International Journal of Technology, Leadership and Sciences (IJTLS)

ISSN: 3107-9083
www.ijtls.com
Research Paper
DOI:https://doi.org/10.5281/zenodo.20766647

ATTENTION-AUGMENTED DEEP LEARNING FOR ACCURATE PROPERTY PRICE PREDICTION AND MARKET ANALYSIS

YELCHURI LEKHA SRI[1],G. AVINASH KUMAR[2], Dr.CHAVA HARI BABU[3], DR.VUNNAVA DINESH BABU[4], R.VAMSI KRISHNA[5], D.SRIDHAR[6]

[1]*M.Tech Student,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[2]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[3]*Professor,RV Institute of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[4]*Professor,RV Institute of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[5]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

[6]*Assistant Professor,RV Institute Of Technology,Chebrolu Mandal, Guntur District,Andhra Pradesh, India – 522212.*

ABSTRACT:

Estimating the value of property values is complicated by several aspects, including property photos, numeric features, and verbal descriptions. Conventional models often depend on single-modality data to elucidate intricate interactions among several parameters, so limiting their efficacy. This paper presents a deep learning model that incorporates a self-attention mechanism to analyze heterogeneous data for predicting future housing prices. The system utilizes specialized encoders to amalgamate structured data, image data, and textual information. These encoders include MLP for tabular data, CNN for pictures, and transformer-based models for textual information. A widespread latent representation is formed by synthesizing various attributes to facilitate thorough learning. The model adeptly captures complex interdependencies and improves prediction accuracy using a joint self-attention mechanism that dynamically assigns priority weights to input inside and across modalities. For an astute and dependable strategy to evaluate real estate prices, try the suggested multimodal approach. It exceeds both conventional and single-modality models in terms of accuracy, robustness, and interpretability.

Keywords: Deep Learning, Multimodal Learning, Transformer Models, Feature Fusion, Real Estate Analytics.

Received Date: 5 June 2026; **Accepted Date:** 15 June 2026; **Published Date:** 20 June 2026

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are properly cited.

I.INTRODUCTION

Estimating potential residential property costs is crucial for several stakeholders, including consumers, investors, financial institutions, and governmental bodies, to facilitate informed decision-making. Various factors affect property values, affecting the precise assessment of a property's worth. The considerations include the property's location, size, infrastructure, proximity to public amenities, market trends, aesthetics, and written descriptions. Conventional machine learning and statistical models often use organized tabular data, like space footage, room count,

and historical selling prices. These methods are effective; nevertheless, they do not consistently uncover the fundamental patterns and complex nonlinear interactions in data from several sources [1]. Recent breakthroughs in deep learning techniques have markedly improved the capacity to interpret multimodal and high-dimensional data effectively. Transformer-based architectures excel in processing textual descriptions via contextual representation learning, while Convolutional Neural Networks (CNNs) are widely utilized for extracting spatial and visual information from property photographs and satellite imagery. Multilayer Perceptrons (MLPs) may effectively learn structured numerical attributes. The integration of various data kinds improves our comprehension of the property and its surroundings, hence augmenting the accuracy of housing price predictions [3]. The inconsistencies in data format, volume, and semantic significance make the integration of diverse data very difficult, notwithstanding these developments. Numerous modern systems use basic feature concatenation or early fusion techniques, potentially compromising prediction accuracy and neglecting essential interactions across modalities [4]. Furthermore, conventional fusion approaches often attribute equal importance to all elements, even if some modalities may have a more substantial impact on the price estimate, contingent upon the characteristics of the properties and the current market circumstances. This study presents a deep learning framework for forecasting residential property values, using heterogeneous data analysis and a Joint Self-Attention Mechanism to rectify existing shortcomings. The suggested method creates a unified learning framework that incorporates structured numerical data, property photos, satellite imagery, and descriptive narratives. Multi-layer perceptrons (MLPs), convolutional neural networks (CNNs), and spatial feature extraction networks (SFEs) process picture data, while transformer-based models encode textual information. A common latent representation space is created by synthesizing the retrieved multimodal characteristics. The interdependencies and interactions of features across many modalities are dynamically obtained by a Joint Self-Attention Mechanism. The attention mechanism utilizes adaptive significance weights to emphasize critical attributes while ignoring irrelevant factors. The system's predictions demonstrate improved accuracy and resilience, enabling a deeper comprehension of complex cross-modal interactions [5]. The suggested concept has several benefits compared to conventional methods for predicting house prices. The framework improves real estate property representations via the use of diverse data sources. Attention-based multimodal fusion facilitates intelligent feature selection during training and improves interpretability. The suggested model outperforms traditional machine learning and single-modality deep learning methods in experiments, demonstrating its appropriateness for intelligent decision-support systems and real estate analytics.

