Comparative Study on Artificial Intelligence Methods in Housing Price Prediction

by Pengolahan Perpustakaan

Submission date: 04-Aug-2025 10:26PM (UTC+0700)
Submission ID: 2626587390
File name: 22747-54954-1-PB_Journal_Paper.pdf (646.12K)
Word count: 5759
Character count: 29304

JCE

Journal of Civil Engineering
ISSN 2086 - 1206 E-ISSN 2579 - 9029

Comparative Study on Artificial Intelligence Methods in Housing Price Prediction

Willy Husada[a*], Ambrosius Matthew Junius Reynaldo[a], Josh F. Hogianto[a], Clarissa A. Putri[a]

Correspondence

[a]Department of Civil Engineering, Petra Christian University, 60236 Surabaya, Indonesia.

Corresponding author email adress: willy.husada@petra.ac.id.

Submitted : 27 Maret 2025
Revised : 15 June 2025
Accepted : 25 June 2025

Abstract

The demand for property, including houses, continues to grow rapidly in Indonesia. The housing price prediction is essential in assisting the stakeholders such as buyers, sellers, and investors to make better decision-making. There are many key factors that influencing the housing prices and it is challenging to identify the most relevant factors. This study provides a comparative analysis of various methods in the housing price prediction that consists of one traditional method, Linear Regression (LR), and three artificial intelligence (AI) methods, including Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Chi-Squared Automatic Interaction Detection (CHAID). The aim is to find the best machine learning method in predicting the housing price in terms of prediction accuracy through the four performance indicators and one combined performance index called the reference index (RI). The main findings of this study is that the AI-based method, the ANN method, has the best accuracy indicated by its highest RI value hence outperforming other methods in predicting the housing prices.

Keywords

Artificial intelligence, housing price, machine learning, prediction

INTRODUCTION

The property industry continues to grow rapidly in Indonesia, establishing its position as an important pillar of economic growth. In 2023, the real estate activities contributed approximately around 2.4 percents to the nation's GDP [1]. The demand for property, including houses, has led to rapid growth in this sector. The high demand for housing is partly driven by its dual purpose, serving both as a living place and as an investment alternative. The housing market plays an important role in the global economy, influencing financial stability, consumer wealth, and investment strategies. Accurate prediction of the housing prices is essential for stakeholders such as buyers, sellers, and investors. Predictive analytics is used in many industries worldwide because it helps handle uncertainties and supports better decision-making. In the real estate industry, it helps home buyers estimate the property prices before making a purchase and assists developers in setting the appropriate prices by considering factors like location, size, and accessibility [2]. When purchasing a property, the stakeholders consider several factors, with house prices being influenced by social, economic, political, and environmental-physical variables [3]. Many studies have explored the factors influencing property prices. Fahirah et al. identified key factors, including physical, economic, social, regulatory, and accessibility factors [4]. Similarly, Rahadi et al. classified price-affecting elements into five categories: marketing concepts, design concepts, location accessibility, location uniqueness, and physical conditions [5]. Olanrewaju et al. highlighted factors such as location, house size, strategic value, land prices, building material quality, and land ownership type [6]. Meanwhile, Zulkifli & Ismail grouped these factors into two main categories: internal and external factors that shape homebuyers' decisions [7]. Although numerous studies have explored house price prediction and its influencing factors, many do not take account for other factors such as electricity capacity, building configuration, distance to the church or mosque, distance to the city centre, distance to the hospital, and distance to the nearest toll gates. Identifying the most relevant factors in prediction cases can be very difficult, but this is where artificial intelligence excel at.

In recent years, the development of big data analysis has made artificial intelligence especially machine learning a crucial predictive tool, widely applied in many scientific fields, including civil engineering. Cheng et al. used machine learning to predict the compressive strength of high-performance concrete [8]. Similarly, Hore et al. used a multilayer perceptron feed-forward network to predict structural failures of multi-story reinforced concrete buildings [9]. Additionally, Cheng et al. performed cash flow prediction for construction projects using an adaptive time-dependent least squares support vector machine [10]. Research on housing price prediction has been conducted with several methods, both traditional and machine learning methods. Traditional methods like linear regression (LR) [11], alongside advanced machine learning

models, including artificial neural network (ANN) [12] and random forest (RF) [13].

RESEARCH SIGNIFICANCE

This study aims to evaluate the performance of four machine learning methods, namely linear regression (LR), artificial neural network (ANN), classification and regression tree (CART), and chi-squared automatic interaction detection (CHAID), by analyzing the key factors that influence the housing price in Surabaya as the case study. The goal is to determine which model provides the most accurate housing price prediction, making it a valuable reference for property investment calculations.

METHODOLOGY

This study has two main objectives, first is to evaluate the four prediction models, namely Linear Regression (LR), Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Chi-Squared Automatic Interaction Detection (CHAID) individually in predicting the housing prices. The second objective is to compare the prediction performance of every method to find the best model in predicting the housing prices. 19 variables or factors are carried out to create the prediction models to predict the housing price. The models will be evaluated by using four performance indicators, Linear Correlation Coefficient (R), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and one performance index, reference index (RI).

