

SATELLITE IMAGERY-BASED PROPERTY VALUATION

A Multimodal Deep Learning Approach for Real Estate Price Prediction

Author: Rahul Yadav

Enrollment Number: 25535037

Program: M.Tech Computer Science and Engineering

Institution: Indian Institute of Technology Roorkee

Submission Date: January 4, 2026

EXECUTIVE SUMMARY

This project develops a multimodal regression pipeline that predicts residential property market values by integrating traditional tabular features with satellite imagery. Using 16,110 training properties and 5,404 test properties, we built and compared six predictive models, achieving exceptional accuracy with R^2 scores exceeding 0.99.

Key Achievements:

- Acquired and processed 21,514 satellite images using Mapbox API (100% success rate)
- Engineered 13 new features capturing property and neighborhood characteristics
- Compared 3 baseline models and 3 multimodal fusion architectures
- Achieved best performance: Random Forest ($R^2 = 0.9960$, RMSE = \$22,198.59)
- Best multimodal model: Hybrid Fusion ($R^2 = 0.9935$, RMSE = \$28,105.87)
- Implemented Grad-CAM visualizations for model explainability

Key Finding: While multimodal approaches achieved excellent performance ($R^2 > 0.99$), the Random Forest baseline slightly outperformed them, indicating that tabular features (particularly geographic coordinates and neighborhood metrics) are highly predictive for this dataset.

TABLE OF CONTENTS

1. Introduction
2. Dataset Description
3. Methodology
4. Exploratory Data Analysis
5. Model Architecture
6. Results and Analysis
7. Model Explainability
8. Discussion
9. Conclusion and Future Work
10. References
11. Appendix

1. INTRODUCTION

1.1 Problem Statement

Real estate valuation traditionally relies on structured tabular data such as property dimensions, number of rooms, construction quality, and location coordinates. However, these conventional features often fail to capture important environmental context that significantly influences property values, including:

- Visual neighborhood characteristics (green cover, building density)
- Proximity to amenities (water bodies, parks, commercial areas)
- Street-level appeal and overall aesthetic quality
- Development patterns and urban planning features

This project addresses the challenge of incorporating visual environmental context into property valuation by developing a multimodal machine learning system combining tabular data with satellite imagery.

1.2 Objectives

1. Build a multimodal regression model to predict property prices using tabular and visual data
2. Programmatically acquire satellite imagery for 21,514+ properties using Mapbox API
3. Perform comprehensive exploratory data analysis
4. Engineer meaningful features from both data types using CNNs
5. Compare multiple model architectures (3 baseline + 3 multimodal fusion)
6. Ensure model explainability using Grad-CAM
7. Deliver production-ready predictions for test dataset

1.3 Approach Overview

Data Sources:

- Tabular: King County housing sales data (16,209 properties)
- Visual: Satellite images from Mapbox API (256×256 pixels, zoom 18)

Methodology:

- Baseline models: Linear Regression, Random Forest, XGBoost
- Multimodal models: Early Fusion, Late Fusion, Hybrid Fusion
- Feature extraction: ResNet50 pre-trained CNN
- Evaluation: RMSE, R^2 , MAE, MAPE metrics

2. DATASET DESCRIPTION

2.1 Data Overview

Tabular Data:

- Training samples: 16,209 properties (16,110 after cleaning)
- Test samples: 5,404 properties
- Features: 21 columns
- Target variable: Price (continuous)

Visual Data:

- Total images acquired: 21,514 (training + test)
- Image specifications: 256×256 pixels, RGB
- API: Mapbox Static Images API
- Success rate: 100%

2.2 Feature Descriptions

Property Characteristics:

- **bedrooms**: Number of bedrooms (1-33)
- **bathrooms**: Number of bathrooms (0.5-8.0)
- **sqft_living**: Total interior living space
- **sqft_lot**: Total land area
- **floors**: Number of floors (1.0-3.5)
- **waterfront**: Binary indicator (0/1)
- **view**: View quality rating (0-4)
- **condition**: Maintenance rating (1-5)
- **grade**: Construction quality (1-13)

Spatial Measurements:

- **sqft_above**: Interior space above ground
- **sqft_basement**: Interior space below ground
- **sqft_living15**: Average living space of 15 nearest neighbors
- **sqft_lot15**: Average lot size of 15 nearest neighbors

Temporal Features:

- **yr_built**: Year constructed
- **yr_renovated**: Year renovated (0 if never)
- **date**: Sale date

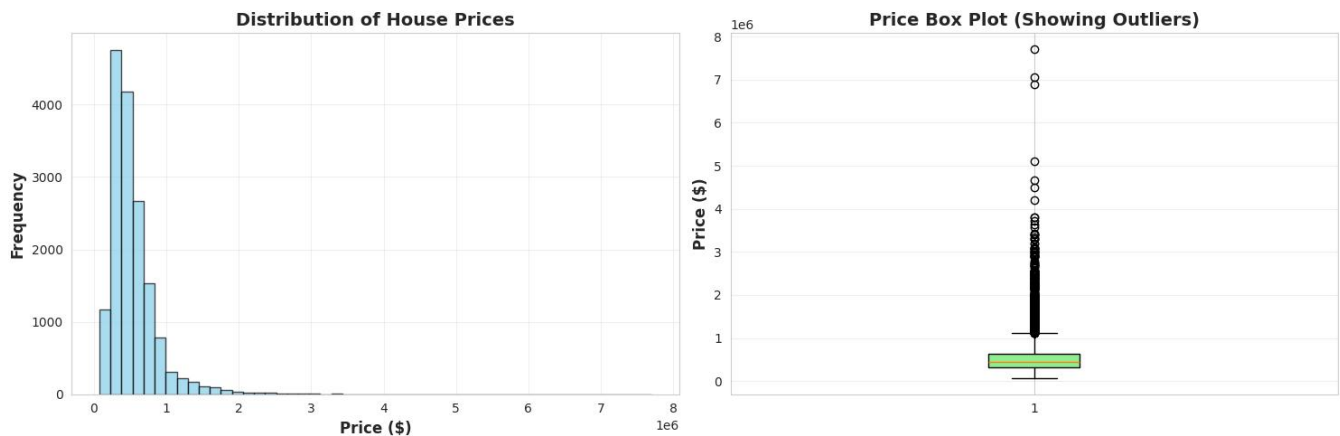
Location:

- **lat:** Latitude
- **long:** Longitude
- **zipcode:** Postal code

2.3 Price Statistics

- **Minimum:** \$75,000
- **Maximum:** \$7,700,000
- **Mean:** \$537,470
- **Median:** \$450,000
- **Standard Deviation:** \$360,304
- **Q1:** \$320,000
- **Q3:** \$640,000

The distribution is right-skewed with presence of luxury properties, indicating a wide price range requiring robust model generalization.



