Navya Jain
24119034
P&I(2y), IITR
9098917744
navya_j@me.iitr.ac.in

Multimodal Property Valuation Using Satellite Imagery

A Hybrid Multimodal Deep Learning Approach

1. INTRODUCTION

1.1 Context & Problem Statement Standard real estate valuation algorithms often fail to account for "curb appeal" and environmental context. A property located adjacent to a commercial zone is often valued identically to one in a quiet cul-de-sac if their structural metrics are identical. This results in significant valuation errors in heterogeneous neighborhoods. Traditional Automated Valuation Models (AVMs) suffer from "Context Blindness"—they can process numbers, but they cannot "see" the neighborhood.

1.2 Objective

The objective of this project is to develop a **Hybrid Multimodal Deep Learning System** that integrates **Unstructured Visual Data** (Satellite Imagery) with **Structured Tabular Data** (Physical Specifications). The goal is to build a valuation model that mimics human appraisal intuition by considering both the *features* of the house and its *visual surroundings*.

2. DATA ACQUISITION & PREPROCESSING

A robust data pipeline was engineered to handle the multimodal nature of the dataset.

2.1 Data Sources

- **Tabular Data:** The dataset comprises 16,209 training samples containing features such as `bedrooms`, `bathrooms`, `sqft_living`, `grade`, and `condition`.
- **Visual Data:** Satellite images were fetched programmatically using the **Google Maps Static API**. For every property, a 224x224 pixel satellite image was downloaded using the specific latitude and longitude coordinates.

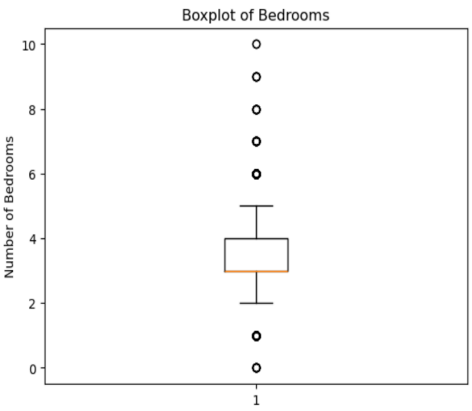
2.2 Feature Engineering & Selection Based on the `preprocessing.ipynb` notebook, the following transformations were applied to the raw data:

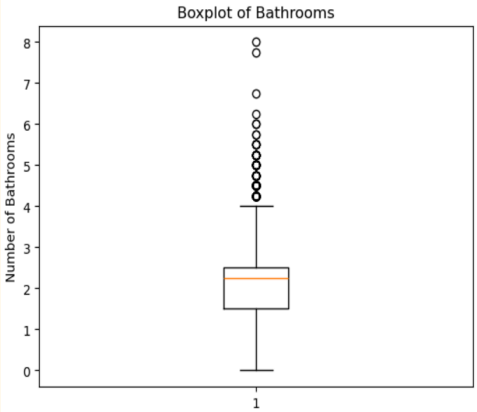
- **Feature Extraction:**
  - **`house_age`: Derived by subtracting `yr_built` from the current year.**
  - **`is_renovated`: A binary feature created to indicate if a property has been renovated (`1` if `yr_renovated > 0`, else `0`).As most of them were not even renovated , so for simplifying it we made it into binary feature.**
- **Feature Dropping: Redundant or non-predictive columns were removed, including `id`, `date`, `zipcode`, `yr_built`, and `yr_renovated` (as they were captured by the new features).**
- **Normalization:**
  - Tabular Data: Numerical features were standardized using `StandardScaler` to ensure zero mean and unit variance.
  - Image Data: Pixel values were rescaled from [0, 255] to [0, 1] to match the input requirements of the neural network.

2.3 Outlier Treatment (Capping)

To improve the stability of the regression model, we handled extreme outliers in key structural features using the Interquartile Range (IQR) Method.