A. FACTORS INFLUENCING HOUSING PRICES

The housing prices are influenced by a combination of various factors, which have been examined in numerous studies using different prediction methods and scopes. Researchers have identified key elements affecting property values from multiple sources, including the Appraisal Institute [3] and studies by Fahirah et al. [4], Rahadi et al. [5], Olanrewaju et al. [6], and Zulkifli & Ismail [7]. These sources highlight a range of variables, such as physical attributes, accessibility, and government regulations all of which shape property prices. In this study, relevant factors from these sources have been selected and summarized in Table 1, serving as the basis for evaluating housing price predictions using different machine learning models.

Table 1 Factors Influencing Housing Prices

Indicator	Unit
X1: Land Area	m^2
X2: Building Area	m^2
X3: Number of Bedrooms	Unit
X4: Number of Bathrooms	Unit
X5: Number of Family Rooms	Unit
X6: Number of Carports	Unit
X7: Electricity Capacity	1: 2200 watt 2: 3500 watt 3: 4400 watt
X8: Building Configuration	0: Not Hook 1: Hook

Table 1 Factors Influencing Housing Prices (cont.)

Indicator	Unit
X9: Distance to Nearest Primary School	km
X10: Distance to Nearest Junior High School	km
X11: Distance to Nearest Senior High School	km
X12: Distance to Nearest University	km
X13: Distance to Nearest Shopping Centre	km
X14: Distance to Nearest Church or Mosque	km
X15: Distance to City Centre	km
X16: Distance to Nearest Hospital	km
X17: Distance to Nearest Toll Gate	km
X18: Region	1: North 2: East 3: West
X19: Land Ownership Type	0: HGB 1: SHM

B. PREDICTION METHODS

- **Linear Regression (LR)**

Identifying the relationship between multiple variables can be difficult due to the technical and analytical complexities. Regression analysis is very effective in solving these situations and widely used across many research fields [14,15]. The LR method purpose is to reduce the distance between data points and the predetermined regression line, ensuring they are as close as possible. In simple LR, a single explanatory variable is used, whereas multiple linear regression (MLR) expands the simple LR by considering the multiple explanatory variables to predict the response variable and influence the mean function [16,17]. The formulation of MLR can be seen in Equation 1. Y is the dependent variable, $X_1, X_2, \ldots, X_i$ are the independent variables, a is constant, and $b_1, b_2, \ldots, b_i$ are the regression coefficients. The equation served as a guideline for predicting the housing prices using linear regression in this study. The dependent variable (Y) represents the house price, while the independent variables represent the key factors such as land area, building area, number of bedrooms, and other relevant attributes.

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_iX_i \quad (1)$$

- **Artificial Neural Network (ANN)**

Artificial Neural Network (ANN) is an Artificial Intelligence (AI) method first introduced by Rosenblatt in 1958. ANN was developed through mathematical modelling of the learning process, inspired by the human brain [18]. It replicates the brain's biological neural networks by learning, retaining information, and generalizing patterns. This method works by simulating data as neural networks in the brain to process all obtained information [19]. ANN is also often referred to as a

Multilayer Perceptron Network (MLPN), which consists of three layers: the input layer, the hidden layer, and the output layer [25]. The input layer will receive all available information or data and then transmit it to the hidden layer, where it is processed with weight and bias values. The output layer represents the final result of the calculations performed. In this research, the input layer represents the indicator component of the housing price, the hidden layer describes the calculation points, and the output layer shows the predicted housing price. The active neurons in the hidden layer can be represented by Equations 2 and 3. I_j is the activation value of neuron j, w_{ij} is the weight between neuron i and j, $f(I_j)$ is the transfer function, x_i, y_j and θ_j are respectively input, output, and bias.

$$I_j = \sum w_{ij}\, x_i + \theta_j \quad (2)$$

$$y_j = f(I_j) \quad (3)$$

- **Classification and Regression Tree (CART)**

Classification and Regression Tree (CART) is a machine learning method used to build predictive models from available data. This method was first proposed by Breiman and demonstrates that the learning trees can be optimized by using a model to prune the saturated trees and select from the remaining trees [21]. As a type of decision tree, CART constructs either a classification tree or a regression tree, depending on whether the target variable is categorical or numerical [22]. The development of classification tree in CART involves three key stages, including tree construction, tree pruning, and determining the optimal tree.

In the tree construction stage, three main processes take place: splitting, terminal node determination, and calculation of misclassification error. The splitting process relies on a heterogeneity function called Gini index where the set with the lowest Gini index is chosen as the root node. The Gini index function can be seen in Equation 4. The terminal nodes are determined through an iterative splitting process until a predefined stopping condition is met. The misclassification error is calculated using Equation 5. p_{mk} is the probability of an object being classified into a particular class

$$R(T) = 1 - \sum_{i=1}^{C}(p_{mk})^2 \quad (4)$$

$$miss = 1 - max\, p_{mk} \quad (5)$$

During the tree pruning stage, managing tree size is crucial to balance the model performance. An excessively large tree can lead to overfitting, while an overly restricted tree may cause underfitting. To address this, the minimum cost complexity method is applied to prune the tree and determine the optimal tree size, as represented in Equation 6. $R_\alpha(T)$ is linear combination of cost and tree complexity (cost complexity), $R(T)$ is Gini index, and $\alpha|\tilde{T}|$ is tree complexity penalty.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad (6)$$

