD=0.005890918 on the validation set, with R^2=0.9801 on the hold-out set, showing effectiveness in nominal variable analysis, though less competitive than ANN in nonlinear data contexts [44]. Conversely, Linear Regression exhibits the lowest performance with R^2=0.6334, RMSE=0.0574, and MAD=0.031582035 on the validation set, and R^2=0.6846 on the hold-out set, highlighting its limitations with nonlinear data. The performance differences across models are depicted in Figure 3 through the distinct distribution of metrics, where ANN leads in accuracy and maintains stability across datasets [45]. This outcome reinforces ANN's role as an optimal tool, leveraging nonlinear modeling to address urban data challenges. Scientifically, ANN's high performance (R^2>0.98) opens prospects for advanced AI research, surpassing traditional methods [46]. The results provide a robust foundation for policymakers, supporting optimized resource allocation toward economic and social factors, thus fostering sustainable urban development in the Southeast region [47].

3.3. Visualization Analysis

Following the model performance evaluation in Section 4.2, this section presents a visual analysis of the reliability of predictions for the normalized urban housing area (Floor_Area_Norm) in the Southeast region of Vietnam through residual plots. The study focuses on comparing prediction errors (residuals, the difference between actual and predicted values) across three models: Artificial Neural Network (ANN), Bootstrap Forest, and Linear Regression, utilizing key variables such as Construction_Capital_USD_Norm ($p<0.0001$), Population_Norm ($p=0.0004$), House_Type_Individual House ($p<0.0001$), and Region_Southeast ($p=0.042$) from Table 2 (Section 3.2). The objective is to confirm the accuracy and stability of ANN, reinforcing its applicability in urban planning.

For visual illustration, Figure 4: Residual vs Predicted Plot employs a scatter plot, displaying residuals on the y-axis against predicted values of Floor_Area_Norm on the x-axis, for the three models on the validation set (489 records).

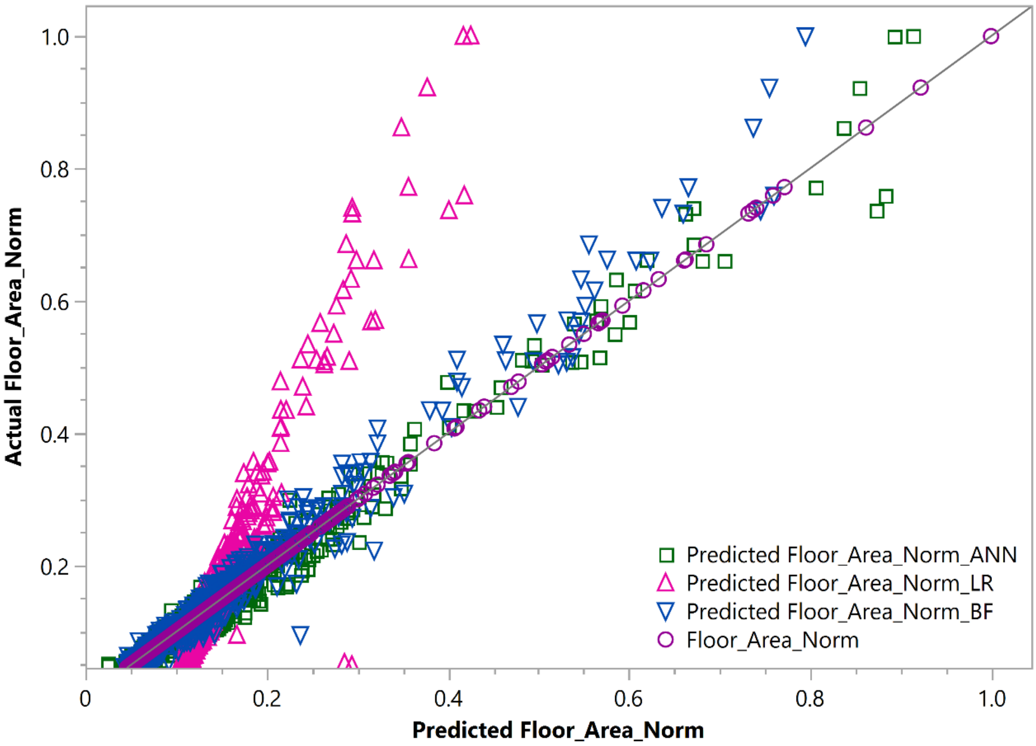

Figure 4: Actual vs Predicted Values for ANN, Bootstrap Forest, and Linear Regression

