Real Estate Predictive Analytics using Multi-Model Regression and Classification Pipeline with Recurrent Learning

B. Ramesh[1]*, Addetla Kavya[1], Bonam Udayasree[1], J Praveen Kumar[1]

[1]Department of Computer Science and Engineering (DS), Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: B. Ramesh

To Cite this Article

B. Ramesh, Addetla Kavya, Bonam Udayasree, J Praveen Kumar, "Real Estate Predictive Analytics using Multi-Model Regression and Classification Pipeline with Recurrent Learning", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 806-818, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp806-818

Submitted: 09-03-2026 Accepted: 16-04-2026 Published: 22-04-2026

ABSTRACT

Real-estate valuation is essential for buyers, sellers, and investors, yet traditional approaches depend heavily on manual inspection, subjective judgment, and limited comparable data. Existing automated methods often frame valuation solely as a regression problem, overlooking classification aspects and failing to address imbalanced datasets. Additionally, conventional models struggle with non-linear relationships and high-dimensional features, leading to poor generalization and inconsistent pricing. This work proposes an integrated machine learning framework that combines preprocessing, Synthetic Minority Over-sampling Technique (SMOTE)–based class balancing, exploratory data analysis, and multi-model training for both regression and classification tasks. The regression module predicts key numerical attributes such as property price (in lakhs) and built-up area (in square feet), ensuring accurate continuous-value estimation across diverse property types. The classification module identifies property orientation into facing categories—North (N), South (S), East (E), and West(W)—important for design, ventilation, sunlight exposure, and valuation. The framework employs three model families: Gradient Boosting Classification and Regression Trees (Gradient Boosting CART), Extreme Gradient Boosting CART (XGBoost-CART), and a hybrid Long Short-Term Memory–TreeNet (LSTM–TreeNet) architecture. The hybrid model captures deep temporal and feature-level patterns from structured and unstructured data, followed by refinement through CART-based decision structures. Results show improved prediction accuracy, enhanced classification precision, reduced human bias, and greater consistency in valuation. The proposed system offers a scalable, data-driven solution for reliable and automated real-estate analytics.

Keywords: Real-estate valuation, property pricing, data preprocessing, SMOTE, exploratory data analysis, regression analysis, classification, imbalanced data, feature engineering, automated valuation system.

1. INTRODUCTION

The real estate sector, as a fundamental pillar of the national economy of India, is integral to macroeconomic stability and sustainable development as shown in Figure 1. Further, India's real estate market has exhibited significant volatility over the past decade, characterized by pronounced disparities among first-, second-, and third-tier cities, thereby increasing the complexity of regulatory interventions [1]. To address these challenges, the Ministry of Housing and Urban–Rural Development has

ISSN:3049-0952
www.jsetms.com

underscored the necessity for cities, particularly first-tier cities, to enhance regulatory autonomy in managing the real estate market and adjust housing purchase restrictions in accordance with local conditions. In first-tier cities, factors such as high population mobility, constrained land resources, and frequent policy interventions have contributed to multiple cycles of housing price surges since the housing reform [2]. These fluctuations have not only influenced the overall real estate market but have also had broader implications for national monetary policy [3]. The impact of first-tier city housing market dynamics is manifested through spillover effects, demonstration effects, and structural market adjustments. The spillover effect arises as elevated housing costs in first-tier cities drive homebuyers to adjacent regions, subsequently exerting upward pressure on property prices in those areas [4].