- **Chi-Squared Automatic Interaction Detection (CHAID)**

The Chi-Squared Automatic Interaction Detection (CHAID) algorithm, introduced by statistician Kass in the late 1970s, is a widely used statistical method for supervised learning in decision tree construction [23]. This technique is used to build a segmentation model, which divides a set of predictor variables into groups of two predictor variables according to a criterion [24]. The indicator used in CHAID is the P-value, which is used to find the best variable among the data and can be calculated with Equation 7.

$$P = Pr\,(x_d^e > x^2) \quad (7)$$

A feature of this algorithm is that it can generate non-binary trees, which means that the number of splits has more than two branches [25]. The CHAID algorithm follows a structured approach to decision tree formation. It begins with a merging process, where a contingency table is created for each categorical variable, and the Chi-square test is applied to evaluate the relationship between pairs of categorical variables. The next step involves splitting, where the independent variable with the most statistically significant result, determined by the smallest p-value, is selected for data partitioning. These partitions are formed based on the comparison of p-values identified earlier in the process. Finally, the most significant value from the previous step is used to split nodes, ensuring that variable categories are appropriately combined. Throughout the process, Pearson's chi-squared test is employed to analyze the data relationships, as shown in Equation 8 and 9.

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} \quad (8)$$

$$n_{ij} = \sum_{n\in D} f_n I(x_n = i \cap y_n = j) \quad (9)$$

C. PERFORMANCE INDICATORS

In this study, the prediction model will be evaluated using four performance indicators and one combined performance index to evaluate the performance of each model including LR, ANN, CART, and CHAID. The indicators and their formulation are presented in Table 2. R is Linear Correlation Coefficient, MAE is Mean Absolute Error, RMSE is Root Mean Squared Error, MAPE is Mean Absolute Percentage Error, n is the sample size, y is the actual output value, and y' is the predicted output value.

Table 2 Performance Indicators

Model Performance Indicators	Mathematical Formula
R	$\frac{n\sum y.y' - (\sum y)(\sum y')}{\sqrt{n(\sum y^2) - (\sum y)^2}\sqrt{n(\sum y'^2) - (\sum y')^2}}$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y - y'\rvert$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y' - y)^2}$
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y - y'}{y}\right\rvert . 100\%$

To enable a comprehensive performance measurement, the three performance indicators used in the testing data are normalized to create a reference index (RI). This normalization process assigns a value of 1 to the model with the best accuracy and 0 to the least accurate one. The RI value is determined by averaging the normalized performance indicators, as outlined in Equation

10. Meanwhile, Equation 11 presents the function used for the data normalization.

$$RI = \frac{RMSE+MAE+MAPE}{3} \quad (10)$$

$$x_{norm} = \frac{(x_{max}-x_i)}{(x_{max}-x_{min})} \quad (11)$$

D. RESEARCH FRAMEWORK

The development of a machine learning model for the housing price prediction involves several steps. First, data preprocessing is carried out by selecting relevant factors that influencing the house prices and removing incomplete data. Then, the dataset is randomized before being divided for five-fold cross-validation, where it is splitted into five partitions of training and testing data. Next, parameter setting and parameter tuning is applied for each method, adjusting variables such as the number of neurons in neural networks or the depth of tree branches in decision trees.

After parameter tuning, the training process will generate a predictive model, which is then employed to the testing data to estimate the housing prices. Subsequently, the accuracy of these prediction models is assessed using the four performance indicators and one combined performance index, ensuring a comprehensive evaluation. Finally, the performance results of the four prediction models is compared to determine the best model in predicting the housing price. The complete research framework is illustrated in Figure 1.

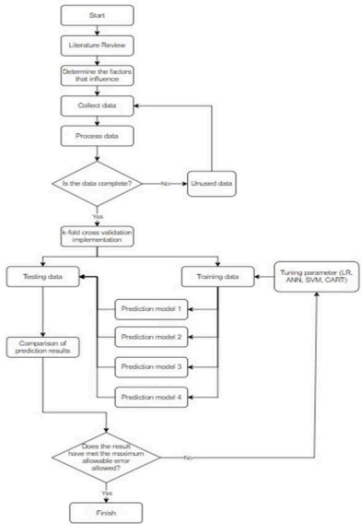

Figure 1 Research Framework

The prediction methods in this study were implemented using SPSS Modeler 18.0 software on a computer equipped with an Intel Core i7-12700H Processor at 2.30 GHz and 16 GB of RAM. To achieve the desired level of prediction accuracy, parameter tuning is necessary. In this study, parameter adjustments were applied to the three prediction models: artificial neural network (ANN), classification and regression tree (CART), and chi-squared automatic interaction detection (CHAID). Meanwhile, the linear regression (LR) method follows the default parameter settings provided by the SPSS Modeler 18.0 software. Table 3 outlines the standard parameter settings for each prediction method used in this research.

Table 3 Parameter Settings in SPSS Modeler 18.0

Method	Parameter	Settings
LR	Singularity Tolerance	1.00E-04
ANN	Neural Network Model	MLP
	Hidden Layer 1	1
	Hidden Layer 2	0
	Overfit Prevention Set (%)	30
CART	Maximum Tree Depth	5
	Maximum Surrogates	5
	Prune Tree	TRUE
	Minimum Change in Impurity	0.0001
	Impurity Measure	Gini
	Overfit Prevention Set (%)	30
CHAID	Significance Level for Splitting	0.05
	Significance Level for Merging	0.05
	Chi-Square for Categorical Targets	Pearson

RESULTS AND DISCUSSIONS

The performance results of the four prediction models, Linear Regression (LR), Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Chi-Squared Automatic Interaction Detection (CHAID) in predicting the housing prices have been obtained and reported. The comparative analysis of the Artificial Intelligence (AI) methods in terms of the prediction accuracy is discussed to find the best prediction model in predicting the housing prices.

