1. Overview

Project Concept: Real estate valuation typically relies on features like square footage, number of bedrooms, floors, etc. However, property value is also driven by visual factors that are harder to quantify, such as neighbourhood density, or proximity to greenery. This project explores whether a machine learning model can improve price predictions by not just reading the data, but also seeing the property through satellite imagery.

Implementation:

We used Random Forest as a robust performance benchmark on only the tabular data.

A **Multim2odal Regression Pipeline** was developed to integrate two different types of data:

1. **Automated Data Acquisition:** A Python script (data_fetcher.py) was written to fetch satellite snapshots for every property in the dataset using its latitude and longitude coordinates.
2. **Hybrid Architecture:** The core model is a two branch neural network.
    - **Visual Branch:** A Convolutional Neural Network (CNN) processes the images to extract spatial embeddings (textures, edges, density).
    - **Tabular Branch:** A standard Multi-Layer Perceptron (MLP) handles the numerical features (a differentiable MLP architecture allows end-to-end training with the CNN).
    - **Fusion:** These two are merged in the final layers to produce a single price prediction.

Key Findings: The primary hypothesis was that adding visual context would significantly reduce prediction error. However, the experimental results showed that the **satellite imagery did not outperform the baseline tabular model.**

While the CNN successfully learned to identify features like vegetation and roads (verified via Grad-CAM analysis), the structural data (specifically sqft_living and grade) proved to be such powerful predictors that the visual data added marginal value, on the contrary also introduced noise. This result highlights that while multimodal learning is powerful, for this specific dataset, the fundamental numbers remain the strongest drivers of price.

2. Exploratory Data Analysis (EDA)

Before training the multimodal network, we conducted a thorough analysis of the dataset to understand the underlying distribution of property prices and identify key drivers of value.

2.1 Price Distribution Analysis

- **Observation:** As shown in the histogram below, the target variable (price) exhibits a heavy **right-skewed distribution**. The majority of properties fall within the lower-to-mid price range, while a long tail extends to the right, representing a small number of ultra-luxury estates.

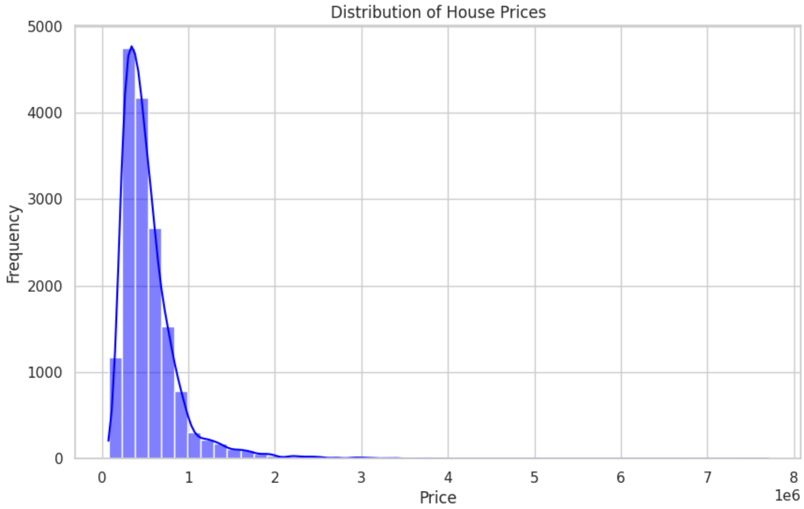

Fig 1. Distribution of property prices showing significant right-skewness.