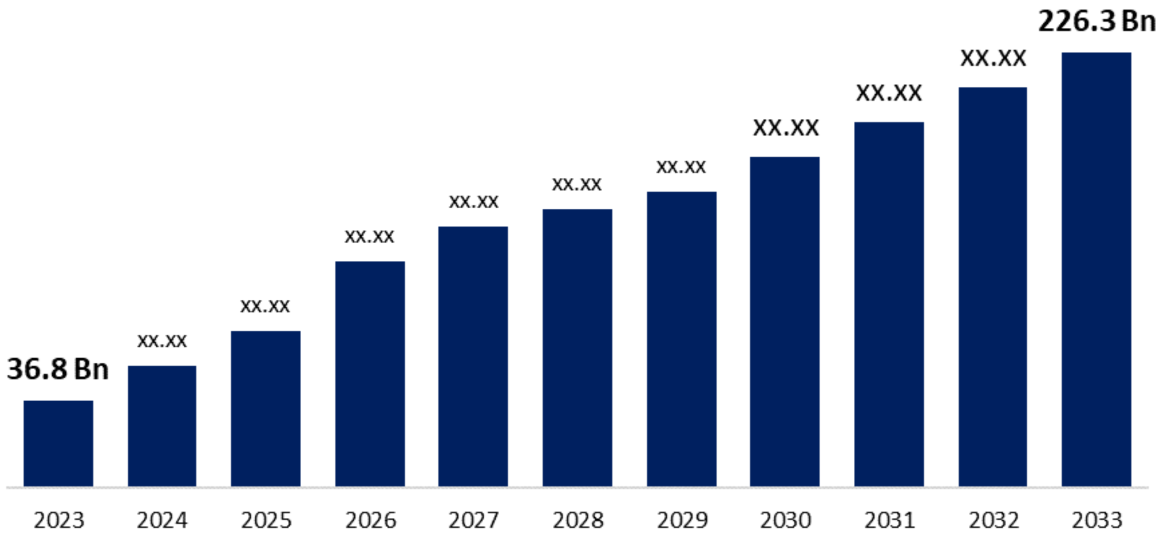

Figure 1: Indian Housing Markets Statistics.

The demonstration effect occurs when rising housing prices in first-tier cities are perceived as indicative of national market trends, reinforcing buyer confidence and precipitating price increases in other cities, particularly strong second-tier cities, at times leading to market overvaluation [5]. Additionally, sustained price appreciation in first-tier cities may prompt real estate firms to adjust their investment strategies, prioritizing high-return urban centers, while third- and fourth-tier cities experience stagnation or decline in housing demand and pricing.

Moreover, fluctuations in first tier-city housing prices may impose “passive constraints” on national monetary policy, compelling the central bank to navigate trade-offs between financial stability in the real estate sector and broader economic growth objectives [6]. This dynamic may, under certain circumstances, result in deviations from an optimal monetary policy trajectory. For instance, India’s economy was undergoing deepening supply-side structural reforms, and the slowing macroeconomic growth theoretically warranted a more accommodative monetary policy to support the real economy [7].

2. LITERATURE SURVEY

Andrade-Giron, et al. [8] examined the effects of dimensionality reduction through Recursive Feature Elimination (RFE), Random Forest (RF), and Boruta on real estate price prediction, assessing ensemble models like Bagging, Random Forest, GB, AdaBoost, Stacking, Voting, and Extra Trees. The results indicate that the Stacking model achieved the best performance with an MAE (mean absolute error) of 14,090, MSE (mean squared error) of 5.338×10^8, RMSE (root mean square error) of 23,100, R^2 of 0.924, and a Concordance Correlation Coefficient (CCC) of 0.960, also demonstrating notable computational efficiency with a time of 67.23 s. GB closely followed, with an MAE of 14,540, R^2 of 0.920, and a CCC of 0.958, requiring 1.76 s for computation. Variable reduction through RFE in both GB and Stacking led to an increase in MAE by 16.9% and 14.6%, respectively, along with slight reductions in R^2 and CCC. The application of Boruta reduced the variables to 16, maintaining performance in Stacking, with an increase in MAE of 9.8% and a R^2 of 0.908.

Song, et al. [9] developed various machine learning methods and interpretability methods like SHAP values are used to explore the impact of supply, demand, policies, and expectations on the real estate market of China's first-tier cities. In terms of commercial housing sales area, adequate housing supply, robust medical services, and high population density boost the sales area, while demand for small units reflects buyers' balance between affordability and education. In terms of commercial housing average sales price, growth is driven by education investment, population density, and income, with loan interest rates serving as a stabilizing tool.

Maselli, et al. [10] proposed a methodological approach that addresses both these issues, comparing the predictive performance of three ML algorithms k-Nearest Neighbors (kNN), Random Forest (RF), and the Artificial Neural Network (ANN) applied to the housing market in the city of Salerno, Italy. For each model, overfitting is preliminarily assessed to ensure predictive robustness. Subsequently, the results are interpreted using explainability techniques, such as Shapley Additive explanations (SHAPs) and Permutation Feature Importance (PFI). This analysis reveals that the Random Forest offers the best balance between predictive accuracy and transparency, with features such as area and proximity to the train station identified as the main drivers of property prices. kNN and the ANN are viable alternatives that are particularly robust in terms of generalization.

Anelli, et al. [11] analysed how normalization techniques influence the outcomes of real estate price regression models using machine learning to uncover complex relationships between urban and economic factors. Six normalization techniques are employed to assess how they affect the estimation of relationships between property value and factors like social degradation, resident population, per capita income, green spaces, building conditions, and degraded neighborhood presence. The study's findings underscore the pivotal role of normalization in shaping the perception of variables, accentuating critical thresholds, or distorting anticipated functional relationships. The work is the first application of a methodological approach to define the best technique based on two criteria: statistical reliability and empirical evidence of the functional relationships obtainable with each standardization technique. Notably, the study underscores the potential of machine-learning-based regression to circumvent the limitations of conventional models, thereby yielding more robust and interpretable results.