A. DATA PREPROCESSING

The housing dataset used for the prediction was collected from various housing developers in Surabaya, East Java, Indonesia consisting of 105 data points. To provide an overview of the dataset, the descriptive statistics were generated to summarize the key indicators. These statistics will help in understanding the distribution and characteristics of the data. Table 4 presents the descriptive statistics of the housing dataset used in this study.

The first step in data preprocessing is implementing the five-fold cross-validation by dividing the 105 data points into 80% for training dataset and 20% for testing dataset. The dataset is then randomized using the random function in Microsoft Excel. Once shuffled, the dataset is split into five subsets (folds) and stored in separate Excel files, with each file contains 84 training data points and 21 testing data points. For the first training dataset, folds 2

Table 4 Descriptive Statistic of The Dataset

Indicator	Unit	Min	Max	Mean	Std
Price	Rupiah	891,300,000	6,244,000,000	2,476,051,374	1,200,864,999
X1: Land Area	m^2	40	210	101.167	41.646
X2: Building Area	m^2	50	284	119.330	51.505
X3: Number of Bedrooms	unit	2 units = 18.095% ; 3 units = 51.429% 4 units = 16.19% ; 5 units = 14.286%			
X4: Number of Bathrooms	unit	1 unit = 6.667% ; 2 unit = 45.714% ; 3 units = 23.81% 4 units = 16.19% ; 5 units = 7.619%			
X5: Number of Family rooms	unit	1 unit = 93.333% ; 2 units = 6.667%			
X6: Number of Carports	unit	1 unit = 40% ; 2 units = 60%			
X7: Electricity Capacity	1: 2200 watt 2: 3500 watt 3: 4400 watt	1 = 52.381% ; 2 = 31.429% ; 3 = 16.19%			
X8: Building Configuration	0: Not Hook 1: Hook	0 = 95.238% ; 1 = 4.762%			
X9: Distance to Nearest Primary School	km	0.432	9.660	2.923	2.618
X10: Distance to Nearest Junior High School	km	0.477	6.674	3.552	2.056
X11: Distance to Nearest Senior High School	km	0.440	9.023	3.920	2.598
X12: Distance to Nearest University	km	0.687	11.512	5.281	3.107
X13: Distance to Nearest Shopping Centre	km	1.661	10.683	5.736	2.238
X14: Distance to Nearest Church or Mosque	km	0.242	5.372	2.244	1.670
X15: Distance to City Centre	km	7.003	30.309	18.552	6.580
X16: Distance to Nearest Hospital	km	0.548	8.819	4.587	1.936
X17: Distance to Nearest Toll Gate	km	1.361	15.370	7.639	3.952
X18: Region	1: North 2: East 3: West	1 = 16.19% ; 2 = 37.143% ; 3 = 46.667%			
X19: Land Ownership Type	0: HGB 1: SHM	0 = 76.19% ; 1 = 23.81%			

dataset. Similarly, the second training dataset consists of folds 1 and 3 through 5, with fold 2 designated for the testing dataset. This process continues until all five datasets are created, as illustrated in Figure 2.

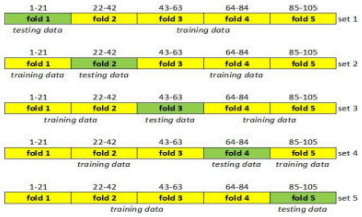

Figure 2 Five-Fold Cross-Validation Model

B. PREDICTION RESULTS

This section presents the prediction results for each machine learning method: LR, ANN, CART, and CHAID. The prediction errors are then calculated using the four performance indicators and one combined performance index. These error values help in determining which model provides the best accuracy in housing price prediction.

✓ **Linear Regression (LR)**

The LR model produces a linear function that defines the relationship between the indicators and the predicted housing price. Each indicator's coefficient influences the predicted price, either increasing or decreasing it, as shown in Table 5. The model demonstrates a strong correlation of 0.980, with the MAE of Rp 180,401,000, the RMSE of Rp 230,253,000, and the MAPE of 8.556%, as shown in Table 6. Figure 3 visualizes the correlation between the predicted and the actual housing prices. The solid linear line represents the model's accuracy, where the data points aligning with the line indicate a precise predictions. While most projections closely follow this trend, some values in the LR method show slight deviations from the linear path.