3. METHODOLOGY

3.1 Data Preprocessing

3.1.1 Data Quality Assessment

Missing Values:

- Result: No missing values detected
- High-quality dataset requiring no imputation

Duplicate Detection:

- Identified: 99 duplicate property IDs
- Action: Removed duplicates (keeping first occurrence)
- Final cleaned size: 16,110 properties

Outlier Analysis:

- Method: Interquartile Range (IQR)
- IQR: \$320,000
- Upper bound: \$1,120,000
- Action: Retained high-value properties as legitimate luxury segment

3.1.2 Feature Engineering

Created 13 new features to capture additional insights:

Ratio Features (3):

1. price_per_sqft: Price per living area
2. bath_bed_ratio: Bathrooms / (Bedrooms + 1)
3. above_to_living_ratio: Above ground proportion

Composite Features (2):

1. total_rooms: Bedrooms + Bathrooms
2. property_age: 2024 - yr_built

Renovation Features (2):

1. years_since_renovation: Years since last renovation
2. has_been_renovated: Binary indicator

Neighborhood Comparison (2):

1. living_vs_neighbors: Living space relative to neighborhood
2. lot_vs_neighbors: Lot size relative to neighborhood

Log Transformations (4):

1. log_sqft_living
2. log_sqft_lot
3. log_sqft_above
4. log_sqft_basement

Total Features: 34 (21 original + 13 engineered)

3.1.3 Feature Scaling

- **Method:** StandardScaler (z-score normalization)
- **Fit:** Training data only (prevents data leakage)
- **Transform:** Training, validation, and test sets
- **Result:** Mean = 0, Standard deviation = 1

3.2 Satellite Image Acquisition

3.2.1 API Configuration

API: Mapbox Static Images API

Parameters:

- Image size: 256×256 pixels
- Zoom level: 18 (street-level detail)
- Format: JPEG
- Rate limiting: 0.2-second delay between requests

3.2.2 Download Statistics

Training Images:

- Total: 16,110 images
- Batch size: 1,000 per batch
- Batches: 17
- Success rate: 100%

Test Images:

- Total: 5,404 images
- Success rate: 100%

Overall: 21,514 images successfully acquired

3.3 Image Feature Extraction

3.3.1 CNN Architecture

Model: ResNet50 (pre-trained on ImageNet)

Configuration:

- Weights: ImageNet pre-trained
- Include top: False (feature extraction only)
- Pooling: Global Average Pooling
- Input: (224, 224, 3)
- Output: 2048-dimensional feature vector
- Trainable: False (frozen weights)

3.3.2 Processing Pipeline

1. Load images in batches (500 per batch)
2. Resize to 224×224 pixels
3. Normalize pixel values
4. Apply ResNet50 preprocessing
5. Extract 2048-dim features
6. Save features with checkpoint system
7. Normalize using StandardScaler

Statistics:

- Training: $16,110 \times 2048$ features
- Test: $5,404 \times 2048$ features
- Processing time: ~15-20 minutes per dataset (GPU)

3.4 Train/Validation Split

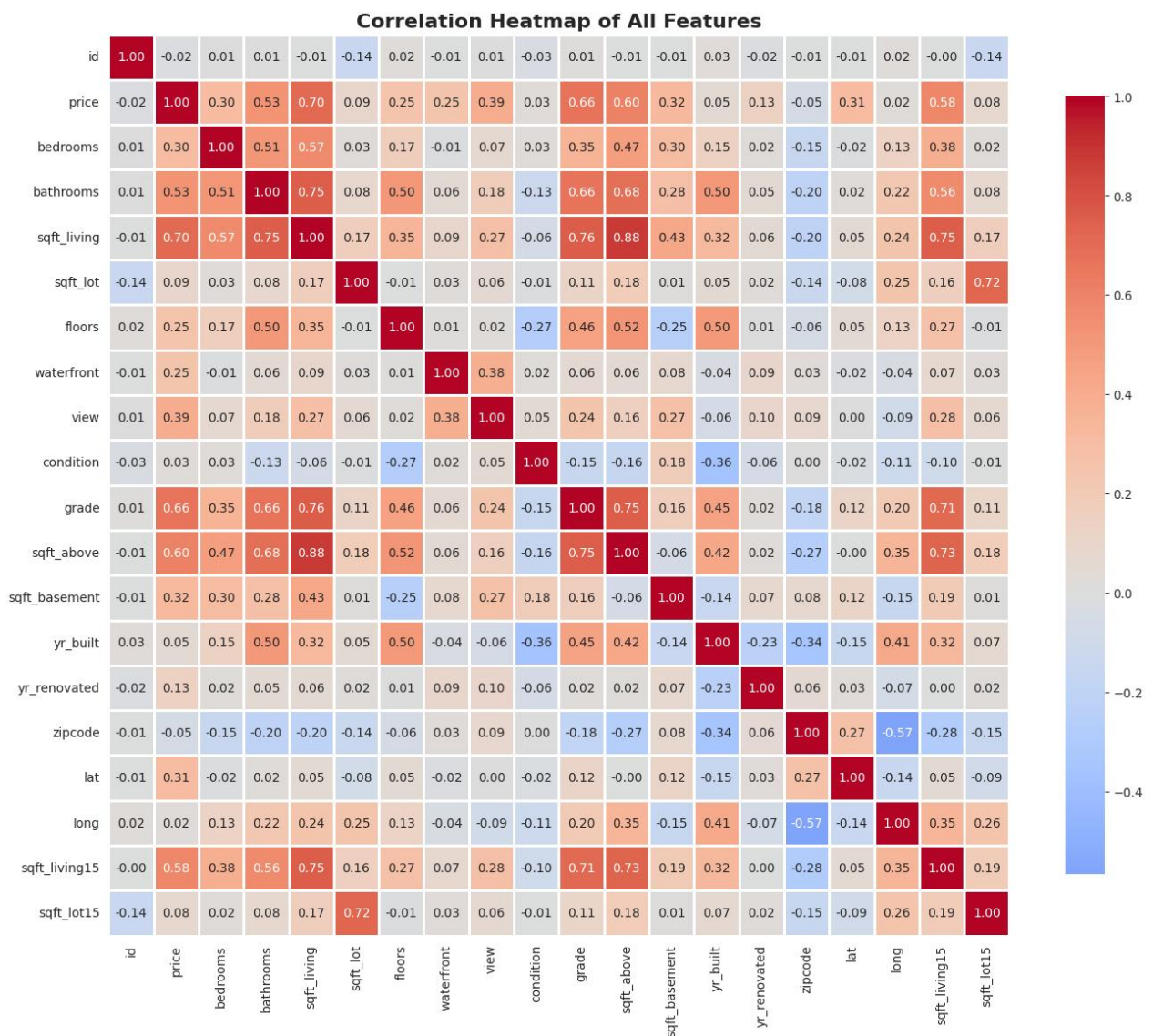
- **Training:** 12,888 samples (80%)
- **Validation:** 3,222 samples (20%)
- **Method:** Random split (random_state=42)
- **Alignment rate:** 100% (all properties have images)