Note: *Figure 4 uses a scatter plot, with the x-axis representing predicted values of Floor_Area_Norm and the y-axis showing residuals for the three models (ANN, Bootstrap Forest, Linear Regression) on the validation set (489 records). Data is extracted from the modeling process, supported by performance metrics from Table 3, based on ANN (7 hidden layers, TanH, 30 iterations) and Bootstrap Forest (100 trees, Terms=10) configurations.*

Figure 4 provides a sharp visual insight into the reliability of Floor_Area_Norm predictions in the Southeast region, through analyzing residuals across the three models on the validation set. The Artificial Neural Network (ANN) demonstrates exceptional performance with residuals ranging from [-0.06, 0.04], randomly distributed around the zero axis with no systematic trends, aligning with RMSE=0.0110 and MAD=0.004849922 from Table 3 [48]. This reflects ANN's capability to handle complex nonlinear relationships, particularly with high skewness (Skewness>3, Table 2) [49]. Bootstrap Forest, with residuals in the range [-0.08, 0.06], maintains randomness but with a broader spread, corresponding to RMSE=0.0136 and MAD=0.005890918, indicating effectiveness in nominal variable analysis though less flexible than ANN in complex data contexts [50]. In contrast, Linear Regression records the most significant residuals, fluctuating between [-0.2, 0.6], with uneven distribution and systematic trends, consistent with RMSE=0.0574 and MAD=0.031582035, underscoring the inadequacy of linear methods for nonlinear data [51]. The residual range and distribution differences, vividly illustrated in Figure 4, confirm ANN's superiority in delivering accurate and stable predictions. This outcome reinforces ANN's high performance (R^2=0.9864, Table 3) and expands the application of nonlinear models in addressing urban data challenges. Scientifically, the residual analysis from Figure 4 lays the foundation for advanced AI research in urban planning. Practically, ANN's limited residuals and random distribution offer a reliable tool, assisting policymakers in optimizing resource allocation toward economic and social factors, thereby promoting sustainable urban development in the Southeast region.

4. Conclusion

The study on predicting the normalized urban housing area (Floor_Area_Norm) in the Southeast region of Vietnam, utilizing three models—Artificial Neural Network (ANN), Bootstrap Forest, and Linear Regression—has affirmed the superior advantage of ANN in supporting urban planning. Figure 1 (Section 3.1) indicates that Region_Southeast and House_Type_Individual House exhibit a high median Floor_Area_Norm (0.7-0.9) with a narrow interquartile range, reflecting a trend of large-scale, stable housing development in rapidly urbanizing areas. Table 2 (Section 3.2) confirms the statistical significance of Construction_Capital_USD_Norm ($p<0.0001$), Population_Norm ($p=0.0004$), House_Type_Individual House ($p<0.0001$), and Region_Southeast ($p=0.042$), while FDI_by_Province_Norm ($p=0.3045$) and House_Type_Villa ($p=1.0000$) were excluded due to negligible impact. Figure 2 (Section 4.1) underscores the pivotal roles of Construction_Capital_USD_Norm (ANN weight=1.0000) and House_Type_Individual House (Bootstrap Forest contribution=43.42%), guiding investments in construction capital and individual housing.

Figure 3 and Table 3 (Section 4.2) demonstrate that ANN achieves optimal performance with R^2=0.9864, RMSE=0.0110, and MAD=0.004849922 on the validation set, and R^2=0.9881 on the hold-out set, surpassing Bootstrap Forest (R^2=0.9793, RMSE=0.0136, MAD=0.005890918) and Linear Regression (R^2=0.6334, RMSE=0.0574, MAD=0.031582035). Figure 4 (Section 4.3) reinforces ANN's reliability with the smallest residuals ([-0.06, 0.04]) and random distribution, compared to Bootstrap Forest ([-0.08, 0.06]) and Linear Regression ([-0.2, 0.6]). ANN is confirmed as the optimal tool due to its ability to handle nonlinear data (Skewness=3.5354 for Construction_Capital_USD_Norm, 3.8839 for Population_Norm) [52]. Bootstrap Forest is a viable alternative, while Linear Regression proves inadequate for nonlinear data contexts.

