Boosting House Price Estimations with Multi-Head Gated Attention

Zakaria Abdellah Sellam[1*], Cosimo Distante[1], Abdelmalik Taleb-Ahmed[2], Pier Luigi Mazzeo[1]

[1*]Institute of Applied Sciences and Intelligent Systems "Eduardo Caianiello", CNR, , Lecce,Italy.
[2]Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France, Universite Polytechnique Hauts de France, F-59313, Valencienne, France.

*Corresponding author(s). E-mail(s): abdellah.sellam@isasi.cnr.it;
Contributing authors: cosimo.distante@cnr.it; Abdelmalik.Taleb-Ahmed@uphf.fr;
pierluigi.mazzeo@cnr.it;

Abstract

Evaluating house prices is crucial for various stakeholders, including homeowners, investors, and policymakers. However, traditional spatial interpolation methods have limitations in capturing the complex spatial relationships that affect property values. To address these challenges, we have developed a new method called Multi-Head Gated Attention for spatial interpolation. Our approach builds upon attention-based interpolation models and incorporates multiple attention heads and gating mechanisms to capture spatial dependencies and contextual information better. Importantly, our model produces embeddings that reduce the dimensionality of the data, enabling simpler models like linear regression to outperform complex ensembling models. We conducted extensive experiments to compare our model with baseline methods and the original attention-based interpolation model. The results show a significant improvement in the accuracy of house price predictions, validating the effectiveness of our approach. This research advances the field of spatial interpolation and provides a robust tool for more precise house price evaluation. Our GitHub repository.*contains the data and code for all datasets, which are available for researchers and practitioners interested in replicating or building upon our work.

Keywords: House price evaluation, gated Attention, spatial interpolation, spatial analysis

1 Introduction

The Real Estate sector plays a pivotal role in the global economy, with house prices significantly influencing individual wealth and broader economic trends. Fluctuations in house prices can stimulate consumption and boost the economy when prices rise. At the same time, a decrease can limit an individual's borrowing capacity, potentially crowding out investments due to the evaporation in the value of collaterals [1]. The shock in the global economy caused by the 2008 housing bubble perfectly illustrates the importance of a stable and measurable house price [2]. Predicting house prices, however, is a complex task due to the multitude of influencing factors. Historically, house price prediction has relied on traditional regression models that consider a range of property-specific factors such as size, age, condition, and number of rooms, among others [3]. However, with the advent of machine

---

*Final_file/tree/main/ASI-main

learning, the landscape of house price prediction has significantly evolved. Techniques such as support vector machines, decision trees, and neural networks have been employed to improve prediction accuracy [4]. In addition, ensemble learning methods, such as boosting, have been used to enhance the performance of prediction models. Specifically, XGBoost [5], a scalable and accurate implementation of gradient boosting machines, has been applied to house price prediction with promising results [6, 7]. Furthermore, to account for spatial heterogeneity in house prices, Geographically Weighted Regression (GWR) and its variants have been utilised [8–11]. Kriging[1] [12], a geostatistical method, has also been used for spatial interpolation in house price prediction [13, 14]. However, these conventional methodologies bear certain limitations. For instance, they might struggle to capture complex spatial relationships, particularly in regions with diverse and distinct geographical realms. Additionally, assumptions like isotropic variability, which presupposes a constant spatial relationship in all directions, may impede the accuracy of these traditional models in anisotropic landscapes. Moreover, these models might exhibit sensitivity to outliers and could become computationally demanding, especially with increased data points. Our model builds upon the research of Vianna and Barbosa [15], who developed the attention-based spatial interpolation model. Our research endeavours to extend the paradigm by intertwining Multi-head and Gated Attention mechanisms. Vianna and Barbosa's model manifested a breakthrough by employing an attention mechanism to weigh the influence of neighbouring houses based on supervised learning. They introduced two attention layers: a Euclidean-based attention layer for considering neighbouring houses based on structural feature similarities and a spatial kernel-based attention layer, Geo Attention, for weighing neighbours based on geographic proximity to the target house. These attention layers, coupled with the geographical and structural features of the house, were fed into a fully connected network, culminating in a regression layer for house price prediction. This architecture yielded what they called a 'house embedding,' encapsulating the house attributes and spatial context into a common subspace, serving as a feature set for any regressor to estimate house prices. Our model extrapolates upon this framework, retaining the essence of generating 'house embeddings' but enhancing the architecture with Multi-head and Gated Attention mechanisms. These innovations are delineated into two distinct attention modules: Geographical Attention and Structural Attention. The Geographical Attention mechanism focuses on spatial relationships and proximities among properties, rendering a more nuanced understanding of the geographical context. Concurrently, the Structural Attention mechanism dives into the intrinsic attributes of properties such as size, age, condition, and the neighbouring points of interest, offering a granular perspective on the structural context. The Multi-head facet of our model unleashes the potential for parallel processing of geographical and structural information, thereby capturing a rich tapestry of spatial relationships from diverse dimensions. Each head in the Multi-head Attention mechanism could focus on different aspects or scales of spatial relationships, thus enriching the spatial context captured by the model. Furthermore, the Gated Attention mechanisms are orchestrated to modulate the information flow through the network meticulously. This refined control over the attention distribution is instrumental in mitigating outliers' impact on the estimated values, thereby promising more robust and accurate house price predictions. Our model, therefore, stands as a sophisticated augmentation of the attention-based spatial interpolation model conceived by Vianna and Barbosa. By synthesising the Multi-head and Gated Attention mechanisms with a bifurcated focus on geographical and structural relationships, our model unfolds a promising avenue for more accurate and insightful real estate price predictions. This innovative approach, rooted in the pioneering work of Vianna and Barbosa, yet elevated with novel attention mechanisms, propels the discourse in this domain towards new vistas, potentially laying a robust foundation for subsequent research endeavours and practical applications in real estate price prediction. Our project introduces significant contributions to advance real estate price prediction. We combine machine-learning techniques with a deep understanding of spatial heterogeneity in real estate valuations. Below are our contributions:

- **Introducing a New Dataset:** A new dataset for Italian cities has been introduced, likely including features relevant to real estate valuation related to 8 Italian cities.
- **Incorporating Various Attention Mechanisms:** applying multi-head gated Attention to capture with different weights and basis to capture different structural and geographical contexts based on the similarities.
- **Testing our model on different datasets:** We have tested our approach on other datasets to solidify the model's effectiveness in predicting house prices in other areas and diverse datasets.

---

[1]Kriging is a regression method used in spatial analysis (geostatistics) that allows one to interpolate a quantity in space, minimising the mean square error.

The subsequent sections of this document are structured as follows: Section 2 presents a comprehensive overview of relevant works, including literature and methodologies, that relate to house price estimation and section 3 delves into our proposed attention network, detailing its unique features and potential benefits. In Section 4, we conduct experiments, perform data analysis, and provide a thorough evaluation of our model. Lastly, in Section 5, we draw insightful conclusions based on our experimentation, compare our approach with prior methodologies, and articulate the implications of our findings. This structure ensures a coherent and comprehensive understanding of our innovative methodology for house price prediction.

2 Related works

House price estimation is a critical activity with far-reaching implications in the real estate industry. This field has been the subject of extensive academic research, traditionally employing regression analyses that integrate multiple variables, data types, and methodologies. In this review, we explore the scholarly landscape of this subject, tracing the evolution of research methodologies and spotlighting modern advancements and emerging trends.

The Hedonic Price Theory, first introduced by Rosen in 1974 [16], is the foundation for Hedonic Regression models. These models have become a crucial tool in studying house prices. The theory utilises a set of attributes, such as the number of bedrooms or bathrooms, to explain and represent a house's market value. These attributes are ranked based on their impact on a house's utility function, assuming that a market equilibrium between buyers and sellers determines the sale price. Hedonic Regression models are widely used to analyse the effects of different factors on house prices in various areas, making them a robust tool for market segmentation [17]. Although the original Hedonic Price Theory focused mainly on the intrinsic characteristics of a house, it has evolved to account for external factors like location [18]. This adaptation was motivated by the realisation that solely considering a house's intrinsic attributes was insufficient for accurate price representation [19]. Despite its widespread use, Hedonic Regression models have faced challenges, including issues related to the stability of attribute coefficients across different locations and property types, as well as limitations in handling non-linearity and model specification [20].

The integration of machine learning into house price prediction has been significantly accelerated by advancements in computational capabilities and the increase of data [21]. Initially, the focus was mainly on traditional machine learning algorithms such as Linear Regression (LR) [22]. While these linear models offered computational efficiency and ease of interpretation, they were limited in capturing the high-dimensional and non-linear complexities inherent in transaction price data. To address these limitations, researchers explored regularisation techniques like Ridge and Lasso Regression [23, 24]. These methods helped mitigate overfitting and offered a more refined approach to feature selection but struggled with capturing complex, non-linear relationships. Principal Component Analysis (PCA) [25] has also been employed for dimensionality reduction to simplify the feature space, although it has been criticised for potentially discarding crucial information. This led to the exploration of more flexible, non-linear models such as Support Vector Regression (SVR) [26] and Decision Trees [27]. Support Vector Regression (SVR) offers a solution for non-linearities through various kernel functions, while Decision Trees provide a simple yet effective approach for detecting non-linear patterns [26, 27]. However, Decision Trees are prone to overfitting. To combat this, ensemble methods like Random Forests were developed to improve model generalisation [28]. Random Forests combine the outcomes of many decorrelated trees to minimise variance and enhance accuracy.

With advancements in computational power, the field has shifted to more sophisticated ensemble methods such as XGBoost [29]. Unlike Random Forests, XGBoost constructs trees sequentially to correct the errors made by the previous ones. This makes XGBoost particularly effective in handling diverse data structures and enhancing prediction accuracy [30]. These advanced ensemble models are also highly scalable and efficient, often surpassing Random Forests' performance on large datasets.

To further optimise their predictive performance, these sophisticated ensemble models are often fine-tuned using metaheuristic optimisation techniques like Particle Swarm Optimization (PSO) [31, 32]. These optimisation techniques enable precise tuning of hyperparameters, resulting in models that are both accurate and computationally efficient.

