Hindawi
Journal of Mathematics
Volume 2022, Article ID 2068353, 16 pages
https://doi.org/10.1155/2022/2068353

Research Article

A Study on the Prediction of House Price Index in First-Tier Cities in China Based on Heterogeneous Integrated Learning Model

Yaqi Mao,[1] **Yonghui Duan**,[1] **Yibin Guo**,[2] **Xiang Wang**,[2] **and Shen Gao**[1]

[1]*Department of Civil Engineering, Henan University of Technology, No. 100, Lianhua Street, Gaoxin District, Zhengzhou 450001, China*
[2]*Department of Civil Engineering, Zhengzhou University of Aeronautics, No. 15, Wenyuan West Road, Zhengdong New District, Zhengzhou 450015, China*

Correspondence should be addressed to Yaqi Mao; 728115479@qq.com

Received 23 May 2022; Revised 30 August 2022; Accepted 8 September 2022; Published 23 September 2022

Academic Editor: Ghous Ali

To address the difficulty of low prediction accuracy, insufficient model stability, and certain lag associated with a single machine learning model in the prediction of house price, this paper proposes a multimodel fusion house price prediction model based on stacking integrated learning. Firstly, web search data affecting house prices were collected by web crawler technology, and Spearman correlation analysis was performed on the attribute set to reduce its complexity and establish a prediction index system for four first-tier cities in China. Secondly, with the goal of improving accuracy, diversity, and generalization ability, the types of base learners as well as metalearners are determined, and the parameters of the base learners are optimized using the grey wolf optimization algorithm to produce the GWO-stacking model, and the experimental results from four datasets demonstrate that the model has high prediction accuracy. Finally, to solve the issue of unintelligible black boxes in machine learning models, we have used the state-of-the-art interpretation method SHAP combined with the LightGBM algorithm to interpret the model, and the result can be used as a basis for real estate policy planning and adjustment and even guide the demand of home buyers, thus improving the efficiency and effectiveness of government policy making.

1. Introduction

Since the reform of the Chinese housing system, the real estate market has experienced rapid growth and has contributed to the growth of the national economy. The impact of housing prices on economic development and people's lives has become increasingly apparent in recent years, and the trend of housing prices has become a focus of attention for all sectors of society. The higher the ranking of the city, the higher the level of economic development, the earlier the timing of the development of the real estate market, the greater the social influence, and the changes in the real estate market in first-tier cities serve as a model and guide for the future price of housing in other cities across the country. Currently, the National Bureau of Statistics (NBS) produces statistics on real estate price trends in the form of price indices, but the release of this data has a lag, so it is imperative to develop an efficient forecasting model to ensure timely and accurate forecasts for price indices in first-tier cities.

Various forecasting methods and forecasting models have been constructed by domestic and foreign scholars to address the problem of real estate price trend forecasting. We have found that the existing forecasting methods can be broadly classified into three categories based on their categories and mechanisms: hedonic model method, econometric methods, and machine learning models. Rosen [1] was the first to conduct the research in this area, who applied the consumer theory proposed by Lancaster [2] to the housing market and devised the hedonic model, which was then widely used as a real estate valuation tool. Chambers [3] investigated the influence of such characteristics as ethnicity

and crime rate on residential home prices. Chica-Olmo [4] assessed the influence of neighbourhood and locational factors on housing prices. Wheeler et al. [5] studied the predictive power of different functional forms in the hedonic model. Noor et al. [6] summarized the research progress in optimizing real estate valuation using big data techniques in hedonic model based on 124 studies. The hedonic model uses the residential transaction price as the dependent variable and regresses it on a set of characteristic variables [7], which allows estimating the implied price of residential housing from an economic perspective [8]. However, hedonic model also has several limitations in terms of the underlying assumptions and estimation, including the choice of the form of the model regression function and the choice of the independent variables. Moreover, since hedonic model is composed of regression, whereas house price forecasting is a nonlinear problem, hedonic model is used more for explaining the degree of influence of independent variables on house prices than for forecasting house prices. The econometric approach uses historical house prices as time series data, while building models based upon their historical prices and making forecasts for the future. Guirguis et al. [9] used the generalized autoregressive conditional heteroskedasticity (GARCH) model to forecast US house prices, with experimental results showing that the GARCH model has high forecasting accuracy in the selected out-of-sample forecasts. A study by Miles [10] used generalized autoregressive (GAR) for house price forecasting and found their proposed model outperformed both the autoregressive moving average (ARMA) and the GARCH models. The study by Zhao et al. [11] reported that the ARIMA model outperformed the comparison model when it came to forecasting New Zealand house prices. However, all the above models use property price data as a longitudinal time series for linear forecasting, which makes it challenging to capture the nonlinearity of the data, resulting in large forecasting errors [12]. Machine learning models are the current trend in house price forecasting. The machine learning models, in comparison to the first two methods, have better self-learning ability, possess a capacity to dig deeper into the data, retain valuable information [13], and are also capable of better nonlinear prediction, so they have become the mainstream method for the prediction of house prices at present. Kauko et al. [14] studied the application of neural network models to the housing market in Helsinki, Finland, and found that neural network models can identify the dimensions of home market formation based on patterns found within the dataset. Fan et al. [15] have proposed a variety of tree-based methods for evaluating the relationship between house prices and housing characteristics. Li et al. [16] developed a model for predicting real estate market prices by using rough sets and wavelet neural networks. Selim [17] developed an artificial neural network (ANN) model for predicting house prices in Turkey, and the experimental results indicated that the ANN model outperformed the comparison model. Dong et al. [18] employed various machine learning methods such as decision tree (DT), random forest (RF), and support vector machine (SVM) to predict the second-hand house price index in 16 cities in China, with the results showing that SVM performed the best. Shah et al. [19] conducted a predictive study of rental prices of apartments in Brazil using various regression models such as AdaBoost, RF, and multilayer perceptron (MLP) based on the factors affecting rental prices. Compared to the hedonic model and the econometric model, the machine learning model has significantly higher prediction accuracy. However, proper parameter selection and setting are critical for the accuracy of the prediction results, and inaccurate parameter settings can severely impact the result [20]. Scholars have also conducted research on this topic. Gu et al. [21] used SVM to establish a house price prediction model. They applied genetic algorithm (GA) to solve the problem of parameter selection for the SVM model, and the model was shown to have a better predictive effect through experimental analysis. Fei and Mingyan [22] employed back propagation neural networks (BPNN) to predict second-hand house prices in Qingdao city and optimize their parameters through modifying lion swarm optimization (LSO). Fang [23] conducted a study on forecasting the price of foreclosed houses in a Chinese province using BP neural network and optimized the parameters of the BP model using GA algorithm in the modeling process. Recently, integrated learning models have been widely used in various fields because of their unique learning approach [24–26]. In the house price prediction problem, Zhu and Li [27] utilized the gradient boosting decision tree (GBDT) model to forecast the price of second-hand houses in China and the particle swarm optimization (PSO) algorithm to determine the model's hyperparameters. Alfaro-Navarro et al. [28] used a plurality of integrated learning algorithms to forecast Spanish house prices. Wang et al. [29] proposed a WOA-SVR model based on Bagging integrated learning approach to forecast house price indices of four Chinese municipalities derived from macroeconomic data.

Machine learning methods have the advantage of being able to be trained using historical data. The prediction effect is determined not only by the performance of the prediction model but also by the selection of valid prediction data. Web search data (WSD) provides a new research idea for addressing prediction problems in the context of big data. The first to propose this study was Ginsberg et al. [30], who used Google's massive user search data to accurately predict the trend of the proportion of influenza-like cases in the United States one week in advance in the "Google Flu Trends" software developed by Google in 2008. Since then, WSD has been widely used in major research fields such as economics and medicine [31–33]. In the house price prediction problem, a study by Wu and Brynjolfsson [34] found that the Google Home search index has good predictive power for real estate market sales and prices. Beracha and Wintoki [35] suggested that the extent of anomalous real estate searches in a particular city may be indicative of future anomalous house price changes in that city. A study by Rizun and Baj-rogowska [36] stated that the primary source of information on predicting house price trends is public government reports and that the data is released with a lag, whereas the use of Google Trends can predict future price changes in advance.

To conclude, numerous attempts have been made in the literature to address the issue of house price forecasting; however, some shortcomings remain. Firstly, the characteristic price models and econometric models are primarily linear forecasts of historical house price data, which make it very difficult to deal with nonlinear features of the data, and their interpretation is limited. Secondly, although machine learning models can effectively extract nonlinear data features from house price data, a single machine learning model tends to have its own limitations, such as low prediction accuracy and poor generalization capabilities. Thirdly, the integrated learning model can integrate the prediction results of multiple base learners to achieve secondary learning of the prediction problem, which can effectively reduce the prediction error, but at present, in house price prediction, most studies use one integrated learning strategy for prediction and fail to further discuss the effect of combining different machine learning models on the house price prediction effect. In addition, much of the data used in existing literature are traditional statistical data, which have deficiencies such as a lengthy acquisition period and limited timeliness, affecting the timeliness of the prediction model.