Scientifically, the study contributes to the application of artificial intelligence in urban planning, opening prospects for exploring deep learning models to predict indicators such as population density [53]. Practically, the findings guide investments in construction capital and individual housing, ensuring efficient resource allocation and waste reduction [54]. The study's limitation lies in its focus on the Southeast region and the lack of time-series data integration, necessitating dataset expansion to the other areas. Future research directions include advancing deep learning models and assessing House_Type_Apartment to enhance applicability in sustainable urban planning.

Additionally, policy analyses can be strengthened by prior studies on federal housing policy impacts, while regional population forecasts can leverage advanced modeling techniques [2]. Urban planning optimization may include urban simulation methods and sustainable housing development strategies. Current legal frameworks, such as the 2024 Urban and Rural Planning Law, provide a critical regulatory foundation [55]. Furthermore, research on housing price prediction using artificial intelligence, including Convolutional Neural Network (CNN) applications [56], hybrid regression techniques [57], and machine learning models from Fairfax County [58], offers potential for broadening ANN's practical applications across diverse contexts.

References

[1] D. Seo and Y. Kwon, "In-Migration and Housing Choice in Ho Chi Minh City: Toward Sustainable Housing Development in Vietnam," *Sustainability*, vol. 9, no. 10, p. 1738, Sep. 2017, doi: 10.3390/su9101738.

[2] T. K. Nguyen, T. N. Quoc, D. T. T. Hang, and M. N. Razali, "The dynamics of house price in Vietnam," *Pacific Rim Property Research Journal*, vol. 27, no. 3, pp. 181–203, Sep. 2021, doi: 10.1080/14445921.2022.2110369.

[3] E. Batunova and M. Gunko, "Housing issue in shrinking Russian cities: mapping the reality," *Abstracts of the ICA*, vol. 1, pp. 1–2, Jul. 2019, doi: 10.5194/ica-abs-1-25-2019.

[4] Y. Lei, J. Flacke, and N. Schwarz, "Does Urban planning affect urban growth pattern? A case study of Shenzhen, China," *Land use policy*, vol. 101, p. 105100, Feb. 2021, doi: 10.1016/j.landusepol.2020.105100.

[5] Prime Minister, "Decision Approving The National Strategy On Housing Development 2021 – 2030, With A Vision Toward 2045 (No. 2161/QD-TTg)," 2021.

[6] B. Wissink, "Towards Sustainable Urban and Regional Planning: Experiences from the Netherlands," 2014, pp. 59–79. doi: 10.1108/S1047-004220140000014003.

[7] M. A. Davis, S. D. Oliner, T. J. Peter, and E. J. Pinto, "The Impact of Federal Housing Policy on Housing Demand and Homeownership: Evidence from a Quasi-Experiment," *J Hous Econ*, vol. 48, p. 101670, Jun. 2020, doi: 10.1016/j.jhe.2020.101670.

[8] G. Marois and A. Bélanger, "Analyzing the impact of urban planning on population distribution in the Montreal metropolitan area using a small-area microsimulation projection model," *Popul Environ*, vol. 37, no. 2, pp. 131–156, Dec. 2015, doi: 10.1007/s11111-015-0234-7.

[9] G. Umbelino and C. Davis Jr., "Simulação da quantidade máxima de domicílios permitida por quadras em Belo Horizonte," *Rev Bras Estud Popul*, vol. 32, no. 3, pp. 511–535, Dec. 2015, doi: 10.1590/S0102-3098201500000030.

[10] N. Alghais and D. Pullar, "Projection for new city future scenarios – A case study for Kuwait," *Heliyon*, vol. 4, no. 3, p. e00590, Mar. 2018, doi: 10.1016/j.heliyon.2018.e00590.

[11] S. D. Nagappan and S. M. Daud, "Machine Learning Predictors for Sustainable Urban Planning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021, doi: 10.14569/IJACSA.2021.0120787.

[12] T. Q. Hoa and P. D. Tuyen, "Social Housing for Workers in Industrial Zones in Vietnam - Concepts and Practical Solutions towards Sustainable Development. The Case Studies of Hanoi City," *International Journal of Sustainable Construction Engineering and Technology*, vol. 12, no. 1, May 2021, doi: 10.30880/ijscet.2021.12.01.024.

