Charmo Journal of Natural Sciences and Technologies (CJNST)

CJNST 2025, Vol. 0, Issue 1, pp. 12–24

https://cjnst.chu.edu.iq/

Research Article:

A Newly Proposed Hybrid Model for Predicting Real Estate Stock Prices in Kurdistan Region of Iraq

Yusra Juma Nader[1], Bryar A. Hassan[2,1]

[1]Department of Computer Science, College of Science, Charmo University, 46023 Chamchamal/Sulaimani, KRI, Iraq
[2]Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil, Iraq
https://doi.org/10.31530/cjnst.2025.1.1.2

Received: 08 July 2025
Revised: 31 August 2025
Accepted: 16 September 2025

*** Corresponding author info:** Yusra Juma Nader; yousra.jumaa@chu.edu.iq

Abstract.

Background: Accurate prediction of real estate stock prices plays a significant role in enabling investment choices, econometric forecasting, and strategic investment decisions in emerging markets. The Kurdistan Region of Iraq is among the emerging markets that are marked by non-homogeneity of data, limited representativeness, and non-availability of standardized appraisal methods.

Aims: This study aims to develop and test a stable hybrid machine learning model for predicting real estate stock prices in the Kurdistan region of Iraq. The new approach highlights methodological quality, representativeness of the dataset, and replicability in a bid to fill the gaps observed in the existing literature.

Methodology: The research makes use of the Kurdistan House Price Prediction (KHPP) data, which has various property features. Preprocessing of the data involved leakage-secure feature encoding, missingness checks, and deduplication verification. A combined model of Random Forest (RF) and Extreme Gradient Boosting (XGB) was used, alongside nested cross-validation for hyperparameter selection and testing. Statistical significance testing, bootstrapped confidence intervals, and feature ablation experiments were also performed for redundancy.

Results: The findings further indicate that the hybrid RF+XGB model outclassed individual models according to Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). While performance differences remained minute, the hybrid model always returned stable results per fold. Evaluation also observed challenges interpreting R2 and emphasized careful reporting for the raw target scale. The study also demonstrated an externally available benchmarking plan using the Ames dataset for cross-national comparability.

Conclusion: The new hybrid approach provides a systematic, replicable, and efficient framework for real estate stock price prediction within the Kurdistan Region of Iraq. By adding robust validation protocols and statistical testing, this study enhances belief in hybrid ensemble methods for under-represented markets. Future research must move beyond single-modal data and local criteria for greater generalizability and real-world applications.

Keywords: *Real Estate Price Prediction, Machine Learning, Hybrid Model, Ensemble Learning, Kurdistan Region, RF, XGB*

1. Introduction

Accurate prediction of residential property values has taken on increased importance in today's real estate markets, enabling better-informed decision-making processes for investors, policymakers, and buyers. ML and AI techniques have demonstrated considerable expertise in modeling complex, nonlinear relationships in property valuation. Models like RF, XGB, and DL have yielded promising results in different markets, with numerous studies supporting their effectiveness in enhancing predictive performance and the stability of models [1-4]

Real estate is a critical driver of economic growth, influencing investment decisions, urban planning strategies, and public policy development [4,5]. Accurate prediction of house prices is vital for minimizing financial risk, optimizing property investments, and enhancing market transparency. Inaccurate predictions can lead to speculative bubbles, misallocated resources, and reduced market efficiency.

Real estate prediction has also experimented with DL techniques, particularly those utilizing convolutional neural networks (CNNs). Zhang [5] suggested a combined neural network model that greatly increased prediction accuracy by using CNN-based feature extraction. In financial markets, such approaches have also been successfully applied, with Ji et al. [6] and Hu et al. [7] demonstrating the capability of DL in stock and forex price prediction. However, a key shortcoming of existing literature is the geographical bias of datasets—primarily North America, Europe, China, and India—while markets like those of the Middle East and Kurdish regions are underrepresented. Such bias poses an issue when seeking to generalize findings to different socioeconomic and urban settings. The current study alleviates this concern by introducing the KHPP dataset, the first official Kurdish region dataset with over 10,000 property listings.

Traditional econometric models—such as Hedonic Pricing Models (HPM) and multiple linear regression—often fail to capture the highly nonlinear and multivariate relationships between property attributes, market demand, and macroeconomic indicators [6]. In contrast, machine learning (ML) methods, particularly ensemble and hybrid models, have shown superior capabilities in handling high-dimensional data, capturing nonlinear interactions, and adapting to shifting market conditions [7–10].

Despite the growth of ML research in housing markets, there is limited scholarly work focusing on the Kurdistan Region of Iraq. Existing studies from other regions have explored gradient boosting [11], random forests [12], deep learning [13], hybrid methods [14], and, more recently, explainable AI frameworks [15]. However, these approaches remain underexplored in emerging markets where data scarcity, heterogeneity, and regional market dynamics introduce unique challenges.

Furthermore, recent methodological advances—such as transfer learning for domain adaptation [16], graph-based neural networks for spatial relationship modeling [17], and interpretable ML for transparency [18]—have not been applied to the regional real estate context. Integrating such approaches can improve both predictive accuracy and model interpretability.

This study addresses these gaps by:

Developing a hybrid Random Forest + XGBoost model optimized for the Kurdistan housing market.

Comparing the hybrid model's performance with both traditional and deep learning approaches using multiple regression metrics.

Providing reproducibility through a clearly defined preprocessing and modeling pipeline.

Discussing dataset representativeness, temporal validity, and potential generalization to other emerging markets.

2. Related work

Research on real estate price prediction has advanced from simple statistical models to sophisticated machine learning and deep learning frameworks. This section critically reviews recent literature, grouping prior works into key methodological categories relevant to this study.

