



Project Proposal

Development of A Hybrid Predictive Model by Using Images and Digital Data for Real Estate Market Value Estimation with Artificial Intelligence

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

Real estate valuation is the process of assessing the value of real property, and it is playing a crucial role in international finance through determination of investment options, mortgage approvals, and proper risk assessments. Recently, the valuation approach had altered due to emergence of digital real estate platforms and the availability of vast property datasets. Automated Valuation Models (AVMs) currently are spread; and providing rapid and scalable valuation estimations rather than traditional valuation methods **(Balila & Shabri, 2024) [1]**.

Despite of automated valuation models advantages, but still have their limits; most of them rely on structured numerical data such as location, property area, property age, and rooms number. Although these details are significant, they don't gather all the important information for property valuation. 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. 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. 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.

Deep learning models have predictive capabilities, but somewhat mysteriously, meaning they do not always provide definitive answers that facilitate decision making for investors and regulators to rely on, Thus, explainability is a major challenge when it comes to huge investment decisions or legal issues. This lack of clarification can be a problem, that make a gap of information especially in sectors where users need to know why they choose this certain value **(Sharma et al., 2024) [3]**.

The aim of this study is to bridge the gaps in current machine learning methods. These gaps are represented in the inability to provide an understandable system and a clear vision that combines the full details of the property and its visual characteristics. Thus, the research proposes to create a reliable and clear predictive model that integrates digital and visual data together to improve the accuracy of predictions to help individuals and decision-makers understand the real estate market and make informed decisions. **(Ezzeldin & Abouagwa, 2025) [4]**.

1.1 Problem Definition

The process of property valuation and assessment is a sensitive and crucial factor in the global economy, as it influences investment decisions and the evaluation of potential financial risks. When investors, institutions, or even individuals make significant financial decisions regarding the purchase or sale of real estate, valuation issues can be extremely costly, as these transactions involve substantial sums. Thus, accurate and precise property valuation is essential to avoid enormous financial losses that could potentially impact and even alter the entire financial system. Although automated valuation models (AVMs), made possible by machine learning and digital advancements, have facilitated digital real estate valuation, these models face three main challenges that limit their accuracy and acceptance by professionals.

First, the qualitative data gap: Traditional automated valuation models rely on quantitative data, such as area, property location, property age, and number of rooms. While these factors provide good initial information, they lack many crucial details for investors and financial decision-makers. Their priority, in addition to quantitative details, is qualitative details such as construction quality, building style, interior finishes, and regular maintenance. These details directly impact property prices. Current models rely on this quantitative data, and therein lies the problem. (Balila & Shebri, 2024) [1].

Second, the challenge of media integration: While deep machine learning and convolutional neural networks (CNNs) offer capabilities for extracting and analyzing property features from images, many systems fail to integrate image analysis and digital data to provide a complete picture. Instead, they handle the two aspects separately, hindering sound and integrated decision-making. To illustrate, a model might evaluate a house of a certain number of rooms based solely on numerical data, disregarding the house's internal condition and maintenance status, which could impact its value. This is due to the lack of integration between numerical and visual binary data sources (Ezzedine and Abu-Aghwa, 2025) [4].

Third, the black box and lack of transparency are key issues in the use of advanced artificial intelligence in real estate. High-performance models, particularly deep neural networks, often operate as "black boxes," providing price estimates without explaining how those estimates were arrived at or what quantitative and visual considerations were used. This makes them difficult for investors and institutional decision-makers to trust (Sharma et al., 2024; Chen, 2025) [3].

Problem Summary: This thesis addresses a dilemma: the failure of current valuation models to integrate structured property data with visual information in a clear, unified, and interpretable manner. This gap leads to property valuations lacking comprehensive interpretation and overall accuracy, both of which are essential for making reliable financial real estate decisions. To address this problem, this research proposes a valuation model that integrates digital and visual data within a single, interpretable framework to generate accurate and reliable predictions.

1.2 Thesis Objectives

- This study aims to develop a new AI-powered model for estimating property prices. The model will integrate the structured numerical property details such as size and location with characteristics extracted from property images; to create a model that is easy to understand and accurate at the same time. to reach our goal, our study will focus on a few key objectives.
- Improving a structured valuation model; by constructing a predictive model using the structured numerical property details.
- Designing a model that uses deep learning to look at pictures and find out what the property's condition looks like.
- Create a two-part method that uses both numbers and visual image features, and then look at how this data combination affects the model's overall performance.
Explaining the model's predictions by integrating numerical and visual characteristics that directly influence real estate market valuations.
- Analyzing and evaluating the model's performance in case of data issues, such as loss, incompleteness, or damage.
- Using the model as a tool to help individuals, investors, and decision-makers in real estate institutions and companies make informed real estate financial decisions.