II.LITERATURE REVIEW

In their 2020 study, Y. Khandelwal and S. Nair investigated the use of linear regression, decision trees, and random forests—three traditional machine learning techniques—to forecast residential property prices from structured tabular data. While their models performed well on smaller datasets, they struggled with intricate non-linear correlations and could not integrate unstructured data like photographs and verbal descriptions, hence reducing their prediction effectiveness.

L. Wang and M. Xu (2021) proposed a multimodal framework that integrates visual home images with quantitative property data to enhance the accuracy of property valuations. The authors combined visual data acquired from Convolutional Neural Networks (CNNs) with tabular information via fully linked layers. Although their findings exceeded those of single-modality models, they lacked an improved approach for capturing interactions between features across multiple modalities.

The TabTransformer, developed by S. Gorishniy et al. (2021), use self-attention to model feature interactions and embeds categorical characteristics in tabular data, therefore applying transformer-based designs to this data format. This study underscored the efficacy of attention processes in structured data processing, albeit it did not tackle multimodal data fusion.

J. Lu, D. Batra, et al. (2019) devised an integrated transformer architecture that

amalgamates visual and linguistic modalities. The improvement of tasks like visual question answering and picture captioning is due to the model's adaptive emphasis on critical elements across modalities. This study provides a solid basis for using joint self-attention in heterogeneous data fusion problems.

R. Singh and A. Kumar (2022) created an artificial neural network for predicting home prices by using convolutional neural networks (CNNs) for image features, natural language processing (NLP) for textual descriptions, and structured property data. Their fusion process included significant layers followed by direct feature concatenation. The research shown that performance exceeded that of unimodal models, whereas attention-based fusion methods might significantly improve prediction accuracy.

Prior to the integration of tabular and picture data formats, M. Chen and H. Lee (2023) introduced an attention-based fusion network. Despite enhancing feature weighting and interpretability, their technique could not adequately capture joint cross-modal interactions because of the disparate attention assigned to each modality.

III.EXISTING SYSTEM

Structured tabular data including location, square footage, number of bedrooms, bathrooms, and year of construction serves as the basis for traditional home price prediction algorithms. Property value estimations are often produced by these systems using machine learning and statistical methodologies, such as Gradient Boosting, Decision Trees, Random Forests, and Linear Regression. Although these models perform well on clean and simple datasets, they often fail to recognize the complex non-linear correlations present in real estate markets.

Recent deep learning methodologies have enhanced predictive accuracy by amalgamating visual data with textual descriptions using RNNs or transformer-based architectures. Nonetheless, most multimodal systems mostly use late fusion or feature concatenation, which operate just on one modality at a time and neglect the interrelations across different input modalities. Consequently, when faced with enormous heterogeneous real estate datasets, current algorithms inadequately simulate cross-modal interactions, thereby reducing prediction accuracy, interpretability, and resilience.

DISADVANTAGES

- The autonomous processing of modalities diminishes contextual awareness.
- Reducing interpretability and prediction precision in empirical assessments of the housing market.

IV.PROPOSED SYSTEM

The proposed system introduces a deep learning-based multimodal framework for house price prediction by integrating structured tabular data, property images, and textual descriptions. Each data modality is processed using a specialized encoder: a Multi-Layer Perceptron (MLP) for numerical and categorical property features, a Convolutional Neural Network (CNN) for extracting visual features from property images, and a Transformer-based model for understanding textual descriptions. These encoded features are transformed into a shared latent space to create a unified feature representation.