As a means of addressing these deficiencies, this paper focuses on the following three aspects in the house price forecasting problem.

To begin with, using Baidu recommendations and references to related literature, the initial word bank of Internet keywords related to house price indexes was created by combining information on eight aspects, including macro regulation, financial policy, tax policy, protection policy, land policy, house price expectation, transaction details, and residential characteristics.

Next, the stacking integration method is utilized for model fusion, based on the excellent performance of the integrated learning method, and an integrated learning regression prediction model based on multimodel fusion is developed. And the grey wolf optimizer (GWO) is used to optimize the parameters of the base learner to prevent overfitting of the metalearner and achieve the goal of improving generalization ability and prediction accuracy of the prediction model.

A third point is that the established datasets for the four cities are input into the model developed in this paper for prediction research, and multiple benchmark models and performance evaluation indexes are set up in order to engage in a more scientific and comprehensive evaluation of the proposed model, and the SHAP method is applied to analyse the keyword features of the four cities in order to aid in the interpretation of the machine learning model.

2. Methodology

2.1. XGBoost Model. XGBoost is a boosting class model developed by Chen and Guestrin [37] in 2016, which is an extension and improvement on the GBDT algorithm. GBDT uses the negative gradient of the loss function as an approximation of the current round of losses and uses it as the optimization objective for the computation [38]. While the traditional GBDT method uses only the first-order derivatives, XGBoost uses a second-order Taylor expansion of the loss function, and to account for the decline of the objective function and the complexity of the model, a regular term is added alongside the objective function to determine the optimal solution overall, avoiding overfitting. In recent years, XGBoost models have demonstrated superior performance in predicting biological, medical, and economic problems. The mathematical principle of the model can be summarized as follows.

Here is the integration model of the definition tree.

$$\widehat{y_i} = \sum_{m=1}^{M} f_m(x_i), f_m \in F. \tag{1}$$

In this equation, $\widehat{y_i}$ is the prediction value, M is the number of trees, F is the range of tree selections, and x_i is the ith input feature.

The loss function for the XGBoost model is shown as follows:

$$L = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{m=1}^{M} \theta(f_m). \tag{2}$$

Here, the first part of the function is the error between the predicted and the actual training values of the XGBoost model, while the second is used to represent the complexity of the tree, which is important when controlling the regularization of the complexity of the model.

$$\theta(f) = \gamma T + \frac{1}{2}\tau\|\omega\|^2. \tag{3}$$

Here, γ and τ represent penalty factors.

It is minimized by adding the incremental function $f_t(x_i)$ to equation (2) to minimize the value of the loss function. Thus, the objective function for the tth time becomes as follows:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{m=1}^{M} \theta(f_m) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{t-1} + f_t(x_i)\right) + \theta(f_t). \tag{4}$$

A second-order Taylor expansion of equation (4) is used to approximate the objective function at this point. Define the set of samples in each subleaf of the jth tree as $I_j = \{i|q(x_i = j)\}$. At this point, we can approximate $L^{(t)}$ as follows:

$$\begin{aligned} L^{(t)} &\cong \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \theta(f_t) \\ &\cong \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\tau\omega^2 \\ &\cong \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \tau\right)\omega_j^2\right] + \gamma T. \end{aligned} \tag{5}$$

Here, $g_i = \partial_{\widehat{y}_i^{t-1}} l(y_i, \widehat{y}_i^{t-1})$ is the first-order derivative of the loss function; $h_i = \partial^2_{\widehat{y}_i^{t-1}} l(y_i, \widehat{y}_i^{t-1})$ is the second-order derivative. The following equation is calculated by defining $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$.

$$L^{(t)} \cong \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2} \left(H_j + \tau \right) \omega_j^2 \right] + \gamma T. \tag{6}$$

The following equation is obtained by taking partial derivatives of ω:

$$\omega_j = -\frac{G_j}{H_j + \tau}. \tag{7}$$

The following equation can be obtained by substituting the weights into the objective function:

$$L^{(t)} \cong -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^{\,2}}{H_j + \tau} + \gamma T. \tag{8}$$

2.2. LightGBM Model. Although the correlation algorithm in XGBoost can reduce the computational effort required to find the optimal segmentation point, it still requires traversal of the dataset. XGBoost faces significant challenges in terms of efficiency as the volume of data continues to increase in the digital era. Microsoft has developed LightGBM [39] which is an open source, efficient gradient boosting framework model based on decision trees. The LightGBM model is also capable of parallel learning, similar to the XGBoost model. Two of the key improvements of LightGBM are the Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) algorithms, which are substantially faster to train than the XGBoost model and use significantly less memory [40].

The GOSS algorithm is a sampling algorithm. In terms of sample reduction, the method utilizes the gradient information of each sample for sampling, keeping samples with larger gradients and selecting random samples with smaller gradients. Additionally, weights are added to the small gradient samples to counteract the effect of sampling on sample distributions. The GOSS algorithm has the following general computation steps: firstly, the feature data to be split in the decision tree model are sorted according to their absolute values; secondly, the first $a * 100\%$ samples with greater absolute values are removed; thirdly, $b * 100\%$ of the remaining small gradient data are selected randomly, and the extracted data are multiplied by the weight value $(1 - a)/b$ in order to give more weight to the untrained samples without too much change to the distribution of the original dataset; finally, $(a + b) * 100\%$ of the data is used to calculate the information gain. Data distribution in a high-dimensional environment is typically sparse, and the EFB method is a lossless approach that uses a feature bundling strategy to reduce the dimensionality of the dataset. During feature processing, to avoid losing information related to features after bundling, the features to be bundled are rarely nonzero values at the same time; i.e., the features must be mutually exclusive. If the conflict ratio of two nonfully mutually exclusive features is low, they can be considered mutually exclusive for the purpose of bundling. An essential idea behind the EFB algorithm is to optimize computation efficiency by reducing the number of training features without affecting prediction accuracy by bundling and combining multiple mutually exclusive features into feature packages.

2.3. Grey Wolf Optimizer. By simulating the predatory behaviour of grey wolf packs, Mirjalili et al. [41] proposed a pack intelligence optimization algorithm, the grey wolf optimizer (GWO), in 2014. The GWO optimization process is carried out by the α, β, and δ wolves, the highest social strata in each generation of the population, who lead the bottom ω wolves by hunting, surrounding, and attacking their prey. GWO has been used to solve optimization problems in many fields due to its simple structure, few adjusting parameters, and easy implementation. It is this algorithm that is used to find the optimal parameters of the base learner model in this paper. The following is a mathematical description of the algorithm.

To begin, we can describe mathematically the process by which a wolf pack searches for and slowly surrounds its prey.

$$\begin{aligned} D &= \left| C \cdot X_p(t) - X(t) \right|, \\ X(t+1) &= X_p(t) - A \cdot D, \\ a &= 2 - \frac{2I}{M}, \\ A &= 2a \cdot r_1 - a, \\ C &= 2 \cdot r_2. \end{aligned} \tag{9}$$

Here, $X(t)$ is the position of the prey after the tth iteration; $X_p(t)$ is the position of the grey wolf at the tth iteration; D denotes the distance between the grey wolf and the prey; $X(t+1)$ denotes the update of the position of the grey wolf; A and C are the coefficient vectors; a is the convergence factor whose value decreases linearly from 2 to 0 with the number of iterations, I is the number of previous iterations, and M is the maximum number of iterations; r_1 and r_2 are the random numbers between $[0, 1]$.

Secondly, the position of the three optimal wolves α, β, and δ is constantly updated to determine the prey. The following is a mathematical description of the hunting process of a wolf pack:

$$\begin{aligned} D_a &= |C_1 \cdot X_\alpha(t) - X(t)|, \\ D_\beta &= \left| C_2 \cdot X_\beta(t) - X(t) \right|, \\ D_\delta &= |C_3 \cdot X_\delta(t) - X(t)|, \\ X_1(t+1) &= X_\alpha(t) - A_1 \cdot D_\alpha, \\ X_2(t+1) &= X_\beta(t) - A_2 \cdot D_\beta, \\ X_3(t+1) &= X_\delta(t) - A_3 \cdot D_\delta, \\ X(t+1) &= \frac{X_1(t+1) + X_2(t+1) + X_3(t+1)}{3}. \end{aligned} \tag{10}$$

Here, $X_\alpha(t)$, $X_\beta(t)$, and $X_\delta(t)$ are the positions of α, β, and δ wolves when the population is iterated to the tth generation; $X(t)$ is the position of individual grey wolves in the t th generation; A_1 and C_1, A_2 and C_2, and A_3 and C_3 are the coefficient vectors of α, β, and δ wolves, respectively; $X_1(t+1)$, $X_2(t+1)$, and $X_3(t+1)$ indicate the positions of α, β, and δ wolves after $(t+1)$ iterations, respectively; $X(t+1)$ is the position of the next generation of grey wolves. Figure 1 illustrates the flow chart of the GWO algorithm.