1.3 Thesis Contribution.

This research aims to make a significant contribution to the use of artificial intelligence (AI) in real estate appraisal and valuation by addressing some critical gaps in current tasks, steps, and practices. It examines both the theoretical and practical aspects of the subject, striving to make AI-assisted real estate appraisal, valuation, and estimation more accurate and reliable. The main contributions can be summarized as follows:-

1. A Unified Hybrid Valuation Approach.

This research proposes a new and more accurate method for prediction using both digital and visual data. Instead of focusing on each separately, this research considers that digital and visual data together form a cohesive framework for achieving and understanding a better valuation.

2. Emphasis on Explainability.

The new models differ from other systems that are difficult to interpret and understand, as they explain things in two ways: with digital data and visual data (images). This makes it easier for people to see what is happening and helps interested companies and real estate sectors to feel comfortable using it.

3. Empirical Evidence on Multimodal Integration.

This research and study examine the use of different methods for pricing property value. We will investigate and test three methods: the first relies solely on numerical data and all its details; the second relies solely on visual data (images); and the third, the most important method in this research, combines both numerical and visual data. Through this combination, the study will demonstrate how using both numerical and visual data together can contribute to understanding property value and pricing.

4. Robustness Analysis.

The model is priced and evaluated in the presence of many conditions that reflect real-world market environments, including the availability of incomplete and heterogeneous data.

5. Practical Orientation.

Practical orientation: The study's contribution to theoretical development. The framework was developed and designed taking into account the extent of the implementation process and contributes to and aims to achieve a real balance between prediction accuracy and interpretability, making it suitable for real-world financial, real estate and investment applications.

1.4 Thesis Organization

Chapter 1: Introduction

This chapter defines the research problem, objectives, and contribution of the research.

Chapter 2: Previous Research and Background

This chapter presents traditional property valuation and appraisal methods, as well as previous research and studies related to machine learning-based and image-based valuation models.

Chapter 3: Proposed Methodology

This chapter presents the proposed hybrid model and explains the data, methodology, and model structure.

Chapter 4: Experimental Results and Their Interpretation

This chapter presents the experimental results, compares the different models, and discusses the expected outcomes.

Chapter 5: Conclusion and Future Research

This chapter presents the main findings of the study and suggests directions for future research.

Chapter 2: Background and Related Works

2.1 Background

For a long time, traditional and outdated methods of property valuation have relied on the opinions and judgments of specialists and comparisons with recent sales in the same area or location. These methods form the basis of real estate appraisal, often depending heavily on the appraiser's accumulated experience and the availability of local data. This can lead to inconsistencies and discrepancies that affect valuations and make it difficult to scale up the process in rapidly changing real estate markets. The emergence of online real estate platforms and massive databases has helped address some of these problems by providing comprehensive information about property details, history, type, and location factors.

This led to the growth of classifications and collections of large real estate datasets, which in turn led to the development of the first Automated Valuation Models (AVMs) (Balila and Shebri, 2024) [1]. These early models employed statistical methods and techniques, such as hedonic pricing models and multiple linear regression, to assess how characteristics like size and location influence price. However, these methods often fall short in addressing the inherently complex and non-linear factors of different real estate markets. With the rapid advancements in machine learning, more sophisticated algorithms emerged, such as decision trees, random forests, vector support machines, and gradient enhancement (Arshad et al., 2024) [7]. These models handle large and complex datasets more effectively and can accurately detect both complex and simple interactions between variables. This can often provide a more accurate valuation of a property's market value compared to traditional or random statistics.

In addition to some digital data, property prices are influenced by other characteristics such as visual features, including interior and exterior design, lighting, maintenance, and other visual real estate attributes. These are among the most important factors that attract or interest buyers. Advances and updates in deep learning, most notably convolutional neural networks (CNNs), have revolutionized how visual property data is analyzed (Gomer et al., 2025) [2]. Convolutional neural networks are now capable of retrieving and capturing the most minute visual details of properties, such as the type and quality of finishes, interior design, and overall condition—details that were previously beyond the reach of automated tools. However, most current real estate valuation systems treat digital data, such as property age and orientation, and visual information, such as interior and exterior design, separately. This approach limits their ability to fully grasp market complexities, as property value depends on both measurable characteristics and visual appeal combined.

