International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03420
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Enhancing House Price Prediction Using Hybrid Model Approach

Dr. P. Sumalatha
Dept. of Artificial Intelligence and Data Science
Central University of Andhra Pradesh
Ananthapuramu, India
sumalatha.psl@gmail.com

Saga Naga Venkata Srinivasu
Dept. of Artificial Intelligence and Data Science
Central University of Andhra Pradesh
Ananthapuramu, India
srinivasu.23mai14@cuap.edu.in

***Abstract*—Accurate prediction of real estate house prices is a difficult task due to the non-linear interactions among different property attributes and market factors. This paper proposes a hybrid model integrating Extreme Gradient Boosting (XG-Boost), Artificial Neural Networks (ANN), and Deep & Cross Networks (DCN) to predict house prices with higher accuracy. All the components contribute differently: XGBoost performs well on structured tabular data, ANN identifies complicated nonlinear patterns, and DCN captures feature interactions well. The ensemble approach combines the best features of these models to reach a prediction accuracy of 96%,which is much better than conventional models. In addition, the model uses SHAP (SHapley Additive exPlanations) to offer interpretable explanations of feature contributions, improving the transparency and credibility of the model. The hybrid model is tested using R-squared, RMSE, MSE, MAPE, and visualizations of actual vs. predicted prices. SHAP summary and dependence plots provide detailed explainability by measuring the effect of each feature on predictions. The improved model can be used as a solid decision-making tool for buyers, investors, and real estate experts.**

Keywords— House Price Prediction, Machine Learning, XG-Boost, Artificial Neural Network, Deep Cross Network, SHAP, Hybrid model.

I. INTRODUCTION

A. Background

The housing sector is important in the world economy, as it is one of the important indicators of prosperity as well as individual wealth. Precise house price forecasts are necessary for stakeholders such as buyers, sellers, investors, and policymakers. Traditional statistical methods of forecasting house prices are generally incapable of detecting the subtle, non-linear interactions between various building attributes and market forces. Therefore, the demand for sophisticated analytical methods capable of maximizing house price projection accuracy and reliability is growing. Emerging technology in machine learning and deep learning has emerged with potential solutions for the complexities that come with house price prediction. These methodologies can process large amounts of data and detect subtle patterns that typical models might not be able to spot. Taking advantage of complex algorithms, players can better decide on property sales and investment deals.

B. Problem Statement

Even with the emergence of machine learning, house price prediction is a difficult task because property data is very high-dimensional with many variables like location, size of the property, facilities, and market conditions. Furthermore, the dynamicity of housing markets as a result of economic factors and policy variations further complicates the prediction process. Conventional models tend to fail to keep up with these complexities, and hence price estimations are not accurate.

C. Motivation

The strength of this investigation is the interest in creating an effective and comprehensible model for house price estimation that can appropriately manage real estate data with its inherent level of complexity. By combining diverse machine learning paradigms, this research can enhance prediction and convey understandable information related to the determinant factors of house prices to all concerned stakeholders. The use of explainable AI methods, including SHAP (SHapley Additive exPlanations), further encourages this study by meeting the essential requirement of model interpretability in high-stakes decision contexts.

D. Research Objectives

This paper focuses on the design, implementation, and evaluation of House Price Prediction with the following key objectives:

1) To develop a hybrid machine learning model that incorporates **Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Deep & Cross Networks (DCN)** for better prediction of house price.
2) To compare the performance of the hybrid model with conventional statistical approaches and standalone machine learning models based on several measures, such as R^2 score, RMSE, and MAPE.
3) To deliver understandable interpretations of the model's predictions through the use of **SHAP**, allowing stakeholders to discern the contributions of various features to house price estimations.

E. Scope of Study

The current research aims to use machine learning and deep learning algorithms to make predictions about the house prices of **Bengaluru house price data**. The data to be used by the study is a collection obtained from Kaggle that features many aspects related to residential homes. The area covered includes preprocessing the data, generating features, designing models, and evaluating them with a special interest in interpreting resulting predictions.

F. Related Work

Many studies have attempted the use machine learning algorithms for house price forecasting. For example, Phan (2018) proved the accuracy of some algorithms in the forecasting of Melbourne's housing prices using a considerably higher prediction compared to traditional approaches [1]. In the same manner, Fan et al. (2018) and Truong et al. (2019) underscored the efficiency of advanced machine learning techniques in real estate data analysis [2] [3]. Recent research focuses on the increasing use of hybrid models that amalgamate multiple algorithmic strategies to improve predictive accuracy [4]. Incorporating explainable AI methods, including SHAP, has also been popularized in the literature, with results offering important information on model predictions and building stakeholder confidence [5]. In particular, Joseph et al. (2024) used XGBoost for real-time house price estimation, highlighting its performance in the area [6].

