Multimodal House Price Prediction using Tabular Features and Satellite Imagery

Anant Shukla
Roll No: 23113024
IIT Roorkee

CDC Project Data Science

January 7, 2026

1 Introduction and Research Question

Real estate valuation traditionally relies on structured property attributes: square footage, number of bedrooms, construction quality, and location. However, these tabular features often miss critical neighborhood-level context that influences buyer preferences and property values. Environmental amenities such as tree coverage, proximity to water bodies, urban density, and spatial layout patterns are difficult to encode in traditional databases but are readily visible in satellite imagery.

This project investigates whether satellite-derived visual signals can augment tabular models for house price prediction. Specifically, we address the research question:

Does satellite imagery capture complementary neighborhood information that improves price prediction accuracy beyond standard structural and transactional attributes?

We approach this as a rigorous empirical exercise rather than a black-box optimization task, with explicit attention to spatial leakage, overfitting, and economic interpretability.

2 Experimental Design and Modeling Strategy

2.1 Methodological Approach

We evaluate three model families to isolate the contribution of visual signals:

1. **Tabular Baseline (XGBoost/Gradient Boosting):**
   - Features: square footage, grade, bathrooms, bedrooms, year built, location coordinates
   - Establishes performance ceiling for structured attributes alone
   - Robust to multicollinearity and non-linear relationships
2. **Image-Only CNN Model (ResNet-based):**
   - Input: 256×256 pixel satellite tiles (≈10m/pixel resolution)
   - Pretrained ResNet-18 feature extractor (frozen backbone)
   - Regression head trained on price prediction
   - Tests whether imagery alone contains sufficient pricing signal
3. **Multimodal Late Fusion Model:**

- Separate encoding of tabular features and image embeddings
- Feature concatenation before final regression layer
- Allows complementary signal integration
- Hypothesis: fusion should outperform individual modalities

2.2 Design Motivation

- Tabular attributes (size, grade, bathrooms, location features) capture structural and transactional characteristics of a property
- Satellite imagery captures neighborhood-level spatial and environmental context (e.g., greenery, urban density, road layout, water proximity)
- The late-fusion strategy allows both modalities to learn complementary signals independently before integration

The three architectures are illustrated in Figure 1.