Battisti, et al, [12] examined the housing market's impact, focusing on how residential affordability affects residential choices, using a case study of the Metropolitan City of Florence. The analysed employs a methodology centered on the Debt-to-Income Ratio (DTI), which cross-references real estate market values with household income brackets to identify affordable areas. The results reveal a clear divide: households with incomes below EUR 26,000 per year (representing about 69% of the population) are excluded from the central urban property market. This evidence confirms regional and national trends, emphasizing a growing mismatch between housing costs and disposable incomes.

Al-Rimawi, et al. [13] aimed to determine the added value that smart city technologies contribute to real estate development one of the key drivers in transforming traditional properties into smart real estate. A total of 16 technologies applied at both the building and urban scales were examined. Using an integrative review methodology, the research evaluated 168 publications. The metadata analysis revealed the current state of each technology in terms of its added value and level of adoption across both scales. In total, 131 distinct added values were identified and systematically categorized according to the sub-phases and processes of the real estate life cycle. The study also highlighted the benefits generated through the integration of these technologies. The findings demonstrate that these technologies have reached sufficient maturity for practical implementation. Consequently, real estate developers, planners, city managers, and industry experts are encouraged to prioritize their adoption to enhance future development practices.

Di Liddo, et al. [14] aimed to empirically assess whether, and to what extent, real estate market dynamics specifically price levels and market vibrancy are influenced by the quality of life within a given reference area. To this end, the analysis compared key parameters of the residential real estate market, including the Real Estate Market Observatory quotations and the real estate market intensity index (used as a proxy for market dynamism), with the Life Quality Index developed by the research centre of the Italian newspaper Il Sole 24 Ore for a selected group of provincial capitals. By further deconstructing the Life Quality Index into the individual indicators used in its formulation, the study identified which dimensions exhibit the strongest associations with real estate market behaviour. This approach made it possible to pinpoint the specific life-quality components most closely linked to market mechanisms and to analyse how these relationships manifest within each urban context.

Çılgın, et al. [15] presented a model architecture that can achieve high accuracy in predicting the current market value of real estates by using a hybrid approach, through clustering models as a preliminary approach, to achieve higher homogeneity with stacking ensemble using multiple machine learning methods. To obtain more homogeneous submarkets, the collected data set was first grouped according to the number of rooms and then each group was divided into clusters by cluster analysis. In this way, more homogeneous submarkets were obtained and predict accuracy was improved. Then, the training process was carried out for 13 different weak learners using fivefold cross-validation for each determined sub-market.

Bastos, et al. [16] analysed the uncertainty in automated property valuations using conformal prediction, a distribution-free procedure for constructing prediction intervals with valid coverage in finite samples. Through an empirical study of property prices in the San Francisco Bay Area, they find that prediction intervals obtained using conformal quantile regression have exact coverage. In contrast, prediction intervals obtained from non-conformal quantile regressions severely undercover the data. They show that the intervals adapt to various characteristics of the dwellings, which is crucial given the heterogeneous nature of real estate data. Indeed, they observed that larger and older properties, those in both low and high-income neighbourhoods, as well as those on the market for less than one year are more challenging to evaluate.

3. PROPOSED SYSTEM

The system architecture as shown in Figure 2 represents an automated real-estate valuation framework designed to predict both numerical and categorical property attributes through an integrated machine learning pipeline. The architecture begins with a user interaction layer where property-related information is provided through a web-based interface or batch input files. The collected data then passes through a structured processing pipeline that includes data preprocessing, class balancing, exploratory analysis, and model training stages. Multiple machine learning models are trained simultaneously to perform both regression and classification tasks. The regression module estimates continuous variables such as property price and built-up area, while the classification module determines the directional orientation of the property. The architecture ensures efficient data flow from input acquisition to prediction generation, enabling accurate, scalable, and automated real-estate analytics.

Input Interface and Data Acquisition: The first stage of the architecture focuses on collecting property-related information through a user-friendly input interface implemented using the Flask web framework. Users can provide property attributes such as location, built-up area, number of bedrooms, amenities, and other relevant structural features. The system supports two modes of data input: individual property entry through a web form and batch input through CSV file uploads. The batch processing capability allows real-estate analysts or agencies to evaluate multiple properties simultaneously. Once the data is submitted, it is transferred to the backend processing module for validation and further analytical operations.

Data Preprocessing and Feature Engineering: In the second stage, the collected data undergoes preprocessing to ensure quality, consistency, and suitability for machine learning models. This step includes handling missing values, removing inconsistencies, normalizing numerical attributes, and encoding categorical variables such as location identifiers or property categories. Feature engineering techniques are also applied to transform raw input attributes into meaningful representations that enhance model learning. For example, composite indicators may be generated by combining multiple structural or geographical attributes, enabling the models to better capture relationships among property features.