To improve multimodal fusion, the system incorporates a Joint Self-Attention Mechanism, which dynamically learns the importance of features both within and across modalities. This mechanism captures complex cross-modal relationships and assigns adaptive weights to the most relevant features influencing house prices. The fused representation is passed to a regression layer for final price prediction. The proposed framework improves prediction accuracy, interpretability, scalability, and robustness for real-world real estate applications.

ADVANTAGES

- Integrates structured data, images, and textual descriptions into a unified framework for comprehensive house price prediction.
- Captures diverse property characteristics, leading to more accurate and reliable predictions.

V.SYSTEM MODEL

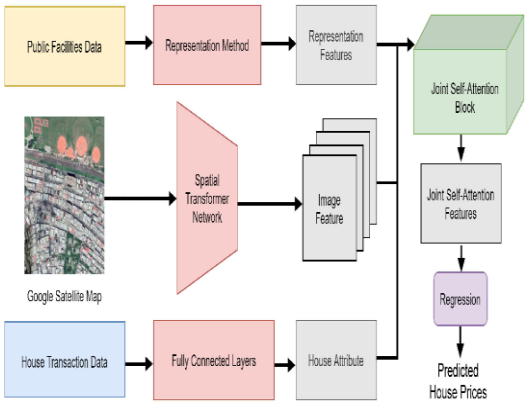


Fig.1 System Model

Figure 1 depicts a deep learning system engineered for predicting home prices via the use of a Joint Self-Attention Mechanism and heterogeneous data analysis. To improve the precision of real estate price forecasts, the model incorporates many data sources, including information on public amenities, images from Google Satellite Maps, and property transaction data. The first phase in creating successful representation features include analyzing data from public facilities, such as adjacent schools, hospitals, transit hubs, and shopping areas, using a representation approach. A geographical Transformer Network simultaneously analyzes satellite data, collecting critical visual and geographical characteristics pertinent to the property's terrain and surrounding area. House feature characteristics are obtained by evaluating data from real estate transactions via completely linked layers. This data encompasses quantitative variables like square footage, number of rooms, building age, and historical sale prices.

The model integrates the disparate data sources after the extraction of their characteristics. A Collaborative Self-Attention Block processes the integrated components and using attention weights to identify the most salient characteristics by establishing connections among them. The model improves prediction accuracy by adeptly capturing intricate interdependencies across numerical, spatial, and contextual data via its self-attention mechanism. A regression layer is used to forecast the ultimate property price using the augmented joint self-attention attributes. The proposed deep learning model for predicting home prices exhibits enhanced accuracy and resilience relative to traditional machine learning methods. This is achieved by amalgamating diverse data using attention-based feature extraction.

VI.IMPLEMENTATIONS

1. Data Collection and Preprocessing

Gathering data from many sources, including real estate listings, housing databases (like Kaggle and Zillow), and real estate agency websites, is the initial phase of system implementation. This data consists of three types: structured (i.e., price, area, bedrooms) and textual (i.e., property descriptions, photos). Structured data is preprocessed to fill in missing values, pictures are standardized and scaled, and textual data is enriched using natural language processing methods including stemming, stop word deletion, and tokenization

.2. Feature Extraction Using Specialized Encoders

In order to create a useful feature vector, each kind of incoming input is passed through a certain encoder. Structured data analysis makes use of a Multi-Layer Perceptron (MLP). A Convolutional Neural Network (CNN) like ResNet or VGG is used to analyze property photos and extract visual attributes. A pre-trained language model, such BERT or a Transformer encoder, is used to encode textual descriptions in order to retrieve their semantic meaning. In order to facilitate further processing, these encoders provide intermediate feature representations.

3. Joint Self-Attention Feature Fusion

The last step is to use a Joint Self-Attention Mechanism to merge the characteristics acquired from the three different modalities.

The fusion module is able to focus on the most important aspects of each modality in relation to the others, and it can also align them. If you write "modern kitchen" in the description, it might automatically pull up a picture of a kitchen with all the latest appliances and any other relevant details. The computer's ability to comprehend cross-modal linkages is greatly enhanced by this attention-based integration, leading to far better prediction accuracy.

