

حسن خليفة

Chapter 3 Methodology-1.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:400516254

Submission Date

Jul 21, 2026, 11:30 AM GMT+3

Download Date

Jul 22, 2026, 3:00 PM GMT+3

File Name

Chapter 3 Methodology-1.docx

File Size

7.4 MB

17 Pages

3,802 Words

22,467 Characters

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **49 Not Cited or Quoted 12%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 7%  Publications
- 8%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Hidden Text**
22 suspect characters on 1 page
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 49 Not Cited or Quoted 12%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 7% Publications
- 8% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	Tilburg University on 2025-01-10	1%
2	Submitted works	Liverpool John Moores University on 2026-02-15	<1%
3	Publication	Amiri, Ahmed. "Rental Property Demand and Supply Analysis Using Machine Lear...	<1%
4	Publication	Bhabani Prasanna Pattanaik, Dharendra Kumar Shukla, Sambit Satpathy, Kamala...	<1%
5	Publication	J. Anitha Menon, Benita Veronica, Ben Milbourn, Mithra S. Gnana Sanga. "Sustain...	<1%
6	Publication	Poonam Nandal, Tapas Kumar, Meeta Singh, Mamta Dahiya. "Progressive Compu...	<1%
7	Submitted works	RV College of Engineering, Bengaluru on 2026-06-04	<1%
8	Internet	www.mdpi.com	<1%
9	Internet	www.numberanalytics.com	<1%
10	Publication	Iacopo Carnacina, Mawada Abdellatif, Manolia Andredaki, James Cooper, Darren ...	<1%

11	Submitted works	York St John University on 2026-05-31	<1%
12	Submitted works	University of Central Lancashire on 2026-07-12	<1%
13	Internet	adab.um.edu.my	<1%
14	Internet	0-www-mdpi-com.brum.beds.ac.uk	<1%
15	Submitted works	South Bank University on 2025-04-17	<1%
16	Internet	journals.orclever.com	<1%
17	Internet	publikationen.bibliothek.kit.edu	<1%
18	Publication	Jianheng Tong, Ying Zhang. "A Real-Time Label-Free Self-Supervised Deep Learnin...	<1%
19	Publication	Zeyao Hu, Long Chen, Yang Song, Zhigang Liu, João Pombo, Pedro Antunes. "A de...	<1%
20	Internet	ir.mu.ac.ke:8080	<1%
21	Internet	norma.ncirl.ie	<1%
22	Internet	www.ncbi.nlm.nih.gov	<1%
23	Publication	Golan Yona. "Introduction to Computational Proteomics", CRC Press, 2010	<1%
24	Submitted works	Multimedia University on 2026-06-24	<1%

25	Submitted works	University of Technology, Sydney on 2025-05-23	<1%
26	Publication	Xing Wang, Yangli Zhu, Wen Li, Zhitao Zuo, Haisheng Chen. "A Surrogate Model fo...	<1%
27	Internet	zg104.github.io	<1%
28	Publication	Ahmet Ali Süzen. "UNI-CAPTCHA: A Novel Robust and Dynamic User-Non-Interacti...	<1%
29	Submitted works	Colorado State University, Global Campus on 2024-09-28	<1%
30	Submitted works	University of Bristol on 2023-12-11	<1%
31	Publication	Haoxiang Tang, Fangyi Li, Xiaonan Hu, Feng Chen, Xin Chen, Xiang Li. "Deep Lear...	<1%

