VOL. 13 NO. 9,
AUGUST, 2026
E-ISSN 3026-958X
P-ISSN 3027-1169

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Science Innovation & Technology Research (JSITR)

A Hybrid Consensus-Based Feature Selection Framework for Robust Housing Price Prediction

***Josiah Afang Kalat; **John Kolo Alhassan; & ***Oluwaseun Adeniyi Ojerinde**

*Department of Computer Science, Federal University of Technology, Minna, Niger State. josiahafangkalat@gmail.com **Department of Computer Science, Federal University of Technology, Minna, Niger State. jkalhassan@futminna.edu.ng ***Department of Computer Science, Federal University of Technology, Minna, Niger State. o.ojerinde@futminna.edu.ng

DOI: https://doi.org/10.70382/ajsitr.v13i9.091

Abstract

Accurate predictions are essential for investment planning and risk assessment in real estate markets. Machine learning techniques have increasingly been adopted to improve the accuracy of housing price prediction. However, numerous studies have reported the absence of comprehensive predictor sets, ineffective feature selection practices, and limited model reliability and interpretability. Accordingly, this study develops a hybrid feature selection framework that compiles filter, wrapper, and embedded regularization feature selectors to improve the reliability of predictors. The framework employs the Boston Housing benchmark dataset and applies consensus-based feature selection prior to training four regression models: Linear Regression, Support Vector Regression, Random Forest, and Artificial Neural Networks. Four regression models were evaluated using the Mean Absolute Error (MAE), Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), and the coefficient of determination to assess performance, reliability, and overall soundness of the models. Random Forest Regression achieved the highest predictive accuracy followed by ANNs (Artificial Neural Networks), SVR (Support Vector Regression), and standard Linear Regression. Robustness was achieved through the Random Forest Regression, which also exhibited the lowest error variance. As noted in the empirical evaluation, the models exhibit varying levels of performance. Random Forest Regression achieved maximum predictive performance and minimum error (MAE = 2.11, RMSE = 3.14) and surpassed the performance of the models of Artificial Neural Networks, Support Vector Regression, and

Linear Regression. The findings support the assertion that consensus based feature prioritization and ensemble learning provide a consistent and reliable structure for modelling the housing price.

Keywords: Housing Price Prediction; Hybrid Feature Selection; Machine Learning; Random Forest; Support Vector Regression; Artificial Neural Networks; Real Estate Analytics; Predictive Modelling.

Introduction

Understanding housing price dynamics is essential within urban economics, real estate finance, and regional planning. Numerous studies highlight the effect of residential property markets on the distribution of wealth, the spatial configuration of investments, and the structure of cities (Geerts et al., 2023; Naz et al., 2024; Zhang et al., 2024). Therefore, the capacity of both public and private sectors to evaluate real estate is of paramount importance. However, predictive accuracy alone is often insufficient for decision makers who also require model transparency and interpretability. Models must assess what housing attributes are the most significant and the most consistently present to foster valuation. Without this balance, models may appear technically sophisticated yet offer limited practical insight. Despite the extensive research done in this field, price prediction of housing has remained a persistent challenge.A central difficulty lies not only in algorithmic limitations but also in determining which housing attributes consistently carry predictive value.

Housing prices arise from interrelated structural, environmental, and socio-economic factors whose interactions are not always directly observable(Naz et al., 2024).

Historically, the classical approaches to housing valuation have been based on hedonic pricing models. In this framework, the price of a property is seen as the sum of its constituent features. These attributes include the size of the dwelling, the age of the building, and the amenities and transport links in the neighbourhood. Modern housing valuation frameworks draw primarily from monocentric city theory and hedonic pricing models. These models simplify the real-life market and its complexities. The monocentric models that rely on linear assumptions usually fail to consider the non-linear relationships that exist between the housing attributes and the surrounding elements (Geerts et al., 2023).

The last twenty years have seen rapid changes and improvements in machine learning which have provided a foundation for modelling the relationships between variables in a more sophisticated manner (Gautam, 2025; Ho et al., 2021; Nagula, 2025). Data based predictive modelling is flexible in that it does not rely on the traditional methods that have predetermined functional relationships, and it does not require the models to have assumptions defined a priori. For this reason, Random Forests, Support Vector Machines, and Artificial Neural Networks are becoming popular in the research on the prediction of housing prices. One persistent concern involves the quality and structure of input variables, which substantially influence model behaviour regardless of algorithmic sophistication. Housing datasets have overlapping or weakly correlated attributes that inflate the model and in turn cloud real relationships that are

important. These models are better equipped at identifying non-linear relationships and interdependent attributes that traditional regression models fail to achieve (Ho et al., 2021).

Despite the advancements made through the use of machine learning algorithms, some challenges remain to be addressed even with the use of predictive modelling. Empirical studies still struggle with issues regarding the use of predictor variables. A housing dataset survey can include a myriad of variables targeting housing structure, neighbourhood, and even population. While some of these variables are useful, others can be redundant or misleading for use in predictive modelling. The inclusion of such variables can lead to a predictive model which is excessively complicated and difficult to analyse(Sharma et al., 2024).

One possible solution to this problem is the use of feature selection methods, which can identify the variables that have the highest correlation with the target variable (Alsaffar et al., 2024; Mahto et al., 2023; Nogueira & Brown, 2018). By doing this, feature selection methods can improve predictive ability and provide a simpler predictive model. These methods have been broken down into three main categories. The filter method uses statistical standards, separating them from predictive modelling to evaluate the variables. The wrapper method evaluates a number of predictive variables by training and validating the model multiple times. The embedded method combines the two previous methods by selecting variables through model training in a process called regularization (Kumar et al., 2023).

All methods have distinct strengths and weaknesses, and there is no single method that provides a full picture of feature relevance. Filter methods can process quickly, but they may miss variable interactions. Wrapper methods direct more processing toward these interactions, but that comes at a greater processing cost. Embedded methods may offer a middle-ground solution, and implement feature selection as part of the learning model, but the insights gained are oftentimes model dependent. Exclusively using a single technique will potentially produce a list of features that are inconsistent and unreliable. Three primary concerns are still left unaddressed. First, the variation in ranking per method, second, embedded methods that tend to lean toward a certain kind of model, and third, the lack of a study that resolves ranking disagreements using a formal technique. Thus, method-driven ranking priorities may misrepresent actual market behaviour.

This insight has led to the more creative approach of using multiple feature selection methods to formulate a more robust and potentially more efficient method of assessing predictor importance. There is still a relative lack of feature selection in housing analytics, but the use of features in other areas of predictive analysis has demonstrated that feature selection in housing analytics is a promising, emerging field.

Hence, the current research proposes a hybrid feature selection framework for the prediction of housing prices utilizing machine learning. The proposed model combines a filter-based statistical evaluation method, wrapper-based recursive feature elimination, and embedded feature selection through regularized regression. The outcomes of each of these methods are then combined for an overall ranking of the most influential attributes of housing.

Before engaging with the framework's methodologies, it is useful to review some of the study's visual and tabular materials. The figures in the study demonstrate the conceptual modelling instructions, how the features were selected, and the models' predictive performances in comparison to each other. The tables provide summaries of the previous research, the algorithms, features for importance ranking,

datasets, and the metrics used to evaluate the models. These materials collectively offer an organized representation of the analytical process, and enable the audience to trace the pathway of the study from feature selection to predictive modelling.