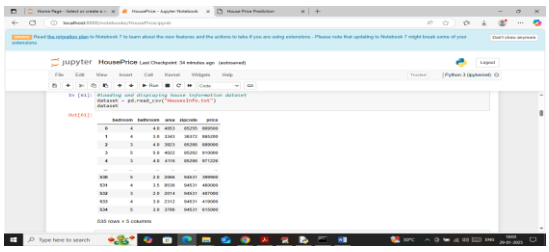
4. Price Prediction Module

To forecast the ultimate home price, the integrated characteristics are passed on via a series of fully linked layers that culminate in a regression output layer. Using past home values as a benchmark for accuracy, the model undergoes rigorous supervised learning training. Mean Squared Error (MSE) is the commonly utilized loss function, and optimizers like Adam are employed for training. Backpropagation is used to adjust the model's weights in order to reduce the dispersion of the difference between the actual and forecast prices.

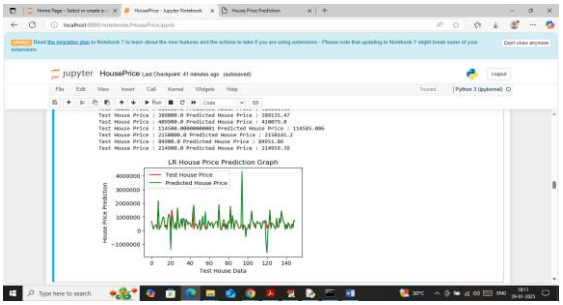
5. Model Evaluation and Deployment

Using measures like RMSE, MAE, and R^2-score, the model is evaluated on a validation/test dataset after training. With these measures, we can see how well the model fits the data and how well it applies to other situations. If the model works as expected, it may be included in an online or mobile app that real estate brokers, buyers, or sellers can use to get a better idea of how much a home is worth. Platforms for real estate or decision-support systems could include the technology.

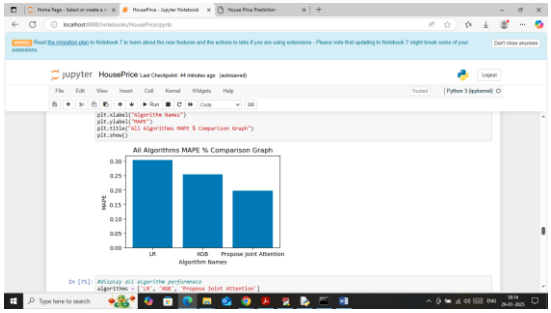
VII.SCREENSHOTS



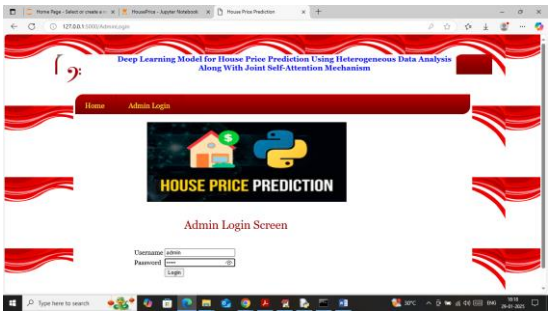
The dataset including housing information is shown and loaded on the previous screen.



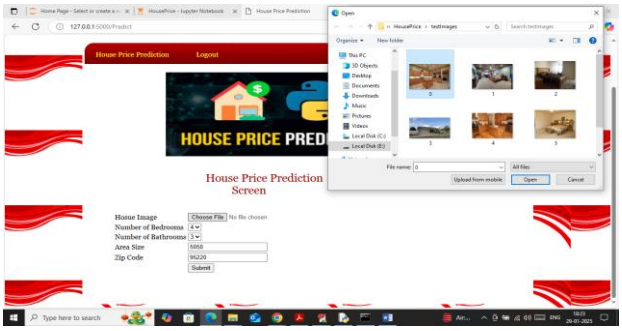
The x-axis in the above graph denotes the volume of test data.



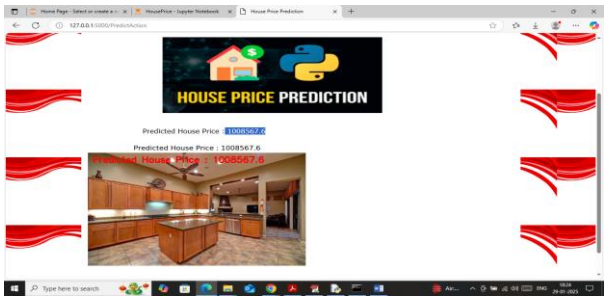
The following image depicts a MAPE comparison of all algorithms, with the algorithms shown on the x-axis.