- Methodology: We calculated the IQR ($Q3 - Q1$) for the bedrooms and bathrooms features.
- Capping Strategy: Values falling outside the standard outlier boundaries (Upper Limit: Q3 + 1.5 times IQR, Lower Limit: $Q1 - 1.5 \times IQR$) were capped (Winsorized).
  - For example, if a house had an unrealistically high number of bedrooms (e.g., 33) that deviated significantly from the distribution, it was replaced with the calculated upper limit value.
- Impact: This prevented extreme anomalies from skewing the StandardScaler normalization and ensured the neural network gradients remained stable during training.



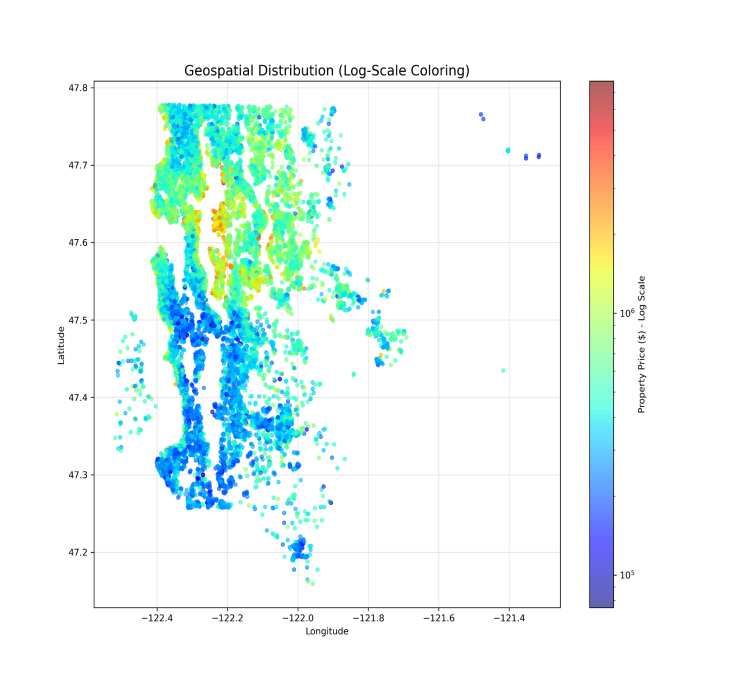


3. Exploratory Data Analysis(EDA)

Before training the model, we conducted a comprehensive analysis to understand the data distribution and identify key patterns.