One of the major challenges facing complex machine learning models, including deep neural networks, is their lack of flexibility and transparency. They often function as "black boxes," providing accurate results and information but failing to explain the decision-making process (Sharma et al., 2024; Chen, 2025) [3]. In many real estate sectors, such as mortgage finance, this lack of clarity or ambiguity is a significant obstacle to building trust and the widespread adoption of these advanced real estate technologies and models. Stakeholders, such as investors, lenders, and regulators, require clear explanations of valuation methods to ensure fairness and accuracy. This need has driven the adoption of explainable artificial intelligence (XAI), which facilitates the clarity and understanding of model decisions. XAI also contributes to building flexibility and trust by enabling users, analysts, and others to see how digital and visual information impacts outputs, thus strengthening the justification for financial decisions.

Some current studies and research are moving towards hybrid, multimedia models that combine digital property data with visual components within a comprehensive system. These models contribute to simulating the complete appraisal and valuation process of a human appraiser or expert by integrating precise quantitative data with in-depth insights derived from images. This study will contribute to developing and proposing a system that not only aims to improve data accuracy and reliability but also promotes transparency and practicality for use across various real estate sectors.

2.2 Related Work

- These recent research studies have focused extensively, continuously, and increasingly on how to use machine learning and deep learning in analyzing, valuing, and predicting real estate prices. All researchers have contributed with different types of data, such as real estate details, digital information such as location information, property descriptions, and also visual data such as pictures of the property, in an attempt to make predictions more accurate.
- **Mora-Garcia et al. (2022)** A large-scale study was conducted to compare the performance of machine learning algorithms in predicting real estate property prices. The study was equally divided between experiments using classical statistical methods and more sophisticated computing techniques such as multiple linear regression (MLR), random forests (RF), gradient enhancement machines (GBM), and artificial neural networks (ANNs). The researchers found that when working with large datasets, modern machine learning methods outperform their older models in detecting more complex patterns, which were often overlooked by traditional practices. The study utilized a geospatially large dataset that included detailed property characteristics as well as precise location data. A key benefit of this research is that it provides empirical evidence that clustering and neural network techniques have the potential to provide more accurate predictions and more effectively process large datasets in the real estate field. However, these models require substantial computational resources and specialized expertise to be implemented correctly, making them less accessible compared to simpler regression tools. From a research perspective, the work of Mora-Garcia et al. (2022) ([8]) exemplifies a deficiency known as "visual omission." Although the study demonstrated the power of machine learning in processing spatial (grid, droplets) and digital (vector) data, it did not include unstructured visual data from real estate images. This omission is vitally important since digital data alone is insufficient to monitor the aesthetic value or the real-life situation of interior home decoration. This points to a fundamental gap in the literature: the need to go beyond purely digital models. In this thesis, we address this gap by combining the high-performance algorithms identified by Mura-Garcia with computer vision techniques to provide a more holistic and accurate evaluation.
- **Wang et al. (2022)** A specialized spatial modeling framework called Geographic Neural Network Weighted Regression (GNNWR) was introduced. This model combines the nonlinear learning capabilities of artificial neural networks (ANNs) with the spatial analysis characteristics of geographic weighted regression (GWR). By integrating these approaches, researchers sought to address the problem of spatial data instability in real estate markets, allowing the model to adjust its weights based on the geographic context of each property. The study used an urban housing dataset enriched with spatial information, linking conventional property characteristics to precise GPS coordinates and environmental variables specific to each neighborhood. The main strength of the GNNWR model lies in its sensitivity to spatial variations; it can detect the effects of unique neighborhoods and local market changes that are often overlooked by larger models, thus improving valuation accuracy in complex urban environments. However, one of the model's major drawbacks is its complexity, requiring vast amounts of

geographic data and fine-tuning to avoid over-focusing on a narrow area or over-allocating a small area such as a street or residential block. From a research perspective, the work of Wang et al. (2022) [9] suffers from what is known as sensory narrowness. While the GNNWR model is adept at understanding the impact of location on housing prices, it primarily treats properties as sets of coordinates and numerical data. It lacks computer vision to assess the actual condition of buildings or the quality of their architectural design. This highlights a significant gap: the need to integrate advanced spatial weighting techniques with visual feature extraction. This thesis aims to address this gap by ensuring that property value is influenced not only by its location on a map but also by its visual condition and physical appearance.

