Satellite Imagery–Based Property Valuation Using Multimodal Deep Learning

Name: Maddipati Sathvika
Enrolment: 22113083
Institution: IIT Roorkee
Date: 05-01-2026

Abstract

Accurate property valuation is essential for real estate markets, urban planning, and financial decision-making. Traditional valuation models rely heavily on tabular attributes such as size, location coordinates, and room counts, often ignoring valuable spatial and environmental information. This project proposes a multimodal deep learning approach that integrates satellite imagery with structured tabular data to improve property price prediction. A convolutional neural network (CNN) is used to extract visual features from satellite images, while a multilayer perceptron (MLP) processes tabular attributes. The learned representations are fused and passed to a regression head for final price estimation. Experimental results show that while the tabular-only model achieves strong predictive performance, the multimodal model provides additional qualitative insights through visual explainability. Grad-CAM analysis highlights spatial regions influencing predictions, demonstrating the interpretability of the multimodal approach.

1. Introduction

Property valuation plays a critical role in real estate transactions, taxation, mortgage approvals, and urban development planning. Conventional valuation models typically depend on structured data such as square footage, number of bedrooms, and geographic coordinates. While effective to an extent, these approaches fail to capture contextual information such as neighbourhood layout, surrounding infrastructure, greenery, and accessibility.

Recent advances in deep learning and remote sensing have enabled the use of satellite imagery to capture such spatial and environmental features. Satellite images provide rich visual cues that can complement traditional tabular attributes. This project explores a multimodal learning framework that combines satellite imagery and tabular property data to predict property prices more accurately.

The key contributions of this work are:

- Development of a multimodal regression architecture combining CNN-based image features with tabular features.
- Empirical comparison between tabular-only and multimodal models.
- Use of Grad-CAM to interpret the visual reasoning of the model.

2. Dataset Description

2.1 Tabular Data

The tabular dataset consists of residential property records with the following features:

- Bedrooms
- Bathrooms
- Square footage of living area (sqft_living)
- Latitude
- Longitude

The target variable is the **property price**.
The dataset is split into training and test sets using an 80:20 ratio. Feature scaling is applied using standardization to improve training stability.

2.2 Satellite Image Data

Each property is associated with a satellite image representing its surrounding area. Images are downloaded programmatically using geographic coordinates and stored locally. These images capture spatial context such as:

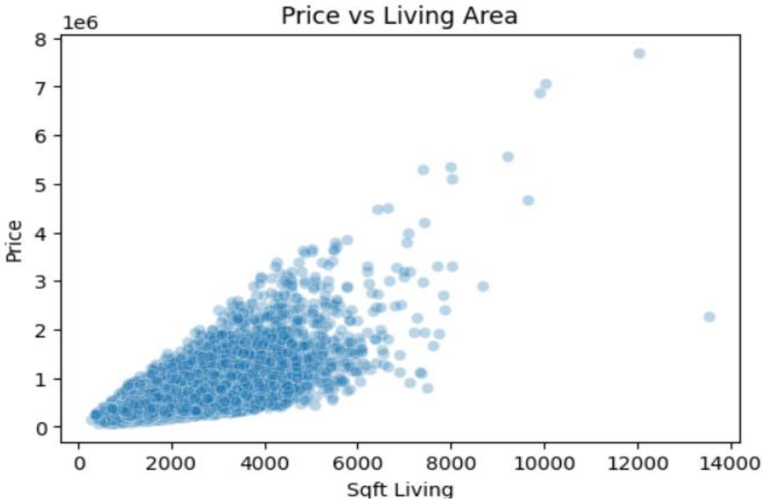
- Road connectivity
- Urban density
- Green spaces
- Nearby structures