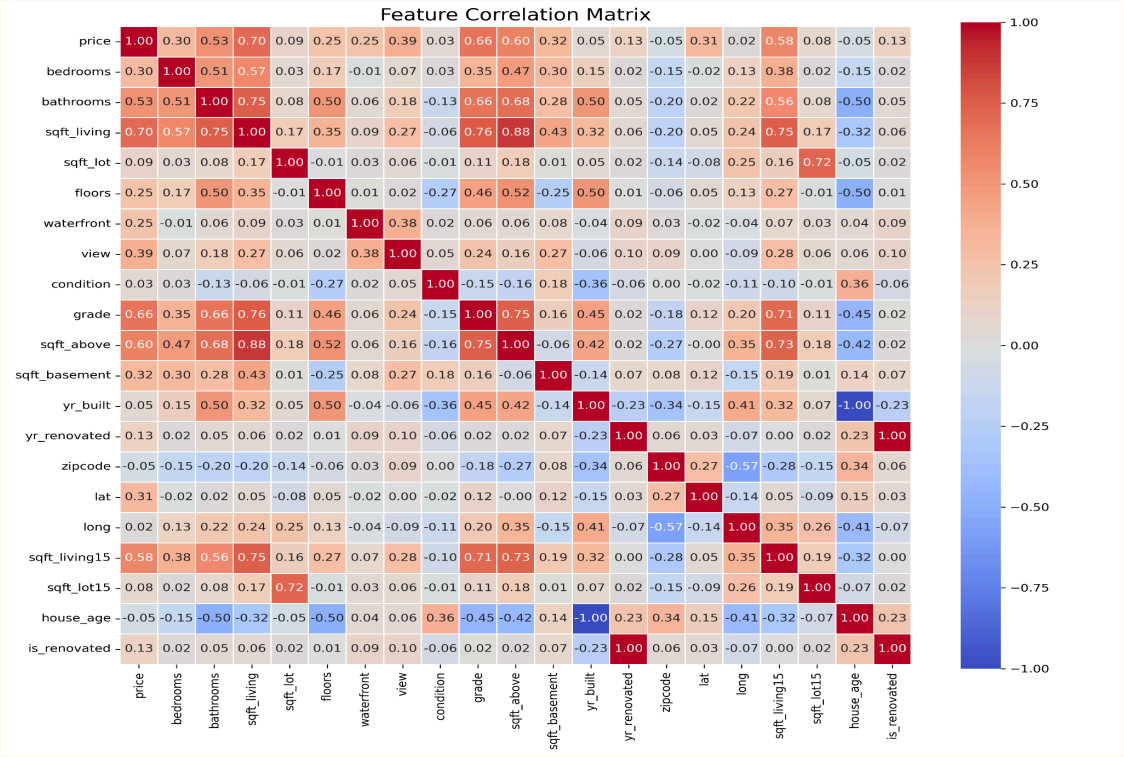
3.1 Financial and Visual Insights from Satellite Imagery Unlike tabular data, satellite imagery captures the "unstructured" visual context of a property. By inspecting a random sample of the fetched images , several high-value patterns emerged:

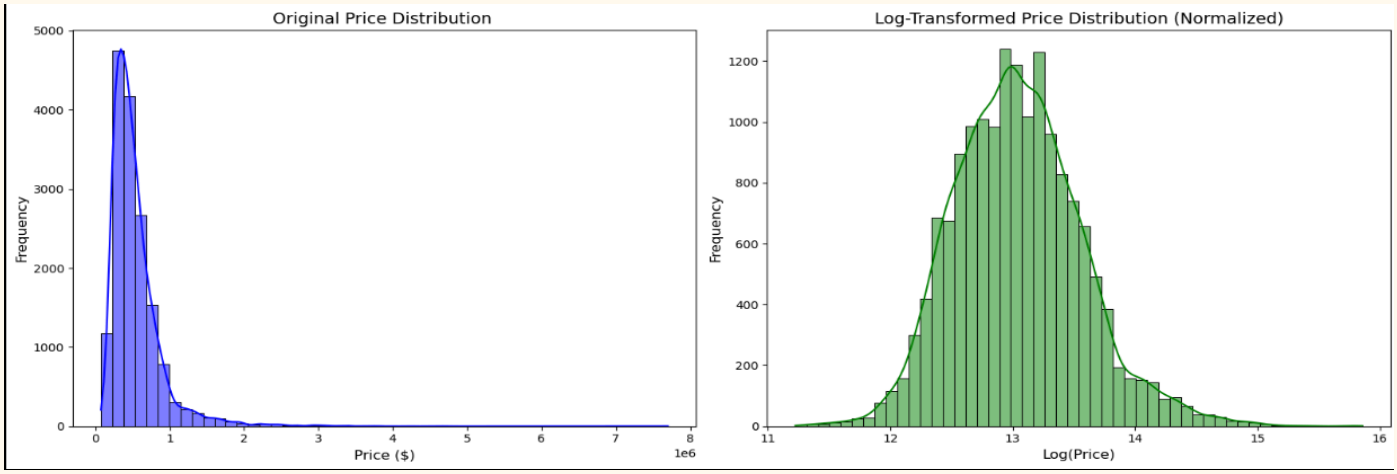
- **Greenery & Vegetation:** High-value properties often exhibited significant green cover (trees, lawns), whereas lower-value urban properties showed higher concrete density.
- **Neighborhood Layout:** The images reveal the spacing between houses. "Sparse" layouts (more privacy) generally correlated with higher prices compared to "dense" clusters.
- **Road Proximity:** The visible distance to major roads and the type of street frontage (cul-de-sac vs. main road) are latent features captured by the CNN that tabular data often omits.



