



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 valuation and assessment of assets is the foundation of the entire world economy and plays a key role in directing investments and defining risks. In other words, having the right valuation for any type of asset—whether commercial or residential—is extremely important when dealing with large figures, either making money or losing it all. In essence, proper valuation helps to avoid financial catastrophes in the system.

Despite technological advancements and innovations that facilitate real estate valuation and enable more advanced automatic valuation models (AVMs), which incorporate artificial intelligence and machine learning algorithms, there remain three main challenges standing in the way of improving the accuracy of the process and gaining professional confidence.

First, it is important to consider the lack of qualitative data that could significantly improve the results. Most AVMs utilize quantitative variables to assess the property, including its area, location, construction age, and number of rooms. Such metrics are essential, but they do not account for all factors considered by potential investors and creditors. (Balila & Shebri, 2024) [1].

The second challenge would be combining different types of media. With the help of advanced machine learning and deep learning neural networks, we can crawl through the picture to identify features representing property attributes. However, since most tools separate processing digital media and analyzing pictures, they do not provide a comprehensive analysis, instead dividing the process into two parts. Thus, in the case of an evaluation of the number of rooms within the house in question, the model may rely purely on numerical indicators, neglecting such factors as the actual condition of the building or its state of repair. (Ezzedine and Abu-Aghwa, 2025) [4].

Third, the issues of black box characteristics and opacity become important for practical applications of sophisticated AI solutions in real estate. The high-performing models, especially those that apply deep learning techniques, have a tendency to act like black-box systems. This means that these systems provide estimates for pricing, but they do not explain how the estimation was made or what variables affected it, either numerically or visually. (Sharma et al., 2024; Chen, 2025) [3].

problem summarization: Valuation models tend to lack an encompassing structure in linking property's quantitative aspects with its qualitative features. The absence of integration between these two aspects makes it difficult to interpret valuation figures as well as increases their inaccuracies. The present study seeks to bridge this gap by introducing a valuation model that incorporates both types of variables into the forecasting process.

1.2 Thesis Objectives

The objective of this study is to develop a machine learning model that will predict the prices of real estate by integrating information about the structure and information inferred from visual data. The successful achievement of this objective will be attained through prioritizing certain goals:-

1. Improving a structured valuation model; by constructing a predictive model using the structured numerical property details.
2. Designing a model that uses deep learning to look at pictures and find out what the property's condition looks like.
3. Develop a dual strategy involving numbers and pictures, and evaluate the impact of this integration on your model's effectiveness
4. Explaining the process through which the model derives predictions, incorporating both numerical data as well as visual data that influence the value of properties in the real estate market.
5. Assess the model's effectiveness on linear or graphed data.
6. Employing the application of AI technology to help humans, investors, and business leaders within real estate companies in making sound decisions regarding property investments.
7. "In carrying out this project, I will generate real-life situations using the amazing data and noise, and after that, show how badly the performance of the model fails, and also show that it cannot be proven to be unacceptable."

1.3 Thesis Contribution.

With the above in mind, the current research attempts to tackle some of the pertinent challenges that exist within the present-day real estate valuation and appraisal exercise by incorporating artificial intelligence (AI) and concentrating on AI specifically in the exercise of the two aspects i.e., theoretical and practical to ensure the successful incorporation of AI within the realm of real estate valuation, appraisal and. The objectives of our research within the current study may be summarized as the following: -

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 procedures outline the fundamental concepts in real estate evaluation, which are always significantly reliant on the skills of the real estate evaluator and the presence of regional data. This normally results in disparities in valuation, making it challenging to achieve economies of scale in the ever-changing world of real estate. The advent of online real estate portals, big data, and computerized databases has made significant strides toward addressing these issues since they have made available all the relevant data regarding the property's characteristics and history being able to tell the type and in what kind of location it is. 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].

Unlike the initial ones, the new model utilizes procedures that were developed in data science in the 1970s. But these approaches seem to be less adept in dealing with the intrinsic nonlinear, multidimensional nature of spread real estate markets. As machine learning technologies started developing fast. Meanwhile, these methods allow working with big, highly complex datasets and identifying complicated relations of high-degree, as well as relatively simple relationships, among the variables involved. Thus, this approach provides a significantly more detailed estimate of the fair market value of a property than any random traditional statistics.