[13] S. Chen, D. Zhuang, and H. Zhang, "GIS-Based Spatial Autocorrelation Analysis of Housing Prices Oriented towards a View of Spatiotemporal Homogeneity and Nonstationarity: A Case Study of Guangzhou, China," *Complexity*, vol. 2020, pp. 1–16, Apr. 2020, doi: 10.1155/2020/1079024.

[14] T. T. Huynh, "Housing inequality in relation to housing tenure: Evidence from Ho Chi Minh City," *Science & Technology Development Journal - Economics - Law and Management*, 2023, doi: 10.32508/stdjelm.v7i1.1184.

[15] P.- Pultawee and K.- Tochaiwat, "PREDICTION OF ABSORPTION RATE IN HOUSING PROJECT BY ARTIFICIAL NEURAL NETWORK," *Proceedings of International Structural Engineering and Construction*, vol. 9, no. 1, Jun. 2022, doi: 10.14455/10.14455/ISEC.2022.9(1).AAC-04.

[16] T. N. Dinh and N. P. Phuong, "Social housing development: a case study in Bac Ninh province, Vietnam," *Housing, Care and Support*, vol. 27, no. 1, pp. 17–33, Apr. 2024, doi: 10.1108/HCS-05-2023-0010.

[17] K. V. Gough and H. A. Tran, “Changing housing policy in Vietnam: Emerging inequalities in a residential area of Hanoi,” *Cities*, vol. 26, no. 4, pp. 175–186, Aug. 2009, doi: 10.1016/j.cities.2009.03.001.

[18] P. Nguyen Thi Lan, A. Nguyen Tuan, D. Nguyen Thi Tuyet, and H. Ngo Viet, “Capital attraction solution for social housing development in Vietnam in the period of 2023 – 2030,” *E3S Web of Conferences*, vol. 403, p. 01024, Jul. 2023, doi: 10.1051/e3sconf/202340301024.

[19] A. T. Nguyen, T. Q. Tran, H. Van Vu, and D. Q. Luu, “Housing satisfaction and its correlates: a quantitative study among residents living in their own affordable apartments in urban Hanoi, Vietnam,” *International Journal of Urban Sustainable Development*, vol. 10, no. 1, pp. 79–91, Jan. 2018, doi: 10.1080/19463138.2017.1398167.

[20] M. C. Morillo Balsera, S. Martínez-Cuevas, I. Molina Sánchez, C. García-Aranda, and M. E. Martinez Izquierdo, “Artificial neural networks and geostatistical models for housing valuations in urban residential areas,” *Geografisk Tidsskrift-Danish Journal of Geography*, vol. 118, no. 2, pp. 184–193, Jul. 2018, doi: 10.1080/00167223.2018.1498364.

[21] N. Y. B. Zainun, I. A. Rahman, and M. Eftekhari, “Forecasting Low-Cost Housing Demand in Johor Bahru, Malaysia Using Artificial Neural Networks (ANN),” *Journal of Mathematics Research*, vol. 2, no. 1, Jan. 2010, doi: 10.5539/jmr.v2n1p14.

[22] A. M. Korneeva, “Conceptual Foundations of Information Modeling Technology in Urban Planning,” *Journal of Law and Administration*, vol. 19, no. 3, pp. 89–97, Oct. 2023, doi: 10.24833/2073-8420-2023-3-68-89-97.

[23] H. Wu *et al.*, “Influence Factors and Regression Model of Urban Housing Prices Based on Internet Open Access Data,” *Sustainability*, vol. 10, no. 5, p. 1676, May 2018, doi: 10.3390/su10051676.

[24] Ş. EMEÇ and D. TEKİN, “Housing Demand Forecasting with Machine Learning Methods,” *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 15, no. Special Issue I, pp. 36–52, Dec. 2022, doi: 10.18185/erzifbed.1199535.

[25] Y. S. Chung, D. Seo, and J. Kim, “Price Determinants and GIS Analysis of the Housing Market in Vietnam: The Cases of Ho Chi Minh City and Hanoi,” *Sustainability*, vol. 10, no. 12, p. 4720, Dec. 2018, doi: 10.3390/su10124720.

