



Project Proposal

Developing A Hybrid Predictive Model by Using Images and Digital Data to Estimate Real Estate Market Value with Ai

Prepared By: Hassan Mohammed Ahmed Alamer Khalifa

ID No: 253002941

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Name Of Supervisor: Dr / Amjad Qtaish

Collage of Informatics, Midocean University, Comoros

Email: 253002941@st.midocean.edu.ki

Chapter 1 Introduction

We always know that real estate valuation is a linchpin of global finance it determines investment choices, mortgage approvals, and feasible risk assessments. Today, the advent of digital real estate platforms and the availability of massive property datasets have transformed the way valuations are done. Automated Valuation Models (AVMs) are now widespread, providing quicker and much more scalable valuation estimates than traditional appraisal methods **(Balila & Shabri, 2024) [1]**.

However, these models still have their limits. Most rely only on structured numerical data such as size, location, property age, and number of rooms. While these details matter, they don't capture everything that affects a property's value. The real estate market is not dominated by just numbers, other less quantifiable items like for example, if property is zoned for residential or commercial, how much of an effect upkeep, quality of interior finishes and even the "first impression" can make on the final price. o These qualities are difficult to measure **(Gümmer et al., 2025) [2]**.

In this context, deep learning has taken some of the work out of the process of gleaning something useful from photos of architecture and interiors in a variety of ways, often with surprisingly insightful results Computer vision has made it po... Valuation is difficult in part because image-based models and numerical data models are entirely different things, and don't usually interact The result is a kind of valuation where the process is broken into parts one model examines numbers one model looks at numbers, another at visuals-without a combined approach to bring the information together.

Another major challenge is explainability Deep learning models can be really good at making predictions, but they can be a bit mysterious. They don't always show how they got to a certain answer, which can make it hard for people like investors and regulators to trust them. When it comes to big financial decisions or legal issues, people need to know why a certain value was chosen. Without a clear explanation, it's only natural that people are going to be hesitant. This lack of transparency can be a major problem, especially in industries where trust and accountability are crucial. **(Sharma et al., 2024) [3]**.

This study aims to tackle those gaps the problem is that current machine learning methods don't provide a clear system that brings together property details and visual features to make reliable market predictions. To fix this, the research suggests a Hybrid Predictive Model that combines numerical and visual data into one easy-to-understand framework. The aim is to improve prediction accuracy and provide transparency to help people understand the real estate market. This model is designed to merge structured property information with visual features, making it easier to make sense of the market. By combining these two types of data, the model can provide more accurate predictions and help people make informed decisions. The goal is to create a system that is both reliable and transparent, allowing users to trust the predictions and understand the reasoning behind them. **(Ezzeldin & Abouagwa, 2025) [4]**.

1.1 Problem Definition

Determining property value plays a crucial role in the global economy, influencing investment decisions, mortgage securitization, and financial risk evaluation. When it comes to buying or selling a house, mistakes can be really costly. This is because these deals usually involve a lot of money. If the price of the property is not valued correctly, it can lead to big financial losses. In some cases, it can even affect the whole system and cause bigger problems.

Although the rise of digital technology and machine learning has brought Automated Valuation Models (AVMs) that offer speed and scalability, these models still face three key issues that limit their accuracy and acceptance by professionals.

The Qualitative Data Gap First, traditional AVMs mainly rely on quantitative data such as square footage, location coordinates, property age, and number of rooms. While these factors provide a starting point, they miss many important details of the real estate market (**Balila & Shabri, 2024**) [1]. In reality, buyers and investors consider qualitative, intangible aspects like the quality of interior finishes, maintenance standards, architectural style, and the overall feel of a property. These elements significantly influence market prices but often remain overlooked because current models focus solely on numerical data.

The Modality Integration Challenge Second, advances in Deep Learning and Convolutional Neural Networks (CNNs) have made it possible to extract features from property images. However, many systems handle image analysis and numerical data separately. This separation stops the AI from combining these different data types into a cohesive understanding. For example, a model might evaluate a “3-bedroom house” based on numbers alone, ignoring visual signs like cracks or wear that should affect its value, simply because the two data streams are not effectively integrated (**Ezzeldin & Abouagwa, 2025**) [4].

