JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 8, No. 2, December 2025, pp. 623~628
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v8i2.4829

House Price Prediction in Surabaya Using Backpropagation Neural Network

Dimas Fajri Pamungkas[1], Abdul Rezha Efrat Najaf[2], Reisa Permatasari[3]
[1,2,3]Department of Information Systems, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia

Article Info

Article history:

Received 08 27, 2025
Revised 11 04, 2025
Accepted 11 28, 2025

Keywords:

House Price Prediction
Artificial Neural Network
Backpropagation
Flask
Surabaya

ABSTRACT

This research develops a house price prediction system in Surabaya using the Backpropagation Neural Network (BPNN) method. The dataset was obtained through web scraping of property listings, resulting in 3,435 records with 52 attributes. To improve stability, the target variable (house price) was transformed using natural logarithms. Several neural network architectures were tested, and the best configuration [32, 64, 32] achieved Mean Absolute Error (MAE) of 0.3125, Root Mean Squared Error (RMSE) of 0.4201, R^2 of 0.7138, and Mean Absolute Percentage Error (MAPE) of 1.46%. A multi-run evaluation of 20 iterations confirmed consistency of results. The model was implemented as a web-based application using Flask, allowing users to predict house prices in real-time. This research shows that BPNN is reliable for property price forecasting and can support decision-making in the housing market. This research not only advances the academic understanding of neural network-based property valuation but also delivers practical insights that can guide policymakers, investors, and urban planners in making data-driven housing market decisions.

Corresponding Author:

Dimas Fajri Pamungkas
Department of Information Systems, Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
Email: dimasfp123@gmail.com

1. Introduction

The property sector represents one of the most dynamic industries globally, with significant economic and social implications. In Indonesia, this sector continues to expand steadily, driven by population growth, urbanization, and the rising demand for both residential and commercial properties [1]. Housing, in particular, serves not only as a basic human necessity but also as an important form of investment, with property values often appreciating over time. Multiple determinants such as land area, building size, location, accessibility, infrastructure, and neighborhood quality strongly influence the formation of property prices [2], [3].

Surabaya, the second-largest metropolitan city in Indonesia, presents a particularly dynamic housing market. Rapid population growth, coupled with urban expansion and infrastructure development, has resulted in fluctuating property prices that vary widely across different districts [2]. Bank Indonesia's Survei Harga Properti Residensial (SHPR) reports regular changes in property indices, with Surabaya consistently ranking as one of the most active property markets nationwide [2]. Accurate prediction of housing prices is therefore vital for prospective buyers, sellers, investors, real estate agents, and policymakers.

Historically, property valuation has relied heavily on statistical models, most notably Multiple Linear Regression (MLR) [4]. While MLR and similar approaches provide interpretability, they often fail to capture the complex, nonlinear interactions between features in housing data [10], [12]. In recent years, Machine Learning (ML) techniques have emerged as powerful alternatives capable of modeling nonlinear relationships and uncovering hidden patterns within large and noisy datasets [5], [6], [8].

Among ML techniques, Artificial Neural Networks (ANNs) have shown particularly strong potential in predictive analytics. The ability of ANNs to approximate nonlinear functions is supported by the universal approximation theorem [18]. The Backpropagation Neural Network (BPNN), first popularized by Rumelhart et al. [4], has become one of the most widely applied variants across domains ranging from healthcare diagnostics [7] to financial forecasting [15]. BPNN is highly suitable for complex regression tasks such as housing price prediction, where interdependent factors collectively determine outcomes.

Several studies have successfully demonstrated the applicability of ANNs in housing markets. Zhang et al. [6] revealed that ANN-based models consistently outperform regression-based models in predicting house prices. Kok et al. [8] and Taşpınar and Yildiz [9] conducted comparative studies across multiple algorithms, concluding that ANNs generally achieve higher accuracy, especially with large and diverse datasets. Furthermore, hybrid approaches combining ANNs with ensemble methods such as Random Forests and Gradient Boosting have been reported to yield promising results in complex prediction problems [15], [17].

More recent research (2023–2025) has further advanced this field by integrating deep learning and explainable AI techniques into real estate prediction. For instance, convolutional neural networks (CNNs) have been applied to capture geographic variations in housing markets [26], while graph neural networks (GNNs) have shown effectiveness in modeling spatial dependencies among properties [27]. Other works emphasize interpretability by combining XGBoost with SHAP values for feature-level insights [28], and adaptive loss functions with feature embedding optimization have demonstrated state-of-the-art accuracy in dynamic residential markets [29]. These developments highlight that AI-based approaches continue to evolve rapidly, reinforcing the importance and timeliness of applying BPNN in the context of Surabaya.