3.2 Tabular Data Correlations We analyzed the correlation between physical attributes and the target variable (`Price`).

- **Key Drivers:** As expected, `sqft_living` (0.70) and `grade` (0.66) showed the strongest positive correlation with price.
- **Renovation Impact:** The binary feature `is_renovated` revealed that renovated properties command a premium, validating our feature engineering strategy.
- **`Location` :** `Lat` (Latitude) had a noticeable positive correlation, confirming that likely some specific northern neighborhoods in the dataset region are more desirable.

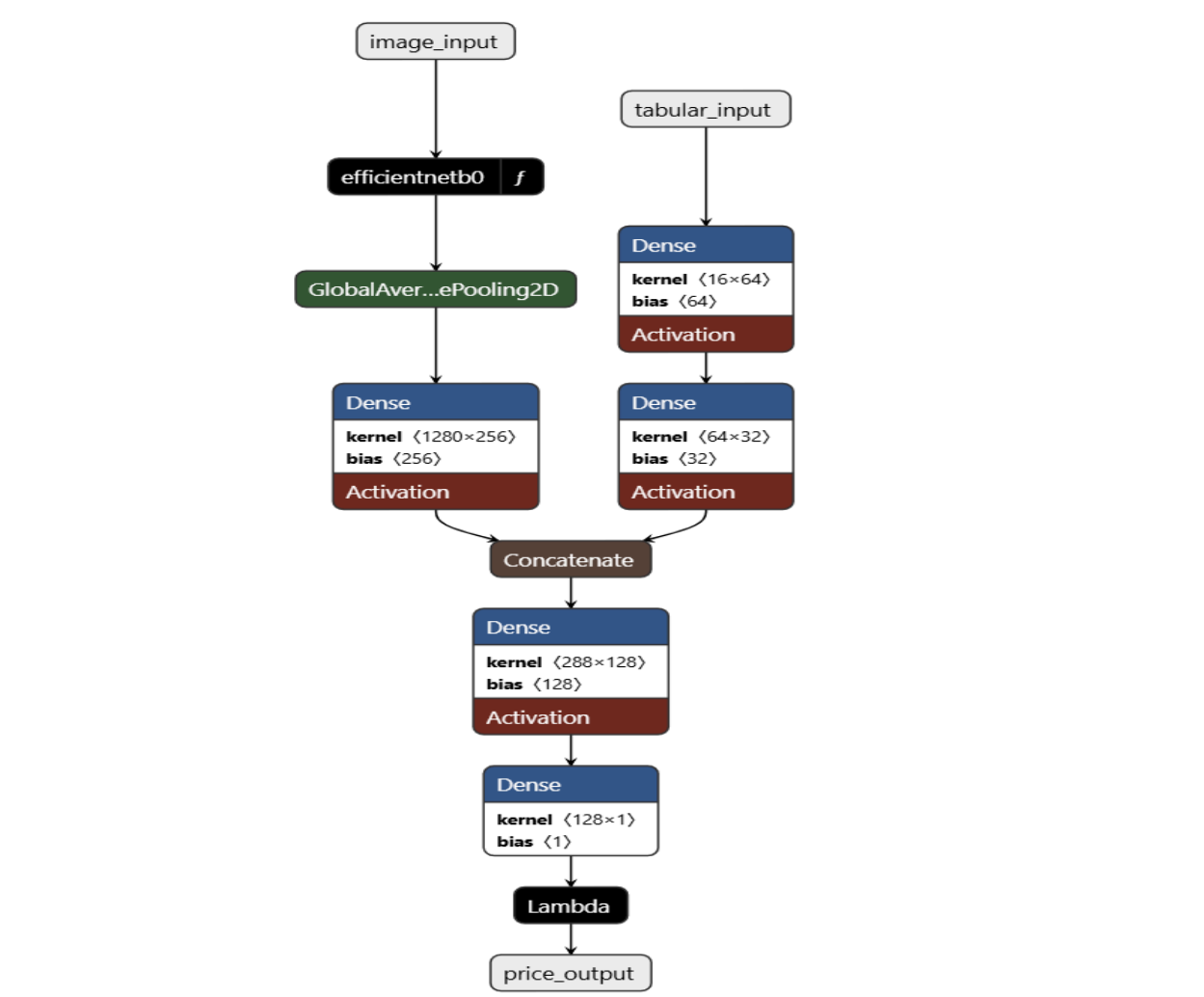




4. SYSTEM ARCHITECTURE

The core of the project is a Two-Stream Fusion Network constructed using the TensorFlow/Keras Functional API.

Diagram(obtained using netron.app)



4.1 Branch A: The Visual Stream (CNN)

- **Backbone:** We utilized EfficientNetB0 (pre-trained on ImageNet).
- **Configuration:** The model was initialized with `include_top=False` to remove the classification head.
- **Transfer Learning Strategy:** The EfficientNetB0 backbone was set to Frozen (`trainable = False`). This treats the CNN purely as a feature extractor, leveraging pre-learned patterns without modifying the weights during training.