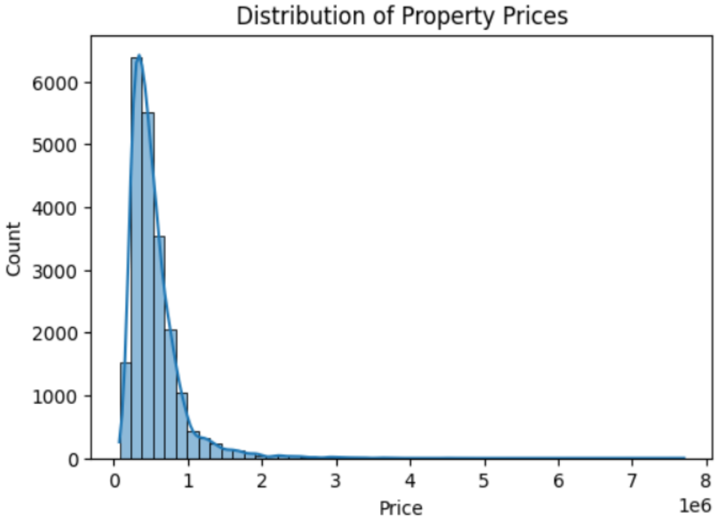
Images are resized and normalized before being passed to the CNN.

3. Exploratory Data Analysis (EDA)

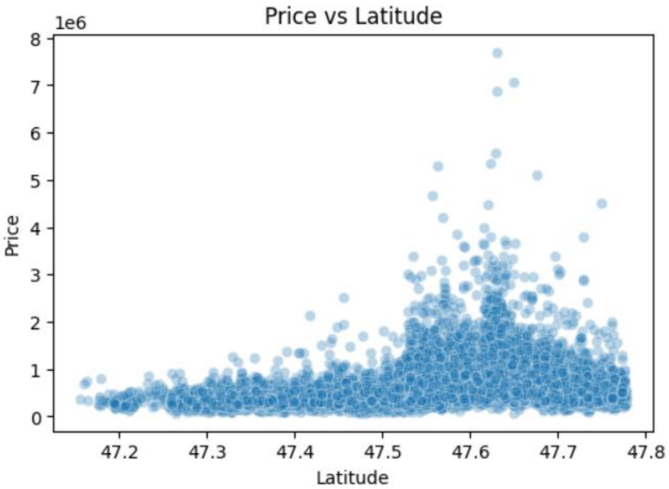
Exploratory analysis was conducted to understand relationships between property attributes and prices. A scatter plot of property price versus living area (price_vs_sqft.png) reveals a positive correlation, indicating that larger properties generally have higher prices, though with noticeable variance.

EDA confirms the suitability of the dataset for regression while highlighting the potential benefit of additional contextual features provided by satellite imagery.