- **Rodriguez-Serrano (2024)** A domain-specific prototype-based learning model was developed to provide explainable predictions for real estate prices. Instead of all the data points, this approach selects a small number of prototypical real models from the dataset as the “anchors”. Concentrating on such concrete instances, the model offers a mode of thought that is closer to human thinking, enabling users to discover which comparable homes had the greatest impact on the estimated value of their own. The study employs a homogenized residential property data set, containing transaction records—rich with information of the property itself. The primary advantage of this model is its interpretability; it enables bridging advanced machine learning outcomes with human intuition by providing case-based explanations. This makes it easier for lenders and homeowners to review and understand the appraisal process. However, the quality and diversity of prototypes have significant influences on the accuracy of model. If the data set does not contain a good prototypical example of a unique or luxury home, the expected value may be less accurate. From a research perspective, Rodriguez-Serrano’s method faces limitations as it relies on a fixed set of features. Although prototype-based explanations are useful for structured data such as number of rooms or size, they have yet to be adapted to visual features. For example, the model can discern prototypes by size but not by modern interior or architectural style, due to the lack of a visual component for deep learning. This is indicative of a fundamental gap: that explanations based on prototypes need to be extended to multiple data types. This thesis work aims to fill this gap with prototypes that integrate digital information and visual components (Rodriguez-Serrano, 2024).[15]
- **Balila and Shabri (2024)** A comprehensive comparative analysis was performed to assess the effectiveness of different ML architectures in the unique real estate market in Dubai. The study considered several algorithms including support vector machines (SVM), gradient boosting machines (GBM), random forests (RF) and artificial neural networks (ANN). The results indicate that ensemble models, combining multiple algorithms, generally result in more accurate predictions, especially when applied to diverse high-value real estate slices that characterize a global city such as Dubai. A local dataset containing details of various transactions, such as type of property, age, location premiums, and proximity to landmarks like Burj Khalifa was employed in the study. Among the key contributions of this study are the regional evidence for the ability of ensemble methods to tackle "luxury home premiums" and rapidly changing market dynamics witnessed in the Middle East real

estate market. Yet, these models require fine-tuning of the hyperparameters, as an inappropriate tuning might lead to overfitting that would disallow their usage for real-time price estimation. From a research point of view, the Bali and Shepper (2024) [10] work is notably wanting, so much as to be called blind. While they did establish strong numerical indicators for the Dubai market, their analysis did not include such elements as the aesthetic quality of properties and the standards of finishing, which oft have an impact on the prices of luxury homes. In fact, in Dubai, the price of two apartments in the same building can vary greatly due to interior design or view, elements not considered by the traditional data sets. This gap indicates the significance of incorporating computer vision techniques to enhance these ensemble models. Based on their findings, this thesis seeks to integrate visual feature extraction to enhance consideration of the 'luxury factor' in real estate evaluation.

- **Hasan et al. (2024)** The study proposes a sophisticated multi-modal deep learning framework that aims to transform appraisal property by integrating three types of data: structured tabular data, high-resolution images, and textual descriptions from property listings. Using a Multi-branch neural network, the researchers demonstrated that the fusion of such heterogeneous data sources results in superior prediction accuracy and lower error rates when compared to models that rely solely on numerical data. The data set has been extracted from various real estate platforms that comprised thousands of listings, each with digital features like area and images of interiors and exteriors, as well as agents' notes. The strength of this process is the wholesale use of data, akin to how a professional real estate appraiser would consider the appearance of a property and its major selling points before making a determination. On the other hand, the integration of such a wide range of data types is a very challenging task. Training a model to balance the importance of an image with a line of text for a name requires significant computing resources and large, clean datasets. From a research perspective, Hassan, Chauhan, Jiang, and Idrees... (2024) [11] confront what they call an interpretative gap. However, although their work emphasizes that integrating more data sources increases accuracy, the model remains a black box, difficult for users to understand regarding how a particular image or phrase influences price predictions. This highlights a pressing need in this area for the development of multimedia interpretive artificial intelligence (XAI). Building on Hassan's multimedia model, this paper seeks to add transparency, enabling the system not only to process images and text but also to clarify how these visual and textual features affect the final user evaluation.
- **Sharma et al. (2024)** The study proposed a sophisticated deep multimodal learning framework, focused on enhancing real estate valuation by integrating three types of data, i.e., structured tabular data, high-resolution images, and textual description from property listings. By leveraging a multi-branch neural network, the authors demonstrate that the fusion of these heterogeneous data sources enhances the prediction accuracies and reduces the error rates over traditional models that rely solely on numeric data. The dataset was collected from multiple real estate platforms and consisted of thousands of listings. Each listing was associated with numerical features such as size, with internal and external photos, and descriptive notes from the agents. The principal advantage of this approach is the very large amount of data it can make use of, similar to how a professional appraiser