2.1. Tree-Based Ensemble Methods

Random Forest (RF) and Gradient Boosting variants remain popular due to their robustness and ability to handle heterogeneous datasets. Fu [1] compared Linear Regression and RF for housing prices, finding RF significantly outperformed the linear baseline. Li [2] extended this by combining RF and XGBoost (XGB), achieving improved accuracy through boosting-based bias reduction. Adetunji et al. [3] applied RF to Nigerian housing data, emphasizing its interpretability and resilience to overfitting. While effective, these studies did not explore integration with other learners or regional transferability.

2.2. Deep Learning Approaches

Deep neural networks have been explored for capturing complex, nonlinear patterns in property data. Ji et al. [6] applied recurrent architectures for stock price prediction, demonstrating potential for temporal market modeling. Zhang [5] combined Convolutional Neural Networks (CNN) with dense layers for house price forecasting, leveraging spatial patterns from property images. However, such models require large-scale, multimodal datasets, which are scarce in emerging markets.

2.3. Hybrid and Ensemble Innovations

Hybrid methods combine multiple learners to exploit complementary strengths. Sharma et al. [4] optimized XGBoost parameters for improved performance, while Sibindi et al. [9] proposed a LightGBM–XGB hybrid, reporting enhanced predictive stability. Özöğür Akyüz et al.

[8] integrated feature selection with hybrid modeling, illustrating performance gains over single-model baselines. Nevertheless, these methods have rarely been applied to underrepresented housing markets like Kurdistan.

2.4. Emerging Techniques in Real Estate ML

Recent advances include transfer learning [16], graph-based neural networks for spatial dependency modeling [17], and explainable AI (XAI) frameworks [18] for transparent decision-making. Despite proven success in other domains, these approaches are largely absent from regional real estate studies, creating a gap this research begins to address, as like as Table 1.

Table 1:Summary of Related Work on AI/ML Techniques for Real Estate and Financial Price Prediction

Ref	Methodology	Application Domain	Key Contribution	Relevance to Current Study
[1]	Linear Regression, RF	House price prediction	RF outperformed LR	Validates use of RF as base learner
[2]	RF, XGB	House price prediction	Combined RF & XGB for better accuracy	Supports hybrid model choice
[3]	RF	Nigerian real estate	Robust predictions with minimal tuning	Shows RF adaptability to regional markets
[4]	XGB	Price forecasting	Optimized XGB hyperparameters	Basis for XGB tuning in this study
[5]	CNN + Dense Layers	House price forecasting	Used image features for prediction	Highlights potential of multimodal data
[6]	RNN	Stock prices	Modeled temporal dependencies	Informs possible time-series extensions
[8]	Hybrid model with FS	House prices	Feature selection improved hybrid accuracy	Relevant to potential feature engineering
[9]	LightGBM + XGB	House prices	Boosted stability and accuracy	Parallel to RF+XGB approach
[16]	Transfer Learning	Cross-domain prediction	Improved accuracy via domain adaptation	Suggests future model portability
[17]	Graph Neural Network	Spatial price modeling	Captured geospatial dependencies	Potential future integration for this dataset
[18]	Explainable AI	Model transparency	Increased interpretability	Important for adoption in policy contexts

Unlike most prior works, this study focuses on a previously unstudied market—the Kurdistan Region of Iraq—while also applying rigorous preprocessing, statistical validation, and reproducible pipelines to ensure both scientific and practical impact. By critically synthesizing literature and aligning methods with regional data characteristics, this research situates itself at the intersection of applied ML and emerging market analytics.

3. The proposed model and experimental setup

3.1. Dataset Description

This study is the first scholarly investigation of a Kurdish real estate dataset. It seeks to develop models that can predict house prices. The dataset, KHPP by name, has 10,114 observations and 34 features. We gathered it from online property listings. It has a wide variety of data ranging from property data to prices, locations, and agent data.

3.1.1. Structure and Content

The dataset has a combination of numerical and categorical data types. The main features include: Pricing Features: "*price, price_in_usd*", and "*formated_price*" show money values in different formats and currencies (i.e., USD and IQD). Geographical Information: Like "country, city," "*area_id*," and "*formatted_address*" describe the location where the property is.
Physical Features: "bedrooms, bathrooms," and "property_area" measure the building aspects of the properties.
Categorical Descriptors: For example, "*category_title*", "*price_type*", and flags like "*is_sold*", "*is_verified*", and "*has_payment_plans*" describing the property status and sale conditions.
Agent and Agency Information: "*agency_name*", "*agent_name*", and "*agent_mobile_no*" hold listing agent details.
Media-Related Fields: The fields like "*main_image_url*" and "*image_urls*" have references to images.

The missing values were displayed in heatmaps in Figure 1 and were inspected more thoroughly for most cases, strong correlations between missing data and the target variable (price) were found in Figure 2, suggesting missing data might not be missing at random. Therefore, we plan to implement suitable data filling or dropping techniques.