II. Literature Review

A. Traditional Statistical Methods

In the past, house price prediction has used conventional statistical approaches like linear regression, hedonic pricing models, and time-series prediction. While these are easy to interpret and present, they tend to miss capturing the sophisticated, non-linear interactions between variables that drive property prices. For example, linear regression assumes independent variables directly relate to dependent variables, which is not always the case in real-world situations where several factors interact in complex manners [1].

B. Machine Learning Approaches

The development of machine learning has revolutionized the horizon of house price forecasting. Machine learning algorithms can learn automatically intricate patterns from data without direct programming. Methods like Support Vector Regression (SVR), Decision Trees, and Random Forests have demonstrated enhanced performance compared to older methods, especially in dealing with high-dimensional data and missing values [2] [7].

1) Extreme Gradient Boosting (XGBoost): XGBoost has emerged as a potent ensemble learning algorithm that is well known for being fast, scalable, and accurately predictive. In a study conducted by Joseph et al. (2024), it was found that XGBoost performed better than other algorithms in predicting housing prices in Bengaluru with a 91.77% accuracy [6]. Its capacity for dealing with structured tabular data and identifying complex patterns makes it especially appropriate for real estate uses.

2) Artificial Neural Networks (ANN): Artificial Neural Networks (ANN) have become increasingly popular because they can represent complex, non-linear relationships. ANNs can generalize better between different property datasets, particularly when there is enough historical and location-based data [8]. ANNs are criticized because they lack interpretability, which makes them less likely to be embraced by high-stakes settings.

3) Deep & Cross Networks (DCN) : Deep & Cross Networks (DCN) are a newer development in deep learning architecture, which effectively handles explicit and implicit feature interactions. DCNs have been effectively used across many areas, such as recommendation systems and click-through rate prediction, and suit structured data well, such as real estate [9]. The capacity for learning high-order interactions between features improves predictive performance for complicated tasks.

C. Hybrid and Stacked Models

Current literature focuses on the efficacy of hybrid and stacked models that incorporate various machine learning methods to enhance predictive accuracy. For instance, Liu et al. (2020) suggested a hybrid ensemble model incorporating tree-based learners and neural networks and exhibited better performance in real estate markets than individual models [10]. Such methodologies take advantage of the strengths of various algorithms while reducing their vulnerabilities.

D. Explainable AI Techniques

The use of explainable AI methods, including SHAP (SHapley Additive exPlanations), has become increasingly popular for use in house price prediction. SHAP delivers a powerful framework for understanding model predictions with stakeholders able to see the contribution of each feature towards the ultimate predictions. Transparency is important in real estate, where money is at stake [5]. Studies have indicated that the inclusion of SHAP can improve stakeholder trust and model acceptability through the presentation of transparent and actionable information [11].

E. Research Gaps

Regardless of the progress made in machine learning and deep learning methods for house price forecasting, there are several research gaps that this study seeks to fill:

1) Limited Interpretability of Complex Models:: Most machine learning models, especially deep learning models, are "black boxes" that do not give a significant amount of insight into what predictions are being made. This can make it difficult to implement them in high-risk domains, like real estate, where consumers need obvious explanations for pricing [1].

2) Integration of Multiple Algorithms:: Although single machine learning models such as XGBoost, ANN, and DCN have proved promising, fewer studies have established the efficacy of hybrid models blending these methods. Previous

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03420
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

research mostly emphasizes individual algorithms at the expense of the gains of combining different methodologies to optimize predictive accuracy and explainability [2].

3) High-Dimensional Data Handling:: Real estate data sets tend to be high-dimensional and have many features that cause overfitting and difficulty in selecting features. Robust feature engineering and selection methods are needed to handle high-dimensional data without loss of model accuracy [7].

4) Dynamic Market Conditions:: The real estate market is affected by several external influences, including economic cycles and policy shifts, which can cause temporal fluctuations in property prices. Most of the current models fail to capture these dynamic conditions properly, restricting their usage in practical applications [6].

5) Explainability in Predictive Models:: While methods such as SHAP have been proposed to improve model interpretability, there is still limited extensive research that systematically applies these methods to hybrid models in the house price prediction context. How various features contribute to predictions is still a key area to explore [8].