evaluates the look and feel of a property and its main points of sale before making his judgement. However, the combination of these different types of data does add complexity. Training the model to weight the importance of the image versus a line of text requires substantial computational power and a large, well-organized dataset. From a research perspective, Hassan et al. (2024) [11] highlight what they call the interpretability gap. While their results show that including additional data sources can improve accuracy, the model itself is still opaque. And that makes it difficult for users to understand why a certain image or phrase has an effect on price predictions. Such a gap points the more general challenge to the field: the need for multi-modal explainable AI (XAI). Based on Hassan's multimodal model and Hassan's approach, this thesis will focus on increasing transparency, by making the system not only capable of analyzing images and texts, but also by allowing it to explain how the features affected the final rating in a way that is understandable for the users.

- **Riveros et al. (2024)** The aim of this paper is to investigate the potential of using the XGBoost (Extreme Gradient Boosting) algorithm to improve the accuracy of property valuation estimates. This machine learning approach has been shown to capture complex, nonlinear relationships between various home characteristics, such as size and location, and their market price. By employing the gradient boosting framework, the study reveals multiple patterns overlooked by most traditional models, demonstrating that property characteristics do not rank in value linearly, even when St is a function of the variables. Organized residential real estate market data, containing thousands of past transactions and information on location and physical characteristics, was used. A key strength of this work is its computational efficiency; XGBoost effectively handles high-dimensional data while minimizing the likelihood of over-allocation, making it suitable for large-scale market forecasting. However, one of its main drawbacks is the model's ambiguity: it does not clearly describe how specific characteristics interact to produce a particular price. This lack of transparency can pose a problem for professional appraisal. From a research perspective, Riveros et al. (2024) [6] suffer from what they call a knowledge gap. Although the XGBoost algorithm provides accurate estimates based on numerical inputs, it cannot assess the visual or qualitative condition of a property. Meta example: If one house has been finished in a high-end style and the other is neglected, you cannot tell that the house in the lower photo is much smaller than the house in the upper photo. This gap points to a pressing need: to combine high-performance gradient enhancement methodologies with computer vision techniques. This thesis aims to address this need by combining the predictive capabilities of XGBoost with visually analyzing features for a better representation of the physical condition of a real estate..
- **Arshad et al. (2024)** This research opened the door of the interpretable artificial intelligence (XAI) in real estate via applying the Cat Boost and LightGBM framework to predict the housing price. To tackle the "black box" issue prevalent in such complicated models, study introduced SHAP (SHapley Additive exPlanations). This approach enabled a neat classification of feature importance and accurately indicated the influence of such features as location and area on the price estimation. The study utilized a well-structured and comprehensive real estate data set consisting of thousands of prior transactions and many tabular features. One of the key

strengths of this work is that it achieved global and local interpretability, providing a mathematical way to justify the valuations to stakeholders such as banks and investors. Cat Boost has also been shown to be very effective in handling categorical data. However, one of its conspicuous disadvantages is that the computation of SHAP values for large datasets is computationally expensive. The model also only operates on structured data and so qualitative factors such as the external appearance of the property, and its condition are ignored. From a research perspective, Irshad et al - [7] (2024) ...They encounter a cognitive blind spot because their model cannot explain the logic underlying visual attractiveness. It can tell how location affects value, but not how modern finishes, which you can see in photos, affect price. This gap indicates the urgent need for a multimodal framework for explainable AI (XAI) that can offer explanations for both numerical data and visual features at the same time. This thesis aims to address this gap by integrating the two types of data in a transparent system.

- **Azam and Rai (2025)** The study looked at how well different ML techniques e.g. random forest (RF), gradient boosting machines (GBM), artificial neural networks (ANN) performed in predicting house prices. It also pointed out that the success of prediction not only depends on the choice of the algorithm, but also depends on the accuracy of data processing and evaluation. By evaluating a few different models, the researchers arrived at the best set of parameters to assess property in today's market. They employed a huge real estate dataset that also contained several previous sales records in addition to different types of property. The main strength of this work lies in its meticulous methodology; It shows that a rigorous data cleaning, feature design, and systematic model testing are the keys to minimising error in Automated Valuation Models (AVMs). However, one of the limitations is this study only utilized the unified tabular data which cannot represent the complex and multimodal aspects often involved in property evaluation in the real world. From a research standpoint, the Azam and Rai (2025) [12] methodology suffers from what can be called 'data one-dimensionality'. Although they demonstrate the importance of data preparation for numerical accuracy, their model is not capable of processing unstructured visual information. This constitutes a fundamental gap as even the most refined numerical data do not reflect the visual state of a property or its architectural style. This suggests a critical need for research that applies the same level of rigor in pre-processing to a multi-modal system incorporating numerical data and images of properties and is the challenge this thesis aims to address.
- **Chen (2025)** The study addressed how the role of machine learning in housing price prediction has changed, confirming that data-driven methods are better at capturing complex market behaviors than traditional statistical methods. The paper compares the random-forest (RF), gradient-boosting machine (GBM), and artificial neural network (ANN) algorithms. Results from Shih Chen revealed that both of these algorithms excel at capturing non-linear patterns and hidden factors in real estate attributes, which many existing models fail to consider. The research employed a long-term dataset on urban housing sales, which included prior prices, building characteristics, and local economic data. One of the strengths of this work lies in its deep forecasting, which highlights the advantages of AI in modeling unpredictable real estate market trends. A drawback of this type of