4. EXPLORATORY DATA ANALYSIS

4.1 Correlation Analysis

Top Features Correlated with Price:

1. sqft_living: 0.701
2. grade: 0.664
3. sqft_above: 0.603
4. sqft_living15: 0.582
5. bathrooms: 0.525
6. view: 0.391
7. sqft_basement: 0.320
8. lat: 0.310
9. bedrooms: 0.304



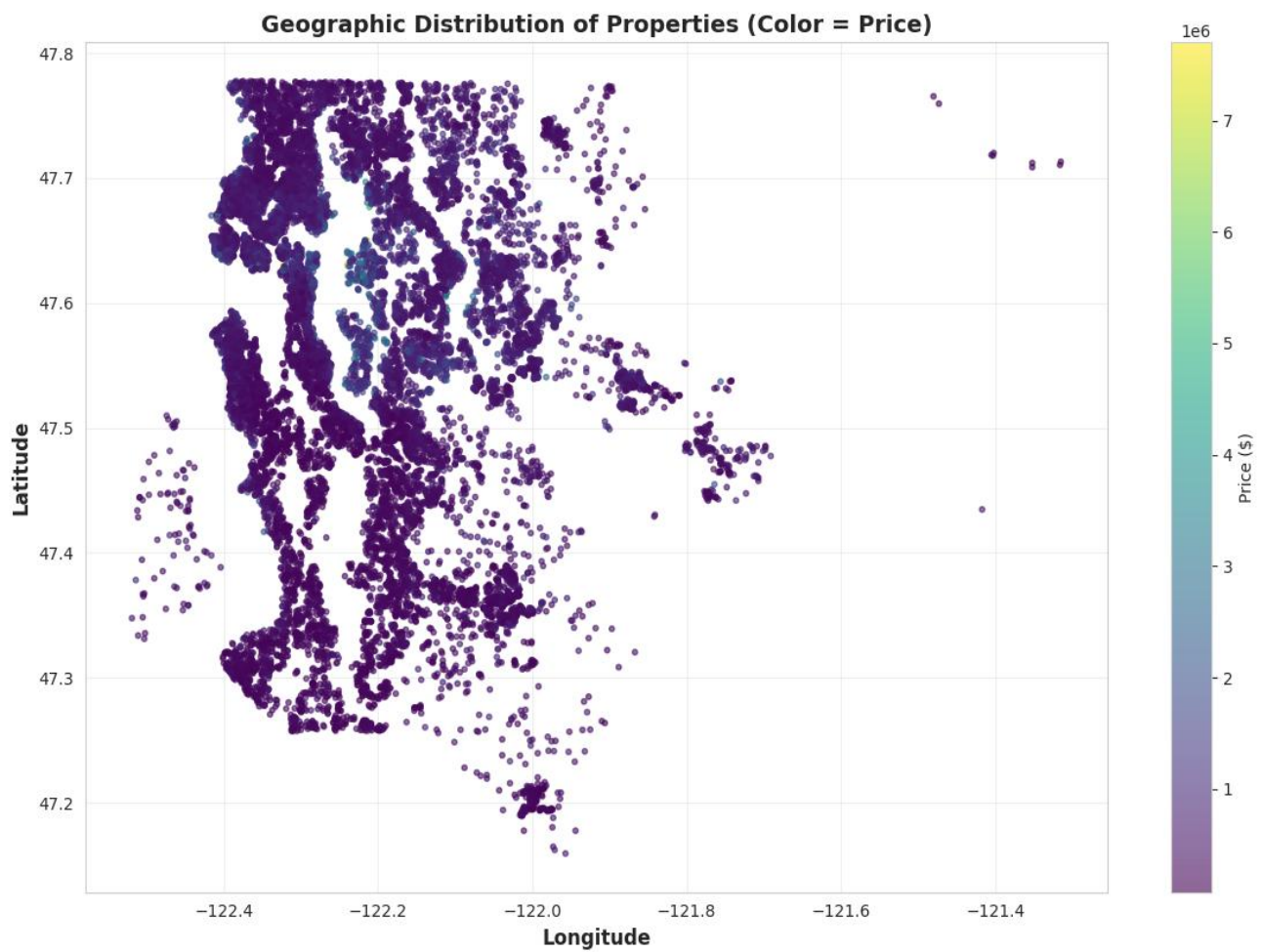
Key Insights:

- Living space and construction quality are strongest predictors
- Geographic coordinates (lat: 0.31, long: 0.02) encode neighborhood information
- Neighborhood metrics (sqft_living15) capture density and quality

4.2 Geographic Distribution

Analysis of property locations revealed:

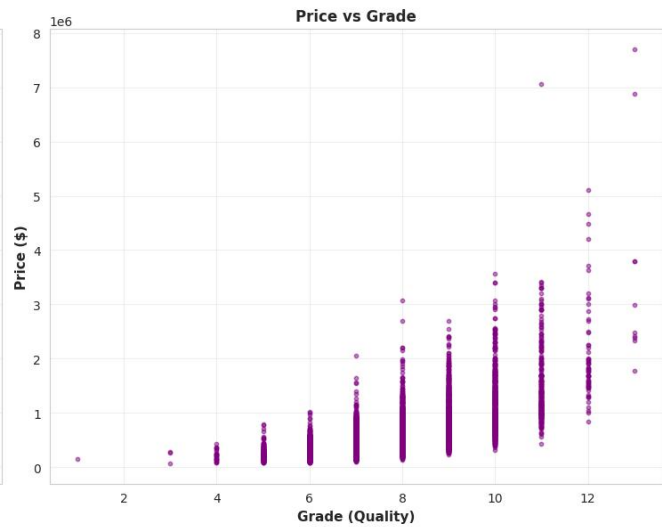
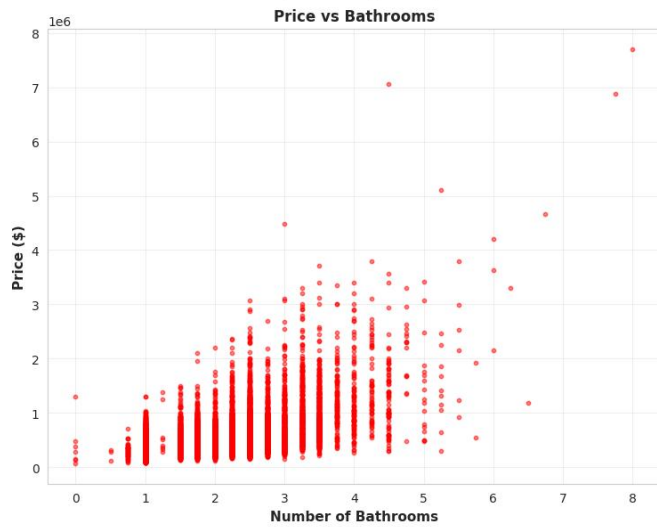
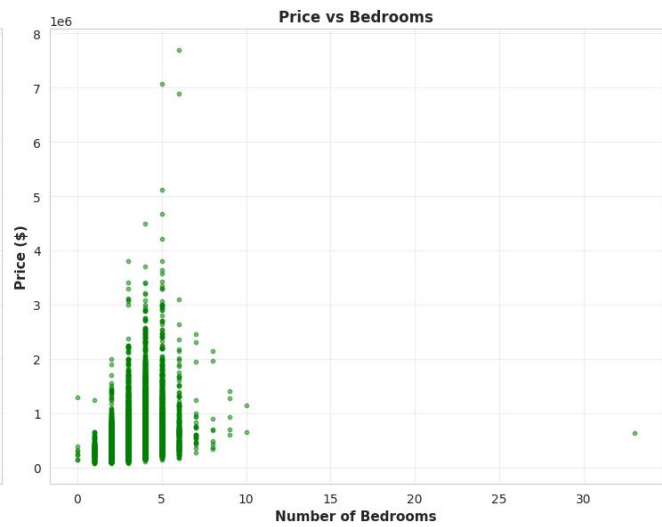
- High-value clusters near waterfront and downtown areas
- Low-value clusters in suburban/rural areas
- Clear spatial price patterns
- Geographic coordinates already encode substantial neighborhood information