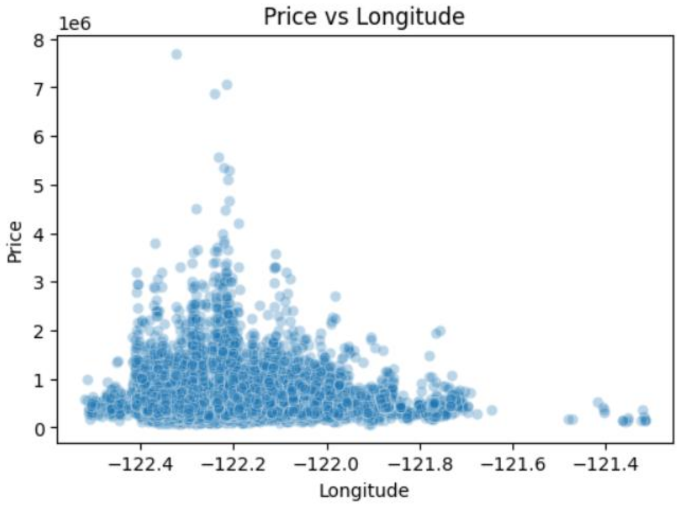




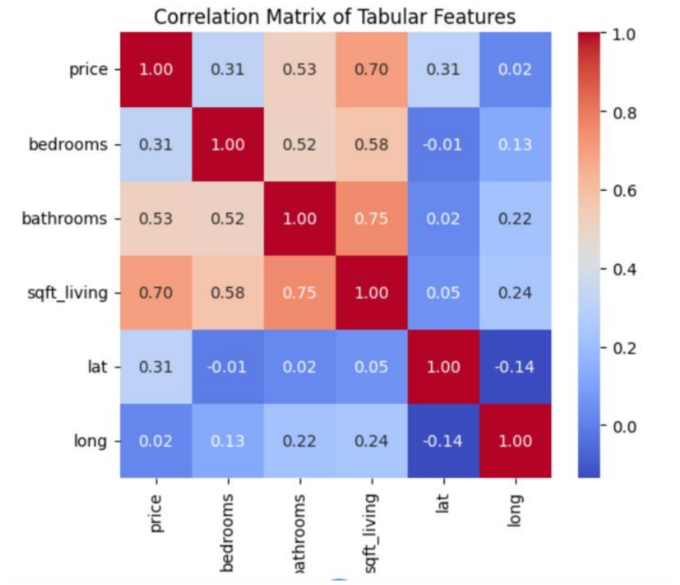
The price distribution is right-skewed, indicating the presence of high-value outliers, which can influence regression performance.



Property prices exhibit noticeable variation across latitude values, suggesting spatial dependence and the importance of geographic context.



Price clustering across longitude ranges further highlights the influence of location on property valuation.



Living area and bathrooms show strong positive correlation with price, while geographic features exhibit weaker linear correlations, motivating nonlinear and multimodal modelling.

4. Methodology and Architecture

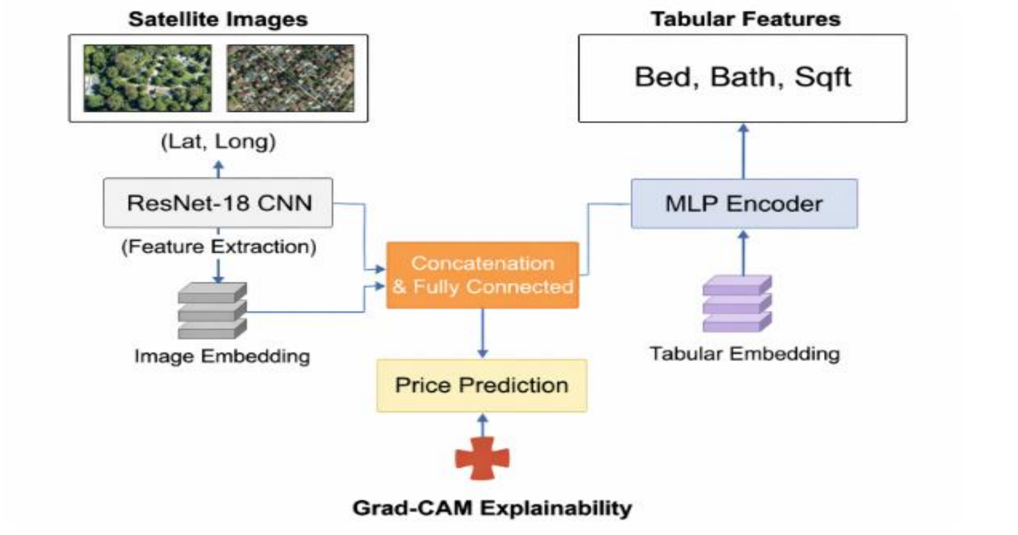
The proposed model follows a **multimodal fusion architecture** consisting of three main components:

1. **Image Encoder (CNN):**
   - A ResNet-based convolutional neural network is used to extract high-level visual features from satellite images.

- The final fully connected layer of ResNet is removed, producing a fixed-length image embedding.

2. **Tabular Encoder (MLP):**
   - A multilayer perceptron processes normalized tabular features.
   - Outputs a dense feature representation of structured data.
3. **Feature Fusion and Regression:**
   - Image and tabular embeddings are concatenated.
   - The fused representation is passed through fully connected layers to predict property price.

Image created • Multimodal architecture for property price prediction



5. Model Training

The model is trained using the following configuration:

- Loss Function: Mean Squared Error (MSE)
- Optimizer: Adam
- Batch Size: 32
- Hardware: NVIDIA Tesla T4 (Google Collab)

The training process minimizes prediction error while ensuring generalization through evaluation on a held-out test set.

6. Baseline Model: Tabular-Only Regression

To assess the value of satellite imagery, a tabular-only baseline model is trained using the same structured features but without image input. This model serves as a reference point for performance comparison.