Some relevant characteristics, which influence the property value and, consequently, its price include various other metrics besides i.e., visual aspects of the property such as (Ex: lighting, flooring etc.) might not be included in the house inventory prior to missing exposure to real estate. Among all of these characteristics, perhaps the most important are visual aspects of the property, which attract or engage the buyers. Deep learning research, especially, developments in the field of convolutional neural networks (CNNs), has dramatically changed the way visual property information has been analyzed, with consequences (n; Gomer et al., 2025) [2]. Specifically, recent developments in CNN technology have enabled analyzing the multi-scale visual information of a property such as its finishes, colors, conditions, etc., in ways that were impossible for previous automated systems. On the other hand, most of the existing systems of real estate valuation analyze the digitalized information about the property (age, location, transport accessibility) as a separate category from the visual information viz, interior and exterior design.

Interpretability issues the primary challenge in employing sophisticated machine learning algorithms like deep neural networks is the requirement for adaptability, which often results in the models being termed as a “black box,” wherein the outcomes and data produced by them are accurate yet cannot be justified with regard to their decision-making process (Sharma et al., 2024; Chen, 2025) [3]. The absence of transparency and clarity in numerous instances when applying the latest technology and advanced models in the real estate sector, specifically within the mortgage financing industry, is a significant constraint in building confidence and making widespread adoption possible. Investors, lenders, and regulators demand detailed explanations regarding the approaches used for property valuation to ensure impartiality and correctness (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). It turned out that current machine learning techniques have an increased efficiency in the detection of such complex patterns that were hardly detectable using traditional methods applied to large data sets. The use of geospatial data for research purposes in the form of a rich set of property characteristics and their exact coordinates was another strong point. The biggest benefit of the experiment is that it proved empirically the possibility of better performance of clustering and neural networks-based models in terms of predictive accuracy as well as ability to process large data sets efficiently in the real estate sector. In turn, application of these models needs special computational capabilities and knowledge, which makes them less accessible than, for instance, regression. From the research perspective, "the visual omission" in machine learning was illustrated by the experiment done by Mora-Garcia et al. (2022) ([8]). The use of clustering and neural networks to process spatial (grid, droplets) and digital (vector) data proved effective in this particular case; however, unstructured visual information from real estate photographs was missing.
- **Wang et al. (2022)** They adopted a spatial modeling technique known as Geographic Neural Network Weighted Regression, GNN-WR. Integrating Neural Networks, a technique for spatial patterns and local effects and a Geographic Weighted Regression possibility. In other words, to solve the issue of real estate market data being stable, the two techniques were combined to have it fit in with the geographical area of the property, allowing the weights to be assigned accordingly.

The urban housing database used in this research consisted of spatial data including not only the usual property markers such as beds count, neighborhood characteristics, but, most importantly, each locality's geographic coordinates and environmental characteristics.

One would argue that the main feature of the GNN model is its spatial variability and ability to detect unique characters of various neighborhoods and changes occurring in local markets while bigger models often miss those. On the other hand, it can be viewed as highly complex due to high spatial variability of the model as criticized by Wang et al., (2022) (9). GNN-WR may be called the victim of what is known as a ubiquity blind spot. Even though it does excel at considering locations when evaluating prices of the property, it fails to take into account that properties cannot be seen just as a set of points on the map or numbers in the database. In other words, this model lacks computer vision to analyze and detect characteristics of the building itself and architectural features of a house with our thesis, we want to bridge that gap by combining complex spatial weighting with visual features of properties.

- **Rodriguez-Serrano (2024)** A prototype-based learning model in a domain-specific context was developed in order to explain the prediction of housing prices. Unlike considering all data points, this approach singled out only a few prototypes of real models among all the points in the data set as "anchors" (Amos et al., 2018). In the view of such concrete examples, the proposed model according to Concretely models represents a way of thinking close to human, involving an assumption of houses similar to the given one having the most influence on its value. The research includes the rationalization of a data set represented by the property record—the property The main advantage of the introduced model in terms of interpretation lies in the possibility to provide explanations of complicated machine learning findings using understandable cases.