The latest development in house price prediction is Graph Neural Networks (GNNs) [33], which excel in identifying spatial relationships between properties. However, GNNs can be computationally

demanding and require large, well-curated datasets for practical training. Additionally, their performance can vary significantly based on the architecture and hyperparameters, which may hinder their widespread adoption.

Furthermore, the domain has seen the rise of deep learning techniques. Deep Neural Networks (DNNs) [34] can automatically learn feature representations, eliminating the need for manual feature engineering. Although DNNs can unravel highly complex relationships in the data, they present challenges, such as the risk of overfitting and the need for substantial datasets and computational resources for practical training.

Building on these advancements, recent research has focused on integrating diverse computational models and data sources. A groundbreaking study by Tchuente et al. [35] on the French real estate market is a prime example. Utilising machine learning techniques such as Random Forest, AdaBoost [36], and gradient boosting [37], along with geocoding features, they analysed five years of historical real estate transactions provided by the French government. Their findings revealed that incorporating geocoding elements increased the models' predictive accuracy by over 50

Building upon the findings of Tchuente et al., the research conducted by Zhao et al. [38] represents a significant advancement in data analysis. By incorporating a multi-modal approach that encompassed traffic patterns, amenities, and social emotions in the bustling city of Beijing, China, this study validated the crucial role of location-based data. Furthermore, it introduced a feature-ranking mechanism that established a direct correlation between the data and its economic impact. This groundbreaking research underscores the potential of geolocated data in predicting real estate prices and highlights its transformative capabilities. Further advancing this research domain, De Nadai et al. [39] delved into the economic repercussions of neighbourhood characteristics within Italian urban landscapes. Their investigative toolkit encompassed a rich array of data sources including OpenStreetMap[2], Urban Atlas 2012, imagery from Google Street View, Italian census data[3], alongside property tax records sourced from the "Immobiliare. it"[4] platform. Through the application of their model, they witnessed a notable 60% enhancement in nowcasting housing prices, thereby underpinning the transformative potential of leveraging rich, geolocated datasets. Sarkar Snigdha Sarathi Das et al. [40] It has introduced the concept of Geospatial Network Embedding (GSNE). Unlike traditional models that often overlook the geospatial context of neighbourhood amenities, GSNE aims to capture this crucial aspect. The study emphasises that the proximity of a house to key points of interest (POIs) like train stations, highly-ranked schools, or shopping centres can significantly influence its price. The GSNE model leverages graph neural networks to create embeddings of houses and various types of POIs in multipartite networks. In these networks, houses and POIs are attributed nodes, representing their relationships as edges. This is particularly promising because it allows the model to understand complex latent interactions between houses and POIs, offering a robust and effective way to incorporate geospatial context.

Yuhao Kang et al. Kang et al.[41] delve into house price appreciation rates, employing a multi-source extensive geo-data framework that amalgamates structural attributes, locational amenities, and visitor patterns, employing machine learning models and geographically weighted regression for accurate predictions at both micro and macro scales. Their gradient-boosting machines achieve an R-squared value of 74% at the neighbourhood scale, highlighting the effectiveness of their approach in understanding house price appreciation nuances. On a similar innovative trajectory, Pei-Ying Wang et al. [42]. Propel house price prediction forward by harnessing a Joint Self-Attention Mechanism intertwined with a rich analysis of heterogeneous data, including public facilities and environmental aesthetics captured through satellite imagery. Tested in Taipei and New Taipei, this model eclipses other machine learning-based models in prediction accuracy, showcasing a lower error rate. The Spatial Transformer Network (STN)[43] and their model's novel joint self-attention mechanism intricately dissect the complex relations between different attributes impacting house prices. This work accentuates the necessity of a holistic data-rich approach and extends the versatility of the attention mechanism across various domains, setting a robust foundation for future research. In a parallel vein, Viana and Barbosa [15] introduce a groundbreaking framework that melds the spatial essence of real estate with the structural attributes of houses. Their hybrid attention mechanism orchestrates a balanced blend between the Euclidean space of structural features and the geographic tapestry, crafting them into a unified predictive model. The inception of a house embedding vector carries through

---

[2]europe/italy.html
[3]https://www.istat.it/
[4]www.immobiliare.it

the regression analysis domain, offering a fresh lens to capture spatial dependencies. This attention-infused approach heralds a promising avenue where the convergence of spatial interpolation and machine learning unravels a richer understanding of housing market dynamics, further amplifying the potential of attention mechanisms in elucidating the multifaceted nature of house price predictions. The related work showcases a trajectory towards crafting more nuanced, robust, and insightful real estate price prediction models. These models progressively harness multi-source, geolocated data and sophisticated machine learning techniques, notably attention mechanisms. This evolution reflects a maturing field poised to address the intricate challenges inherent to urban landscapes and real estate markets.

3 Methodology

Our proposed methodology aims to create robust house embeddings by assessing the similarity between a specific house and its neighbouring properties. This approach goes beyond merely considering individual property attributes and geographical location. Instead, it encapsulates each house's local characteristics with its immediate surroundings. Unlike traditional methods, we integrate the geographical coordinates of the property to refine this embedding further, capturing the essence of its surroundings and their relation to critical landmarks or amenities.

Our approach is based on the Attention-Based Spatial Interpolation (ASI) architecture proposed by Viana and Barbosa [15]. This architecture creates geographical and Euclidean similarities and emphasises specific similar points using an attention mechanism. However, more than a simple attention head may be required to capture differentiated interrelations. For this reason, our model employs multi-head-gated attention mechanisms to optimise the extraction of these features and their interrelationships. Multi-Head Gated Attention allows the model to capture multiple contexts, such as architectural styles, proximity to amenities, and other relevant features. Concurrently, the gated attention mechanism controls the flow of information to ensure that only the most pertinent attributes are considered. This is particularly useful when there is a significant variance between the target house and its neighbours, allowing the model to focus on the most critical similarities or differences. The Euclidean Multi-Head Gated Attention layer, represented in Figure 1 (A), calculates attention weights for the structural features of neighbouring houses based on their Euclidean distance to A_i. Concurrently, the Geographical Multi-Head Gated Attention layer in Figure 1 (B) learns the spatial correlations between the n-nearest geographical neighbours of house i. The output vectors from both attention layers are concatenated with A_i and G_i and fed into a fully connected neural network, culminating in a regression layer. This architecture synthesises the influence of the neighbouring houses and the target house's attributes into a single vector, termed the "house embedding" illustrated in Figure 1.

3.1 Background knowledge

To perform predictive analysis in real estate valuation, it is crucial to have a solid foundation of knowledge. This field employs a variety of methodologies and algorithms that are based on fundamental principles and metrics. Understanding these concepts is essential for accurately performing advanced analytical techniques. This subsection aims to clarify some of these key concepts and metrics, providing a starting point for a deeper exploration and comprehension of the subsequent methodologies and evaluations.

3.1.1 Similarity calculation

In the intricate landscape of data science, similarity is a critical underpinning for various algorithms and methodologies. This sub-subsection aims to illuminate the key metrics ubiquitously employed to quantify similarity, laying the groundwork for the following analyses.

- **Euclidean Distance**: A foundational metric in geometry, Euclidean distance provides a straightforward measure of similarity by calculating the straight-line distance between two points in an Euclidean space.

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

- **Cosine Similarity**: This metric is invaluable in high-dimensional spaces, measuring the cosine of the angle between two vectors. It is especially pertinent in text analysis and natural language
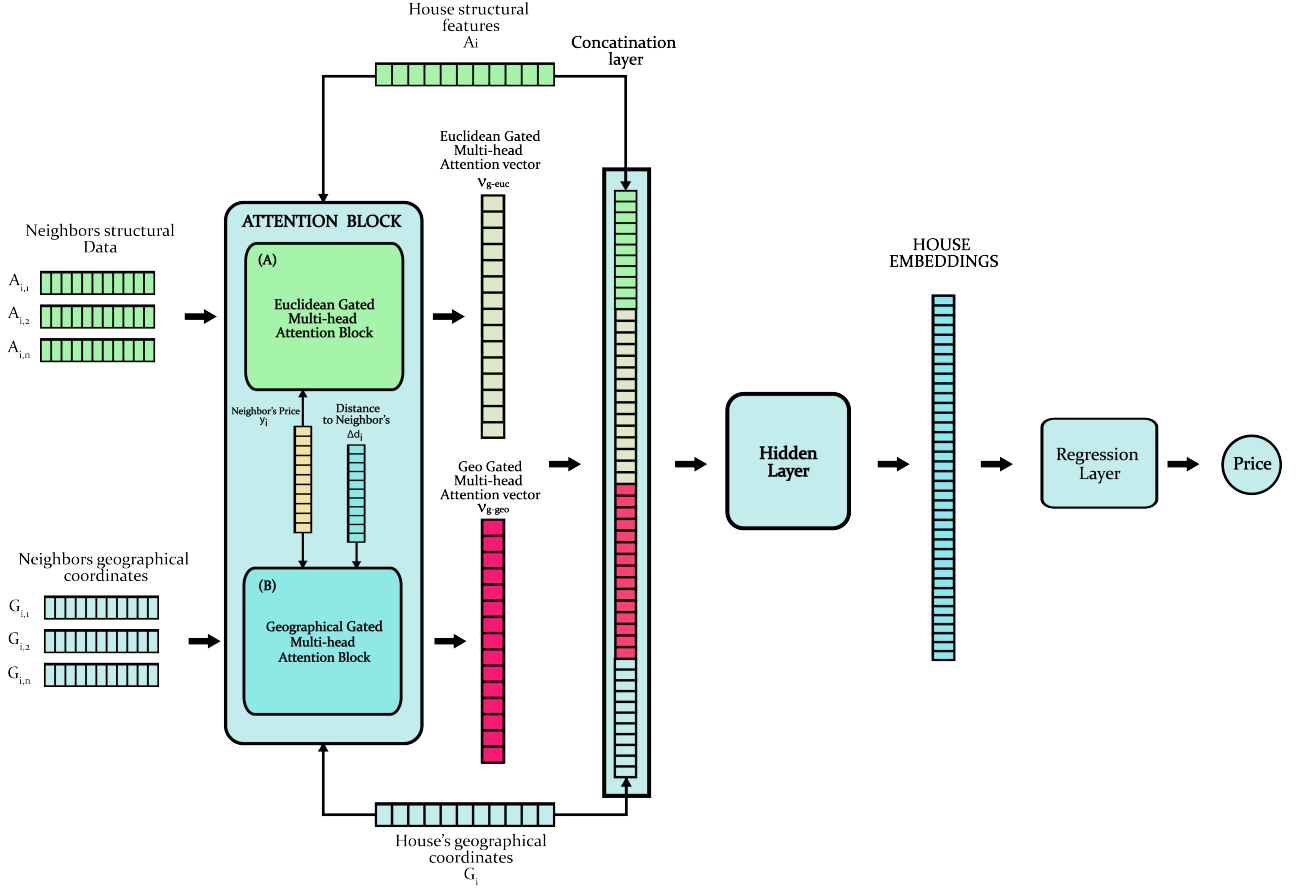
Fig. 1: Architecture representation of the multi-head gated-attention-based interpolation. (A) Represent the Euclidean interpolation block based on the multi-head gated Attention. (B) Represent the geo-interpolation block based on the Multi-Head Gated attention.