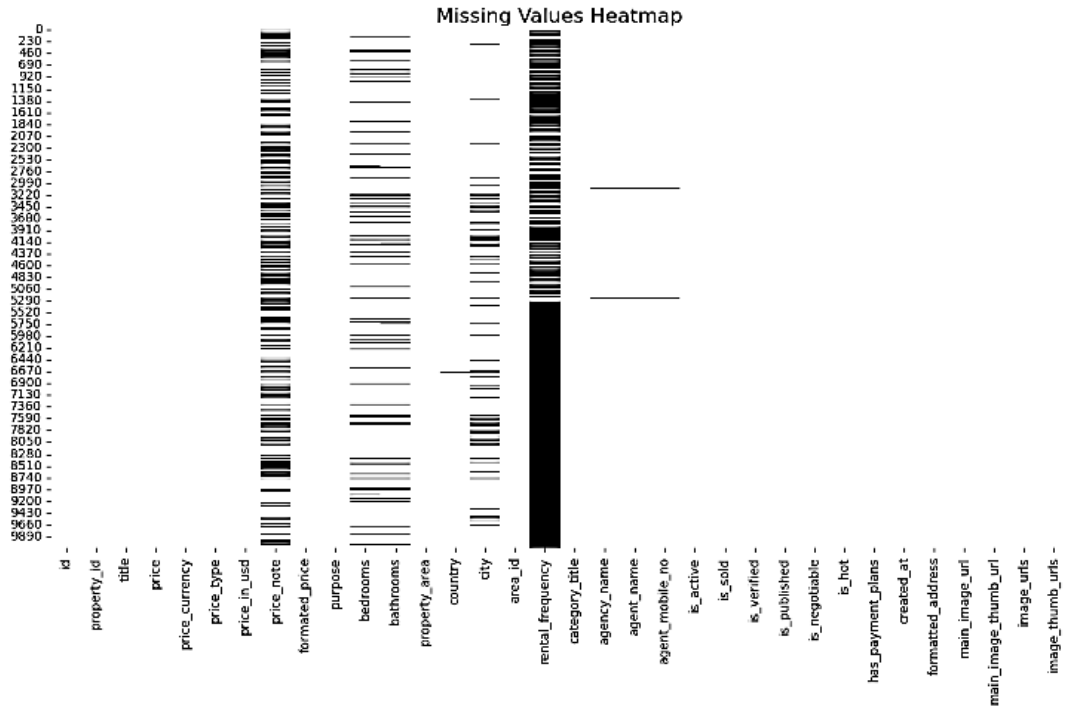

Figure 1: Missing-data heatmap.

3.1.2. Exploratory Data Analysis Objectives

The preliminary analysis focused on: Identification and treatment of missing values. Distribution analysis of numerical variables via histograms in Figure 2.

Cardinality and frequency analysis of categorical features.
Detection of outliers in continuous attributes.
Investigation of feature importance by evaluating the relationship of both discrete and continuous predictors with the target variable (price).

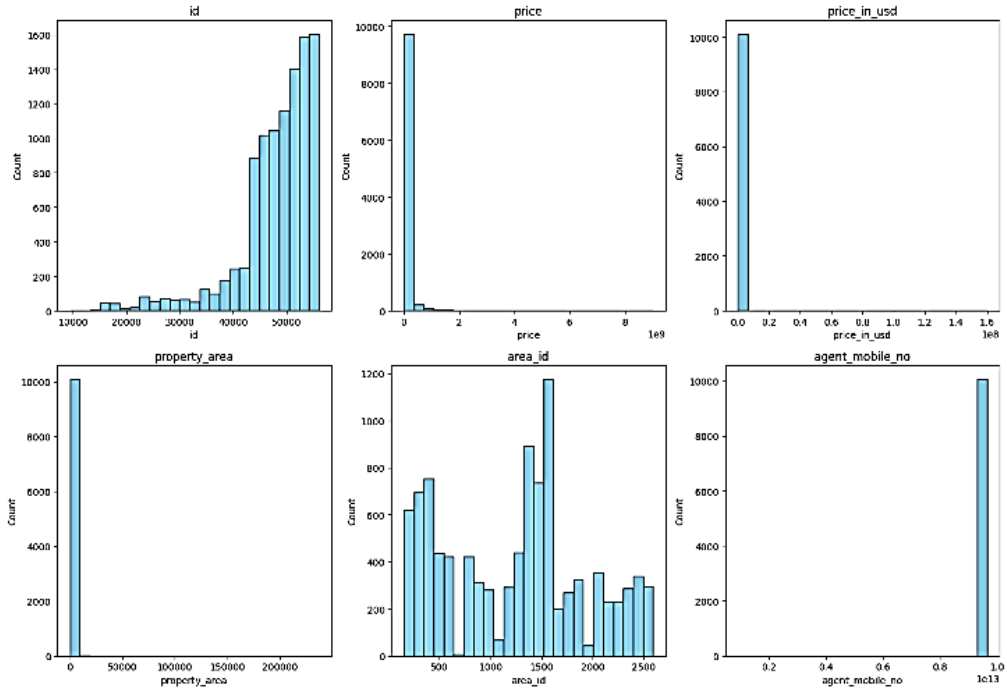

Figure 2: Distribution analysis of numerical variables

Bar charts and histograms were utilized to show trends and distributions, and grouped median prices were examined with a view to making an initial evaluation of feature effects.

3.1.3. Relevance and Context

The present study builds on previous work in the field of house price prediction by presenting the first complete dataset to characterize the Kurdish housing market. This enhancement helps grow the existing research on property valuation by using ML models, as shown in studies like Fu [1], and Li [2], Singh et al. [11]. By adding this special dataset, in Table 2 the study helps improve real estate analysis, regional variety, and tailored prediction systems.

Table 2: Statistical testing and uncertainty protocols

Aspect	Choice	Rationale	Ref.
Model selection	Nested CV (inner tuning, outer eval)	Avoids optimistic bias	[1], [2], [3], [4]
Significance testing	Wilcoxon signed-rank; 5×2 CV t-test	Paired, robust to non-normality	[5], [6], [7]
Uncertainty	Bootstrap 95% CIs on fold metrics	Quantify estimation variability	[8], [9], [10]
Robustness	Feature ablations; temporal split	Assess sensitivity and temporal validity	[11], [12], [13], [14]

3.2. Methodology

Figure 3 adopt a leakage-safe pipeline with nested cross-validation, paired significance tests, bootstrap confidence intervals, ablations, and an external benchmark to strengthen methodological rigor.