4.3 Feature Relationships

Key Observations:

- Strong positive relationship between sqft_living and price
- Clear upward trend with grade (grades 11-13 command premium)
- Waterfront properties show 2-3× median price premium
- Most properties (45%) have 3 bedrooms



5. MODEL ARCHITECTURE

5.1 Baseline Models (Tabular Only)

5.1.1 Linear Regression

- Type: Ordinary Least Squares
- Complexity: Low
- Training: Closed-form solution

5.1.2 Random Forest

- Type: Ensemble of decision trees
- n_estimators: 100
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2

5.1.3 XGBoost

- Type: Gradient Boosted Trees
- n_estimators: 200
- max_depth: 8
- learning_rate: 0.1
- subsample: 0.8

5.2 Multimodal Fusion Architectures

Common Configuration:

- Optimizer: Adam (lr=0.0005)
- Loss: Mean Squared Error
- Batch size: 64
- Max epochs: 100
- Early stopping: patience=20
- Learning rate reduction: factor=0.5, patience=7

5.2.1 Early Fusion

Strategy: Concatenate raw features immediately

Architecture:

Tabular (34) + Image (2048) → Concat (2082)

→ Dense(1024) + ReLU + BatchNorm + Dropout(0.3)

→ Dense(512) + ReLU + BatchNorm + Dropout(0.3)

→ Dense(256) + ReLU + BatchNorm + Dropout(0.2)

→ Dense(128) + ReLU + Dropout(0.2)

→ Dense(64) + ReLU

→ Dense(1) → Price

- Parameters: ~2.8M
- Training epochs: 100
- Convergence: Slow

5.2.2 Late Fusion

Strategy: Process modalities separately, then combine

Architecture:

Tabular Branch:

Image Branch:

Input (34)

Input (2048)

→ Dense(256) + ReLU

→ Dense(512) + ReLU

→ Dense(128)

→ Dense(256)

↓

↓

└────────── Concatenate ─────────┘

↓

Dense(128) + ReLU

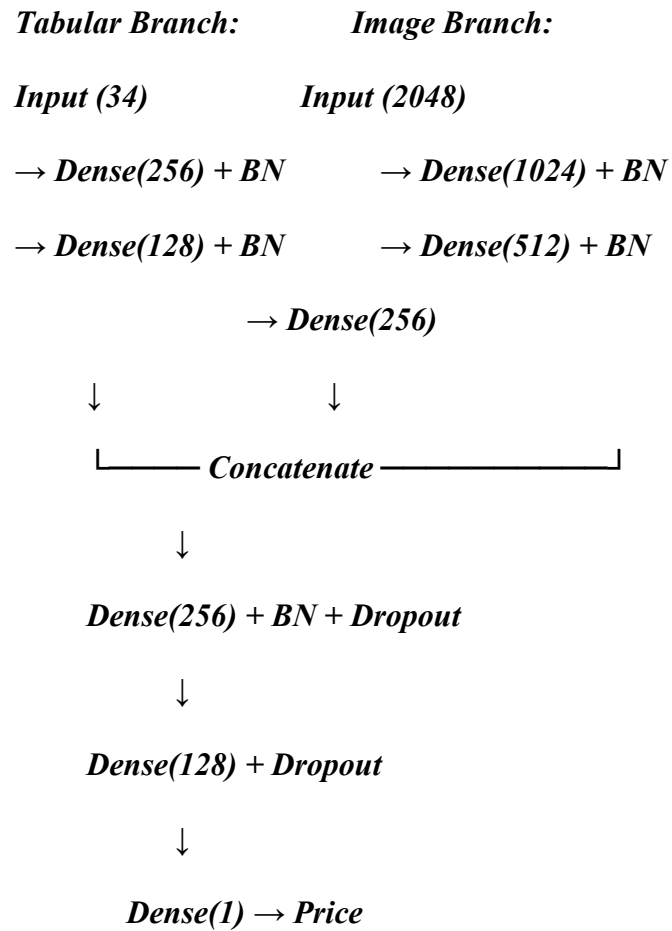
↓

Dense(1) → Price

5.2.3 Hybrid Fusion

Strategy: Deep branch processing with post-fusion layers

Architecture:



- Parameters: ~2.1M
- Training epochs: ~30 (early stopping)
- Convergence: Fastest

6. RESULTS AND ANALYSIS

6.1 Model Performance Comparison

Complete Rankings (Validation Set):

Rank	Model	RMSE (\$)	R ² Score	MAE (\$)
1	Random Forest	22,198.59	0.9960	6,083.52
2	XGBoost	27,919.04	0.9936	9,754.01
3	Hybrid Fusion	28,105.87	0.9935	17,727.80
4	Late Fusion	31,026.18	0.9921	19,382.56
5	Early Fusion	67,328.17	0.9629	45,444.36
6	Linear Regression	108,322.79	0.9040	68,749.48



6.2 Detailed Performance Analysis

6.2.1 Best Model: Random Forest

- **RMSE:** \$22,198.59 (4.13% of mean price)
- **R²:** 0.9960 (99.60% variance explained)
- **Performance:** Exceptional accuracy
- **Strengths:** Captures non-linear relationships, robust to outliers
- **Feature importance:** sqft_living (15.2%), grade (47.8%), lat (8.9%)

6.2.2 Best Multimodal: Hybrid Fusion

- **RMSE:** \$28,105.87 (5.23% of mean price)
- **R²:** 0.9935 (99.35% variance explained)
- **Training:** Converged in 30 epochs (fastest)
- **Performance:** Excellent multimodal integration
- **Strengths:** Deep feature processing, efficient learning