model is that it can be difficult to interpret; it detects complex patterns but does not explain the reasons behind specific price changes, which poses a problem for policymakers. Additionally, the study only considered Structured data. From the perspective of analysis, Chen's work overlooks an important aspect, which is the physical condition of the properties. Although the data-driven approach can recognize the market trends, it excludes consideration of factors such as internal decay or modernization, which are not included in the structured data. This exposes a radical flaw: the need to combine data-based prediction with visual information. This thesis aims to address this shortcoming by incorporating the learning-based techniques of Chen with a computer vision component to develop a more comprehensive and accurate pricing model (Chen, 2025) [13].

- **Li (2025)** [14] The study presented a comprehensive and detailed examination on the growth of AI in the real state industry, demonstrating the transition from the conventional models to more advanced deep learning techniques. The study included a range of techniques such as random forests (RF), gradient boosting machines (GBM), artificial neural networks (ANN). A particular emphasis is laid on the explosive development of convolutional neural networks (CNN) for image analysis, recurrent neural networks (RNN/LSTM) for time series data of the financial markets. Li's investigation advocates hybrid machine learning models and ensemble deep learning models as a suitable approach to capture the complex nature of real estate valuation. The research utilized a global descriptive dataset that provides data from multiple international real estate platforms, containing structured transaction data, architectural images, and historical price trends. The main strength of this work is its comprehensive perspective, as it provides a full map of the technical environment for recent AVM techniques, and shows how different types of neural networks can handle different data formats. However, the report also reveals a key challenge: the difficulty of integrating such models. Li recognized their significance but he also pointed out their immense computational requirements and the need for specialized skills to orchestrate these diverse algorithms into a single system. From the perspective of research, Lee's work is subject to the so-called application-explanation gap. While it does offer substantial evidence for the applicability of hybrid and deep learning models, the focus is largely on "can we predict prices" rather than "how are these predictions made". The study emphasizes that the more CNNs and LSTMs are combined in a model, the more complex the models are to be interpreted. 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 research focused on comparing several methods for predicting house prices in Riyadh, Saudi Arabia. The study tested three approaches: traditional hedonic regression, pseudo-neural models, and random forest models. The results showed that the random forest model outperformed both hedonic regression and neural networks, making it the most suitable for estimating property value in the Riyadh housing market. The researchers relied on a comprehensive local dataset containing information on property location, size, age, and distance to major urban services. The strength of this study lies in its regional focus, which provides evidence for the

best algorithms applicable to Riyadh's economics and structure. Nevertheless, the scope of the study was somewhat limited. While the random forest model performed best among the three tested methods, the study did not cover advanced techniques such as XGBoost or models with tabular data. In terms of research, the work of Ezz El-Din and Abu Ajwa suffers from structural deficiencies. Although they identified the best-performing model for tabular data, they did not consider important qualitative factors such as architectural design and interior design luxury, which are crucial in the Saudi Arabian luxury real estate market. This creates a gap that necessitates moving beyond purely digital approaches. Based on their findings, this thesis aims to leverage the advantages of clustering models, both analog and image analysis, to explore the overall market value of real estate in similar regional markets (Ezzedine and Abu Ajwa, 2025) [4].