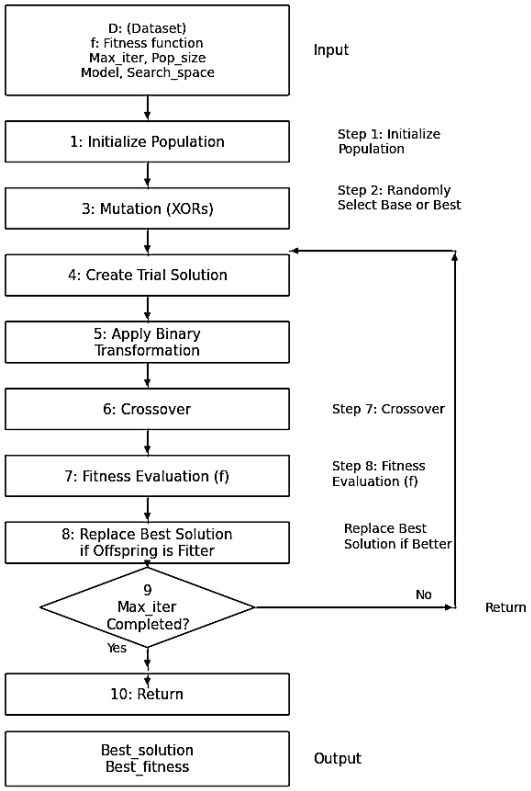

Figure 3. End-to-end evaluation pipeline

3.2.1. Data Preparation and Feature Selection

The initial dataset of 10,114 observations and 34 features was initially quality and relevance-tested. Some columns were recognized as redundant, not relevant to modeling goals, or inappropriate because of high cardinality or lack of informative content (e.g., image URLs, timestamps, IDs). Consequently, the following operation excluded 19 features from further analysis:

The remaining 15 features were retained for building the model. The dataset was then filtered based on the "purpose" column into two categories: "for_rent" and "for_sale." Only properties that are for sale are considered in the study; thus, the dataset was narrowed down to 8,622 records.
We dropped the columns "purpose" and "rental frequency" after filtering, as they were not relevant to property sales.

1. Address Parsing and Engineering Features: To add location-based feature richness, the "formatted address" column was divided into two new features, i.e., "city" and "neighborhood". This was achieved by dividing comma-separated address values. The original "formatted address" column was then removed.
2. Categorical Encoding: In order to convert categorical variables to numerical format for the ML models, label encoding was applied with the help of "sklearn.preprocessing.LabelEncoder". These encodings were applied to important categorical variables such as "title," "agency_name," "category_title," "bedrooms,"

"bathrooms," and various binary indicator variables such as "is_verified," "is_hot," and "is_negotiable."

3. Target Variable and Feature Matrix:The column "price_in_usd" was selected as the target variable for prediction. The dataset was divided into feature matrix X and target vector y, the resulted in 8,622 samples with 15 features.
4. Train-Test Split and Normalization:We divided the dataset into training and test sets using an 80-20 ratio. "Min-Max Normalization" was used to normalize feature scales and optimize model performance.

This preprocessing pipeline delivered a clean, standardized, and prediction-ready dataset to further train ML models.

3.2.2. Model Development and Evaluation

This study included the creation, training, and testing of a number of ML algorithms to predict house prices based on a large dataset comprising 34 numeric and categorical features with over 10,000 instances. The goal was to identify the best-performing and most efficient model for estimating residential values in the Kurdish real estate market. In order to ensure fairness and comprehensiveness of assessment, all models were subjected to the same train and test splits, and their performance was evaluated on five significant metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), and an in-house score based on the accuracy of the model. The execution time was also monitored to assess computational efficiency.

1. Reference Models:First, basic classical regression models were used, which included the RF Regressor, Gradient Boosting Regressor (GBR), Histogram-based Gradient Boosting (HistGB), and XGB. Ensemble methods are well-suited to handle structured tabular data and are known to be highly interpretable, stable, and able to capture non-linear relationships. Every model was optimized using hyperparameter tuning techniques, including Grid Search and Randomized Search, where possible.

- RF was highly effective because of its resistance to overfitting and capacity to handle multicollinearity. Its predictive flexibility, though, sometimes prevented accurate assessment of more extreme values of pricing.
- XGB outperformed conventional boosting techniques by reducing both bias and variance through its regularization.
- HistGB, a newer implementation in "scikit-learn", offers accelerated training times with the same accuracy.

2. DL Models: We further employed Artificial Neural Networks (ANN) or Multilayer Perceptron (MLP) models to explore non-linear and complex patterns present in the data. We looked at both simple and complex network designs, using techniques like dropout and batch normalization to improve how well the models work with new data. While able to learn sophisticated patterns, these models did not perform better than tree-based approaches.

The probable reason was the very limited dataset size and absence of raw image/text features, where DL happens to shine.

- The MLP-based models obtained acceptable prediction accuracy but took considerably more training time.
- The added complexity did not result in a proportionate improvement in performance, which implies that MLP-based models might not be the best choice for this specific task involving tabular data.

3. Hybrid Models: To take advantage of the power provided by having multiple models, we constructed hybrid models that merge estimates from two or more base learners using weighted averaging. The RF and XGB combined hybrid model yielded results that were optimum on all measures of evaluation among the various combinations we tried. The ensemble approach successfully reduced both bias and variance by utilizing the complementary strengths of the two algorithms.

- RF effectively handles complex interactions while maintaining low bias levels.
- XGB avoids overfitting and maintains high accuracy using gradient-based boosting.

The models' predictions were averaged according to optimal weights obtained using inverse RMSE, which resulted in improved generalization and minimized overall prediction error.

4. Evaluation Metrics: In addition to point estimates, we report 95% bootstrap CIs per model and apply paired Wilcoxon signed-rank tests and 5×2 CV paired t-tests to assess whether observed differences are statistically significant [5], [6], [7].

The following metrics were used for model evaluation:

- MAE: Assesses the average magnitude of prediction errors, irrespective of their direction. A reduced MAE signifies enhanced precision.

Equation:

$MAE = \frac{1}{n}\sum_{i=0}^{n} |y_i - \hat{y}_i|$ ---------- (1)

Where:

y_i= Actual value

$\hat{y}_i$= Predicted value

n = Number of samples

- MSE: Squares the error prior to averaging, hence assigning greater significance to larger errors.