Table 5 LR Coefficient for Each Indicator

Indicator	Coefficient
X1	14,147,728.0
X2	11,219,869.6
X3	-41,467,157.4
X4	-11,762,760.3
X5	-229,002,938.8
X6	70,939,539.3
X7	94,903,241.2
X8	-36,362,221.4
X9	-37,222,134.9
X10	-59,075,776.9
X11	57,902,683.8
X12	23,102,658.3
X13	-48,670,411.3
X14	109,216,964.5
X15	-10,006,753.6
X16	-14,681,789.4
X17	67,574,005.4
X18	81,145,303.7
X19	-60,648,430.4
Constant	-622,930,714.8

Table 6 The Prediction Performance of LR Method

Performance Indicators	Result
R	0.980
MAE (million Rupiahs)	180.401
RMSE (million Rupiahs)	230.253
MAPE (%)	8.556

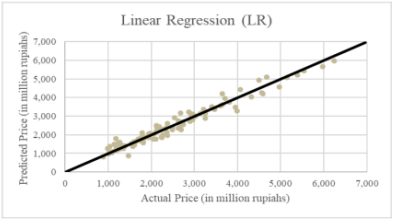

Figure 3 Relationship Between the Predicted and the Actual Housing Prices Using the LR Method

✓ **Classification and Regression Tree (CART)**

In this method, the tree depth is adjusted to find the best prediction result based on the given indicators. Among the tested tree depths, a tree depth of 10 delivers the best performance, as it has the highest RI value. The CART model with a tree depth of 10 achieves a correlation of 0.936, the MAE of Rp 395,034,000, the RMSE of Rp 370,577,000, and the MAPE of 11.671%, as seen in Table 7. Figure 4 illustrates the relationship between the predicted and the actual housing prices using the CART method. Several projected data points show significant deviations from the linear reference line, indicating variations in the model's prediction accuracy.

Table 7 The Prediction Performance of CART Method

Tree Depth	Performance Indicator				
	R	MAE (million Rupiahs)	RMSE (million Rupiahs)	MAPE (%)	RI
3	0.877	434.504	536.128	18.402	0.000
4	0.918	428.194	412.032	13.836	0.529
5	0.930	413.104	387.121	13.042	0.746
6	0.932	411.087	388.833	12.763	0.773
7	0.935	396.707	371.450	12.127	0.961
8	0.936	395.370	370.418	11.779	0.992
9	0.936	395.034	372.729	11.717	0.993
10	0.936	395.034	370.577	11.671	1.000

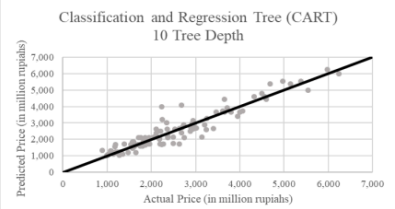

Figure 4 Relationship Between the Predicted and the Actual Housing Prices Using the CART Method

Chi-Square Automatic Interaction Detection (CHAID)

In this method, the tree depth is adjusted to identify the most accurate prediction result based on the given indicato[9] Among the tested tree depths, a tree depth of 7 delivers the best performance, as it has the highest RI value. The CHAID model with a depth of 7 obtains a correlation of 0.923, the MAE of Rp 306,075,000, the RMSE of Rp 425,346,000, and the MAPE of 11.739%, as seen in Table 8. Figure 5 illustrates the relationship between the predicted and the actual housing prices using the CHAID method. Several projected points show significant deviations from the linear reference line, indicating variations in the model's prediction accuracy.

Table 8 The Prediction Performance of CHAID Method

Tree Depth	Performance Indicator				
	R	MAE (million Rupiahs)	RMSE (million Rupiahs)	MAPE (%)	RI
3	0.919	332.784	470.799	12.009	0.021
4	0.920	315.000	436.117	12.027	0.476
5	0.922	308.924	426.615	11.778	0.910
6	0.923	306.075	425.484	11.739	0.999
7	0.923	306.075	425.346	11.739	1.000

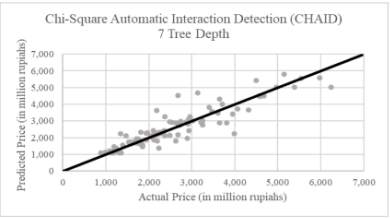

Figure 5 Relationship Between the Predicted and the Actual Housing Prices Using the CHAID Method

Artificial Neural Network (ANN)

In the ANN method, the number of neurons is adjusted to get the best prediction accuracy based on the given indicators. Among the tested number of neurons, the ANN model with 12 neurons delivers the best performance, as it has the highest RI value. The ANN model with 12 neurons achieves a correlation of 0.984, the MAE of Rp 160,681,000, the RMSE of Rp 216,822,000, and the MAPE of 7.401%, as seen in Table 9.

Figure 6 illustrates the relationship between the predicted and the actual housing prices using the ANN method. Several projected points show some deviations from the linear reference line, indicating the variations in prediction accuracy of the model.

Table 9 The Prediction Performance of ANN Method

Number of Neurons	Performance Indicator				
	R	MAE (million Rupiahs)	RMSE (million Rupiahs)	MAPE (%)	RI
10	0.980	174.751	235.434	7.584	0.585
12	0.984	160.681	216.822	7.401	1.000
20	0.968	185.319	281.989	7.870	0.000

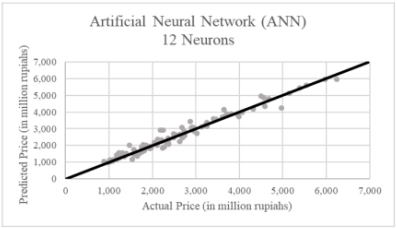

Figure 6 Relationship Between the Predicted and the Actual Housing Prices Using the ANN Method