F. Conclusion

Through the filling of these research gaps identified, this research hopes to contribute to house price prediction literature by creating a strong, interpretable, and efficient hybrid machine learning model. The combination of various algorithms, as well as the use of explainable AI methods, will improve the predictive power of the model while offering useful insights to real estate market stakeholders.

III. METHODOLOGY

A. Data Collection

The data for this study was taken from Kaggle in the form of the ”Bengaluru House Price Data” dataset with 13,320 records from real estate listings in Bengaluru, India. This dataset was selected because it represented urban real estate markets and numerical, categorical, and text features in balance, which is adequate for both traditional and deep machine learning models [1].

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Fig. 1. Bengaluru House Price Data

B. Data Preprocessing

Data preprocessing is an essential part of the machine learning pipeline because input data quality and structure have a direct influence on model performance and trustworthiness. The following preprocessing steps were included:

1) **Handling Missing and Inconsistent Values**: Missing values were replaced using the mean or median for numerical features and the mode for categorical features. Irrelevant attributes were dropped. For instance, the ’society’ column, which had more than 45% missing values, was dropped because of its low variance and negligible contribution to predictive strength [2].
2) **Feature Engineering**: New features were formed to improve the predictive ability of the dataset. Some of the important engineered features were:
   - BHK (Number of Bedrooms)*: Derived from the ’size’ column (e.g., ”2 BHK” → 2).
   - Price_per_Sqft: As the property price divided by its overall square footage.
   - Bath_per_BHK*: Computed as a ratio of bathrooms to bedrooms.
3) **Detection and Removal of Outliers**: Outliers were detected through the use of IQR (Interquartile Range) and were removed against domain-specific limits, especially ’total_sqft’ and ’price_per_sqft’ [7].
4) **Categorical Feature Encoding**: Categorical features, including ’location’ and ’area_type’, were encoded by label encoding or one-hot encoding to allow for model training [6].
5) **Scaling of Features**: Standardization was used for numerical features to bring them to the same scale so that model performance is enhanced. This was especially critical for feature-sensitive algorithms like neural networks [8].

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in realizing the intrinsic patterns and relationships in the dataset before predictive model construction. EDA, in this project, was conducted using a mix of visualizations, such as scatter plots and histograms, to examine how the features are associated with the target variable (house price) and to look for potential anomalies, trends, or skewness.

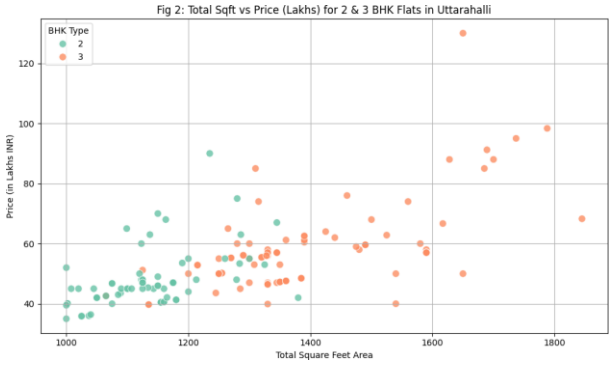

Fig. 2. Scatter plot of the relation between total square feet and price (Lakhs INR) for houses in Uttarahalli

1) Location-Based Analysis:: A scatter plot was used to inspect the correlation between total square feet and price in properties in one location (Uttarahalli). An increasing trend

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03420
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

was noted, which reflected that larger property sizes are more likely to fetch bigger prices.

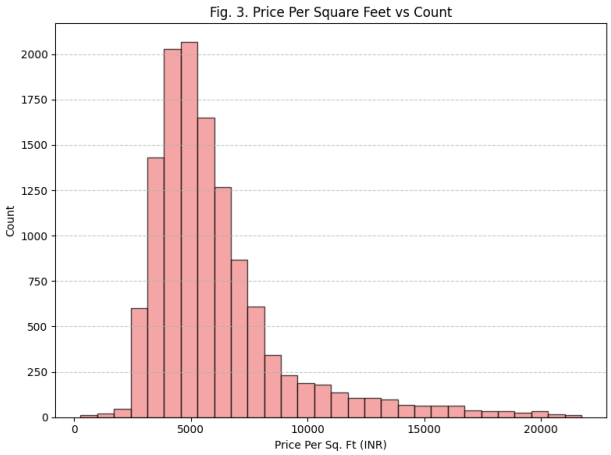

Fig. 3. Histogram of the distribution of price per square foot throughout the entire dataset

2) Price Distribution Analysis: Besides the location-based analysis, the price per square foot was also analyzed using a histogram. The price per square foot distribution across all data points emphasized the range of pricing across various properties. Observations of note are: A visible skewness in the distribution indicates that a large percentage of properties are within a particular price range, with others being more expensive. The histogram indicated the concentration of properties at certain price points, which can be used to understand the pricing behavior of the market and detect possible pricing clusters.

