Journal of Housing and the Built Environment (2022) 37:2079–2114
https://doi.org/10.1007/s10901-022-09937-1

ARTICLE

Comparison of machine learning models to provide preliminary forecasts of real estate prices

Jui-Sheng Chou[1] · Dillon-Brandon Fleshman[1] · Dinh-Nhat Truong[1,2]

Received: 24 January 2021 / Accepted: 31 January 2022 / Published online: 15 March 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

Real estate is one of the most critical investments in the household portfolio, and represents the greatest proportion of wealth of the private households in highly developed countries. This research provides a succinct review of machine learning techniques for predicting house prices. Data on dwelling transaction prices in Taipei City were collected from the real price registration system of the Ministry of the Interior, Taiwan. Four well-known artificial intelligence techniques—Artificial Neural Networks (ANNs), Support Vector Machine, Classification and Regression Tree, and Linear Regression- were used to develop both baseline and ensemble models. A hybrid model was also built and its predictive performance compared with those of the individual models in both baseline and ensemble schemes. The comprehensive comparison indicated that the particle swarm optimization (PSO)-Bagging-ANNs hybrid model outperforms the other models that are proposed herein as well as others that can be found in the literature. The provision of multiple prediction models allows users to determine the most suitable one, based on their background, needs, and comprehension of machine learning, for predicting house prices.

Keywords House price forecasting · Data mining · Machine learning · Ensemble method · Particle swarm optimization · Hybrid model

✉ Jui-Sheng Chou
jschou@mail.ntust.edu.tw

Dillon-Brandon Fleshman
M10605105@mail.ntust.edu.tw

Dinh-Nhat Truong
D10605806@mail.ntust.edu.tw

[1] Department of Civil and Construction Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[2] Department of Civil Engineering, University of Architecture Ho Chi Minh City, Ho Chi Minh City, Vietnam

1 Introduction

The Taipei metropolitan area is located in Northern Taiwan and is one of the most important economic centers in Taiwan. It covers an area of 2224.1 km^2, accounting for 6.1% of the area of Taiwan (Wang et al., 2018). The area includes a population of approximately 6.7 million, accounting for 30% of Taiwan's total population (Wang et al., 2018). Taipei is one of the world's most densely populated cities, with an average population density of 9942 people per square kilometer (Li et al., 2016). Taipei is also one of the world's most expensive cities. Taipei's house price-to-income ratio has risen sharply from just 6.4 in 2004 to about 15.8 in 2020, according to the country's Ministry of Interior (MOI)—higher than London (8.6), New York (5.9), Toronto (9.9), Sydney (11.8) or Vancouver (13) (Delmendo, 2021).

Real estate is one of the most crucial investments in the household portfolios in Taiwan. (Chiang et al., 2015). Housing not only provides a place to live but is also a critical investment that affects quality of life. The price of a house is of considerable interest to its stakeholders (homeowners, appraisers, developers, investors, and others). However, due to the information asymmetry in house markets, people commonly experience financial losses as a result of their investment. The forecasting of house prices has a significant role in the making of decision by stakeholders to actualize the potential of their investment (Kouwenberg & Zwinkels, 2014).

Traditionally, appraisers estimate the value of housing from explicit market information, ultimately assessing prices based on their subjective evaluations. Accordingly, the appraised price will inevitably be influenced by manual interference. To resolve this issue, a widely used technique, machine learning (ML), has been introduced (Fan et al., 2018; Liu & Liu, 2019). Several studies have proposed effective and efficient machine learning techniques for predicting house prices. However, to the best of our knowledge, a comprehensive comparison among the available machine learning models to provide preliminary forecasts of real estate prices has not been reported earlier in the literature.

To fill the research gap, this work investigates the popular machine learning models for real estate valuation. Four single (baseline) machine learning models, linear regression (LR), classification and regression tree (CART), support vector regression (SVR), and artificial neural network (ANNs), as well as three ensemble models, involving voting, bagging, and stacking, are compared for house valuation. Moreover, hybrid models are developed by introducing a metaheuristic algorithm—particle swarm optimization (PSO) to possibly improve upon the performance of aforementioned machine learners.

Two software packages (WEKA and MATLAB) are utilized to develop above models for the potential real estate estimators. WEKA (Waikato Environment for Knowledge Analysis), available at https://www.cs.waikato.ac.nz/ml/weka/, is an advanced collection of machine learning algorithms and preprocessing techniques that has been designed to implement practical applications; it provides, among other features, various methods for transforming and preprocessing the input data and for making an attribute selection, as well as for classification, clustering and regression tasks and for a statistical evaluation of the resulting learning schemes. WEKA is a very user-friendly tool and it can also be easily interfaced with some of the most commonly used programming languages for machine learning tasks, such as Java, Python and R (Merlini & Rossini, 2021). MATLAB, the language of technical computing platform, is available in academia, industry and research laboratories easily (Job et al., 2021).

Particularly, WEKA is aimed at practitioners who do not have a programming background. MATLAB is used by those with an adequate background in machine learning, and it offers metaheuristic coding that provides more accurate estimation of house prices. The main objective of this research is to find the most efficient and accurate predictive models for the current dwelling dataset by carrying out multiple machine learning techniques.

2 Literature review

2.1 Traditional methods for valuing real estate

The cost approach, the income approach and the sales comparative approach are three traditional methods of valuing real estate. The cost approach determines the value of real estate in the following steps. Firstly, the site is appraised as if it were vacant. Secondly, the cost of the improvements on the site are assessed. Finally, the value of the site under its optimal use is determined with depreciation subtracted. The method is suitable for properties that do not generate incomes, such as schools, and can be compared with only a few buildings in the surrounding area.

The sales comparative approach values properties by comparison with similar properties that were recently involved in open market transactions. As every property is unique, any characteristic may impact its estimated value. The superiority of this method is that it reflects market value using a simple computation (Adetiloye & Eke, 2014; Kontrimas & Verikas, 2011).

The income approach derives market value from the estimated future income that would be generated by the property, discounted to obtain the present value. The gross income multiplier (GIM) and discounted cash-flow analysis (DCF) are commonly used techniques to relate estimate income to market value. The GIM method estimates the value of an investment by dividing the sale price by gross annual income. DCF analysis estimates the value of an investment based on its future cash flow. By utilizing the discount rate, DCF analysis seeks the present value of expected future cash flow, which is used to assess the potential investment. This method, unlike the cost approach, is suitable for commercial offices or leased apartments that are used to generate income.

2.2 Hedonic price theory

Rosen (1974) (Rosen, 1974) defined hedonic prices as “the implicit prices of attributes and are revealed to economic agents from observed prices of differentiated products and the specific quantities of characteristics associated with them”. In other words, an item can be valued by using its characteristics, allowing appraisers to obtain its hedonic price by adding up the implicit values of its homogeneous characteristics. This theory implies a regressive relationship between the item and its attributes, in which each attribute distinctly contribute to the hedonic price of the item. Compressing distinguished characteristics into one dimension to avoid complexity and the intractability of a multi-commodity is one of the advantages of hedonic price theory (Xiao, 2017).

Hedonic price theory has been widely used in real estate evaluation. It allows appraisers to decompose house price into values of individual attributes (such as living area, size of plot, number of rooms, number of bathrooms, location, parking facilities, and location). House price normally fluctuates with the size of the plot, which is one of the attribute that most strongly affects the property value. Despite the wide use of the hedonic house price model and the traditional methods introduced above, some issues such as multi-linearity and explanatory variable interactions limit its accuracy in property valuation (Limsombunchai et al., 2004). Therefore, many researchers have introduced techniques, such as those based on artificial intelligence, to solve these problems.

2.3 Use of machine learning in house price prediction

The development of machine learning accelerated with increases in computing power and its availability (Chaphalkar & Sandbhor, 2013). Machine learning is implemented to construct models that mimic human reasoning; consequently, it can deduce new facts from historical data and respond adaptively to changes of previously obtained information (Chou & Bui, 2014; Chou & Truong, 2021). The practicality of machine learning has been established in various fields and machine learning has been much used in academic fields and forecasting. The use of machine learning for forecasting residential values has been studied in the literature since the 1990s (Chaphalkar & Sandbhor, 2013). Owing to its ability to efficiently solve problems that humans are not able to answer right away, machine learning-based models are feasible alternatives to stakeholders for predicting house prices given the nonlinearity of the relevant data (Barzegar et al., 2016).

In recent years, numerous machine learning-based methods have been used to forecast in property values, as they can find functional relationships in historical data. ANNs, SVR, CART and LR are the most fundamental and commonly used machine learning algorithms for prediction (Chou et al., 2018). In 2004, Limsombunchai et al. (2004) compared the predictive accuracy of the hedonic price model with that of the ANNs model in house price prediction. ANNs has been proven to be superior to the hedonic house price model (Limsombunchai et al., 2004). Wu (2017) utilized various feature extraction algorithms and feature selection methods to build an SVR predicting model for a similar purpose (Wu, 2017). Fan et al. (2006) applied CART to avoid such potential problems as market disequilibrium that affect the hedonic-based regression approach (Fan et al., 2006). Varna et al. (2018) proposed LR and ANNs models for house price prediction, enhanced using a boosting algorithm (Varma et al., 2018). Therefore, this research exploits ANNs, SVR, CART and LR to develop baseline and ensemble models in WEKA.

Various studies have used metaheuristic methods to find the optimal hyperparameters of machine learning techniques and ultimately enhance the predictive accuracy (Claesen & De Moor, 2015). PSO has been proven to yielding more satisfactory results than other optimization algorithms, and it is based on uncomplicated concepts and is easy to implement (Alfiyatin et al., 2017; Wang et al., 2014). Wang et al. (2014) introduced the PSO algorithm to optimize the hyperparameters of the support vector machine (SVM) model. Their proposed PSO-SVM model had better forecast real estate values than did SVM or back propagation neural network (Wang et al., 2014). Alfiyantin et al. (2017) utilized the PSO algorithm to select influential variables and to determine their optimal coefficients in prediction. Their research proved that PSO can be combined with regression analysis to minimize the prediction error (Alfiyatin et al., 2017). This work introduces PSO, which is simple and efficient and provides robust optimization for building hybrids of baseline and ensemble models.

3 Methodology

3.1 Data visualization and preprocessing

The box plot is a common technique for visualizing and interpreting data (Williamson et al., 1989), which are transformed to five statistical values—upper extreme, upper quartile, median, lower quartile, and lower extreme. Numerical data are thus plotted graphically, allowing researchers to comprehend quickly the broad pattern of the data and to see the patterns that are concealed in the dataset, such as the outliers (Potter, 2006). Outliers are observations that are assumed to follow a different distribution from the majority of the dataset. They are also likely to have a profound impact on the data analysis, resulting in erroneous prediction (Schwertman et al., 2004).

The threshold of outliers is based on four boundaries—two inner fences and two outer fences. The data between inner fences and outer fences are designated as mild outliers, while the data that exceed the outer fences are designated as extreme outliers (Dawson, 2011). Since a boxplot not only enables the easy identification of outliers, but also provides decisive information to eliminate them or not so as to prevent distortion of the measures of spread and sensitivity in prediction (Schwertman et al., 2004), this research employs a boxplot technique to identify and exclude the extreme outliers to make more accurate and stable predictions for approximately normal distributed house price datasets.

Feature selection is another statistical technique used to enhance predictive performance and the time and cost-effectiveness of prediction (Guyon & Elisseeff, 2003). The three categories for feature selection are filters, wrappers and embedded techniques. Filters extract features without involving learning techniques, leading to high computational efficiency. Wrappers utilize learning algorithms to evaluate the influential attributes in the feature space. Despite the fact that wrappers tend to outperform filters, as the feature space grows, the computational burden increases, so wrappers are rarely used in practice. Finally, the embedded techniques are the combination of filters and wrappers (Hira & Gillies, 2015; Urbanowicz et al., 2018).