2.4. Stacking Integration Learning. A single prediction model, in general, shows a decreasing marginal utility as its prediction accuracy increases. Ensemble learning (EL) [42] is a multialgorithm fusion machine learning method that uses statistical learning theory. Stacking [43] integrated learning provides greater predictive performance by combining different machine learning algorithms together and utilizing the strengths of each algorithm. Stacking stacks multiple algorithms in different layers [44], and its number of layers can be freely set, but from the research and applications in various fields, the general two-layer structure of stacking can strengthen the learning effect without causing the model to be too complex. By applying K-fold cross-validation to the original dataset, this integration idea divides it into subsets, which are then input to each base learner of the layer 1 prediction model, and each base learner then generates its own prediction result. Following this, the output of layer 1 is then used as input to layer 2 to train the metalearner of layer 2's prediction model, and the final prediction results are derived from the model located in layer 2. As illustrated in Figure 2, the stacking learning framework generalizes the output of multiple models to improve prediction accuracy as a whole.

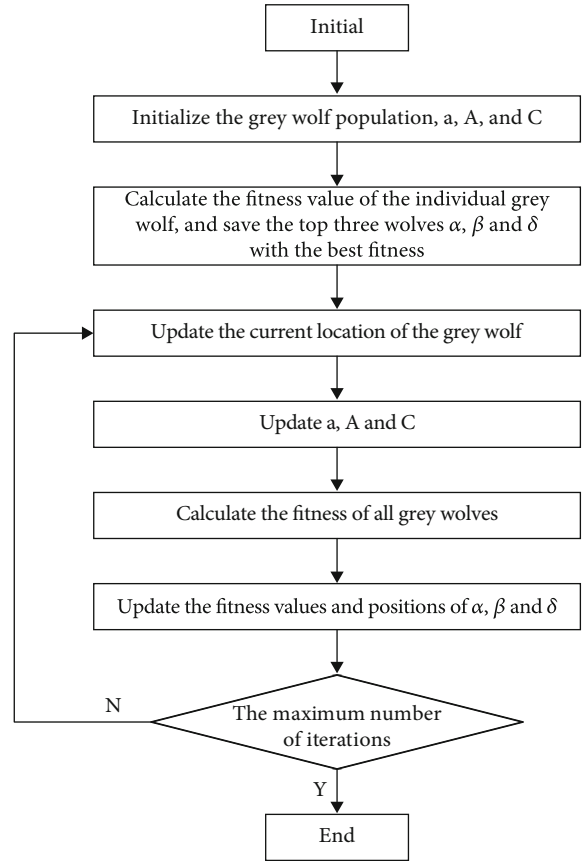

Figure 1: Flow chart of the grey wolf optimization.

3. Residential Price Prediction Algorithm Using Fusion Stacking Integrated Learning

Stacking integrated learning coupled with multivariate learners to predict residential prices is essentially a regression model that uses historical data such as changes in house prices as input features and future house prices as output of the prediction.

3.1. Selection of Base Learners and Metalearners. The choice of the right base learner is crucial to the forecasting process because it allows data mining for factors affecting house prices from a variety of spatial and structural views. Additionally, it will address problems arising from unbalanced high-dimensional data categories and the tendency for models to overfit to achieve complementary advantages between different learners as well as improve their adaptability. And the selection process for individual metalearners is more inclined to optimize the overall regression process than that for base learners.

To construct the stacking integration model, we must first identify the type of learners that will be used as base learners, and the type of learners should be selected regarding both accuracy and diversity. The improvement of the overall prediction of the model is aided by choosing a base learner with a higher level of ability to learn. The use of boosting in recent years has developed rapidly as a method for reducing model bias. AdaBoost, GBDT, and others are representative algorithms. Among them, gradient boosting-based tree models are very popular because of their excellent performance. Considering the prediction performance of the model, two algorithms, XGBoost and LightGBM, that improve GBDT from a set of different perspectives, are selected as the base learners in this paper. On the other hand, the purpose of using different algorithms as base learners is to explore the relationships existing between historical house price data from different spatial and structural perspectives. And combine the principles of each algorithm to build different prediction models, to make use of the advantages of different machine learning algorithms and make each model complement each other's strengths. Support vector machine (SVM) [45] is a typical machine learning method based on statistical theory, with a solid theoretical foundation and strong generalization capability. Support vector regression (SVR) model is a prediction algorithm for SVM for regression modeling, which shows better learning performance for solving regression problems with small samples, nonlinearity, and high dimensionality [46]. MLP [47], as a typical representative of neural networks, has very good nonlinear mapping capability, high parallelism, and global optimization and nowadays has achieved good success in image processing, regression prediction,

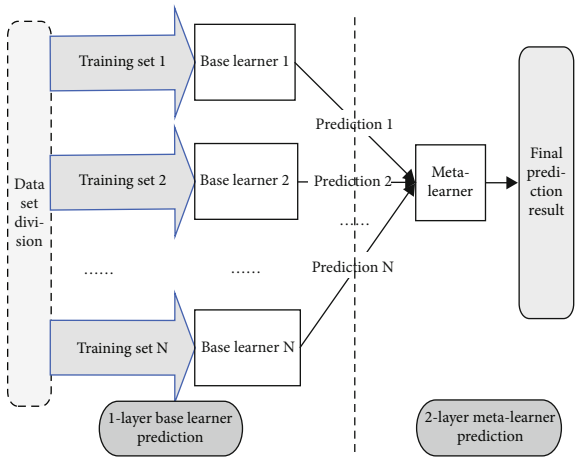

Figure 2: Ensemble learning method based on stacking.

and pattern recognition. As a result, to obtain the best prediction, SVR and MLP, which have the largest degree of difference from XGBoost and LightGBM, are selected as the first layer of stacking's base learners. Additionally, since the predictions of the base learners differ and each has its own characteristics, selecting a metalearner can be used to improve the bias of each learner while ensuring generalization capability, thus preventing the overfitting problem. We use multiresponse linear regression (MLR) [48] as the metalearner for the second layer of stacking to optimize the prediction of the final integrated stacking. MLR is characterized by high generalization capacity and no need to adjust parameters, which allows it to effectively correct the bias of multiple learning algorithms in the first layer of the training set.

3.2. Cross-Validation of 5-Fold. An output generated by the base learner is used as the training set for the metalearner. To reduce the risk of overfitting, it is necessary to divide the data usage process in an appropriate manner so that the data is not repeatedly learned by the two-layer learner. It is first necessary to divide the original house price training dataset into five subdatasets according to time and ensure that none of the datasets overlap with each other. In a single base learner, one block of data serves as the test set, and the remaining four blocks serve as the training set. As each base learner produces a prediction for its own test set, the five results are eventually combined into a metatraining set that is the same size as the original training set. In Figure 3, the detailed process of generating the metatraining set and metatest set can be seen. The labels of the original training set and the original test set are not changed; only the feature matrix is altered from the original house price index for each keyword influencing factor to the prediction results of each base learner for these factors. Despite the advantages of machine learning algorithms from both a theoretical and an application standpoint, it is important to note that proper hyperparameter settings are necessary to realize these advantages. In the stacking integration strategy, the selection of the hyperparameters of the base learner effect the overall learning effect and prediction performance of the model, which is a research hotspot and a challenge for the base learner. In existing studies, the selection of hyperparameters for the base learner is usually achieved by utilizing cross-validation [49] and grid search methods [50, 51]; however, when there are many parameters or a wide range of parameter values, these methods generate a large amount of computation and reduce the efficiency of the model training [52]. Swarm intelligence is an iterative search algorithm with the following advantages: flexibility, global search, self-organization, and capability for parallel processing. In this paper, we select the hyperparameters of the classical GWO optimization stacking algorithm in the swarm intelligence algorithm that will serve as the base learner.

3.3. Residential Price Prediction Algorithm Process Based on Multivariate Learners Fused with Stacking Integrated Learning. The stacking integrated house price forecasting model in this paper is composed of the following seven steps.

Step 1. Divide the dataset into a training set Y_{train} and a test set Y_{test}, and divide Y_{train} into five subsets of equal size Y_1, Y_2,…,Y_5 using a 5-fold cross-validation.