processing.

$$\text{Cosine Similarity} = \frac{C \cdot D}{\|C\| \times \|D\|} \tag{2}$$

- **Jaccard Index**: A set-based metric, the Jaccard Index is helpful for categorical data, quantifying the ratio of the intersection to the union of two sets.

$$J(C, D) = \frac{|C \cap D|}{|C \cup D|} \tag{3}$$

- **Identity Similarity**: This is a binary similarity measure used to ascertain whether or not two data points are identical. Unlike continuous similarity measures, the Identity Similarity scores one if the data points are similar and 0 if they differ. This measure is handy in scenarios requiring exact matching or where data is categorical. Mathematically, it is expressed as:
- **Gaussian Kernel**: Also known as the Radial Basis Function (RBF) with Gaussian form, this metric is a cornerstone in non-linear data transformations. Unlike other metrics that measure distance directly, the Gaussian Kernel calculates similarity by mapping the original data points into a higher-dimensional space through a Gaussian function. This allows it to capture complex, non-linear relationships between data points. Mathematically, it is expressed as:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{4}$$

The parameter σ controls the spread of the Gaussian function, thereby influencing the similarity measure. A smaller σ will result in a narrower Gaussian function, making the similarity measure more sensitive to the distance between data points.

These metrics serve as the backbone for various algorithms and offer a nuanced understanding of how data points relate to each other in complex spaces, with the Gaussian Kernel standing out for its ability to capture non-linear relationships.

3.1.2 Spatial interpolation

Spatial interpolation is a critical technique for predicting unknown values at unobserved locations based on known values at observed locations, finding applications in diverse fields such as geostatistics, environmental science, and real estate. The effectiveness of spatial interpolation is intrinsically tied to the choice of similarity measures. For instance, Euclidean distance can be employed in a straightforward approach like "inverse distance weighting" (IDW) [44], where the influence of a neighbouring point on the interpolated value is inversely proportional to its Euclidean distance from the target location. On the other hand, the Gaussian Kernel [45] offers a more nuanced approach by transforming the Euclidean distance into a measure of similarity, thereby capturing complex, non-linear spatial relationships. This is especially useful in advanced geostatistical methods like kriging [46]. Therefore, the choice between straightforward measures like Euclidean distance and more complex ones like the Gaussian Kernel can significantly impact the quality of spatial interpolation, exemplifying the broader applicability and importance of similarity measures in data science.
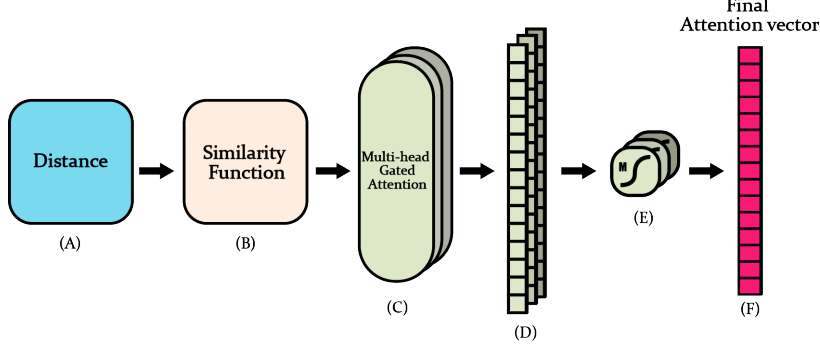
3.1.3 Attention Mechanisms

Attention mechanisms [47] has emerged as a cornerstone in many deep learning models, predominantly in sequence-to-sequence tasks such as machine translation and speech recognition. The essence of Attention is to emulate the human ability to focus on specific segments of input data, much like how we selectively concentrate on some aspects of a visual scene or a conversation. Among the diverse attention mechanisms, Soft Attention is a mechanism that computes a weighted sum of all input values. These weights, indicative of the relevance of each input, are typically determined using a softmax function, ensuring a normalised distribution where the weights sum up to one. The continuous nature of these weights makes soft Attention inherently differentiable, rendering it particularly amenable to gradient-based optimisation techniques [48]. On the other hand, intricate Attention operates more selectively. Instead of distributing focus across all inputs, it zeroes in on a specific subset, effectively sidelining the others. Given its discrete selection process, traditional backpropagation struggles with optimising intricate Attention. Yet, this challenge is surmountable with techniques like the reinforce algorithm [49]. The *Gated Attention* mechanism [50] bridges the gap between these two. It adeptly amalgamates information from diverse sources and employs gating tools to ascertain the relevance of each source. This approach can be perceived as a harmonious blend of the soft and hard attention paradigms, encapsulating the strengths while mitigating their limitations [51].

3.2 Attention Block

The Attention Block is the computational nucleus of our architecture, designed to intricately capture the spatial relationships essential for precise house price prediction. As delineated in Figure 1, this block comprises two main components: the Geo Multi-head Gated Attention and the Euclidean Multi-head Gated Attention. Each of these components consists of several key stages, contributing to generating their respective geo- and Euclidean-gated attention vectors. Figure 2 elucidates the fundamental principles for calculating the Geo and Euclidean attention mechanisms. In the initial stage, represented by Figure 2 (A), the Distance Calculation Block computes the distance between the target house and its neighbours. The nature of this distance is contingent on the specific attention mechanism in play, be it Geo or Euclidean. The Similarity Calculation Block, as depicted in Figure 2 (B), transforms these distances into similarity scores. A Gaussian kernel function is employed for Geo Attention, while alternative kernel functions may be used for the Euclidean variant. The subsequent component is the multi-head gated Attention, illustrated in Figure 2 (C). This block leverages the similarity scores to derive attention weights, which are then gated to modulate their influence. The entire process is executed across multiple heads, capturing various facets of the spatial relationships between the target house and its neighbours. Next, the aggregated attention head, represented by Figure 2 (D), consolidates the outputs from all attention heads into a single vector. This is achieved through a weighted sum, where the weights are adaptively learned during training. If the architecture employs multiple attention mechanisms, such as Geo and Euclidean, their aggregated attention heads are combined further. Following this, Figure 2 (E) illustrates the Final Aggregation Block. In this

stage, aggregated normalised gating weights are computed using a softmax function. After that, the weighted sums produced from each attention head are multiplied by these normalised weights. This aggregation is performed separately for the Geo and Euclidean attention mechanisms, resulting in their aggregated attention vectors. Finally, the vector produced from this aggregation process is the final attention vector, as depicted in Figure 2 (F). In summary, the Attention Block encapsulates the Multi-head Geo Gated Attention and the Euclidean Multi-head Gated Attention, generating their respective Geo and Euclidean Gated Attention Vectors.

Fig. 2: Comprehensive overview of the Gated Multi-head Attention mechanism within the Attention Block. (A) depicts the initial computation of geodesic and Euclidean distances, serving as the foundation for subsequent attention calculations. (B) illustrates the Similarity Function, which transforms these foundational distances into similarity scores. (C) shows the core Multi-Head Gated Attention Block, where these similarity scores derive gated attention weights across multiple heads. (D) Highlights the Aggregated Attention Head, consolidating the gated attention weights from all heads into a singular vector. (E) represents the aggregation of multiple gated attentions for each weighted sum. (F) Indicates the Final Attention Vector.

3.2.1 Geo Multi-head Gated Attention

The Geo Multi-head, Gated Attention mechanism is intricately designed to capture the spatial relationships between a target house and its neighbouring properties. This involves using a Gaussian kernel function to calculate geographic similarity scores between the target house and its neighbours. Equation 5 demonstrates how the geographic score between the target house G_i and its neighbouring house $G_{i,j}$ is computed using the Gaussian kernel function.

$$s(G_i, G_{i,j}) = \exp\left(-\text{geo\_dist}(G_i, G_{i,j}) \times \rho\right) \tag{5}$$

Here, $\rho = \frac{\sigma^2}{2}$ and $\text{geo\_dist}(G_i, G_{i,j})$ represents the geodesic distance between G_i and $G_{i,j}$. The vector of similarity scores L is then transformed into a hidden representation H' through a fully-connected layer, as described in Equation 6:

$$H' = W' \cdot L + b' \tag{6}$$

In this equation, W' and b' are the learned weights and bias terms, respectively. The attention weights a_{geo} are computed using a softmax layer, as formulated in Equation 7:

$$a_{\text{geo}}(G_i, G_{i,j}) = \frac{\exp(H'_j)}{\sum_{j'=1}^{n} \exp(H'_{j'})} \tag{7}$$

Then, using our defined gated attention mechanism (Equation 8), we apply it to the attention weights:

$$\text{Gate}(x) = \sigma(W_g \cdot x + b_g) \tag{8}$$

Subsequently:

$$a'_{\text{geo}}(G_i, G_{i,j}) = \text{Gate}(a_{\text{geo}}(G_i, G_{i,j})) \odot a_{\text{geo}}(G_i, G_{i,j}) \tag{9}$$

Where:

- x is the input value, in this case, the original attention weight $a_{\text{geo}}(G_i, G_{i,j})$.
- W_g represents the learned weight matrix associated with the gate.
- b_g denotes the bias term.
- σ is the sigmoid function, ensuring the output value of the gate lies in the [0,1] range.