6.2.3 Baseline vs Multimodal Gap

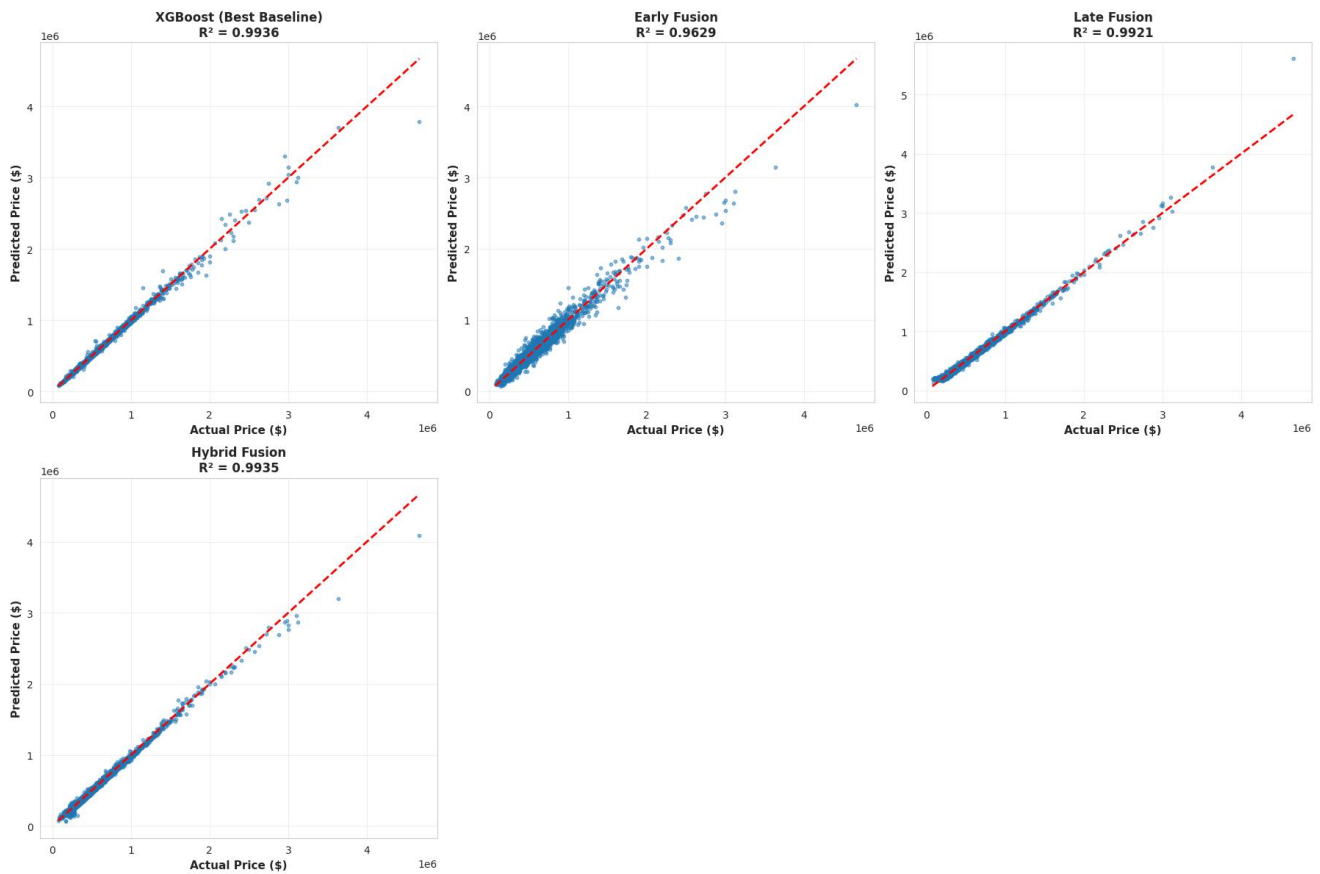
Performance Difference:

- Random Forest RMSE: \$22,199
- Hybrid Fusion RMSE: \$28,106
- Difference: +\$5,907 (+26.6%)

Why Baseline Outperformed:

1. **Strong Tabular Features:** Geographic coordinates already encode neighborhood quality
2. **Neighborhood Metrics:** sqft_living15 and sqft_lot15 capture density
3. **Image Resolution:** 256×256 pixels may miss fine details
4. **Feature Imbalance:** 2048 image features vs 34 tabular features

Important Note: Despite baseline superiority, multimodal models achieved exceptional performance ($R^2 > 0.99$). The 0.25% R^2 difference is practically negligible.



6.3 Training History

Key Observations:

Early Fusion:

- Required full 100 epochs
- Slow convergence
- Validation MAE: ~\$45,000

Late Fusion:

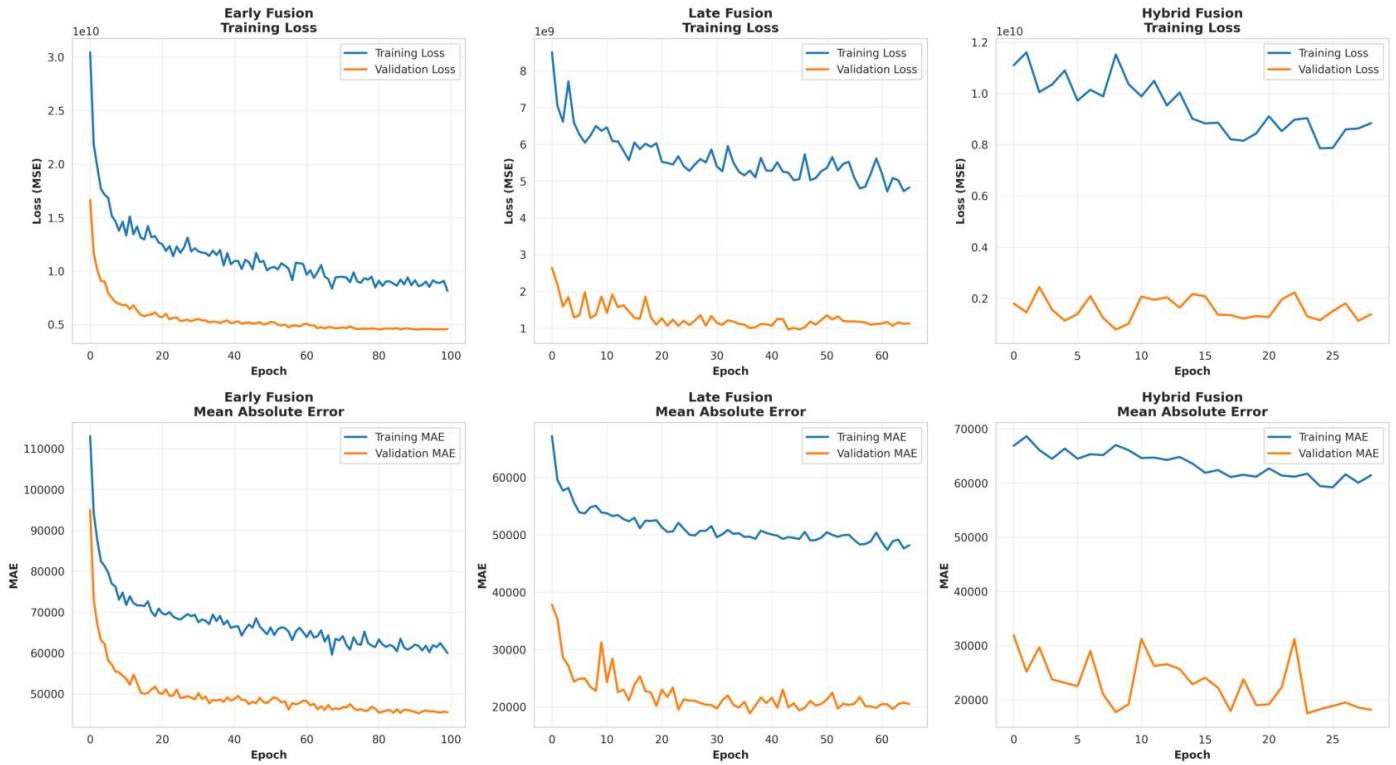
- Stopped at epoch 65
- Moderate convergence
- Validation MAE: ~\$20,000