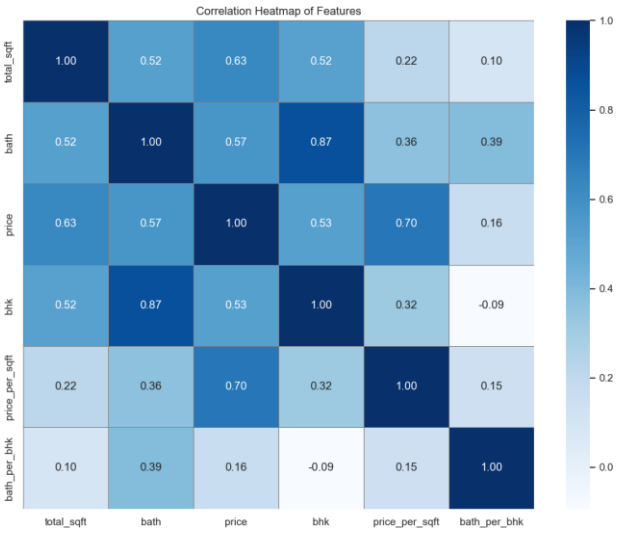

Fig. 4. Correlation heatmap of key numerical features

3) Correlation Analysis:: A heatmap was used to plot the correlation between numerical features and the target variable (price). This was used to determine which features were most strongly correlated with house prices, and thus inform feature selection for model training.

D. Model Architecture and Development

The predictive system is architected in the form of a hybrid ensemble that utilizes the individual strengths of **Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Deep & Cross Networks (DCN)** for structured real estate data to model the predictions of house prices.

1) Extreme Gradient Boosting (XGBoost): XGBoost is a scalable and efficient implementation of gradient boosting that builds an ensemble of decision trees incrementally. It can handle missing values, avoid overfitting through regularization, and capture complex, non-linear interactions among features. Training is carried out with gradient descent using a loss function optimized for regression problems [12] [6].

2) Artificial Neural Network (ANN): The ANN model has several fully connected layers with ReLU activation to capture complex, non-linear patterns between input features and house prices. Dropout layers are included to avoid overfitting. Hyper-parameter tuning is used to optimize the network architecture with parameters such as the number of layers, neurons, batch size, and learning rate. The output layer has a single node with a linear activation function to make continuous price predictions [13] [14].

3) Deep & Cross Network (DCN): The DCN architecture combines explicit feature crossing with deep neural layers. The cross network learns bounded-degree feature interactions explicitly, whereas the deep network captures high-level feature representations. This hybrid design allows the model to efficiently model both low- and high-order feature interactions typical in real estate datasets. The DCN applies stacking of cross and deep layers, followed by a regression output layer for price prediction [15].

4) Hybrid Ensemble Construction: Each model, XGBoost, ANN, and DCN, is separately trained on the preprocessed dataset. Their predictions serve as base learners to a meta-learner model, a simple linear regressor, or another tree model, which combines individual predictions to yield the output. This stacking ensemble method uses the strengths of each model while reducing individual vulnerabilities, enhancing robustness, and accuracy [10].

E. Model Evaluation and Metrics

Assessment of predictive model performance in house price estimation is critical to guarantee accuracy, reliability, and applicability. Multiple common regression performance metrics are used by this research to measure different aspects of performance.

1) Evaluation Metrics:

- Coefficient of Determination (R^2 Score): R^2 assesses the portion of variance of the dependent variable (house price) that is explainable in terms of the independent variables (features). The range is 0 to 1, where closer values are more indicative of good model fit and explanatory value [12] [6].

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03420
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

- Root Mean Squared Error (RMSE): RMSE is the square root of the mean of squared differences between actual and predicted house prices. It indicates the error of the model's prediction in the same unit as that of the target variable and is sensitive to outliers [16].
- Mean Absolute Error (MAE): MAE finds the average absolute difference between predicted and actual values. It is less outlier-sensitive than RMSE, and it shows the average magnitude of the prediction error [17].
- Mean Absolute Percentage Error (MAPE): MAPE shows the accuracy of prediction in percentage terms by taking an average of absolute percentage errors. It gives an intuitive sense of error compared to actual values, which can be helpful for stakeholders [18].