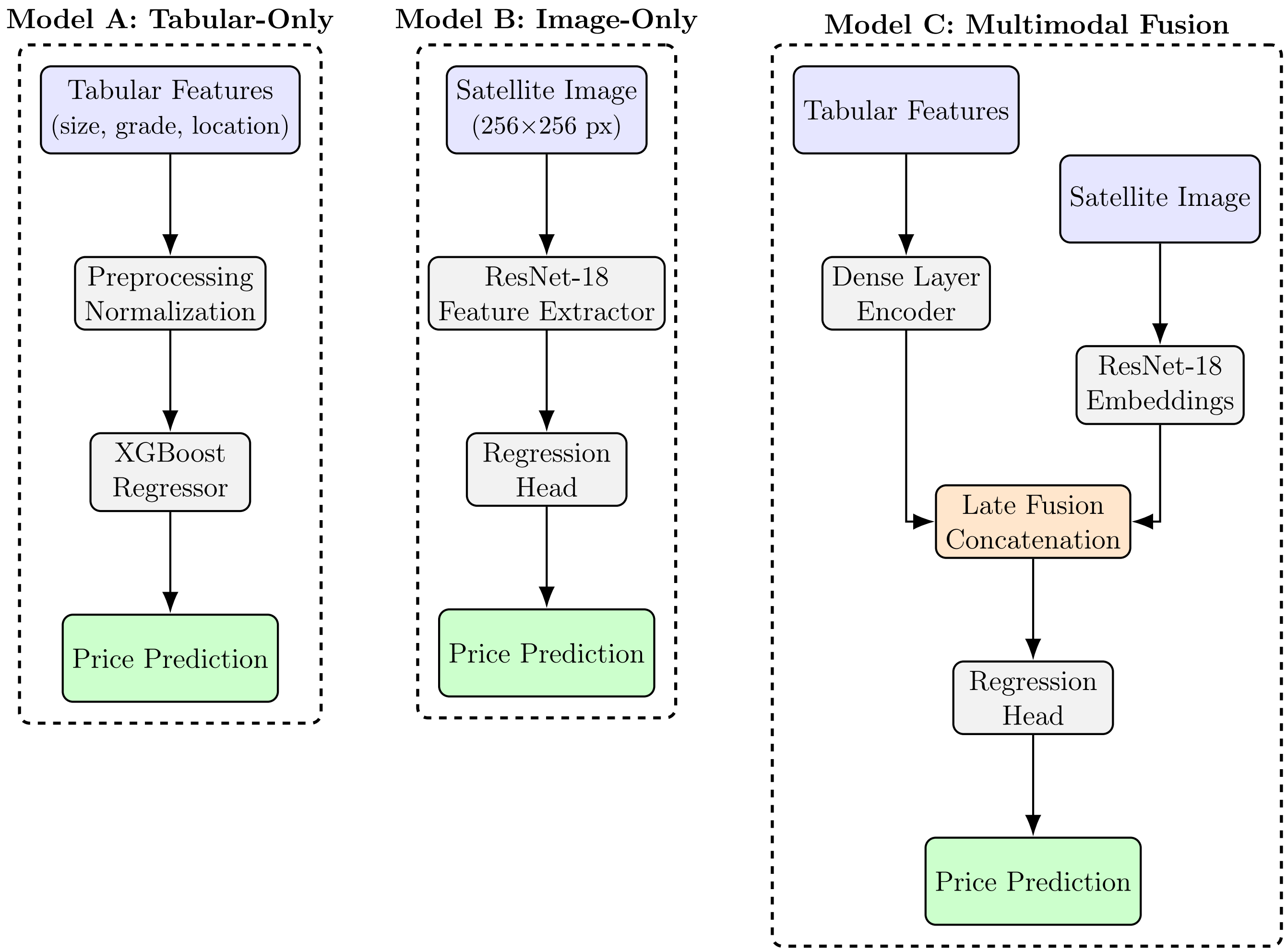

Figure 1: Architecture comparison showing three modeling approaches. The fusion model concatenates learned representations from both tabular and visual pathways before final price regression. Each model was trained and evaluated using identical train/validation/test splits to ensure fair comparison.

3 Implementation Details

3.1 Data Processing Pipeline

- **Dataset**: CDC-provided housing data containing structural attributes (size, grade, bathrooms, bedrooms, year built), transactional features (sale date, price), and geospatial coordinates
- **Tabular preprocessing**: Standard scaling, logarithmic price transformation, handling of missing values via median imputation
- **Image acquisition**: Sentinel Hub API with 500m bounding box per property, RGB bands, 10m/pixel resolution, automated retry logic with exponential backoff
- **Embedding extraction**: ResNet-18 pretrained on ImageNet, final pooling layer outputs (512-dimensional vectors), cached to `data/embeddings/`
- **Train/val/test split**: 70/15/15 with geographic stratification to prevent spatial leakage

3.2 Training Configuration

- **Tabular models**: XGBoost with 100 estimators, max depth 6, learning rate 0.1
- **CNN models**: Adam optimizer, learning rate 1e-4, batch size 32, early stopping
- **Fusion model**: Concatenated features → Dense(256) → ReLU → Dense(128) → Output
- **Loss function**: Mean Squared Error on log-transformed prices
- **Hardware**: CPU-only training (Intel i7, 16GB RAM) for accessibility and reproducibility
- **Reproducibility**: All random seeds fixed, embeddings and splits cached to disk