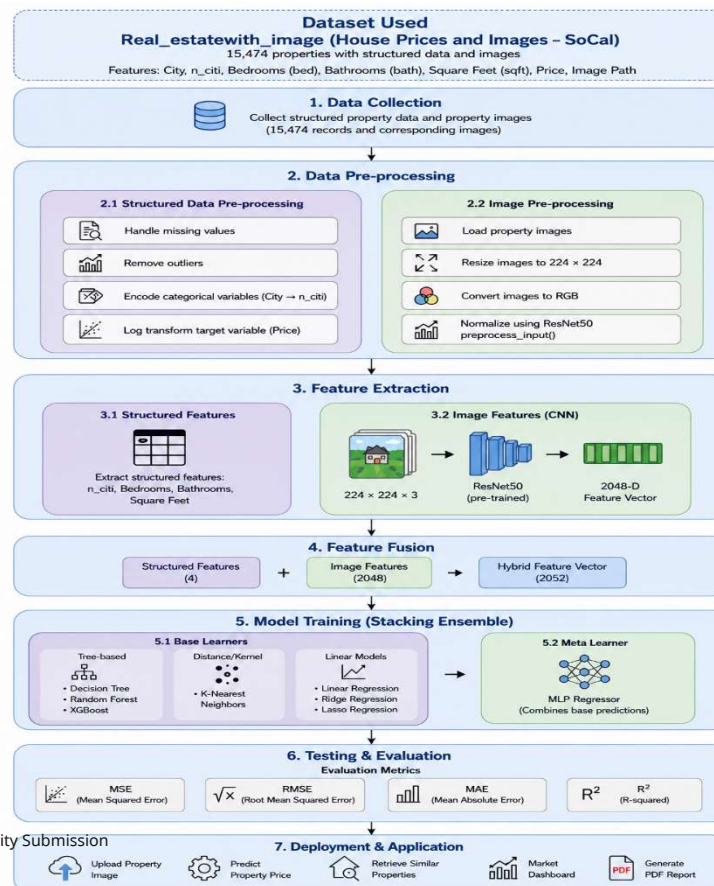
Chapter 3 — Research Methodology

3.1 Introduction

This chapter presents the research methodology adopted to develop the proposed Clarivision AI system for real estate valuation. The methodology describes the complete workflow followed throughout the study, beginning with data collection and preprocessing, followed by image feature extraction and structured data processing. The chapter then explains the development of the proposed hybrid artificial intelligence model, which combines deep image features extracted using ResNet50 was selected because it offers an excellent balance between feature extraction capability, computational efficiency, and transfer learning performance for image-based applications. with structured property attributes and employs the XGBoost regression algorithm for property price prediction. Finally, the evaluation methodology and performance metrics used to validate the proposed model are presented.

The proposed methodology aims to improve the accuracy and reliability of automated property valuation by integrating visual property characteristics with structured real estate information into a unified prediction framework. In addition, the developed system incorporates an intelligent recommendation module that retrieves visually similar properties within the same city, thereby providing practical support for real estate professionals and financial institutions during property valuation and decision-making.

Figure 3.1: General Methodology of the Proposed Clarivision AI Framework.



shows the entire flow of operations of the proposed hybrid real estate appraisal model, which comprises data preprocessing, image features extraction, hybrid feature fusion, price prediction, similarity-based real estate property search, and decision-making support.

3.2 Proposed Hybrid Framework.

- The proposed Clarivision AI framework was designed for enhancing the accuracy of predicting real estate prices using the combination of structured property attributes and visual information contained in property images. Differently from other existing Automated Valuation Models (AVM), the Clarivision AI framework considers not only structured property attributes but also deep image features that are extracted from property pictures.
- The presented framework includes several consecutive steps. Firstly, the dataset named Real_estatewith_image is obtained and processed. Structured property attributes are subject to preprocessing that comprises such steps as data cleansing, city encoding, feature selection, and logarithmical transformation of property prices. At the same time, all property images are resized, converted into RGB format, and then preprocessed. After that, the resized property images are fed into the pre-trained convolutional neural network ResNet50. ResNet50 network generates a 2048-dimensional deep feature vector that characterizes visual property features.
- Next step is the concatenation of deep image features and structured property attributes, that is city, number of bedrooms, number of bathrooms, and square footage. Obtained hybrid feature vector is used as an input to the XGBoost Regressor that estimates property market value.
- After the prediction of prices, the suggested framework extracts visually similar houses from the database. In this extraction phase, properties that belong to the same city are filtered out initially, followed by calculating cosine similarity of the deep image feature vectors. The five most visually similar properties are then displayed along with their property information to help users understand the valuation process.

Finally, the predicted value of the property, analysis of the model, market dashboard, and PDF report of valuation of the property are generated using Clarivision AI decision-making system.

3.3 DataSet Description

The suggested Clarivision AI framework is built based on the (Real_estatewith_image) dataset. This data set was chosen due to the presence of structured data and images related to properties, which makes it appropriate for building an artificial intelligence framework, which combines numerical and visual features for property valuation.

The dataset comprises 15,474 property records each containing a picture of a property and a set of structured features characterizing the property. Structured features include the city, number of bedrooms, number of bathrooms, square footage, street, and the price of a property. Both the picture and structured data make it possible to incorporate numerical and visual features into the suggested framework.

Even though the original dataset includes various variables, only those that are necessary for the prediction process have been included into the model development stage. The following structured features have been used:

- ✓ City(cit) (numerical representation)
- ✓ Bedrooms
- ✓ Bathrooms
- ✓ Square Feet (Sqft)

The target variable for this analysis is the market price of property that was subjected to transformation through a logarithmic function while building the model.

Besides structured property features, each image of the property was fed to the pre-trained ResNet50 deep learning convolutional neural network. Images were encoded into 2048 dimensional vectors representing features extracted from the image. The resulting image features were then concatenated with structured property features to form hybrid feature vector.

Table 3.1: Structured Features Used in the Proposed Model

Feature	Description	Type
City	Encoded city identifier	Categorical
Bedrooms	Number of bedrooms	Numerical
Bathrooms	Number of bathrooms	Numerical
Square Feet	Property area (sqft)	Numerical
Property Image	Exterior property image	Image
Property Price	Target variable	Numerical

3.4 Data Collection.

Data collection constitutes the initial phase of building the suggested Clarivision AI system. The (Real_estatewith_image) real estate dataset was chosen as the primary source of data for this study. This publicly accessible dataset was selected from Kaggle because it offers both structured property information and property images, making it ideal for the development and testing of hybrid artificial intelligence systems for automated property valuation.

This dataset consists of 15,474 records of residential properties characterized by each property real estate image as well as several descriptive characteristics. These characteristics include city, street, number of rooms, number of bathrooms, square footage, and market price. Availability of structured data and corresponding property image allows the suggested approach to utilize complementary information provided by multiple data types.

The collected dataset was verified prior to modeling, ensuring the presence of a valid property image for each structured record. This check was necessary due to the nature of the hybrid approach, which requires synchronization of structured and image data for each training sample.

After the data collection, the dataset was processed through multiple preprocessing steps prior to the use for training and evaluation.

21 3.5 Data Pre-processing.