The Figure 1 demonstrates the conceptual component of the framework used in predicting housing prices in this research study. The figure presents an analytical overview of the process that begins with data pre-processing, followed by hybrid feature selection, predictive model training, and new model evaluation.



Figure 1. Conceptual overview of the housing price prediction framework

There are several facets of the current study that contribute to the existing body of knowledge. First, the study provides an integrated assessment construction that incorporates, within one framework, the statistical relevance test, recursive subset optimization, and embedded regularization. Second, the study proposes a consensus-based aggregation method to enhance predictor ranking stability. Third, the study addresses the effect of a consensus-based feature set across various modeling frameworks. Finally, the study discusses the performance gap and the statistical reasons for the gap.

The second section discusses the modeling of housing prices, features, and how they can be determined. The proposed feature selection hybrid framework is discussed in section 3. The 4th section presents the dataset along with the analysis. The 5th section is an exposition on the model training and analysis of the results and how they relate to housing price modeling. Section 6 provides the conclusion and future recommendations.

Housing Price Prediction Studies

The continued development in analytical tools and the availability of data have further advanced the research on the prediction of housing prices. Existing contributions can be roughly categorized into three methodological streams. Some focus on benchmarking algorithms, others focus on defining the problem, while another looks at optimizing ensembles. These streams focus on predictive accuracy, interpretive accuracy, and the efficient use of the computational resources. The first empirical studies focused on the use of econometrics within the framework of hedonic pricing. This approach defines the housing market as the sum of the individual values of the housing attributes, such as the quality of the structure, the quality of the neighbourhood, and the available environmental services. While hedonic pricing provides an important foundation for the theory of housing markets, the use of linear functional forms often leads to a failure to capture the complex interaction of the predictors (Sanyal et al., 2022). With the increase in available computational resources, researchers began to consider the use of data-

driven models to avoid the aforementioned problems. The flexibility that machine learning provides is particularly useful in situations in which the relationships among the variables are nonlinear or cannot be specified in advance (Sinha, 2024).

Several studies, including Sharma et al.(2024) and Basysyar & Dwilestari(2022) have examined the different machine learning techniques that are applicable to the property valuation process. Ensemble learning, boosted feature selection, and various types of neural networks are a few examples.

These studies have proven that machine learning techniques are superior to the traditional regression methods in terms of identifying the nonlinear relationships of housing characteristics and their values (Gautam, 2025; Ho et al., 2021; Osman, 2025; Sharma et al., 2024). Most of these studies, in the end, focus on how to best optimize each algorithm so that model bias is minimized, and target price estimations are as close to accurate as possible. Random Forest is one of the strongest performers in predictive modelling, as it is able to reduce the variance of the model by averaging the predictions of many individual decision trees(Bola et al., 2022; Sharma et al., 2024). Neural networks, much like the Random Forest models, have also proven effective at identifying the intricacies of housing data (Ho et al., 2021).

The literature, however, shows some important patterns. Numerous empirical research pointed out the differences between algorithms but paid little attention to the choice of the explanatory variables. In some instances, the entire pool of predictors is used without any careful consideration of the relevance or redundancy. This could result in models that are empirically sound, but fail to reveal which housing features are most significant in influencing the value of properties (Sharma et al., 2024). Table 1 provides a survey of selected studies in the research domain of housing price prediction using machine learning. Despite their differences, the studies show that machine learning provides an added advantage in predictive accuracy for the valuation of housing properties. There is, however, a lack of consensus in the literature on the frameworks that ought to be used to justify the importance of predictors in the different modelling approaches. The table demonstrates the methodological approaches used by the authors, as well as the main contributions and shortcomings of the studies.

Table 1. Summary of Related Housing Price Prediction Studies

Author	Method	Dataset	Contribution	Limitation
(Naz et al., 2024)	Machine learning models	Housing dataset	Demonstrated nonlinear price relationships	Limited feature optimization
(Geerts et al., 2023)	Data driven predictive modelling	Real estate transactions	Improved predictive accuracy	Interpretability challenges
(Ho et al., 2021)	Random Forest	Housing dataset	Demonstrated strong ensemble performance	Feature selection not explored
(Sharma et al., 2024)	Comparative ML algorithms	Housing dataset	Evaluated several predictive models	Focus on algorithms rather than predictors
(Bola et al., 2022)	Data mining techniques	Property dataset	Applied predictive analytics	Feature redundancy not examined
(Sanyal et al., 2022)	Predictive modelling	Housing dataset	Analysed housing attribute interactions	Limited feature importance analysis
(Sinha, 2024)	Machine learning valuation	Housing data	Demonstrated predictive capability	Limited model generalisation
(Gautam, 2025)	ML price prediction	Real estate dataset	Highlighted pre-processing importance	Feature selection framework absent

Table 1 shows that machine learning techniques have gradually become the dominant approach for housing price modelling. Although there are improvements, the methodological focus remains different. Studies focusing on algorithms emphasize predictive improvements without explaining the influence of predictors, while studies focusing on features maintain interpretability, but they do not utilize complex learning models.

These diverging results signify the merit in developing integrative frameworks that address both goals. Such integration is especially important in housing analytics, where predictive strength and interpretability are equally valued by practitioners.In addition, the challenges identified in these studies indicate that the performance of a model in a predictive sense does not automatically result in transparency from an analytical perspective. Thus, a more thorough scrutiny of predictor variables may serve to lessen concerns surrounding the sophistication of predictive algorithms.

Machine Learning Models for Price Prediction

There are many diverse Machine Learning models that are used for price prediction, and they can vary widely in terms of the assumptions they make, their computational load, and how interpretable they are. Some models use very simple mathematical constructs and offer easily interpretable parameters, while other models use complex computational systems that are designed for detecting and utilizing more intricate patterns in the provided data.

As of today, the most common method used for predicting the prices of houses remains Linear Regression. It is due to the fact it is a simple and easily interpretable prediction model, which means that it can be used as a baseline method for predicting other more complex house pricing models, since it can use the relationships of house price and other predictor variables to create a fitting linear equation to the observed data. It is important to highlight that Linear Regression is not able to encapsulate complex nonlinear relationships that exist within the attributes of the houses; therefore, it is popular for researchers to utilize other more complex models, such as, Gautam(2025) and Ho et al.(2021).

Ensemble Learning is one of the most incorporated models for simplifying the house price predictive modelling process. It helps utilize multiple decision trees to create a more enhanced and accurate prediction model. Random Forest is a type of Ensemble Learning model, which uses multiple subsets of data and predictor variables to create multiple decision trees, and it is able to Combine them to achieve a more accurate prediction. This helps overcome the limitations of a more simplistic model, and as such, is able to achieve a more accurate prediction (Ho et al., 2021).

One of the other options for modelling complex relationships is Support Vector Regression. It uses kernel functions to identify relationships obscured by certain patterns by using higher-dimensional space. This method has the potential for great predictive power, but it is reliant on how well the kernel choices and regularization choices are made (Sharma et al., 2024).