Thus, the process becomes much easier for lenders and property owners. Nevertheless, the quality and variety of prototypes considerably affect the accuracy of a model for example, if there is no good prototypical sample among the properties representing a unique/luxury house, the expected value might be not accurate enough. 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. This research paper considered several algorithms such as support vector machine (SVM), gradient boosting machine (GBM), random forest (RF), and artificial neural network (ANN). The results indicated that the combination of different algorithms within an ensemble model tends to improve the performance of the model for high-value real estate segments in the international city of Dubai. The data used for this analysis was local data on property sales and included the type of property, age, premium, and distance to landmark buildings like the Burj Khalifa tower. Some of the notable findings from this research were the application of ensemble methods to solve the issue of luxury homes' premiums and changes in the market dynamics of Middle Eastern real estate markets. 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 research point of view, the Bali and Shipper (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)** Thus, the proposed study involves a highly advanced deep learning approach to transforming appraisal property by taking into account the following data sources: structured tabular data, images, and text description of property listing. In addition, using a multi-branch neural network, the authors proved that the combination of such different data sources led to improved accuracy and decreased errors in price prediction, compared with the models utilizing only numeric data. The database is gathered based on data extracted from various real estate platforms comprising thousands of real estates where every listing has numeric data related to the area, and photos both of interior and exterior parts of houses, as well as agents' notes about them. The benefit of such an approach lies in the usage of all possible data, similarly to the approach chosen by a professional appraiser who evaluates a house based on its looks and main selling points. The downside of this process is that it is a highly complex and resource-intensive process. To train the model properly by weighting such elements as the importance of the image and a sentence of text for the name would require considerable computing powers and large amounts of data. As for the research problems identified by Hassan, Chauhan, Jiang, and Idrees... (2024), they face the interpretational gap problem. However, while focusing on the idea of increasing the number of data sources for achieving increased accuracy, this model remains a black box with a

high level of opacity. Therefore, one of the major problems to solve in this area is developing interpretive artificial intelligence in multimedia systems. Thus, continuing from the ideas introduced by Hassan in his multimedia model, the current paper aims at adding transparency to the model.

- **Sharma et al. (2024)** The research presented a complex multimodal machine learning architecture designed specifically for improving real estate price prediction using structured tabular data, images, and natural language descriptions of listings. According to Hassan et al. (2024) [11], the proposed algorithm based on a multi-branch neural network proved the positive impact of using multiple sources of heterogeneous data and outperformed conventional approaches relying exclusively on numeric variables. The dataset included thousands of listings taken from various real estate sites and comprised numerical attributes such as area and photographs (internal and external) with descriptions provided by real estate agents. The most obvious benefit is that the proposed solution allows one to process enormous amounts of data – similarly to a professional appraiser who evaluates visual properties and highlights critical selling factors before providing his or her assessment. At the same time, the use of multiple information sources also increases the complexity of building the model, as training the algorithm to consider the relevance of the image compared to the textual attribute requires significant computational resources and a vast dataset. Hassan et al. (2024) [11] point to the interpretability gap inherent in the results obtained during the experiment conducted. Even though adding images and textual descriptions improved the accuracy of real estate price prediction, the lack of interpretability remains an essential issue since users cannot comprehend why a specific picture or phrase impacts the outcome. Therefore, the current study proves the necessity for multi-modal XAI solutions, which is the primary direction for the work proposed in this thesis. Specifically, based on Hassan's multimodal model and Hassan's methodology, this project aims to increase the transparency of the solution, allowing it to analyze visual and textual features of the data while explaining the impact of individual factors on the outcome rating to the user.
- **Riveros et al. (2024)** The objective of this article is to analyze the feasibility of applying XGBoost algorithm for the purpose of improving precision in calculating property value. As is known, this approach enables capturing the complex non-linear relations between many factors characterizing a house and its price in the market. Using gradient boosting techniques, the article demonstrates that there are several trends missed by almost all existing approaches, revealing that characteristics of a property do not rank linearly with respect to each other, even when (St) is expressed through them. The organized residential real estate market dataset, comprising of thousands of previous transactions along with data on location and physical properties, was utilized. The main advantage of this approach lies in its computational efficiency, enabling it to process high-dimensional input data and reducing the probability of over-allocation. Consequently, XGBoost algorithms are

well-suited for use in market forecasting. On the contrary, the main disadvantage of the technique in question is the opacity of the model, as it fails to reveal how certain factors interact with each other to yield an exact price. This drawback poses a serious obstacle to conducting professional appraisal. With regard to scientific studies, Riveros et al. (2024) [6] suffer from a phenomenon called a knowledge gap. In particular, although the algorithm in question provides accurate price prediction based on numerical parameters, it lacks the capacity to evaluate the visual and qualitative state of a property. Meta example: While one house has been renovated in a luxury manner and the other has remained unattended, you will not be able to ascertain that the house in the bottom picture is significantly smaller than the one on top. Thus, there arises an urgent necessity to merge efficient gradient enhancement techniques with computer vision methods. This is precisely the task that is planned to be solved in the present thesis.