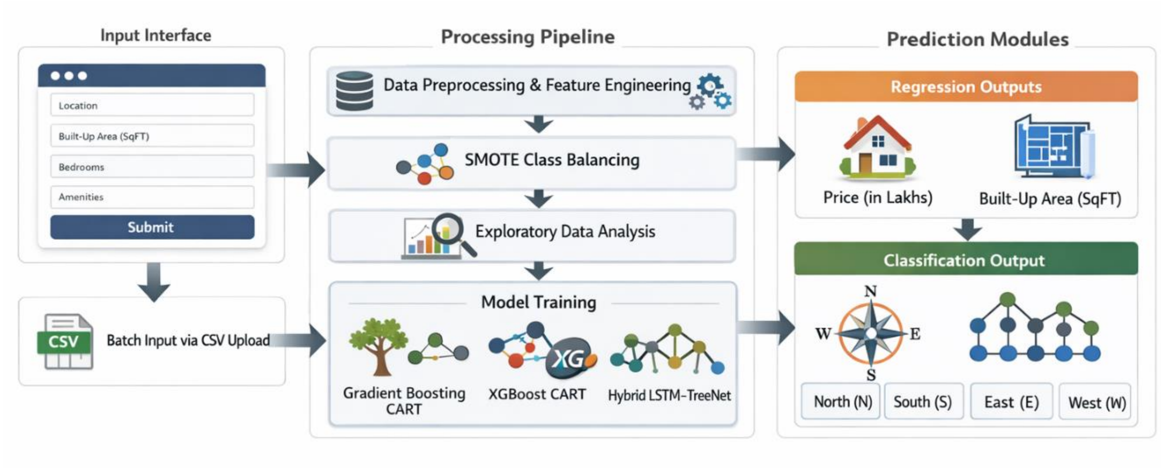

Figure 2: System architecture for real estate price prediction.

SMOTE-Based Class Balancing: The third stage addresses the issue of class imbalance in the dataset, particularly for categorical attributes such as property orientation (North, South, East, and West). In many real-estate datasets, certain directional categories may appear less frequently, which can bias the classification model. To overcome this limitation, the SMOTE is applied to the training data. SMOTE generates synthetic samples for minority classes by interpolating between existing observations, thereby balancing the dataset and improving the classifier's ability to learn representative patterns for all categories.

Exploratory Data Analysis: After balancing the dataset, EDA is performed to understand the statistical characteristics and relationships among property attributes. This step involves analyzing distributions, correlations, and feature interactions that influence property pricing and structural attributes. Visualization techniques such as histograms, scatter plots, and correlation matrices help reveal important patterns, such as how built-up area affects price or how certain property features correlate with directional orientation. The insights gained during EDA guide model training and feature selection decisions.

Model Training Using Existing Algorithms: The fifth stage involves training existing machine learning algorithms that serve as baseline models for the valuation system. Two major ensemble learning approaches are implemented: Gradient Boosting CART and XGBoost-CART. These algorithms can handle structured tabular data effectively and are widely used in real-estate analytics. Separate models are trained for regression tasks such as predicting property price and built-up area, and for classification tasks such as predicting property orientation. Performance metrics are calculated to evaluate the accuracy and reliability of each algorithm.

Proposed Hybrid LSTM–TreeNet Model Training: The sixth stage introduces the proposed hybrid LSTM–TreeNet model designed to improve prediction performance. In this architecture, LSTM networks are used to learn complex feature relationships and sequential dependencies present in the dataset. The learned representations are then passed to a TreeNet-based decision structure that refines predictions using CART-style decision trees. This hybrid approach combines the deep representation learning ability of neural networks with the interpretability and structured decision-making capability of tree-based models. As a result, the model achieves improved accuracy and robustness for both regression and classification tasks.

Regression Prediction Module: The regression module generates numerical predictions related to real-estate properties. Specifically, it estimates the property price in lakhs of rupees and predicts the built-up area measured in square feet. These predictions are essential for buyers, sellers, and investors to assess property value and compare listings in the market. The regression models analyze relationships among structural features, location characteristics, and other property attributes to produce precise numerical outputs.

Classification Prediction Module: The final stage of the architecture performs categorical prediction of property orientation. The trained classification models categorize the property facing direction into one of four classes: North (N), South (S), East (E), or West (W). Directional orientation is an important factor in property valuation because it influences sunlight exposure, ventilation, and architectural design preferences in many regions. The classification output provides additional insight into the property characteristics, enhancing the overall usefulness of the valuation framework.

4. RESULTS ANALYSIS

This section presents the graphical user interface and visual outputs of the Real Estate Sale Price Prediction System. The figures illustrate the end-to-end workflow of the application, including dataset upload, preprocessing, exploratory data analysis, model evaluation, and prediction generation. Each interface component is designed to support efficient user interaction and systematic analysis of real-estate data. The figures collectively demonstrate the functionality, usability, and output presentation of the proposed system.