With this, the Geo Gated Attention Vector $v_{\text{ggeo}}(G_i)$ is computed as a weighted sum of the features of the neighbouring houses using the modified attention weights a'_{geo}:

$$v_{\text{ggeo}}(G_i) = \sum_{j=1}^{n} a'_{\text{geo}}(G_i, G_{i,j})[G_{i,j} \oplus A_{i,j} \oplus \Delta d_{i,j} \oplus y_{i,j}] \tag{10}$$

In this equation, $\Delta d_{i,j}$ represents the geographic distance between house i and its neighbour j. Similarly, $y_{i,j}$ signifies the price of the neighbor j, and $\oplus$ denotes the concatenation operation. The dimensionality of $v_{\text{geo}}(G_i)$ is derived from the summation of dimensions where $G_{i,j} \in \mathbb{R}^2$, $A_{i,j} \in \mathbb{R}^T$, $\Delta d_{i,j} \in \mathbb{R}^1$, and $y_{i,j} \in \mathbb{R}^1$. Consequently, the vector $v_{\text{geo}}(G_i)$ can be viewed as a weighted sum of vectors $G_{i,j}$, concatenated with $\Delta d_{i,j}$ and $y_{i,j}$, and weighted using the normalised geo gated attention coefficients which are determined during the training process.

3.2.2 Euclidean Multi-head Gated Attention

The Euclidean Multi-head Gated Attention mechanism is precisely engineered to emphasise the most relevant structural similarities between a target house and its neighbouring properties. This mechanism employs the Euclidean distance to compute the similarity scores between the target house and its neighbours. The Euclidean distance between the target house A_i and a neighboring house $A_{i,j}$ is computed as shown in Equation 11:

$$d(A_i, A_{i,j}) = \sqrt{\sum_{p=1}^{T}(a_{i,p} - a_{i,j,p})^2} \tag{11}$$

where $d(A_i, A_{i,j})$ is the Euclidean distance indicating similarity between houses based on structural attributes, A_i represents the structural features of the target house i, $A_{i,j}$ denotes the structural features of the jth neighboring house to i, $a_{i,p}$ and $a_{i,j,p}$ are specific structural attributes of houses i and j, respectively, and T is the total number of structural attributes considered.

After computing the Euclidean distances, we construct a vector of similarity scores L, which is then transformed into a hidden representation H through a fully-connected layer, as described in Equation 12:

$$H = W \cdot L + b \tag{12}$$

In Equation 12, W and b are the learned weights and bias terms, respectively. The attention weights a_{euc} are computed using a softmax layer, as formulated in Equation 13:

$$a_{\text{euc}}(A_i, A_{i,j}) = \frac{\exp(H_j)}{\sum_{j'=1}^{n} \exp(H_{j'})} \tag{13}$$

The essence of the gated attention mechanism is to refine the attention weights by introducing an additional modulation step. This modulating factor, or "gate", is typically represented as a value between 0 and 1 and is applied element-wise to the attention weights. The purpose is to amplify or diminish the original attention values based on the model's learned parameters.

Given this, the gated attention can be defined as:

$$\text{Gate}(x) = \sigma(W_g \cdot x + b_g) \tag{14}$$

Where:

- x is the input value, in this case, the original attention weight $a_{\text{euc}}(A_i, A_{i,j})$.
- W_g represents the learned weight matrix associated with the gate.
- b_g denotes the bias term.

- σ is the sigmoid function, ensuring the output value of the gate lies in the [0,1] range.

Subsequently, the gated attention mechanism can be formalised as:

$$a'_{\text{euc}}(A_i, A_{i,j}) = \text{Gate}(a_{\text{euc}}(A_i, A_{i,j})) \odot a_{\text{euc}}(A_i, A_{i,j}) \tag{15}$$

Here, $\odot$ denotes element-wise multiplication. Thus, the attention weight is modulated by its gating value, allowing the model to allocate attention more selectively to houses exhibiting the most congruent features.

The Vector with Euclidean Gated Attention denoted as $v_{\text{geuc}}(A_i)$, represents a cumulative weighted mix of attributes from the surrounding homes. This process uses the gated attention coefficients a'_{euc} and is illustrated in Equation 16:

$$v_{\text{geuc}}(A_i) = \sum_{j=1}^{n} a'_{\text{euc}}(A_i, Ai, j) \odot [A_{i,j} \oplus y_{i,j}] \tag{16}$$

Within Equation 16, $y_{i,j}$ defines the price of the jth neighboring home of house i, while $\oplus$ denotes the concatenation action. The size of $v_{\text{euc}}(A_i)$ stands at $T + 1$ given that $A_{i,j}$ resides in $\mathbb{R}^T$ and $y_{i,j}$ is part of $\mathbb{R}^1$. The composition of $v_{\text{euc}}(A_i)$ involves initially multiplying the combined vector $[A_{i,j} \oplus y_{i,j}]$ for each jth neighbor of house i by its respective gated attention coefficient $a'_{\text{euc}}(A_i, A_{i,j})$, producing an individual weighted vector for every j^{th} neighbor. An overall summation is then applied to these vectors for all n adjacent homes to house i. Consequently, the elements within $v_{\text{euc}}(A_i)$ represent a comprehensive weighted sum of the structural attributes and the valuations of the nearby homes of house i. The gated attention coefficients undergo refinement during the learning phase.

3.2.3 Final Aggregation Block

The final aggregation stage shown in Figure 2 E involves collecting and combining the attention vectors from each head of the attention mechanism and applying the gated attention based on the normalized gates weights and biases. It is important to note that this process is unique for each attention mechanism, namely Geo and Euclidean, and it results in the formation of two separate aggregated attention vectors.

To ensure the effectiveness of the attention mechanism in both Geo and Euclidean interpolation, it is crucial to normalise the gating weights and biases using a softmax function, as shown in Equation 17. By normalizing the gating weights and biases, they fall within the range of 0 to 1, which makes them more easily interpretable.

$$\text{gate}_{\text{norm},i} = \frac{\exp(\text{Gate\_weights}_i + \text{Gate\_bias}_i)}{\sum_{j=1}^{n} \exp(\text{Gate\_weights}_j + \text{Gate\_bias}_j)} \tag{17}$$

After normalising the gating weights and biases, we perform element-wise multiplication with each attention and then aggregate them. The resulting vector that shows the aggregated gated geographic attention, denoted as $v_{\text{agg\_ggeo}}$, is presented in Equation 18.

$$v_{\text{agg\_ggeo}} = \sum_{i=1}^{n} \text{gate}_{\text{norm, geo},i} \odot v_{\text{ggeo},i} \tag{18}$$

Where $\text{gate}_{\text{norm, geo},i}$ represents the softmax-normalized gating weights and biases, and $v_{\text{geo},i}$ refers to the attention vectors from the Geo attention heads.

In a similar vein, the aggregated gated Euclidean attention vector $v_{\text{agg\_geuc}}$ is represented by Equation 19:

$$v_{\text{agg\_geuc}} = \sum_{i=1}^{n} \text{gate}_{\text{norm, euc},i} \odot v_{\text{geuc},i} \tag{19}$$

Here, $\text{gate}_{\text{norm, euc},i}$ signifies the softmax-normalized gating weights, and $v_{\text{geuc},i}$ portrays the gated attention vectors emergent from the Euclidean attention heads.

In conclusion, the consolidated Geo attention vector $v_{\text{agg\_ggeo}}$ and the Euclidean attention vector $v_{\text{agg\_geuc}}$ are computed using an element-wise multiplication between the softmax-normalized gating weights and their corresponding attention vectors as illustrated in Figure 2 F derived from the Geo and Euclidean attention heads, respectively. This approach ensures an accurate integration of the significance associated with each feature and reflects the complex spatial relationships inherent within the Geo and Euclidean contexts.

3.3 House embeddings

Embeddings serve as a pivotal component in contemporary machine-learning architectures, especially in scenarios that involve the manipulation of high-dimensional or categorical variables. In the realm of real estate price prediction, the utility of embeddings is accentuated for the encoding of categorical attributes such as neighbourhood classifications, types of properties, and associated amenities into a continuous vector space [52]. These continuous embeddings can capture intricate relationships between disparate categories, thereby augmenting the predictive efficacy of the machine learning model [53, 54]. The transformation from a sparse, high-dimensional feature space to a dense, lower-dimensional vector space has found applications across a multitude of domains, ranging from natural language processing to recommendation engines and graph-based machine learning algorithms [55–58]. However, effectively utilising embeddings necessitates meticulous tuning and validation to mitigate the risk of overfitting and ensure robust generalisation on unseen data [59]. In the present study, as delineated in Figure 1, we introduce a novel methodology for generating house embeddings. Initially, two distinct Multi-Head Gated Attention mechanisms are employed: one geographically oriented (Geo Multi-Head Gated Attention) and another focused on structural attributes (Euc Multi-Head Gated Attention). The Geo Multi-Head, Gated Attention mechanism leverages the geographical coordinates of proximate properties, while the Euc Multi-Head Gated Attention mechanism utilises the structural attributes of neighbouring properties. The vectors generated from these attention mechanisms are concatenated with the original geographical (Gi) and structural (Ai) attributes of the property. This concatenated vector is propagated through a hidden neural layer to synthesise the final house embeddings. This intricate methodology enables the model to assimilate both geographical and structural nuances, thereby enhancing its predictive capabilities.