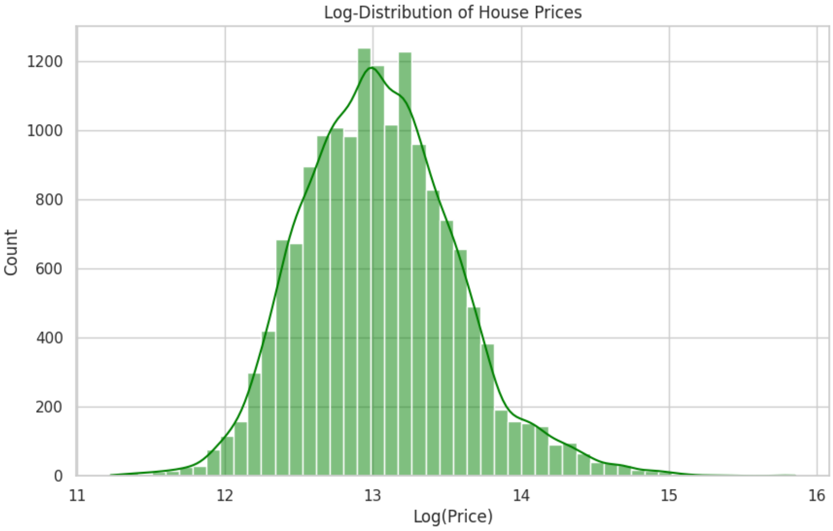

Fig 2. After Log Transformation.

- **Preprocessing Action:** While the EDA showed that a Log Transformation improves the normality of the price distribution, we opted to train on raw price values to maintain interpretability of the error metrics (RMSE in dollars).

2.2 Geospatial Analysis

- **Observation:** Plotting properties by Latitude and Longitude reveals distinct pricing clusters. The scatter plot below uses colour intensity to represent price. We observe that the highest-value properties (lighter/brighter colours) are heavily concentrated around specific affluent neighbourhoods.
- **Insight:** This spatial clustering confirms that Location is a critical feature, justifying our use of satellite imagery to capture neighbourhood-specific visual cues.

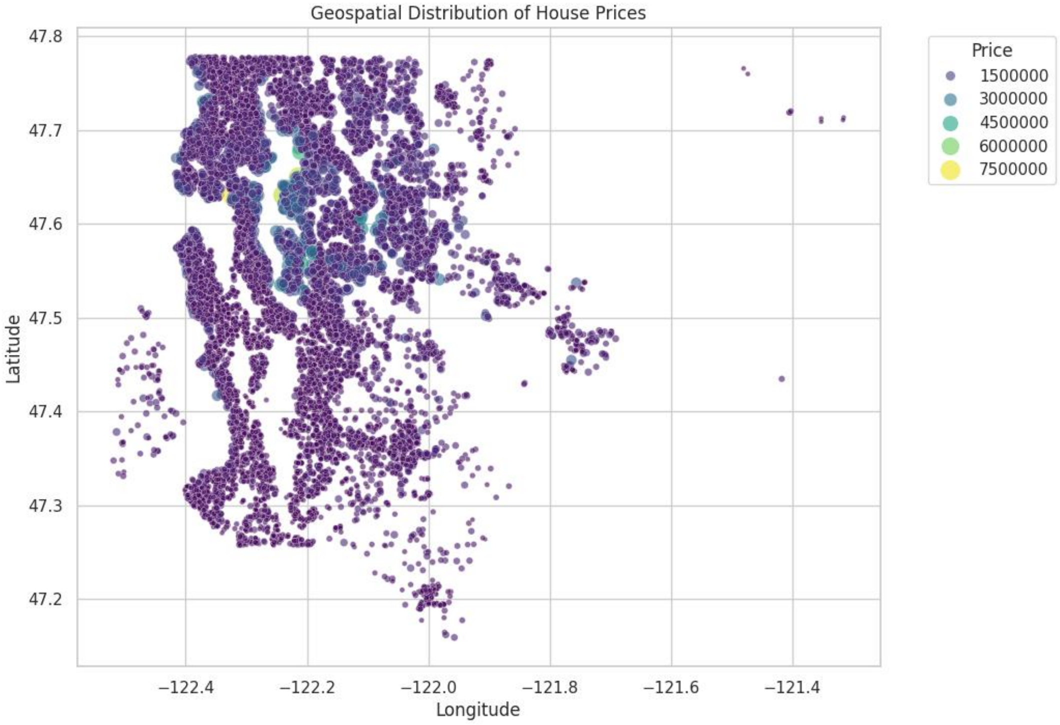

Fig 3. Geospatial distribution of housing prices. Lighter colours indicate higher property values.