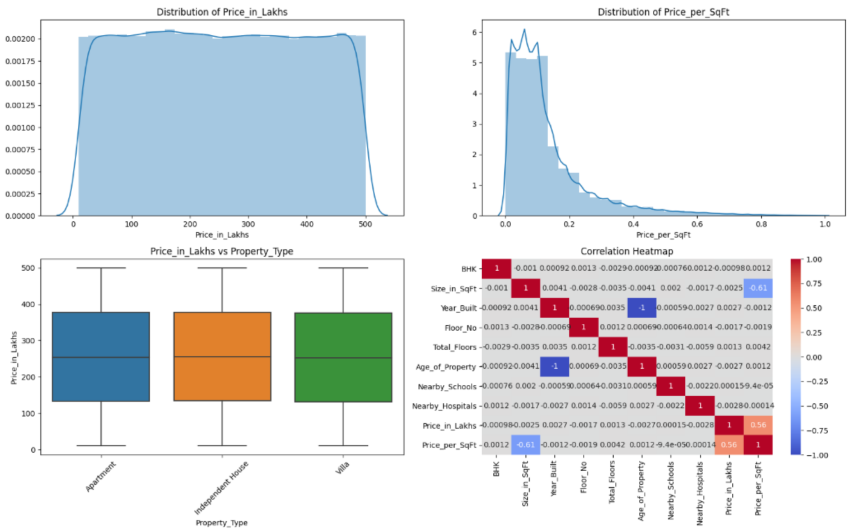

Figure 3: EDA Results Interface.

Figure 3 presents the EDA results, where the system visualizes key statistical patterns and relationships within the dataset. The screen displays multiple plots generated automatically from the uploaded data, including the distribution of Price_in_Lakhs, the distribution of Price_per_SqFt, a boxplot showing price variations across different property types, and a correlation heatmap illustrating the strength of relationships among numerical features. These visualizations help users quickly understand data spread, skewness, feature behaviour, and potential dependencies important for model training. The interface provides a clear and organized layout, enabling users to interpret dataset characteristics without manual plotting.

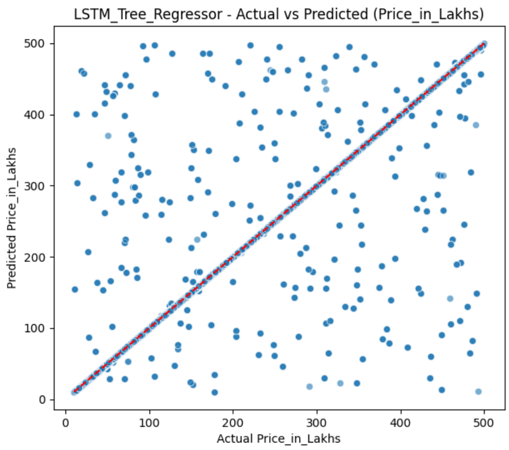

Figure 4. Scatter Plots of Price in Lakhs Target of Various Regressors - LSTM-TreeNet.

The scatter plot for the LSTM–TreeNet Regressor presented in Figure 4 demonstrates a strong alignment between predicted values and the ideal diagonal reference line. Most of the predicted points follow the diagonal trend closely, indicating that the model accurately captures the relationship between input features and property price. This improved performance is attributed to the hybrid architecture, where the LSTM network extracts deep feature representations from the dataset, and the decision-tree-based regression model refines the final predictions. As a result, the LSTM–TreeNet model provides more consistent and accurate predictions across both low and high price ranges, demonstrating superior regression performance compared to the other models.

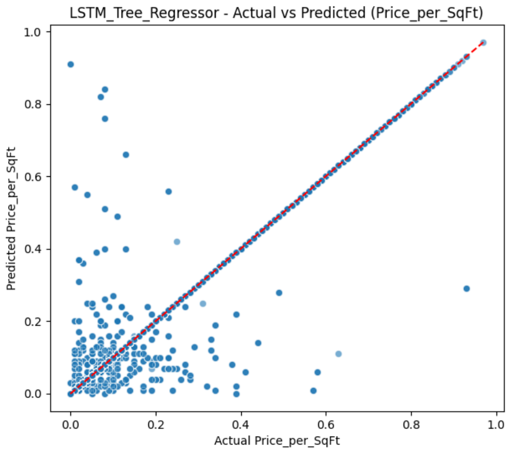

Figure 5 Scatter Plots of Price Per SQFT Target of Various Regressors - LSTM-TreeNet.

Figure 5 demonstrates predictions that closely align with the diagonal reference line across the entire range from 0.0 to 1.0, indicating a strong correspondence between actual and predicted values. This alignment shows that the hybrid model effectively captures the underlying relationships between property features and price per square foot, producing more accurate and consistent predictions compared to the traditional Gradient Boosting and XGBoost models.