Data pre-processing is one of the most critical steps in building an artificial intelligence model since the quality of the data used influences the accuracy of the predictions. In preparing the proposed hybrid model, the obtained data set was processed to enhance consistency, reduce data quality challenges, and ensure that there was proper representation of both the structured property data and images.

There were two major processes during the pre-processing of the data set, namely data cleaning and feature engineering.

3.5.1 Data Cleaning

A preliminary investigation of the structured dataset of real estate properties was made in order to determine any possible problems related to data quality. Missing data, duplicated data, and inconsistent data were assessed in order to be sure that only valid data is used in the learning process. Properties with missing data or invalid links to images were omitted in order to avoid inconsistencies between structured data and its corresponding images.

8 In addition, data analysis was conducted in order to find any abnormal data which might have a negative impact on the performance of the model.

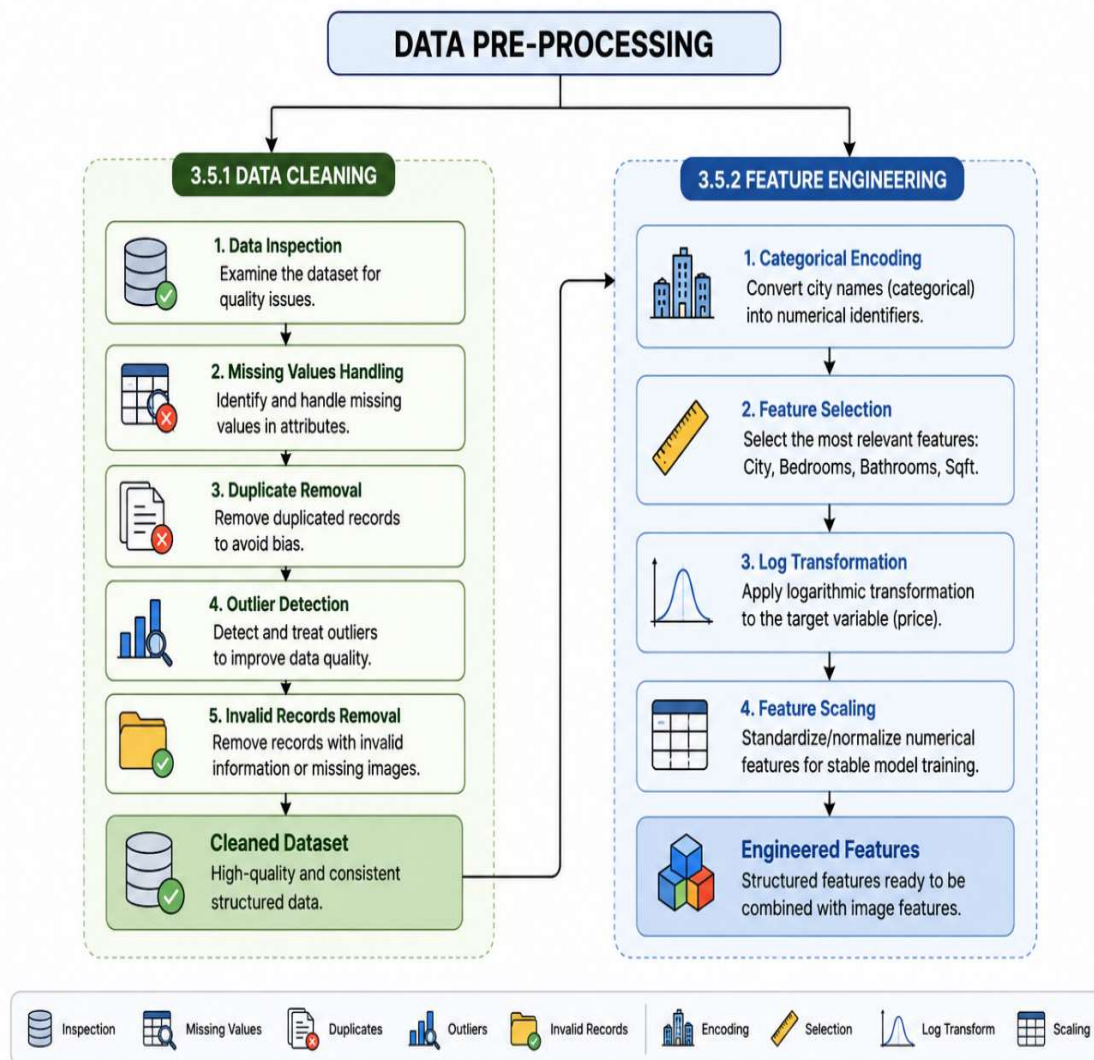
3.5.2 Feature Engineering

- 25 After performing data cleansing, a number of feature engineering steps were carried out in order to enhance the learning capacity of the developed model. The only attributes used for estimating the price of the property are the city of residence, the number of bedrooms, the number of bathrooms, and the area.

As machine learning algorithms do not work with categorical attributes, city names were coded numerically using categorical encoding. Additionally, the target attribute – price of property – is also coded using a log function, which will reduce the skewness of the price.

Finally, this structured feature vector is merged with the deep image features from ResNet50 to create the hybrid feature representation.

- 7 Figure 3.2: Data preprocessing workflow adopted in the proposed Clarivision AI framework.



- 7 summarizes the complete preprocessing workflow adopted before training the hybrid prediction model.