Step 2. Fit the XGBoost model using the training set Y_{train}, while using the GWO algorithm to find the hyperparameters of the model.

Step 3. For the 5 subsets in Step 1, one subset Y_i is chosen as the subtest set in order and the remaining 4 subsets $Y_{-i} = Y_{\text{train}} - Y_i$ as the subtraining set.

Step 4. Fit the XGBoost model using the subtraining set Y_{-i}, use the GWO algorithm to find the hyperparameters of XGBoost in the fitting process, and use the fitted model to predict the subtest set Y_i to get the prediction result α_i, while use the fitted model to predict the test set Y_{test} to get the prediction result β_i.

Step 5. The prediction results $\{\alpha_1, \alpha_2, \cdots, \alpha_5\}$ and $\{\beta_1, \beta_2, \cdots, \beta_5\}$ are obtained by repeating Step 4 five times, where the 5 predictions in α_i are combined to obtain vector $A1$ with the same length as the training set Y_{train}. The prediction results β_i of the 5 test sets Y_{test} are weighted and averaged to obtain vector $B1$ with the same length as the test set Y_{test}.

Step 6. Performing Step 2 to Step 5 above for LightGBM, MLP, and SVR models to obtain $A2$, $A3$, and $A4$ and $B2$, $B3$, and $B4$, respectively.

Step 7. The dataset train = $\{A1, A2, A3, A4, y\}$ is obtained by combining $A1$ to $A4$ with the label y of Y_{train}. After obtaining the dataset train, the two-layer metalearner MLR model is fitted using train. At the same time, the dataset test = $\{B1, B2, B3, B4\}$ is input to the fitted metalearner as a new test set for prediction, and the obtained prediction result is the final prediction result of stacking integrated learning.

Figure 3: Schematic diagram of base model generation metadataset.

4. Empirical Analyses

4.1. Data Source. The paper focuses on four Chinese first-tier cities, Beijing, Shanghai, Guangzhou, and Shenzhen, because they have a guiding effect on house price movements in other cities. With the growth of information technology and the penetration of networks, the Internet is increasingly able to generate new data in real time. In the context of big data, it has been observed that Internet search data contains a wide range of predictable information [53–55]. In China, Baidu search is currently the most popular search engine [56]. In this study, we used a combination of predictor variables based on web search keywords related to house prices, and the data were gathered from the Baidu search index (http://index.baidu.com), and the period was daily data collected between January 2011 and February 2022 and summed up and organized into monthly data. The forecast label is the monthly chain data of the sales price index of new commodity residential units released by the NBS, denoted by the symbol y_t, which is obtained from the NBS website (http://www.stats.gov.cn/), and the time frame is from January 2011 to February 2022.

Web search data reflects objectively the possible relevant demands of web users. Due to the large number of web search terms related to the prediction labels, selecting the most effective keywords is essential to determining the accuracy of the prediction. Initial collection methods of keywords include direct word selection, technical word selection, and range word selection. Comparing the advantages and disadvantages of these three methods, as well as reviewing the relevant literature, this study adopts the method of range word selection in conjunction with direct word selection for the initial network keyword lexicon for house price prediction. These aspects include macroeconomic regulation, financial policy, tax policy, security policy, land policy, house price expectations, transaction details, and residential characteristics, as shown in Table 1.

4.2. Data Preprocessing. A high-dimensional dataset can cause problems such as high computational complexity and slow running time of the model, and it is likely that network keyword data will contain a certain amount of noise due to its own characteristics. A basic idea behind feature selection is to select the most effective variables from the original data to minimize covariances between data and to reduce the dimensionality of the dataset. Feature selection before modelling can not only minimize noise and overfitting but also improve training efficiency and prediction performance. We select features using the Spearman correlation analysis, which is commonly used in statistics, and its mathematical principles are as follows:

$$\rho_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2, \tag{11}$$

where ρ_s is the Spearman correlation coefficient between web search data and real estate prices, n is the number of samples, and d_i is the set of ranking difference obtained from the corresponding subtraction of the descending ranking of the web search keyword X_i and the house price index y_t.

Taking into consideration the fact that the whole process from the emergence of consumers' intentions to purchase a house to the final purchase decision generally does not last longer than 12 months, and the online attention of Internet users varies with location. This study therefore calculates the correlation coefficients between the lagged 12 periods and y_t for each keyword in the initial keyword lexicon of the four cities separately. A significance level of 0.01 was also set, and only the data with the highest absolute value of correlation coefficient were used, to develop a system of web keyword prediction variables for each city. Moreover, since y_t at lag 1 is highly correlated with the predicted variable, the house price index at lag 1 is also added to the predictor variables, as shown in Table 2.

Data from the web searches of each of the four cities identified in Table 2 are used as the input features of the model, and the commodity residential sales price index data for each city is used as the prediction label. We used equation (12), which was applied to standardize the experimental dataset to reduce outlier interference and to avoid the impact of the magnitude of input variables on the predictive power of the model.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{12}$$

where x^* is the normalized data value, x is the input data value before normalization, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the input data. Normalized values fall within the [0,1] interval, and this method of data treatment increases the predictive power of the model to some extent [57]. Normalized data are divided proportionately into a training set and a test set (the training set represents 80 percent of the sample data, and the test set represents 20 percent), and after the model is trained, the prediction results are then back-normalized in order to obtain the predicted values.

4.3. Evaluation Functions. In this paper, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are selected as the evaluation functions of the prediction models. Among the three performance measures, smaller values of RMSE, MAE, and MAPE indicate better model prediction performance, as shown in the following equations:

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}, \\ \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|, \\ \mathrm{MAPE} &= \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|, \end{aligned} \tag{13}$$

where n denotes the number of months of the house price index in the test set, i denotes the number of months, y_i is the true value of the house price index in month i of the test set, and $\widehat{y_i}$ is the predicted value of the house price index in month i of the test set.

5. Experimental Results and Discussion

5.1. Analysis and Comparison of the Results of the Integrated Model. On the sales price index of new commodity residential units in four first-tier cities in China, Figure 4 illustrates the prediction effect of a single learner with the default Sklearn parameters and the stacking integrated model optimized by the GWO algorithm. Table 3 presents the hyperparameter settings for the GWO-stacking-based learner applied to four datasets. To allow for an accurate comparison and analysis of the prediction accuracy of different models, Table 4 provides the prediction results of each model for the four indicators. Table 4 and Figure 4 demonstrate that the stacking integrated model outperforms the single-base learner in all three-prediction metrics for the Beijing, Shanghai, and Shenzhen datasets, whereas it lags the SVR for the Guangzhou dataset. This proves that SVR has better prediction for small samples of high-dimensional

Table 1: Web search keyword phrase database.

Category	Web search keywords
Macrocontrol	Real estate regulation (X1), house price regulation (X2), home buying policy (X3), purchase restriction policy (X4), purchase restriction (X5), removal of purchase restriction (X6)
Financial policy	Down payment percentage (X7), down payment loan (X8), down payment (X9), mortgage (X10), mortgage loan interest rate (X11), mortgage to buy a house (X12), deposit prime rate (X13), deposit interest rate (X14), loan (X15), loan interest rate (X16), loan interest rate cut (X17), second mortgage policy (X18), mortgage (X19), mortgage interest rate calculator (X20), home loan (X21), mortgage interest rate (X22), latest mortgage interest rate (X23), home loan (X24), interest rate increase (X25), interest rate reduction (X26), bank loan (X27), personal loan (X28), home loan (X29), term loan (X30), equal principal (X31), equal principal interest (X32), is it a good idea to pay off your mortgage early? (X33), home loan calculator (X34), housing fund loan (X35), housing fund loan interest rate (X36), housing fund loan amount (X37)
Tax policy	Home sale tax (X38), property transaction tax (X39), property tax (X40), real estate sales tax (X41), property deed tax (X42), deed tax (X43), commercial property deed tax (X44), second home tax (X45), land value added tax (X46), property tax (X47), new property tax (X48), how to calculate property tax (X49)
Protection policy	Limited-price housing (X50), guaranteed housing (X51), guaranteed housing (X52), public rental housing (X53), public rental housing application requirements (X54), affordable housing (X55), low-rent housing (X56), low-rent housing application requirements (X57), shantytown renovation (X58), property rights law (X59), housing subsidies (X60), housing provident fund (X61), provident fund loan conditions (X62), personal housing provident fund inquiry (X63), housing provident fund inquiry (X64), housing provident fund withdrawal (X65)
Land policy	Land auction (X66), land sale (X67), land reserve (X68), land use rights (X69), idle land (X70), compensation for land acquisition (X71)
House price expectation	House price (X72), speculation (X73), house price trend (X74), house price trend chart (X75), China house price future trend (X76), will house price fall (X77), real estate bubble (X78)
Transaction details	House sale contract (X79), property certificate (X80), house sale agreement (X81), house sale agreement (X82), house purchase contract (X83), house purchase contract (X84), commodity house sale contract (X85), existing house (X86), term house (X87), property (X88), property fee (X89), second suite (X90), small property house (X91), housing registration method (X92), net signature (X93), what is the meaning of net signature (X94)
Residential characteristics and others	Real estate (X95), plot ratio (X96), what does plot ratio mean (X97), green ratio (X98), useful life of commercial houses (X99), five certificates (X100), years of ownership (X101), common area (X102), proportion of common area (X103), commercial houses (X104), vacant houses (X105), houses for rent (X106), rent-sales ratio (X107), real estate market (X108), real estate network (X109), real estate transaction network (X110), real estate information network (X111), real estate transaction center (X112), second-hand house website (X113), property market policy (X114), buying a house (X115), house buying procedures (X116), house buying process (X117), house buying notes (X118), school district housing (X119)

Table 2: Final forecast variables for the four cities.