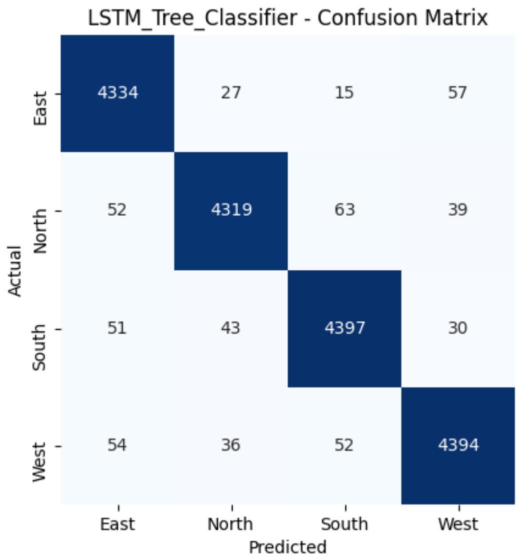

Figure 6: Confusion Matrices of Various Classifiers - LSTM-TreeNet.

In contrast, the LSTM–TreeNet classifier (Figure 6) demonstrates significantly improved classification performance, with very high correct predictions along the diagonal such as 4334 (East), 4319 (North), 4397 (South), and 4394 (West), and only minimal misclassifications including 27 East predicted as North, 63 North predicted as South, 43 South predicted as North, and 52 West predicted as South. These results indicate that the proposed hybrid LSTM–TreeNet model achieves substantially higher classification accuracy and better class separation compared to the traditional Gradient Boosting and XGBoost classifiers.

t_Accessibility	Parking_Space	Security	Amenities	Facing	Owner_Type	Availability_Status	Price_in_Lakhs	Price_per_SqFt	Facing
h	No	No	Playground, Gym, Garden, Pool, Clubhouse	West	Owner	Ready_to_Move	489.76	0.10	East
w	No	Yes	Playground, Clubhouse, Pool, Gym, Garden	North	Builder	Under_Construction	195.52	0.08	East
w	Yes	No	Clubhouse, Pool, Playground, Gym	South	Broker	Ready_to_Move	183.79	0.05	East

Figure 7: Prediction Results from Test CSV File.

Figure 7 shows the Prediction Results interface, where the system displays the predicted outputs for the uploaded test dataset. The page presents a clean, tabular view of the input features such as State, City, Locality, Property_Type, BHK, Size_in_SqFt, Year_Built, furnished_Status, Floor_No, and Total_Floors, along with the model-generated predictions appended to the dataset. At the top of the page, the user is given the option to download the complete prediction results in CSV format or initiate another prediction cycle. This screen allows users to verify sample predictions directly before exporting the full results.

4.1 Comparative Analysis

Table 1 presents the comparative performance evaluation of different classification methodologies used for predicting the property facing direction, including the GB Classifier, XGBoost Classifier, and the proposed LSTM–TreeNet Classifier.

The Gradient Boosting (GB) Classifier achieved an accuracy of 0.2794, with precision and recall values also equal to 0.2794, and an F1-score of 0.2786. These results indicate that the model struggles to accurately classify the property facing directions across the dataset. The relatively low performance suggests that the GB classifier is unable to effectively capture the complex relationships between property features and directional orientation. Additionally, the near-equal values of precision and recall indicate that the model performs uniformly poorly across different classes, likely due to overlapping feature patterns and the limited capability of shallow boosting structures in handling complex feature interactions present in real-estate datasets.

The XGBoost Classifier demonstrates slightly improved performance compared to the Gradient Boosting model, achieving an accuracy of 0.2880, precision of 0.2881, recall of 0.2880, and an F1-score of 0.2872. XGBoost incorporates advanced gradient boosting techniques such as regularization and second-order gradient optimization, which typically enhance prediction accuracy. However, despite these improvements, the classification performance remains relatively low. This indicates that while XGBoost captures some additional patterns within the dataset, it still struggles to effectively distinguish between the four directional classes of property orientation. The results suggest that traditional tree-based boosting algorithms alone may not be sufficient to model the complex nonlinear dependencies within the dataset.

In contrast, the proposed LSTM–TreeNet Classifier significantly outperforms the other models, achieving an accuracy, precision, recall, and F1-score of 0.9711. These results demonstrate a dramatic improvement in classification capability compared to the traditional ensemble models. The high and consistent values across all evaluation metrics indicate that the model not only predicts the correct classes with high accuracy but also maintains balanced performance across all property orientation categories. This superior performance is attributed to the hybrid architecture, where the LSTM network extracts deep feature representations from the structured property dataset, capturing complex relationships among variables, and the decision-tree-based classifier refines the final predictions using hierarchical decision boundaries. As a result, the LSTM–TreeNet classifier effectively learns both latent feature patterns and interpretable decision rules, leading to highly accurate and reliable classification outcomes.