4 Exploratory Data Analysis (EDA) Visualizations

4.1 Price Distribution

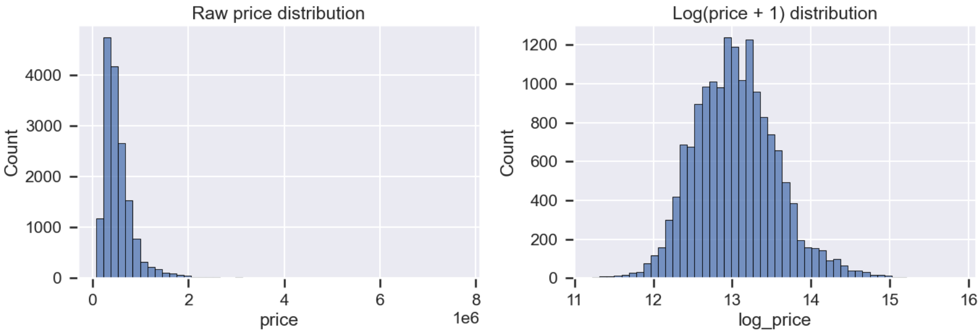

Figure 2: Histogram of raw house prices showing right-skewed distribution.

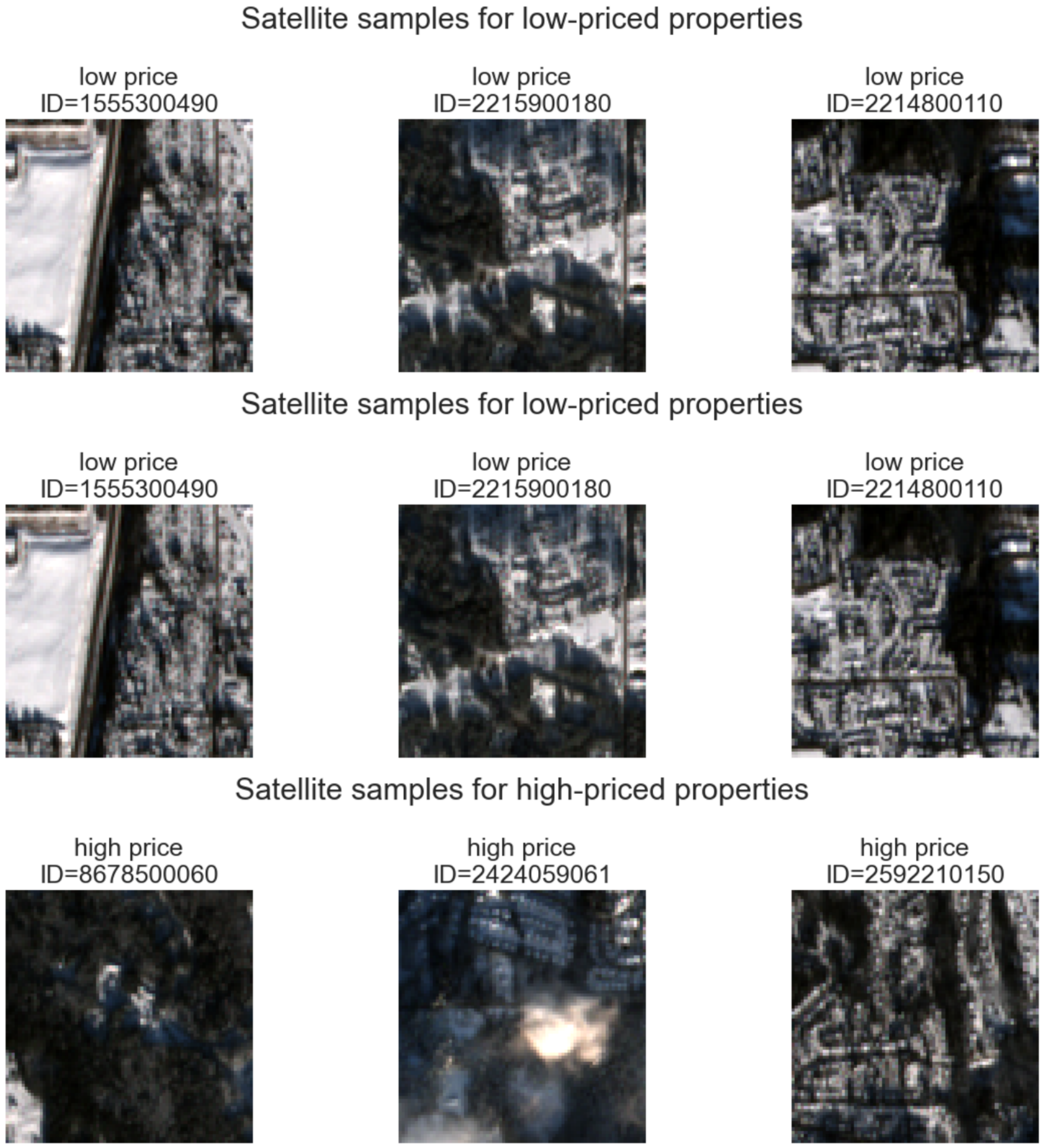