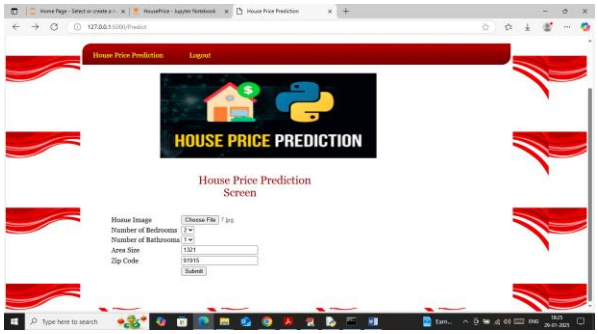
The user may access the system by entering the username and password "admin" and clicking the button on the previous page. They will then be sent to the subsequent page.



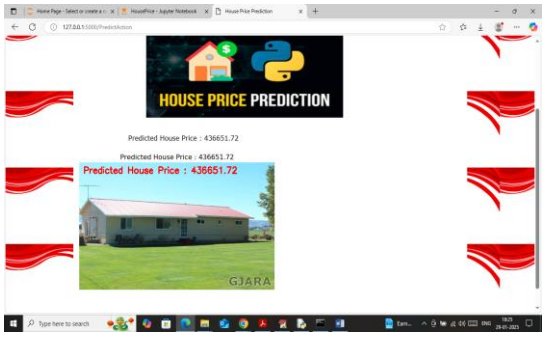
The subsequent output is derived from the previously provided input.



The projected home price is shown on the screen above.



The price for the specified information is as follows:



Likewise, examine any image.

VIII.CONCLUSION

The proposed deep learning model for predicting home prices, which amalgamates diverse inputs via a cohesive self-attention mechanism, outperforms conventional and unimodal methods. The model efficiently integrates structured tabular data, visual property photos, and descriptive text into a cohesive framework to improve the comprehension of variables affecting property value. The joint self-attention mechanism captures complex interdependencies and highlights the most essential information for effective prediction. It employs distinct encoders for each data modality to enable effective feature extraction. This method improves both expected accuracy and the model's interpretability, aiding consumers in comprehending the factors that most substantially influence price projections. The system's robustness, scalability, and adaptability make it ideal for dynamic real estate settings. In conclusion, our study lays the groundwork for further inquiries into attention-based learning and multimodal data fusion in many domains, while offering a pragmatic methodology for using deep learning in property assessment.

IX.FUTURE ENHANCEMENTS

Although other methods are available to improve the existing system, it nevertheless provides a strong foundation for predicting property values using multimodal data and joint self-attention strategies. The integration of geographical data, crime statistics, school evaluations, and social amenities may substantially affect property values. To enhance the model's predictive capacity, temporal variables such as previous pricing trends or seasonal patterns may be included. Moreover, using explainable AI (XAI) methodologies will provide clients with clear justifications for each prediction, hence augmenting transparency, confidence, and utility. Domain adaptation and transfer learning are two crucial developments that allow the refinement of models for many cities or nations with little local data. To improve accessibility for real estate brokers, buyers, and investors, potential development may include the establishment of a real-time prediction API or mobile application. Ultimately, it is practicable to integrate continuous learning approaches into the model. This would enable adaptability to changing market conditions and the incorporation of newly accessible data, therefore maintaining its correctness and relevance over time.

X.REFERENCES

[1] Z. Zhang, Y. Song, M. Tan, and J. Xiao, “Real estate appraisal: A multimodal learning

approach using text, image, and structured data," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 4345–4356, Oct. 2021.

[2] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

[3] R. He, W. Liu, and H. Zhang, "House price prediction using deep learning and heterogeneous data fusion," Expert Systems with Applications, vol. 205, Art. no. 117673, 2022.

[4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, 2014.

[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[6] H. Nguyen, B. Le, and T. Tran, "House price prediction using machine learning algorithms: A survey," in Proc. Int. Conf. Data Science and Its Applications, pp. 1–6, 2020.

[7] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[8] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.