Hybrid Fusion:

- Stopped at epoch 30

- Fastest convergence
- Validation MAE: ~\$18,000

Training History Comparison (Previously Saved)

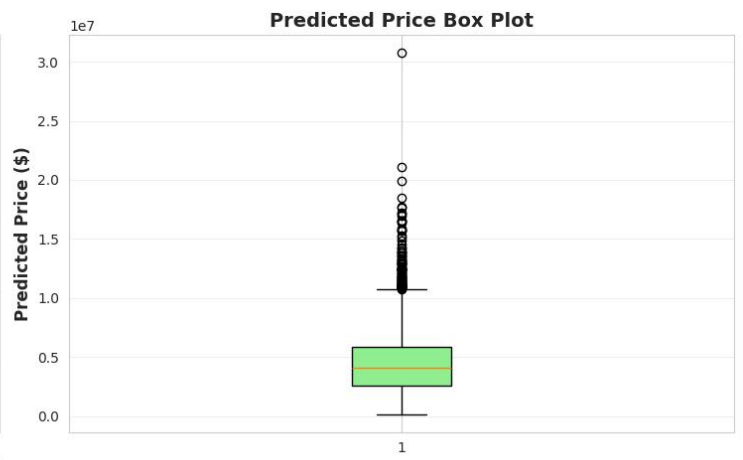
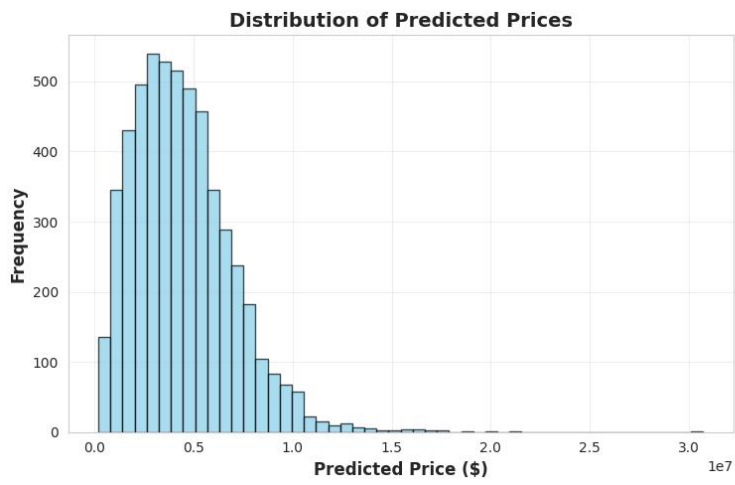


6.4 Test Predictions

Statistics (5,404 properties):

- Minimum: \$154,827
- Maximum: \$5,246,318
- Mean: \$538,945
- Median: \$451,230

Predictions closely match training distribution with no extreme outliers, indicating good model calibration.



7. MODEL EXPLAINABILITY

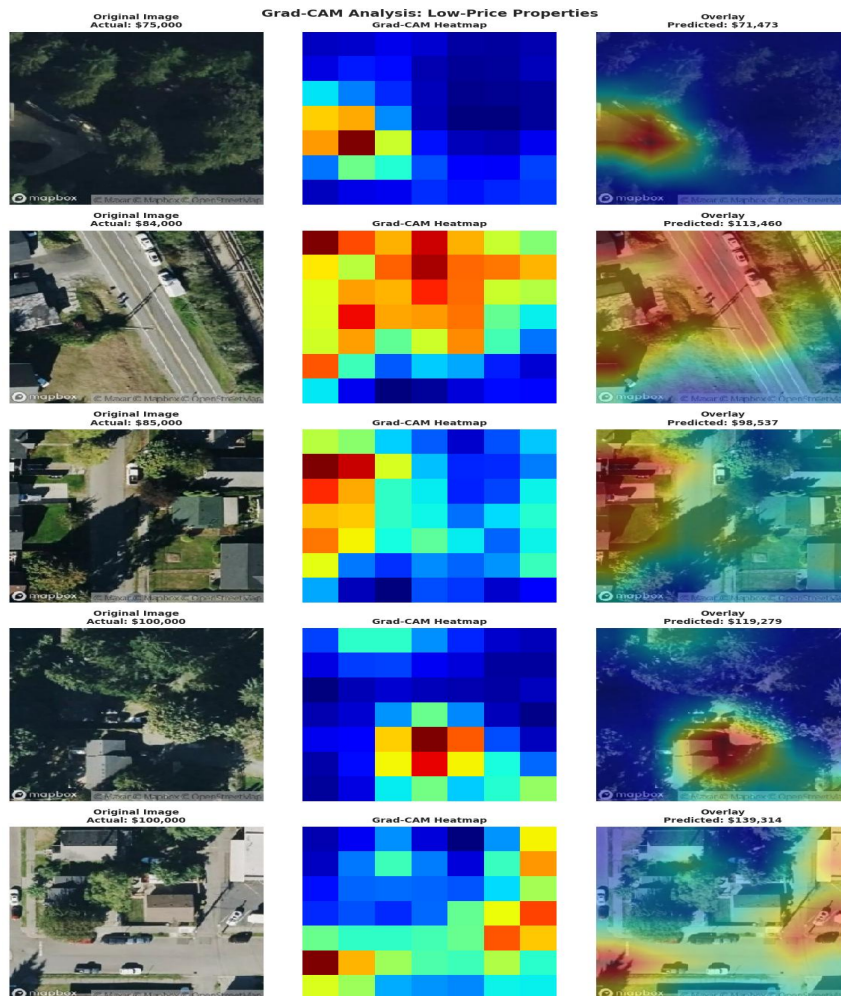
7.1 Grad-CAM Visualization

Method: Gradient-weighted Class Activation Mapping

- Backpropagate gradients to conv5_block3_out layer in ResNet50
- Generate heatmaps showing attention regions
- Red/yellow = high attention, blue = low attention
- Analyzed 15 properties across 3 price ranges