In accordance with our research objectives and considering computational efficiency, this work uses a feature selection method in WEKA, called ReliefF. ReliefF is a weight-based algorithm, called ReliefFAttributeEval in the WEKA platform, and is used in conjunction with Ranker search algorithm to obtain a rank list (Fallahi & Jafari, 2011). It updates the weight by randomly sampling instances from training data, and subsequently evaluates the performance by distinguishing the nearest sample instances of selected attributes from the same and different classes (Huang et al., 2009). ReliefF can handle both nominal and continuous features, identifying their interactions, and managing missing data; it is therefore considered one of the most beneficial and significant filter-based algorithms (Palma-Mendoza et al., 2018).

3.2 Baseline prediction models

3.2.1 Artificial neural networks

Artificial neural networks are inspired by human brains, and were originally designed to duplicate human learning. A basic ANN structure has three main layers—an input data layer, a hidden layer, and an output layer (Bahia, 2013; Peter & Raglend, 2017). The input data layer is the layer to which collected data are input; the data are then transferred through the hidden layer, and finally the training result or prediction is output to the output layer. More specifically, a neural network is built from numerous interconnected neurons with an activation mechanism by which information is transferred between them. Each neuron is associated with a weight and bias adjustment from the input connection. An activation function sums the weighted inputs and biases, and the value thus obtained is transferred to the output neuron (Bahia, 2013; Dey et al., 2019; Khamis & Kamarudin, 2014).

3.2.2 Support vector machine for classification and regression

Support Vector Machine (SVM) is a supervised learning algorithm, which was originally developed by Vapnik (1995) (Vapnik, 1995). It can both be utilized to solve classification and regression problems. When target variables include categorical data, the support vector machine for classification can be used to determine the optimal hyper-plane to distinguish among different classes (Cheng et al., 2019; Chou et al., 2018, 2021a). When the target variables are continuous data, the support vector machine for regression (SVR) is used. SVR is a variation of SVM that maps its inputs onto n-dimensional feature space, and then fits them by a nonlinear kernel function (Chou et al., 2016). The general model for feature space can be given by the following equation.

$$f(x,\omega) = \sum_{j=1}^{n} \omega_i g_i(x) + b \tag{1}$$

where w_i is a weight vector; $g_i(x)$is a set of non-linear transformations; $b$is a bias term.

To measure the quality of estimation, a loss function $L_\varepsilon = \left[y, f(x,\omega)\right]$ is applied, where

$$L_\varepsilon = \left[y, f(x,\omega)\right] = \begin{cases} 0 & if\,|y - f(x,\omega)| \le \varepsilon \\ |y - f(x,\omega)| - \varepsilon & otherwise \end{cases} \tag{2}$$

SVR implements an ε-insensitive loss function in the hyperplanes to concurrently minimize $||\omega||^2$ and reduce the complexity of the model. Kernel functions such as linear, radial basis, polynomial, and sigmoid functions can be used to find support vector on the function surface during training. Owing to its superiority, the radial basis function (RBF) [i.e., $K(x, x_k) = \exp(-\gamma \left\|x - x_k\right\|^2)$] has been proven effective for SVR (Chou et al., 2017) and was used as the kernel function in this study.

3.2.3 Classification and regression tree

Classification and regression tree (CART) is a popular tree-structured analysis technique for use with either categorical variables (classification tree) or numerical variables (regression tree). Particularly, CART reacts sensitively to slight changes in the training dataset (Erdal & Karakurt, 2013). Each internal node in CART has a partitioning rule for testing an attribute. Each branch is represented by the previous result of the test. Eventually, the terminal nodes display a class label or prediction. CART is pruned to optimize the total error by reducing the number of branches through a repeating operation that is determined by the Gini impurity (Chou et al., 2016, 2021b).

The Gini impurity indicates the likelihood that a randomly selected sample is labeled incorrectly, as follows.

$$Gini = 1 - \sum_{i=1}^{J} p_i^2 \tag{3}$$

where i = categorical variable; J = number of categories and p_i = probability that *i-th* category is chosen.

3.2.4 Linear regression

The purpose of linear regression is to find functional relationship between response variables and explanatory variables (Fumo & Rafe Biswas, 2015). When only one explanatory variable is considered in the prediction process, it is called univariate linear regression. When more the one explanatory variables is considered, the process is called multivariate or simple linear regression. The house price dataset involves a numerical response variable, Y, and multiple explanatory variables, X_i, which will be the inputs of the multivariate regression model during the prediction process. The general form of multivariate linear regression is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \varepsilon \tag{4}$$

where Y = response variable; X_1, X_2,….., X_i = explanatory variables; i = number of variables; β_0 = constant variable; $\beta_1, \beta_2, \ldots, \beta_i$ = regression coefficient; and finally, ε = error term.

3.3 Ensemble prediction models

3.3.1 Voting

Voting is the simplest ensemble prediction model, both conceptually and with respect to implementation (Chou & Tran, 2018). In prediction problems, voting averages the outputs of target models for further analysis. In this study, the voting method was managed to generate 11 ensemble models using two, three or four individual techniques based on the aforementioned ANNs, SVR, CART, and LR models. The two-technique ensembles were ANNs + LR, ANNs + SVR, ANNs + CART, SVR + CART, SVR + LR, and CART + LR. The three-technique ensembles were ANNs + SVR + CART, ANNs + SVR + LR, ANNs + CART + LR, and SVR + CART + LR. The four-technique ensemble was ANNs + SVR + CART + LR.

3.3.2 Bagging

Bagging is an acronym for bootstrap aggregation, which is a universal ensemble method and one of the earliest to be developed (Erdal & Karakurt, 2013). Bagging promotes model diversity by applying bootstrap technique to generate randomly several independent subdata from the original training dataset, and trains each model individually using the targeting machine learning technique. Eventually, they are aggregated to build a homogeneous model for prediction by either averaging the outputs of the models or outputting the results with majority votes. This work presents four homogeneous bagging ensembles that use machine learning techniques; they are ANNs ensemble, SVR ensemble, CART ensemble and LR ensemble.

3.3.3 Stacking

Stacking is also called stacked generalization; it is a hierarchically structured ensemble learning method that was proposed by Wolpert (1992). In the first hierarchy, the assemblies of base-learners can be customized to train the samples. Subsequently, the output of all base-learners are "stacked" and input to the meta-learner. The meta-learner is utilized to aggregate the patterns extracted from the base-learners and a prediction is ultimately made (Qian & Rasheed, 2007). Since Wolpert had proven that SVR outperforms other supervised learning techniques in recognizing patterns (Wolpert, 1992), SVR is used as the meta-learner for all 11 heterogeneous combinations of base-learners as described in the voting scheme.

3.4 Hybrid models

Hybrid models have been proved to outperform markedly single and ensemble learners both empirically and theoretically, especially when they are used to deal with classification problems, complicated regression, and high-dimensional data. Hybrid models are widely applied to a diverse range of real world issues, such as financial forecasting, medical diagnosis, and person recognition. They can significantly improve adaptability as well as the quality of reasoning (Kazienko et al., 2013). Various studies have found that hybrid models, which incorporate metaheuristic methods and machine learners, are likely to have a favorable predictive accuracy. However, its effectiveness in house price prediction is still unknown. Therefore, in this study, metaheuristic algorithms are introduced to improve the performance of predictive models.

A metaheuristic algorithm can be regarded as a general-purpose heuristic method, and it assists a particular heuristic to escape from local optima and guide it toward favorable regions in the search domain, which contain robust solutions (Hammouche et al., 2010). Countless metaheuristic algorithms for optimizing hyperparameters of machine learners have been implemented. The most commonly used metaheuristic techniques, to name a few, are ant colony, genetic algorithm and grid search. Each has its weaknesses. The disadvantage of ant colony is it frequently traps in local optimums (Wang & Yao, 2001). The genetic algorithm requires intricate implementation and a distinct crossover or mutation to be designed for each situation (Elbeltagi et al., 2005). Grid search has a very large computation time (Do et al., 2007).

Another popular metaheuristic algorithm is particle swarm optimization, which can be used for comprehensive global optimization; conceptually simple; is easily implemented and converges rapidly (Wang et al., 2014). Hence, rather than using the aforementioned metaheuristic techniques, PSO is utilized to optimize the machine learners herein to develop hybrid models. PSO is an adaptive algorithm that was first proposed by Kennedy and Eberhart in 1995 (Kennedy & Eberhart, 1995), and based on the social-psychological metaphor of bird flocking or fish schooling (Ghasemi et al., 2019).

In the PSO algorithm, a population of particles are placed stochastically in the search space, and each particle represents a feasible solution. Every particle evaluates its objective function at its present location. Subsequently, the velocities of particles are updated in a manner that depends on its best position and the best position of the swarm; new positions of all particles are ultimately obtained from the average of their velocities. The aforementioned procedure iterates until the stopping criterion is met (Armaghani et al., 2017). The velocity and location of the particles are determined and updated using Eqs. (5) and (6).

$$v_i(t+1) = \varphi \times v_i(t) + (c_1 \times rand() \times (p_i^{best} - p_i(t))) + (c_2 \times rand() \times (p_{gbest} - p_i(t))) \tag{5}$$

$$p_i(t+1) = p_i(t) + v_i(t) \tag{6}$$

where v_{i} $(t+1)$ = updated velocity of the *ith* particle; φ = inertia weight; c_1 and c_2 = weighting coefficients for individual best and global best positions respectively; p_i (t) = position of the *ith* particle at time t; p_i^{best} = best known position of *ith* particle so far, and p_{gbest} = best position of any particle in the swarm until now. The *rand ()* = a function that generates the value of a uniformly distributed random variable ∈ [0, 1].

3.5 Model reliability and performance evaluation

3.5.1 Cross-fold validation

Cross-fold validation is utilized to compare the predictive accuracy of two or more models with reliability. Initially, *k*-fold cross-validation randomly splits the dataset into *k* individual and equally sized subsets. One of the partitioned subsets is used for validation while the rest of them are applied to generate the training model. The aforementioned procedure is repeated *k* times to complete the cross-validation. Kohavi (1995) proposed that when the value of *k* is 10, the cross—validation is analytically valid, and the variance, bias and computation efficiency are all optimal (Kohavi, 1995).

3.5.2 Statistical indicators

3.5.2.1 Pearson correlation coefficient (R) The Pearson correlation coefficient is a measure of the correlation between two variables; it has value from −1 to 1. A value closer to 1 indicates a more positive correlation between the variables. A value closer to −1 indicates a higher negative correlation between the variables, representing the two variables have a higher negative correlation. A correlation coefficient of 0 indicates no linear relationship between the variables. Equation (7) defines Person's R.

$$\mathrm{R} = \frac{n\sum y.y' - \left(\sum y\right)\left(\sum y'\right)}{\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}\sqrt{n\left(\sum y'^2\right) - \left(\sum y'\right)^2}} \tag{7}$$

where y = actual value; y' = predicted value; and n = number of samples.

3.5.2.2 Mean absolute percentage error (MAPE) MAPE is also known as the mean absolute percentage deviation, which is a statistical measure of predictive accuracy. By calculating the ratio between the relative differences and the baseline value, it is unaffected by sample size or units. A lower MAPE indicates greater accuracy of the model. Equation (8) displays the mathematical formula of MAPE.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y' - y}{y}\right| \tag{8}$$

where y = actual value; y' = predicted value; and n = number of samples.

3.5.2.3 Root mean squared error (RMSE) RMSE represents the degree of dispersion between predicted values and actual values; it is the square root of the average sum of the errors between them. The RMSE is given by Eq. (9).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y' - y)^2} \tag{9}$$

where y = actual value; y' = predicted value; and n = number of samples.

3.5.2.4 Mean absolute error (MAE) MAE is the absolute average error between actual and predicted values; and so its unit is the same as the observed values. The MAE can be computed using Eq. (10).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y - y'| \tag{10}$$

where y = actual value; y' = predicted value; and n = number of samples.