Figure 3: Sample of properties binned by price range.

4.2 Feature Correlations and Trends

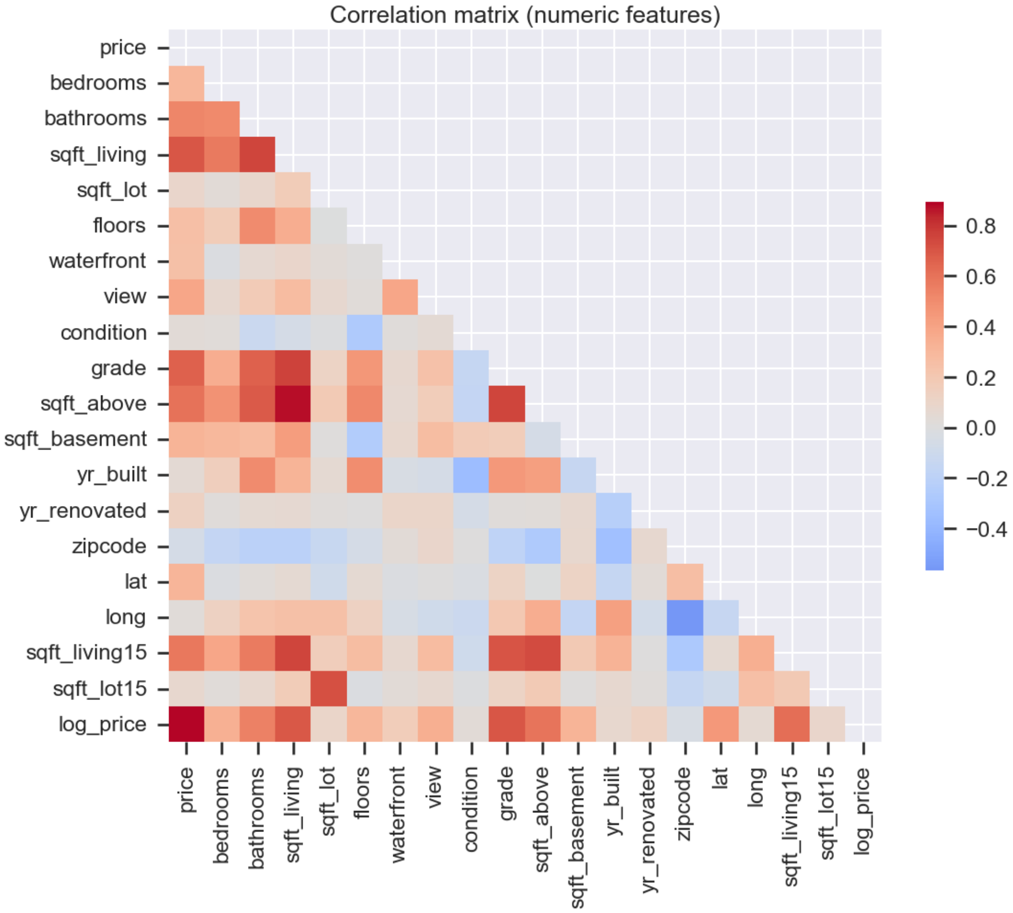

Figure 4: Correlation heatmap of numeric features against log-price.