3.6 Property Image Processing

Property images contain valuable visual information that is impossible to encode in structured numerical attributes. Thus, an image processing pipeline was designed to extract high-level visual features describing the physical aspects of each property. The resulting deep features were combined with structured property attributes in order to increase prediction performance of the suggested hybrid approach.

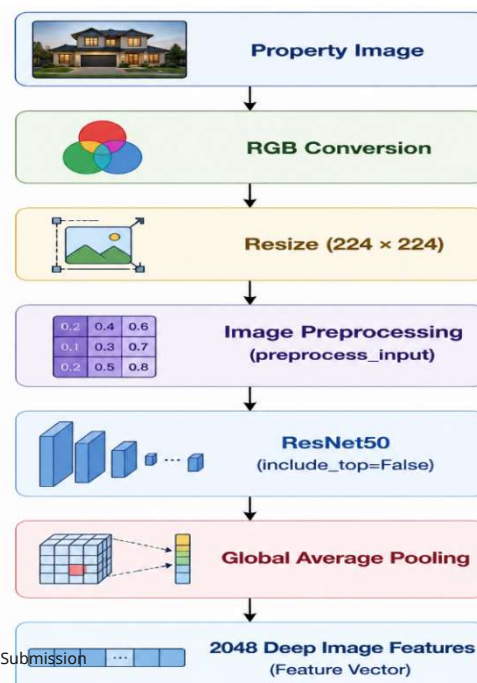
First of all, all property images that were uploaded were converted to the RGB color scheme to make sure that they would be compatible with the pre-trained convolutional neural network.

- 6 After that, all property images were resized to the size of (224 × 224 pixels), which corresponds to the standard input size for ResNet50 architecture. Before performing feature extraction,
- 8 images were normalized via the (ResNet50 preprocessing function) **ResNet50**, a convolutional neural network (CNN) pre-trained on the ImageNet dataset.

- 4 After preprocessing, the images were fed to the pre-trained (ResNet50) convolutional neural network. Instead of using the output of the final classification layer, the classification head was
- 22 omitted by using parameter (include_top=False), while the global average pooling (pooling='avg') was used to produce a compact image representation. In result, each property image was converted into the (2048-dimensional deep feature vector).

The extracted deep image features carry high-level semantics such as architectural appearance, façade type, exterior quality, among others, that are learned by the convolutional neural network. The feature vectors from this process were then merged with the property attributes in the hybrid feature fusion phase.

Figure 3.3: Pipeline for Property Image Processing.



3.7 Hybrid Feature Fusion.

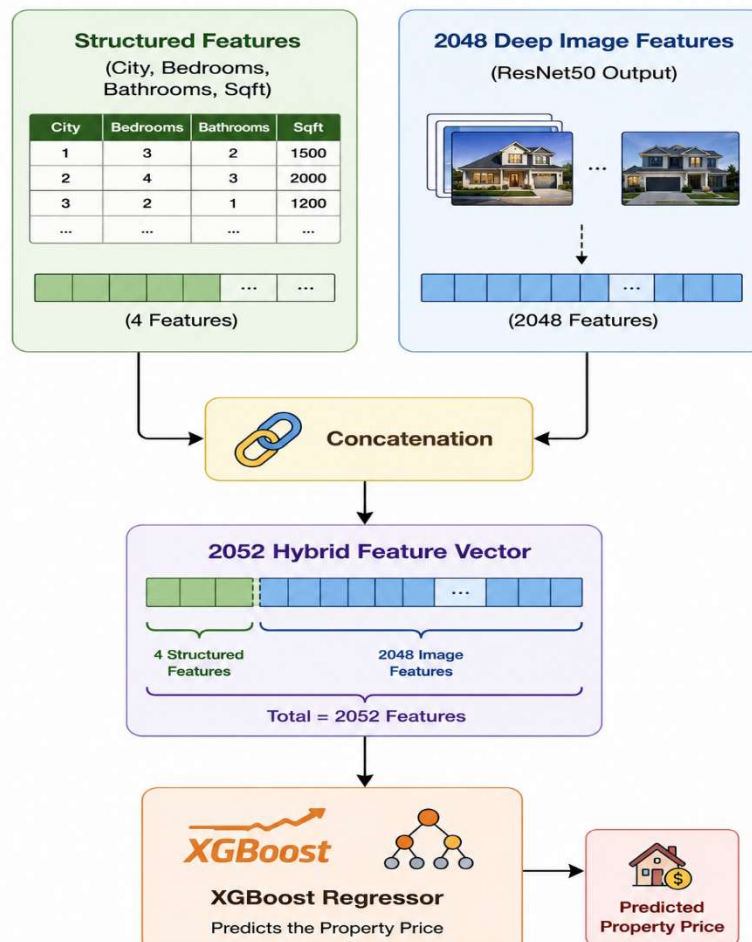
The proposed Clarivision AI system employs a combination of structured property attributes and deep image representations to form a unified hybrid feature vector for property value estimation. The hybrid representation ensures that the prediction model can make use of both numerical and visual property information simultaneously.

The structured data comprise four property attributes: encoded city name, number of bedrooms, number of bathrooms, and size of the property in square feet. The four variables represent the physical and geographical aspects of the property.