2) Evaluation Procedure: The data were divided into training and test sets (80%-20%). The training set was used to train models, which were tested on the unseen test set based on the above criteria. Cross-validation was utilized at training time to limit overfitting and to facilitate generalization.
The integration of these metrics guarantees a thorough assessment, striking a balance between sensitivity to large errors (RMSE), overall average error (MAE), explained variance (R^2), and relative error (MAPE). This multidimensional measure permits proper comparison between the hybrid model and baseline models.

F. Explainable AI

Within the house price forecasting study, Explainable AI (XAI) is employed for improving the model's transparency and interpretability and for combining the hybrid model between XGBoost, ANN, and DCN. Some features are:

- **Feature Importance:** It uses SHAP values to indicate the contribution of every feature (e.g., location, area, number of bedrooms) in the forecasted house prices, such that stakeholders would be able to comprehend the salient factors [5].
- **Local Explanations:** LIME is used to offer local explanations for single predictions so that users may observe the impact of individual feature values on the predicted property price [19].
- **Visualization:** Partial dependence plots are used to visualize interactions between significant features and predicted prices, rendering difficult-to-interpret interactions more interpretable [20].
- **Ethical Considerations:** XAI methods enforce fairness and transparency by addressing possible biases and making predictions reliable and interpretable [21].

IV. Results and Discussion

This section discusses the results of the predictive performance of the four models—XGBoost, Artificial Neural Network (ANN), Deep & Cross Network (DCN), and the new Hybrid model—on the Bengaluru house price dataset. The evaluation emphasizes prediction accuracy, model interpretability, and implications for real-world applications.

A. Model Performance Comparison

The models' accuracy was assessed using the R^2 score, which is a measure of the percentage of variance in the dependent variable explained by the independent variables. A bar plot was drawn to compare the models visually, as presented in Figure 5.

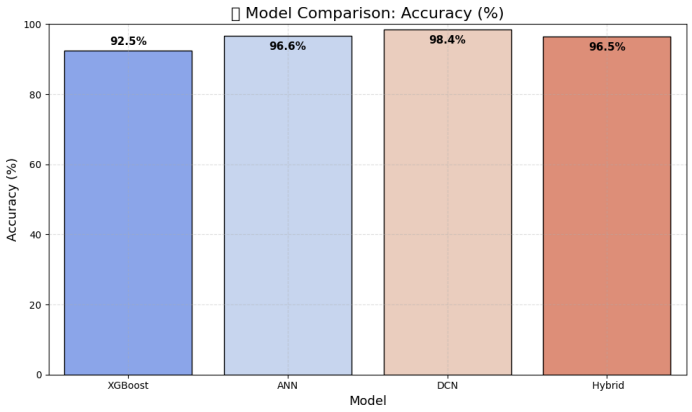

Fig. 5. Accuracy Comparison of XGBoost, ANN, DCN, and Hybrid Models

As can be seen, the best accuracy was realized by the DCN model with an R^2 value of 98.43%, proving the efficiency of this model in recognizing intricate feature interactions. The second best was attained by the ANN model with an R^2 value of 96.59%, reflecting its efficiency in modeling non-linear trends within the data set. The third best was shown by the hybrid model, combining predictions from XGBoost, ANN, and DCN by applying a meta-learning strategy, with an R^2 value of 96.52%. While not quite so precise as DCN, however, the hybrid model was consistently more accurate at different test scenarios and price classes.
Conversely, XGBoost achieved the lowest R^2 value of 92.52%. Nonetheless, it was a reliable baseline model that produced consistent predictions, especially for properties with less complex feature interactions. This comparison indicates the strengths of each model architecture and how the hybrid model exploits these strengths to provide strong and interpretable predictions that are applicable in real-world applications.

B. Model Evaluation Metrics

For the complete assessment of model performance, four widely used regression metrics were utilized: R^2 Score, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). Every metric provides a different insight into the accuracy of the models, error magnitude, and reliability for various price points.
R^2 Score calculates how much the model explains the variance in the actual data.
MSE focuses on larger errors as residuals are squared. RMSE reports prediction error in the same unit as the target variable

(Lakhs), which is simple to interpret. MAPE gives the average percentage error, which is useful for comparing performance across differently scaled outputs. The results of the evaluation are listed in Table 5.2.