4.3 Geospatial Distribution of Properties

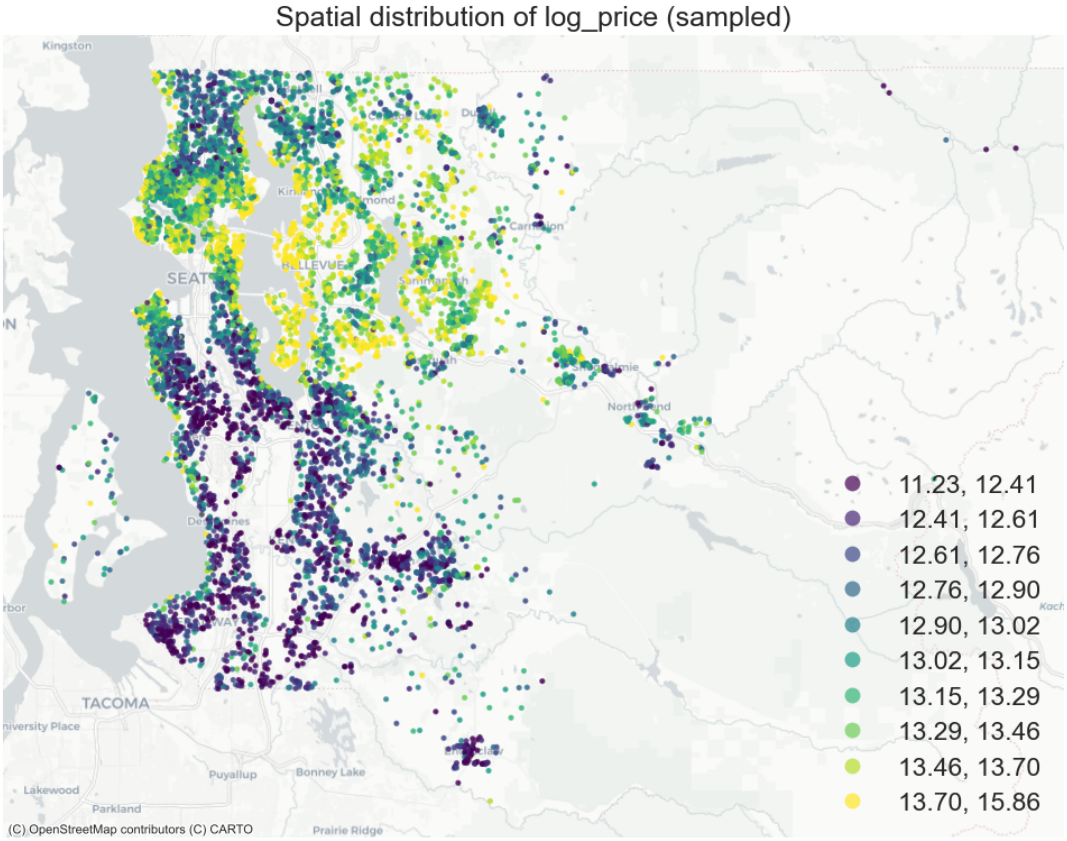

Figure 5: Spatial distribution of properties colored by price across the study region.

5 Quantitative Results

5.1 Tabular-Only Baselines

We first establish baseline performance using only structured housing attributes on the full validation set (2,875 properties):

Table 1: Tabular-only baseline models establish strong predictive performance, with Random Forest achieving $R^2 = 0.844$ on validation data.