Beijing			Shanghai			Guangzhou			Shenzhen		
Features	Number of lags	Correlation factor	Features	Number of lags	Correlation factor	Features	Number of lags	Correlation factor	Features	Number of lags	Correlation factor
X1	9	-0.405	X1	9	-0.515	X1	10	-0.436	X1	1	-0.375
X2	8	-0.315	X2	9	-0.472	X2	10	-0.338	X2	7	-0.406
X3	11	-0.320	X3	12	-0.251	X3	1	0.270	X3	10	-0.262
X5	4	-0.240	X9	1	0.255	X16	2	0.294	X6	6	0.282
X10	9	-0.254	X12	9	-0.395	X17	8	0.485	X8	12	-0.243
…	…	…	…	…	…	…	…	…	…	…	…
X106	11	-0.242	X111	9	-0.273	X113	1	0.251	X83	11	-0.310
X111	6	-0.293	X112	1	0.333	X114	7	0.336	X90	1	0.235
X113	12	-0.286	X113	12	-0.290	X115	1	0.303	X112	3	0.301
X114	6	0.269	X114	6	0.309	X118	1	0.281	X113	12	-0.273
X120	1	0.737	X120	1	0.817	X120	1	0.838	X120	1	0.789

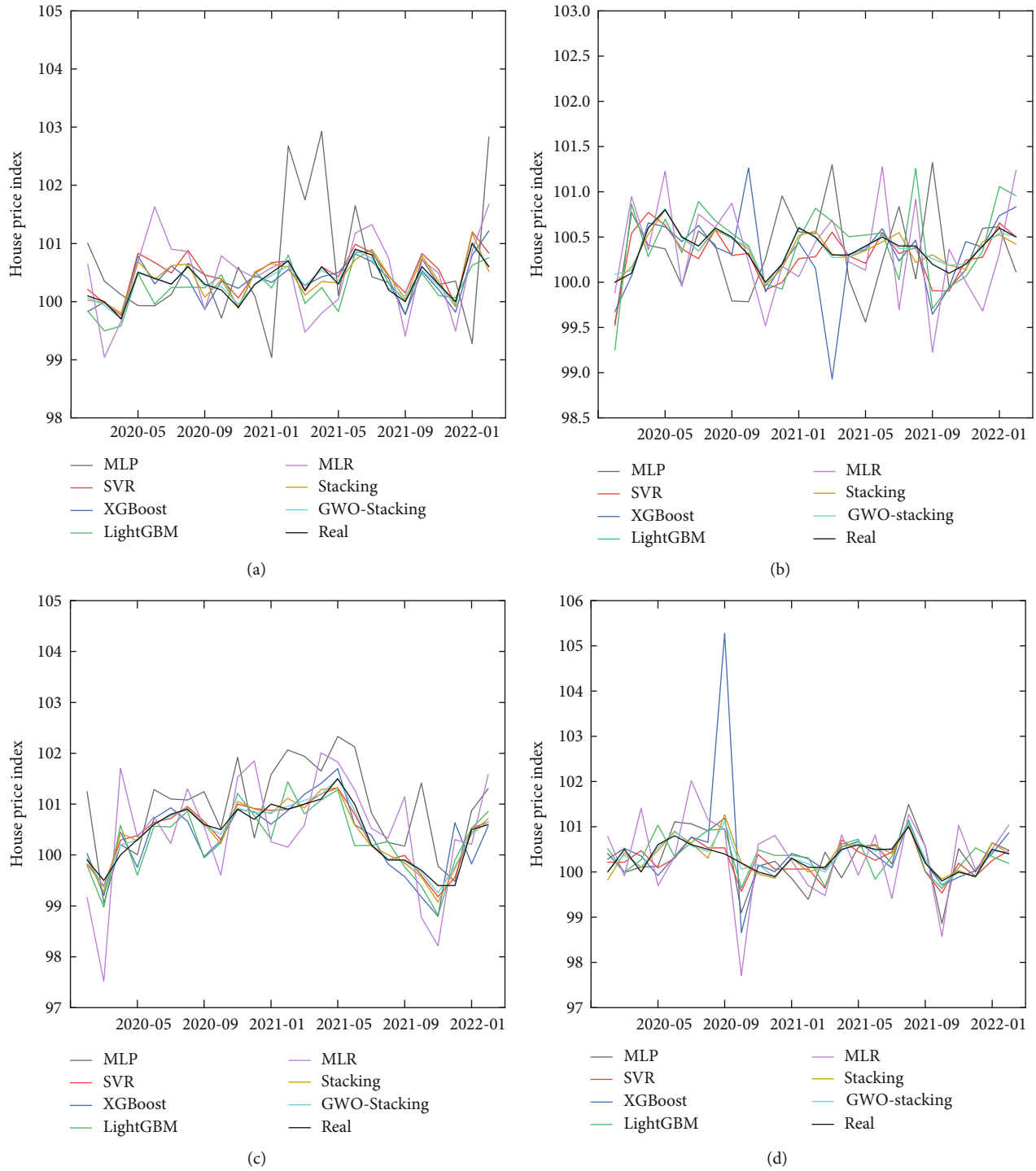

Figure 4: The effect of different model predictions for four cities.

data. In addition, while the prediction performance can be improved by stacked integration, its base learner parameters are difficult to determine, and direct prediction in a variety of datasets may lead to larger prediction errors, which may in turn increase the instability of the model. As observed, the GWO-stacking model proposed in this paper performs the best among all the compared models for the four datasets, which can be attributed to two main reasons. In one respect, the model integrates several different algorithms, reducing the prediction variance of the model and enhancing its generalization ability. On the other hand, the parameters of the base learner are intelligently calculated using the GWO algorithm, which enhances the overall stability and precision of the model.

5.2. Analysis of the Impact of the Integration Method on the Integration Model. In order to verify the effects of stacking integration method on the integration model, arithmetic

Table 3: Model parameter setting.

Algorithm category	Name	Beijing	Shanghai	Guangzhou	Shenzhen
Base learner	LightGBM	n_estimators = 5 Max_depth = 5 Num_leaves = 49 Learning_rate = 0.472	n_estimators = 12 Max_depth = 8 Num_leaves = 28 Learning_rate = 0.076	n_estimators = 8 Max_depth = 5 Num_leaves = 21 Learning_rate = 0.1805	n_estimators =2 Max_depth = 2 Num_leaves = 25 Learning_rate = 0.01
	XGBoost	n_estimators = 4 Max_depth = 2 Learning_rate = 0.818	n_estimators = 37 Max_depth = 3 Learning_rate = 0.0667	n_estimators = 20 Max_depth = 2 Learning_rate = 0.0661	n_estimators = 4 Max_depth = 2 Learning_rate = 0.6827
	SVR	$C = 3$ Gamma = 5	$C = 20$ Gamma = 4	$C = 43$ Gamma = 85	$C = 22$ Gamma = 3
	MLP	Hidden_layer_sizes = 298 Alpha = 0.8422	Hidden_layer_sizes = 123 Alpha = 0.6496	Hidden_layer_sizes = 277 Alpha = 0.8941	Hidden_layer_sizes = 21 Alpha = 0.8091
Optimization algorithm	GWO	Number of iterations: 100 Population size: 30	Number of iterations: 100 Population size: 30	Number of iterations: 100 Population size: 30	Number of iterations: 100 Population size: 30

Table 4: Comparison of the prediction accuracy of different models in four cities.