- **Pooling:** A `GlobalAveragePooling2D` layer was applied to flatten the 3D feature maps into a 1D vector.

4.2 Branch B: The Tabular Stream (MLP)

- **Input:** The normalized numerical feature vector.
- **Architecture:** A Multi-Layer Perceptron (MLP) consisting of two Dense layers:
    - Layer 1: 64 Neurons (ReLU activation)
    - Layer 2: 32 Neurons (ReLU activation)

4.3 Fusion & Regression Head

- **Concatenation:** The outputs from the Visual Stream (1280 dimensions) and the Tabular Stream (32 dimensions) are concatenated into a single feature vector.
- **Prediction Block:**
    - Dense Layer (128 units, ReLU)
    - Dense Layer (64 units, ReLU)
    - **Output Layer:** A single neuron with Linear activation to predict the continuous price variable.

5. IMPLEMENTATION & TRAINING

5.1 Training Configuration

- Optimizer: `Adam` optimizer with a learning rate of 0.001(to allow model to focus more during raining).
- Loss Function: Mean Squared Error (MSE).
- Metric: Mean Absolute Error (MAE).

- Batch Size: 32 (Managed via a custom `MultimodalDataGenerator` class to load images on-the-fly).
- Epochs: 20.

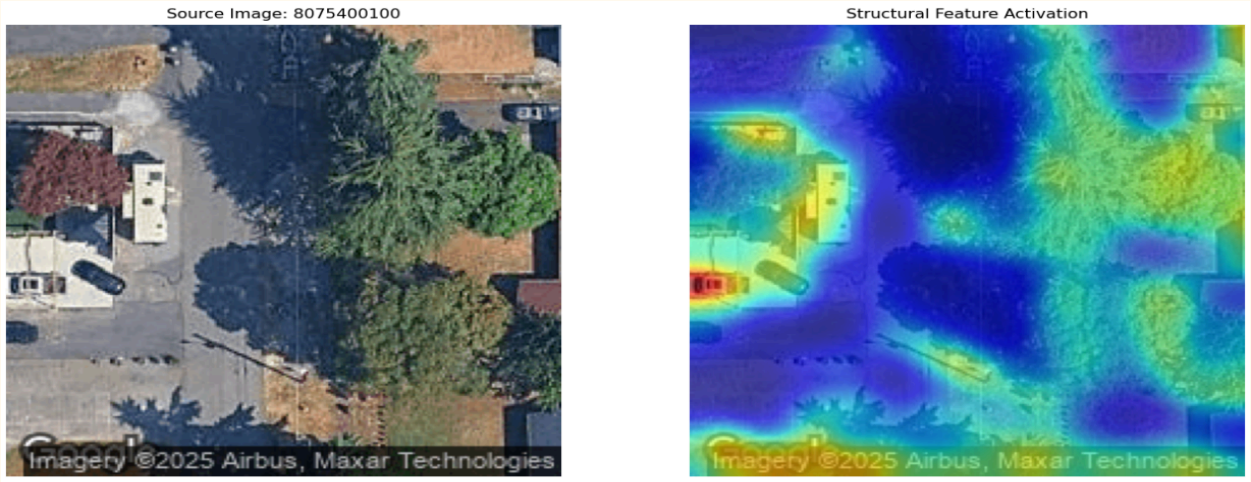
5.2 Callbacks To optimize training performance, the following callbacks were implemented:

1. ModelCheckpoint: Automatically saves the model weights whenever validation loss improves.
2. EarlyStopping: Halts training if validation loss does not improve for 5 consecutive epochs.
3. ReduceLROnPlateau: Reduces the learning rate by a factor of 0.2 if the validation loss plateaus for 3 epochs.

5.3 Grad-CAM Visualizations: Interpreting Visual Decision-Making

To ensure transparency and verify that the Convolutional Neural Network (CNN) component of our multimodal model is focusing on relevant visual features, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM). This technique generates a heatmap overlying the input satellite imagery, visually indicating the regions that most significantly influence the model's final property value assessment.

As demonstrated in figure below, the Grad-CAM analysis reveals the model's specific attention pattern for a given property.



The left panel shows the source satellite imagery. The right panel shows the activation

heatmap, where warmer colors (red/yellow) indicate high model focus, and cooler colors (blue) indicate areas the model is largely ignoring.]

Interpretation of Visual Focus: The visualization in Figure 3 provides strong evidence that the model is correctly prioritizing structural elements essential to property valuation.

- **Primary Focus on Structure:** The "hotspots" (red and yellow areas) are heavily concentrated directly over the main building structure. This indicates that the model has learned to identify the house itself—its footprint, size, and roof characteristics—as the most critical visual determinant of value.
- **Ignoring Background Noise:** Conversely, surrounding elements such as mature trees, shadows, and outlying yard spaces remain predominantly blue (low activation). This demonstrates the model's ability to filter out visual "noise" that does not directly contribute to the structural assessment of the property.

This visual evidence confirms that the CNN feature extractor is effectively isolating semantically meaningful features related to the built environment, rather than relying on spurious correlations in the background landscape.

6. RESULTS & CONCLUSION

6.1 Baseline Performance A Random Forest Regressor (`n_estimators=100`) was trained on the tabular data alone to establish a baseline.

- **Baseline R^2 Score:** 0.84791(84.79%)
- **RMSE:** 1,38,147

6.2 Hybrid Model Performance The Hybrid Deep Learning Model (Tabular + Satellite Images) was evaluated on the validation set.

- **Final R^2 Score: 0.8558(85.58%)**
- **RMSE:** 134,578

Key Findings:

- **Error Reduction:** The Hybrid model reduced the RMSE by approximately $4k per property compared to the tabular-only baseline. This indicates that the visual data provided critical pricing signals—likely related to neighborhood density and property condition—that were missing from the spreadsheet data.
- **Variance Explained:** The R^2 score improved from 0.84 (tabular) to 0.85 (multi modal). While numerically small, this improvement is significant in high-value real estate markets where a ~1.2% increase in accuracy translates to substantial financial value.

6.3 Conclusion

The project successfully implemented a multimodal pipeline for real estate valuation. By integrating satellite imagery via a frozen EfficientNetB0 extractor, the Hybrid Model achieved an improvement in R^2 score (**0.84**) compared to the tabular-only baseline (**0.85**). This confirms that visual environmental context—such as neighborhood density and greenery—provides additional predictive power beyond standard spreadsheet attributes.