The Black Box and Trust Deficit Third, one major hurdle in using advanced AI in real estate is its lack of transparency. High-performing models, especially deep neural networks, often act as “black boxes,” offering a price estimate without explaining how it arrived at that figure. For financial institutions, investors, and regulators, this lack of clarity is a problem. They need to be able to justify valuation results for auditing and compliance. Without a system that explains which features-numerical or visual-affected the decision, it's hard to build the trust required for important financial uses (**Sharma et al., 2024; Chen, 2025**) [3].

Summary of the Problem The main issue this thesis tackles is that existing valuation models do not combine structured property data with visual information in a way that is both unified and interpretable. This gap leads to property value estimates that lack transparency and comprehensive accuracy, which are essential for reliable financial decisions. To address this, the research proposes a hybrid valuation model that merges numerical and

visual data within a single, explainable framework, aiming to produce accurate and trustworthy market value predictions.

1.2 Thesis Objectives

- This project aims to develop a new model that uses artificial intelligence to estimate real estate prices. The model will combine property details, such as size and location, with features extracted from images of the property. The goal is to create a model that is both understandable and accurate. To achieve this, the study will focus on several objectives:-
 - Develop a valuation model based on structured data. This involves building a predictive model using key property attributes, including area, location, age, property type, and other relevant factors.
 - Design a model that analyzes images. This will use deep learning to capture important visual cues related to the property's condition, finishing quality, and overall appearance.
 - Create a hybrid approach to integrate both structured data and image features into one predictive framework. The study will also explore how combining these different types of data affects the model's performance.
- Incorporate methods to explain the model's predictions. This includes clarifying the role of numerical features and identifying the visual areas or characteristics that influence the estimates.
- Incorporate methods to explain the model's predictions. This includes clarifying the role of numerical features and identifying the visual areas or characteristics that influence the estimates.
- Assess the model's robustness by examining how it performs when data are incomplete, missing, or noisy. Investigate the practical use of the model as a decision-support tool for investors, financial institutions, and digital real estate platforms.

1.3 Thesis Contribution.

This research aims to improve the way we use artificial intelligence to value real estate by filling in some important gaps in the current methods and practices being used. It looks at both the theoretical and practical sides of the issue, trying to make AI-driven real estate valuation more accurate and reliable. The key contributions can be summarized as follows: the key contributions can be summarized as follows:

1. A Unified Hybrid Valuation Approach.

This research suggests a new way of predicting things by using both numbers and pictures together. Instead of looking at them separately, it thinks they can help each other to give a better understanding (Ezzeldin & Abouagwa, 2025) [4].

2. Emphasis on Explainability.

The new model is different from other systems that are hard to understand. It makes things clear in two ways: with numbers and with pictures. This makes it easier for people to see what's going on and helps organizations feel comfortable using it (Sharma et al., 2024) [3].

3. Empirical Evidence on Multimodal Integration.

The study looks at how different methods can help figure out how much a property is worth. It tests three ways: using just numbers and details, using just pictures, and combining both. By doing this, it shows how using both numbers and pictures together can add value to understanding the worth of a property (Li, 2025; Arshad et al., 2024) [14][5].

4. Robustness Analysis.

The model is evaluated under conditions that reflect real-world market environments, including incomplete and heterogeneous data availability.

5. Practical Orientation.

Practical Orientation Beyond theoretical development, the framework is designed with implementation feasibility in mind. It aims to balance predictive accuracy and interpretability, making it suitable for real-world financial and investment applications (Riveros et al., 2024) [6].

1.4 Thesis Organization

Chapter 1: Introduction

Introduces the research problem, objectives, and contributions of the study.

Chapter 2: Background and Related Work

Reviews traditional real estate valuation methods and previous research related to machine learning and image-based valuation models.

Chapter 3: Proposed Approach

Presents the proposed hybrid model and describes the data, methodology, and model architecture.

Chapter 4: Experimental Results and Discussion

Reports the experimental results, compares different models, and discusses the findings.

Chapter 5: Conclusion and Future Work

Summarizes the main results of the study and suggests directions for future research.

Chapter 2: Background and Related Works

2.1 Background

Traditional methods of real estate appraisal have long depended on the judgment of professionals and comparisons with recent sales in the area. While these approaches form the basis for valuation, they often rely heavily on the appraiser's personal experience and the availability of local data. This can cause inconsistencies in value estimates and makes it hard to scale the process in markets that change quickly. The rise of online real estate platforms and large databases has helped address some of these challenges by offering extensive information on property details, transaction history, and location factors.