Nevertheless, in the Indonesian context, particularly in Surabaya, there is still a limited number of empirical studies applying ANN-based models for property valuation. Most research to date has focused on regression or basic ML algorithms, without extensive experimentation on deep or multi-layer neural networks. This study therefore contributes by filling that research gap.

The primary objective of this research is to develop and evaluate a Backpropagation Neural Network model for predicting house prices in Surabaya. A dataset of 3,435 housing records was collected from an online property listing platform through web scraping. After comprehensive preprocessing including normalization, encoding, and outlier treatment several ANN architectures were tested to identify the best-performing configuration. Finally, the optimized model was deployed into a Flask-based web application for real-time prediction.

The significance of this study lies in two key aspects. From an academic perspective, it provides empirical evidence regarding the applicability of BPNN in property valuation in Indonesia. From a practical perspective, it offers a functional tool that can support informed decision-making for buyers, sellers, and investors.

2. Research Method

This research followed a structured methodology consisting of six main stages: data collection, data preprocessing, data splitting, neural network design, model training and evaluation, and system implementation.

2.1. Data Collection

The dataset was obtained using web scraping from Pinhome, one of the largest property listing platforms in Indonesia. Web scraping is an established approach for extracting structured data at scale from online sources [13]. A total of 3,435 housing records were collected, each comprising 52 attributes. These attributes included land area, building area, number of bedrooms, number of bathrooms, number of floors, electricity capacity, land ownership type, and property price.

The inclusion of diverse attributes is crucial because property price formation is inherently multifactorial, influenced not only by physical characteristics but also by legal, infrastructural, and locational aspects [1], [2], [3].

2.2. Data Preprocessing

Raw housing data is often noisy and incomplete; therefore, preprocessing was carried out through several steps to improve data quality. First, missing values were addressed by applying mean or mode imputation depending on the attribute type, while records with excessive incompleteness were removed to maintain reliability [10], [11]. Second, categorical attributes such as land ownership were transformed into numerical form using one-hot encoding to enable compatibility with the neural network input layer [7]. Third, continuous variables including land area and building area were normalized into a 0–1 range, which helps reduce scale dominance and stabilize the learning process [8] In addition to normalization, dimensionality reduction techniques such as Principal Component Analysis (PCA) can be considered for future studies to reduce feature redundancy and improve computational efficiency [16]. While PCA was not applied in this research, its integration could further optimize the learning process by capturing the most informative features in housing datasets. Finally, outliers were detected using the Interquartile Range (IQR) method and quantile thresholds, and extreme values were trimmed to minimize potential distortion during training [14]. Through these steps, the dataset was refined to ensure it was clean, structured, and suitable for training the Artificial Neural Network.

2.3. Data Splitting

The dataset was divided into training and testing subsets using an 80:20 ratio. This produced 2,748 training samples and 687 testing samples. Such a ratio follows best practices in machine learning, striking a balance between model training capacity and evaluation robustness [12].

2.4. Neural Network Design

The ANN model was designed as a feedforward Backpropagation Neural Network. The input layer size corresponded to the number of predictor attributes, while the output layer contained a single node representing the predicted log-transformed house price.

Several architectures were tested, including:

- [32] (single hidden layer)
- [64, 32] (two hidden layers)
- [32, 64, 32] (three hidden layers)

The activation function applied in the hidden layers was the sigmoid function, defined as:

$$\sigma(x) = \frac{1}{1 + e^{\{-x\}}} \tag{1}$$

The network was trained using the backpropagation algorithm, which iteratively adjusts weights by propagating error gradients backward through the network [4], [5]. The loss function used was Mean Squared Error (MSE). Training was implemented using TensorFlow/Keras libraries [19], [21].

Backpropagation Algorithm (Simplified Mathematical Overview)

1. Forward Propagation: Input features are propagated through the network to generate predicted outputs.
2. Error Calculation: The difference between predicted and actual outputs is computed using MSE.
3. Backward Propagation: Partial derivatives of the error with respect to weights are computed using the chain rule.
4. Weight Update: Weights are updated using gradient descent, defined as:

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta \frac{\partial E}{\partial w_{ij}} \tag{2}$$

where η is the learning rate and E is the error function [5], [18].

2.5. Evaluation Metrics

The performance of the model was evaluated using four standard regression metrics that collectively provide a comprehensive assessment of prediction accuracy. The Mean Absolute Error (MAE) measures the average magnitude of absolute differences between predicted and actual values, offering an intuitive sense of error in the same units as the target variable. The Root Mean Squared Error (RMSE) emphasizes larger errors by squaring the differences before averaging, making it particularly sensitive to outliers. The Coefficient of Determination (R^2) quantifies how well the model explains the variance in the target variable, with values closer to one indicating a stronger fit. Finally, the Mean Absolute Percentage Error (MAPE) expresses prediction errors as percentages, allowing for easy interpretation of relative accuracy across different scales. Together, these metrics ensure a balanced evaluation of both absolute and relative prediction accuracy as well as the overall goodness of fit of the model [12].