3.5.2.5 Synthesis index (SI) To understand the performance of each model comprehensively, the synthesis index (SI) is introduced. SI is a statistical measure that is based on the four aforementioned statistical indicators, and is computed using 1-R, MAPE, RMSE, and MAE. The range of SI is between 0 and 1; a value close to 0 indicates that the model has a high accuracy in average perspective. The equation for SI is as follows.

$$\mathrm{SI} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{P_i - P_{\min,i}}{P_{\max,i} - P_{\min,i}}\right) \tag{11}$$

where m = number of performance measures; $P_i = i^{th}$ performance measure.

4 House price dataset

The dataset was purchased with granted funds from the Ministry of Science and Technology, Taiwan, and obtained from the real price registration system (RPRS) of the Ministry of the Interior, Taiwan. The RPRS includes the prices and many attributes in all housing-related transactions in Taipei City, Taiwan, during the period from 2013 to 2017, with a total of 209,402 samples. The samples undergo a data cleaning in which a box plot is used to eliminate extreme outliers that were not considered normal appraisal processes. The features based on literature suggestions were selected from the RPRS for subsequent analysis. In particular, the actual distance between locations of real estate and objectives, which are Yes In My Back Yard (YIMBY) and Not In My Back Yard (NIMBY), is calculated by transforming the coordinate TWD97, obtained from the real price registration system, to WGS84. The end of this section will define prestigious elementary and junior high school districts.

4.1 Data preprocessing

4.1.1 Data cleaning

The transaction data that were provided by the RPRS include a number of transaction targets, such as land only and parking space transactions. Also, information is missing in cases in which the transaction is abnormal. To ensure that each datum can be used, information that does not meet specified conditions is deleted. Only information that is complete and consistent with normal conditions is retained.

4.1.1.1 Screening based on transaction target The purpose of this study is to predict house prices using both internal and external features. Therefore, only transaction samples for "room + land" and "room + land + parking space" are used. A total of 13,695 samples were removed to maintain the consistency of the dataset.

4.1.1.2 Screening based on transaction period The law on registering real prices came into effect in August, 2012. Earlier registration data were deficient. To ensure continuity of the dates associated with valid prices as well as completeness of the data, transaction data for Taipei City from January 2013 to December 2017 are used. A total of 8365 samples are excluded for this reason.

4.1.1.3 Screening based on total construction floors Most building types in Taipei City are residential buildings with various numbers of floors—under six, between 6 and 11 or over 12 floors. Such buildings are less heterogeneous than other kinds of real estate. Residential buildings whose numbers of floors fell within these three categories were used, while others were expunged. A total of 44,977 samples thus eliminated.

4.1.1.4 Screening based on area Property can be registered as shared upon registration of the price, and each co-owner is free to dispose of his or her due portion. The share may be involved in legal auctions, mergers or other transactions that are authorized under Article 23 of the Technical Rules for the Valuation of Real Estate; such transactions cannot be identified in the remarks column of RPRS database. To guarantee that all the transaction data are associated with normal trades, a total of 183 targets with areas of under three pings (floor area unit particularly used in Taiwan), which may involve a partial transfer of rights of commonly held properties, were thus removed.

4.1.1.5 Screening based on building age The value of real estate is reduced by physical, functional or economic depreciation as its age increases. Therefore, data that are not associated with a specified building age are removed. Furthermore, this study focuses on pre-owned housing, and homes with an age of less than one year may be pre-sold. Therefore, data for dwellings that are less than a year old are deleted. A total of 39,786 samples were thus excluded.

4.1.1.6 Screening based on transaction floor If the transaction involves property on the first floor, it involves a commercial target, which falls outside the scope of our study. In addition, if a transaction involves more than one floor or the basement, the total transaction price will not reflect the market. Therefore, data concerning transactions that involve the first floor, multiple floors or contains the basement are removed. A total of 15,619 samples are eliminated.

4.1.1.7 Screening based on number of rooms Owners of real estate in Taipei City often divide their home into several rental suites, which can be classified as income-type real estate. The value of this category of homes is difficult to compare with that of ordinary ones. To avoid including possible rental suites, transactions that involve more than six rooms are excluded. A total of 312 samples are thus expunged.

4.1.1.8 Screening based on main usage To ensure the homogeneity of the massive valuation models that will be developed in this study, only transactions that involve "residential usage" and "residential and commercial usage" as the main utilization are retained, while the rest are eliminated. A total of 58,504 are thus removed.

4.1.1.9 Screening based on parking space In the RPRS, some parking spaces of independent property owners are mandatorily registered, while others are "openly registered", so their cost and area need not be specified. To maintain the consistency of the data, transactions without the registered price and the area of the parking space are excluded. A total of 3318 samples are thus eliminated.

4.1.1.10 Screening based on special transaction Article 23 of the Technical Regulations for Real Estate Valuation requires that real estate appraisers consider underlying trading conditions when selecting a target for comparison. If a special transaction status is specified, an appropriate adjustment should be made. If the circumstances that affect the transaction price cannot be effectively grasped, then the transaction should not be considered. Consequently, to ensure that the transaction price satisfies both criteria for so-called "normal price" in the valuation technical rules, only normal transactions specified in the remarks column are considered. A total of 11,022 samples were thus expunged.

4.1.2 Removing outliers

A box plot is used to detect and eliminate extreme outliers that fall outside the outer fences in the original dataset. Figure 1 shows the box plot of the original dataset after cleaning. A total of 401 extreme outliers that exceed the upper outer fence are then excluded. Following the box plot elimination, 13,220 samples remain, with minimum = 1.00E+06, maximum = 7.41E+07, average = 2.14E+07 and standard deviation = 1.33E+07. Figure 2 further presents the density of samples in each district of Taipei City. The resulting samples that fit to the study scope are subsequently utilized to develop house price prediction models.

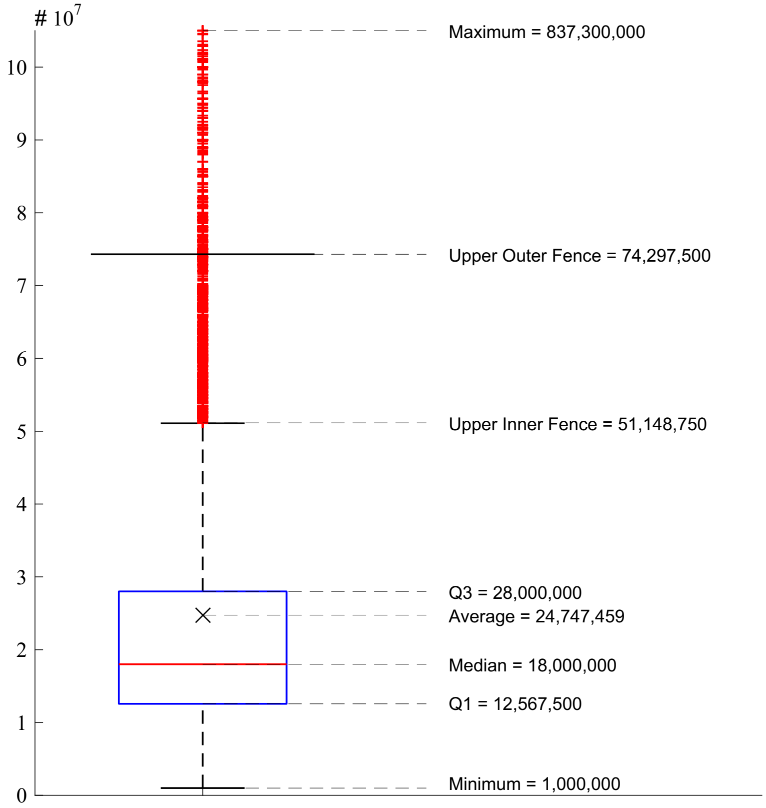

Fig. 1 Box plot of original dataset (Unit: NTD)

4.2 Feature description and acquisition

4.2.1 Feature identification

House price are affected by geopolitical features, regional features and individual features of the dwelling. Geopolitical features involve economic, political, social and environmental conditions, which can be implicitly reflected by the date of the house transaction. Individual features designated are obtained from the information available in the RPRS. Regional factors include Amenities and Disamenities. The features in RPRS were compared with the influential features addressed in the literature. The selected factors will be described below. Table 1 defines Amenities and Disamenities. Table 2 lists the 54 ultimately determined features. Table 3 provides comprehensive information about the collection of features from the literature (Chiang et al., 2017; Chau & Chin, 2003; Du et al., 2013; Geng et al., 2015; Huang et al., 2017a; Huang et al., 2017b; Lin, 2004; Rong & Sun, 2014; Wang, 2011; Wei et al., 2010; Wen et al., 2005; Wu et al., 2014; Yang & Su, 2011; Zhang & Zhang, 2010). The response variable and individual features are described as below.

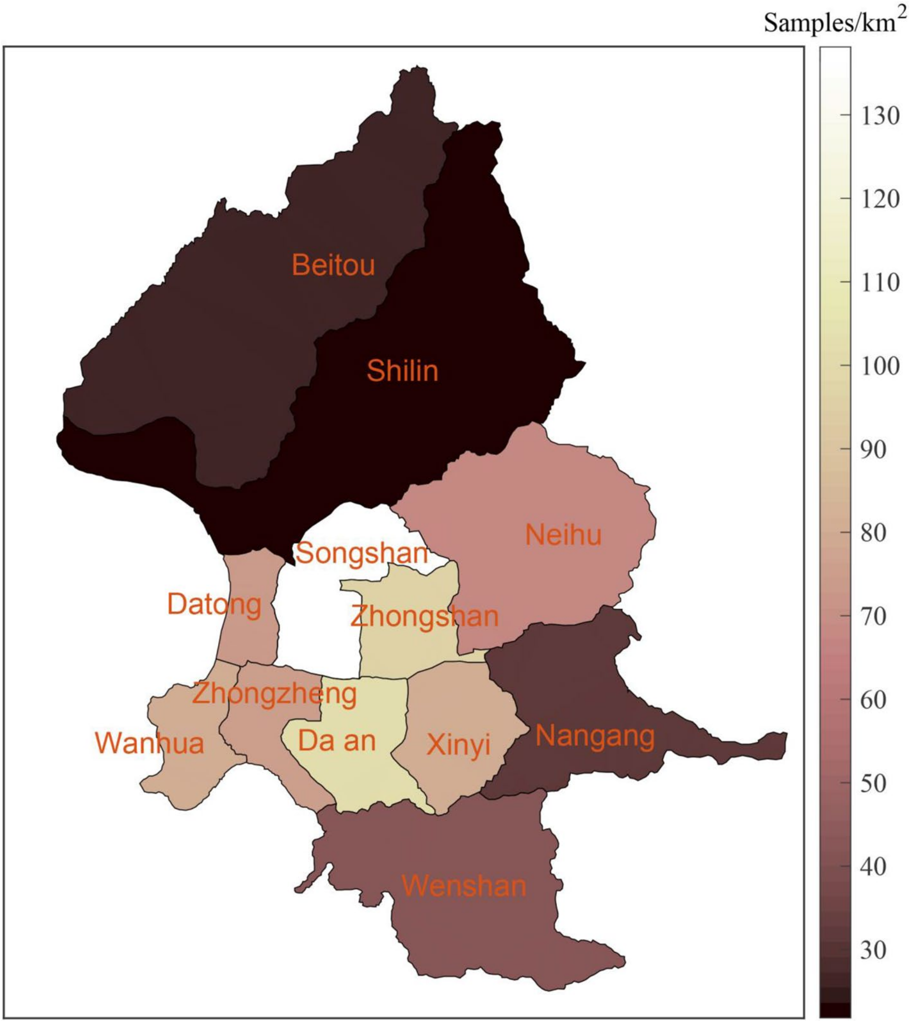

Fig. 2 Density of samples at each district of Taipei City

4.2.1.1 Response variable The response variable is the *acquisition price of the real estate (house price)*, which is provided in RPRS.