On the other hand, every property image is analyzed using the pre-trained ResNet50 convolutional neural network. Through feature extraction, every property image is transformed into a 2048-dimensional (4 structured features + 2048 deep image features) deep feature vector capturing the high-level visual content of the image learned from the property image.

Lastly, the structured data are combined with the deep features extracted from the image to obtain a 2052-dimensional hybrid feature vector to be used as the input to the XGBoost regression model.

Figure 3.4: Hybrid Feature Fusion Process.



3.8 Model Training

- 5 With the construction of the hybrid feature vectors, the dataset was split into the training set and the testing set for developing and evaluating the proposed prediction model. Hybrid feature vector, which includes structured property attributes and deep image features by ResNet50, was selected as the input of the XGBoost regression algorithm.

3.8.1 Data Splitting

- 6 For a more unbiased assessment of the proposed model, the dataset was split into two subsets:
 - 4 • **80% Training Set:** Used to train the XGBoost regression model to identify the relationships between the hybrid features and property prices.
 - 26 • **20% Testing Set:** Used to assess the prediction performance of the trained model on new data.

This train/test splitting process makes the evaluation of the model to be conducted under realistic conditions and gives an objective measure of the generalization capability of the model.

3.8.2 Cross-Validation

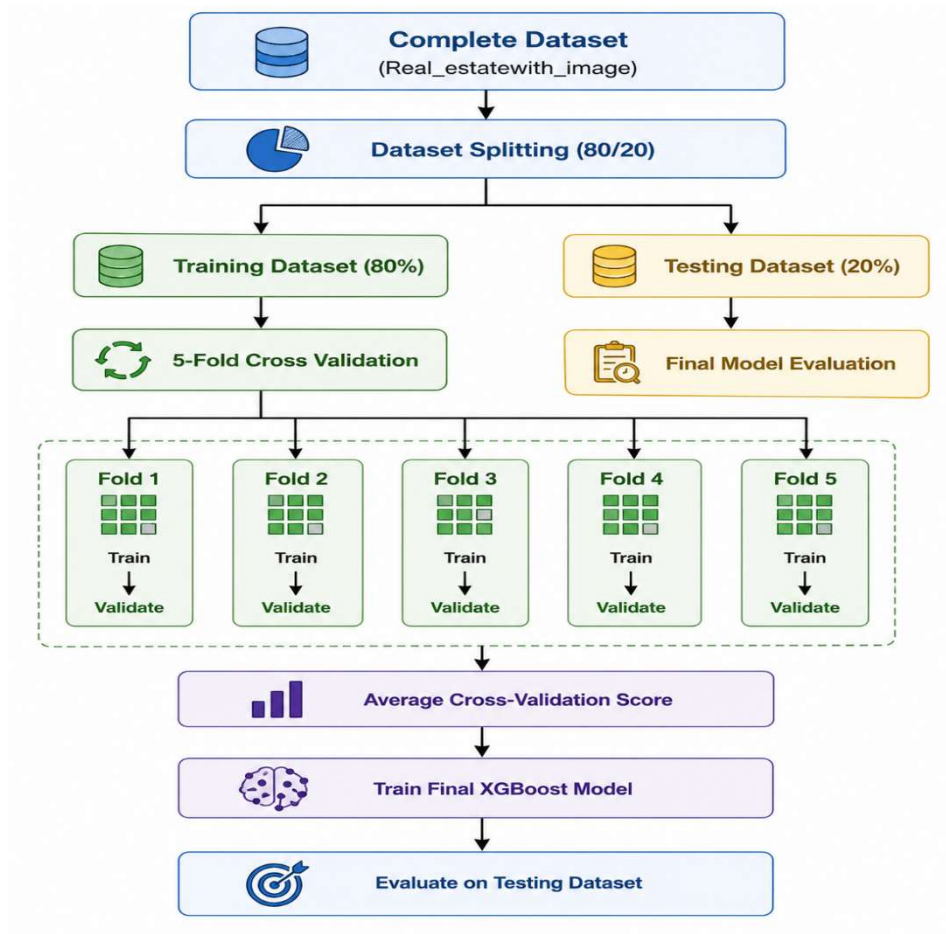
- 19 In addition to the data splitting process, 5-Fold Cross-Validation technique was also used during the modeling process to test the robustness and stability of the proposed prediction model.

- 2 The entire dataset for training was partitioned into five equally sized sets. In every run, four of the sets were used for training the model, whereas the other one was used for validation. This process was performed five times so that every subset was eventually used for the purpose of validation at least once. The overall performance was computed as the average of the five performances obtained by the validation process.

Cross-validation helps in avoiding overfitting of the model and gives a better estimate of its prediction performance.

Shuffle=True and Random State=42 were applied to ensure reproducibility.

Figure 3.5: Dataset Splitting and Cross-Validation Process



3.9 Hybrid Prediction Model

After the feature fusion, the created 2052-dimensional hybrid feature vector became the input of the prediction model. The proposed Clarivision AI uses the Extreme Gradient Boosting (XGBoost) Regressor to predict the market price of residential houses.

It was chosen due to its effectiveness among the algorithms of ensemble learning that is used for structured regression tasks. As compared to traditional regression models, XGBoost offers increased accuracy of predictions, fast training, built-in regularization functions, and proper consideration of nonlinear dependencies between variables. All these properties of XGBoost made this algorithm the best fit for the problem of real estate valuations when the property price depends on the interaction of its attributes and visual characteristics.