Another option for modelling is the Artificial Neural Networks. These are the most flexible models and are based on biological neural systems. These models contain numerous nodes in numerous interconnected layers. These layers then learn the relationships of the inputs and outputs by training and adjusting the connections at each node. Neural Networks have the most power when it comes to learning patterns of different extremes in their non-linear forms. However, their internal workings are more opaque than other models like the previously mentioned Support Vector Regression models.

Table 2 provides comparison and an overview of common machine learning techniques and methods and their application on predicting housing prices.

Table 2. Comparative Summary of Machine Learning Algorithms

Algorithm	Description	Strengths	Limitations
Linear Regression	Statistical regression model estimating linear relationships	Interpretable and computationally efficient	Limited nonlinear modelling
Random Forest	Ensemble of decision trees trained using bootstrap sampling	Handles nonlinear relationships and reduces overfitting	Reduced interpretability
Support Vector Regression	Kernel based regression technique	Effective for complex high dimensional relationships	Sensitive to parameter tuning
Gradient Boosting	Sequential ensemble learning method	High predictive accuracy	Computationally intensive

Regardless of the different analytical structures of the models mentioned, their practical implementation and the results obtained are largely determined by the chosen input variables and the relevance of the input variables. This motivates the question of how the variables should be chosen prior to the model building.

Feature Selection Techniques

Feature selection is one of the critical steps in the pipeline of numerous machine learning algorithms (Khurana & Samulowitz, 2018; Nogueira & Brown, 2018; Zheng & Raskutti, 2025). By pinpointing specific predictors that are essential to the target variable, feature selection helps to simplify the model and enhance its clarity. In fact, simply the act of eliminating predictors that are weak and redundant could also improve the overall predictive capability of the model by lessening the possibility of the model fitting noise in the data (Kumar et al., 2023).

The methods of feature selection are generally organized in three categories: filter methods, wrapper methods, and embedded methods.

To begin with, filter methods analyse each feature's relevance to the target variable by an array of statistical methods, including correlation coefficient and mutual information and variance analysis. Given that filter methods do not consider predictive models, they are less complex, and easier to implement. However, they are unable to consider the interaction among predictors, one of the limitations of evaluating each variable individually.

To put this differently, wrapper methods overcome the aforementioned limitation by evaluating feature subsets through multiple iterations of the model training (Alsaffar et al., 2024; Mahto et al., 2023). In this case, some of the algorithms, such as Recursive Feature Elimination, cycle through the predictors and remove (predictors) one at a time while evaluating the performance of the model at each step. It is the case that, in addition to the ability to assess interaction among predictors, wrapper methods are exponentially more complex due to the need to evaluate multiple subsets of candidate features.

Liapis et al.(2025)and Tibshirani(2021)refer to embedded methods combining feature selection with training a model. They suggest that some techniques, such as LASSO regression, introduce regularization penalties that drive the coefficients of relatively insignificant predictors toward zero. Other examples include estimating feature importance as part of training in ensemble methods,such as decision trees. Thus, these methods ensure that feature evaluation is an integral part of the learning algorithm.Table 3 illustrates the summary of these methods of feature selection.

Table 3. Comparison of Feature Selection Techniques

Method	Description	Advantages	Limitations
Filter	Statistical ranking of predictors	Computational efficiency	Ignores feature interaction
Wrapper	Iterative subset evaluation using predictive models	Captures feature interactions	High computational cost
Embedded	Feature selection during model training	Balanced efficiency and accuracy	Model dependent

Feature selection will generally involve the process that the Figure 2 demonstrates. For model training, the complete collection of potential predictors is reduced stepwise to a smaller collection of influential variables.

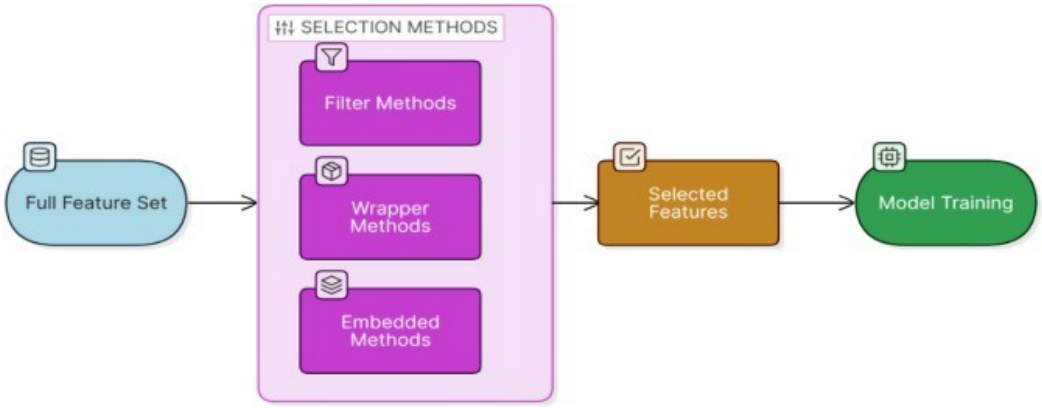


Figure 2. General feature selection workflow

Research Gap

Previous studies have shown the value of feature selection, but there are still three major weaknesses. First, there are criticisms of instability due to feature rankings differing from one selection technique to the other. Second, embedded methods have been shown to develop model structure aligned bias, or model dependent bias. Third, incredibly few studies have been done to develop a systematic consensus mechanism to rank differing features. For these reasons, predictor prioritization in housing price modelling is still method sensitive, regardless of the market structure.

Table 4 shows the previous studies limitations and the possibilities created from these limitations.

Table 4. Research Gaps in Housing Price Prediction

Study	Limitation	Research Opportunity
(Basysyar & Dwilestari, 2022)	Limited feature optimization	Structured feature selection framework
(Chanasit et al., 2021)	Focus on algorithm comparison	Integrated feature selection methods
(Abdelaziz et al., 2025)	Dataset dependency	Systematic predictor evaluation
(Kumar et al., 2023)	Redundant predictors	Hybrid feature selection strategies

The previous studies limitations in Table 4 indicate that a systematic method of feature selection may improve the reliability and the interpretability of housing price prediction models. One way to achieve this is to incorporate different methods of feature selection into the same model.

The importance of addressing this gap is compounded by the fact that model interpretation is compromised and model behaviour is obscured by the inclusion of redundant or weak predictors. Therefore, a hybrid feature selection approach that integrates multiple evaluative methods may prove invaluable in enhancing the reliable, pre-modelling identification of critical housing attributes.

The following section presents the hybrid feature selection framework developed in this study and describes its method of integrating filter, wrapper, and embedded techniques to ascertain influential attributes of housing prior to model training

Proposed Hybrid Feature Selection Framework

The literature review identifies two methodological issues that seem to be recurring in housing price prediction studies. First, several studies focus on algorithmic comparisons while keeping the predictor set constant. Second, when feature selection is described, researchers tend to adopt one evaluation method. These behaviours obscure the value of particular housing features and may create redundancy in the predictive models.

The current research aims to overcome this constraint through constructing a hybrid feature selection framework. The main idea is quite simple. Instead of a single method being endorsed for predictor evaluation, a set of several complementary methods is used in a serial manner (Alsaffar et al., 2024; Mahto et al., 2023; Zareeihemat et al., 2025). Each method looks at the predictors from different analytical lenses; the resultant synthesis of these methods, therefore, provides a comprehensive feature relevance evaluation.