4.2.1.2 Characteristic features of the housing *Building type*: Most buildings in Taipei City are residential buildings with fewer than six and 11, or more than 12 floors, resulting in great homogeneity. The construction costs of various types of building, and their functions must be considered.

Area: A building can be divided into three segments: the main building area, the auxiliary building area and a public-shared facilities area. The sum of these three areas, over which the owners have land rights, is usually proportional to the house price in a nonlinear pattern.

Table 1 Description of amenities and disamenities

Category	Feature	Description
Amenity YIMBY	MRT (Mass Rapid Transit) exit	Exits of all MRT stations in Taipei City
	Neighborhood park	The parks with an area of over 1 hectare in Taipei City
	Large park	The parks with an area of over 10 hectares in Taipei City
	Prestigious school	The prestigious elementary and junior high schools with full recruitment during the period from 2013 to 2017
	Department store	33 department stores in Taipei City on Google Maps
	Large gymnasium	12 Sports Centers in Taipei City and Taipei Gymnasium
Disamenity NIMBY	Funeral parlor	The First and Second Funeral Parlor in Taipei City
	Airport	Taipei Songshan Airport
	Substation	Substations located in Taipei City
	High voltage tower	High voltage towers located in Taipei City
	Waste incineration plant	Neihu, Muzha, and Beitou Waste Incineration Plants
	Sewage treatment plant	Neihu and Dihua Sewage Treatment Plants
	Gas station	Gas stations located in Taipei City

Age: As time goes on, buildings lose value owing to the weather, negative effects of natural disasters and internal use. Hence, the age of a house is an important factor.

Number of rooms: The number and areas of living rooms, bedrooms and bathrooms have a significant impact on the price of real estate. Different families have different needs.

Number of total floors: More floors in a building correspond to greater cost of project development. Under ordinary circumstances, more floors correspond to a higher house price.

Transaction floor: The transaction floor is the floor on which the real estate is located. A higher floor of the residence provides better views.

Parking space: The presence or absence of parking spaces affects, and the number of parking spaces is usually nonlinearly proportional to, the value of a dwelling.

Types of parking space: Whether a parking space is on the ground or uses an elevator affects house prices.

4.2.1.3 Disamenity NIMBY *Funeral parlor*: Funeral ceremonies sometimes generate loud noises, which reduce tranquility and folk customs and religious beliefs make most people unwilling to live in the neighborhood.

Airport: In areas around airports, the noise of aircraft during take-off and landing can be heard, and the construction and height of surrounding buildings are limited.

Substation: Residents prefer to keep their distance from the powerful electromagnetic waves around substations.

High voltage tower: Residents prefer to keep their distance from the powerful electromagnetic waves around high voltage towers.

Waste incineration plant: Although large-scale waste incineration plants have pollution control standards, their pollution of the air has a negative impact on willingness to purchase houses nearby.

Table 2 Original features available from RPRS for house price prediction

Category	Feature	Description
Characteristics of house	Area	Ping
	Age	Year
	Building under 6 floors	Yes = 1, No = 0
	Building between 6-11 floors	Yes = 1, No = 0
	Building over 12 floors	Yes = 1, No = 0
	Brick building	Yes = 1, No = 0
	RC building	Yes = 1, No = 0
	SRC building	Yes = 1, No = 0
	Number of living rooms	Value
	Number of bedrooms	Value
	Number of bathrooms	Value
	Number of total floors	Value
	Number of parking spaces	Value
	Transaction floor	Value
	Ground parking space	Yes = 1, No = 0
	Elevator parking space	Yes = 1, No = 0
	Other types of parking space	Yes = 1, No = 0
Transaction period	The first half of 2013	Yes = 1, No = 0
	The second half of 2013	Yes = 1, No = 0
	The first half of 2014	Yes = 1, No = 0
	The second half of 2014	Yes = 1, No = 0
	The first half of 2015	Yes = 1, No = 0
	The second half of 2015	Yes = 1, No = 0
	The first half of 2016	Yes = 1, No = 0
	The second half of 2016	Yes = 1, No = 0
	The first half of 2017	Yes = 1, No = 0
	The second half of 2017	Yes = 1, No = 0
Administrative region	Beitou District	Yes = 1, No = 0
	Shilin District	Yes = 1, No = 0
	Datong District	Yes = 1, No = 0
	Songshan District	Yes = 1, No = 0
	Zhongshan District	Yes = 1, No = 0
	Neihu District	Yes = 1, No = 0
	Wanhua District	Yes = 1, No = 0
	Zhongzheng District	Yes = 1, No = 0
	Da’an District	Yes = 1, No = 0
	Xinyi District	Yes = 1, No = 0
	Nangang District	Yes = 1, No = 0
	Wenshan District	Yes = 1, No = 0

Table 2 (continued)

Category	Feature	Description
Amenity YIMBY	MRT exit	Distance in meter
	Park within 1 hectare	Distance in meter
	Park over 10 hectares	Distance in meter
	Gymnasium	Distance in meter
	Prestigious elementary school	Distance in meter
	Prestigious junior high school	Distance in meter
	Department store	Distance in meter
Disamenity NIMBY	Funeral parlor	Distance in meter
	Airport	Distance in meter
	Substation	Distance in meter
	High voltage tower	Distance in meter
	Waste incineration plant	Distance in meter
	Sewage treatment plant	Distance in meter
	Hospital	Distance in meter

Sewage treatment plant: Sewage treatment plants may produce odor affecting the prices of nearby dwellings.

Gas station: Gas stations may have environmental impacts by causing traffic jams, noise, air pollution and a risk of explosion in nearby areas.

Hospital: Since diseases and viruses may spread, residents are not willing to dwell adjacent to a hospital.

4.2.1.4 Amenity YIMBY *MRT*: A number of studies have confirmed that the distance between a building and an MRT station influences the value of real estate. While most studies consider the straight line distance between the two, actual walking distances is used herein as the input distance factor.

Park: Green resources in metropolitan areas are not easy to access. Closeness to park corresponds to higher value of a dwelling. Actual walking distance is the input distance factor herein.

Table 3 Collection of features from literature

Literature	Region	Response variable	Feature
Yang & Su (2011)	Taipei City, Taiwan	House price	Distance to schools, distance to large parks, distance to department stores, distance to MRT stations, distance to large stadiums, distance to funeral parlor, distance to sewage treatment plants, distance to temples, distance to substations, distance to waste incineration plants
Lin (2004)	Taipei City, Taiwan	House price	Building area, age of the house, floor, building structure, administrative division, school district to which the building belongs
Huang et al. (2017a)	Taipei City, Taiwan	House price	Administrative division, transaction season, house type, floor, total floors, age of the house, number of pings, whether it is close to MRT stations, whether it is close to school, whether it is close to park, distance to city center
Chiang et al. (2017)	Taipei City, Taiwan	House price	Building area, age of the house, trading floor, residential building, trading season, distance to Youbike sites, distance to city center, distance to urban area, distance to hospitals, distance to parks, distance to schools, distance to MRT stations, distance to NIMBY
Wen et al. (2005)	Hangzhou City, China	House price	Building area, distance to West Lake, internal environment, distance to business districts, traffic conditions, garage, loft, decoration, environment, community management, history of the house, entertainment facilities, transaction period, nearby universities
Wei et al. (2010)	Harbin City, China	House price	Building area, distance to the business districts, living function, age of the house, neighboring cultural or sports facilities, floors, whether there is a basement or building, educational facilities, decoration
Zhang & Zhang (2010)	Jilin Province, China	House price	Building area, number of rooms, air heating system, age of the house, distance to business districts, air pollution level within the residential area, distance to polluted rivers
Wang (2011)	Shanghai City, China	House price	Administrative division, traffic conditions, building area, age of the house, orientation of the house, decoration, floor, surrounding environment, management organization, library, educational facilities, management fee, vacant period
Rong & Sun (2014)	Kunming City, China	House price	Distance to railways, property costs, floor area ratio, distance to educational facilities, distance to bus stops, decoration, building area
Geng et al. (2015)	Beijing City, China	House price	Distance to high-speed railway stations, the school district of elementary and junior high school, distance to the nearest university, distance to hospitals, distance to supermarkets, distance to parks, distance to administrative agencies, distance to MRT stations, distance to the nearest sightseeing spot
Wu et al. (2015)	Shenzhen City, China	House price	Distance to the Central Business District (CBD), distance to parks, distance to schools, distance to main roads, distance to subways
Huang et al. (2017b)	Shanghai City, China	House price	Building area, population density of administrative division, whether it is close to schools, decoration, sex ratio, orientation of the house, floor, GDP, average wage, population density, distance to urban area

Table 3 (continued)

Literature	Region	Response variable	Feature
Chau & Chin (2003)	Worldwide	–	Distance to business area, whether it has sea, river or lakeshore landscape, whether there are hills, valleys or golf course landscape, whether the viewing of house is facing obstacles, land lease period, number of living rooms, bedrooms and bathrooms; building area, parking space, basement, courtyard, elevator or air-conditioning system, floor, structural materials, facilities (pool, playground, etc.), residents' income, whether it is close to prestigious schools, whether it is close to hospitals, whether it is close to religious gathering places, crime rate, noise disturbance, distance to shopping centers, distance to forests, environmental quality
Du et al. (2013)	Taiwan	–	Age of the house, building area, the administrative division, type of the building, total number of floors, whether it is the first floor, the date of transaction, the number of bathrooms, structural type

Table 4 Default parameter settings of filter-based algorithm in WEKA

Algorithm	Parameter	Default value
ReliefFAttributeEval	doNotCheckCapabilities	False
	numNeighbours	10
	sampleSize	− 1
	Seed	1
	Sigma	2
	weightDistance	False
Ranker	generateRanking	True
	numToSelect	− 1
	startSet	N/A
	threshold	− 1.7976931348623157E308