As stated above, the hybrid feature vector includes four structured attributes (city, bedrooms, bathrooms, and square feet) along with the 2048 deep image features obtained with the help of pre-trained ResNet50 model. The final 2052-dimensional feature vector allows XGBoost to take advantage of both numerical and visual features at once.

In order to improve the learning accuracy of the regression model and minimize the effect of outliers in property prices, the target variable (property price) was log-transformed prior to training. When predicting the property price, the logarithmic value obtained from the training process was reversed to the original scale by means of exponentiation.

The learned XGBoost regression model was saved as `true_hybrid_model.pkl` and then incorporated into the Clarivision AI application. The final trained model was stored as `true_hybrid_model.pkl` and deployed within the Clarivision AI application.

3.9.1 Transformation of Target Variable

The distribution of property prices had a positive skew, which could harm the learning procedure for the regression algorithms. In order to minimize skewness and normalize the variance, the target variable (property price) was log-transformed prior to training the models.

The formula for log-transformation is:

$$\begin{aligned} &[\\ &\text{Price}_{\{\log\}} = \log(1 + \text{Price}) \\ &] \end{aligned}$$

where:

- * (Price) is the actual property price.
- * (Price_{log}) is the transformed target value used during the training of the model.

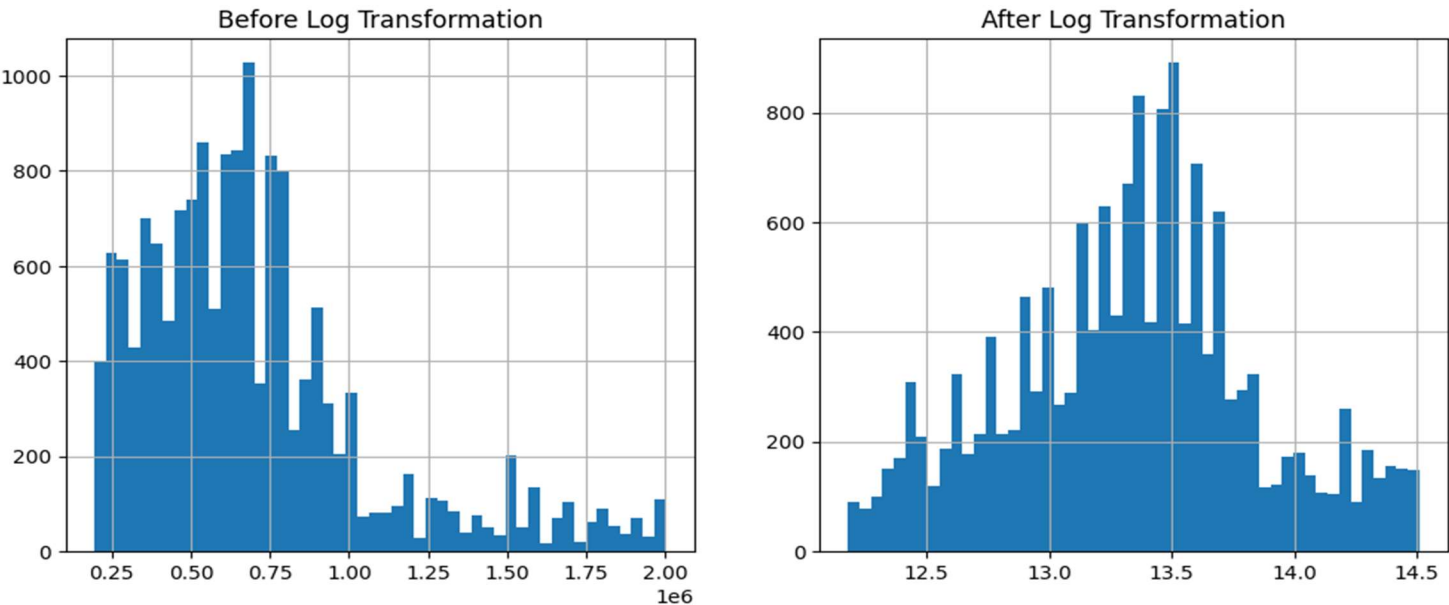
Once the XGBoost model provided the logarithmic prediction, the actual property price was reversed to its original scale through the inverse of the exponential function:

$$\begin{aligned} &[\\ &\text{Price} = e^{\{\text{Prediction}\}} - 1 \\ &] \end{aligned}$$

where:

- * (Prediction) is the logarithmic property price predicted by the XGBoost model.
- * (Price) is the final estimated property price that is shown to the user.

Figure 3.6 Distribution of Property Prices Before and After Log Transformation.



The distribution of property prices before and after the logarithmic transformation is shown in Figure 3.6 The property prices before the logarithmic transformation had a very skewed distribution on the right side. However, after the logarithmic transformation, the distribution had a symmetrical pattern and was close to the normal distribution. This reduced the impact of outliers on the learning capability of XGBoost regression model.

1 Table 3.2 Hyperparameters of the XGBoost Model.

Hyperparameter	Value
Objective	reg:squarederror
Number of Estimators (n_estimators)	300
Learning Rate (learning_rate)	0.05
Maximum Depth (max_depth)	6
Subsample (subsample)	Default (<i>Not specified</i>)
Colsample_bytree (colsample_bytree)	Default (<i>Not specified</i>)
Random State (random state)	42

The XGBoost regression model was configured using a set of hyperparameters to optimize prediction performance. The selected hyperparameters used in the final hybrid prediction model are summarized in Table 3.2.