	Beijing			Shanghai			Guangzhou			Shenzhen		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
GWO-stacking	0.0476	0.0380	0.3481	0.0397	0.0293	0.2193	0.1129	0.0779	0.6445	0.1612	0.0893	0.3688
Stacking	0.1529	0.1272	0.3759	0.0721	0.0585	0.2229	0.1969	0.1518	0.6589	0.1931	0.0981	0.3742
LightGBM	0.2453	0.1957	0.4027	0.3749	0.2913	0.3962	0.4093	0.3277	0.7100	0.3755	0.3232	0.4145
XGBoost	0.2316	0.1914	0.3873	0.3904	0.2415	0.3620	0.4199	0.3224	0.7191	1.0517	0.4347	0.5986
SVR	0.1771	0.1514	0.4003	0.2007	0.1566	0.2732	0.1278	0.1031	0.6481	0.2888	0.2357	0.3751
MLP	1.0049	0.7172	0.7649	0.5019	0.4037	0.4091	0.7346	0.6085	0.9484	0.4889	0.4103	0.5283
MLR	0.5617	0.4578	0.6258	0.4837	0.3993	0.4695	0.8641	0.7209	0.9944	0.8887	0.7186	0.7434

mean (AM) and weighted average (WA) are used to compare them. The AM method assigns the same weights to the prediction results of the base learners, while WA assigns different weights according to the performance of the base learners, and both methods analyse the linear relationship between the predictions of the base learners. Figure 5 shows the prediction errors of the three integrated methods based on the four datasets. Based on the prediction error curves of the model in Figure 5, one can see that the prediction error of the GWO-stacking integrated model is closest to 0 in the four datasets as compared to the simple average method and the weighted average method, which indicates that this strategy is the most effective. By introducing nonlinear relationships and integrating the results of the base learner's prediction, the stacking integration method can provide more accurate insights into the relationship between the predicted values and the real house price data.

5.3. Feature Importance Analysis. As machine learning develops rapidly, models can achieve high levels of prediction accuracy, and the need for interpretable machine learning is also growing to ensure the reasons why models make decisions are reliable, which is essential for using machine learning to find truly novel scientific results [58]. The SHAP approach utilizes the Shapley value proposed by Lundberg and Lee [59] in 2017 using a game theory approach, treating each feature variable in the dataset as a player, and using the dataset to train the model to obtain predictions. It is the value created when many players work together to complete a project, considering the contribution each player makes and allocating the benefits of cooperation fairly through SHAP. The SHAP method is a model interpretation tool that applies to tree-based algorithms and quantifies the contribution of each feature to the prediction and reveals the relationship between the specific values of the features and the predictions [60]. A prediction value is generated for each sample in the model, and the SHAP value is the value assigned to each of the features in that sample. If the ith sample is x_i, the jth feature of x_i is x_{ij}, the predicted value of the model for that sample is y_i, and the baseline (usually the mean of all sample target variables) of the whole model is y_{base}; then, the SHAP value is calculated as shown in

$$y_i = y_{\text{base}} + f(x_{i1}) + f(x_{i2}) + f(x_{i3}) + \cdots + f(x_{ik}), \tag{14}$$

where $f(x_{ij})$ is the SHAP value of x_{ij} and $f(x_{i1})$ is the contribution of the 1st feature in the ith sample to the final predicted value y_i. In SHAP analysis, the magnitude, positive or negative, of the SHAP value for each feature is analysed to estimate the change in expected model prediction. The higher the SHAP value of a feature, the greater the impact

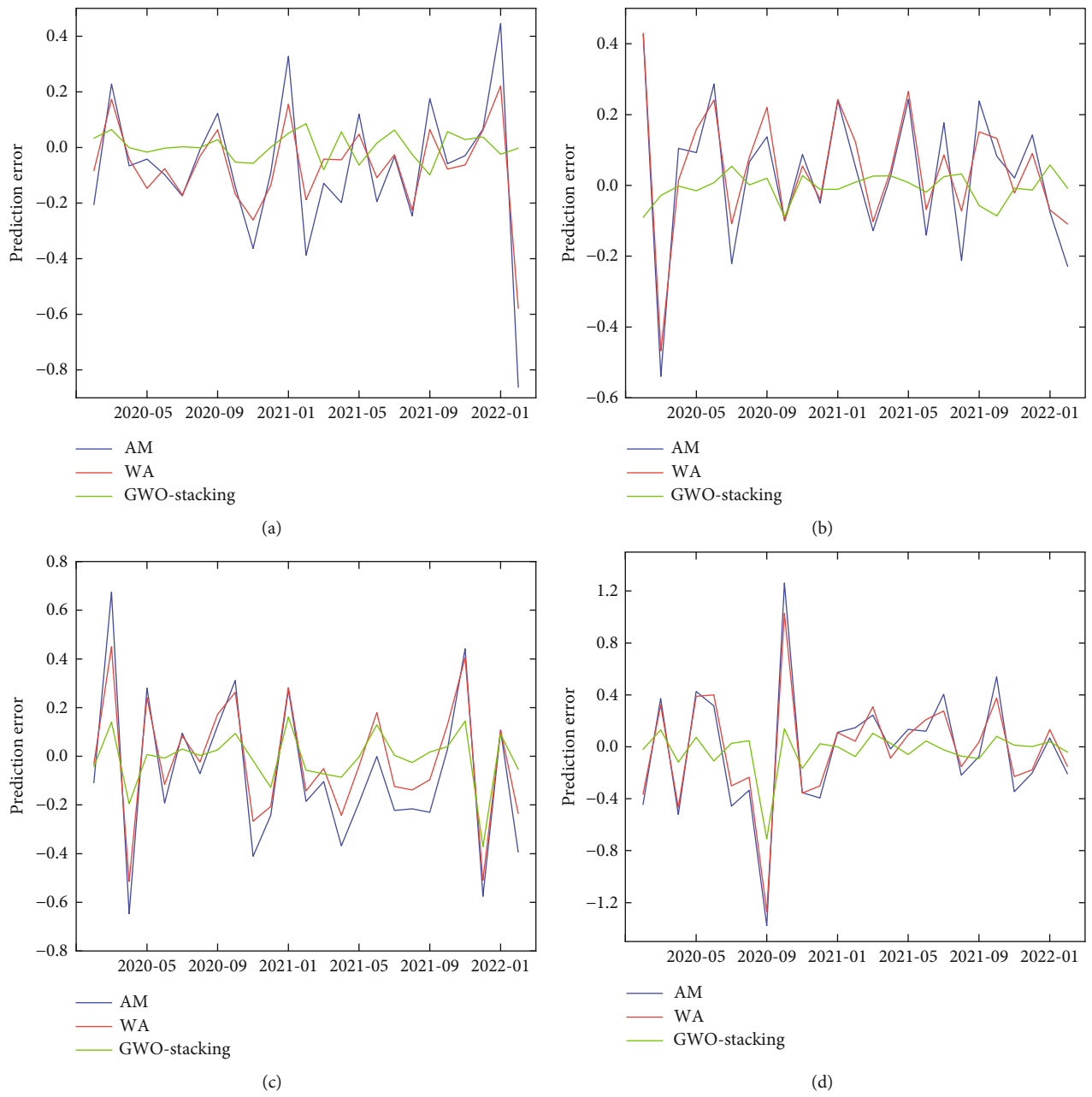

Figure 5: Comparison of prediction errors of three integration methods.

of that feature on the model, and the lower the SHAP value, the smaller the impact, with positive and negative values representing the positive and negative impact of a feature.

This paper uses the method in combination with the LightGBM algorithm based on the tree model for feature importance analysis with the aim of discovering factors which have a strong influence on the prediction of the house price index in the four cities. Figure 6 shows the SHAP summary plots of the top 10 features of the four cities. Each point represents a sample, and redder indicates a greater value for the feature, while blue indicates a lesser value for the feature. Based on the average of the absolute value of the SHAP value of each feature, the significance ranking of each feature was produced based on the influence of the predicted features on the house price index, as shown in Table 5.