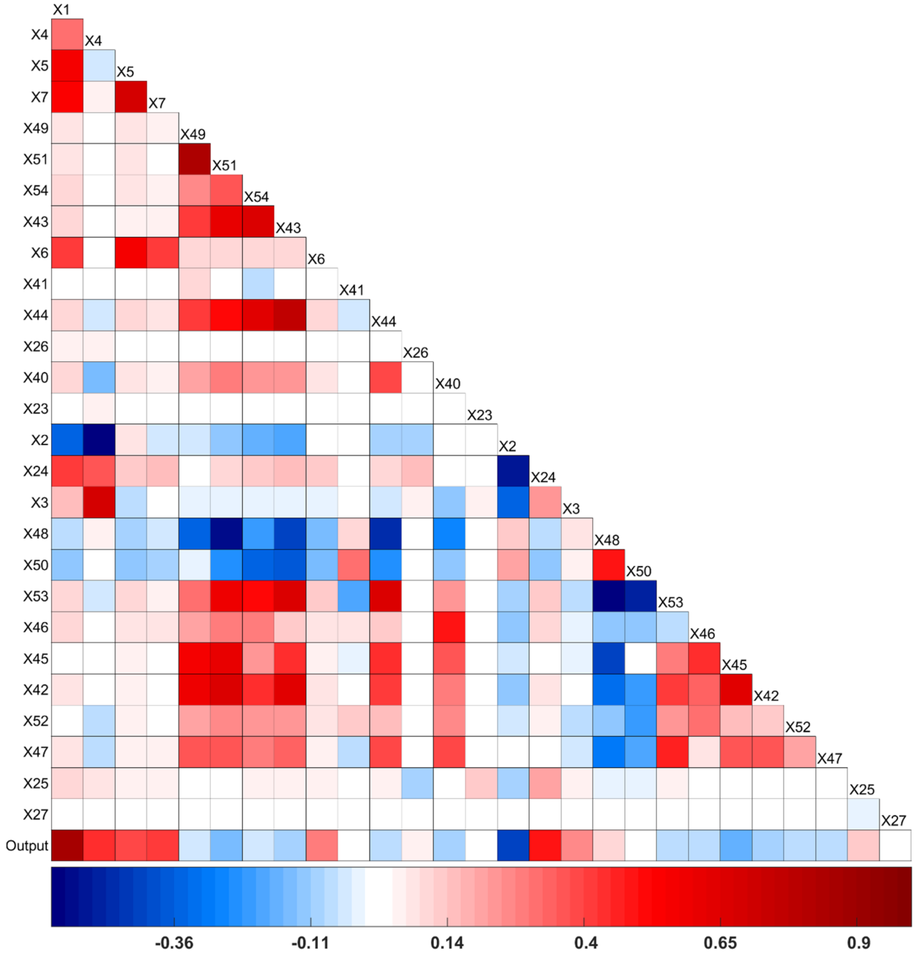

Fig. 3 Feature correlation matrix for the house price dataset

Table 5 Ultimately filtered features for house price prediction

No.	Rank	Var.	Feature	Min	Max	Mean	Std. dev.
1	0.0713702	X1	Area	3.09155	132.9336	34.13144	14.9292
2	0.041458	X2	Age	2	55	24.22065	13.78737
3	0.0323912	X4	Number of total floors	2	33	8.821558	4.701371
4	0.0305644	X24	Number of parking spaces	0	6	0.253555	0.534043
5	0.030357	X5	Number of rooms	0	6	2.672542	0.980382
6	0.0189563	X3	Transaction floor	2	28	5.353631	3.378328
7	0.0158378	X7	Number of bathrooms	0	6	1.611422	0.674213
8	0.0143893	X48	High voltage tower	24	3622	1312.73	837.1312
9	0.010155	X49	Sewage treatment plant	131	9634	4000.833	1,872.908
10	0.0087934	X50	Waste incineration plant	286	9640	4193.257	1983.481
11	0.0071486	X51	Airport	633	10,972	5043.304	2267.859
12	0.0061982	X53	Funeral parlor	70	10,330	4115.926	2432.05
13	0.0055773	X54	Hospital	28	4598	1235.662	769.8108
14	0.0054296	X46	Prestigious elementary school	23	9110	2187.164	1630.969
15	0.0052015	X43	Department store	0	8744	2134.945	1602.708
16	0.0044702	X45	Prestigious junior high school	76	7356	1900.686	1198.961
17	0.0040854	X6	Number of living rooms	0	6	1.69357	0.575074
18	0.0040759	X42	Park over 10 hectares	88	9480	2635.467	1597.882
19	0.0040426	X41	Park within 1 hectare	3	3999	758.643	411.7716
20	0.0035121	X52	Substation	11	3135	749.7481	421.4679
21	0.0032671	X44	Gymnasium	66	7860	2004	1209.426
22	0.003145	X47	Gas station	11	2761	726.3085	491.2171
23	0.0029717	X26	Elevator parking space	0	1	0.057716	0.233214
24	0.001303	X25	Ground parking space	0	1	0.16354	0.369872
25	0.0009398	X40	MRT exit	0	7773	831.9968	710.485
26	0.0004644	X27	Other types of parking space	0	1	0.002648	0.051388
27	0.0002376	X23	SRC building	0	1	0.023071	0.150135
–	–	Y	House price (NTD)	1,000,000	74,100,000	21,421,931	13,319,624

School: Some parents purchase houses in prestigious school districts. Closeness to prestigious schools increases the probability of being a member of the corresponding prestigious school district. Distance between residence and prestigious school is therefore one of the factors used in modeling.

Department store: Department stores in Taipei City are often located in business districts, and homes next to such districts tend to have a higher price due to their greater convenience.

Public sports facility: Public sports centers or gymnasiums provide sports and leisure services. Residents who live adjacent to large sports centers have more convenient access to leisure. Considering the homogeneity of facilities, this study is concerned only indoor sports centers.

4.2.2 Feature selection using filter-based algorithm

In feature selection, a filter-based algorithm, ReliefFAttributeEval, is used. WEKA, which stands for Waikato Environment for Knowledge Analysis, was developed at the University of Waikato, New Zealand. It is a machine learning package with free accessibility and a user-friendly graphical interface; it provides diverse algorithms for data mining, including for data pre-processing, regression, clustering, classification and visualization.

For simplicity, ReliefFAttributeEval and Ranker are set to their default parameters, which are exhibited in Table 4. Features with a rank of greater than zero are selected as ultimate attributes for the following studies, in which 27 features were extracted for model construction. Figure 3 shows the feature correlation matrix for the house price dataset. Table 5 presents the ultimately filtered 27 features for house price prediction.

4.2.3 Distance measurement method

Several ways can be used to calculate the actual walking distance between a building and nearby amenities or disamenities. One is to input the house address or coordinates into Google Maps and to use the path calculation API, including the Directions API or the Distance Matrix API, to find the best route to public facilities such as MRT, parks, and others. However, at present, Google Maps only offers this service in web-browsing mode. More than 10,000 samples were collected from RPRS, and manually collecting distance information from the webpage would be very time-consuming and probably result in human error. Therefore, pre-collected coordinates of various facilities, such as the 364 exits of Taipei MRT stations, 33 department stores, 13 large parks, 129 neighborhood parks and sports centers and 13 Taipei Gymnasiums are used herein to calculate actual walking distances.

Table 6 Prestigious elementary and junior high schools in Taipei from 2013 to 2017

Elementary school	Junior high school
Taipei Municipal DunHua Elementary School	Taipei Municipal Lishan Junior High School
Taipei Municipal Jian-Kang Elementary School	Taipei Municipal DunHua Junior high School
Taipei Municipal BoAi Elementary School	Taipei Municipal Xingya Junior High School
National Taipei University of Education Experimental Elementary School	Taipei Municipal Jinhua Junior High School
Taipei Municipal XinSheng Elementary School	Taipei Municipal Long Men Junior High School
Taipei Municipal Jinhua Elementary School	Taipei Municipal Zhongzheng Junior High School
Taipei Municipal RenAi Elementary School	Taipei Municipal Shipai Junior High School
Taipei Mandarin Experimental Elementary School	Taipei Municipal Ming-Hu Junior High School
Affiliated Experimental Elementary School of University of Taipei	The Junior High Division of the Affiliated Senior High School of NTNU
Taipei Municipal PengLai Elementary School	Taipei Municipal Chengyuan Junior High School
Taipei Municipal NanGang Elementary School	Taipei Municipal Datong Junior High School
Taipei Municipal WenHua Elementary School	The Affiliated High School of National Chengchi University
Taipei Municipal Xing-An Elementary School	Taipei Municipal Yangming Junior High School
Taipei Municipal Ming Hu Elementary School	Taipei Municipal NanGang Junior High School
Taipei Municipal Shipai Elementary School	Taipei Municipal ZhongShan Junior High School

Table 7 Default parameter settings of baseline models in WEKA

Baseline model	Parameter	WEKA Default value
Artificial neural network	Name	Multilayer perceptron
	Hidden layer specification	N/A
	Number of hidden layers	1
	Number of hidden nodes	(attributes + classes)/2
	Number of learning iterations	500
	Momentum	0.2
	Type of normalize	[− 1,1]
	Shuffle examples	True
	Learning rate	0.3
Support vector machine	Name	SMOreg
	Kernel type	RBF
	Regularization parameter (C)	1
	ε (epsilon)	0.001
	RBF γ (gamma)	0.01
	Stopping criteria	Tolerance: 0.001
Classification and regression tree	Name	REPTree
	Tree depth	Infinite
	Number of random splits/leaf node	N/A
	Min number of samples/leaf node	2
	Create trainer mode	Single parameter
	Minimum variance for split (V)	1.0E−3
	Number of folds for reduced error pruning (N)	3
	Seed for random data shuffling (S)	1
	Pruning (P)	True
Linear regression	Name	Linear regression
	Fitting regression line method	Ordinary least squares
	L2 regularization weight	1.0E−8
	Features selection	M5-prime
	Use bias	N/A
	Eliminate collinear features	True
	Tolerance	N/A

Initially, the straight-line distances between buildings and each facility were calculated, and then the three coordinates with the shortest distance were selected. Finally, information from an Excel file with the three coordinates was inputted into the Python computing platform and the coordinate data were loaded to the Google Maps server through the Google Maps Distance Matrix API. The server typically returned three optimal values of distances and the shortest was chosen, and input as the actual walking distance.

Table 8 Default parameter settings of ensemble models in WEKA

Ensemble model	Parameter	WEKA Default value
Voting	Name	Voting
	batchSize	100
	Classifiers	Customized
	combinationRule	Average of probabilities
	numDecimalPlaces	2
	Seed	1
Bagging	Name	Bagging
	bagSizePercent	100
	batchSize	100
	Classifier	Customized
	numDecimalPlaces	2
	numExcutionSlots	1
	numIterations	10
	Seed	1
Stacking	Name	Stacking
	batchSize	100
	Classifiers	Customized
	metaClassifier	ZeroR
	numDecimalPlaces	2
	numExcutionSlots	1
	numFolds	10
	Seed	1

4.2.4 Location of residential school

Many studies have suggested that prestigious school districts positively influence housing prices. The definition of “school district” for the elementary and junior high schools in Taipei City is not based on the shortest distance between buildings and the schools but on the location of the residences. The government makes minor adjustments to the school district every year, affecting the residences that belong to it, so the school districts of elementary and junior high schools may vary.

Residences that are closer to prestigious elementary and junior high schools are more likely to belong to the prestigious school district. Therefore, the shortest walking distance from the building to the prestigious school is used herein as a variable. Prestigious school used herein include elementary and junior high schools that were full during the period from 2013 to 2017. Table 6 presents all prestigious elementary and junior high schools that meet this criterion.

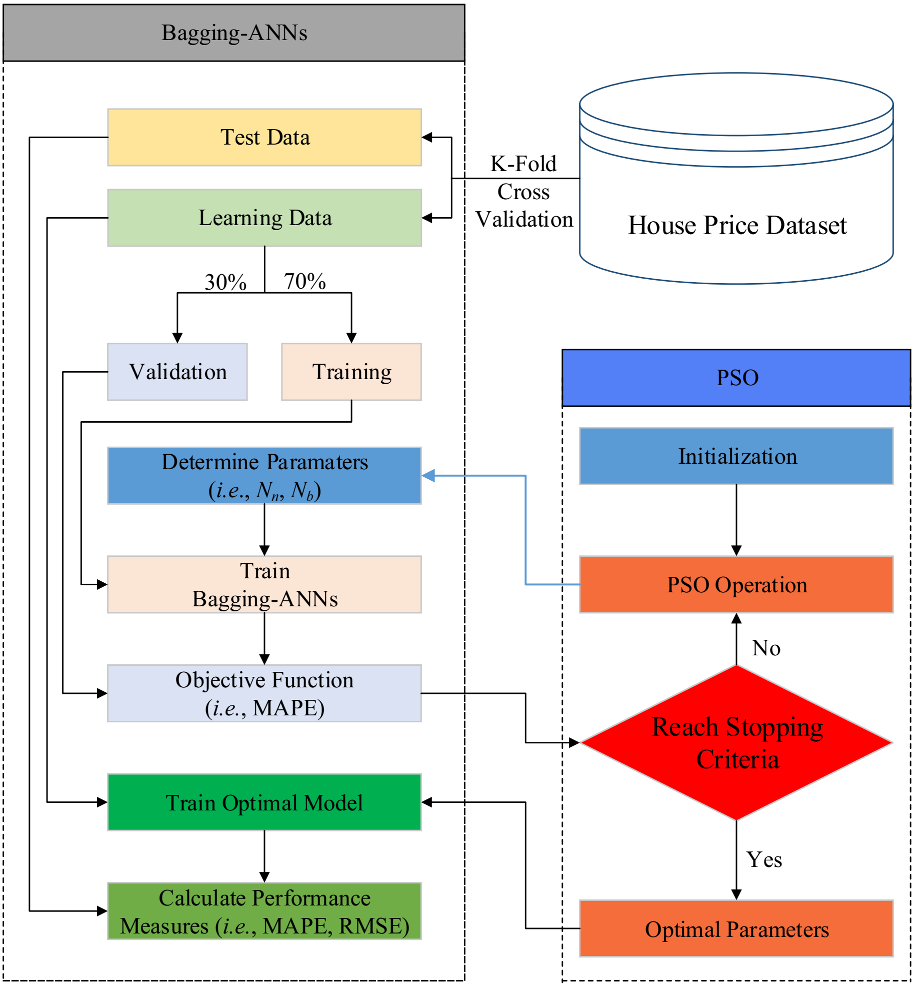

Fig. 4 Architecture of hybrid model

5 Model development

5.1 Baseline prediction models

For ease of use and efficiency, all of the main parameters of the baseline models that are developed in WEKA, excluding SMOreg (SVR), are set to their default values; RBFkernel is used instead of PolyKernel as the kernel function of SMOreg (SVR). Table 7 presents the default parameter settings for all baseline models.