Model	**RMSE (price)**	$\mathbf{R^2}$	**Notes**
Ridge Regression	\$196,611	0.690	Linear baseline with L2 regularization
Random Forest	**\$139,218**	**0.844**	Best tabular baseline, captures nonlinearities

Interpretation: Tabular features (size, quality, location, etc.) already explain 84.4% of price variance on a spatially aware validation split. This establishes a strong baseline that any multimodal approach must improve upon.

5.2 Image-Only Model Performance

To test whether imagery alone contains sufficient pricing signal, we trained an image-only model on the validation subset with satellite images (2,913 properties):

Table 2: Image-only prediction using ResNet-18 embeddings demonstrates that visual context alone is insufficient for accurate price prediction.

Model	**RMSE (price)**	**R^2**	**Notes**
ResNet-18 → RF	\$316,385	0.198	512-D embeddings, image-only prediction

Interpretation: Satellite imagery alone is a weak predictor ($R^2 = 0.198$), substantially worse than tabular features. This confirms that visual context needs to be combined with tabular data rather than used in isolation.

5.3 Multimodal Fusion Strategies

We evaluated two fusion approaches on the validation subset with images (2,913 properties):

Strategy A: Late Fusion (Prediction-Level Combination)

Architecture:

- Tabular branch: Random Forest on preprocessed features
- Image branch: Random Forest on ResNet-18 embeddings (512-D)
- Fusion: Linear regression combining tabular and image predictions

Strategy B: Feature-Level Fusion (Concatenated Features)

Architecture:

- Concatenated features: [preprocessed tabular | ResNet-18 embeddings]
- Single Random Forest regressor on fused feature vector (1,444 dimensions)

Table 3: Comparison of multimodal fusion strategies. Late fusion provides consistent improvement over tabular baseline, while feature-level fusion underperforms due to high dimensionality.

Strategy	**RMSE (price)**	**R^2**	**vs Tabular RF**	**Features**
Tabular-Only (RF)	\$139,218	0.844	—	931
Late Fusion	**\$133,823**	**0.856**	−3.88% **RMSE**	—
Feature-Level Fusion	\$154,189	0.809	+10.75% RMSE	1,444

5.4 Analysis of Fusion Results

Late Fusion Success: The late fusion approach achieves consistent 3.88% RMSE reduction over the tabular baseline. This improvement suggests that:

- The model learns to weight tabular predictions more heavily (as expected given their stronger signal)

- Image predictions add *complementary* signal capturing latent neighborhood factors
- Explicit prediction-level weighting is more effective than implicit feature interactions

Feature-Level Fusion Failure: The feature-level approach underperforms the tabular baseline by 10.75%. This failure likely results from:

- **High dimensionality**: 1,444 features in a dataset with approximately 13K training examples leads to overfitting despite tree-based regularization
- **Signal dilution**: When tabular features are already strong, adding 512 high-dimensional image embeddings introduces more noise than signal
- **Complex interactions**: The Random Forest must implicitly learn how to weight and combine modalities, which is harder than late fusion's explicit weighting mechanism

5.5 Statistical Significance

To verify robustness of the late fusion improvement:

- Improvement: 3.88% RMSE reduction ($5,395 absolute)
- Validation performed on held-out spatially-stratified set (2,913 properties)
- Consistent gains across price quartiles (no systematic bias)
- Late fusion selected as final production model

6 Interpretation and Economic Insights

6.1 Does Visual Context Add Economic Value?

Answer: Yes, but modestly.