The growth of these large property datasets led to the creation of the first Automated Valuation Models (AVMs) (Balila & Shabri, 2024) [1]. Early versions used statistical methods like Hedonic Pricing Models and multiple linear regression to estimate how features such as size and location affect price. Yet, these approaches often fall short when dealing with the complex, non-linear nature of real estate markets. Advances in machine learning have brought about more advanced algorithms like Decision Trees, Random Forests, Support Vector Machines, and Gradient Boosting (Arshad et al., 2024) [7]. These models handle large and complex data better and can capture the complicated interactions among variables, often providing a more accurate estimate of a property's market value than traditional statistics.

Beyond numbers, a property's value is influenced by visual factors like design, lighting, and upkeep that affect buyer perception. Recent breakthroughs in Deep Learning, especially Convolutional Neural Networks (CNNs), have transformed how property images are analyzed (Gümmer et al., 2025) [2]. CNNs can now detect subtle visual details—such as the quality of finishes, interior styling, and structural condition—that were previously beyond the reach of automated tools. However, most current valuation systems treat numerical data and visual information separately. This division limits their ability to fully understand market complexity, since property value depends on both measurable features and visual appeal combined.

A major issue with complex machine learning models, particularly deep neural networks, is their lack of transparency. They often function as “black boxes,” delivering accurate results but failing to explain how decisions are made (Sharma et al., 2024; Chen, 2025) [3]. In areas like finance and real estate, this opacity hinders trust and wider use. Stakeholders such as investors, lenders, and regulators need clear explanations of valuation methods to ensure fairness and accuracy. This need has pushed the adoption of Explainable AI (XAI), which aims to make model decisions understandable. XAI helps build trust by allowing users to see how both numerical and visual information drive the outputs, supporting stronger justification of financial decisions.

Current research is moving toward Hybrid Multimodal Models that combine numerical property data with visual features within one integrated system. These models try to replicate the well-rounded judgment of a human expert

by merging precise quantitative data with the nuanced insights drawn from images. This study contributes by proposing a system that aims not only to improve accuracy and reliability but also to be transparent and practical for real estate industry use.

2.2 Related Work

Recent research has been paying more attention to how machine learning and deep learning can be used in valuing real estate and predicting housing prices. Scholars have worked with different types of data—like property details, location information, written descriptions, and images of properties—to try to make predictions more accurate

- **Mora-Garcia et al. (2022)** An extensive investigation was carried out to compare how different machine learning algorithms perform in predicting residential property prices. The study focused on the transition from traditional statistical methods to more advanced computational approaches, such as Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN). The researchers found that with large datasets, modern machine learning methods tend to outperform older models by capturing complex patterns that traditional techniques often miss. The study used a large geospatial dataset combining detailed property features with precise location information. One key benefit of this research is its empirical evidence showing that ensemble methods and neural networks can provide more accurate predictions and handle big data more effectively in the real estate field. However, these models demand significant computational power and specialized expertise to deploy properly, which limits their accessibility compared to simpler regression tools. From a research standpoint, the work by Mora-Garcia et al. (2022) [8] has a limitation described as Visual Omission. Although the study demonstrated the strengths of machine learning in processing spatial and numerical data, it did not include unstructured visual data from property images. This omission is important because numerical data alone cannot capture the aesthetic premium or the true condition of a home's interior. This points to a crucial gap in the research: the need to go beyond purely numerical models. The present thesis plans to fill this gap by integrating the high-performance algorithms identified by Mora-Garcia with computer vision techniques to achieve a more comprehensive and accurate valuation.
- **Wang et al. (2022)** A specialized spatial modeling framework called Geographical Neural Network Weighted Regression (GNNWR) was introduced. This model combines the non-linear learning abilities of Artificial Neural Networks (ANN) with the spatial analysis features of Geographically Weighted Regression (GWR). By merging these methods, the researchers aimed to address the issue of spatial non-stationarity seen in real estate markets, allowing the model to adjust its weighting based on each property's geographic context. The study used an urban housing dataset enriched with spatial information, linking traditional property features with exact GPS coordinates and neighborhood-specific environmental variables. The main strength of the GNNWR model lies in its sensitivity to spatial differences; it can effectively capture unique neighborhood effects and local market changes that broader models often overlook, which improves valuation accuracy in complex urban settings. However, a notable drawback is the model's complexity, requiring large amounts of geographic data and careful

tuning to avoid becoming too narrowly focused or overfitting to a small area like a street or block. From a research standpoint, the work of Wang et al. (2022) [9] faces what's called Sensory Narrowness. While the GNNWR model is good at understanding how location impacts housing prices, it treats properties mainly as sets of coordinates and numerical data. It lacks computer vision to evaluate the actual condition of buildings or the quality of their architecture. This highlights an important gap: the need to combine advanced spatial weighting methods with visual feature extraction. This thesis aims to address that gap by making sure a property's value is influenced not just by its map location but also by its visual condition and physical appearance.