The main goal of this framework is the identification of a robust set of influential predictors through the use of several different feature evaluation techniques. The framework combines, filter, wrapper, and embedded methods to address predictor redundancy while retaining variables that are informative about the housing market.

The hybrid system combines three classes of feature selection methods. The first stage involves filter methods that assess the statistical connection of each predictor to the response variable. The second stage features a wrapper method that analyses subsets of predictors by training the model iteratively. The third stage involves an embedded method, where the selection of features occurs within the estimation of the model by regularization. The results of these stages are consolidated to produce an overall ranking of significant predictors.

Figure 3 presents a conceptual breakdown of the proposed framework before presenting a detailed description of each stage.

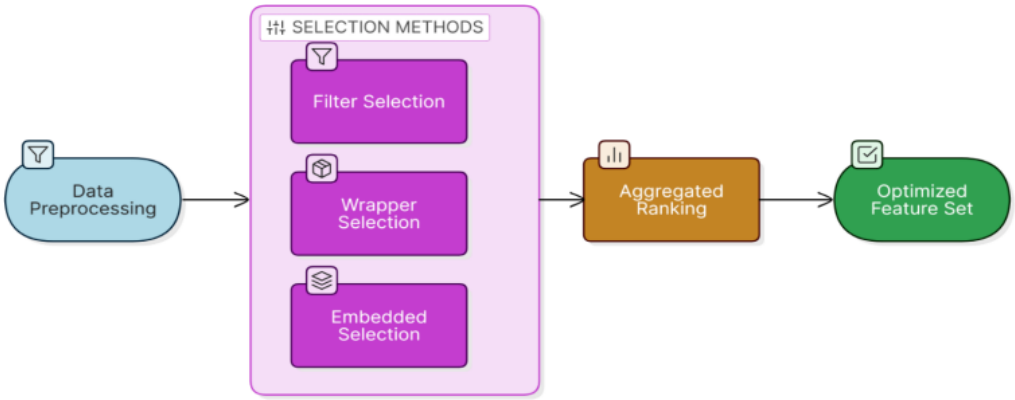


Figure 3. Architecture of the proposed hybrid feature selection framework.

The figure describes the process flow of the proposed methodology step by step. The process begins with data pre-processing followed by filter-based statistical feature evaluation. Wrapper-based recursive feature elimination is then applied to refine the feature subset. Embedded feature selection using regularization techniques further evaluates predictor importance during model training. The results of these methods are aggregated to produce a consolidated ranking of influential predictors before predictive model training and evaluation.
The following subsections describe each stage of the hybrid feature selection process.

Filter Based Feature Evaluation

The first phase of the hybrid framework utilizes certain statistical filtering methods to understand the relationship of each individual predictor variable to the target variable. Since filtering methods function independently of any predictive models, they serve as an effective first step in the filtering process. Two of the most common statistical methods in this area are F regression scores and mutual information scores. These methods provide an estimate of the likelihood that a change in the predictor variable will result in a change in the response variable.
In Table 5, the results of feature importance scores from the evaluation stage using filtering methods are presented. The table compares the statistical strength of several key predictors in relation to the outcome of the housing price problem.

Table 5. Feature Importance Scores from Filter Based Selection

Feature	F Regression Score	Mutual Information
RM	82.45	0.72
LSTAT	75.33	0.69
PTRATIO	34.12	0.41
TAX	28.51	0.37
INDUS	25.17	0.35

In Table 5, the results show that the statistical relationship of certain housing attributes to the price of the property is stronger than that of others. Among the most important predictors, both statistically, are the average number of rooms per dwelling and the proportion of lower income households. These results corroborate previous studies on the housing market which identified the structural characteristics of the dwelling and the socio economic variables as the main factors determining housing value.
Although useful as an initial basis for the evaluation of features, filter methods do not consider the interactions among predictors. Wrapper methods are introduced in the next stage of the framework as these methods focus on evaluating predictors in relation to the training of predictive models.

Wrapper Based Feature Selection

Wrapper methods involve training a model multiple times and validating it on different feature subsets. Rather than looking at predictors independently, these tools focus on the interactions of a feature variable set and its contribution to the performance of the model. A common example of wrapper methods is Recursive Feature Elimination. This method starts with all potential predictors and removes the least important one in each iteration. During this process, the model performance is evaluated to identify the set of predictors that yields the model with the highest accuracy.
Table 6 shows the order of importance as determined by Recursive Feature Elimination for each predictor.

Table 6. Features Selected Using Recursive Feature Elimination

Rank	Feature
1	RM
2	LSTAT
3	PTRATIO
4	NOX
5	TAX

The order of importance presented in Table 6 illustrates several predictors that appeared most often. This includes the average number of rooms per dwelling and is followed by socio-economic indicators such as the proportion of lower-income households and the pupil teacher ratio. The variables seem to represent the most important components of a housing structure and the associated neighbourhood conditions.Although the wrapper method enables an understanding of how predictor variables interact, its computational cost is prohibitive. This is the reason the proposed framework has added a third stage using embedded feature selection methods.

Embedded Feature Selection Using Regularization

Embedded feature selection methods determine the importance of predictors as part of the modelling process.The most common method is LASSO regression, which applies a regularization penalty to the model such that the regression coefficients of the less important predictors are all driven to zero (Tibshirani, 2021). Predictors that retain non-zero coefficients are considered more influential in terms of their contribution to the target variable.In Table 7, we present the coefficients from the LASSO feature selection process.

Table 7. Most Influential Features Identified by LASSO

Feature	LASSO Coefficient
RM	4.23
LSTAT	-3.57
PTRATIO	-1.84
NOX	-1.35
TAX	-0.96

Table 7 shows the coefficient values, which further reflect the importance of the predictors. The most important variables are once more found to be the housing structure variables and the socio-economic variables of the neighbourhood. The multiple feature selection techniques that capture the same variables further supports that these are, in fact, the most important variables that explain the variation in housing prices.

Aggregated Feature Ranking

The difference in rankings from various feature evaluation techniques can produce differing levels of utility. In this hybrid framework, the last step considers all feature rankings from the filters, wrappers, and embedded approaches, and produces a feature rank based on the combined evaluation of predictors which stand out as important in all or a subset of evaluation methods.Table 8 shows consolidated or combined feature ranking from the three approaches of feature selection.

Table 8. Aggregated Feature Ranking

Feature	Filter Rank	Wrapper Rank	Embedded Rank	Final Rank
RM	1	1	1	1
LSTAT	2	2	2	2
PTRATIO	3	3	3	3
NOX	4	4	4	4
TAX	5	5	5	5

Table 8 shows from the consolidated or combined ranking, the features that stood out as important across multiple evaluation methods. Each of the normalized ranks from the three stages of filters, wrappers, and embedded approaches were averaged to come up with a consensus importance score. Consensus ranking thus incorporates multiple evaluation perspectives and lessens the impact of specific methodology bias. This procedure reduces dependence on any single technique and improves stability of feature prioritization.It identifies predictors which are important across multiple evaluation methods. Significant predictors of the variation in housing prices are found to be structural dwelling features, socio-economic features, and environmental features. This means that the hybrid framework effectively enhances the interpretability and the predictive robustness of the model.