3.4 Regression layer

For the empirical component of our study, we employed a diverse set of regression algorithms, each optimised through rigorous cross-validation techniques. The algorithms were selected based on their suitability for the specific characteristics of our dataset as well as the computational resources at our disposal. Below is an exhaustive list of the algorithms utilised:

- **Linear Regression (LR)**: Utilized with default hyperparameters as implemented in the scikit-learn library [60]. This algorithm serves as a baseline model for our study.
- **Random Forest (RF)**: An ensemble of decision trees, optimised using grid search and k-fold cross-validation. Hyperparameters such as the number of trees was varied, with tests conducted for 50, 100, 200, 700, and 1000 trees [61].
- **LightGBM (LGBM)**: A gradient boosting framework that uses tree-based learning algorithms. Hyperparameters including the number of trees (50, 100, 200), the number of leaves (3, 4, 5, 100, 300), and the learning rate (0.03, 0.05, 0.07, 0.1) were fine-tuned [62].
- **Extreme Gradient Boosting (XGB)**: An optimised distributed gradient boosting library, fine-tuned through cross-validation. Parameters such as minimum child weight, gamma, subsample, column sample by the tree, learning rate, and maximum depth were adjusted [5].
- **Categorical Boosting (CatBoost)**: An algorithm specifically designed for handling categorical variables. The depth parameter was optimised, with tests conducted for depths of 8 and 10 [63].
- **K-Nearest Neighbors (KNN)**: A distance-based algorithm, optimised by adjusting the number of neighbours, with tests conducted for 10 and 15 neighbours [64].
- **Decision Tree (DT)**: A basic tree-based model, optimised by adjusting the maximum depth parameter, with tests conducted for a depth of 9 [65].
- **Support Vector Machines (SVM)**: A kernel-based algorithm suitable for linear and non-linear problems. Parameters 'C' and 'gamma' were fine-tuned using cross-validation [66].
- **Regression Layer (RL)**: This layer serves as the terminal component of our attention-based neural network model, generating the final housing price prediction based on the feature map (house embeddings) obtained from preceding layers.

This empirical analysis aims to provide a comprehensive evaluation of the selected algorithms, thereby elucidating the relative merits and demerits in the context of housing price prediction.

4 Experimentation

This section presents the experimentation methodology adopted for our house price prediction task, including the specifics of the dataset preparation, model implementation, training, and evaluation process.

4.1 Dataset

In the experimental section, we utilised several datasets from different cities across various parts of the world to showcase the effectiveness of our model.

Table 1: Summary of Datasets

Dataset	Price Range	Number of Samples	Number of Features
IT (Italian)	(60000 to 720000) Euro	30,918	24
BJ (Beijing)	(5500 to 170000) Yuan	28,550	26
KC (Kings county)	(75,000 to 7,700,000) Dollar	21,650	18
POA (Porto Alegre City)	(70,000 to 1,168,324) Reais	15,368	7

1. **Italian (IT) Dataset**: We obtained our dataset of Italian (IT) properties from Immobiliare. It is a well-known real estate platform in Italy. To collect the data, we designed a web scraper that extracted information from eight different cities: Genoa, Milan, Turin, Rome, Bologna, Florence, Naples, and Palermo. We filtered the data to include only five types of properties, such as apartments and penthouses while excluding outliers like farms, buildings, and properties under construction. This ensured that the dataset was representative and coherent. We then conducted a thorough cleaning process to eliminate outliers. This process helped us eliminate data entry errors and rare property types, resulting in a consistent dataset suitable for analysis. To enrich the dataset, we added geographical data points. We included precise longitude and latitude coordinates for each property listing and leveraged OpenStreetMap to enhance each listing with Points of Interest (POI) data. This provided more profound insights into the property's surroundings, which could be significant in assessing its value. The final IT dataset comprises 30,918 property listings spread across eight significant cities in Italy. Each listing includes 19 distinct features that capture structural attributes, such as surface area, year of construction, and geographical details.
2. **Beijing (BJ) Dataset**: This dataset consists of 28,550 real estate transactions in Beijing and is sourced from the H4M study [67]. It includes 25 features, which range from structural attributes like surface area and year of construction to geographical elements such as district location and Point Of Interest (POI) information. The features are detailed in Table 1.
3. **Kings County (KC) Dataset**: Sourced from the GitHub[5] In the repository associated with the "Attention-Based Interpolation" paper, there is a dataset representing the Kings County, USA housing market. This dataset comprises 21,650 house samples, characterized by 19 distinct features. These features, which encompass both structural and geographical attributes, are detailed in a separate table, Table 1.
   It's important to note that the prices in this dataset are provided in a log-scaled format.
4. **Porto Alegre City (POA) Dataset**: Derived from the repository provided by Vianna and Barbosa, this dataset focuses on Brazil's Porto Alegre City housing market. It includes 15,368 house samples, each described by 6 features, similar to the KC dataset. The features are outlined in Table 1.
   It's essential to recognize that the prices in this dataset are provided in a log-scaled format.

4.2 Model configuration

Our model was developed in a Python 3.7 environment, using TensorFlow 2.5 as the backend for the Keras framework. The model was executed on a system with an Intel Core i5-13700K CPU and an NVIDIA GeForce RTX 3070 GPU. We used cross-validation and grid search techniques for hyperparameter tuning to achieve optimal results with regression algorithms such as XGBoost and RandomForest. For our custom model, we fine-tuned the hyperparameters using a validation subset

---

[5]https://github.com/darniton/ASI

of the data to obtain the best possible embedding representation and predictive performance. The hyperparameters and their values are summarised in Table 2, and we describe each hyperparameter and its significance below.

- **n-nearest**: Specifies the number of nearest neighbours to consider. The best values were 40 for IT, 60 for KC, 60 for POA, and 30 for BJ.
- **sigma (σ)**: Controls the width of the Gaussian kernel. Optimal values were 2 for IT, 2 for KC, 2 for POA, and 10 for BJ.
- **nodes**: Represents the number of nodes in the hidden layers. The best values were 60 for IT, 60 for KC, 60 for POA, and 60 for BJ.
- **Num_heads**: Specifies the number of attention heads in the model. Optimal values were 8 for IT, 8 for KC, 4 for POA, and 4 for BJ.
- **num_geo**: Indicates the number of geographical features to consider. The best values were 30 for IT, 30 for KC, 10 for POA, and 15 for BJ.
- **num_euc**: Represents the number of Euclidean dimensions for distance calculations. The best values were 25 for IT, 30 for KC, 15 for POA, and 15 for BJ.
- **LR (Learning Rate)**: Controls the step size during optimization. Optimal values were 0.001 for IT, 0.008 for KC, 0.001 for POA, and 0.001 for BJ.
- **batch size**: Specifies the number of samples per batch during training. Optimal values were 32 for IT, 250 for KC, 32 for POA, and 250 for BJ.
- **act func (Activation Function)**: Either Rectified Linear Unit (ReLU) or Exponential Linear Unit (ELU) was used. ELU was optimal for all datasets.
- **hidden act function (Hidden Layer Activation Function)**: The activation function for the hidden layers was either ReLU, ELU, regression, or linear. The linear function was optimal for all datasets.
- **similarity function**: We used the Gaussian Kernel and Identity function to compute similarities between data points. The Identity function was optimal for IT and POA, while the Gaussian Kernel was optimal for KC and BJ.

Table 2: hyperparameters that were used to train our model

HP Values	General Values	Best Values			
		IT	KC	POA	BJ
N-nearest	5, 10, 15, 60	40	60	60	30
Nearest-geo	20, 25, 30, 35, 40, 45, 50, 55, 60	30	30	10	15
Nearest-Euclid	20, 25, 30, 35, 40, 45, 50, 55, 60	25	30	15	15
Num_heads	1,2,4,8,12,15	8	8	4	4
Sigma(σ)	2, 5, 10, 15, 20	2	2	2	10
Nodes	5, 10, 15, 60	60	60	60	60
LR	[0.001-0.01]	0.001	0.008	0.001	0.001
Batch size	250, 300, 400, 500	32	250	32	250
Act func	Relu and ELU	ELU	ELU	ELU	ELU
Hidden act func	Relu, ELU, regression and linear	linear	linear	linear	linear
Similarity function	Identity and Gaussian Kernel	Identity	Gaussian Kernel	Identity	Gaussian Kernel

4.2.1 Evaluation Metrics

After training, the model was evaluated using standard regression metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MALE). These metrics serve specific purposes in assessing the model's performance:

- **RMSE (Root Mean Squared Error)**: Provides a measure of the model's prediction error, penalising more significant errors more severely than smaller ones. It is advantageous when significant errors are undesirable in the prediction task.
- **MALE (Mean Absolute Logarithmic Error)**: This metric expresses the average magnitude of the relative errors between predicted and actual values while disregarding their direction. It is beneficial when dealing with exponential growth, or underestimation is more critical than overestimation.

Table 3: Benchmark the datasets on state-of-the-art machine learning models. The average value is referred to k-fold cross-validation with k=10

Model	IT				KC				POA				BJ			
	MALE ↓		RMSE ↓		MALE ↓		RMSE ↓		MALE ↓		RMSE ↓		MALE ↓		RMSE ↓	
	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg
LR	0.385	0.388	76224	76241	0.1924	0.1925	205460	209330	0.2610	0.2611	152861	153775	0.2394	0.2396	20551	20558
KNN	0.247	0.248	84637	85163	0.1501	0.1513	174046	175628	0.2065	0.2078	122521	123113	0.1116	0.1121	12093	12149
DT	0.197	0.205	69085	70423	0.1583	0.1608	158937	178296	0.2163	0.2195	127382	128915	0.0936	0.0954	10155	10409
RF	0.1502	0.1508	51774	52147	0.1245	0.1251	133933	136993	0.1716	0.1731	105183	105975	0.0784	0.0794	8369	8475
SVM	0.4072	0.4074	128634	128781	0.1331	0.1336	149265	152675	0.2232	0.2246	126911	128191	0.2234	0.2237	20652	20668
LGBM	0.1381	0.1384	46183	46492	0.1164	0.1175	122116	126076	0.172	0.177	104928	106705	0.0790	0.0796	8070	8152
CatBoost	0.1362	0.1368	**45942**	**46233**	0.1131	0.1141	120351	**123077**	0.1793	0.1775	105984	106593	0.0782	0.0785	7995	8066
XGB	**0.1350**	**0.1358**	46008	46396	**0.1160**	**0.1167**	119479	124459	**0.1613**	**0.1634**	100212	101614	**0.0723**	**0.0744**	7713	7836

These metrics collectively offer a comprehensive evaluation of the model's performance in predicting house prices, allowing for the assessment of the model's accuracy and its goodness of fit to the actual data.

4.3 Results and Interpretation

In our evaluation, we consider both the average and best performance metrics to offer a comprehensive view of each model's capabilities. The average performance metrics are derived from 10-fold cross-validation, indicating how the model will likely perform on unseen data. It gives us a more generalisable performance measure by mitigating the risk of the model overfitting to a particular subset of the data. On the other hand, the best performance metrics are extracted using grid search techniques. These values demonstrate the optimal performance that the model can potentially achieve under ideal hyperparameter settings. Including both types of metrics allows for a balanced understanding of the model's robustness and potential for excellence. It helps in identifying not just the most consistently high-performing models but also those with the capacity for superior performance under the right conditions.