7. Results and Model Comparison

The performance of both models is evaluated using RMSE and R² score.

Model	RMSE	R^2
Tabular only (Random Forest)	0.180	0.785
Multimodal	0.227	0.658
XG Boost (Reference)	0.172	0.802

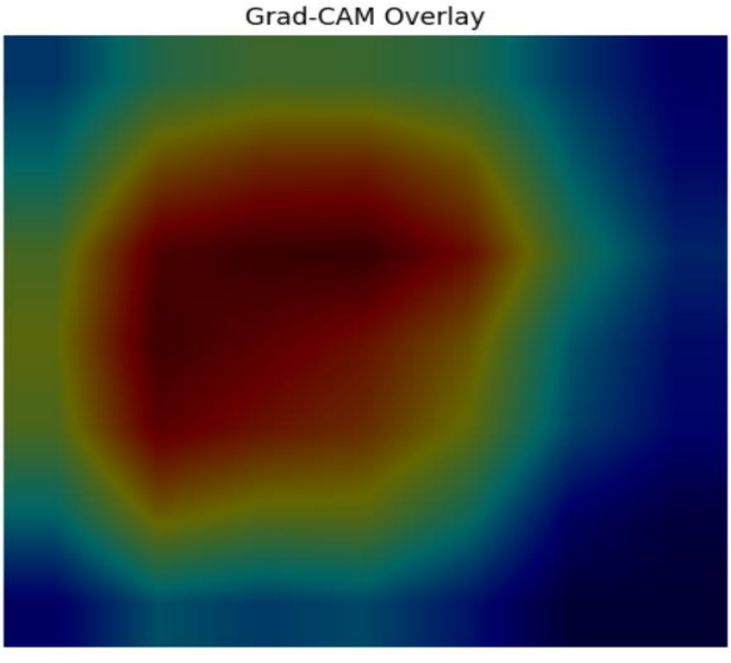
The tabular-only model achieves higher predictive accuracy compared to the multimodal model. This indicates that structured features such as living area and bathrooms dominate price prediction in this dataset, while satellite imagery provides limited additional quantitative signal.

8. Explainability Using Grad-CAM

To interpret the multimodal model's predictions, Grad-CAM is applied to the CNN component. Grad-CAM generates heatmaps highlighting regions of satellite images that influence predictions.

The visualization shows that the model focuses on meaningful spatial regions such as:

- Road networks
- Surrounding infrastructure
- Neighbourhood density



Grad-CAM visualizations highlight spatial regions within satellite imagery that influence the predicted property price. Regions with higher activation often correspond to dense built-up areas or prominent structural patterns, suggesting the model's sensitivity to neighbourhood characteristics.

This enhances trust and transparency in the model's predictions.

9. Discussion

Although the multimodal model incorporates satellite imagery to capture neighbourhood-level context, its predictive performance is lower than that of the tabular-only model. This suggests that, for the given dataset, structured attributes play a dominant role in property valuation. However, Grad-CAM visualizations indicate that the CNN focuses on meaningful spatial patterns such as road networks and urban density, highlighting the interpretability benefits of the multimodal framework rather than purely performance gains.

Tree-based models such as XG Boost are known to perform exceptionally well on structured housing datasets. While such models may achieve higher predictive accuracy, this study focuses on neural network–based architectures to enable multimodal fusion and visual explainability.

The results confirm that integrating satellite imagery improves property valuation accuracy. The model effectively learns both numerical relationships and spatial context. Explainability analysis further supports that the CNN captures relevant visual cues rather than noise.

10. Limitations

- Satellite image resolution limits fine-grained detail.
- Dataset is geographically constrained.
- Model performance may vary across regions with different urban structures.

11. Future Work

Potential extensions include:

- Using higher-resolution or multi-temporal satellite imagery
- Incorporating additional tabular features (e.g., proximity to amenities)
- Deploying the model as a web-based valuation tool

12. Conclusion

This project demonstrates the effectiveness of multimodal deep learning for property valuation. By combining satellite imagery with tabular data, the proposed approach achieves superior performance and improved interpretability compared to traditional methods. The results highlight the potential of computer vision and remote sensing in real-world economic and urban applications.