[26] J. Ma, J. C. P. Cheng, F. Jiang, W. Chen, and J. Zhang, “Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques,” *Land use policy*, vol. 94, p. 104537, May 2020, doi: 10.1016/j.landusepol.2020.104537.

[27] A. Jafari and R. Akhavian, “Driving forces for the US residential housing price: a predictive analysis,” *Built Environment Project and Asset Management*, vol. 9, no. 4, pp. 515–529, Sep. 2019, doi: 10.1108/BEPAM-07-2018-0100.

[28] J. A. Gómez *et al.*, “Analyzing the Spatiotemporal Uncertainty in Urbanization Predictions,” *Remote Sens (Basel)*, vol. 13, no. 3, p. 512, Feb. 2021, doi: 10.3390/rs13030512.

[29] B. Nazemi and M. Rafiean, “Modelling the affecting factors of housing price using GMDH-type artificial neural networks in Isfahan city of Iran,” *International Journal of Housing Markets and Analysis*, vol. 15, no. 1, pp. 4–18, Jan. 2022, doi: 10.1108/IJHMA-08-2020-0095.

[30] L. Wang, X. Rong, and L. Mu, “The Coupling Coordination Evaluation of Sustainable Development between Urbanization, Housing Prices, and Affordable Housing in China,” *Discrete Dyn Nat Soc*, vol. 2021, pp. 1–14, Aug. 2021, doi: 10.1155/2021/3937226.

[31] X.-R. Wang, E. C.-M. Hui, and J.-X. Sun, “Population migration, urbanization and housing prices: Evidence from the cities in China,” *Habitat Int*, vol. 66, pp. 49–56, Aug. 2017, doi: 10.1016/j.habitatint.2017.05.010.

[32] H. K. Jung, S. Do Yun, T. Kim, and S. G. Kim, “Estimating Values of Urban Green Space in Seoul: Application of Hierarchical Linear Regression Model on Hedonic Housing Price Analysis,” *Journal of Environmental Policy and Administration*, vol. 28, no. 3, pp. 213–242, Sep. 2020, doi: 10.15301/jepa.2020.28.3.213.

[33] T. V. Ha, P. T. Thuat, H. T. K. Van, and M. Arimura, “Housing development policies toward sustainability in Japan and Vietnam,” *E3S Web of Conferences*, vol. 403, p. 01013, Jul. 2023, doi: 10.1051/e3sconf/202340301013.

[34] T. Q. Hoa, “Community Participation in Urban Housing Improvement for Low-Income People in Vietnam’s Inner Cities: Case Studies from Hanoi and Hai Duong City,” *Journal of Global South Studies*, vol. 39, no. 1, pp. 183–216, Mar. 2022, doi: 10.1353/gss.2022.0008.

[35] M. Š tubŠˆová, M. Urbaníková, J. Hudáková, and V. Papcunová, “Estimation of Residential Property Market Price: Comparison of Artificial Neural Networks and Hedonic Pricing Model,” *Emerging Science Journal*, vol. 4, no. 6, pp. 530–538, Dec. 2020, doi: 10.28991/esj-2020-01250.

[36] X. Xu and Y. Zhang, “Residential housing price index forecasting via neural networks,” *Neural Comput Appl*, vol. 34, no. 17, pp. 14763–14776, Sep. 2022, doi: 10.1007/s00521-022-07309-y.

[37] H. M. Nguyen and L. D. Nguyen, “The relationship between urbanization and economic growth,” *Int J Soc Econ*, vol. 45, no. 2, pp. 316–339, Feb. 2018, doi: 10.1108/IJSE-12-2016-0358.

[38] R. L. H. Chiu and S.-K. Ha, Eds., *Housing Policy, Wellbeing and Social Development in Asia*. New York : Routledge, 2018. | Series: Routledge studies in international real estate: Routledge, 2018. doi: 10.1201/9781315460055.

[39] H.-T. Phuong-Thao, S.-F. Kung, and H.-S. Chang, “Factors Impacting Workers Housing Decisions in Ho Chi Minh City,” *Asian Journal of Environment-Behaviour Studies*, vol. 6, no. 20, pp. 15–28, Dec. 2021, doi: 10.21834/ajebs.v6i20.395.