- **Rodriguez-Serrano (2024)** A specialized prototype-based machine learning model was developed to provide explainable predictions of real estate prices. Unlike traditional black-box methods, this approach selects certain representative properties from the dataset as benchmarks. By focusing on these tangible examples, the model offers a more human-like way of reasoning, letting users understand which similar houses most influenced the estimated value of their own property. The study used a standardized dataset of residential real estate, rich in transaction records and property details. The main benefit of this model lies in its interpretability; it helps connect advanced machine learning results with human intuition by providing case-based explanations. This makes it easier for financial institutions and homeowners to review and understand the valuation process. However, the model's accuracy depends greatly on the quality and variety of prototypes. If the dataset does not include a good representative example for a unique or luxury home, the predicted value may be less accurate. From a research standpoint, Rodriguez-Serrano's method faces limitations due to its reliance on a fixed set of features. While prototype-based explanations work well for structured data like room counts or square footage, they have not yet adapted to visual characteristics. For instance, the model can find prototypes by size but cannot identify ones by modern interior design or architectural style because it lacks a deep learning visual component. This points to a key gap: the need to extend prototype-based explainability to cover multiple types of data. This thesis aims to address that by using prototypes that combine both numerical information and visual elements **Rodriguez-Serrano (2024)** [15].

- **Balila and Shabri (2024)** A comprehensive comparative analysis was carried out to assess different machine learning architectures for their effectiveness in Dubai's unique real estate market. The study examined several algorithms, including Support Vector Machines (SVM), Gradient Boosting Machines (GBM), Random Forests (RF), and Artificial Neural Networks (ANN). The results showed that ensemble models, which combine multiple algorithms, generally delivered the most accurate predictions, especially when dealing with the diverse, high-value property segments typical of a global city like Dubai. The research used a localized dataset that included a variety of transactional details such as property type, age, location premiums, and proximity to major landmarks like the Burj Khalifa. One key contribution is the regional evidence supporting ensemble methods' ability to manage the "luxury premiums" and the rapid market changes seen in Middle Eastern real estate. However, these models require careful hyper-parameter tuning, as improper adjustment can lead to complexity that hinders their

use in real-time price evaluation. From a research standpoint, the work by Balila and Shabri (2024) [10] has a notable limitation, which could be called Visual Blindness. While they identified strong numerical predictors for the Dubai market, their analysis overlooks factors like aesthetic qualities and finishing standards that often influence prices in luxury properties. In Dubai, two apartments within the same building can vary widely in price due to interior design or views—elements not captured in conventional datasets. This gap points to the importance of adding Computer Vision (CV) techniques to enhance these ensemble models. Building on their findings, this thesis aims to incorporate visual feature extraction to better account for the “luxury factor” in property valuation.

- **Hasan et al. (2024)** The study introduced a complex multimodal deep learning framework aimed at transforming property valuation by combining three types of data: structured tabular data, high-resolution images, and textual descriptions from property listings. Using a multi-branch neural network, the researchers showed that integrating these different data sources leads to better prediction accuracy and lower error rates compared to traditional models that rely solely on numerical data. The dataset came from various online real estate platforms and included thousands of listings, each linking numerical features like square footage with interior and exterior photos, as well as descriptive agent notes. The main strength of this approach lies in its broad use of data, similar to how a professional appraiser assesses both the look of a property and its key selling points before making a decision. On the downside, combining these diverse data types is quite complex. Training a model to balance the importance of an image with a line of text demands significant computing resources and a large, clean dataset. From a research angle, Hasan et al. (2024) [11] face what they call an Interpretability Gap. Although their work confirms that adding more data sources improves accuracy, the model itself remains opaque, making it hard for users to understand exactly why a particular image or phrase affected the price prediction. This reveals a crucial gap in the field—the development of Multimodal Explainable AI (XAI). Building on Hasan’s multimodal model, this thesis aims to add transparency, enabling the system to not only process images and text but also clarify how those visual and textual features influence the final valuation for users.
- **Sharma et al. (2024)** The study proposed a complex multimodal deep learning framework designed to improve property valuation by combining three types of data: structured tabular data, high-resolution images, and textual descriptions from property listings. By using a multi-branch neural network, the researchers demonstrated that integrating these varied data sources results in better prediction accuracy and lower error rates than traditional models that rely only on numerical data. The dataset was collected from multiple online real estate platforms and included thousands of listings. Each listing linked numerical features like square footage with interior and exterior photos, along with descriptive notes from agents. The main advantage of this method is its broad use of data, akin to how a professional appraiser evaluates both a property’s appearance and its key selling points before making a judgment. However, combining these different data types adds complexity. Training a model to weigh the significance of an image against a line of text requires substantial computing power and a large, well-