2.6. System Implementation

The trained ANN was deployed in a web-based application using the Flask micro-framework [20]. The application interface allows users to input housing attributes and instantly receive predicted prices. This implementation demonstrates how machine learning models can be operationalized for practical decision-making in the property sector.

3. Result and Discussion

3.1. Data Characteristics

Exploratory Data Analysis (EDA) revealed substantial variation in property prices. The raw distribution was right-skewed, with a small number of high-priced properties disproportionately raising the mean. A logarithmic transformation of prices was therefore applied to stabilize variance and approximate normality, consistent with prior studies on real estate modeling [6], [8].

Correlation analysis indicated that land area, building area, and the number of bedrooms exhibited strong positive relationships with property price, confirming their role as primary predictors [1], [2].

3.2. Model Performance

Among the tested ANN architectures, the [32, 64, 32] configuration achieved the best performance with the following metrics:

- MAE = 0.3125
- RMSE = 0.4201
- $R^2 = 0.7138$
- MAPE = 1.46%

These results indicate strong predictive accuracy, surpassing many regression-based benchmarks. The model's stability was confirmed through 20 independent runs, which yielded consistent results.

3.3. Comparative Analysis with Prior Studies

The achieved R^2 value of 0.71 is comparable with findings from Zhang et al. [6] and Taşpınar and Yildiz [9], who reported R^2 values between 0.65 and 0.72. Kok et al. [8] also demonstrated similar performance with ANN-based models in housing prediction.

While ensemble approaches such as Random Forests [15] and Gradient Boosting [17] have been shown to yield competitive results, the current study highlights that carefully designed ANNs can achieve similar or better accuracy in contexts such as Surabaya.

3.3.1 Comparative Summary of ANN-Based House Price Prediction Results with Prior Studies

To further contextualize the results, Table X Model Performance Comparison Across Neural Network Architectures (MAE, RMSE, R^2, MAPE) presents a narrative comparison of this study's findings with selected prior research. Zhang et al. [6] reported that ANN models could reach R^2 values of approximately 0.68 when predicting property prices in China using a dataset of more than 10,000 entries. Taşpınar and Yildiz [9] demonstrated that ANNs achieved higher accuracy than Decision Trees and Support Vector Regression, reporting R^2 values between 0.65 and 0.72 on Turkish housing data. Similarly, Kok et al. [8] highlighted the effectiveness of ANNs when applied to housing data from multiple European cities.

Compared with these works, the present study achieved an R^2 of 0.71 with only 3,435 records, indicating that the ANN was able to generalize effectively even with a smaller dataset. The low MAPE of 1.46% demonstrates strong robustness, which is crucial in financial applications.

This comparison underscores that although ensemble methods such as Random Forests [15] and boosting algorithms [17] have demonstrated competitive performance in various contexts, a carefully designed ANN remains capable of delivering results that are on par with, or superior to, those of ensemble models. In the context of Surabaya, where real estate data may be fragmented or limited, this robustness is particularly valuable.

3.4 System Implementation

The Flask-based web system was successfully developed and tested. The system's interface is user-friendly and accessible even to non-technical users, making it suitable for adoption by real estate stakeholders. Similar studies integrating ML models into web applications have reported that practical usability significantly enhances stakeholder engagement [20], [25].

3.5. Discussion

This research reinforces the growing consensus that ANNs outperform traditional regression methods in housing price prediction tasks [4], [6], [9]. However, limitations remain. The dataset was limited to a single platform, which may introduce sampling bias. Additionally, external macroeconomic factors such as interest rates, inflation, and regional development policies were not included, despite their potential impact on property prices [22], [23]. For instance, rising interest rates generally reduce mortgage affordability and dampen housing demand, while inflation and regional development policies can directly influence construction costs and investment flows. The absence of these variables may limit the model's ability to fully capture market volatility and long-term pricing trends.

Future studies could address these limitations by:

- Incorporating datasets from multiple listing platforms.
- Integrating macroeconomic and spatial features.
- Comparing ANN performance with ensemble and deep learning methods such as Random Forests, XGBoost, and Convolutional Neural Networks [15], [17], [24].

4. Conclusion

This research developed and evaluated a Backpropagation Neural Network for predicting house prices in Surabaya. Using 3,435 housing records obtained via web scraping, the model was trained and tested across multiple architectures. The best configuration, [32, 64, 32], achieved an R^2 of 0.71 with low error rates across MAE, RMSE, and MAPE.