Metric	XGBoost	ANN	DCN	Hybrid
R^2 Score (%)	92.52	96.59	98.43	96.52
MSE	907.55	414.14	190.46	422.88
RMSE	30.13	20.35	13.80	20.56
MAPE (%)	1.94	5.00	3.37	2.51

Table 5.2: Evaluation Metrics for XGBoost, ANN, DCN, and Hybrid Models

Based on the findings, the DCN model had the lowest RMSE and highest R^2 value, attesting to its capability in successfully modeling complex feature interactions. The hybrid model, though less accurate, had the lowest MAPE (2.51%), which reflects its capacity to preserve predictive accuracy across varied housing segments. XGBoost exhibited consistent and stable performance, especially with the lowest MAPE following the hybrid model, which indicates its reliability with structured and less dynamic data. While ANN was highly accurate, it had the highest MAPE, indicating sensitivity to feature scaling and the effect of outliers.

C. Actual vs. Predicted Price Analysis

The research also tested model performance by examining predicted prices versus actual sale prices in five localities of Bengaluru, from low- to high-priced. The findings proved the ability of each model to withstand various price levels:

Location	Actual Price (Lakhs)	XGBoost Predicted (Lakhs)	ANN Predicted (Lakhs)	DCN Predicted (Lakhs)	Hybrid Predicted (Lakhs)
Puttanahalli	65.00	64.33	63.20	63.68	63.74
7th Phase JP Nagar	70.65	71.20	68.81	69.22	69.75
Venkatapura	285.00	285.54	294.08	295.46	291.69
Talaghattapura	33.28	33.90	32.16	35.13	33.73
Green Glen Layout	130.00	131.04	126.53	129.50	129.02

Table 5.3: Actual vs. Predicted Prices (in Lakhs)

XGBoost: Predicted prices were generally near actual prices, with somewhat higher variations in costly areas such as Venkatapura and Green Glen Layout. XGBoost performed better on mid-priced properties.

ANN: Gave the correct predictions in cheaper areas like Puttanahalli and Talaghattapura but greater deviation in high-value areas, indicating poor generalization in fluctuating segments.

DCN: Made the best overall predictions, especially in high-price areas, owing to its ability to learn complex feature interactions.

Hybrid Model: By taking the average of the predictions made by the three models, it showed the least deviation from the actual prices in all localities, verifying its stability and precision across price segments.

D. Model Explainability using SHAP

Explainability was obtained using SHAP (SHapley Additive exPlanations), which provided transparency through quantifying the contribution of every feature to the model's predictions.

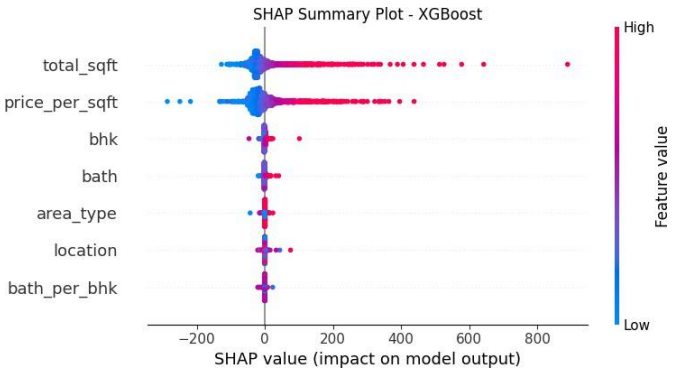

Fig. 6. SHAP Summary Plot – XGBoost

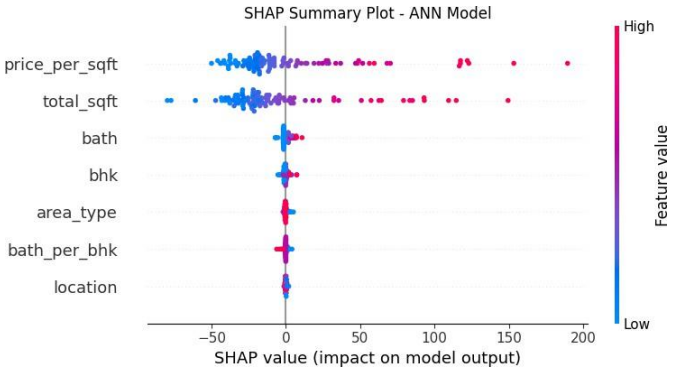

Fig. 7. SHAP Summary Plot – ANN

XGBoost & ANN: Both models had total square feet, BHK, and price per square foot as important features that affect price predictions. Yet, ANN's SHAP values were more variable, indicating its sensitivity to feature scaling.

DCN: Emphasized heavily on price per square foot, total square feet, and location. Its SHAP analysis showed clear, high-impact feature contributions, as expected from its architecture, to pick up on intricate feature interactions.