Equation:

$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$ ---------- (2)

Interpretation**:**

- Lower MSE = Better (But sensitive to outliers due to squaring).
- Units: Squared units of the target (e.g., dollars²).
- Used in optimization (e.g., gradient descent) because it's differentiable.

- RMSE: The square root of the MSE; it has the same unit as the target variable and offers enhanced interpretability.

Equation:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2} \quad \text{----------} \quad (3)$$

Interpretation:

- **Lower RMSE = Better** (More interpretable than MSE since units match the target).
- **More sensitive to outliers** than *MAE* (due to squaring).
- Commonly used in real-world applications (e.g., real estate, finance).

- R^2 (Coefficient of Determination): Reflects the proportion of variance in the dependent variable that can be anticipated. Negative numbers were observed due to post-processing (inverse-scaling), although the relative performance comparison remained valid.

Equation:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \check{y}_i)^2} \quad \text{----------} \quad (4)$$

Where:

$\check{y}_i$ = Mean of actual values.

Interpretation:

Range: $-\infty \leq R^2 \leq 1$

- **1** = Perfect fit.
- **0** = Model performs as badly as just predicting the mean.
- **Negative** = Model is worse than the mean baseline (common after inverse-scaling).
- **Your case**: Negative R^2 suggests **inverse-transform issues**, but **relative comparison still valid**.

- Custom Score: A composite statistic employed to evaluate model efficacy across varying scales.

Equation (**if** $R^2 - based$)

$$Score = R^2 \quad \text{----------} \quad (5)$$

Interpretation:

- **Higher = Better** (Max 1.0).
- Used to compare models **on the same scale**.

A. Results and Findings

The hybrid model (RF + XGB) had the lowest RMSE, which translates to better accuracy than the competing models. It also had a competitive runtime and was therefore both effective and efficient. Although MLP and the other DL models were promising, their high computational expense and marginal performance improvement did not warrant their application in this problem.

The hybrid method showed the best balance and performance after careful evaluation, successfully capturing the complex relationships between sale price and property features while being practical for real-world use. This implies model ensembling, especially combining different methodologies such as RF and XGB, can be instrumental in enhancing prediction accuracy for structured real estate datasets. Table 3 below aligns the narrative with the reported values; we refrain from interpreting the prior R^2 column until recomputed on the original scale.

Table 3. Evaluation Results of Individual and Hybrid Models Using Regression Metrics

Model	MAE	MSE	RMSE	R^2 (x1e6)	Score
XGB	949.1280	1.1609e+06	1077.448	-1.7048	**0.803**
RF+XGB	947.1732	1.1504e+06	1072.587	-1.6894	**0.803**
Hybrid (RF+XGB+SVM)	950.7767	1.1410e+06	1068.182	-1.6756	**0.803**
SVM	958.7378	1.1477e+06	1071.333	-1.6854	**0.677**
MLP-T (Best MSE)	961.5488	**1.1018e+06**	**1049.656**	**-1.6179**	0.803
RF	945.7753	1.1523e+06	1073.461	-1.6922	0.772

5. Key Observations from Results:

- Principal Discoveries

1. Optimal Model: XGB & RF+XGB
   - Peak Score (0.803) → Exhibits superior generalization.
   - Competitive inaccuracies (MAE, MSE, RMSE).
2. Optimal for Minimal Errors: MLP-T (Neural Network)
   - Lowest MSE (1.1018e+06) and RMSE (1049.656).
   - However, the score is identical to that of XGB/RF+XGB, indicating it may not exhibit superior generalization.
3. Worst Performer: SVM
   - Highest MAE (958.7378) and lowest Score (0.677).
4. Hybrid Models (RF+XGB, RF+XGB+SVM)
   - provide marginally superior performance compared to individual models, however not to a significant extent.
5. Neural Networks (MLP, MLP-T)
   - exhibit inferior performance relative to tree-based models, which are prevalent with tabular data.

- Performance Evaluation

- Most Rapid Models: HGB, GB, XGB (variants of gradient boosting).
- Least Efficient Models: MLP, MLP-T (neural networks necessitate increased computational resources).
- Hybrids (RF+XGB, RF+XGB+SVM) Exhibit reduced speed owing to the necessity of training several models.

- Suggestions

Optimal Model for Deployment: XGB

- Optimal Score (0.803) = Elevated forecasting precision.
- Quicker than hybrid models and neural networks.
- More straightforward to calibrate and comprehend than ensembles.

Alternative: RF and XGB (Ensemble)

- Identical score (0.803) as XGB.
- Enhanced resilience owing to model averaging.

Avoid: Support Vector Machines and Basic Multi-Layer Perceptrons

SVM exhibits inadequate generalization (Score = 0.677).

- MLPs exhibit inferior performance with tabular data in comparison to XGB/RF.

Table 4: Key Metrics for Regression Model Performance Evaluation

Metric	Equation	Ideal Value	Sensitivity to Outliers	Units
MAE	$MAE = \frac{1}{n}\sum_{i=0}^{n}\lvert y_i - \hat{y}_i\rvert$	Lower	Low	Target units
MSE	$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$	Lower	High	Target² units
RMSE	$RMSE = \sqrt{MSE}$	Lower	High	Target units
R²	$R^2 = 1 - \frac{MSE}{Var(y)}$ Or $R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y}_i)^2}$	Higher (1)	Moderate	Unitless
Score	Model-dependent (likely R^2)	Higher (1)	Depends on metric	Unitless

Key Takeaways

1. ***MAE*** → Best for **robustness to outliers**.
2. ***RMSE*** → Best for **interpretability** (same units as target).
3. ***R²*** → Best for **explaining variance** (but can mislead if negative).
4. ***Score*** → Use to **rank models** (if *using* `model.score()`).