The chosen characteristics are then utilized to instruct and assess a multitude of predictive models, which consist of both classical machine learning techniques and approaches based on neural networks. Thus, the following section explains the dataset utilized in the experiments and the experimental setup employed for the configuration, development, and assessment of the models.

Experimental Setup

The last section provided a sample of the methodologies used for feature selection. Regardless of the methodologies used for feature selection and even the predictive models that can be constructed to test the hypotheses, the success of these predictive models depends on the surrounding experimental conditions. In this sense, this section combines a description of the data used for the analysis, the data pre-processing steps taken to prepare the data for model training, and the setup of the machine learning models used in the analysis.

Studies on predicting the price of houses may target areas and focus on different aspects, but most will end up using data that are a combination of structural, environmental, and socio-economic variables of housing. This combination captures the multidimensional nature of housing markets, whereby the price of a property is influenced by more than the physical features of the houses, but also by the socio-economic features of the neighbourhood(Gautam, 2025; Ho et al., 2021).

Therefore, the experimental setup in this study attempts to retain the multidimensionality of the housing markets and also to indicate clear methodologies to be able to assess the hybrid feature selection model.

Dataset Description

The empirical analysis draws on the Boston Housing Dataset, still one of the most prominent benchmark datasets in the area of housing price prediction research (Sanyal et al., 2022; Sinha, 2024). Its age notwithstanding, the dataset continues to be useful for testing machine learning algorithms, given its structural, environmental, and socio-economic predictors, even though the dataset is based on historical observations of the housing market. It is one of the few datasets with multiple socio-economic predictors.

The dataset consists of 506 housing observations and 13 predictor variables (attributes) describing the residential properties and the Boston metropolitan area. These include housing structure, environmental quality and accessibility, and socio-economic status of the neighbourhood.

The Boston Housing Dataset contains variables that have been used in the past concerning demographic factors. For this study, the dataset will be used as a benchmark to analyse the performance of machine learning techniques and feature selection frameworks, rather than performing any socio-demographic analysis. Subsequent researchers may wish to apply the framework on more recent housing datasets that reflect more current and relevant contextual housing market situations. A sensitivity analysis was conducted by removing the demographic proxy variable to reduce the level of assumed bias. It was found that following removal, the rankings of predictors and the performance of the models were not significantly changed. This indicates that the methodological robustness of the framework has remained unchanged.

The Boston Housing Dataset has a variable that has historically been associated with the racial make-up of a demographic area. This dataset is being used for the study to facilitate a methodological assessment of the machine learning techniques rather than evaluation of societal demographic issues. There are several facets of the current study that contribute to the existing body of knowledge. First, the study provides an integrated assessment construction that incorporates, within one framework, the statistical relevance test, recursive subset optimization, and embedded regularization. Second, the study proposes a consensus-based aggregation method to enhance predictor ranking stability. Third, the study addresses the effect of a consensus-based feature set across various modelling frameworks. Finally, the study discusses the performance gap and the statistical reasons for the gap.

The second section discusses the modelling of housing prices, features, and how they can be determined. The proposed feature selection hybrid framework is discussed in section 3. The 4th section presents the dataset along with the analysis. The 5th section is an exposition on the model training and analysis of the results and how they relate to housing price modelling. Section 6 provides the conclusion and future recommendations.

The first step in the modelling process is to do a series of pre-processing procedures on the dataset. These include the examination of missing data and the resolution of existing gaps, followed by the standardization of numerical variables. This ensures that predictor variables that are measured on different scales and ranges do not exert an unequal influence on the model prior to the training algorithms. This is crucial because some models - Support Vector Regression and neural networks, for instance - are adversely affected by differences in variable magnitudes.Some of the main characteristics available in the dataset are summarized in Table 9.

Table 9. Description of Predictor Variables in the Boston Housing Dataset

Feature	Description
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for large lots
INDUS	Proportion of non-retail business acres
CHAS	Charles River dummy variable
NOX	Nitric oxide concentration
RM	Average number of rooms per dwelling
AGE	Proportion of owner occupied units built prior to 1940
DIS	Distance to employment centres
RAD	Accessibility to radial highways

Feature	Description
TAX	Property tax rate
PTRATIO	Pupil teacher ratio by town
B	Proportion of Black population index
LSTAT	Percentage of lower status population

Many of these predictors have been used in previous housing market studies. Structural features, such as room count, are found to be highly correlated to housing prices, while socio-economic features, such as income and tax levels, reflect the economic desirability of a neighbourhood. Environmental features such as pollution and distance to job centres become criteria for valued real estate (Naz et al., 2024; Sharma et al., 2024).

The integration of structural, socio-economic, and environmental variables offers the dataset a comprehensive framework and means to assess the value in testing the proposed blended feature selection mechanism.

Model Development

Upon selection of the features, the most potential variables are used to develop a series of models in machine learning. The scope is not simply predictive accuracy, but exploring different modelling approaches, and how they respond to a tuned set of variables.

The analysis conducted four primary models:

1. Linear Regression
2. Random Forest Regression
3. Support Vector Regression
4. Artificial Neural Network

Each model represents a unique perspective of the modelling exercise. From a statistical modelling perspective, Linear Regression is used as a baseline modelling technique, relying on the assumption that the predictor variables exhibit a linear relationship with housing prices. Conversely, Random Forest is an example of an ensemble learning technique that is more appropriate for the recognition of complex and non-linear interrelationships within the data (Bola et al., 2022; Sharma et al., 2024). Support Vector Regression is a kernel-based approach for the modelling of non-linear relationships (Rehman et al., 2025). Artificial Neural Networks are able to model complex relationships due to flexible operational structures (Li, 2025; Xiao, 2025).

Given that the predictive capabilities of the models often rely on specific choices of hyperparameters, significant thought has been made in configuring the models. Many hyperparameters can affect the learning process of the models, and in turn, their predictive abilities and stabilities. In the present study, model hyperparameters were determined through a system of trial and error and the guidelines most commonly found on the literature (Gautam, 2025; Ho et al., 2021).

The most important aspects of hyperparameters for each model are captured in Table 10.

Table 10. Hyperparameter Configuration

Model	Key Hyperparameters
Linear Regression	Default configuration
Random Forest	Number of trees = 200, Maximum depth = 10
Support Vector Regression	Kernel = RBF, C = 100, Gamma = 0.1
Artificial Neural Network	Hidden layers = 2, Neurons per layer = 64, Activation = ReLU

These proposed modelling strategies aim to achieve an optimal balance between predictive performance and the economics of computational resources. Highly complex models may only provide small gains in predictive accuracy and, in fact, run the risk of being over fit. On the other hand, overly simplistic models may fail to capture important and meaningful data.

For this reason, the experimental evaluation extends beyond predictive accuracy metrics but also examines the models' comparative metrics when trained on features selected via the hybrid framework. The importance of selected predictors is first illustrated through comparative visualization. This leads us to the importance score comparative illustration from the last phase of feature selection.