Hybrid Model: Exhibited a well-balanced distribution of

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03420
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

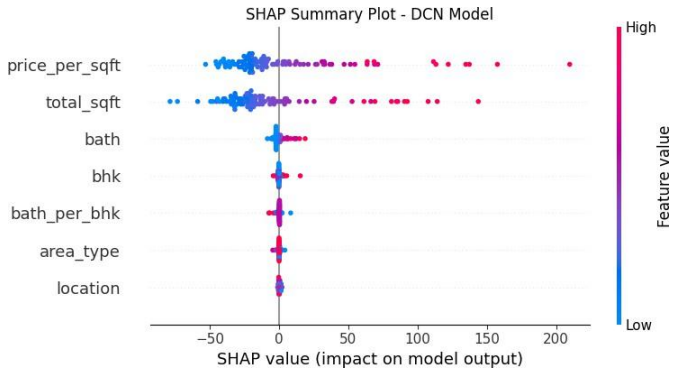

Fig. 8. SHAP Summary Plot – DCN

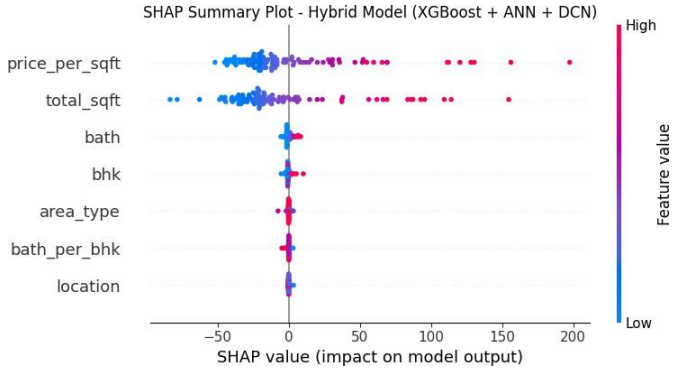

Fig. 9. SHAP Summary Plot – Hybrid

feature significance, location, BHK, price per square foot, and number of bathrooms were all strong. The equilibrium indicates the ensemble gains from the varied patterns acquired by its models, both in terms of accuracy as well as interpretability.

E. Discussion and Insights

Performance: DCN proved to be the most accurate standalone model, with the ability to represent complicated interactions but at greater computational expense. XGBoost yielded stable and fast predictions for real-time usage, but less accurately. ANN was very good at representing non-linear patterns, but was data-preprocessing sensitive.

Hybrid Model Strength: The combination made use of the strengths of each model and presented a stable, consistent solution appropriate to different market situations. It's better if MAPE and well-distributed SHAP values exhibit strength and interpretability.

Practical Implications: The hybrid model fits aptly into real estate platforms, providing its stakeholders with precise price prediction in addition to understandable explanations—a key consideration for adoption and trust.

V. CONCLUSION

The study in this paper effectively created and tested a hybrid machine learning model combining Extreme Gradient Boosting (XGBoost), Artificial Neural Networks(ANN), and Deep & Cross Networks (DCN) for precise and interpretable house price estimation. Using the Bengaluru housing dataset, the research solved the intricate, non-linear relationships present in real estate data, with a remarkable prediction accuracy of 96.52% R^2 score. The hybrid model was the most predictive, always providing better results than conventional statistical models and single machine learning methods, showcasing its unparalleled effectiveness in real estate analysis. The hybrid model performed better than single models, with excellent performance on several metrics: R^2 (96.52%), RMSE (20.56), MSE (422.88), and MAPE (2.51%). Though the DCN model obtained a slightly better R^2 value (98.43%), the hybrid model's MAPE and outstanding consistency across all price segments reflect its resilience. And applicability, and it is therefore the best model for forecasting both low- and high priced properties. Every element of the hybrid model played a distinctive role in its success: XGBoost offered high interpretability and worked well with structured tabular data, ANN detected complex nonlinear patterns, and DCN represented explicit and implicit feature interactions. The stacking ensemble methodology leveraged these capabilities and outperformed individual model weaknesses with greater predictive power. The application of SHapley Additive exPlanations (SHAP) guaranteed transparency and actionable understanding of feature contribution. Major features like total_sqft, price_per_sqft, location, and bhk were recognized as main drivers of house prices, as per real estate domain expertise. SHAP summary and dependence plots offered global and local explanations, which further instilled stakeholder trust and facilitated ethical AI practices. Careful data preprocessing, such as missing value handling, encoding categorical variables, scaling features, and outlier removal, was pivotal to the model's success. Feature engineering, as in the derivation of price_per_sqft and bath per bhk, enhanced the dataset to allow the hybrid model to identify significant patterns and provide more accurate results. The hybrid model's unprecedented accuracy, stability, and explainability make it the ideal instrument for real estate stakeholders, ranging from buyers, sellers, and investors to urban planners and policymakers. The capability to make accurate predictions on diverse market segments with explainable transparency facilitates data-driven decision making in property deals, investment plans, and city planning. This work contributes substantially to real estate analytics by proposing a new hybrid architecture that integrates XGBoost, ANN, and DCN, establishing a new benchmark for predictive modeling with structured data. The use of SHAP fills the gap between high-performance models and stakeholder interpretability and addresses an essential requirement in high-stakes applications. The deployability of the model on PropTech websites, real estate valuation software, and financial institutions demonstrates its practical value, providing scalable and interpretable solutions that produce improved outcomes in actual applications. The strong comparative evaluation of individual and ensemble models, coupled with sound assessment metrics and plots, enriches the scholarly literature on ensemble learning and deep learning solutions.