WEKA provides a user-friendly graphical interface for executing an analytic procedure, enabling users to input data; adjust the parameters of the algorithm, and eventually construct models merely by clicking buttons.

```
Begin
1. Initialization stage
        Subdivide the data into k subsets as learning data (training data and validation data) and test data
        Initialize number of particle, maximum iteration, inertial weight and the boundary of optimized parameters
2. Perform K folds such that, for each fold, the following steps are performed
    while ($t$ < Max_Iteration) do
    2.1 PSO operator
        (1) Find individual best and global best of each particle
        (2) Calculate velocity of each particle
        (3) Update new location for each particle
    end while
    2.2 Bagging- ANNs function for validation
        (1) Set the hyperparameters: $N_n, N_b$
        (2) Train model with hyperparameters $(N_n, N_b)$ for Bagging- ANNs
        (3) Evaluate the trained (optimized) model using validation data by Eq. (8)
        (4) Determine the fitness function $f(N_n, N_b)$ by Eq. (12) and go to step 2.1
    2.3 Have the stopping criteria been met?
        (1) If the criteria have been met, go to step 3
        (2) Otherwise, go to step 2.1
3. Optimized Bagging- ANNs model
        (1) Incorporate the optimized parameters $(N_n, N_b)$ into the model using learning data
        (2) Calculate the average accuracy over the k test folds from Eq. (7) to Eq. (10)
        (3) Save K optimal models
4. Plot stage
        (1) Evaluate the post-process results and visualize the results and confirm the best solution
End
```

Fig. 5 Pseudocode of PSO-Bagging-ANNs

Table 9 Parameter settings of PSO-Bagging-ANNs prediction model in MATLAB

Algorithm	Parameter	Value
PSO-Bagging-ANNs	Max iteration	40
	Number of particle	30
	Inertia weight	0.9
	c_1	2
	c_2	2
	Number of bags	[1, 10]
	Number of neurons	[1, 25]
	Activation function	tansig
	Solver	trainbr

5.2 Ensemble prediction models

In WEKA, the implementation of ensemble models is analogous to that of single models. Thus, only the parameters of ensemble models, which are voting, bagging, and stacking (listed in Table 8), are considered in this section.

5.3 Hybrid prediction model

The hybrid prediction model is developed in MATLAB. MATLAB, created by MathWorks, allows users to plot functions, manipulate matrixes, execute algorithms and create customized user interfaces. Owing to its ability to perform advanced mathematical computations, MATLAB was used herein to integrate the PSO algorithm with the best model in baseline and ensemble schemes, which is bagging ANNs, to develop a hybrid model, namely, PSO-Bagging-ANNs. Figures 4 and 5 show the architecture and pseudo-code of this hybrid model, respectively. The algorithmic parameters of PSO include the number of particles, the maximum number of iterations, the inertia weight, and c_1 and c_2. Based on literature reviews, a rule of thumb and numerical experiments (Kamaruddin & Ravi, 2016; Yu & Xiaohui, 2011; Zhan et al., 2009), the following parameters were set; number of particles = 30, maximum number of iterations = 40, inertia weight = 0.9 and $c_1 = c_2 = 2$. After the parameters of PSO were set, the algorithm was used to optimize bagging ANNs. The objective function is presented as

$$f\left(N_n, N_b\right) = MAPE^{Training\,process}_{Validation\,data} \tag{12}$$

where N_n is the number of neurons and N_b is the number of bags.

Optimizing the number of bags allows the bagging algorithm to search the best number of bags and make the best predictions. However, to prevent overlapping and overfitting when the number of bags is very large, the upper bound on the number of bags was set to ten. The literature suggests that for ANNs, the hyperbolic tangent activation function, as well as Bayesian regularization backpropagation solver, which are tansig and trainbr in MATLAB respectively, are more likely than other functions to yield superior predictions (Isa et al., 2010; Karlik & Olgac, 2011; Kaur & Salaria, 2013; Methaprayoon et al., 2007). With respect to the number of neurons in ANNs, Armaghani et al. (2017) recommended that the use of $(N_i + N_o)/2$ hidden nodes, where N_i = number of inputs and N_o = number of outputs (Armaghani et al., 2017). Thus, considering the aforementioned threshold as well as the affordability of hardware devices,

Table 10 Prediction performance of baseline models in WEKA

Model	Performance measure				
	R	RMSE (NTD)	MAE (NTD)	MAPE (%)	$SI_{overall}$
ANNs	**0.948**	4,927,394	**3,018,614**	**16.20**	**0.000 (1)**
SVR	0.929	5,124,475	3,426,277	17.25	0.530 (3)
CART	0.932	**4,837,640**	3,214,792	16.21	0.160 (2)
LR	0.921	5,203,429	3,747,564	21.47	1.000 (4)

Bold value denotes the best performance in all baseline models; (.) indicates the overall ranking

Table 11 Prediction performance of ensemble models in WEKA

	Model	Performance measure				
		R	RMSE (NTD)	MAE (NTD)	MAPE (%)	$SI_{overall}$
Voting	ANNs + CART	0.956	3,930,235	2,697,942	14.02	0.141 (4)
	ANNs + SVR	0.951	4,173,897	2,887,190	14.91	0.275 (11)
	ANNs + LR	0.948	4,232,145	3,015,086	16.47	0.373 (17)
	CART + SVR	0.945	4,426,900	2,945,696	14.56	0.353 (15)
	CART + LR	0.943	4,445,766	3,056,225	16.07	0.438 (21)
	SVR + LR	0.925	5,101,488	3,552,547	19.16	0.846 (25)
	ANNs + CART + SVR	0.956	3,969,050	2,704,909	13.71	0.140 (3)
	ANNs + CART + LR	0.954	3,994,342	2,772,285	14.53	0.191 (7)
	CART + SVR + LR	0.942	4,531,509	3,100,485	16.10	0.468 (22)
	ANNs + SVR + LR	0.945	4,392,866	3,078,435	16.41	0.431 (20)
	ANNs + CART + SVR + LR	0.951	4,140,885	2,859,455	14.85	0.260 (10)
Bagging	ANNs	**0.961**	**3,681,032**	**2,518,849**	**13.00**	**0.000 (1)**
	SVR	0.929	5,113,746	3,418,603	17.20	0.739 (23)
	CART	0.954	3,981,850	2,600,441	13.10	0.110 (2)
	LR	0.921	5,204,110	3,747,883	21.46	1.000 (26)
Stacking	ANNs + CART	0.951	4,138,478	2,811,604	14.62	0.246 (9)
	ANNs + SVR	0.948	4,278,884	2,978,455	15.70	0.351 (14)
	ANNs + LR	0.947	4,330,566	3,055,567	16.54	0.409 (19)
	CART + SVR	0.945	4,399,824	2,917,823	14.42	0.343 (13)
	CART + LR	0.942	4,488,062	2,989,652	14.99	0.404 (18)
	SVR + LR	0.928	5,080,933	3,443,707	17.67	0.758 (24)
	ANNs + CART + SVR	0.953	4,030,459	2,728,454	13.92	0.175 (5)
	ANNs + CART + LR	0.953	4,056,317	2,756,743	14.19	0.197 (8)
	CART + SVR + LR	0.948	4,282,978	2,984,935	15.76	0.357 (16)
	ANNs + SVR + LR	0.945	4,393,659	2,916,369	14.42	0.342 (12)
	ANNs + CART + SVR + LR	0.953	4,030,031	2,728,187	13.91	0.175 (6)

Bold value denotes the best performance in all ensemble models; (.) indicates the overall ranking

Table 12 Optimal PSO-Bagging-ANNs parameter values in each fold

Fold	Number of bags (N_b)	Number of neurons (N_n)
1	5	15
2	7	16
3	6	19
4	5	17
5	6	19
6	5	17
7	6	22
8	5	22
9	7	19
10	6	19

Table 13 Prediction performance of PSO-Bagging-ANNs in MATLAB

Fold	Learning data				Test data			
	R	RMSE (NTD)	MAE (NTD)	MAPE (%)	R	RMSE (NTD)	MAE (NTD)	MAPE (%)
1	0.972	3,153,492	2,169,272	11.29	0.961	3,656,805	2,429,986	12.15
2	0.972	3,106,810	2,144,081	11.08	0.963	3,638,305	2,321,698	11.88
3	0.974	3,044,098	2,087,657	10.87	0.969	3,315,239	2,266,778	11.53
4	0.973	3,086,624	2,117,159	11.09	0.973	3,339,196	2,315,346	11.67
5	0.974	3,020,690	2,089,355	10.90	0.974	3,473,455	2,235,193	11.11
6	0.972	3,116,705	2,141,568	11.21	0.972	3,117,154	2,181,177	11.22
7	0.976	2,917,312	2,036,154	10.78	0.976	3,675,984	2,365,362	12.18
8	0.974	3,007,578	2,073,954	10.93	0.974	3,231,708	2,214,954	11.54
9	0.974	3,034,277	2,089,044	10.91	0.974	3,436,821	2,322,935	11.76
10	0.973	3,073,726	2,107,042	11.01	0.973	3,015,495	2,085,231	10.85
Mean	0.973	3,056,131	2,105,529	11.01	0.970	3,390,016	2,273,866	11.59
Max	0.976	3,153,492	2,169,272	11.29	0.976	3,675,984	2,429,986	12.18
Min	0.972	2,917,312	2,036,154	10.78	0.961	3,015,495	2,085,231	10.85
Std.	0.001	67,051	38,942	0.16	0.005	228,937	99,080	0.43

the upper bound on the number of neurons is set to 25. Table 9 demonstrates the parameter settings of PSO-Bagging-ANNs.