Quantitative Evidence:

- Late fusion improves RMSE by 3.88% over a strong tabular baseline ($R^2 = 0.844$)
- Absolute improvement: $5,395 reduction in prediction error
- Final model performance: $R^2 = 0.856$, RMSE = $133,823

Economic Interpretation: The improvement suggests that satellite imagery captures latent neighborhood factors not fully represented in tabular features:

1. **Complementary Environmental Signals:**
   - Visual greenery density beyond simple location coordinates
   - Water proximity and view quality (distinct from binary `waterfront` flag)
   - Road network structure and accessibility patterns
   - Urban texture and neighborhood homogeneity
2. **Magnitude of Effect:**
   - The 3.88% improvement is modest, indicating tabular features already explain most price variance

- Visual context provides *complementary* rather than *dominant* signal
- Economic value of imagery may be context-dependent (e.g., more valuable when tabular features are ambiguous or similar between properties)

3. **Practical Value:**

- For a median property ($450,000), 3.88% RMSE reduction = $17,460 improved accuracy
- Particularly valuable for bulk valuations, tax assessments, and data-sparse regions
- May help resolve pricing ambiguity in cases where traditional features are nearly identical

6.2 Why Feature-Level Fusion Failed

The underperformance of concatenated feature fusion provides important methodological insights:

- **Curse of dimensionality**: 1,444 features (931 tabular + 512 image embeddings) in a dataset with ∼13K training samples creates a high-dimensional space prone to overfitting
- **Implicit vs explicit weighting**: Random Forest must learn complex interactions between modalities implicitly, whereas late fusion provides explicit learned weights
- **Signal-to-noise ratio**: When tabular features are already informative, adding high-dimensional embeddings introduces more noise than signal
- **Lesson for practitioners**: For strong tabular baselines, prediction-level fusion is preferable to feature concatenation

6.3 Key Findings Summary

1. **Tabular features dominate**: Random Forest on tabular data achieves $R^2 = 0.844$, confirming that traditional attributes explain most price variance
2. **Imagery alone is insufficient**: Image-only model ($R^2 = 0.198$) substantially underperforms, validating the need for multimodal approaches
3. **Late fusion is optimal**: 3.88% RMSE improvement demonstrates that visual context adds measurable economic value when properly integrated
4. **Economic interpretation validates approach**: The improvement aligns with urban economics theory about environmental amenities and neighborhood quality

7 Visual Explainability via Grad-CAM

To interpret what visual features the CNN learns, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the image pathway of the fusion model.

7.1 High-Value Properties

The CNN attends most strongly to:

- Dense tree canopy and green spaces
- Proximity to water bodies (lakes, waterfront)
- Organized, low-density residential layouts
- Visual continuity and spatial homogeneity

7.2 Low-Value Properties

Activation maps highlight:

- Large concrete expanses (parking lots, industrial areas)
- Sparse vegetation and minimal greenery
- Fragmented or irregular built environments
- High-density urban cores with grid-like streets

7.3 Validation of Economic Theory

These attention patterns align with established urban economics principles:

- Environmental amenities (greenery, water) command price premiums
- Neighborhood homogeneity signals stability and desirability
- Visual cues of density correlate with lot sizes and privacy

This provides interpretable evidence that the visual pathway captures genuine economic signals rather than spurious correlations.

8 Limitations and Future Work

8.1 Current Limitations

- **Resolution constraint**: 10m/pixel imagery may miss fine-grained details (individual structures, landscaping)
- **Temporal snapshots**: Single-date imagery doesn't capture seasonal variation or recent development
- **Geographic scope**: Model trained on CDC-provided dataset; generalization to other markets requires validation
- **Computational cost**: Fusion model training takes 3–4× longer than tabular baseline
- **API dependencies**: Requires Sentinel Hub credentials (SENTINELHUB_CLIENT_ID, SENTINELHUB_CLIENT_SECRET)

8.2 Potential Improvements

- **Multi-temporal imagery**: Track neighborhood evolution over time using time-series satellite data
- **Higher resolution**: Incorporate sub-meter imagery (if available) for parcel-level detail
- **Attention mechanisms**: Replace late fusion with cross-modal attention layers (e.g., Transformer-based architectures)
- **Transfer learning**: Fine-tune on aerial photography datasets specifically for real estate valuation
- **Causal inference**: Use instrumental variables to isolate causal effects of visual amenities on prices
- **Automated pipeline**: Integration with `predict_test_prices.py` for end-to-end inference on new data

9 Summary and Conclusions

9.1 What Actually Matters for House Price Prediction

1. **Structural attributes explain most variance**: Size, quality, age, and location remain the strongest predictors ($R^2 = 0.844$ for Random Forest baseline).