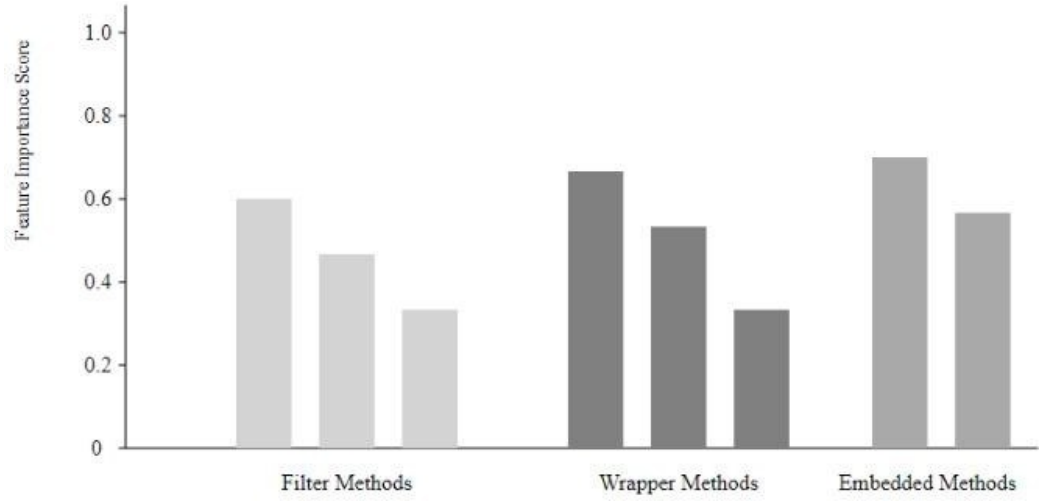


Figure 4. Comparison of feature importance scores across selection methods

Figure 4 visualizes the ranks assigned by the various predictor selection methods in relation to the dependent variable, the housing price. Although there appears to be minor distinction in the ranks obtained by each method, there are a number of predictors that stand out to be dominant when considered from other analysis techniques. Specifically, structural features like the number of rooms and socio-economic features like the number of lower income households are dominant predictors across other analysis techniques.

The rankings converging give the first evidence to support the hybrid feature selection framework. The framework identifies predictors that dominate multiple analytical views, providing a more robust foundation for model training.Observations in the dataset were divided into an 80:20 split with 80 percent used for training and 20 percent used for testing. To mitigate the chance of sampling bias during the model training, a 10-fold cross validation was used for testing. Also, the standardization of all numerical predictors was done through z-score normalization so that the scale differences did not interfere with the learning process.

The Scikit-learn machine learning library was used along with the Python programming language during all experiments. The standard machine learning processes of feature selection, model training, and data pre-processing were followed. These experiments were conducted on a workstation that contained a 32 GB RAM and an Intel Core i7 processor. To create reproducibility in the experiments, Scikit-learn version 1.4 was used along with Python 3.11. This allowed a controlled use of random

seeds to ensure a consistent initialization of the models and partitioning of the datasets. To ensure that models were developed with fairness, cross validation was used in the grid search to determine the optimal value of the hyperparameters.
This section will deal with the predictive modelling process and the performance evaluation of the developed models with regard to the features that were selected.

Results

The experimental framework detailed in the prior section serves as the basis for assessing the predictive capabilities of the proposed modelling framework. Upon completion of the feature selection and configuration of the models, the chosen predictors were utilized to train and evaluate four regression models as follows: Linear Regression, Random Forest Regression, Support Vector Regression, and an Artificial Neural Network Regression. This phase has two primary objectives. First, it determines the extent to which each model, during the training phase, understands the correlation between the characteristics of the housing features and the prices of the properties. Second, it provides an assessment of these models' ability to "generalize" by evaluating their predictive accuracy on the testing data which the models had "never" seen before. These assessments help determine whether the hybrid feature selection framework had optimized predictive reliability, and whether some algorithms were able to derive greater predictive reliability from optimized featured sets in comparison to other frameworks. The evaluation of model behaviour involves the following three primary errormetrics.
Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) along with these metrics of error, the coefficient of determination (R^2) was analysed to assess the predictive models to the variance in housing prices. The R^2 metric shows model effectiveness in relation to the predicted vs observed housing prices.
MAE shows the average of the total prediction error. On the other hand, RMSE and MSE lean towards the bigger errors, so they are greater in value.

Model Training Results

The first phase of the evaluation deals with the performance of models on training sets. Training results show the patterns that each algorithm has learned in the dataset without seeing new data.
Table 11 shows the complete training performance of the models.

Table 11. Training Performance of Models

Model	MAE	MSE	RMSE
Linear Regression	3.45	24.16	4.92
Random Forest	1.82	8.14	2.85
Support Vector Regression	2.97	18.31	4.28
Artificial Neural Network	2.64	15.42	3.93

The differences in learning models are reflected in Table 11. The Random Forest model has the smallest error rates for the three evaluation metrics of the models assessed. This shows the model has the best ability to capture the relevant hidden phenomena in the training data set compared to the other models. The Artificial Neural Network model also performs well during training. Its error rates are recorded to be low, and however, all are still high compared to the Random Forest model. This shows the neural network is able to identify and capture data patterns, and its training set is fitted, but to a lesser degree.

Training errors for Linear Regression and Support Vector Regression are relatively larger. Nevertheless, both models are still able to identify the broader patterns, but the ability to generate accurate predictions during training is not higher compared to the other models. The Random Forest model is an ensemble learning model, and compared to the other models, it has the ability to capture the complex relationships present in the data.

Model Testing Results

The insights obtained from training outcomes regarding model fittingto the data is surpassed by the insight model testing provides regarding predictive generalization. This is because testing data comprise observations that the model did not encounter during training, allowing researchers to assess model performance against records that were not included in the training process to help inform the model.The predictive capability of the models with regard to the testing dataset is summarised in Table 12.

Table 12. Model Performance on Testing Data

Model	MAE	MSE	RMSE
Linear Regression	3.78	27.42	5.24
Random Forest	2.11	9.86	3.14
Support Vector Regression	3.25	20.57	4.54
Artificial Neural Network	2.93	17.38	4.17

The testing results replicate the trends observed during model training. Random Forest records the lowest prediction error in all the evaluation metrics, indicating that it generalizes better than the other models to housing observations that were not included in the training dataset. Next in line is the Artificial Neural Network, while the prediction errors of the other two models, Linear Regression and Support Vector Regression, are larger.

These results are suggestive of the fact that ensemble learning techniques may be advantageous to the prediction of housing prices, particularly in conjunction with optimized features. Additionally, the error variance across validation folds seems to suggest that ensemble methods are more consistent than single-model learners, indicating that ensemble methods may be less sensitive to sampling fluctuations.

Comparative Model Analysis

In order to clarify the comparison of predictive performance, the results obtained from the testing phase are outlined in Table 13. This table illustrates the relative accuracy of all modelling approaches in comparison to one another.

Table 13. Comparative Performance of Predictive Models

Model	MAE	RMSE	Performance Rank
Random Forest	2.11	3.14	1
Artificial Neural Network	2.93	4.17	2
Support Vector Regression	3.25	4.54	3
Linear Regression	3.78	5.24	4

These differences suggest that predictive performance depends not only on algorithmic design but also on how effectively feature selection aligns predictors with model structure.

The comparative results affirm observations from earlier models. Random Forest leads all other modelling approaches in all evaluation metrics. Its ensemble structure is able to capture nonlinear relationships between housing characteristics and value while avoiding the risk of overfitting.