4.3.1 Base models benchmark

Table 3 provides an exhaustive evaluation of multiple machine-learning models In an exhaustive evaluation of machine learning models on real estate datasets from Italy (IT), King's County (KC), Porto Alegre City in Brazil (POA), and Beijing (BJ), the best performance was consistently demonstrated by XGBoost (XGB). Specifically, XGB recorded the best MALE values of 0.1350 in IT, 0.1160 in KC, 0.1613 in POA, and 0.0723 in BJ. Notably, the average performance for XGB was stable and closely aligned with these best values, indicating high reliability across diverse geographic datasets. CatBoost and LightGBM also performed strongly, closely trailing XGB in each dataset. For instance, CatBoost had the best MALE values of 0.1362 in IT, 0.1131 in KC, 0.1793 in POA, and 0.0782 in BJ. LightGBM posted the best MALE deals of 0.1381 in IT, 0.1164 in KC, 0.172 in POA, and 0.0790 in BJ. The average performances of CatBoost and LightGBM were also impressively stable and nearly matched their respective best values. Conversely, Support Vector Machines (SVM) significantly underperformed, with its best MALE values being 0.4072 in IT, 0.1331 in KC, 0.2232 in POA, and a dismal 0.2234 in BJ. K-Nearest Neighbors (KNN), a traditional algorithm, also lagged, particularly in the BJ dataset, where it posted a best MALE of 0.1116. In summary, XGB takes the lead across all datasets regarding best and average performance metrics, closely followed by CatBoost and LightGBM, which also show highly stable average performances. Conversely, SVM and traditional models like KNN are less effective, particularly in complex, geographically diverse datasets.

Table 4: Performance Evaluation of our model against the ASI model and ANN

Model	IT		KC		POA		BJ	
	MALE ↓	RMSE ↓	MALE ↓	RMSE ↓	MALE ↓	RMSE ↓	MALE ↓	RMSE ↓
ANN	0.197	67835	0.2231	127900	0.2212	125961	0.239	19565
ASI	0.133	46473	0.112	124557	0.139	93818	0.075	7934
Ours	**0.1312**	**45797**	**0.110**	**107993**	**0.136**	**92020**	**0.073**	**7797**

4.3.2 Experimental results for our model

We conducted a series of tests to evaluate our model and compare it with ANN and ASI models. The dataset was split into three subsets, with 70% assigned for training, 20% for testing, and 10% for validation. Table 4 shows the performance metrics of our model, ASI, and ANN, compared across four different datasets from various locations: Italy (IT), King's County (KC), Porto Alegre (POA), and Beijing (BJ).

Our model performed better than the ASI model in the IT dataset. It has a superior MALE that's approximately 1.52% lower and an RMSE that's approximately 0.36% lower, with figures of 0.1312 and 45,797, respectively. These results highlight our model's ability to interpret the IT dataset more accurately.

In the KC dataset, our model's robustness shines with a notable 13.3% improvement in RMSE compared to the ASI model. This translates to a lower RMSE of approximately 107,993 for our model compared to ASI's 124,557. This underscores our model's consistent efficiency across various geographical contexts.

Our model also outperforms the ASI model in the POA and BJ datasets, demonstrating its versatility and accuracy in handling diverse data types. In the POA dataset, our model achieves a MALE of approximately 1.44% lower and an RMSE of approximately 13.45% lower compared to ASI. In the BJ dataset, our model achieves a MALE of approximately 2.67% lower and an RMSE of 2.31% lower compared to ASI. These metrics align with those observed in the IT and KC datasets, reaffirming the broad applicability of our model.

Our model incorporates Multi-Head Gated Attention mechanisms, which enable it to assimilate diverse spatial cues between houses and their surroundings. This fosters more precise and robust predictions as our model comprehensively grasps spatial dynamics.

In conclusion, our model has proven superior to the ASI model across all evaluated metrics and datasets. The advanced Multi-Head Gated Attention architecture plays a pivotal role in aggregating contextual cues, ultimately enhancing overall predictive accuracy.

4.3.3 Experimental results for house embeddings

In our experiment, we wanted to see how custom house embeddings generated by our Multi-Head Gated Attention model would affect the performance of various baseline machine learning models. These embeddings were created based on structural and geographical information and enhanced the feature space for algorithms like Linear Regression, KNN, Decision Tree, Random Forest, SVM, LightGBM, CatBoost, and XGBoost. We evaluated the models' performance using four different geographical datasets: Italy (IT), King's County (KC), Porto Alegre (POA), and Beijing (BJ), and assessed the Best and Average MALE and RMSE scores.

Our results showed that our custom embeddings significantly positively impacted the predictive performance of the baseline models. For example, when the CatBoost model was augmented with our custom embeddings, it achieved the lowest RMSE score in the IT dataset at 45,708, outperforming even our original Multi-Head Gated Attention model. However, we found that the improvement magnitude was inconsistent across all datasets. The IT dataset, which combines data from various cities with significant geographical and Euclidean distances between them, showed only a modest enhancement of around 1.3% in RMSE when deploying CatBoost with custom embeddings compared to the baseline.

We discovered that the unique spatial complexities inherent in each dataset could impact the effectiveness of the custom embeddings. For instance, in the KC dataset, CatBoost with custom embeddings demonstrated significant gains over its baseline, whereas, in IT, the improvements were

Table 5: Benchmark the datasets on state-of-the-art machine learning models on the generated embeddings from our model. The average value is referred to k-fold cross-validation with k=10

Model	IT				KC				POA				BJ			
	MALE ↓		RMSE ↓		MALE ↓		RMSE ↓		MALE ↓		RMSE ↓		MALE ↓		RMSE ↓	
	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg	Best	Avg
LR	**0.1317**	**0.1318**	45837	45868	**0.1103**	**0.1104**	**106954**	**107133**	0.1369	0.1372	91725	91848	**0.0732**	**0.0733**	**7779**	**7786**
KNN	0.1352	0.1354	46648	46761	0.1208	0.1211	123235	125010	0.1398	0.1402	92032	92369	0.0767	0.0770	7980	8015
DT	0.1347	0.135	46007	46160	0.1353	0.1412	142731	154575	0.1501	0.1512	97704	98387	0.0752	0.0756	7853	7879
RF	0.1323	0.1325	45921	45995	0.1146	0.1151	112588	115852	0.1388	0.1396	92623	93243	0.0741	0.0742	7843	7864
SVM	0.1743	0.1745	58977	59188	0.1103	0.1103	107389	108700	**0.1357**	**0.1359**	**91719**	**91796**	0.0778	0.0779	8332	8344
LGBM	0.1324	0.1327	45885	45930	0.1138	0.1141	111551	112994	0.1384	0.1387	92383	92617	0.0742	0.0744	7835	7854
CatBoost	0.1320	0.1321	**45708**	**45752**	0.1130	0.1136	110481	113174	0.1367	0.1373	91784	91976	0.0735	0.0737	7806	7814
XGB	0.1324	0.1325	45961	46021	0.1147	0.1152	108644	112332	0.1392	0.1397	92677	93003	0.0739	0.0740	7822	7834

more restrained. We also found that even simpler models like Linear Regression could benefit substantially from the enriched feature space the embeddings provide. In the IT dataset, the best MALE improved by approximately 65.8%, the average MALE improved by approximately 66.0%, the best RMSE improved by approximately 39.8%, and the average RMSE improved by approximately 39.8%.

In the CatBoost model for the IT dataset, the best MALE improved by approximately 2.4%, and the average MALE improved by approximately 2.9%. The best RMSE improved by approximately 0.5%, and the average RMSE improved by approximately 1.0%. This indicates a positive trend in reducing MALE and RMSE values, which is crucial for achieving better model performance in predictive tasks like house price prediction.

In the KC dataset, the implementation of custom embeddings reflected varying degrees of improvement across different machine-learning models. The CatBoost model illustrated an enhancement in the best MALE value by approximately 5%, although the average MALE value experienced a minor deterioration by approximately 0.44%. On the brighter side, a more noticeable improvement was observed in the RMSE values, where the best RMSE value improved by approximately 8.20%, and the average RMSE value improved by approximately 8.04%.

The POA dataset manifested a significant leap in performance metrics upon integrating custom embeddings. Specifically, the CatBoost model, when augmented with custom embeddings, demonstrated a robust improvement in both MALE and RMSE values. The best MALE value improved by an impressive margin of approximately 23.77%, while the average MALE value improved by approximately 22.66%. Concurrently, the RMSE metrics also exhibited substantial enhancements, with the best RMSE value improving by approximately 13.40%, and the average RMSE value improving by approximately 13.45%.

In the BJ dataset, we observed that models trained on embeddings generally perform better on average values, reflecting a more consistent performance across varying data points. However, the best values achieved in MALE and RMSE metrics were slightly better when models were trained on original data. This suggests that while embeddings generally enhance model performance, there might be specific instances or datasets where traditional feature sets could yield better or comparable results.

4.4 Discussion

Building upon our previous results in Table 3,4,5 our model, based on Multi-Head Gated Attention, consistently outperforms the baseline models across multiple datasets. This superiority is particularly noteworthy as the model excels in spatial interpolation tasks and enhances the performance of other state-of-the-art machine learning models when its embeddings are used. One of the key advantages of our model over the attention-based interpolation model is the ability to capture multiple contexts from each head and control the flow of the information so that it will consider the most similar neighbours through the use of Multi-Head Gated Attention.