Table 1: Classification Model Performance Comparison of Various Methodologies.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
GB Classifier	0.2794	0.2794	0.2794	0.2786
XGBoost Classifier	0.2880	0.2881	0.2880	0.2872
LSTM–TreeNet Classifier	0.9711	0.9711	0.9711	0.9711

Table 2 presents the comparative performance analysis of different regression models used for predicting property price in lakhs, including the GB Regressor, XGBoost Regressor, and the proposed LSTM–TreeNet Regressor. The evaluation metrics considered in this comparison are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score, which collectively measure the accuracy and reliability of the predicted price values. The Gradient Boosting Regressor achieved an MAE of 121.4966, MSE of 19729.6618, RMSE of 140.4623, and a very low R^2 score of 0.0018, indicating that the model is unable to effectively capture the relationship between input features and property prices, resulting in large prediction errors and almost negligible explanatory power.

Similarly, the XGBoost Regressor produced an MAE of 169.3848, MSE of 42885.8985, and RMSE of 207.0891, along with a negative R^2 score of −1.1698, which indicates that the model performs worse than a simple mean-based baseline predictor, demonstrating significant difficulty in modeling the complex patterns within the dataset. In contrast, the proposed LSTM–TreeNet Regressor significantly outperforms the other regression models, achieving a much lower MAE of 6.5626, MSE of 1608.3563, and RMSE of 40.1043, along with a high R^2 score of 0.9186. These results indicate that the hybrid model accurately predicts property prices with substantially reduced prediction errors and strong explanatory capability. The improved performance can be attributed to the integration of LSTM-based deep feature extraction, which captures complex nonlinear relationships among property attributes, and the decision-tree regression mechanism, which refines the final predictions through structured decision boundaries. Consequently, the LSTM–TreeNet model demonstrates superior regression performance and provides more reliable price estimation compared to traditional boosting-based regression approaches.

Table.2: Price in Lakhs Target Performance Comparison of Various Regression Models.

Model	**MAE**	**MSE**	**RMSE**	**R^2 Score**
GB Regressor	121.4966	19729.6618	140.4623	0.0018
XGBoost Regressor	169.3848	42885.8985	207.0891	-1.1698

LSTM–TreeNet Regressor	6.5626	1608.3563	40.1043	0.9186

Table 3 presents the comparative performance evaluation of different regression models used for predicting the Price per Square Foot of properties, including the GB Regressor, XGBoost Regressor, and the proposed LSTM–TreeNet Regressor. The evaluation metrics used for comparison, which collectively measure the accuracy, error magnitude, and predictive capability of each regression model. The Gradient Boosting Regressor achieved an MAE of 0.0667, MSE of 0.0093, and RMSE of 0.0966, with an R^2 score of 0.4710, indicating moderate prediction capability but still considerable deviation from the actual values.

The XGBoost Regressor shows significantly poorer performance with an MAE of 0.2324, MSE of 0.0601, and RMSE of 0.2451, along with a negative R^2 score of −2.4042, which indicates that the model performs worse than a baseline model predicting the mean value and fails to effectively capture the relationship between input features and price per square foot. In contrast, the proposed LSTM–TreeNet Regressor demonstrates substantially superior performance, achieving a very low MAE of 0.0046, MSE of 0.0014, and RMSE of 0.0369, along with a high R^2 score of 0.9227. These results indicate that the hybrid model produces highly accurate predictions with minimal error and strong explanatory power. The improved performance can be attributed to the combination of LSTM-based deep feature extraction, which captures complex nonlinear relationships among property attributes, and the decision-tree regression mechanism, which refines predictions using structured decision boundaries. Consequently, the LSTM–TreeNet model provides significantly more reliable and precise price-per-square-foot predictions compared to the traditional Gradient Boosting and XGBoost regression mode

Table. 3: Price Per SQFT Target Performance Comparison of Various Regression Models.

Model	MAE	MSE	RMSE	R^2 Score
GB Regressor	0.0667	0.0093	0.0966	0.4710
XGBoost Regressor	0.2324	0.0601	0.2451	-2.4042
LSTM–TreeNet Regressor	0.0046	0.0014	0.0369	0.9227

5. CONCLUSION

The proposed real-estate valuation framework demonstrates the effectiveness of integrating advanced machine learning and deep learning techniques for predicting both property orientation and price-related attributes. The system incorporates a complete analytical pipeline including dataset preprocessing, exploratory data analysis, SMOTE-based class balancing, and the training of multiple predictive models. Experimental evaluation was performed using three different model families: GB-CART, XGBoost-CART, and the proposed hybrid LSTM–TreeNet architecture. The results indicate that traditional ensemble models such as Gradient Boosting and XGBoost show limited capability in accurately capturing the complex relationships present in the real-estate dataset, particularly in regression tasks where higher prediction errors were observed. Similarly, the classification performance of these baseline models remained relatively low due to their limited ability to model nonlinear and high-dimensional interactions among property features.