- **Yazdani and Raissi (2023)** This paper has introduced a novel real estate appraisal method with self-supervised vision transformer (ViT) models. Rather than relying on traditional convolutional neural networks, this approach applies a design based on transformer models to extract detailed visual features from real estate images and combines them with structured meta-data. The researchers demonstrated that, by enabling it to learn visual patterns without human labels, the model outperforms traditional automated valuation model (AVM) that relies solely on digitized or tabular-based data. The study employed a massive-scale multi-modal dataset consisting of thousands of real estate listings in which high-resolution images of the properties are paired with structured and location-relevant information about the property. The key strength of this approach is its computational complexity, which allows for the capture of fine global patterns in images - such as how lighting, material quality, and architectural design correlate - that finer models tend to ignore. As with drawbacks, the training and inference of such advanced transformer models require significant computational power and specialized hardware such as graphics processing units (GPUs). From a research perspective, work by Yazdani and Risi faces a challenge known as the interpretability gap of transformers. Whereas vision transformers provide high accuracy, they are often more difficult to interpret than standard deep learning models and that makes it difficult for raters to understand why some visual features have an impact on price. These points to a pressing research need: the development of an adequate visual fidelity level for explainable AI (XAI) And from their multi-modal approach, this thesis strives to introduce a layer of transparency in the senses that the visual insights provided by the transformer can be clearly explained to the end-users **Yazdani and Raissi (2023)** [16].

- **Gummer et al. (2025)** The researchers provided a two-stage binning approach to improve the accuracy and clarity of automated home price predictions. In the first stage, properties are ranked within similar geographic and locational groups. Then, within each group, local regression models are used. This divide and conquer approach based on spatial intelligence enables the AI to detect local market variations which a single global model might miss, resulting in more customized and explainable price estimates. The study employs a geospatial enriched housing data set which marries traditional structural details with neighborhood

environmental data and access to urban areas. The key advantage of this approach lies in its simple logic: by segmenting the market into understandable groupings such as high density urban or residential suburbs, the model provides an interpretive layer that stakeholders can readily understand. But the drawback of this association method is the cluster boundaries problem, where houses located near the borders of regions might be given different evaluations depending on the cluster it belongs to. From a research perspective, Gomer et al. Although their model does a good job of classifying homes by location and numeric data, it does not account for the wide differences in appearance and condition among homes in the same area. For example, a luxury and a dilapidated home might be treated similarly in a regression analysis, which could distort the results. What this demonstrates is a fundamental blind spot: the need to incorporate visual features into the clustering process. This thesis attempts to fill this gap by clustering houses not only by location but also by appearance.

- **Tariq et al. (2025)** The researchers developed an advanced architecture to integrate multi-modal data to enhance the accuracy of automated evaluation models. This approach employed a “late fusion” strategy: three separate NN components were trained - a CNN to process the images, an NLP transformer to process the text description, and a gradient boosting machine (GBM) for the structured tabular data. Then, their outputs were merged into a final decisions layer. This architecture allowed the model to simultaneously process heterogeneous data streams and produce a single price estimate. The dataset utilized in this study consisted of a comprehensive collection of over 40,000 real estate ads, gathered from various urban markets and featured different types of data. Every advert links physical feature such as size and number of rooms to HD interior and exterior images plus a sentiment analysis written by the agents. One of the key advantages of this research is that it is experimentally demonstrated that multi-modal models are indeed capable of ‘perceiving’ the value of a property in a human-like way. They can also detect “hidden value” in trendy renovations and high-end finishes that pure digital data may miss. However, a significant drawback that becomes apparent is the high computational cost and large memory requirements to synchronize these three different model architectures during training. From a research perspective, Tarek et al. (2025) [17] tackled the so-called “integration ambiguity” by presenting a restriction. Although the model integrates heterogeneous data types efficiently, it is challenging for the model to capture interactions among modalities So he cannot say, was the evaluation mainly swayed by a particular pictorial element or by a structural feature? This presents a fundamental inadequacy: the need for “inter-modality interpretability” In this thesis, we aim at addressing this limitation by proposing a framework not only for the integration of such data types, but also to exploit attention mapping and integrated gradients to visualize the contribution of each visual or numerical feature in the final evaluation.