It can be seen from the SHAP summary chart and Table 5 that the closest previous period house price index (X120) to the forecast month has the greatest influence on the forecast results and that as this value increases, the probability of an increase in the next period house price index increases. In addition, the higher the SHAP values of real estate regulation variables such as real estate regulation (X1), house price regulation (X2), and house buying policy (X3), the lower the house price index indicates that real estate macrocontrol policies are effective in curbing the rise in house prices, but the implementation effect of these policies is delayed. There is a strong correlation between

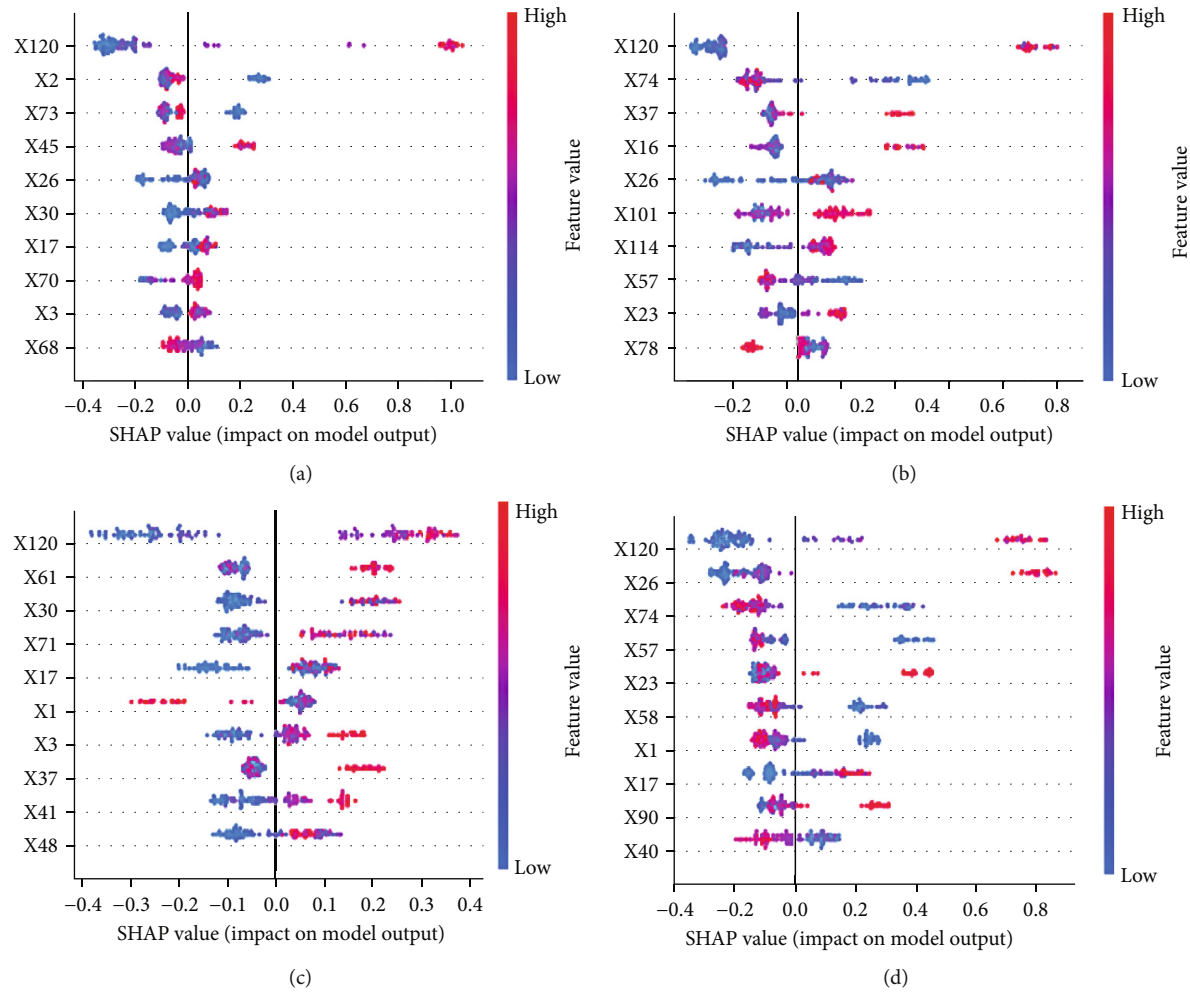

Figure 6: Four cities SHAP summary chart.

Table 5: Four cities in order of importance of keyword features.

Beijing		Shanghai		Guangzhou		Shenzhen	
Feature	Score	Feature	Score	Feature	Score	Feature	Score
X120	0.396209586	X120	0.377720893	X120	0.260471533	X120	0.30129077
X2	0.111398957	X74	0.182622338	X61	0.122077055	X26	0.279506534
X73	0.098570217	X37	0.119484245	X30	0.118872507	X74	0.190158678
X45	0.069979145	X16	0.119447621	X71	0.097751492	X57	0.151601933
X26	0.06726668	X26	0.118716092	X17	0.097546289	X23	0.150099089
X30	0.06367437	X101	0.117109957	X1	0.084826217	X58	0.11913004
X17	0.057235422	X114	0.100917925	X3	0.075693502	X1	0.117133864
X70	0.052079315	X57	0.079666529	X37	0.0749995	X17	0.109632074
X3	0.049618967	X23	0.067915002	X41	0.073965233	X90	0.101216581
X68	0.047516662	X78	0.066385166	X48	0.071451516	X40	0.084089772

keywords such as term loans (X30), latest mortgage rates (X23), interest rate cuts (X26), loan rate cuts (X17), and the house price index, because interest rate cuts lower the cost of owning monetary capital, which results in a large amount of capital being poured into the property market, driving up property prices. The influence of keywords such as house price trend (X74), real estate bubble (X78), property market policy (X114), and length of ownership (X101) is also significant, showing that residents are concerned about the real estate market, leading to a strong demand for home ownership, with a high demand, therefore, leading to an increase in house prices. The keyword second-hand

house tax (X45), real estate business tax (X41), and new property tax (X48) are positively related to the house price index, because of the government proposing various tax policies in order to curb housing prices, but these policies do not take effect immediately but are gradually implemented. If potential home buyers are informed through the Internet that the tax on property transactions will increase or be introduced, it will likely cause some panic, and this in turn will lead to the purchase of houses before the policy is implemented, which in turn will increase the prices of homes.

6. Conclusions

In the context of big data, traditional forecasting data and forecasting techniques are increasingly unable to meet the needs of realistic forecasting work. Aiming at the nonlinear variation characteristics of real estate prices, this paper proposes a GWO-stacking integrated learning house price index forecasting model combined with web search data. Following a comparative analysis, the following conclusions were drawn:

(1) This paper constructs an initial keyword phrase database for web search from eight aspects related to property prices, macroeconomic regulation, financial policy, tax policy, security policy, land policy, house price expectation, transaction details, residential characteristics, and others, and uses Spearman correlation analysis to filter out the final keyword prediction variables. The results show that the prediction using the dataset built in this paper is about two weeks ahead of the official release of the house price index, and the prediction results not only make up for the relative lag in the release of traditional statistical data information but also can be used as an effective supplement and reference to traditional real estate price statistics

(2) Simulation experiments on four Chinese first-tier city datasets in Beijing, Shanghai, Guangzhou, and Shenzhen show that the stacking integration strategy of using the XGBoost, LightGBM, MLP, and SVR models as base learners and the MLR model as a metalearner achieves better prediction results compared to a single prediction model; optimizing the hyperparameters of the base learner using the GWO algorithm can further improve the prediction accuracy and stability of the hybrid model; compared to other integration methods, the stacking method has a smaller generalization error. This experimental result proves that the GWO-stacking model proposed in this paper is a reasonable and effective model and can be applied to the field of house price prediction with high prediction accuracy

(3) To enhance the interpretability of the machine learning model, this study calculates and visualizes the SHAP values of each predictor variable during the prediction process and then performs feature importance ranking. The results show that the most influential variable on all four cities is the house price index in the previous period, while the financial policy category keywords are also an important factor influencing the change in the house price index

The shortcoming of this study is the manual selection of different combinations of base learners based on previous studies, and this method is not efficient. In future research, it is hoped to build a more intelligent prediction system by building a base learner candidate library and then combining it with an intelligent optimization algorithm to achieve automatic combination of base learners.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No. 81973791).

References

[1] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[2] K. J. Lancaster, "A new approach to consumer theory," *Journal of Political Economy*, vol. 74, no. 2, pp. 132–157, 1966.

[3] D. N. Chambers, "The racial housing price differential and racially transitional neighborhoods," *Journal of Urban Economics*, vol. 32, no. 2, pp. 214–232, 1992.

[4] J. Chica-olmo, "Prediction of housing location price by a multivariate spatial method: cokriging," *Journal of Real Estate Research*, vol. 29, no. 1, pp. 91–114, 2007.

[5] D. C. Wheeler, A. Páez, J. Spinney, and L. A. Waller, "A Bayesian approach to hedonic price analysis," *Papers in Regional Science*, vol. 93, no. 3, pp. 663–683, 2014.

[6] N. M. Noor, M. Z. Asmawi, and A. Abdullah, "Sustainable urban regeneration: GIS and hedonic pricing method in determining the value of green space in housing area," *Procedia-Social and Behavioral Sciences*, vol. 170, pp. 669–679, 2015.

[7] X. Liang, Y. Liu, T. Qiu, Y. Jing, and F. Fang, "The effects of locational factors on the housing prices of residential communities: the case of Ningbo, China," *Habitat International*, vol. 81, pp. 1–11, 2018.

[8] C. Wei, M. Fu, L. Wang, H. Yang, F. Tang, and Y. Xiong, "The research development of hedonic price model-based real estate appraisal in the era of big data," *Land*, vol. 11, no. 3, p. 334, 2022.