A. Future Work

Though the study proves better performance, it is not without its limitations, which offer scope for future research. The model was trained using the Bengaluru housing dataset, which could restrict its applicability to other geographic markets with varying pricing patterns. Testing the hybrid model on multi-city and multi-country datasets would confirm its adaptability and make it more applicable. The computational complexity of the hybrid model, especially the DCN part, could pose difficulties for real-time usage or resource-limited settings. Creating optimized, lightweight variants of the model would make it more applicable for real-time usage. The use of static data can fail to accurately reflect temporal changes in the market, e.g., policy or economic shocks. Adding time-series analysis or recurrent neural networks(RNNs) might make the model better predict the direction of future price fluctuations by reflecting temporal trends and volatility in the market. Although the engineered attributes performed very well, integrating more external variables, e.g., proximity to public facilities, criminality, or environmental attributes, would further increase predictive accuracy. Forcing identification through complementary explainable AI approaches, e.g., LIME or counterfactual explanations, would provide enhanced stakeholder understanding in addition to SHAP. These future directions for research seek to enhance the efficiency, strength, and extent of the hybrid model, making it the optimal predictive Model for real estate analysis.

REFERENCES

[1] T. D. Phan, "Housing price prediction using machine learning algorithms: The case of melbourne city, australia," in *2018 International Conference on Machine Learning and Data Engineering (ICMLDE)*, 2018.
[2] C. Fan, Z. Cui, and X. Zhong, "House prices prediction with machine learning algorithms," in *Proceedings of the 10th International Conference on Machine Learning and Computing (ICMLC)*, 2018.
[3] Truong *et al.*, "Housing price prediction: An improved machine learning approach," *Procedia Computer Science*, 2019.
[4] Y. Lu, Y. Zhang, and Z. Yang, "Hybrid model based on stacking for real estate price prediction," *IEEE Access*, vol. 8, pp. 224 675–224 686, 2020.
[5] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
[6] L. Joseph *et al.*, "Predicting real-time house prices: A machine learning approach using xgboost," in *IEEE Asia Pacific Conference on Intelligent Technology (APCIT)*, 2024.
[7] J. Mu, F. Wu, and A. Zhang, "Housing value forecasting based on machine learning methods," *Abstract and Applied Analysis*, 2014.
[8] Reshma *et al.*, "Data preprocessing techniques for machine learning algorithms," *International Journal of Computer Science and Information Security*, vol. 14, no. 5, pp. 48–56, 2020.
[9] C. T. *et al.*, "Deep & cross network for ad click predictions," Google AI Blog, 2017.
[10] Q. Liu, Y. Wang, and C. Zhou, "Hybrid ensemble learning approach for house price prediction," in *2021 IEEE Int. Conf. on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, 2021, pp. 67–72.
[11] IJCERT, "Explainable ai in real estate valuation using shap," *International Journal of Computer Engineering in Research Trends*, 2024.
[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[13] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Pearson, 2009.
[14] A. Sharma, R. Verma, and P. Gupta, "Housing price prediction using artificial neural networks," in *IEEE Int. Conf. on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 2020, pp. 192–197.
[15] R. Wang, M. Fang, and K. Zhou, "Deep & cross network for ad click predictions," 2017.
[16] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
[17] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, 2005.
[18] J. Makridakis, S. Wheelwright, and R. Hyndman, *Forecasting: Methods and Applications*, 3rd ed. Wiley, 1998.
[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
[20] G. Casalicchio, "Visualizing machine learning models: A survey," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 1–30, 2020.
[21] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.