6 Model performance and discussion

6.1 Performance in baseline models

In the first stage of the analytical process, the implemented baseline models are four single models in WEKA; they are Multilayer Perception (ANNs), SMOreg (SVR), REPtree (CART) and Linear Regression (LR). Their predictive accuracies are compared in terms of R, RMSE, MAE, MAPE and SI. Table 10 presents the performance of all four baseline models. Table 11 shows that the ANNs model in WEKA outperformed the other baseline models. It yielded three of the best values of the measures of interest- R value

Table 14 Performance comparison among the best models in various schemes

Scheme	Model	Performance measure				
		R	RMSE (NTD)	MAE (NTD)	MAPE (%)	$SI_{overall}$
Baseline	ANNs	0.948	4,927,394	3,018,614	16.20	1.000 (3)
Ensemble	Bagging ANNs	0.961	3,681,032	2,518,849	13.00	0.307 (2)
Hybrid	PSO-Bagging-ANNs	**0.970**	**3,390,016**	**2,273,866**	**11.59**	**0.000 (1)**

Bold value denotes the best performance in all models; (.) indicates the overall ranking

Table 15 Comprehensive Comparison of house price prediction models in literature

Year	Author	Data description	Model	Performance measure				Exchange rate
				R^2 (%)	RMSE (NTD)	MAE (NTD)	MAPE (%)	
2004	Limsombunchai et al. (2004)	Dataset = 200	HPM	74.99	*13,070,078	–	–	1 NZD = 20.34 NTD
			NN	84.08	*9,134,927	–	–	
2011	Kontrimas & Kamarudin (2011)	Dataset = 100	OLS	–	–	*2,213,627	15.02	1 EUR = 33.92 NTD
		Max Price = 29,680,000	MLP	–	–	*3,061,469	23.30	
		Min Price = 6,071,680	SVR	–	–	*2,003,172	13.62	
2013	Núñez Tabales et al. (2013)	Dataset = 10,124	ANNs	86.47	*1,023,615	*747,409	–	1 EUR = 33.92 NTD
			HPM	79.61	*1,072,247	*769,427	–	
2014	Khamis et al. (2014)	Dataset = 1,047	MLR	64.40	*1,229,082	–	–	1 USD = 30.42 NTD
			NN	81.70	*1,093,672	–	–	
2015	Sarip & Hafez (2015)	Dataset = 350	ANNs	–	–	*1,367,650	-	1 MRY = 6.45 NTD
			FIS			*1,174,570		
			FLSR			*1,158,480		
2018	Muralidharan et al. (2018)	–	NN	–	–	–	30.65	–
			DTs	–	–	–	22.84	
2022	This study	Dataset = 13,220	PSO-Bagging-ANNs	94.09	3,390,016	2,273,866	11.59	–
		Max Price = 74,100,000						
		Min Price = 1,000,000						
		Average Price = 21,421,931						
		Std. Price = 13,319,624						

The value with * indicates the original value is converted according to the exchange rate April/09/2019

The values of MAE extracted from Sarip & Hafez (2015) are approximately recorded from their graphs

The values of RMSE extracted from Khamis et al. (2014) are calculated by their original MSE values

The values of R^2, RMSE, and MAE extracted from Núñez Tabales et al. (2013) are the average performance

This table only records the best value of RMSE from the research of Limsombunchai et al. (2004)

(0.948), MAE (3,018,614 NTD), and MAPE (16.20%)—along with the second best RMSE (4,927,394 NTD), so ANNs had the lowest SI.

6.2 Performance in ensemble models

To improve predictive performance, three ensemble models in WEKA-voting, bagging and stacking-were used. Each method aggregates the aforementioned four baseline models during the procedure. Table 11 displays the performance results that were obtained from the ensemble models in WEKA. The bagging ANNs ensemble model achieved the best results of the ensemble models in WEKA, with the best values of all evaluation measures—R value (0.961), RMSE (3,681,032 NTD), MAE (2,518,849 NTD), and MAPE (13%).

6.3 Performance of hybrid model

In the hybrid model, PSO-Bagging-ANNs, the number of bags and the number of neurons are optimized by PSO with 30 particles and a maximum of 40 iterations. Table 12 provides the optimal PSO-Bagging-ANNs parameter values. Table 13 displays the prediction performance made using PSO-Bagging-ANNs in MATLAB.

6.4 Comprehensive comparison and discussion

The best baseline, ensemble, and hybrid models were identified in this investigation. The best machine learning baseline model is ANNs, whereas the best ensemble model is bagging ANNs. The hybrid model, PSO-Bagging-ANNs, was compared with ANNs and bagging ANNs. Table 14 shows the performance of the three models in predicting house prices.

The baseline models have the advantage of being effortless to implement, requiring users only to select the models and parameters in the software package WEKA herein. However, their accuracies are lower than those in the other schemes. Consequently, the baseline models are particularly suitable for the users with very little background knowledge of machine learning. For users who understand fundamental machine learning techniques, ensemble models are better because of their superior predictive performance.

The PSO-Bagging-ANNs hybrid model that was demonstrated in this research is normally more accurate than the single and ensemble models because it has the advantages of both and the disadvantages of neither. PSO-Bagging-ANNs yielded the best results with respect to R (0.970), RMSE (3,390,016 NTD), MAE (2,273,866 NTD) and MAPE (11.59%). Table 15 comprehensively compares models that have been proposed for predicting house prices in the literature. The PSO-Bagging-ANNs hybrid model, with R (0.970) and MAPE (11.59%), outperforms those models in literature.

The proposed PSO-Bagging-ANNs model is built in the MATLAB environment, and requires a decent understanding of optimization technique and machine learning concepts for execution. Although the PSO-Bagging-ANNs model performs best in house price prediction, it is more time-consuming than the other models. Therefore, users must determine their own needs and their comprehension of machine learning to select the most appropriate technique for house appraisal.

7 Conclusion and recommendation

Real estate is one of the most essential investments in a household portfolio. House prices considerably vary across locations and individual features so forecasting prices of dwellings is significant in decision-making concerning such a portfolio. Appraisers often estimate the value of housing based on explicit market information, but ultimately assess the price subjectively, causing the appraised price inevitably to be influenced. Consequently, the hedonic price theory was proposed. The ultimate hedonic price is obtained by valuing and summing the values of the implicit characteristics of the objective. Multi-linearity or explanatory variable interactions may limit the accuracy of the hedonic price model in property valuation. Machine learning techniques are developed to solve the aforementioned problems.

Samples of dwelling transactions in Taipei City were collected from RPRS. To eliminate noise in the original dataset, the data were cleaned and preprocessed using a box plot. To select features that affect house price, filter-based algorithm, executed using ReliefFAttributeEval in the WEKA platform, was carried out to improve the predictions and reduce computational cost. The resulting samples and features were used to develop predictive models. The machine learning schemes that are used to predict the house prices are baseline, ensemble and hybrid techniques. Four machine learning baseline techniques (namely ANNs, SVR, CART and LR) were executed in their orginal form. All baseline models were then combined distinctly to generate three types of ensembles, including voting, bagging and stacking. The predictive accuracy of the hybrid model—PSO-Bagging-ANNs was compared with those of baseline/ensemble models and prior studies.

Specifically, WEKA was introduced to develop baseline and ensemble models and the PSO-Bagging-ANNs hybrid model was constructed in the MATLAB environment. The results demonstrated that the ANNs and bagging ANNs yielded the best predictive performance of any of the baseline and ensemble models, respectively. Subsequently, the ANNs and bagging ANNs were compared with the PSO-Bagging-ANNs hybrid model. The comprehensive comparison of all three models indicated that the PSO-Bagging-ANNs model outperformed the others with respect to R (0.970), RMSE (3,390,016 NTD), MAE (2,273,866 NTD) and MAPE (11.59%). A comparison of previously proposed models demonstrated that the proposed PSO-Bagging-ANNs had outstanding predictive power. This investigation comprehensively compares various popular baseline and ensemble machine learning techniques for predicting house prices, as well as a hybrid model developed in this study. To the best of our knowledge, no work has included such an extensive comparison. Users can determine their most suitable method for forecasting the values of dwellings.

This study only involves features that are available in RPRS, but several investigations have noted that views of the neighborhood and the orientation of the buildings are critical features in evaluating the price of housing. Further studies are recommended to establish a 3D model or conduct an on-site observation to quantify the aforementioned features using omputer vision via deep learning techniques, ultimately to obtain more comprehensive information for predicting house prices. This work proved the feasibility of using ANNs with a single hidden layer for predicting house prices. It is likely to achieve greater predictive accuracy with more hidden layers, which allow for what is known as deep artificial neural networks. Using a larger house price dataset than other studies, deep learning techniques are reasonably expected to reduce the likelihood of overfitting. Further works with a larger dataset and more hidden layers along with more unstructured features (e.g., text, image, voice, and video) as well as applying newly available methods (e.g., deep learning and gradient boosting decision tree-based algorithms) are needed to possibly enhance the predictive accuracy of house valuation.

Funding Funding was provided by Ministry of Science and Technology, Taiwan.

Data availability The fundamental codes and data that support the findings of this study are available from the corresponding author under reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Adetiloye, K. A., & Eke, P. O. (2014). A review of real estate valuation and optimal pricing techniques. *Asian Economic and Financial Review, 4*(12), 1878–1893.

Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization. *International Journal of Advanced Computer Science and Applications, 8*(10), 323–326. https://doi.org/10.14569/IJACSA.2017.081042

Armaghani, D. J., Raja, R. S. N. S. B., Faizi, K., & Rashid, A. S. A. (2017). Developing a hybrid PSO–ANN model for estimating the ultimate bearing capacity of rock-socketed piles. *Neural Computing and Applications, 28*(2), 391–405. https://doi.org/10.1007/s00521-015-2072-z

Bahia, I. S. H. (2013). A data mining model by using ANN for predicting real estate market: Comparative study. *International Journal of Intelligence Science, 3*(4), 162–169. https://doi.org/10.4236/ijis.2013.34017

Barzegar, R., Adamowski, J., & Moghaddam, A. A. (2016). Application of wavelet-artificial intelligence hybrid models for water quality prediction: A case study in Aji-Chay River, Iran. *Stochastic Environmental Research and Risk Assessment, 30*(7), 1797–1819. https://doi.org/10.1007/s00477-016-1213-y

Chaphalkar, N., & Sandbhor, S. (2013). Use of artificial intelligence in real property valuation. *International Journal of Engineering and Technology, 5*(3), 2334–2337.

Chau, K. W., & Chin, T. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications, 27*(2), 145–165.

Cheng, M.-Y., Prayogo, D., & Wu, Y.-W. (2019). A self-tuning least squares support vector machine for estimating the pavement rutting behavior of asphalt mixtures. *Soft Computing, 23*(17), 7755–7768. https://doi.org/10.1007/s00500-018-3400-x

Chiang, C., Han, C.-C., Chiang, Y.-M., Tsai, T.-C., Wu, F.-S., & Seng, D. (2015). Funding liquidity in the news and housing price. *Market Liquidity,* 1–25. https://doi.org/10.2139/ssrn.2565340

Chiang, Y.-H., Chuang, Y.-T., & Chang, C.-O. (2017). The impact of public bike station on residential housing price in Taipei City. *Transportation Planning Journal, 46*(4), 399–428. https://www.AiritiLibrary.com/Publication/Index/10177159-201712-201802050017-201802050017-399-428

Chou, J.-S., & Bui, D.-K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings, 82*, 437–446. https://doi.org/10.1016/j.enbuild.2014.07.036

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics, 44*, 57–75. https://doi.org/10.1016/j.ecoinf.2018.01.005

Chou, J.-S., Ngo, N.-T., & Chong, W. K. (2017). The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. *Engineering Applications of Artificial Intelligence, 65*, 471–483. https://doi.org/10.1016/j.engappai.2016.09.008

Chou, J.-S., & Tran, D.-S. (2018). Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy, 165*, 709–726. https://doi.org/10.1016/j.energy.2018.09.144

Chou, J.-S., & Truong, D.-N. (2021). Multistep energy consumption forecasting by metaheuristic optimization of time-series analysis and machine learning. *International Journal of Energy Research, 45*(3), 4581–4612. https://doi.org/10.1002/er.6125

Chou, J.-S., Truong, D.-N., Le, T.-L., & Thu Ha Truong, T. (2021a). Bio-inspired optimization of weighted-feature machine learning for strength property prediction of fiber-reinforced soil. *Expert Systems with Applications, 180*, 115042. https://doi.org/10.1016/j.eswa.2021.115042

Chou, J.-S., Truong, D.-N., & Tsai, C.-F. (2021b). Solving regression problems with intelligent machine learner for engineering informatics. *Mathematics, 9*(6), 686. https://doi.org/10.3390/math9060686

Chou, J.-S., Yang, K.-H., & Lin, J.-Y. (2016). Peak shear strength of discrete fiber-reinforced soils computed by machine learning and metaensemble methods. *Journal of Computing in Civil Engineering, 30*(6), 04016036. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000595

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. In *MIC 2015: The XI Metaheuristics International Conference in Agadir, Morocco,* pp. 1-5. Retrieved September 7, 2021, from https://arxiv.org/abs/1502.02127

Dawson, R. (2011). How significant is a boxplot outlier? *Journal of Statistics Education, 19*(2), 1–13. https://doi.org/10.1080/10691898.2011.11889610

Delmendo, L. C. (2021). Taiwan's house prices surging, amidst strong economic growth. Retrieved September 7, 2021 from https://www.globalpropertyguide.com/Asia/Taiwan/Price-History

Dey, A., Miyani, G., & Sil, A. (2019). Application of artificial neural network (ANN) for estimating reliable service life of reinforced concrete (RC) structure bookkeeping factors responsible for deterioration mechanism. *Soft Computing, 24*(3), 2109-2123. https://doi.org/10.1007/s00500-019-04042-y

Do, H., Silverman, H. F., & Yu, Y. (2007). A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07,* pp. I-121–I-124. https://doi.org/10.1109/ICASSP.2007.366631

Du, Y.-S., Song, F.-C., Zeng, Y.-S., Ge, J.-N., & Chen, F.-Y. (2013). Retrospective analysis of hedonic price model in Taiwan. *Quarterly Research on Land Issues, 12*(2), 44–57.