organized dataset. From a research perspective, Hasan et al. (2024) [11] highlight what they call an Interpretability Gap. While their findings show that adding more data sources boosts accuracy, the model itself remains difficult to interpret. This makes it challenging for users to understand why a specific image or phrase influenced the price prediction. This gap points to a key challenge in the field—the need for Multimodal Explainable AI (XAI). Building on Hasan’s multimodal model, this thesis aims to increase transparency by enabling the system not just to analyze images and text, but also to explain how these features affect the final valuation in a way users can understand.

- **Riveros et al. (2024)** investigated applying XGBoost (Extreme Gradient Boosting) to enhance the accuracy of property value estimates. Their research showed that this machine learning approach can effectively capture the complex, non-linear relationships between various house features—like size and location—and their market prices. By using the gradient boosting framework, the study identified patterns that traditional models often overlook, emphasizing that property features rarely relate to value in a simple, linear way. The study used a structured dataset of residential real estate, including thousands of past transactions with detailed physical and location information. One strength of this work lies in its computational power; XGBoost handles high-dimensional data well and reduces the chances of overfitting, making it suitable for large-scale market predictions. However, a key drawback is the model’s black box nature—it doesn’t clearly explain how specific feature interactions lead to a given price. This lack of transparency can be problematic for professional appraisals. From a research standpoint, Riveros et al. (2024) [6] face what they call a Perceptual Gap. While XGBoost offers precise valuations based on numerical inputs, it cannot assess the visual or qualitative condition of a property. For example, it fails to differentiate between two houses of similar size and location if one has been luxuriously renovated and the other is neglected. This gap points to a clear need: combining high-performing gradient boosting methods with computer vision techniques. This thesis aims to address this by integrating XGBoost predictive capabilities with visual feature analysis to better reflect a property’s physical state.

- **Arshad et al. (2024)** This research advanced Explainable AI (XAI) in real estate by applying Cat Boost and LightGBM frameworks to predict housing prices. To address the black box issue common in these complex models, the study incorporated SHAP (Shapley Additive Explanations). This method provided a clear ranking of feature importance and showed exactly how variables like location and square footage influenced the final price estimate. The study used a detailed structured real estate dataset with thousands of past transactions and many tabular features. One key strength of this work is that it achieved both global and local interpretability, offering a mathematically grounded way to justify valuations to stakeholders such as banks and investors. Cat Boost was effective in handling categorical data with good efficiency. However, a notable drawback is that calculating SHAP values for large datasets demands high computational resources. Also, the model only works with structured data and ignores qualitative factors like the property's visual appeal and condition. From a research standpoint, Arshad et al. (2024) [7] encounter a Perceptual Blind Spot because their model cannot explain the reasoning behind visual attractiveness. While it can quantify how Location affects value, it cannot clarify

how modern finishes, visible in photos, impact price. This gap points to an important need for a Multimodal XAI framework that can provide explanations for both digital data and visual features at once. This thesis aims to address this gap by integrating these two data types into a single transparent system.