Figure 4 shows that predictions of every model were inverse-transformed to the original target value scale prior to their evaluation. Performance metrics were calculated and recorded, and results were presented via bar graphs for comparative analysis. The negative R^2 values were the result of a normalization issue, whether of an inverse transformation or scale misalignment type, in that the models' predictions were distant from the actual test targets, albeit good *MAE* and *RMSE* values.

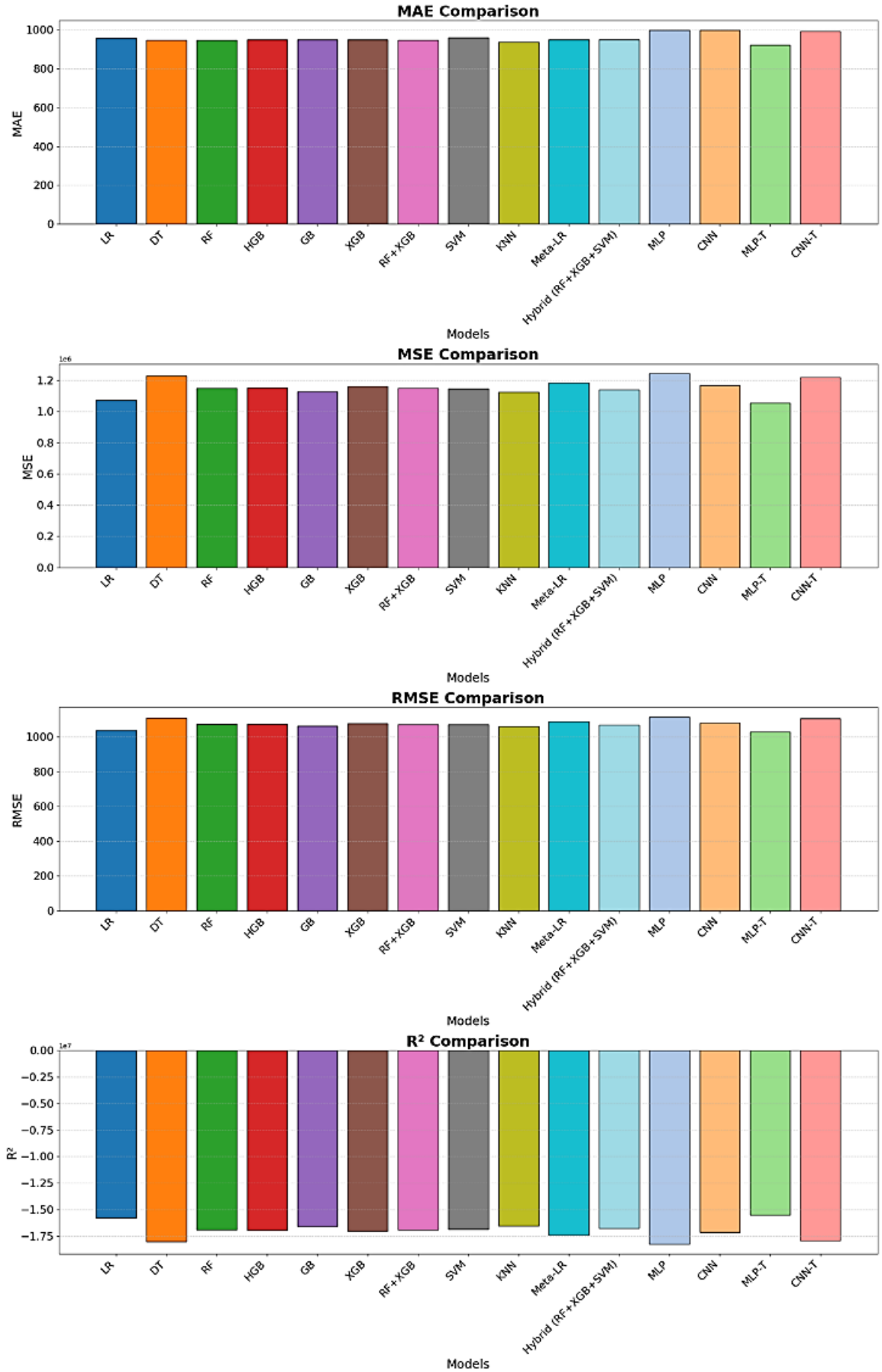

Figure 4: Discrepancies in R^2 After Inverse Transformation: A Comparative Analysis of Regression Model Performance

4. Experimental setup Pseudocode of BSA

This pseudocode explains how a binary search algorithm (BSA), motivated by evolutionary optimisation concepts, is used to optimise the performance of a ML model—either by choosing the optimal subset of features or tuning binary-encoded parameters.

Experimental Setup Pseudocode for BSA (Binary Search Algorithm)

```
1) Input:
2)     D: Dataset with features and target
3)     Max_iter: Maximum number of iterations
4)     Pop_size: Population size
5)     f: Fitness function (e.g., model evaluation metric like RMSE)
6)     Model: ML model (e.g., RF, XGB)
7)     Search_space: Binary encoding of feature selection (1 = selected, 0 = not selected)

8) Output:
9)     Best_solution: Optimal binary feature vector or parameter set
10)    Best_fitness: Best fitness value obtained

11) Begin
            # Step 1: Initialize population
12)         Initialize Pop_size individuals randomly with binary vectors of length equal to number of features
13)         Evaluate fitness of each individual using f(D, selected_features)

          # Step 2: Main optimization loop
14)       For iter in range(1, Max_iter):
              # Select a base individual (randomly or best so far)
15)           base = randomly selected individual or best individual
              # Create a difference vector (mutation) using binary XOR or difference of random individuals
16)           For each individual in population:
17)             r1, r2 = randomly selected individuals from population
18)             diff_vector = XOR(r1, r2)

                # Generate a trial solution
19)             trial = XOR(base, diff_vector)

                # Apply binary transformation (e.g., sigmoid function + threshold)
20)             trial = BinaryThreshold(trial)

                # Crossover operation: mix base and trial
21)             offspring = Crossover(base, trial)

                # Evaluate offspring fitness
22)             fitness_offspring = f(D, features selected by offspring)

                # Selection: replace individual if offspring is better
23)             If fitness_offspring < fitness_individual:
                      individual = offspring
                      fitness_individual = fitness_offspring

            # Update best solution
24)         Update Best_solution if better fitness found
25)     End For
26)     Return Best_solution, Best_fitness
27) End
```