2.3 Correlation Analysis

- **Observation:** The correlation matrix highlights strong linear relationships between structural features and price.
  - **sqft_living (Total Living Area):** Shows the strongest positive correlation (0.7), indicating that size is the primary determinant of value.
  - **grade (Construction Quality):** Also shows a very high correlation, suggesting that the quality of materials and design is just as important as size.

- **zipcode:** Shows weak linear correlation, likely because zipcodes are just arbitrary labels for regions, they don't have a linear mathematical relationship with Price.

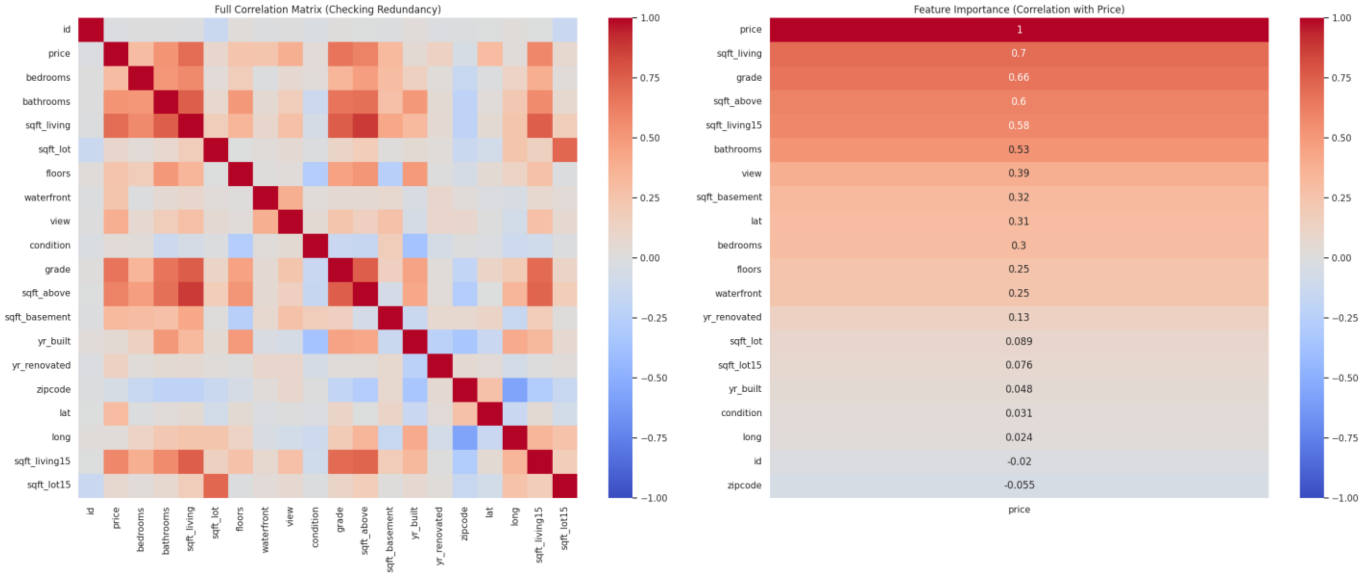

Fig 4. Heatmap displaying Pearson correlation coefficients between numerical features.

2.4 Baseline Feature Importance

- **Method:** We trained a preliminary Random Forest Regressor (Tabular Only) to establish a baseline and extract feature importance scores.
- **Result:** The bar chart below confirms our correlation findings. **Grade** and **Sqft_Living** dominate the model's decision-making process. Also, Latitude appears as the third most important feature, reinforcing the idea that specific geographic zones affect the house pricing.