- **Arshad et al. (2024)** This research gave birth to the door of the interpretable artificial intelligence (XAI) in real estate through the use of the Cat Boost and LightGBM framework in predicting the housing price. As far as resolving the "black box" phenomenon, which is common in complex models, SHAP (SHapley Additive exPlanations) was introduced. Such a method facilitated the neat categorization of the feature importance and clearly stated the impact of the mentioned feature, including the location and area, on the price estimate. For the study, an organized and detailed real estate dataset with thousands of historical data entries and numerous tabular features was used. One of the major advantages of the research is that it provided global and local explainability, allowing mathematical justifications of the evaluations for the interested stakeholders like banks and investors. Another advantage of Cat Boost is that it works extremely efficiently with categorical data. On the other hand, one of the most evident disadvantages of the research is the expensive computation of SHAP values of big data. Also, the algorithm operates exclusively with structured data and therefore completely ignores any qualitative feature, such as the external look of the property, and its condition. From the standpoint of research, Irshad et al - [7] (2024) There is a limitation related to the perception of cognition, as the model does not account for the rationality behind the aesthetic appeal of the property. The model is able to determine how the location impacts the value, but not how the modern finishing material, visible in the images, impacts the pricing of the house. This reflects the critical need for a multimodal XAI explanation method capable of providing rationale for both numeric and visual attributes simultaneously. This study seeks to fill that void.
- **Azam and Rai (2025)** The study investigated the efficiency of various machine learning algorithms such as Random Forest (RF), Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN) in estimating house prices. At the same time, the study highlighted the role of both the algorithm used and accuracy of data preprocessing and validation in the process. By comparing the results from several different models, the authors

managed to derive the best configuration of parameters for the evaluation of property in today's market. They have worked with an enormous real estate dataset that included information about multiple past sales cases, as well as varied kinds of property. What makes this paper so valuable is its rigorous approach to methodology. Specifically, it shows that a proper data cleaning procedure, adequate feature design and model selection are essential steps towards obtaining high accuracy of Automated Valuation Models (AVMs). One of the limitations of this paper is that it used a purely tabular type of data that cannot capture all aspects of real property evaluation. Methodologically speaking, there is a problem of so-called "data one-dimensionality" in the work of Azam and Rai (2025) [12]. While it highlights the importance of correct data preprocessing for numeric accuracy, the presented model does not allow working with unstructured data such as pictures of the property. Such a drawback is critical since no matter how accurate the data might be, it cannot reflect the condition of the object visually. Therefore, further research should involve the development of a multi-modal system of data preprocessing combining numeric data and property images.

- **Chen (2025)** The study identified the changing role of machine learning in the domain of housing price prediction, proving that data-based techniques have proven to outperform conventional statistics in predicting complex dynamics of the market. In this work, the author used two machine-learning algorithms (random forest [RF] and gradient boosting machine [GBM]) and compared their results with artificial neural networks (ANN). As was found by Shih Chen, both machine learning algorithms were able to detect non-linearity and hidden variables that impact the attributes of real estate. This work used a long historical dataset including past values, features of buildings themselves, as well as local economy. An advantage of this study is its deep forecasting that demonstrates how data-driven algorithms outperform conventional prediction in capturing the unexpected dynamics of the real estate market. However, one of the weaknesses of this model is difficulty in interpreting results since even though the model is able to find complex patterns, it cannot explain particular price variations, which creates problems for policy makers. In addition, the authors have only examined Structured Data. In terms of analyzing the research, there is a lack of taking into consideration of one of the main issues in the field, namely, the physical state of property. While the algorithm is capable of recognizing the market dynamics, it ignores other variables, such as whether there has been any decay or upgrading of houses, because these data are not included in the structured data. This creates a radical weakness in the algorithm that requires the incorporation of data analysis together with the usage of computer vision techniques. This is what I am going to examine in my thesis. (Chen, 2025) [13].
- **Li (2025)** [14] In this way, the study performed a comprehensive analysis of the growth of AI in the real estate industry through transition from classical approaches to more sophisticated deep learning methods. Such techniques were mentioned in the paper as random forest (RF), gradient boosting machine (GBM), and artificial