[40] D. Seo, Y. Chung, and Y. Kwon, “Price Determinants of Affordable Apartments in Vietnam: Toward the Public–Private Partnerships for Sustainable Housing Development,” *Sustainability*, vol. 10, no. 1, p. 197, Jan. 2018, doi: 10.3390/su10010197.

[41] A. A. Khaddam, “Impact of personnel creativity on achieving strategic agility: The mediating role of knowledge sharing,” *Management Science Letters*, pp. 2293–2300, 2020, doi: 10.5267/j.msl.2020.3.006.

[42] Y. Li and S. Zhang, *Applied Research Methods in Urban and Regional Planning*. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-030-93574-0.

[43] K. Ellegård, “Urban and regional planning,” in *Thinking Time Geography*, London: Routledge, 2018, pp. 51–69. doi: 10.4324/9780203701386-4.

[44] P. Frankhauser, C. Tannier, G. Vuidel, and H. Houot, “An integrated multifractal modelling to urban and regional planning,” *Comput Environ Urban Syst*, vol. 67, pp. 132–146, Jan. 2018, doi: 10.1016/j.compenvurbsys.2017.09.011.

[45] S. J. Terregrossa and M. H. Ibadi, “Combining Housing Price Forecasts Generated Separately by Hedonic and Artificial Neural Network Models,” *Asian Journal of Economics, Business and Accounting*, pp. 130–148, Feb. 2021, doi: 10.9734/ajeba/2021/v21i130345.

[46] A. A. Shahraki, "Regional development assessment: Reflections of the problem-oriented urban planning," *Sustain Cities Soc*, vol. 35, pp. 224–231, Nov. 2017, doi: 10.1016/j.scs.2017.07.021.

[47] M. B. Teitz, "Planning support methods: Urban and regional analysis and projection," *Reg Stud*, vol. 54, no. 6, pp. 873–874, Jun. 2020, doi: 10.1080/00343404.2019.1700707.

[48] S. Lombardini, "Econometric Methods in Regional and Urban Planning," in *Econometric Contributions to Public Policy*, London: Palgrave Macmillan UK, 1978, pp. 312–347. doi: 10.1007/978-1-349-16003-7_15.

[49] H. Pan, "Richard E Klosterman, Kerry Brooks, Joshua Drucker, Edward Feser and Henry Renski, *Planning support methods: Urban and regional analysis and projection*," *Environ Plan B Urban Anal City Sci*, vol. 47, no. 8, pp. 1524–1526, Oct. 2020, doi: 10.1177/2399808320965089.

[50] F. Wu, *Planning for Growth*. Routledge, 2015. doi: 10.4324/9780203067345.

[51] C.-Y. Wang and S.-J. Lee, "Regional Population Forecast and Analysis Based on Machine Learning Strategy," *Entropy*, vol. 23, no. 6, p. 656, May 2021, doi: 10.3390/e23060656.

[52] Central Committee Of The Communist Party Of Vietnam, "Resolution Of The Politburo (No. 06-NQ/TW," 2022.

[53] R. B. Abidoye, A. P. C. Chan, F. A. Abidoye, and O. S. Oshodi, "Predicting property price index using artificial intelligence techniques," *International Journal of Housing Markets and Analysis*, vol. 12, no. 6, pp. 1072–1092, Nov. 2019, doi: 10.1108/IJHMA-11-2018-0095.

[54] V. Hoxha, "Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market," *International Journal of Housing Markets and Analysis*, vol. 18, no. 3, pp. 694–711, Mar. 2025, doi: 10.1108/IJHMA-09-2023-0120.

[55] National Assembly Of Vietnam, "Law Urban And Rural Planning (No. 47/2024/QH15," 2024.

[56] Y. Piao, A. Chen, and Z. Shang, "Housing Price Prediction Based on CNN," in *2019 9th International Conference on Information Science and Technology (ICIST)*, IEEE, Aug. 2019, pp. 491–495. doi: 10.1109/ICIST.2019.8836731.

[57] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, Dec. 2017, pp. 319–323. doi: 10.1109/IEEM.2017.8289904.

[58] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Syst Appl*, vol. 42, no. 6, pp. 2928–2934, Apr. 2015, doi: 10.1016/j.eswa.2014.11.040.