- **Azam and Rai (2025)** The study examined how effective different machine learning methods—Random Forest (RF), Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN)—are at improving house price predictions. It highlighted that success in prediction depends not only on the choice of algorithm but also heavily on the thoroughness of data preprocessing and evaluation. By comparing several models, the researchers found the best setups needed for accurate property valuation in today’s market. They used a large real estate dataset that included many past sales records along with various property features. The main strength of this work lies in its careful methodology; it shows that detailed data cleaning, feature design, and systematic testing of models are key to lowering errors in Automated Valuation Models (AVMs). Still, one limitation is that the study focuses only on uniform tabular data, which doesn’t capture the complex, multi-sensory aspects that real-world property appraisal often involves. From a research standpoint, the approach by Azam and Rai (2025) [12] is limited by what could be called “Data One-dimensionality.” Although they demonstrated the importance of data preparation for numerical accuracy, their model cannot handle unstructured visual information. This is a critical gap because even the most refined numerical data won’t reflect a property’s visual condition or architectural style. This suggests a clear need for research that brings the same rigor in preprocessing to a multimodal system combining both numerical data and property images, which is a challenge this thesis aims to tackle.
- **Chen (2025)** The study examined how machine learning’s role in predicting housing prices has changed, arguing that data-driven methods tend to capture complex market behaviors better than traditional statistical approaches. It compared Random Forest (RF), Gradient Boosting (GBM), and Artificial Neural Networks (ANN). Chen’s results showed that these algorithms are particularly good at spotting non-linear patterns and hidden links in property features that older models often miss. The research used a long-term dataset of urban housing sales, including detailed past prices, building characteristics, and local economic data. One clear strength of this work is the depth of its predictions, making a strong case for AI’s advantages in modeling unpredictable real estate trends. On the downside, these models can be hard to interpret; while they detect complex patterns, they rarely explain **why** certain price changes happen. Also, the study only used structured data. From a research angle, Chen’s work misses an important aspect—the physical condition of the properties. Although the data-driven method captures market trends, it doesn’t account for visible factors like interior wear or recent renovations that aren’t part of the structured data. This reveals a key gap: the need to combine data-driven forecasting with visual information. This thesis aims to fill that gap by combining Chen’s machine learning techniques with a computer vision component to create a more complete and accurate pricing model. **Chen (2025)** [13].

- **Li (2025)** [14] The study provided a detailed overview of how artificial intelligence has evolved in the real estate sector, showing the shift from traditional models to more advanced deep learning techniques. It covered a range of methods, including Random Forest (RF), Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN). Particular attention was given to the rising role of Convolutional Neural Networks (CNN) for image analysis and Recurrent Neural Networks (RNN/LSTM) for handling time-based market data. Li's research supports using Hybrid and Ensemble Machine Learning–Deep Learning models as an effective way to capture the complex nature of property valuation. The research used a global meta-dataset combining information from multiple international property platforms. This included structured transaction data, architectural images, and historical price trends. The main strength of this work is its comprehensive view, mapping the entire tech environment of modern Automated Valuation Models (AVMs) and showing how different neural network types can address various data formats. However, the review also points out a major challenge: the difficulty in integrating these models. Li acknowledges their importance but also notes the heavy computational demands and the need for specialized skills to coordinate these diverse algorithms into one system. From a research standpoint, Li's work faces what is called the Implementation-Explainability Gap. While it offers a solid guide for applying hybrid and deep learning models, it mainly addresses the technical ability to predict prices rather than explaining how those predictions are made. The study highlights that as models blend CNNs and LSTMs, they become harder for people to understand. This is why the focus of this thesis is not only to build the hybrid ML-DL architecture Li describes but also to add an Explainable AI (XAI) layer to make sure the model's results are both accurate and transparent for practical use **Li (2025)** [14].
- **Ezzeldin and Abouagwa (2025)** The study focused on comparing different methods for predicting real estate prices in Riyadh, Saudi Arabia. It tested three approaches: traditional Hedonic Regression, Artificial Neural Networks (ANN), and Random Forest (RF) models. The results showed that the Random Forest model performed better than both the regression and neural network methods, making it the most reliable for estimating property values in Riyadh's housing market. The researchers used a detailed local dataset that included information on property location, size, age, and distance from major urban services. The main strength of this study lies in its regional focus, offering evidence about which algorithms suit the economic and structural conditions specific to Riyadh. However, the research was somewhat limited in scope. Although Random Forest was the best among the three methods tested, the study did not examine more advanced techniques like XGBoost or models that incorporate visual data. From a research standpoint, the work by Ezzeldin and Abouagwa has a Structural Rigidity limitation. While they identified the top-performing model for tabular data, they overlooked important qualitative factors such as architectural design and interior luxury, which matter a lot in Saudi Arabia's high-end real estate market. This reveals a gap that calls for moving beyond purely numerical approaches. Building on their findings, this thesis intends to combine the strengths of ensemble models with visual analysis to better capture the full market value of properties in comparable regional markets **Ezzeldin and Abouagwa (2025)** [4].