C. COMPARATIVE ANALYSIS OF THE ARTIFICIAL INTELLIGENCE (AI) METHODS

A summary of the prediction performance of each model is provided in Table 10. These results are compared to identify which artificial intelligence (AI) method delivers the highest accuracy in the [10]using price prediction. Among all four tested methods, the ANN method gives the best performance in predicting the housing price, as indicated by its highest RI value of 1.00 and its training time of 3 seconds only. The ANN model with 12 neurons achieves a correlation R of 0.984, the MAE of Rp 160,681,000, the RMSE of Rp 216,822,000, and the MAPE of 7.401%. The flexibi[6]y of the ANN model allows it to build better prediction model. On the other hand, CART has the lowest RI value. The CART model with a tree depth of 10 achieves a correlation R of 0.936, the MAE of Rp 395,034,000, the RMSE of Rp 370,577,000, and the MAPE of 11.671%. These results indicating that the CART method is the least effective method for predicting the housing prices. In every prediction model, two indicators which are Land Area (X1) and Building Area (X2) are the top indicators that much more impactful or important than the others in predicting the housing price. Overall, it can be determined that the ANN method yields the best prediction accuracy compared to the other three methods used in this case study.

Table 10 The Comparison of Prediction Performance of Four Artificial Intelligence (AI) Method

AI Method	Performance Indicator				RI	Rank	Training Time (seconds)
	R	MAE (million Rupiahs)	RMSE (million Rupiahs)	MAPE (%)			
LR	0.980	180.401	230.253	8.556	0.862	2	17
CART	0.936	395.034	370.577	11.671	0.093	4	10
CHAID	0.923	306.075	425.346	11.739	0.127	3	6
ANN	0.984	160.681	216.822	7.401	1.000	1	3

CONCLUSIONS

This study conducts a comparative analysis of several methods in the housing price prediction, consists of one traditional approach, which is Linear Regression (LR), and three artificial intelligence (AI) methods, which are Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Chi-Squared Automatic Interaction Detection (CHAID). The housing dataset is used to create the prediction models, which will be further tested to determine the accuracy in predicting the housing prices. The model accuracy is evaluated using the four performance indicators—R, MAE, RMSE, and MAPE—together with the Preference Index (RI) to find the best prediction method. Based on the case study results, it can be concluded that the AI-based methods, the ANN model, outperform the other methods in predicting the housing prices.

REFERENCES

[1] M. Siahaan, "Real estate in Indonesia - statistics & facts," statista. [Online]. Available: https://www.statista.com/topics/8596/real-estate-in-indonesia/#topicOverview

[2] E. Z. Teoh, W.-C. Yau, T. S. Ong, and T. Connie, "Explainable housing price prediction with determinant analysis," *IJHMA*, vol. 16, no. 5, pp. 1021–1045, Aug. 2023, doi: 10.1108/IJHMA-02-2022-0025.

[3] Appraisal Institute, Ed., *The appraisal of real estate*, Fifteenth edition. Chicago, IL: Appraisal Institute, 2020.

[4] F. Fahirah, A. Basong, and T. Hermansah, "Identifikasi Faktor yang Mempengaruhi Nilai Jual Lahan dan Bangunan pada Perumahan Tipe Sederhana.pdf," *Jurnal SMARTek*, vol. 8, no. 4, pp. 251–269.

[5] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartoto, and I. B. Syamwil, "Factors influencing the price of housing in Indonesia," *International Journal of Housing Markets and Analysis*, vol. 8, no. 2, pp. 169–188, Jun. 2015, doi: 10.1108/IJHMA-04-2014-0008.

[6] A. Olanrewaju, X. Y. Lim, S. Y. Tan, J. E. Lee, and H. Adnan, "Factors Affecting Housing Prices in Malaysia: Analysis of The Supply Side," *Planning Malaysia Journal*, vol. 16, no. 6, pp. 225-235, Sept. 2018, doi: 10.21837/pm.v16i6.477

[7] F. Zulkifli and H. Ismail, "Factors Influencing House Buyer's Decision in Malaysia. Case Study: Sepang, Selangor," *PM*, vol. 21, Jul. 2023, doi: 10.21837/pm.v21i27.1293.

[8] M.-Y. Cheng, P. M. Firdausi, and D. Prayogo, "High-performance concrete compressive strength prediction using Genetic Weighted Pyramid Operation Tree (GWPOT)," *Engineering Applications of Artificial Intelligence*, vol. 29, pp. 104–113, Mar. 2014, doi: 10.1016/j.engappai.2013.11.014.

[9] Hore, Sirshendu *et al.*, "Neural-based prediction of structural failure of multistoried RC buildings," *Structural Engineering and Mechanics*, vol. 58, no. 3, pp. 459–473, May 2016, doi: 10.12989/SEM.2016.58.3.459.

[10] M.-Y. Cheng, N.-D. Hoang, and Y.-W. Wu, "Cash Flow Prediction for Construction Project using A Novel Adaptive Time-Dependant Least Square Support Vector Machine Inference Model," *Journal of Civil Engineering and Management*, vol. 21, no. 6, pp. 679–688, Jun. 2015, doi: 10.3846/13923730.2014.893906.

[11] A. Saiful, "Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression," *JATISI*, vol. 8, no. 1, pp. 41–50, Mar. 2021, doi: 10.35957/jatisi.v8i1.701.

[12] A. R. Hutami, "Aplikasi Neural Network untuk Prediksi Harga Rumah di Yogyakarta Menggunakan Backpropagation," Universitas Islam Indonesia, Yogyakarta, 2018.