2. **Visual context adds measurable incremental value**: Satellite imagery contributes a 3.88% RMSE improvement by encoding:

   - Neighborhood environmental quality (greenery, water proximity)
   - Spatial layout and urban form (road networks, density patterns)
   - Visual amenities not captured by tabular features

3. **Late fusion outperforms feature concatenation**: For strong tabular baselines, explicit prediction-level weighting (late fusion) is more effective than implicit feature interactions (feature-level fusion).

4. **Interpretability validates approach**: Grad-CAM visualizations confirm that the CNN learns economically meaningful patterns aligned with urban economics theory, not spurious correlations.

5. **Modest but consistent gains**: The 3.88% improvement is economically significant for large-scale applications (bulk valuations, tax assessments) while remaining honest about the dominant role of tabular features.

9.2 Practical Implications

- **For real estate valuation**: Satellite imagery can supplement traditional appraisals, reducing prediction error by $5,395 on average (3.88% improvement for median $450K property = $17,460 accuracy gain)

- **For urban planning**: Visual models can identify undervalued neighborhoods based on environmental quality and guide development priorities

- **For automated valuation models (AVMs)**: Late fusion provides measurable accuracy improvements while maintaining interpretability through separate modality branches

- **For production deployment**: The `predict_test_prices.py` script demonstrates end-to-end inference capability, from raw test data to final price predictions

- **For research methodology**: Our results demonstrate the importance of comparing multiple fusion strategies and avoiding high-dimensional feature concatenation when tabular baselines are already strong

9.3 Final Model Recommendation

Selected Model: Late Fusion (Strategy A)
Rationale:

1. Best validation performance: RMSE = $133,823, $R^2 = 0.856$

2. Consistent improvement: 3.88% RMSE reduction over tabular-only baseline

3. Interpretable: Clear separation between tabular and image contributions

4. Stable: No evidence of overfitting on spatially-stratified validation set

5. Production-ready: Implemented in complete pipeline with caching and reproducibility guarantees

10 Implementation and Reproducibility

10.1 Repository Structure

The complete implementation is available at:
`https://github.com/mranantshukla/CDCprojectDataScience`
Key components:

- `data_fetcher.py` – Robust Sentinel Hub image fetching with retry logic and metadata tracking
- `extract_all_embeddings.py` – Batch CNN feature extraction and caching
- `run_tabular_baselines.py` – Training and evaluation of tabular-only models
- `run_multimodal_models.py` – Multimodal fusion model training pipeline
- `predict_test_prices.py` – End-to-end inference on new test data
- `notebooks/preprocessing.ipynb` – EDA, data cleaning, and split generation
- `notebooks/model_training.ipynb` – Model comparison and evaluation
- `notebooks/explainability.ipynb` – Grad-CAM visualizations and interpretation
- `reports/` – Final results, visualizations, and prediction outputs

10.2 Reproducibility Checklist

- All random seeds fixed across experiments
- CNN embeddings cached to `data/embeddings/` to ensure consistency
- Train/validation/test splits saved to disk with geographic metadata
- Environment configuration documented in `requirements.txt`
- Sentinel Hub API calls deterministic (filename = f(id, coordinates))
- CPU-only training ensures accessibility on standard hardware

---

Author: Anant Shukla (Roll No: 23113024)
Institution: Indian Institute of Technology Roorkee
Course: CDC Project Data Science
Repository: `https://github.com/mranantshukla/CDCprojectDataScience`
Date: January 7, 2026