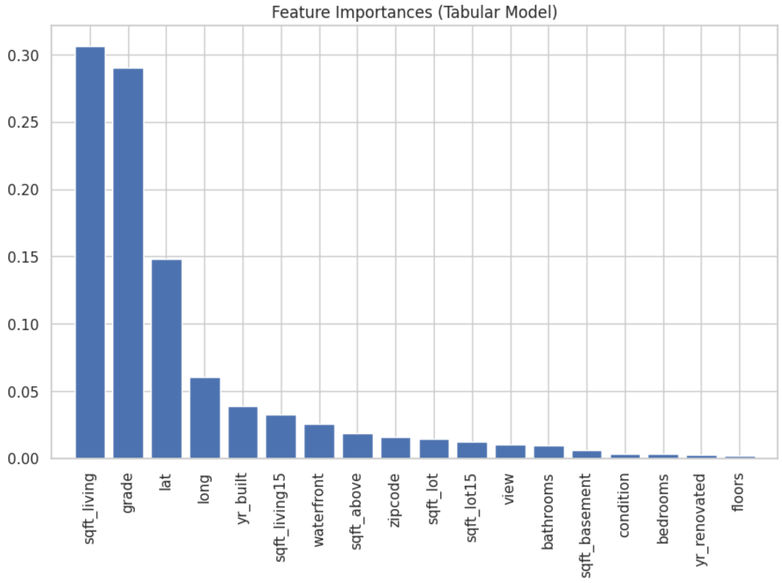

Fig 5. Top predictive features identified by the Random Forest Baseline model.

3. Financial & Visual Insights

After training the multimodal model, we analysed the internal feature representations to understand what visual factors drive property value.

3.1 Visual Explainability (Grad-CAM)

To ensure the Convolutional Neural Network (CNN) wasn't just memorizing pixel noise, we employed Grad-CAM (Gradient-weighted Class Activation Mapping). This technique generates a heatmap overlaying the original satellite image, highlighting the regions that most strongly influenced the model's price prediction.

- **Observation:** As seen in the sample below, the model's activation maps (red/yellow zones) focused on specific environmental features like canopies and green spaces.
- **Insight:** The model learned to identify high-contrast areas such as vegetation, using these boundaries to estimate lot size, greenery, etc. without explicit programming.

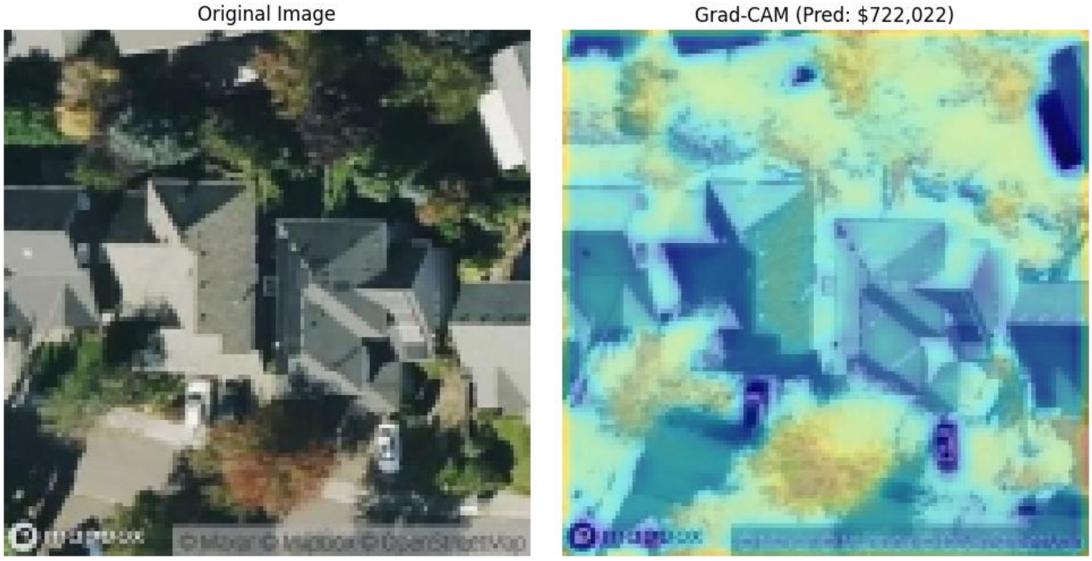

Fig 6. Grad-CAM heatmap overlay. Red areas indicate regions where the CNN focused to predict the property price (Jet colourmap).