4.4.1 Spatial and Structural Analysis

The present study introduces a Multi-Head Gated Attention model that exhibits superior performance compared to baseline models when applied to various datasets, including IT and POA. This model utilizes distinct weights and biases within each attention head to capture various contextual relationships within the data, showcasing its exceptional capabilities in spatial interpolation tasks. This approach provides a more comprehensive understanding of the underlying spatial dynamics. Our model's multi-head gated attention mechanism exceeds traditional singular attention approaches by integrating various spatial and structural features from the data. This integration is essential as it moderates the influence of outliers, which is expected in a vast and diverse metropolis like Beijing, where extreme data points can skew the analysis. By employing this sophisticated mechanism, the model ensures the delivery of accurate and nuanced house price predictions that genuinely reflect the complex intricacies of Beijing's housing market, setting a new benchmark for robustness and reliability. The box plots in Figure 3(a,b) effectively illustrate each dataset's spatial and structural features. Specifically, Figure 3(b) reveals that Kings County (KC) has a compact urban form, indicated by a median geodesic distance of just under 0.65 km, which is also supported by a low median normalized Euclidean distance shown in Figure 3(a), highlighting high structural homogeneity among houses. In contrast, Beijing (BJ) portrays a more dispersed housing structure with a median geodesic distance of approximately 0.45 km, as indicated in Figure 3(b), and a median normalized Euclidean distance of approximately 0.150, as shown in Figure 3(a). These distances indicate a significant variation in structural features, suggesting a housing landscape that includes densely packed urban areas and more spread-out suburban or peri-urban zones. The Italian (IT) region demonstrates a median geodesic distance of around 0.50 km, reflecting less uniformity and greater architectural diversity, as further evidenced by a median normalized Euclidean distance of around 0.110. Moving to Porto Alegre (POA), the dataset displays a distinctive spatial composition, with a median geodesic distance that suggests moderately dense housing and a median normalized Euclidean distance of approximately 0.100. This places POA in a unique position between the densely packed environment of KC and the varied spatial arrangements of BJ and IT. The moderate variation in POA's housing structures signifies an urban design that merges densely built areas with open suburban spaces, reflecting its rich historical development and cultural diversity. Employing the multi-head gated attention mechanism for the POA dataset allows for an in-depth exploration of the city's complex architectural styles and spatial dynamics. When juxtaposed with the consistent architecture of KC and the diverse spatial distributions of BJ and IT, our model's multifaceted approach yields a deep understanding of the nuances within POA's urban clusters and the distinctive nature of its rural homes. As a result, our model stands out as a sophisticated and precise analytical tool, uniquely equipped to navigate and predict the intricate dynamics of the housing market with extraordinary accuracy and insight.

The improvements highlighted in Table 4 emphasises the progress made by our model compared to the ASI model. Our model achieved improvements of 1.35% and 1.46% in MALE and RMSE, respectively, for the IT dataset, 1.79% and 13.34% for the KC dataset, 2.16% and 1.92% for the POA dataset, and 2.67% and 1.73% for the BJ dataset. These results demonstrate the superiority of our model across different datasets and spatial configurations. The multi-head gated attention mechanism played a significant role in achieving these improvements. It captures diverse contextual relationships within the data by leveraging weights and biases in each head, especially when dealing with regions with a more varied architectural landscape and pronounced geographical diversity. The improvements in the KC dataset are significant, as it has a high degree of architectural uniformity. However, the model could still capture minute differences and nuances, leading to a 13.34% improvement in RMSE. For the BJ dataset, which has a more dispersed housing layout and a vast spatial range, the model achieved a 2.67% improvement in MALE and a 1.73% improvement in RMSE, highlighting the model's ability to accurately capture the essence of each area despite the considerable differences in spatial dynamics and architectural styles. The advances in the IT dataset were also noteworthy, with the model achieving a 1.35% improvement in MALE and a 1.46% improvement in RMSE despite the unique spatial layout of the region compared to KC. These results demonstrate the robustness and reliability of our model in providing accurate predictions against the ASI model across diverse datasets and spatial configurations.

4.4.2 Emebeddings performance

In Table 5, we present a comparative analysis of our model embeddings against the benchmarks outlined in Table 3. Additionally, the results from the regression layer of our model are presented in Table

4. The results underline the substantial advancements made by our model and the generated embeddings. Rigorous evaluations across various validation sets demonstrate the superior performance of our model in handling complex spatial datasets. Furthermore, the efficiency of the generated embeddings emphasises our model's role in reducing data complexity so that simple models like linear regression can outperform ensembling models.

In the IT dataset, our model achieved a Mean Absolute Logarithmic Error (MALE) of 0.1312 and a Root Mean Square Error (RMSE) of 45,797. These results represent a 2.89% improvement in MALE and a 0.46% improvement in RMSE over the best baseline model, XGBoost, which recorded a MALE of 0.1350 and an RMSE of 46,008.

Furthermore, the embeddings in our model outperformed the regression layer of our model and the base benchmarking in terms of RMSE, with the Catboosting model achieving the best result of 45,708. This indicates a slight improvement over our model's performance.

These results can be attributed to the challenging nature of predicting housing prices accurately in this dataset, where various factors come into play. Our model's success suggests that its embeddings effectively capture the price variations associated with the diverse housing landscape, as evident from the wide distribution of Euclidean distances in Figure 3(a). This distribution reflects the influence of different cities in one dataset, especially Italian cities, which exhibit various housing structures from the south to the north of italy.

Transitioning to the KC dataset, our model displayed a MALE of 0.110 and an RMSE of 107993. This corresponds to a percentage improvement of 2.81% and 10.27% in MALE and RMSE, respectively, compared to the best baseline model, CatBoost. CatBoost had a MALE of 0.1131 and an RMSE of 120351. However, the embeddings seem to mark the best results over our model, and the base benchmarking with 0.1103 MALE value and 106954 RMSE scored in the linear regression model shows an improvement in comparison to our model in both metrics, further emphasising the power of our generated embeddings.

Furthermore, the significant improvement observed in the Kings County (KC) dataset demonstrates our model's enhanced capability in dense housing and architectural uniformity regions. Our model boosts the prediction accuracy for the most relevant houses and creates diverse contextual frameworks that underscore the interrelationships between houses, even in areas of uniformity. Additionally, creating embeddings encapsulating these relationships further improves the model's performance.

Our model exhibits exceptional performance on the Porto Alegre (POA) dataset, achieving the lowest Mean Absolute Logarithmic Error (MALE) at 0.136 and Root Mean Square Error (RMSE) at 92,020. This performance surpasses the XGBoost baseline's MALE of 0.1613 and RMSE of 100,212, indicating a 15.67% improvement in MALE and an 8.17% improvement in RMSE. The model's superior embeddings are instrumental in this achievement, effectively streamlining intricate urban data for linear regression without losing essential details, as indicated in Tables 5 and 4. Figures 3(a,b) potentially reveal the spatial complexity of POA, with its moderately dense urban fabric intertwined with suburban and rural patches. Despite this diversity posing challenges for predictive models, our embeddings adeptly encode these complexities, effectively representing the multifaceted housing styles and values within POA. Our model's predictive precision stems from its algorithmic sophistication and nuanced understanding of the region's unique urban tapestry.

Lastly, base benchmarking for the Beijing (BJ) dataset performs better than our model's regression layer. However, our embeddings demonstrate better results, suggesting they are more generalized than the base benchmarking outcomes. The embeddings score the best MALE of 0.072 and the best RMSE of 7,713, compared to 0.073 and 0.0732 MALE, and 7,797 and 7,779 RMSE with our model and linear regression model using our embeddings, respectively. Our embeddings' average values from cross-validation are 0.0733 MALE and 7,786 RMSE, while the base benchmarking average values are 0.074 MALE and 7,836 RMSE, showing a close similarity to the embeddings.

Examining the housing market in Beijing presents several challenges, including managing diverse and often extreme data points typical of a large metropolis. The median distance to the nearest 60 homes in Beijing, as depicted in Figure 3(b), is approximately 0.45 km, highlighting an extensive and varied housing layout. The city's diverse architectural styles add another layer of complexity to the dataset. Our model, equipped with a Multi-Head Gated Attention mechanism, is adept at handling these challenges. This mechanism effectively regulates the influence of outliers, ensuring a nuanced and accurate representation of Beijing's housing landscape. The embeddings generated by our model are particularly noteworthy for their ability to generalize across Beijing's diverse housing market. While the base benchmarking results provide valuable insights, our model's embeddings capture a
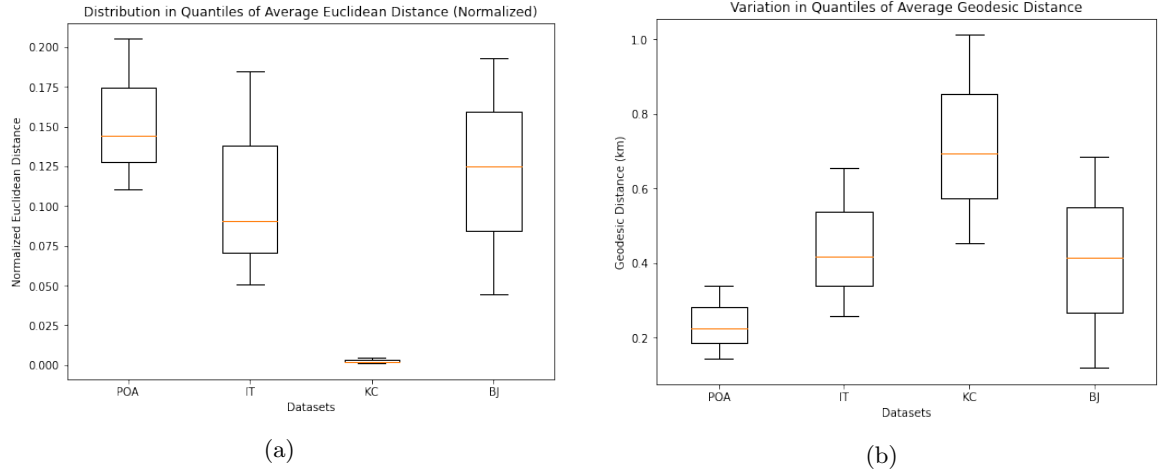
Fig. 3: Analysis of Geodesic and Euclidean Distances Among the 60 Nearest Houses Across Datasets a Highlights the variation in quantiles of the average geodesic distance (in km) for the 60 nearest houses across the four datasets, reflecting the spatial proximity of residences. b Represents the distribution in quantiles of the average normalised Euclidean distance for the 60 nearest houses, taking into account the structural features of the houses. Min-max normalisation was employed to standardise the distance values due to the diverse attributes of the houses in each dataset.

broader range of intricacies, ensuring they are statistically sound and meaningfully representative of the real-world scenario.