[13] P. Choirunisa, "Implementasi Artificial Inteligence Untuk Memprediksi Harga Penjualan Rumah Menggunakan Metode Random Forest dan Flask," Universitas Islam Indonesia, Yogyakarta, 2020.

[14] K. M. A. Hossain and M. Lachemi, "Strength, durability and micro-structural aspects of high performance volcanic ash concrete," *Cement and Concrete Research*, vol. 37, no. 5, pp. 759–766, May 2007, doi: 10.1016/j.cemconres.2007.02.014.

[15] S. A. A. Karim and N. F. Kamsani, *Water Quality Index Prediction Using Multiple Linear Fuzzy Regression Model: Case Study in Perak River, Malaysia*. in SpringerBriefs in Water Science and Technology. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-3485-0.

[16] S. A. Eslamian, S. S. Li, and F. Haghighat, "A new multiple regression model for predictions of urban water use," *Sustainable Cities and Society*, vol. 27, pp. 419–429, Nov. 2016, doi: 10.1016/j.scs.2016.08.003.

[17] P. Pandit, P. Dey, and K. N. Krishnamurthy, "Comparative Assessment of Multiple Linear Regression and Fuzzy Linear Regression Models," *SN*

COMPUT. SCI., vol. 2, no. 2, p. 76, Apr. 2021, doi: 10.1007/s42979-021-00473-3.

[18] A. Yağmur, M. Kayakuş, and M. Terzioğlu, "House price prediction modeling using machine learning techniques: a comparative study," *Aestimum*, vol. 81, Feb. 2023, doi: 10.36253/aestim-13703.

[19] Y. Wu and J. Feng, "Development and Application of Artificial Neural Network," *Wireless Pers Commun*, vol. 102, no. 2, pp. 1645–1656, Sep. 2018, doi: 10.1007/s11277-017-5224-x.

[20] H. Yoon, S.-C. Jun, Y. Hyun, G.-O. Bae, and K.-K. Lee, "A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer," *Journal of Hydrology*, vol. 396, no. 1–2, pp. 128–138, Jan. 2011, doi: 10.1016/j.jhydrol.2010.11.002.

[21] K. A. Grajski, L. Breiman, G. V. Di Prisco, and W. J. Freeman, "Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART," *IEEE Trans. Biomed. Eng.*, vol. BME-33, no. 12, pp. 1076–1086, Dec. 1986, doi: 10.1109/TBME.1986.325684.

[22] J.-S. Chou, C.-F. Tsai, A.-D. Pham, and Y.-H. Lu, "Machine learning in concrete strength simulations: Multi-nation data analytics," *Construction and Building Materials*, vol. 73, pp. 771–780, Dec. 2014, doi: 10.1016/j.conbuildmat.2014.09.054.

[23] M. Milanović and M. Stamenković, "CHAID Decision Tree: Methodological Frame and Application," *Economic Themes*, vol. 54, no. 4, pp. 563–586, Dec. 2016, doi: 10.1515/ethemes-2016-0029.

[24] M. Gunduz and I. Al-Ajji, "Employment of CHAID and CRT decision tree algorithms to develop bid/no-bid decision-making models for contractors," *ECAM*, vol. 29, no. 9, pp. 3712–3736, Nov. 2022, doi: 10.1108/ECAM-01-2021-0042.

[25] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Systems with Applications*, vol. 67, pp. 239–251, Jan. 2017, doi: 10.1016/j.eswa.2016.08.065.

Comparative Study on Artificial Intelligence Methods in Housing Price Prediction

ORIGINALITY REPORT

19%	14%	14%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	2%
2	nottingham-repository.worktribe.com Internet Source	1%
3	Submitted to University of Florida Student Paper	1%
4	Jui-Sheng Chou, Dac-Khuong Bui. "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design", Energy and Buildings, 2014 Publication	1%
5	www.scirp.org Internet Source	1%
6	Manoj Khandelwal, Danial Jahed Armaghani, Ramesh Murlidhar Bhatawdekar, Pijush Samui, Saffet Yagiz. "Advancements in Underground Infrastructures", CRC Press, 2025 Publication	1%
7	Submitted to Macquarie University Student Paper	1%
8	Moloud Abdar, Mariam Zomorodi-Moghadam, Resul Das, I-Hsien Ting. "Performance analysis of classification algorithms on early detection of liver disease", Expert Systems with Applications, 2017 Publication	1%

9	jurnal.itscience.org Internet Source	1 %
10	www.researchgate.net Internet Source	1 %
11	Murat Gunduz, Ibrahim Al-Ajji. "Employment of CHAID and CRT decision tree algorithms to develop bid/no-bid decision-making models for contractors", Engineering, Construction and Architectural Management, 2021 Publication	1 %
12	Qianru Niu, Shuangyin Ren, Wei Gao, Chunjiang Wang. "A Dynamic Threat Assessment Method for Multi-Target Unmanned Aerial Vehicles at Multiple Time Points Based on Fuzzy Multi-Attribute Decision Making and Fuse Intention", Mathematics, 2025 Publication	1 %
13	mail.planningmalaysia.org Internet Source	1 %
14	Hafiz, Rezwana Binte. "Characterization and Modeling of Properties of Unsaturated and Partially Damaged Cement-Based Materials.", Missouri University of Science and Technology Publication	<1 %
15	www.frontiersin.org Internet Source	<1 %
16	eprints.cihanuniversity.edu.iq Internet Source	<1 %
17	repozitorij.unios.hr Internet Source	<1 %
18	Raymond Elikplim Kofinti, Damiano Kulundu Manda, Martine Odhiambo Oleche, Germano Mwabu. "Consumption Inequality and	<1 %