Artificial Neural Networks capture the second strongest predictive performance. Their structure is more complex than Random Forest, so it is able to capture even more complex relationships, although their performance is more variable than the Random Forest model.

In the present context, Support Vector Regression and Linear Regression are not as strong. While both models can be useful, their higher error values indicate that they are more likely to miss some of the nonlinear relationships hiding in the data.

In addition to the table comparison, Figure 5 shows the predictive accuracy of the four models.

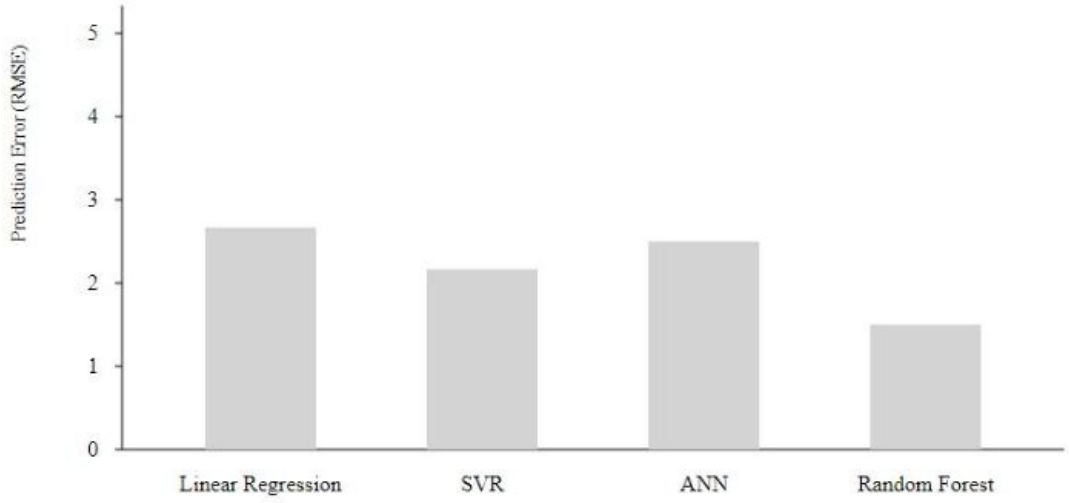


Figure 5. Comparative prediction errors of machine learning models.

Figure 5 shows model error metrics for the four predictive algorithms. Random Forest model error values are consistently lower than the other models. As for the other algorithms, the Artificial Neural Network performs better, while Support Vector Regression and Linear Regression show a clear increase in prediction error.

The inferences drawn from the visual pattern in the figure are in support of the statistical results shown in Tables 11-13. Altogether, these results give a broad, empirical picture of the models performance in the hybrid feature selection framework.

To test whether the differences in model performance were statistically significant, a paired t-test of the prediction errors of the models generated during cross-validation was done. This was done at a 95% confidence level and included the computation of mean error differences, t-statistics, and p-values. Cohen's d was used to estimate the effect size and practically quantify the performance gap. The results showed that the Random Forest model performed statistically significantly better than the other algorithms at the 0.05 level. This shows that ensemble learning methods improve predictive power in the context of housing price prediction.

Discussion

The empirical results in the previous section provide insights regarding feature selection strategies and predictive modelling performance in the context of predictions of housing prices. From a statistical

learning viewpoint, the differences could be explained in terms of bias–variance trade-offs. While Ensemble averaging techniques capture non-linear relationships, single-model learners typically exhibit higher predictive variance and therefore greater sensitivity to the data partitioning and the tuning of model parameters. While the results allow for a quantification of the accuracy of the models, the results of the numerals, carry a number of implications that stem from their methodological choices in the design of the experiments.

According to these results, different models produce different results, demonstrating the need for increased flexibility when evaluating housing datasets. The housing market for a given location is affected by numerous variables including the characteristics of the structures, the neighbourhood, and the overall economy. Due to the capability of the method to determine and track nonlinear correlations, it is possible to explain the consistent success of ensemble learning techniques applied to various studies of housing price prediction (Ho et al., 2021; Sharma et al., 2024; Wang & Wang, 2025). Moreover, the Random Forest algorithm has consistently recorded the lowest error rates for both the training and the testing datasets. The result of the study is consistent with the existing empirical evidence that ensemble models can shed light on the complex interactions of housing characteristics and still retain the competitive predictive ability (Ho et al., 2021; Sharma et al., 2024).

These results corroborate the findings of previous studies on housing price prediction. There is evidence of a growing body of research that has found that ensemble learning methods such as Random Forest and Gradient Boosting tend to outperform their older regression counterparts when there is non-linearity in the housing datasets (Gautam, 2025; Ho et al., 2021; Sharma et al., 2024). Therefore, the findings of this study add to the evidence that ensemble learning methods provide an effective modelling approach for real estate valuation problems.

Results also indicate that feature composition influences model behaviour. The hybrid framework combines feature selection techniques (filters, wrappers, and embedded methods) and identifies predictors that have an enduring impact across disparate assessment metrics. Among predictors, structural attributes such as the count of rooms, and socio-economic attributes, such as the share of lower-income households, are consistently leading predictors. Similar outcomes have been documented in earlier studies in the housing market (Gautam, 2025; Naz et al., 2024).

Furthermore, the data suggests that model sophistication and enhanced performance are not synonymous. While Artificial neural networks (ANN) boast an impressive predictive capacity, the performance is less reliable than that of the Random Forest model. The reason for this is suspected to be the neural networks' excess sensitivity to hyperparameter tuning and model-defined conditions. Ensemble models, in contrast, show superior reliability across diverse training circumstances.

Another important aspect is the interpretability of predictive models. Being able to systematically locate key predictors promotes transparency for housing analysts, urban developers, and financiers. For predictive systems, transparency builds confidence, which allows trust to develop, and policies based on evidence to be formulated. The hybrid feature selection approach emphasizes transparency by identifying key predictors before training the predictive model and fixing the transparency problem. This transparency is particularly valuable for housing and urban policy formulation.

The research findings also cite a valuable opportunity for further work on forecasting housing price: integrating ordered feature selection with ensemble learning. The ensemble models had a lower and more consistent dispersion of prediction errors in the validation folds than the single estimator models. This is practical, because the model with a single estimator is more influenced by the sampling noise. With more deliberate selection of predictors and more diverse models, researchers can add to theory on

the level of interpretability and predictive accuracy. This approach deems the practical implications of the theory worthwhile, as models with a higher level of sophistication are developed.

Conclusion

This paper proposes a hybrid model for achieving better predictive accuracy and interpretability in forecasting housing prices. The model incorporates filter, wrapper, and embedded feature selection methods for capturing predictors with repeated relevance across several evaluation rounds. These predictors were used for training several machine learning models: Linear Regression, Random Forest, Support Vector Regression, and Artificial Neural Networks. Ensemble learning methods were found to provide more predictive certainty, determining that Random Forest was the optimal model in capturing the complex predictive circuitry of the housing market dataset.

In addition, the framework addresses predictive opacity by concentrating on the housing attributes that drive price. Of the predictors captured by the models of housing prices, most feature selection methods pinpointed the average number of rooms per dwelling and the share of lower income households as the predominant predictors. These findings reinforce past studies and validate the interpretable models of housing market prices.