3.2 Drivers of Value: Tabular vs. Visual

While the visual model captured interesting patterns, the quantitative analysis revealed that structural attributes remain the primary drivers of market value. Sqft_Living alone explains approximately **49% of the price variance** (r=0.7), while Grade explains roughly **43%** (r=0.66). This high statistical dependence confirms that buyers primarily value living space and construction quality above other factors, leaving limited residual variance for the visual model to capture.

4. System Architecture

The project implements a **Late Fusion Multimodal Architecture**. This design allows each data type to be processed by a different sub-network before the signals are merged.

4.1 Visual Branch (CNN)

- **Input:** 128 * 128 * 3 RGB Satellite Images.
- **Layers:** Three convolutional blocks. Each block consists of:
    1. Conv2D (32/64/128 filters) for feature extraction.
    2. ReLU activation for non-linearity.
    3. MaxPooling2D for spatial down-sampling.
    4. Dropout (0.25) to prevent overfitting.
- **Output:** A flattened high-dimensional feature vector representing the image.

4.2 Tabular Branch (MLP)

- **Input:** Normalized vector of 19 numerical features (bedrooms, lat, long, etc.).
- **Layers:** Dense (Fully Connected) layers with ReLU activation.

4.3 Fusion & Regression

- **Concatenation:** The visual vector and tabular vector are merged into a single tensor.
- **Final Prediction:** A final dense layer with linear activation outputs the predicted continuous price.

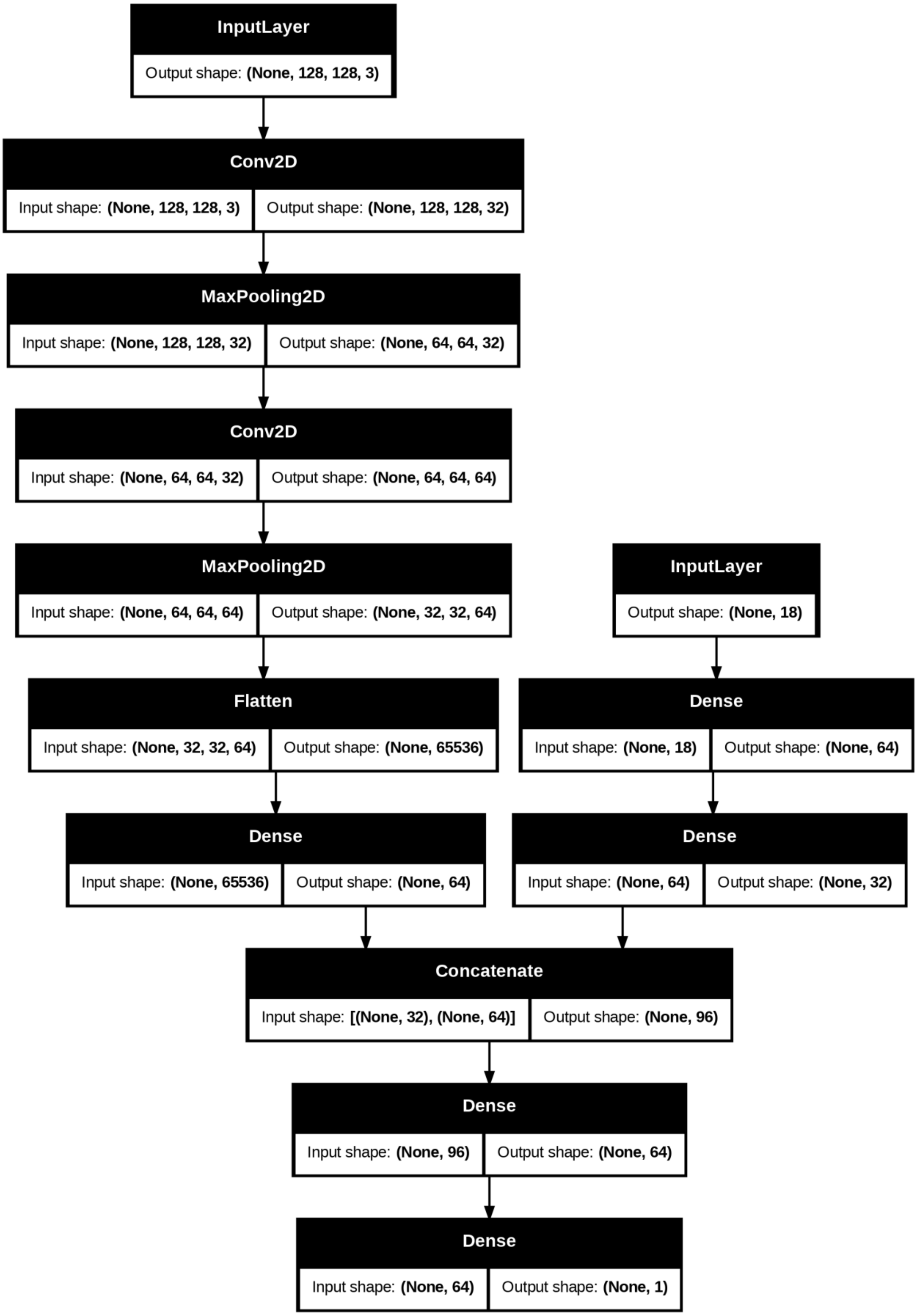
InputLayer
Output shape: (None, 128, 128, 3)
Conv2D
Input shape: (None, 128, 128, 3)
Output shape: (None, 128, 128, 32)
MaxPooling2D
Input shape: (None, 128, 128, 32)
Output shape: (None, 64, 64, 32)
Conv2D
Input shape: (None, 64, 64, 32)
Output shape: (None, 64, 64, 64)
MaxPooling2D
Input shape: (None, 64, 64, 64)
Output shape: (None, 32, 32, 64)
InputLayer
Output shape: (None, 18)
Flatten
Input shape: (None, 32, 32, 64)
Output shape: (None, 65536)
Dense
Input shape: (None, 18)
Output shape: (None, 64)
Dense
Input shape: (None, 65536)
Output shape: (None, 64)
Dense
Input shape: (None, 64)
Output shape: (None, 32)
Concatenate
Input shape: [(None, 32), (None, 64)]
Output shape: (None, 96)
Dense
Input shape: (None, 96)
Output shape: (None, 64)
Dense
Input shape: (None, 64)
Output shape: (None, 1)