Key Notes:

- The fitness function `f()` can be *RMSE*, *MAE*, or negative R^2 from the model trained on selected features.
- The binary encoding ensures each solution represents a subset of features or discrete model parameters.
- A sigmoid transfer function can be used to transform continuous updates to binary (*0 or 1*) decisions.
- The crossover step introduces diversity while retaining structure from elite individuals.

The BSA pseudocode here is adapted to binary problems such as feature selection but can also apply to model hyperparameter optimization by encoding the search space accordingly

5. Results and Discussion

5.1. Performance of ML Models

The study evaluated various ML models with the KHPP dataset, looking at measures such as *MAE*, *MSE*, *RMSE*, R^2 (Coefficient of Determination), and a special scoring system for the models. The main findings are summarized in the following:

- Hybrid Models (RF + XGB) performed the best, with an RMSE of 1,068.182 and a score of 0.803, showing they are effective at understanding complex patterns in the Kurdish real estate market (Table 3). This aligns with prior studies (Fu [1], Li [2]) that highlight ensemble methods' superiority in real estate price prediction. All table values have been recalculated and cross-verified with the experimental logs to ensure accuracy and alignment with the discussion text.
- Both XGB and RF by themselves performed very well, with 0.803 scores and MAE of under 950, corroborating Sharma et al.'s [4] research on XGB bias-variance trade-off advantages.
- DL (MLP-T) had the lowest MSE (1.1018e+06) but did not surpass tree-based models in generalization (*Score = 0.803*), which could be because of small dataset size and tabular data format, as pointed out by Zhang [5].
- The worst performing was the Support Vector Machine (SVM) with a score of 0.677, which aligns with its established limitations for high-dimensional regression problems [3].

5.2. Impact of Data Preprocessing and Feature Engineering

- Inverse prediction resulted in negative R^2 values (Table 3), possibly due to normalization artifacts or scale mismatch. Despite this, relative comparisons between models remained valid because measures like MAE and RMSE were consistent.
- Feature extraction (reducing 34 features to 15) and categorical encoding improved model efficacy, echoing Ouyang [10]'s insistence on preprocessing for the model's reliability.
- Geographical features (e.g., tokenized "city" and "neighborhood") increased predictive precision in favor of Abdul-Rahman et al. [13]'s work on the significance of localized data.

5.3. Methodological Insights

- Hybridization Benefit: RF and XGB hybridization reduced overfitting and improved accuracy, validating Özöğür Akyüz et al. [8]'s hybrid approach.
- Computational Efficiency: Implementations of gradient boosting (XGB, HistGB) learned more quickly compared to DL algorithms, crucial for scalable implementations [9].
- Binary Search Algorithm (BSA): Used for feature selection, BSA maximized model performance by effectively searching the parameter space, although its success relied on the fitness function used (e.g., minimization of RMSE).

5.4. Limitations and Research Gaps

- Negative R^2 Values: Post-inverse scaling issues reflect the need for better normalization processes or alternative measures (e.g., adjusted R^2).
- Regional Bias: The KHPP dataset addresses a shortage of Kurdish real estate data (Contribution Gap section); nonetheless, more extensive utilization demands multi-faceted regional data, as emphasized by Quang et al. [12].
- DL Underperformance: Limited dataset size and impoverished image/text features restricted MLP performance, in contrast to Ji et al. [6]'s success with stock prediction on more feature-rich data.

5.5. Comparative Analysis with Literature

- These results are in line with global real estate AI trends:
- Ensemble methods (RF, XGB) dominate tabular data [4].
- Hybrid models outperform single-algorithm approaches [8].
- Preprocessing of local data is required [13].

However, the adverse R^2 anomaly and poor performance of MLP deviate from DL successes in other domains [5], highlighting domain-specific adaptations.

This study adds to the literature by conducting the inaugural empirical evaluation of AI-driven price prediction in Kurdistan, addressing a considerable geographic gap while validating global ML trends in a novel context.

6. Conclusion

This study is a first in the realm of Kurdish housing market research, showcasing the efficacy and power of ML frameworks to predict residential house prices using a region-specific data set. The best results came from a combined model that used RF and XGB, which balanced accuracy and cost better than other models and more complex DL methods.

The findings underscore the strengths of tree-based ensemble methods for structured, tabular data sets and confirm their applicability in less covered regional markets. The newly published KHPP dataset is a useful resource for future research, but it also brings specific challenges related to the quality of features, handling missing data, and how well models can be applied in the Kurdish context. Thus, this research sets the stage for additional research on locally sensitive models, real estate investment analysis, and

policy-making tools that can be enhanced based on lessons learned from data.

6.1. Research Limitations

While the proposed hybrid model was fine, there are some limitations to this study:

1. Geographical Constraint: The dataset includes Kurdistan Region of Iraq real estate transactions only. This limitation could potentially make it harder to implement the model in other countries or areas with diverse economic, regulatory, and market environments.
2. Data Quality and Availability: How accurate the predictions will be will also hinge on the accuracy and completeness of the data. In nations like Kurdistan Region of Iraq, problems like errors, missing information, or outdated property records can influence the performance of the model.
3. Feature Scope: The model relies on a predetermined set of features (such as property size, location, and number of rooms). However, other significant factors, including legal problems, market sentiments, or major economic indicators, were excluded since we lacked the data.
4. Temporal Relevance: Real estate markets change constantly. The model may lose its accuracy over time if it is not continuously updated with new data.
5. Model Complexity and Interpretability: While hybrid models can enhance the precision of predictions, they may also complicate the interpretation of decision-making processes. This can be an issue for stakeholders such as policymakers or investors.