- **Yazdani and Raissi (2023)** This study introduced a new way to value properties by using Self-Supervised Vision Transformers (ViT). Instead of relying on typical convolutional neural networks, this method applies a transformer-based design to extract detailed visual features from property images and combine them with structured metadata. The researchers showed that by letting the model learn visual patterns without human-labeled categories, it performed better than traditional Automated Valuation Models (AVMs) that only use numerical or tabular data. The research used a large multimodal dataset that included thousands of real estate listings, pairing high-resolution photos of properties with related structural and geographic information. The main strength of this approach lies in its computational complexity, which helps capture subtle global patterns in images—like how lighting, material quality, and architectural design relate to each other—that simpler models tend to overlook. On the downside, training and running these advanced transformer models require a lot of computing power and specialized hardware such as GPUs. From a research standpoint, the work by Yazdani and Raissi faces the challenge known as the Transformer Interpretability Gap. While Vision Transformers offer strong accuracy, they are often more difficult to interpret than standard deep learning models. This makes it hard for appraisers to understand why certain visual features influence the price. This points to an important research need: finding a balance between high visual accuracy and Explainable AI (XAI). Building on their multimodal approach, this thesis aims to add a transparency layer, so the visual insights found by the transformer can be clearly explained to users **Yazdani and Raissi (2023)** [16].

- **Gummer et al. (2025)** They introduced a two-stage clustering method to improve both the accuracy and clarity of automated house price predictions. First, properties are grouped based on shared geographic and location features. Then, within each group, localized regression models are applied. This divide and conquer approach helps the AI capture local market differences that a single global model might miss, resulting in more customized and transparent price estimates. The study used a housing dataset enriched with geospatial information, combining traditional structural details with neighborhood environment and urban accessibility data. The main benefit of this method is its clear logic: by dividing the market into understandable groups like urban high-density or suburban residential, the model offers an interpretability layer that stakeholders can easily grasp. But one downside is the Cluster Boundary Problem, where homes near the edges of zones might get inconsistent valuations depending on their assigned group. From a research angle, Gummer et al. (2025) [2] face a limitation called Visual Exclusion. While their model groups homes by location and numerical data well, it ignores the big visual and condition differences between properties in the same area. For example, a luxury home and a distressed one might be treated the same way in the regression, which can cause inaccuracies. This reveals a key gap: the need to include visual features in clustering. This thesis aims to address that gap by grouping homes not only by location but also by appearance.

- **Tariq et al. (2025)** Researchers developed an advanced multimodal fusion architecture to improve the accuracy of Automated Valuation Models (AVMs). This approach used a "late-fusion" strategy: three distinct neural network components—a Convolutional Neural Network (CNN) for image analysis, a Natural Language Processing (NLP) transformer for textual descriptions, and a Gradient Boosting Machine (GBM) for structured tabular data—were trained separately. Their outputs were then integrated in a final decision layer. This configuration enabled the model to process various data streams concurrently to generate a single price estimate. The dataset for this study consisted of a comprehensive collection of over 40,000 property listings, compiled from various metropolitan markets, and included multiple data types. Each listing entry connected physical characteristics, such as square footage and room count, with high-resolution interior and exterior images, alongside agent-written sentiment analysis. A primary benefit of this research is its empirical demonstration that multimodal models can indeed "perceive" property value in a manner analogous to human appraisal. They are capable of identifying "hidden value" in modern renovations and luxury finishes that purely numerical data would overlook. Nevertheless, a significant drawback arises from the substantial computational cost and memory demands required to synchronize these three distinct model architectures during the training phase. From a research perspective, Tariq et al. (2025) [17] encounter a limitation referred to as Fusion Opacity. Although the model effectively integrates disparate data types, it struggles to elucidate the cross-modal influence. This means it cannot specify whether a valuation was predominantly influenced by a particular photographic element or a specific structural feature. This situation highlights a critical deficiency: the necessity for "Cross-Modal Interpretability." This thesis intends to address this deficiency by developing a system that not only fuses these data types but also employs attention-mapping and integrated gradients to precisely clarify the contribution of each visual or numerical feature to the final valuation.