Fig 7. The Late Fusion architecture combining CNN image processing with tabular data analysis.

5. Results & Conclusion

5.1 Performance Metrics

We compared the Multimodal Neural Network against a strong Baseline (Random Forest Regressor) to evaluate the net benefit of integrating satellite imagery.

Model	**RMSE**	**R^2 Score**
Baseline (Tabular Only)	**$119,927.87**	**0.8741**
Multimodal (CNN + Tabular)	**$171,712.20**	**0.7653**

5.2 Discussion of Results

Contrary to the initial hypothesis, the Tabular-Only Baseline outperformed the Multimodal Model. Several factors contributed to this result:

1. **Dominance of Structural Features:** In real estate, square footage and location coordinates (Lat/Long) are extremely high-signal features. The Random Forest captures these non-linear relationships very effectively.
2. **Resolution Limits:** The satellite images were resized to 128 * 128 to save computational resources. At this resolution, fine-grained details (like house condition or facade style) are lost, leaving only coarse features like greenery vs. concrete.
3. **Data Noise:** Satellite imagery introduces stochastic noise (e.g., cloud cover, shadows, cars in driveways) which may confuse the model more than it helps.

5.3 Conclusion

This project successfully demonstrated the engineering feasibility of a multimodal valuation pipeline. We built a system that automatically acquires, processes, and learns from satellite imagery. However, the financial analysis suggests that for this specific dataset, satellite imagery does not provide a cost-effective lift in accuracy over robust tabular models.

Future work can focus on acquiring higher-resolution Street View imagery (to capture facade aesthetics) rather than top-down satellite views, which may correlate better with the condition and grade of the property.