- **2.3 Summarization of Related Work.**

Author	Study Name	Method Used	Dataset	Advantages	Disadvantages	Restrictions / Research Gap
Mora-Garcia et al.2022	Housing price prediction using machine learning: a comprehensive study	MLR, RF, GBM, ANN	Large-scale geospatial property attributes	Proved ML superiority in handling "big data."	High computational cost; technical complexity.	Visual omission: This is the result of not integrating visual data such as unclear and scattered images for additional aesthetic purposes.
Wang et al.2022	A hybrid deep learning approach for housing price prediction	GNNWR (Hybrid ANN & GWR)	Spatially-enriched urban data (GPS)	Superior sensitivity to local "neighborhood effects."	Risk of overfitting to specific local blocks.	Sensory narrowness: The property is dealt with only within the framework of coordinates; its architectural quality is not evaluated.
Yazdani & Raissi 2023	Real Estate Property Valuation using Self-Supervised Vision Transformers	Self-Supervised Vision Transformers	Multimodal: Photos + Structured metadata	Captures intricate global patterns in images.	Extreme hardware requirements (High-end GPUs).	Distraction gap: Vision transformers are "black boxes" and they provide no explanation for visual features.
Rodriguez-Serrano 2024	Prototype-based learning for real estate valuation: a machine learning model that explains prices	Prototype-Based ML	Standardized residential (Transactional)	Human-centric reasoning using tangible benchmarks.	Accuracy depends heavily on prototype diversity.	Limitations of the feature set: not yet adapted to explaining visual models (e.g., modern style vs. old style)
Balila & Shabri 2024	Balila & Shabri 2024 real estate valuation study	SVM, GBM, RF, ANN	Localized Dubai listings (Landmarks/Type)	High accuracy in luxury/high-value global hubs.	Sensitive to complex hyper-parameter tuning.	Blindness prevails: The quality of interior finishes cannot be taken into account, which is crucial in the luxury product markets.
Hasan et al. 2024	A Multi-Modal Deep Learning Based Approach for House Price Prediction	Multimodal Deep Learning	Images + Text + Tabular data	Mimics human appraisers by using "curb appeal."	Massive complexity in data fusion and processing.	Interpretation gap: high accuracy but there is no XAI layer to explain why the image was changed to price.
Sharma et al. 2024	An Optimal House Price Prediction Algorithm: XGBoost	XGBoost	Structured tabular portal data	Exceptional at capturing non-linear relationships.	Opacity; difficult for non-technical auditing.	Visual exclusion: There is no comparison between a modernized house and a dilapidated one.
Riveros et al. 2024	Scalable Property Valuation Models via Graph-based Deep Learning	XGBoost	Structured records (Physical/Location)	Robustness against overfitting in large datasets.	Limited inherent explanation for feature weighting.	The cognitive gap: He lacks the visual intelligence to process the qualitative state of the attribute.
Arshad et al. 2024	From Predictive Importance to Causality: Which Machine Learning Model Reflects Reality	Cat Boost, LightGBM, SHAP	Comprehensive structured property records	Achieved global and local interpretability.	SHAP calculation is computationally expensive.	Cognitive blindness: math/numbers can be explained but not the logic behind the visual gravity.
Azam & Rai 2025	Predictive Analytics for Housing Market Trends and Valuation	RF, GBM, ANN	Historical sales (Tabular)	Proved that data rigor reduces error rates.	Limited to homogeneous tabular data formats.	One-dimensional analysis of the data: The processed data is still unable to take visual/architectural appeal into account.

Author	Study Name	Method Used	Dataset	Advantages	Disadvantages	
Chen 2025	Multimodal Air-Quality Prediction: A Multimodal Feature Fusion Network Based on Shared-Specific Modal Feature Decoupling	RF, GBM, ANN	Longitudinal urban housing transactions	Strong predictive depth for market volatility.	Opacity regarding the "why" of valuation shifts.	Contextual blindness: Not realizing the physical consumption or the renewals that are not included in schedules.
Li 2025	"Estimating the market value of real estate using a hybrid model that integrates service and structural characteristics with visual analysis in light of interpretable artificial intelligence."	Hybrid/Ensemble (CNN, LSTM, RF)	Multi-source global meta-dataset	Maps the entire technical ecosystem of AVMs.	Massive integration and synchronization complexity.	Executorial Gap and XAI: They Focus on Constructing Hybrid Models, Rather Than Making Them Interpretable.
Ezzeldin & Abouagwa 2025	Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal	Hedonic, ANN, RF	Riyadh residential listings (Size/Age)	Regional empirical evidence for the Saudi market.	Did not explore advanced boosting or vision.	Robustness: The qualitative visual factors influencing the high-end Saudi real estate were ignored.
Gümmer et al. 2025	Unveiling Location-Specific Price Drivers: A Two-Stage Cluster Analysis for Interpretable House Price Predictions	Two-stage Clustering + Regression	Geospatially enriched data (Environmental)	High logical clarity through market segmentation.	Cluster boundary errors for "edge" properties.	Visual coverage: Spatial grouping overlooks the fact that houses in the same area vary greatly in appearance.
Tariq et al. (2025)	"A Multi-modal Explainable AI Framework for Property Valuation: Integrating Structural Data and Visual Features"	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.	Multimedia explainability gap – the need for XAI to show how each specific type of data influences the outcome.[17]