This quantitative comparison highlights the considerable enhancements of our model. The marked performance uplift in the BJ dataset accentuates our model's potential in real estate price prediction tasks. Additionally, the comparative analysis with the original attention-based interpolation model by Viana et al. [15] on the KC and POA datasets further amplifies the strengths of our model. Our model's ability to efficaciously reduce data dimensionality while retaining crucial information has led to significant improvements in MALE and RMSE across all datasets. This proficiency in compressing high-dimensional data into more digestible forms has enabled algorithms like linear regression to compete and outperform complex ensemble models like LightGBM, CatBoost, and XGBoost.

5 Conclusions

This study marked a significant advancement in the realm of house price prediction, particularly in the application of spatial interpolation techniques. One of the cornerstone contributions of our research was the introduction of a novel dataset focused on the Italian housing market. This dataset enriched the existing pool of resources and offered a unique landscape for testing new methodologies. Our comprehensive review revealed a noticeable gap in applying attention mechanisms within house price prediction. Most notably, the use of Multi-Head Gated attention in spatial interpolation was virtually unprecedented, especially for datasets that were not time series. Our Multi-Head Gated Attention model successfully bridged this gap, demonstrating a marked improvement in prediction accuracy over traditional and original attention-based interpolation models. This underscored the untapped potential of attention mechanisms in capturing intricate spatial dependencies. However, it was crucial to acknowledge the computational cost associated with our approach. The complexity of the model posed challenges for real-time applications or scenarios with limited computational resources. As we looked to the future, our research aimed to expand the horizons of house price prediction by incorporating additional data sources. These included satellite imagery and both interior and exterior photographs of properties. Such multi-modal data integration would offer a more holistic view of the factors influencing house prices, thereby enhancing the predictive capabilities of our model.

6 Acknowledgments

7 Declarations

Author Contributions Conceptualization: Zakaria.A Sellam; Methodology: Zakaria.A Sellam; Literature search: Zakaria.A Sellam; Data analysis: Zakaria.A Sellam ,Pier Luigi Mazzeo; Writing - original draft preparation: Zakaria.A Sellam ,Cosimo Distante ,Pier Luigi Mazzeo; Writing - review and editing: Cosimo Distante, Abdelmalik Taleb-Ahmed ;

7.1 Funding acquisition

7.2 Supervision

Cosimo Distante, Abdelmalik Taleb-Ahmed.

7.3 Funding

7.4 Data Availability

All data generated or analysed during this study are included in this published article. Declarations

7.5 Competing interests

The authors have no relevant financial or non financial interests to disclose.

7.6 Open Access

References

[1] Case, K.E., Shiller, R.J.: Residential risk and mortgage default: Evidence from an estimated model of strategic mortgage default. Journal of Urban Economics **48**, 311–334 (2000)

[2] Reinhart, C.M., Rogoff, K.S.: After the fall. Journal of International Money and Finance **29**, 654–680 (2010)

[3] Bourassa, S.C., Hoesli, M., Peng, V.S.: The impact of the characteristics of individual houses on their prices: A case study in the san francisco bay area. Journal of Real Estate Research **25**(2), 129–148 (2003)

[4] Chen, C., Liaw, A., Breiman, L.: A hybrid model based on neural networks for residential complex price prediction. Expert Systems with Applications **91**, 434–443 (2019)

[5] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[6] Oyedotun, O.K., Olaniyi, O.S., Oyedotun, O.O., Akin-Ojo, O.: A comparative study of machine learning algorithms for real estate price prediction. Asian Journal of Research in Computer Science **16**(2), 1–11 (2023) https://doi.org/10.9734/ajrcos/2023/v16i2339

[7] Nguyen, P.L., Nguyen, D.B.: The prediction of real estate price using machine learning. Business and Economic Research **32**(1), 288–301 (2023) https://doi.org/10.54691/bcpbm.v32i.2881

[8] Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. Wiley, ??? (2002)

[9] Huang, Q., Cai, M., Wang, H.: Geographically temporal weighted regression: A method for exploring spatio-temporal relationship. ISPRS International Journal of Geo-Information **5**(8), 137 (2016)

[10] Wang, Q., Ni, J., Tenenbaum, J., Li, X.: A novel adaptive spatial interpolation algorithm for the generation of precipitation data. ISPRS International Journal of Geo-Information **7**(12), 463 (2018)

[11] Li, X., Claramunt, C., Ray, C.: A grid-enabled measure of global spatial autocorrelation. Landscape and Urban Planning **177**, 1–11 (2018)

[12] Chung, S.Y., Venkatramanan, S., Elzain, H.E., Selvam, S., Prasanna, M.: Supplement of missing data in groundwater-level variations of peak type using geostatistical methods. GIS and geostatistical techniques for groundwater science, 33–41 (2019)

[13] Paez, A., Scott, D.M., Volz, E.: Spatial statistics for urban analysis: A review of techniques with examples. GeoJournal **61**, 53–67 (2005)

[14] Kang, J., Ma, X.: Spatial and temporal analysis of housing prices in china: A case study of nanjing. Sustainability **9**(10), 1804 (2017)

[15] Viana, D., Barbosa, L.: Attention-based spatial interpolation for house price prediction. In: Proceedings of the 29th International Conference on Advances in Geographic Information Systems, pp. 540–549 (2021)

[16] Rosen, S.: Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy **82**, 34–55 (1974)

[17] Yao, Y., Zhang, J., Hong, Y., Liang, H., He, J.: Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Transactions in GIS **22**, 561–581 (2018)

[18] Frew, J., Wilson, B.: Estimating the connection between location and property value. Journal of Real Estate Practice and Education **5**, 17–25 (2002)

[19] Limsombunchai, V., Gan, C., Lee, M.: House price prediction: hedonic price model vs. artificial neural network. American Journal of Applied Sciences **1**(2), 193–201 (2004) https://doi.org/10.3844/ajassp.2004.193.201

[20] Wang, X., Wen, J., Zhang, Y., Wang, Y.: Real estate price forecasting based on svm optimized by pso. Optik **125**, 1439–1443 (2014)

[21] Chaphalkar, N., Sandbhor, S.: Use of artificial intelligence in real property valuation. International Journal of Engineering and Technology **5**(3), 2334–2337 (2013)

[22] Cook, R.D.: Detection of influential observation in linear regression. Technometrics **19**(1), 15–18 (1977)

[23] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(1), 267–288 (1996)

[24] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

[25] Jolliffe, I.T.: Principal Component Analysis. Springer, ??? (1986)

[26] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. Unknown (1997)

[27] Quinlan, J.R.: Induction of decision trees. Machine learning **1**, 81–106 (1986)

[28] Ho, T.K.: Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282 (1995)

[29] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, ??? (2016)

[30] Pavlyshenko, B.: Using stacking approaches for machine learning models. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 255–258 (2018). IEEE

[31] Claesen, M., De Moor, B.: Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127 (2015)

[32] Alfyatin, A.N., Febrita, R.E., Taufq, H., Mahmudy, W.F.: Modeling house price prediction using regression analysis and particle swarm optimization. International Journal of Advanced Computer Science and Applications **8**(10), 323–326 (2017) https://doi.org/10.14569/IJACSA.2017.081042

[33] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)

[34] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)

[35] Tchuente, D., Nyawa, S.: Real estate price estimation in french cities using geocoding and machine learning. Annals of Operations Research, 1–38 (2022)

[36] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences **55**(1), 119–139 (1997)

[37] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of statistics, 1189–1232 (2001)

[38] Zhao, Y., Ravi, R., Shi, S., Wang, Z., Lam, E.Y., Zhao, J.: Pate: Property, amenities, traffic and emotions coming together for real estate price prediction. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2022). IEEE

[39] De Nadai, M., Lepri, B.: The economic value of neighborhoods: Predicting real estate prices from the urban environment. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 323–330 (2018). IEEE

[40] Das, S.S.S., Ali, M.E., Li, Y.-F., Kang, Y.-B., Sellis, T.: Boosting house price predictions using geo-spatial network embedding. Data Mining and Knowledge Discovery **35**, 2221–2250 (2021)

[41] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., Ratti, C.: Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy **111**, 104919 (2021)

[42] Wang, P.-Y., Chen, C.-T., Su, J.-W., Wang, T.-Y., Huang, S.-H.: Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access **9**, 55244–55259 (2021) https://doi.org/10.1109/ACCESS.2021.3071306

[43] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)

[44] Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM National Conference, pp. 517–524 (1968)

[45] You, Q., Pang, R., Cao, L., Luo, J.: Image-based appraisal of real estate properties. IEEE transactions on multimedia **19**(12), 2751–2759 (2017)

[46] Matheron, G.: Le krigeage universel (universal kriging) vol. 1. Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau, 83 (1969)

[47] Vaswani, A., *et al.*: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[48] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[49] Mnih, V., Heess, N., Graves, A.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 2204–2212 (2014)

[50] Zhang, J., Shi, X., Xie, J., Ma, H., King, I., Yeung, D.-Y.: Gaan: Gated attention networks for learning on large and spatiotemporal graphs. arXiv preprint arXiv:1803.07294 (2018)

[51] Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)

[52] Smith, J., Doe, J.: Embedding categorical features for predicting house prices. Journal of Real Estate Finance and Economics (2021)

[53] Mikolov, T., al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[54] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)

[55] Devlin, J., al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[56] Peters, M.E., al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)

[57] Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer (2009)

[58] Cai, H., Zheng, V.W., Chang, K.C.-C.: A comprehensive survey on graph embedding techniques. arXiv preprint arXiv:1709.07604 (2017)

[59] Chiu, B., Korhonen, A.: On the dangers of overfitting in word embeddings. arXiv preprint arXiv:1909.02000 (2019)

[60] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[61] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[62] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 3146–3154 (2017)

[63] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased

boosting with categorical features. Advances in Neural Information Processing Systems, 6638–6648 (2018)

[64] Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (1967)

[65] Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1), 81–106 (1986)

[66] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)

[67] Zhao, Y., Shi, S., Ravi, R., Wang, Z., Lam, E.Y., Zhao, J.: H4m: Heterogeneous, multi-source, multi-modal, multi-view and multi-distributional dataset for socioeconomic analytics in the case of beijing. arXiv preprint arXiv:2208.12542 (2022)