[9] H. S. Guirguis, C. I. Giannikos, and R. I. Anderson, "The US housing market: asset pricing forecasts using time varying coefficients," *The Journal of Real Estate Finance and Economics*, vol. 30, no. 1, pp. 33–53, 2005.

[10] W. Miles, "Boom–bust cycles and the forecasting performance of linear and non-linear models of house prices," *The Journal*

of Real Estate Finance and Economics, vol. 36, no. 3, pp. 249–264, 2008.

[11] L. Zhao, J. Mbachu, and Z. Liu, "Exploring the trend of New Zealand housing prices to support sustainable development," *Sustainability*, vol. 11, no. 9, p. 2482, 2019.

[12] L. Hanyue, Z. Jianping, and W. Chengrong, "A model for online forum traffic prediction integrated with multiple models," *Computer Engineering*, vol. 46, no. 12, pp. 60–66, 2020.

[13] Q. Zhao, W. Xu, Y. Ji, G. Liu, and W. Zhang, "Application of machine learning to financial asset price forecasting and allocation: a literature review," *Chinese Journal of Management*, vol. 17, no. 11, pp. 1716–1728, 2020.

[14] T. Kauko, P. Hooimeijer, and J. Hakfoort, "Capturing housing market segmentation: an alternative approach based on neural network modelling," *Housing Studies*, vol. 17, no. 6, pp. 875–894, 2002.

[15] G. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: a decision tree approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006.

[16] X. Li Daying and C. R. Wei, "Real estate price forecast using rough sets and wavelet neural networks," *Management Review*, vol. 21, no. 11, pp. 18–22, 2009.

[17] H. Selim, "Determinants of house prices in Turkey: hedonic regression versus artificial neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2843–2852, 2009.

[18] D. Qian, S. Nana, and L. Wei, "Real estate price prediction based on web search data," *Statistical Research*, vol. 31, no. 10, pp. 81–88, 2014.

[19] K. Shah, H. Shah, A. Zantye, and M. Rao, "Prediction of rental prices for apartments in Brazil using regression techniques," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT),*, pp. 1–7, Kharagpur, India, 2021.

[20] J. Andre, P. Siarry, and T. Dognon, "An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization," *Advances in Engineering Software*, vol. 32, no. 1, pp. 49–60, 2001.

[21] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3383–3386, 2011.

[22] D. Fei and J. Mingyan, "Housing price prediction based on improved lion swarm algorithm and BP neural network model," *Journal of Shandong University (Engineering Science)*, vol. 51, no. 4, pp. 8–16, 2021.

[23] Y. Fang, "Forecast of foreclosure property market trends during the epidemic based on GA-BP neural network," *Scientific Programming*, vol. 2022, Article ID 3220986, 7 pages, 2022.

[24] X. Deng, A. Ye, J. Zhong et al., "Bagging-XGBoost algorithm based extreme weather identification and short-term load forecasting model," *Energy Reports*, vol. 8, pp. 8661–8674, 2022.

[25] Y. Duan, Y. Mao, Y. Guo, X. Wang, and S. Gao, "COVID-19 propagation prediction model using improved grey wolf optimization algorithms in combination with XGBoost and bagging-integrated learning," *Mathematical Problems in Engineering*, vol. 2022, Article ID 1314459, 13 pages, 2022.

[26] Z. Li and S. Lv, "Performance analysis and optimization of packed-bed TES systems based on ensemble learning method," *Energy Reports*, vol. 8, pp. 8165–8176, 2022.

[27] H. Zhu and H. Li, "Predict prices of second-hand house using GBDT algorithm and PSO algorithm," *Frontiers in Economics and Management*, vol. 2, no. 11, pp. 513–524, 2021.

[28] J. Alfaro-navarro, E. L. Cano, E. Alfaro-cortés, N. García, M. Gámez, and B. Larraz, "A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems," *Complexity*, vol. 2020, Article ID 5287263, 12 pages, 2020.

[29] X. Wang, S. Gao, S. Zhou, Y. Guo, Y. Duan, and D. Wu, "Prediction of house price index based on bagging integrated WOA-SVR model," *Mathematical Problems in Engineering*, vol. 2021, Article ID 3744320, 15 pages, 2021.

[30] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[31] Y. Lin, R. Wang, X. Gong, and G. Jia, "Cross-correlation and forecast impact of public attention on USD/CNY exchange rate: evidence from Baidu index," *Physica A: Statistical Mechanics and its Applications*, vol. 604, article 127686, 2022.

[32] X. Zhu and C. Xia, "Visual network analysis of the Baidu-index data on greenhouse gas," *International Journal of Modern Physics B*, vol. 35, no. 8, article 2150115, 2021.

[33] K. T. Roberto, R. D. G. Jamora, K. M. C. Moalong, and A. I. Espiritu, "Infodemiology of autoimmune encephalitis, autoimmune seizures, and autoimmune epilepsy: an analysis of online search behavior using Google Trends," *Epilepsy & Behavior*, vol. 132, article 108730, 2022.

[34] L. Wu and E. Brynjolfsson, *Economic Analysis of the Digital Economy*, University of Chicago Press, 2015.

[35] E. Beracha and M. B. Wintoki, "Forecasting residential real estate price changes from online search activity," *Journal of Real Estate Research*, vol. 35, no. 3, pp. 283–312, 2013.

[36] N. Rizun and A. Baj-rogowska, "Can web search queries predict prices change on the real estate market?," *Ieee Access*, vol. 9, pp. 70095–70117, 2021.

[37] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, 2016.

[38] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[39] G. Ke, Q. Meng, T. Finley et al., "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[40] A. Sharma and B. Singh, "AE-LGBM: sequence-based novel approach to detect interacting protein pairs *via* ensemble of autoencoder and LightGBM," *Computers in Biology and Medicine*, vol. 125, article 103964, 2020.

[41] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[42] O. Sagi and L. Rokach, "Ensemble learning: a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, article e1249, 2018.

[43] A. I. Naimi and L. B. Balzer, "Stacked generalization: an introduction to super learning," *European Journal of Epidemiology*, vol. 33, no. 5, pp. 459–464, 2018.

[44] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren, "StackGenVis: alignment of data, algorithms, and models for

stacking ensemble learning using performance metrics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1547–1557, 2021.

[45] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 1999.

[46] M. Awad and R. Khanna, *Efficient Learning Machines*, Springer, 2015.

[47] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, vol. 43, no. 1, pp. 3–31, 2000.

[48] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999.

[49] M. G. Meharie, W. J. Mengesha, Z. A. Gariy, and R. N. Mutuku, "Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects," *Engineering, Construction and Architectural Management*, vol. 29, no. 7, pp. 2836–2853, 2021.

[50] Y. Xu, R. Meng, and X. Zhao, "Research on a gas concentration prediction algorithm based on stacking," *Sensors*, vol. 21, no. 5, p. 1597, 2021.

[51] J. Yu, R. Pan, and Y. Zhao, "High-dimensional, small-sample product quality prediction method based on mic-stacking ensemble learning," *Applied Sciences*, vol. 12, no. 1, p. 23, 2022.

[52] Y. Bao and Z. Liu, "A fast grid search method in support vector regression forecasting time series," in *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 504–511, Springer, 2006.

[53] S. Park and J. Kim, "The effect of interest in renewable energy on US household electricity consumption: an analysis using Google Trends data," *Renewable Energy*, vol. 127, pp. 1004–1010, 2018.

[54] M. Syamsuddin, M. Fakhruddin, J. T. M. Sahetapy-engel, and E. Soewono, "Causality analysis of Google Trends and dengue incidence in Bandung, Indonesia with linkage of digital data modeling: longitudinal observational study," *Journal of Medical Internet Research*, vol. 22, no. 7, article e17633, 2020.

[55] M. Costola, M. Iacopini, and C. R. Santagiustina, "Google search volumes and the financial markets during the COVID-19 outbreak," *Finance Research Letters*, vol. 42, article 101884, 2021.

[56] W. Zhou, L. Zhong, X. Tang, T. Huang, and Y. Xie, "The early warning and monitoring of Covid-19 by using Baidu search index in China," *Journal of Infection*, vol. 84, no. 5, pp. e82–e84, 2022.

[57] J. Singh, H. V. Knapp, J. Arnold, and M. Demissie, "Hydrological modeling of the Iroquois river watershed using HSPF and SWAT[1]," *Jawra Journal of the American Water Resources Association*, vol. 41, no. 2, pp. 343–360, 2005.

[58] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42200–42216, 2020.

[59] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[60] A. P. Carrieri, N. Haiminen, S. Maudsley-barton et al., "Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences," *Scientific Reports*, vol. 11, no. 1, pp. 1–18, 2021.