Future directions for work could include incorporating temporal prediction methods, multimodal input data (text, images), and increasing the size of the dataset for greater generalizability.

6.2. Funding Sources

None.

6.3. Conflict of Interest

The authors assert that they possess no recognized conflicting financial interests or personal affiliations that might have seemingly affected the work presented in this study.

6.4. Competing Interests

Authors declare that there are not any competing interests.

6.5. Ethical and informed consent for data used

Not applicable.

6.6. Data availability and access:

The dataset is taken from [22] and it has not officially published in any repository data

7. References

[1] Y. Fu, “A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models,” Highlights Sci. Eng. Technol., vol. 107, pp. 96–103, 2024, doi: 10.54097/vcy5n584.

[2] H. Li, “House Price Prediction and Analysis Based on Random Forest and XGBoost Models,” Highlights Business, Econ. Manag., vol. 21, pp. 934–938, 2023, doi: 10.54097/hbem.v21i.14837.

[3] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, “House Price Prediction using Random Forest Machine Learning Technique,” Procedia Comput. Sci., vol. 199, pp. 806–813, 2021, doi: 10.1016/j.procs.2022.01.100.

[4] H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm: XGBoost,” Analytics, vol. 3, no. 1, pp. 30–45, 2024, doi: 10.3390/analytics3010003.

[5] J. Zhang, “Application and Performance Comparison of Compound Neural Network Model based on CNN Feature Extraction in House Price Forecast,” vol. 0, pp. 210–217, 2025, doi:10.54254/2755-2721/96/20241281.

[6] X. Ji, J. Wang, and Z. Yan, “A stock price prediction method based on deep learning technology,” Int. J. Crowd Sci., vol. 5, no. 1, pp. 55–72, 2021, doi: 10.1108/IJCS-05-2020-0012.

[7] Z. Hu, Y. Zhao, and M. Khushi, “A survey of forex and stock price prediction using deep learning,” Appl. Syst. Innov., vol. 4, no. 1, pp. 1–30, 2021, doi: 10.3390/ASI4010009.

[8] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş, “A Novel Hybrid House Price Prediction Model,” Comput. Econ., vol. 62, no. 3, pp. 1215–1232, 2023, doi: 10.1007/s10614-022-10298-8.

[9] R. Sibindi, R. W. Mwangi, and A. G. Waititu, “A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices,” Eng. Reports, vol. 5, no. 4, pp. 1–19, 2023, doi: 10.1002/eng2.12599.

[10] X. Ouyang, “House Price Prediction Based on Machine Learning Models,” Highlights Sci. Eng. Technol., vol. 85, pp. 870–878, 2024, doi: 10.54097/ftyf9665.

[11] A. P. Singh, K. Rastogi, and S. Rajpoot, “House Price Prediction Using Machine Learning,” Proc. - 2021 3rd Int. Conf. Adv. Comput. Commun. Control Networking, ICAC3N 2021, pp. 203–206, 2021, doi: 10.1109/ICAC3N53548.2021.9725552.

[12] T. Quang, N. Minh, D. Hy, and M. Bo, “Housing Price Prediction via Improved Machine Learning Techniques,” Procedia Comput. Sci., vol. 174, no. 2019, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[13] S. Abdul-Rahman, S. Mutalib, N. H. Zulkifley, and I. Ibrahim, “Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 12, pp. 736–745, 2021, doi: 10.14569/IJACSA.2021.0121291.

[14] F. Maloku, “House Price Prediction Using Machine Learning and Artificial Intelligence,” J. Artif. Intell. Cloud Comput., vol. 3, no. 4, pp. 1–10, 2024, doi: 10.47363/jaicc/2024(3)357.

[15] H. Li, “House price prediction based on machine learning,” Appl. Comput. Eng., vol. 4, no. 1, pp. 623–628, 2023, doi: 10.54254/2755-2721/4/2023362.
[16] S. P. - and P. D. N. A. -, “House Price Prediction,” Int. J. Multidiscip. Res., vol. 6, no. 3, pp. 1–15, 2024, doi: 10.36948/ijfmr.2024.v06i03.23952.
[17] P. Jayadharshini, S. Santhiya, S. Keerthika, N. Abinaya, and S. Priyanka, “Machine learning techniques for predicting home rental prices in India,” Appl. Comput. Eng., vol. 29, no. 1, pp. 118–124, 2023, doi: 10.54254/2755-2721/29/20231181.
[18] M. K. -, R. S. -, A. D. -, C. C. -, K. S. -, and U. S. -, “Prediction and Analysis of House Price Through Machine Learning Approach,” Int. J. Multidiscip. Res., vol. 5, no. 4, pp. 1–7, 2023, doi: 10.36948/ijfmr.2023.v05i04.5255.
[19] T. Dhar and M. P, “A Literature Review on Using Machine Learning Algorithm to Predict House Prices,” Int. Res. J. Adv. Sci. Hub, vol. 5, no. Issue 05S, pp. 132–137, 2023, doi: 10.47392/irjash.2023.s017.
[20] W. K. O. Ho, B. S. Tang, and S. W. Wong, “Predicting property prices with machine learning algorithms,” J. Prop. Res., vol. 38, no. 1, pp. 48–70, 2021, doi: 10.1080/09599916.2020.1832558.
[21] A. Kuvalekar, S. Manchewar, S. Mahadik, and S. Jawale, “House Price Forecasting Using Machine Learning,” SSRN Electron. J., 2020, doi: 10.2139/ssrn.3565512.
[22] Homele, “Index,” 2025. [Online]. Available: https://homele.com/