3.9.2 Explanation of the Selected Hyperparameters

The final XGBoost regression model was designed utilizing an appropriately chosen set of hyperparameters that achieved the best trade-off between accuracy of predictions and model complexity. Description of the chosen parameters is provided below:

- **Objective (reg: squared error).**

- 23 As the objective function in XGBoost, the squared error objective function is used due to the fact that the proposed system is designed to perform regression analysis, where the task is to predict the continuous price values of the properties instead of classification of categorical labels.

- **Number of Estimators (300).**

- 1 The parameter "number of estimators" indicates the total amount of boosting trees generated by the algorithm during its work. Value of ****300**** was chosen since it allows the model to have enough learning capacity without losing stability and not causing computational burden at the same time.

- **Learning Rate (0.05).**

- 1 The learning rate determines the contribution of the boosting tree into the training process. Small value of ****0.05**** was chosen for this parameter to allow gradual learning of the model.

- **Maximum Depth (6):**

Maximum depth represents the complexity of each decision tree. Maximum depth of 6 was chosen as a proper tradeoff between making the model capture complicated patterns of the real estate data and avoid overfitting due to high complexity.

- **Subsample (Default):**

- 1 This hyperparameter is responsible for the proportion of training samples being used for each tree construction. As there was no manual configuration, the default settings were used.

- **Colsample_bytree (Default):**

It is the proportion of features which are randomly selected during each tree construction process. Default settings were used as it has been proven by preliminary experiments that the model predicts sufficiently well with them.

• Random State (42):

A constant random state of 42 was used for the purpose of reproducing the results of the experiment. It makes sure that dataset splitting and model initialization would be the same when running the system repeatedly.

3.10 Similar Property Retrieval.

Aside from estimating the value of the property in the market, the proposed Clarivision AI framework also comes with a property retrieval module which helps users compare their property estimate with other similar-looking properties available in the database. The said function aids in providing better interpretation of the estimation results as well as market references.

28

Once the price of the property has been estimated, the uploaded property image will be processed through the same ResNet50 feature extraction pipeline used during the training process of the model. This will result in the generation of a 2048-dimensional deep feature vector for the visual features of the uploaded property image.

In order to ensure more relevant property retrieval results, the search will initially be done based on properties within the same city chosen by the user. The reason for this is that the geographical filtration guarantees that the retrieved properties are within the same market as the property whose price was estimated.

Finally, after city-based property filtration, the extracted deep feature vector will be compared with the stored feature vector of all properties within the chosen city using Cosine Similarity.

Cosine Similarity was selected because it effectively measures the similarity between high-dimensional feature vectors extracted from ResNet50 regardless of their magnitude.

9

The Cosine Similarity is computed as below:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}| \times |\mathbf{B}|}$$

Where:

A = Feature Vector of the uploaded image of the property.

B = Feature Vector of the stored property image in the database.

13

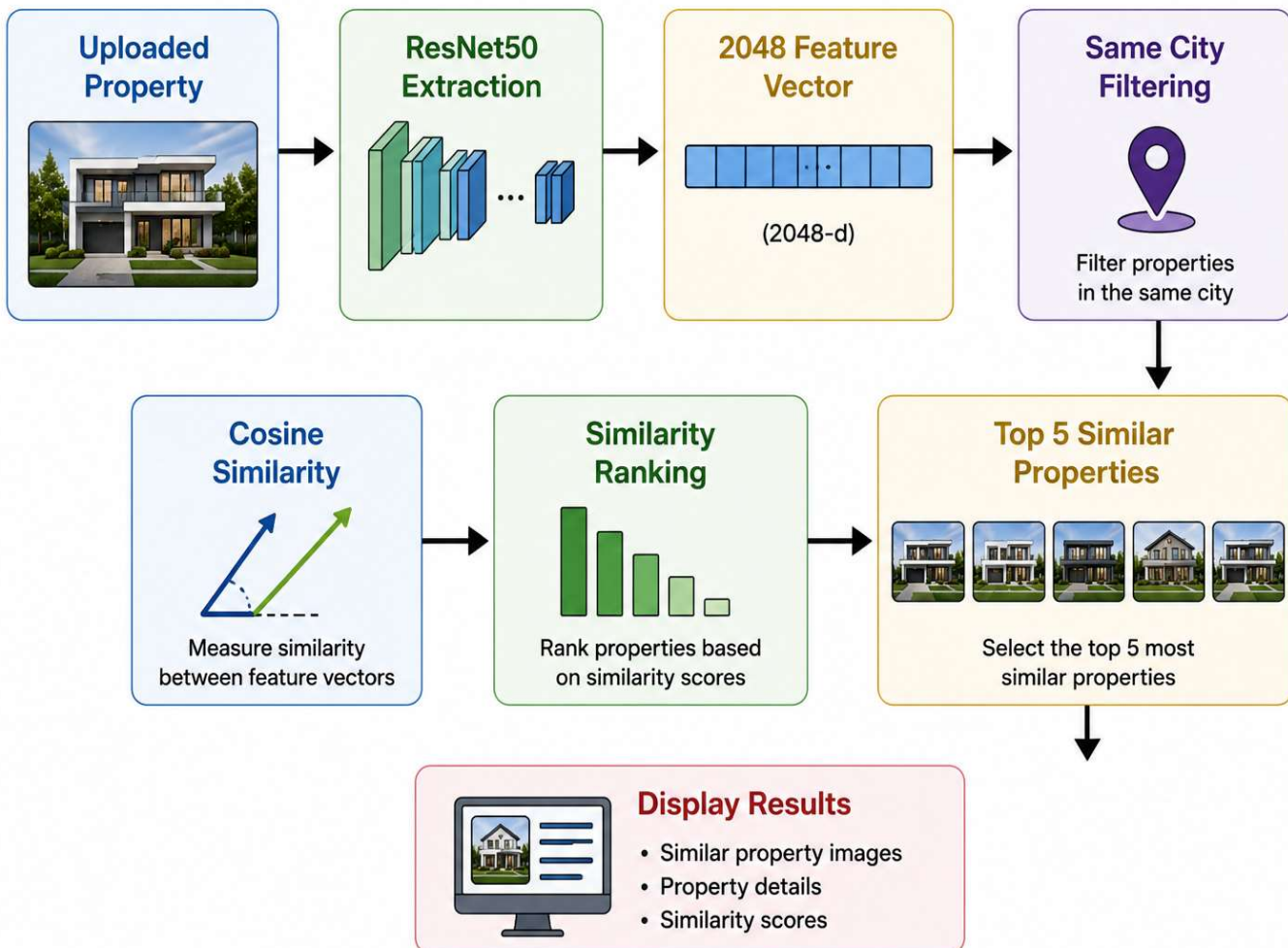
The similarity score ranges from 0 to 1, where higher values indicate greater visual similarity between