	Multidimensional Child Poverty in Ghana: Does Access to Communal Services Matter?", Child Indicators Research, 2024 Publication	
19	journal.mediadigitalpublikasi.com Internet Source	<1 %
20	Ali Soltani, Chyi Lin Lee. "The non-linear dynamics of South Australian regional housing markets: A machine learning approach", Applied Geography, 2024 Publication	<1 %
21	Submitted to Liverpool John Moores University Student Paper	<1 %
22	www.tandfonline.com Internet Source	<1 %
23	Kanchana Vishwanadee Mathotaarachchi, Raza Hasan, Salman Mahmood. "Advanced Machine Learning Techniques for Predictive Modeling of Property Prices", Information, 2024 Publication	<1 %
24	www.sbp-journal.com Internet Source	<1 %
25	Jui-Sheng Chou, Chih-Fong Tsai, Anh-Duc Pham, Yu-Hsin Lu. "Machine learning in concrete strength simulations: Multi-nation data analytics", Construction and Building Materials, 2014 Publication	<1 %
26	www.atlantis-press.com Internet Source	<1 %
27	Chiloane, Nhleko Monique. "Effects of Layering on the Mechanical Properties of Cemented Tailings Backfill Under Unconfined	<1 %

	Compression", University of South Africa (South Africa) Publication	
28	Hadi Salehi, Rigoberto Burgueño. "Emerging artificial intelligence methods in structural engineering", Engineering Structures, 2018 Publication	<1 %
29	Min-Yuan Cheng, Pratama Mahardika Firdausi, Doddy Prayogo. "High-performance concrete compressive strength prediction using Genetic Weighted Pyramid Operation Tree (GWPOT)", Engineering Applications of Artificial Intelligence, 2014 Publication	<1 %
30	ejurnal.politeknikpratama.ac.id Internet Source	<1 %
31	ia601004.us.archive.org Internet Source	<1 %
32	www.journal.uestc.edu.cn Internet Source	<1 %
33	Meng Zhang, Koki Hibi, Junya Inoue. "GPU-accelerated artificial neural network potential for molecular dynamics simulation", Computer Physics Communications, 2023 Publication	<1 %
34	eprints.nottingham.ac.uk Internet Source	<1 %
35	eprints.usq.edu.au Internet Source	<1 %
36	irjet.net Internet Source	<1 %
37	link.springer.com Internet Source	<1 %

38	van Lille, Adele. "The Identification and Composition of Social Media Post Characteristics for Optimal Stakeholder Engagement of Participation Sports Events", University of South Africa (South Africa) Publication	<1 %
39	www.bis.org Internet Source	<1 %
40	www.researcherslinks.com Internet Source	<1 %
41	www.sciencegate.app Internet Source	<1 %
42	www.techscience.com Internet Source	<1 %
43	Ali Behnood, Venous Behnood, Mahsa Modiri Gharehveran, Kursat Esat Alyamac. "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm", Construction and Building Materials, 2017 Publication	<1 %
44	Johanna Ärje, Salme Kärkkäinen, Kristian Meissner, Alexandros Iosifidis, Türker Ince, Moncef Gabbouj, Serkan Kiranyaz. "The effect of automated taxa identification errors on biological indices", Expert Systems with Applications, 2017 Publication	<1 %
45	docplayer.net Internet Source	<1 %
46	isda2001.softcomputing.net Internet Source	<1 %
47	lutpub.lut.fi Internet Source	<1 %

48	www.citcglobal.com Internet Source	<1 %
49	www.hindawi.com Internet Source	<1 %
50	Maedeh Mahmoudi, Amin Mahdavi-Meymand, Ammar AlDallal, Mohammad Zounemat-Kermani. "Improving groundwater quality predictions in semi-arid regions using ensemble learning models", Environmental Science and Pollution Research, 2025 Publication	<1 %
51	Saravanan Krishnan, Ramesh Kesavan, B. Surendiran, G. S. Mahalakshmi. "Handbook of Artificial Intelligence in Biomedical Engineering", Apple Academic Press, 2021 Publication	<1 %
52	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022 Publication	<1 %
53	Aleksandar Đukić, Milorad K. Banjanin, Mirko Stojčić, Tihomir Đurić, Radenka Đekić, Dejan Anđelković. "An Ensemble of Machine Learning Models for the Classification and Selection of Categorical Variables in Traffic Inspection Work of Importance for the Sustainable Execution of Events", Sustainability, 2024 Publication	<1 %
54	Amani Gomaa Shaaban, Mohamed Helmy Khafagy, Mohamed Abbas Elmasry, Heba El-Beih, Mohamed Hasan Ibrahim. "Knowledge discovery in manufacturing datasets using	<1 %

data mining techniques to improve business performance", Indonesian Journal of Electrical Engineering and Computer Science, 2022
Publication

55 S. Sahoo, T. A. Russo, J. Elliott, I. Foster. "Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S.", Water Resources Research, 2017 <1%
Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		