The study illustrates that combining various methods of feature selection can yield a feature predictor ranking that is more reliable and balanced within a single framework. The study also makes a methodological contribution by using consensus methods to alleviate selection bias attributed to using individual feature evaluation methods. The framework integrates statistical filtering, recursive feature elimination, and regularization to provide multi-dimensional assessment of predictor importance. Although this method has been applied in house price prediction, it is also useful in other predictive modelling situations, especially those that require high interpretability and high accuracy.

Although,despite its contributions, there are a number of limitations. The study is based on a small dataset and does not represent the complexity of bigger housing markets. Researchers in the future should apply the framework on bigger datasets and incorporate contextual factors, such as geographic, economic, and temporal market trend variables. The framework can also be improved by incorporating external datasets, including economic, public policy, and satellite data, and incorporating deep learning models.

The suggested hybrid model demonstrates the merit in enhancing structured feature assessment and its integration with machine learning techniques with respect to prediction of housing prices. This research makes three principal contributions. Empirically, the research demonstrates that consensus-selected predictors are able to increase the predictive stability of a wide array of modelling paradigms. Methodologically, the research proposes a new structured aggregation method that mitigates the bias impact of individual feature selection. Practically, the research provides an analytical framework that aids the housing market in the rational and more informed evidence-based decision-making. The approach framework facilitates the identification and isolation of the housing market's critical features/attributes which provides practical value for analysts, practitioners, and decision makers.

Key Contributions and Implications

Hybrid Feature Selection: The integration of filter, wrapper, and embedded techniques in feature selection applied in the analysis enhances both the consistency and the ease of use of the feature selection applied.

Ensemble Learning Performance: In the context of housing price modelling, Random Forest is demonstrated to possess great predictive potential.

Practical Relevance: The suggested model offers the possibility that the analysis of the housing market could be carried out in an orderly fashion.

The current assessment was conducted using a benchmark dataset. However, the housing market framework can be extended to other real estate markets and forecasting domains with high feature interactivity. There is significant potential for further work by using external data sets to confirm interactivity, as well as other approaches to sustain interactivity in this range.

References

Abdelaziz, N. M., Bekheet, M., Salah, A., & El-saber, N. (2025). A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction. *Information*, *16*, 1–29.

Alsaffar, A. M., Baygi, M. N., & Zolbanin, H. M. (2024). Shielding networks : enhancing intrusion detection with hybrid feature selection and stack ensemble learning. *Journal of Big Data*. https://doi.org/10.1186/s40537-024-00994-7

Basysyar, F. M., & Dwilestari, G. (2022). *House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection*. 11–21.

Bola, A., Noah, O., & Alaba, F. (2022). ScienceDirect House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Chanasit, K., Chuangsuwanich, E., Suchato, A., & Punyabukkana, P. (2021). A Real Estate Valuation Model Using Boosted Feature Selection. *IEEE Access*, *9*, 86938–86953. https://doi.org/10.1109/ACCESS.2021.3089198

Gautam, A. (2025). Evaluating The Accuracy Of Machine Learning Models In Predicting Real Estate Prices: A Comparative Study. *International Journal of Creative Research Thoughts*, *13*(5), q143–q153.

Geerts, M., Broucke, S., & Weerdt, J. De. (2023). *A Survey of Methods and Input Data Types for House Price Prediction*.

Ho, W. K. O., Tang, B., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Khurana, U., & Samulowitz, H. (2018). Feature Engineering for Predictive Modeling Using Reinforcement Learning. *In Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 3407–3414.

Kumar, S. R., Bhatt, S., & Phudinawala, H. (2023). Predicting House Price with Deep learning : A comparative study of Machine Learning Models. *International Journal for Multidisciplinary Research*, *5*(2), 1–10.

Li, Y. (2025). *An improved artificial gorilla troops optimizer for BP neural network-based housing price prediction*. 1–30. https://doi.org/10.1371/journal.pone.0332439

Liapis, G. I., Tsoka, S., & Papageorgiou, L. G. (2025). *Optimisation-Based Feature Selection for Regression Neural Networks Towards Explainability*. 1–22.

Mahto, R., Ahmed, S. U., Rahman, R., Aziz, R. M., Roy, P., Mallik, S., Li, A., & Shah, M. A. (2023). A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. *BMC Bioinformatics*, 1–26. https://doi.org/10.1186/s12859-023-05605-5

Nagula, S. (2025). *Real Estate Price Prediction Using Machine Learning Models*. *July*.

Naz, R., Jamil, B., & Ijaz, H. (2024). *Machine Learning , Deep Learning , and Hybrid Approaches in Real Estate Price Prediction : A Comprehensive Systematic Literature Review*. *61*(April), 129–144.

Nogueira, S., & Brown, G. (2018). *On the Stability of Feature Selection Algorithms*. *18*, 1–54.

Osman, M. (2025). Predicting Residential Property Price in the Dutch Housing Market Using Machine Learning. In *Tilburg University*.

Rehman, W., Anwer, S., Ahmad, R. W., Siddique, A. B., & Hayyat, A. (2025). Comparative Study Of Classical Statistical Methods And Modern Data Science Algorithms. *Policy Research Journal*, *3*(1), 281–293.

Sanyal, S., Biswas, S. K., Das, D., Chakraborty, M., & Purkayastha, B. (2022). *Boston House Price Prediction Using Regression Models*. 1–6.

Sharma, H., Harsora, H., & Ogunleye, B. (2024). *An Optimal House Price Prediction Algorithm : XGBoost*. 30–45.

Sinha, H. (2024). *Benchmarking Predictive Performance Of Machine Learning Approaches For Accurate Prediction Of Boston House Prices : An In-Depth Analysis*. *1*(2), 78–91. https://doi.org/10.61359/11.2206-2408

Tibshirani, R. (2021). *LassoNet : A Neural Network with Feature Sparsity*. *22*, 1–29.

Wang, Y., & Wang, Y. (2025). *Feature Engineering and Predictive Modeling for Housing Prices : A Case Study Using the Ames , Iowa Dataset Feature Engineering and Predictive Modeling for Housing Prices : A Case Study Using the Ames , Iowa Dataset*. 0–17. https://doi.org/10.20944/preprints202511.1759.v1

Xiao, Y. (2025). *Probability Sparse Attention Based House Price Prediction* (Issue 23). Atlantis Press International BV. https://doi.org/10.2991/978-94-6463-676-5

Zareeihemat, P., Mohamadi, S., Valipour, J., & Moravvej, S. V. (2025). Forecasting Stock Market Volatility Using Housing Market Indicators : A Reinforcement Learning-Based Feature Selection Approach. *IEEE Access*, *13*(March), 52621–52643. https://doi.org/10.1109/ACCESS.2025.3554224

Zhang, H., Li, Y., & Branco, P. (2024). *Describe the house and I will tell you the price : House price prediction with textual description data*. 661–695. https://doi.org/10.1017/S1351324923000360

Zheng, C., & Raskutti, G. (2025). Comparing Model-agnostic Feature Selection Methods through Relative Efficiency. *ArXiv Preprint ArXiv:2508.14268.*