From an academic perspective, this study contributes to the literature by providing empirical evidence on the applicability of ANN-based models in the Indonesian housing market, which has been underexplored. From a practical perspective, it delivers a working web-based system that allows stakeholders to perform real-time property valuation.

Limitations of this research include reliance on a single data source and the exclusion of macroeconomic variables. Future research should expand the dataset, incorporate additional external factors, and experiment with hybrid or ensemble learning techniques to further improve prediction accuracy.

In summary, the study confirms that Backpropagation Neural Networks can effectively model complex, nonlinear relationships in real estate data, providing a reliable tool for supporting decision-making in the property sector of Surabaya.

Beyond its academic and technical contributions, this study also holds practical implications. For policymakers, the ability to predict housing prices accurately can support urban planning and housing affordability programs. For real estate developers, such predictive systems may assist in determining optimal pricing strategies and investment decisions. Finally, for individual buyers and sellers, the system provides an accessible tool to evaluate whether a property is fairly priced relative to market conditions. These implications highlight that beyond theoretical accuracy, predictive models can play a crucial role in supporting sustainable housing markets in rapidly developing cities such as Surabaya.

Acknowledgement

The authors would like to express their gratitude to the Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jawa Timur, for providing academic support and research facilities. We also extend our appreciation to colleagues who contributed valuable feedback during the development of this work. Finally, sincere thanks are addressed to our supervisors and peers for their guidance and encouragement throughout the completion of this research.

References

[1] D. Geltner, N. G. Miller, J. Clayton, and P. Eichholtz, Commercial Real Estate Analysis and Investments, 3rd ed. Mason, OH: Cengage Learning, 2013.

[2] Bank Indonesia, "Survei Harga Properti Residensial (SHPR)," Jakarta, 2024.

[3] Badan Pusat Statistik (BPS), "Statistik Perumahan dan Permukiman Indonesia," 2023.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[5] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ: Pearson, 2009.

[6] X. Zhang, Y. Li, and H. Wang, “Prediction of housing prices based on neural networks,” *Procedia Computer Science*, vol. 199, pp. 883–890, 2022.

[7] A. Khashei and M. Bijari, “An artificial neural network (p,d,q) model for time series forecasting,” *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, 2010.

[8] J. Kok, A. L. Polak, and S. C. Wong, “Machine learning for house price prediction,” *Computers, Environment and Urban Systems*, vol. 91, p. 101710, 2022.

[9] S. B. Taşpınar and Y. Yildiz, “Comparison of machine learning algorithms for house price prediction,” *Sustainability*, vol. 13, no. 11, p. 6232, 2021.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2012.

[11] G. Shmueli, N. R. Patel, and P. C. Bruce, *Data Mining for Business Intelligence*. Hoboken, NJ: Wiley, 2007.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[13] M. Mitchell, *Web Scraping with Python*, 2nd ed. Sebastopol, CA: O’Reilly Media, 2018.

[14] P. C. Austin and J. V. Tu, “Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality,” *J. Clin. Epidemiol.*, vol. 57, no. 11, pp. 1138–1146, 2004.

[15] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Chapman & Hall/CRC, 2012.

[16] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[17] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[19] F. Chollet, *Deep Learning with Python*, 2nd ed. Shelter Island, NY: Manning, 2021.

[20] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed. Sebastopol, CA: O’Reilly Media, 2018.

[21] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA: O’Reilly Media, 2023.

[22] T. H. Davenport and D. Patil, “Data scientist: The sexiest job of the 21st century,” *Harvard Business Review*, vol. 90, no. 10, pp. 70–76, 2012.

[23] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[25] B. Marr, *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Hoboken, NJ: Wiley, 2016.

[26] H. Lee, H. Han, C. Pettit, et al., “Machine learning approach to residential valuation: a convolutional neural network
model for geographic variation,” The Annals of Regional Science, vol. 72, pp. 579–599, 2024, doi: 10.1007/s00168 023-01212-7.

[27] E. Riveros, C. Vairetti, C. Wegmann, S. Truffa, and S. Maldonado, “Scalable Property Valuation Models via Graph based Deep Learning,” arXiv preprint arXiv:2405.06553, 2024.

[28] T. Kee and W. K. O. Ho, “eXplainable Machine Learning for Real Estate: XGBoost and Shapley Values in Price Prediction,” Civil Engineering Journal, vol. 11, no. 5, pp. 2116–2133, 2025, doi: 10.28991/CEJ-2025-011-05-022.

[29] H. Zhang, “Residential real estate price prediction based on adaptive loss function and feature embedding optimization,” Humanities and Social Sciences Communications, vol. 12, p. 832, 2025, doi: 10.1057/s41599-025 05217-9.