2.3 Summarization of Related Work.

Author	Method Used	Dataset	Advantages	Disadvantages	Restrictions / Research Gap
Mora-Garcia et al.2022	MLR, RF, GBM, ANN	Large-scale geospatial property attributes	Proved ML superiority in handling "big data."	High computational cost; technical complexity.	Visual Omission: Fails to incorporate unstructured imagery for aesthetic premiums.
Wang et al.2022	GNNWR (Hybrid ANN & GWR)	Spatially-enriched urban data (GPS)	Superior sensitivity to local "neighborhood effects."	Risk of overfitting to specific local blocks.	Sensory Narrowness: Treats property as coordinates only; no architectural quality assessment.
Yazdani & Raissi 2023	Self-Supervised Vision Transformers	Multimodal: Photos + Structured metadata	Captures intricate global patterns in images.	Extreme hardware requirements (High-end GPUs).	Interpretability Gap: Vision Transformers are "black boxes"; lacks explanation for visual features.
Rodriguez-Serrano 2024	Prototype-Based ML	Standardized residential (Transactional)	Human-centric reasoning using tangible benchmarks.	Accuracy depends heavily on prototype diversity.	Feature-Set Restriction: Not yet adapted to explain visual prototypes (e.g., modern vs. old style).
Balila & Shabri 2024	SVM, GBM, RF, ANN	Localized Dubai listings (Landmarks/Type)	High accuracy in luxury/high-value global hubs.	Sensitive to complex hyper-parameter tuning.	Visual Blindness: Cannot account for interior finishing quality, which is vital in luxury markets.
Hasan et al. 2024	Multimodal Deep Learning	Images + Text + Tabular data	Mimics human appraisers by using "curb appeal."	Massive complexity in data fusion and processing.	Explainability Gap: High accuracy but no XAI layer to explain why an image changed the price.
Sharma et al. 2024	XGBoost	Structured tabular portal data	Exceptional at capturing non-linear relationships.	Opacity; difficult for non-technical auditing.	Visual Exclusion: Cannot distinguish between a renovated home and a distressed one.
Riveros et al. 2024	XGBoost	Structured records (Physical/Location)	Robustness against overfitting in large datasets.	Limited inherent explanation for feature weighting.	Perceptual Gap: Lacks visual intelligence to process the qualitative state of a property.
Arshad et al. 2024	Cat Boost, LightGBM, SHAP	Comprehensive structured property records	Achieved global and local interpretability.	SHAP calculation is computationally expensive.	Perceptual Blind Spot: Can explain math/numbers but not the logic behind visual appeal.
Azam & Rai 2025	RF, GBM, ANN	Historical sales (Tabular)	Proved that data rigor reduces error rates.	Limited to homogeneous tabular data formats.	Data One-dimensionality: Prepared data still cannot account for visual/architectural appeal.
Chen 2025	RF, GBM, ANN	Longitudinal urban housing transactions	Strong predictive depth for market volatility.	Opacity regarding the "why" of valuation shifts.	Contextual Blindness: Unaware of physical depreciation or renovations not in spreadsheets.
Li 2025	Hybrid/Ensemble (CNN, LSTM, RF)	Multi-source global meta-dataset	Maps the entire technical ecosystem of AVMs.	Massive integration and synchronization complexity.	Implementation-XAI Gap: Focuses on building hybrid models, not on making them interpretable.
Ezzeldin & Abouagwa 2025	Hedonic, ANN, RF	Riyadh residential listings (Size/Age)	Regional empirical evidence for the Saudi market.	Did not explore advanced boosting or vision.	Structural Rigidity: Ignores qualitative visual factors influential in high-end Saudi real estate.
Gümmer et al. 2025	Two-stage Clustering + Regression	Geospatially enriched data (Environmental)	High logical clarity through market segmentation.	Cluster boundary errors for "edge" properties.	Visual Exclusion: Grouping by location ignores that houses in one area vary wildly in look.
§ Tariq et al. (2025)	combine CNN (images) + GBM (tabular)	Comprehensive Multimodal Dataset	Mimics human appraisal by "perceiving" aesthetic premiums and renovations that aren't in the tax records.	High Computational Cost: Running CNNs and Transformers simultaneously requires significant GPU power and memory.	The Cross-Modal Interpretability Gap—the need for XAI to show the influence of each specific data type on the result.[17]