Top five similar properties in the same city based on the computed similarity values are shown to the user along with the following details:

- 2 Image of the retrieved property, Similarity Percentage (Calculated), City, Number of Bedrooms, Number of Bathrooms, Square Footage, and Market Price.

The integration of geographic filtering and visual similarity has made the proposed search/retrieval module capable of providing highly relevant similar properties to validate the estimated market value.

Figure 3.7 Similar Property Retrieval Process



3.11 Model Evaluation.

- 4 For the purpose of measuring the efficiency of the proposed hybrid predictive model, a number of commonly accepted regression performance criteria were utilized. These criteria help in obtaining a thorough evaluation of the accuracy of predictions and the reliability of the model as a whole.

This evaluation was carried out based on the testing dataset which had been used after the stage of model training had been accomplished. Given the nature of the task which involves predicting

- 18 continuous property prices, regression criteria were chosen to measure the performance of the model. The following evaluation criteria were used within this research:

16 ○ Mean Absolute Error (MAE)

Mean Absolute Error (MAE) represents an average absolute deviation of the predicted property prices from the real market prices. The lower the MAE, the higher prediction accuracy and model performance are.

14 ○ Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the average magnitude of prediction errors but with larger weight being assigned to the larger errors. RMSE criterion is especially helpful in identifying those models which make large prediction mistakes.

○ Coefficient of Determination (R^2)

- 17 The coefficient of determination (R^2) provides an index that indicates the proportion of the variability in the property price data set that can be explained by the prediction model. The value of R^2 close to one shows the success of the model in describing the relationship between the input variables and the output variable.

○ Cross-Validation

- 30 In order to test the robustness and generalization ability of the developed model, the process of k-Fold Cross-Validation was used while developing the model. This method tests the model performance on several subsets of data sets and minimizes the possibility of over-fitting the model. The results derived from these evaluation methods are described in Chapter 4.

Table 3.3 Evaluation Metrics Used in This Research.

Metric	Purpose
MAE	Measures the average absolute prediction error
RMSE	Measures the magnitude of prediction errors
R ²	Measures the proportion of explained variance
Cross Validation	Evaluates model robustness and generalization

3.12 System Workflow.

The complete workflow of the proposed Clarivision AI is shown in Figure 3.X. The process starts with the user providing structured information about the property's attributes in terms of the city name, number of bedrooms, number of bathrooms, and size of the property. Furthermore, a property image is provided by the user using the graphical user interface.

After the inputs have been obtained, the structured attribute data are encoded, and the uploaded property image is preprocessed with operations that include converting the image into RGB, resizing the image, and normalizing the image. Subsequently, the preprocessed image is fed into the pre-trained ResNet50 convolutional neural network model, which extracts 2048-dimension vector for the property.

The image-based feature vector is concatenated with the structured attributes of the property to form a 2052-dimension hybrid feature vector. Then, the feature vector is used as an input for the trained XGBoost regression model to predict the market price of the property.

After the prediction of the property price, the system identifies visually similar properties in the database. In order to retrieve visually similar properties, first, all the properties which fall in the same city as chosen by the user are filtered out. Cosine similarity is then used to calculate the similarities between the uploaded property photo and all the photos of the property available in the same city. The top five similar properties based on the similarity score are then displayed to the user along with their respective details.

Finally, the entire valuation report provided by the Clarivision AI system includes predicted property price, Reliability Score, market category, similar properties, feature importance analysis, market dashboard, and the automatic generation of PDF property valuation report.

Figure 3.8: Overall Workflow of the Proposed Clarivision AI System.