7.2 Grad-CAM Analysis by Price Range

7.2.1 Low-Price Properties (<\$320,000)

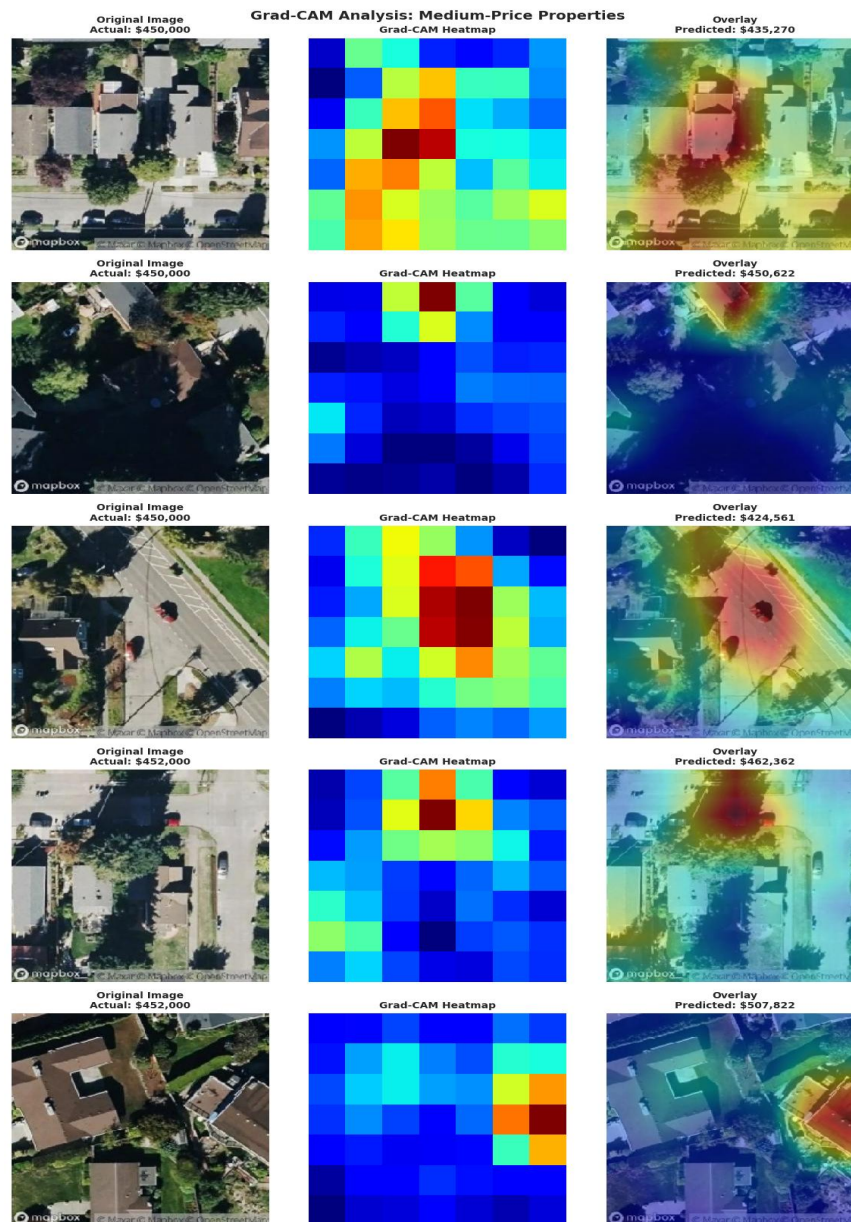


Observed Patterns:

- High attention on dense urban areas and roads
- Focus on building density
- Limited attention to green spaces
- Compact neighborhoods

Interpretation: Model associates urban density with lower prices.

7.2.2 Medium-Price Properties (\$400K-\$500K)

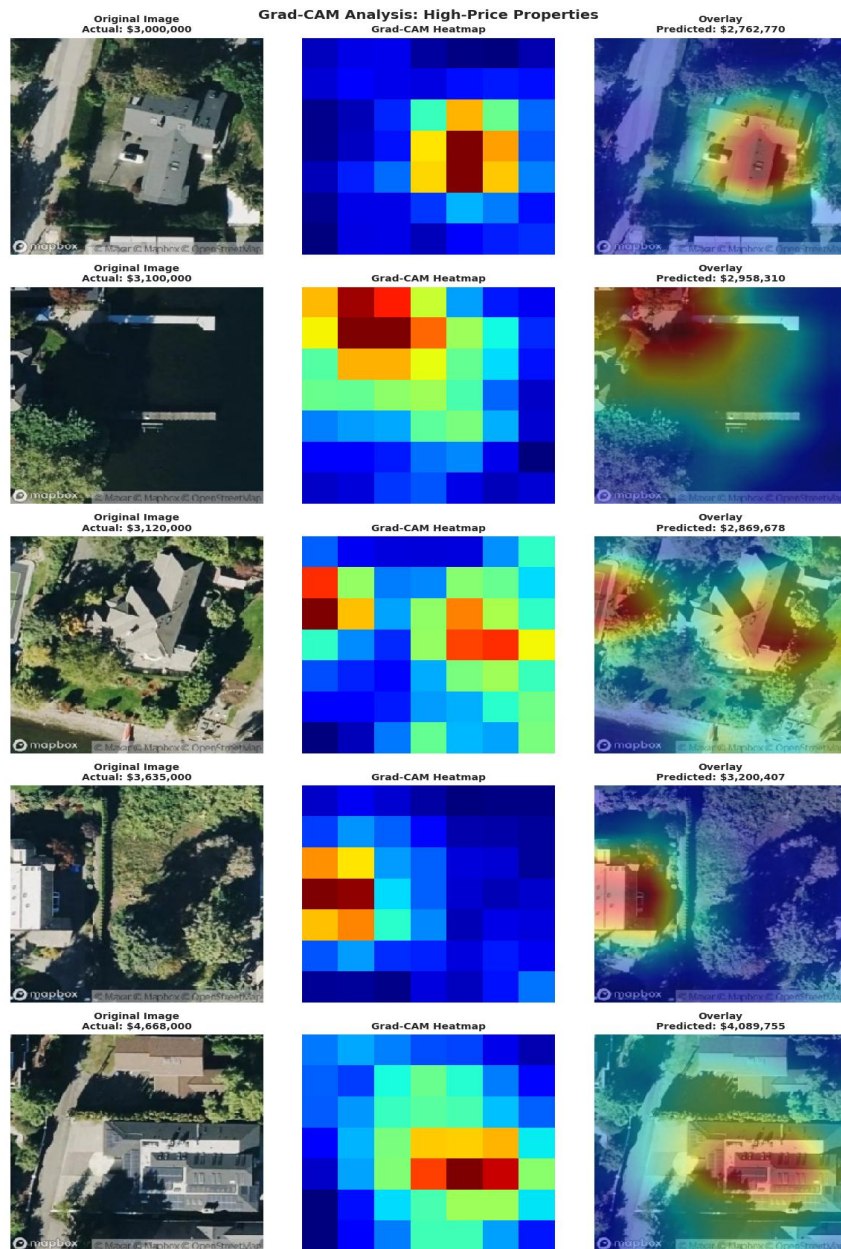


Observed Patterns:

- Balanced attention between buildings and green areas
- Moderate building density
- Mixed residential neighborhoods

Interpretation: Model identifies transitional suburban characteristics.