In contrast, the proposed LSTM–TreeNet model significantly outperformed the conventional approaches across both classification and regression tasks. The hybrid architecture leverages the deep feature extraction capability of the LSTM network to learn complex patterns from structured property data, while the decision-tree component refines predictions through interpretable decision boundaries. This combination resulted in highly accurate predictions, achieving approximately 97.11% classification accuracy for property orientation and strong regression performance with R^2 scores above 0.91 for price-related targets. The model also demonstrated lower prediction errors for both property price in lakhs and price per square foot, confirming its ability to effectively capture nonlinear dependencies among real-estate attributes. So, the developed system provides a scalable and automated solution for real-estate valuation, reducing reliance on manual estimation methods and improving decision-making for buyers, sellers, and investors.

REFERNCES

[1] Chen, X.; Chen, Y. Evaluation and Improvement Strategies of the Long-Term Mechanism for Real Estate: A Perspective of the "Three-in-One" Macro Policy Approach. Stud. Explore. 2022, 8, 99–112.

[2] Chen, X.; Cheng, S.; Chen, K.; Xiao, Z.Y. A Study on the Influencing Factors of Housing Prices in First-Tier Cities Based on Machine Learning Methods. Nankai J. Philos. Soc. Sci. 2023, 6, 146–163.

[3] Gong, J.; Zheng, T. Investor Attention and Intercity Housing Price Spillover Effects: A Study Based on Baidu Search Index Between Pair Cities. J. Financ. Econ. 2023, 6, 55–70.

[4] Liu, S.; Chen, M. Study on the Transmission Effect of Chinese Urban Housing Price Fluctuation Diffusion Level. Reg. Res. Dev. 2021, 40, 45–50.

[5] Cui, Z.; Zhou, M.; Kong, L. Study on the Heterogeneity of Influencing Factors of Urban Housing Prices in China. Tax Econ. 2022, 6, 65–74.

[6] Paulus, N.M.; Lautenschlaeger, L.; Schaefers, W. Social Media and Real Estate: Do Twitter Users Predict REIT Performance? J. Real Estate Res. 2024, 1, 1–34.

[7] Alhefnawei, A.M.M.; Al, E. Population Modelling and Housing Demand Prediction for the Saudi 2030 Vision: A Case Study of Riyadh City. Int. J. Hous. Mark. Anal. 2024, 17, 1558–1572.

[8] Andrade-Girón, D.C.; Marin-Rodriguez, W.J.; Zuñiga-Rojas, M.G. Intelligent Feature Selection Ensemble Model for Price Prediction in Real Estate Markets. Informatics 2025, 12, 52. https://doi.org/10.3390/informatics12020052

[9] Song, D.; Hu, G.; Li, H.; Zhao, H.; Wang, Z.; Liu, Y. Real Estate Market Forecasting for Enterprises in First-Tier Cities: Based on Explainable Machine Learning Models. Systems 2025, 13, 513. https://doi.org/10.3390/systems1307051

[10] Maselli, G.; Nesticò, A. Machine Learning Algorithms and Explainable Artificial Intelligence for Property Valuation. Real Estate 2025, 2, 12. https://doi.org/10.3390/realestate2030012

[11] Anelli, D.; Morano, P.; Tajani, F.; Guarini, M.R. The Interpretative Effects of Normalization Techniques on Complex Regression Modelling: An Application to Real Estate Values Using Machine Learning. Information 2025, 16, 486. https://doi.org/10.3390/info16060486

[12] Battisti, F.; Campo, O.; Forte, F.; Menna, D.; Perdonò, M. Residential Mobility: The Impact of the Real Estate Market on Housing Location Decisions. Real Estate 2025, 2, 9. https://doi.org/10.3390/realestate2030009

[13] Al-Rimawi, T.; Nadler, M. Leveraging Smart City Technologies for Enhanced Real Estate Development: An Integrative Review. Smart Cities 2025, 8, 10. https://doi.org/10.3390/smartcities8010010

[14] Di Liddo, F.; Amoruso, P.; Morano, P.; Tajani, F.; Locurcio, M. A Data Analysis of the Relationship Between Life Quality Indicators and the Real Estate Market in Italian Provincial Capitals. Real Estate 2025, 2, 4. https://doi.org/10.3390/realestate2020004

[15] Çılgın, C., Gökçen, H. A Hybrid Machine Learning Model Architecture with Clustering Analysis and Stacking Ensemble for Real Estate Price Prediction. Comput Econ 66, 127–178 (2025). https://doi.org/10.1007/s10614-024-10703-4

[16] Bastos, J. A., & Paquette, J. (2024). On the uncertainty of real estate price predictions. Journal of Property Research, 42(1), 1–19. https://doi.org/10.1080/09599916.2024.2403998