Elbeltagi, E., Hegazy, T., & Grierson, D. (2005). Comparison among five evolutionary-based optimization algorithms. *Advanced Engineering Informatics, 19*(1), 43–53. https://doi.org/10.1016/j.aei.2005.01.004

Erdal, H. I., & Karakurt, O. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology, 477*, 119–128. https://doi.org/10.1016/j.jhydrol.2012.11.015

Fallahi, A., & Jafari, S. (2011). An expert system for detection of breast cancer using data preprocessing and bayesian network. *International Journal of Advanced Science and Technology, 34*, 65–70.

Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. In *ICMLC 2018: Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 6–10). https://doi.org/10.1145/3195106.3195133

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43*(12), 2301–2315. https://doi.org/10.1080/00420980600990928

Fumo, N., & Rafe Biswas, M. A. (2015). Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews, 47*, 332–343. https://doi.org/10.1016/j.rser.2015.03.035

Geng, B., Bao, H., & Liang, Y. (2015). A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model. *Habitat International, 49*, 333–339. https://doi.org/10.1016/j.habitatint.2015.06.005

Ghasemi, M., Akbari, E., Rahimnejad, A., Razavi, S. E., Ghavidel, S., & Li, L. (2019). Phasor particle swarm optimization: A simple and efficient variant of PSO. *Soft Computing, 23*(19), 9701–9718. https://doi.org/10.1007/s00500-018-3536-8

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hammouche, K., Diaf, M., & Siarry, P. (2010). A comparative study of various meta-heuristic techniques applied to the multilevel thresholding problem. *Engineering Applications of Artificial Intelligence, 23*(5), 676–688. https://doi.org/10.1016/j.engappai.2009.09.011

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics, 2015*, 198363. https://doi.org/10.1155/2015/198363

Huang, Y., McCullagh, P. J., & Black, N. D. (2009). An optimization of ReliefF for classification in large datasets. *Data & Knowledge Engineering, 68*(11), 1348–1356. https://doi.org/10.1016/j.datak.2009.07.011

Huang, Y.-J., Chiang, Y.-H., & Chang, C.-O. (2017a). Impact of public housing on nearby residential property values in Taipei city. *Journal of City and Planning, 44*(3), 277–302. https://doi.org/10.6128/CP.44.3.277

Huang, Z., Chen, R., Xu, D., & Zhou, W. (2017b). Spatial and hedonic analysis of housing prices in Shanghai. *Habitat International, 67*, 69–78. https://doi.org/10.1016/j.habitatint.2017.07.002

Isa, I., Saad, Z., Omar, S., Osman, M., Ahmad, K., & Sakim, H. M. (2010). Suitable MLP network activation functions for breast cancer and thyroid disease detection. In *2010 second international conference on computational intelligence, modelling and simulation* (pp. 39–44). IEEE. https://doi.org/10.1109/CIMSiM.2010.93

Job, F., Mathew, D. S., Meyer, D. T., & Narbey, S. (2021). An investigation on the experimental analysis and MATLAB simulation for dye-sensitized solar cell. *Materials Today: Proceedings,* 1–7. https://doi.org/10.1016/j.matpr.2021.07.225

Kamaruddin, S., & Ravi, V. (2016). Credit card fraud detection using big data analytics: Use of PSOANN based one-class classification. In *Proceedings of the International Conference on Informatics and Analytics* (pp. 1–8). ACM. https://doi.org/10.1145/2980258.2980319

Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems, 1*(4), 111–122.

Kaur, H., & Salaria, D. S. (2013). Bayesian regularization based neural network tool for software effort estimation. *Global Journal of Computer Science and Technology, 13*(2), 45–50.

Kazienko, P., Lughofer, E., & Trawiński, B. (2013). Hybrid and ensemble methods in machine learning J. UCS special issue. *Journal of Universal Computer Science, 19*(4), 457–461.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95— International Conference on Neural Networks, 4*, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Khamis, A. B., & Kamarudin, N. K. K. B. (2014). Comparative study on estimate house price using statistical and neural network model. *International Journal of Scientific & Technology Research, 3*(12), 126–131.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence, 2,* 1137–11452.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing, 11*(1), 443–448. https://doi.org/10.1016/j.asoc.2009.12.003

Kouwenberg, R., & Zwinkels, R. (2014). Forecasting the US housing market. *International Journal of Forecasting, 30*(3), 415–425. https://doi.org/10.1016/j.ijforecast.2013.12.010

Li, J., Liu, X., Liu, J., & Li, W. (2016). City profile: Taipei. *Cities, 55*, 1–8. https://doi.org/10.1016/j.cities.2016.03.007

Limsombunchai, V., Gan, C., & Lee, M. (2004). House price prediction: hedonic price model vs. artificial neural network. *American Journal of Applied Sciences, 1*(3), 193-201. https://doi.org/10.3844/ajassp.2004.193.201

Lin, S.-J. (2004). The marginal willingness-to-pay of star public elementary and junior high school districts in Taipei City. *Journal of Housing Studies*, *13*(1), 15–34. https://doi.org/10.6375/JHS.200406.0015

Liu, R., & Liu, L. (2019). Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm. *Soft Computing, 23*(22), 11829-11839. https://doi.org/10.1007/s00500-018-03739-w

Merlini, D., & Rossini, M. (2021). Text categorization with WEKA: A survey. *Machine Learning with Applications, 4*, 100033. https://doi.org/10.1016/j.mlwa.2021.100033

Methaprayoon, K., Yingvivatanapong, C., Lee, W.-J., & Liao, J. R. (2007). An integration of ANN wind power estimation into unit commitment considering the forecasting uncertainty. *IEEE Transactions on Industry Applications, 43*(6), 1441–1448. https://doi.org/10.1109/TIA.2007.908203

Muralidharan, S., Phiri, K., Sinha, S. K., & Kim, B. (2018). Analysis and prediction of real estate prices: A case of the Boston housing market. *Issues in Information Systems, 19*(2), 109–118. https://doi.org/10.48009/2_iis_2018_109-118

Núñez-Tabales, J., Rey Carmona, F., & Caridad, J. (2013). Implicit prices in urban real estate valuation. *Revista de la Construcción, 12*(2), 116–126. https://doi.org/10.4067/S0718-915X2013000200009

Palma-Mendoza, R.-J., Rodriguez, D., & De-Marcos, L. (2018). Distributed ReliefF-based feature selection in Spark. *Knowledge and Information Systems*, *57*(1), 1–20. https://doi.org/10.1007/s10115-017-1145-y

Peter, S. E., & Raglend, I. J. (2017). Sequential wavelet-ANN with embedded ANN-PSO hybrid electricity price forecasting model for Indian energy exchange. *Neural Computing and Applications, 28*(8), 2277–2292. https://doi.org/10.1007/s00521-015-2141-3

Potter, K. (2006). Methods for presenting statistical information: The box plot. *Visualization of Large and Unstructured Data Sets, GI-Edition Lecture Notes in Informatics (LNI), S-4*, 97–106.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence, 26*(1), 25–33. https://doi.org/10.1007/s10489-006-0001-7

Rong, L. H., & Sun, Y. M. (2014). The analysis of second-hand housing price influencing factors based on hedonic model and WEB information. *Applied Mechanics and Materials, 587–589,* 2285–2289. https://doi.org/10.4028/www.scientific.net/AMM.587-589.2285

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169

Sarip, A. G., & Hafez, M. B. (2015). Fuzzy logic application for house price prediction. *International Journal of Property Sciences*, *5*(1), 24–30. https://doi.org/10.22452/ijps.vol5no1.3

Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis, 47*(1), 165–174. https://doi.org/10.1016/j.csda.2003.10.012

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: introduction and review. *Journal of biomedical informatics, 85,* 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer. https://doi.org/10.1007/978-1-4757-2440-0

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1936–1939). IEEE. https://doi.org/10.1109/ICICCT.2018.8473231

Wang, L. V., & Yao, G. (2001). Ultrasound-modulated laser tomography. In *Saratov Fall Meeting 2000: Optical Technologies in Biophysics and Medicine II*, *4241*, pp. 1–5. Retrieved September 7, 2021, from https://doi.org/10.1117/12.431526

Wang, X. (2011). The application of SPSS in empirical research of housing hedonic price. In *2011 International Conference on Multimedia Technology* (pp. 3262–3265). IEEE. https://doi.org/10.1109/ICMT.2011.6003072

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik, 125*(3), 1439–1443. https://doi.org/10.1016/j.ijleo.2013.09.017

Wang, Y., Chen, P.-C., Ma, H.-W., Cheng, K.-L., & Chang, C.-Y. (2018). Socio-economic metabolism of urban construction materials: A case study of the Taipei metropolitan area. *Resources, Conservation and Recycling, 128*, 563–571. https://doi.org/10.1016/j.resconrec.2016.08.019

Wei, W., Guang-ji, T., & Hong-rui, Z. (2010). Empirical analysis on the housing price in Harbin City based on hedonic model. In *2010 International Conference on Management Science & Engineering 17th Annual Conference Proceedings* (pp. 1659–1664). IEEE. https://doi.org/10.1109/ICMSE.2010.5720005

Wen, H.-Z., Sheng-hua, J., & Xiao-yu, G. (2005). Hedonic price analysis of urban housing: An empirical research on Hangzhou, China. *Journal of Zhejiang University-Science A, 6*(8), 907–914. https://doi.org/10.1631/jzus.2005.A0907

Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: A simple visual method to interpret data. *Annals of Internal Medicine, 110*(11), 916–921. https://doi.org/10.7326/0003-4819-110-11-916

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wu, J., Wang, M., Li, W., Peng, J., & Huang, L. (2015). Impact of urban green space on residential housing prices: Case study in Shenzhen. *Journal of Urban Planning and Development, 141*(4), 05014023. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000241

Wu, J. Y. (2017). *Housing price prediction using support vector regression* (pp. 1–56). San José State University. https://doi.org/10.31979/etd.vpub-6bgs

Xiao, Y. (2017). *Urban morphology and housing market*. Springer. https://doi.org/10.1007/978-981-10-2762-8

Yang, C.-H. & Su, S.-H. (2011). The impacts of housing price in YIMBY and NIMBY. *Journal of Housing Studies*, *20*(2), 61–80. https://doi.org/10.6375/JHS.201112.0062

Yu, H., & Xiaohui, W. (2011). PSO-based energy-balanced double cluster-heads clustering routing for wireless sensor networks. *Procedia Engineering, 15*, 3073–3077. https://doi.org/10.1016/j.proeng.2011.08.576

Zhang, H., & Zhang, M. (2010). Environment hedonic price analysis: Evidence from Jilin city. In *2010 Second International Conference on Communication Systems, Networks and Applications* (pp. 354–356). IEEE. https://doi.org/10.1109/ICCSNA.2010.5588741

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.