neural network (ANN). Special attention was paid to the fast development of Convolutional Neural Networks for the processing of the images and Recurrent Neural Networks, in particular LSTM for time-series financial market data analysis. As a result, Li proposed hybrid and ensemble approaches with the usage of both classical machine learning and deep learning as a potential technique for real estate valuation analysis. In terms of datasets used, the research relied on a global descriptive one which contains transactions data, architectural images, and information on the historical changes of prices in multiple international real estate platforms. The key benefit of the paper lies in its overall scope because it creates an informative picture of current technical situation in the field of AVM, and demonstrates the ability of each type of NN to process different data formats. However, the report also highlights a potential limitation. Even though the author acknowledged the importance of the studied technologies, he still mentioned the significant computational power required by them, and special skills needed to properly integrate them. On the research side, there is a risk of so-called application-explanation gap for Lee's study. Despite all the arguments supporting the possibility of usage of hybrid and ensemble techniques, the focus of the research is primarily on the question "can we predict prices" rather than "how are predictions made". Moreover, the study states that as more CNNs and LSTMs are included in the model, it becomes less understandable. That is why, the primary goal of this thesis is not only to create a hybrid ML-DL architecture suggested by Li, but also to implement an Explainable AI approach to provide additional transparency of the resulting model **Li (2025)** [14].

- **Ezzeldin and Abouagwa (2025)** As an object of study, this research compared several models for predicting the prices of houses in Riyadh, Saudi Arabia. In total, three different models were investigated, including hedonic regression model, pseudo-neural network, and random forest. It was found that a random forest is more effective than both hedonic regression and neural networks, therefore, it may be used in the calculation of prices for real estate in the Riyadh market. Researchers used an extensive set of data describing the characteristics of properties such as their location, size, age, and proximity to key facilities of urban life. The value of the research is that it has a regional focus and gives evidence of the best algorithms for calculating prices in Riyadh based on its economics and structure. However, some limitations were identified in relation to the investigation. Despite the fact that the most successful model out of three considered was the model of a random forest, the authors failed to analyze other algorithms like XGBoost and others related to tabular data. At the same time, the work of Ezz El-Din and Abu Ajwa in relation to research has structural problems. They have revealed the best performing model for tabular data but ignored essential quality parameters such as architectural and luxury interior designs in Saudi Arabia 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. Like the disadvantages, the training and inference processes of such state-of-the-art transformer models demand considerable computation power and specialized hardware like GPUs. As far as the literature is concerned, research by Yazdani and Risi is confronted with what is called the interpretability gap problem of transformers. While vision transformers offer high accuracy levels, they tend to be harder to interpret than traditional deep learning techniques; hence it becomes hard for human raters to understand why certain visual attributes affect prices. This leads to a critical research requirement; that is, the establishment of an adequate level of visual fidelity for explainable AI (XAI). And through their multimodal framework, this study aims to add an element of transparency where the visual insights obtained from the transformer become comprehensible to end-users **Yazdani and Raissi (2023)** [16].
- **Gummer et al. (2025)** The authors proposed a method that involves two-step binning of property prices in order to increase the accuracy and explanation ability of predictions made by automated home valuation models. At the initial stage, properties should be ranked in similar geographic and locational clusters. Later, on each cluster local regression models should be applied. With such "divide and conquer" approach, taking advantage of spatial intelligence, the machine will be able to discover local market anomalies, which will make the results more customized and interpretable. In their study, the authors use geospatial enhanced housing data set combining traditional information about structures' characteristics with information regarding environmental and urban amenities in nearby neighborhoods. And while the strength of this approach in interpretation of results is evident from the simple logic that follows the segmentation of the market into understandable clusters such as high density urban and residential suburbs, the main weakness of the association approach comes from the cluster boundaries dilemma when the same houses located at the border of regions may receive different ratings. From the scientific standpoint, the authors have done well in creating a model that segments the market according to geographic and locational factors and numeric data; however,

this model fails to take into consideration the variety of the appearance and conditions of houses in the same location. For instance, a fancy house and a rundown one would be considered equally in the regression analysis. Thus, the importance of the visual element of a property was revealed.

- **Tariq et al. (2025)** The authors designed an innovative architecture for fusing multi-modal information to improve the performance of automated models in terms of accuracy. Specifically, they applied a “late fusion” technique, where each NN component was individually trained on different types of information - a CNN for processing images, an NLP transformer for the text description, and a GBM for the structured tabular data. Afterward, the results were aggregated into one decision layer. Thus, the proposed architecture permitted analyzing multiple streams of heterogeneous information and generating a unified price estimate. In this study, the used dataset comprised a set of over 40,000 real estate listings containing various types of data collected from diverse urban areas. Each advertisement correlates physical characteristics (such as the square footage and number of rooms) with high-resolution pictures and agent’s sentiment analysis. An essential strength of this study is that multi-modal models have been proven to be able to ‘perceive’ the value of real estate in a human-like manner. 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.

Author	Study Name	Method Used	Dataset	Advantages	Disadvantages	
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.
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.	Executional 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.