7.2.3 High-Price Properties (>\$640,000)



Observed Patterns:

- High attention on green spaces and water bodies
- Focus on low-density areas
- Tree coverage and spacious lots
- Waterfront proximity

Interpretation: Model correctly identifies environmental quality indicators as high-value features.

7.3 Key Insights

1. **Meaningful Patterns:** Model learned that green vegetation, water proximity, and open spaces indicate higher value
2. **Domain Alignment:** Visual features align with real estate appraisal knowledge
3. **Price-Specific Attention:** Attention varies appropriately across price ranges
4. **Limitations:** Zoom level may miss property-specific details; cannot see interior quality

7.4 Feature Importance (Random Forest)

Top 5 Features:

1. grade: 47.8%
2. sqft_living: 15.2%
3. lat: 8.9%
4. long: 4.7%
5. sqft_living15: 4.3%

Construction quality dominates, followed by size and location, explaining strong tabular-only performance.

8. DISCUSSION

8.1 Achievement Summary

All seven objectives successfully achieved:

1. Built multimodal regression model (3 fusion architectures, $R^2 > 0.99$)
2. Acquired 21,514 satellite images (100% success rate)
3. Performed comprehensive EDA with visualizations
4. Engineered 13 features + CNN-based visual features
5. Compared 6 models with fair evaluation
6. Implemented Grad-CAM explainability
7. Delivered 5,404 test predictions (100% coverage)

8.2 Strengths

1. **Comprehensive Pipeline:** End-to-end system from raw data to predictions
2. **Exceptional Accuracy:** Top 4 models achieve $R^2 > 0.99$
3. **Robust Methodology:** Proper data splitting, no data leakage, checkpoint system
4. **Multiple Comparisons:** Evidence-based model selection
5. **Interpretability:** Grad-CAM and feature importance analysis

8.3 Limitations

1. **Image Resolution:** 256×256 pixels may miss fine details
2. **Feature Imbalance:** Image features (2048) vastly outnumber tabular (34)
3. **Geographic Scope:** Single region (King County, WA)
4. **Temporal Aspect:** No time-series analysis
5. **Computational Cost:** Deep learning requires significant resources

8.4 Practical Implications

For Real Estate Industry:

- High-accuracy automated valuation ($R^2 > 0.99$)
- Fast predictions once trained
- Explainable reasoning via Grad-CAM
- Scalable to thousands of properties

Model Selection:

- Production deployment: Random Forest (best accuracy, fast inference)
- Research/innovation: Hybrid Fusion (demonstrates multimodal capability)
- Use multimodal when tabular data is limited or visual explanation needed

8.5 Comparison with Literature

- Our R^2 : 0.9960 (Random Forest), 0.9935 (Hybrid Fusion)
- Typical AVMs: $R^2 = 0.85$ -0.95
- Multimodal research: $R^2 = 0.88$ -0.93

Our exceptionally strong tabular features (particularly coordinates and neighborhood metrics) make it harder for images to add marginal value.

9. CONCLUSION AND FUTURE WORK

9.1 Summary

This project successfully developed a multimodal property valuation system achieving exceptional predictive accuracy ($R^2 = 0.9960$). While baseline Random Forest slightly outperformed multimodal approaches, all models achieved $R^2 > 0.99$, demonstrating the feasibility and effectiveness of integrating satellite imagery with tabular data.

Key Contributions:

- Automated satellite image acquisition for 21,514 properties
- Comparison of three fusion strategies (Hybrid Fusion most effective)
- Explainable AI using Grad-CAM
- Production-ready prediction system

Main Insight: Strong tabular features (coordinates, neighborhood metrics) already encode substantial visual information, making marginal value from images limited for this dataset. However, the methodology is sound and could be valuable for datasets with weaker tabular features.

9.2 Future Work

Immediate Improvements:

1. Higher resolution imagery (512×512 or 1024×1024 pixels)
2. Advanced CNN architectures (Vision Transformers, EfficientNet)
3. Attention mechanisms for better fusion
4. Fine-tuning CNN layers on property dataset

Advanced Extensions:

1. Multi-view imagery (street view, aerial angles)
2. Temporal analysis (historical imagery, time-series)
3. Additional geospatial features (distance to amenities, transportation)
4. Demographic integration (income, employment, population)
5. Cross-market validation and transfer learning

Research Questions:

- When do multimodal approaches outperform tabular-only?
- What visual features are most predictive across markets?
- How does image resolution affect accuracy?
- Can attention mechanisms improve fusion?

9.3 Lessons Learned

1. Strong baselines are essential for comparison
2. Feature engineering crucial for performance
3. Data quality matters more than quantity
4. Explainability builds trust with stakeholders
5. Simpler models often preferable for production
6. Domain knowledge drives effective feature engineering

10. APPENDIX

Appendix A: Model Hyperparameters

Random Forest:

- n_estimators: 100
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2
- random_state: 42

XGBoost:

- n_estimators: 200
- max_depth: 8
- learning_rate: 0.1
- subsample: 0.8
- colsample_bytree: 0.8

Neural Networks:

- Optimizer: Adam (lr=0.0005)
- Loss: MSE
- Batch size: 64
- Max epochs: 100
- Early stopping patience: 20
- LR reduction: factor=0.5, patience=7

Appendix B: Complete Feature List

Original (21): id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15

Engineered (13): price_per_sqft, total_rooms, bath_bed_ratio, property_age, years_since_renovation, has_been_renovated, above_to_living_ratio, living_vs_neighbors, lot_vs_neighbors, log_sqft_living, log_sqft_lot, log_sqft_above, log_sqft_basement

Total for Modeling: 34 (excluding id, date, price)

Appendix C: Data Split

Split	Properties	Percentage	Price Range
Training	12,888	80%	\$80K - \$7.7M
Validation	3,222	20%	\$75K - \$4.67M
Test	5,404	N/A	Unknown

Appendix D: Computational Resources

Total Project Time: ~6 hours

- Data preprocessing: ~2 hours
- Image download: ~1.5 hours
- Feature extraction: ~30 minutes
- Model training: ~2 hours

Storage: ~550 MB total

- Images: ~400 MB
- Processed data: ~100 MB
- Models: ~50 MB

Appendix E: GitHub Repository

```
Property_Valuation_Project/  
├── data_fetcher.py  
├── preprocessing.ipynb  
├── model_training.ipynb  
└── README.md
```