Real Estate Price Prediction

Rabia Naz[1*], Bushra Jamil[2], Humaira Ijaz[2]
[1]Department of Software Engineering, University of Sargodha, Sargodha, Pakistan.
[2]Department of IT, University of Sargodha, Sargodha, Pakistan.
[*]**Correspondence:** rabianaz935@gmail.com

Citation | Naz. R, Jamil. B, Ijaz. H, "Real Estate Price Prediction", IJIST, Vol. 6 Issue. 2 pp 1031-1044, July 2024
Received | July 05, 2024 **Revised |** July 23, 2024 **Accepted |** July 24, 2024 **Published |** July 25, 2024.

Real estate price predictions are critical for stakeholders, including investors and developers, because they have a considerable impact on investment decisions and market stability. In order to fill in the shortcomings in earlier approaches, this work presents a novel methodology by utilizing Deep Learning (DL) and Machine Learning (ML) techniques to improve real estate price forecast accuracy. We used the "House Prices 2023 Dataset" from Kaggle, which contains 168,000 entries of Pakistani property data. Our methodology included extensive data preparation, feature engineering, and the use of various algorithms, including Linear Regression, Gradient Boosting, Random Forest, Convolutional Neural Networks (CNN), and K-Nearest Neighbors (KNN). The models were tested using MSE, RMSE, R-squared, and accuracy. KNN outperformed the other models, with a lower RMSE of 13.79 and a higher R-squared value of 0.85, indicating improved predictive accuracy. RF also produced impressive results, with an accuracy of 80%. Handling complicated feature interactions, guaranteeing model scalability, and controlling hardware resources were all challenges that suggested possibilities for future improvement. As a result, our research offers a solid foundation for raising forecasting accuracy in fluctuations in the market and emphasizes the possibility of utilizing ML approaches for better real estate price prediction.

Keywords: Real Estate; Machine Learning; Deep Learning; Market Dynamics; Investment Analysis

Introduction:

Real property encompasses all interests, benefits, liberties, and constraints associated with owning real estate, such as land, permanent improvements, and associated equipment [1]. The real estate market experienced significant fluctuations and steady expansion during the pandemic [2]. A generic index of real estate total returns rose about seventeen times between 1947 and 1982 [3]. In 2019, real estate contributed 7.62% of GDP, potentially reaching 13.6% when considering indirect impacts on the construction industry [2]. Shiller observed that residential investment reached its highest level since 1950 in the fourth quarter of 2005, at 6.3% of GDP [4].

The success of the real estate industry hinges on core competencies that bolster long-term competitive advantages. These competencies include specialization, adaptability, client value, and responsiveness [5]. The real estate industry is unique because its "product" is not convertible [6]. The real estate market is defined by extreme heterogeneity due to physical characteristics and location, resulting in large transaction costs, carrying expenses, illiquidity, tax issues, and high search fees [7]. Real estate investment can be direct, involving tangible properties, or indirect, involving financial exposure through REITs traded on stock markets [8]. Market efficiency concerns lenders, investors, politicians, homeowners, and scholars. The joint hypothesis issue and the Grossman and Stiglitz paradox question the reliability of market efficiency and the possibility of total informational efficiency [9].

The research question of our study is the absence of real estate price forecasting models that are capable of addressing uncertainty and heterogeneity in the markets. The major limitation that affects most current forecasting models is their failure to integrate multiple sources of data and adjust to market volatilities. This study aims at filling this gap through applying progressive techniques of ML and DL together with data preprocessing and feature selection to improve the predictive performance in the real estate market. We are interested in increasing the forecasting precision of real estate prices and making it more resistant to changes in market conditions and input data types.

Objectives of the Study:

The main research question focuses on the need to have better real estate price forecasting models that can effectively look at numerous and diverse markets, and combined datasets. We aim to enhance forecasting accuracy by utilizing Machine Learning (ML) and Deep Learning (DL) techniques, alongside thorough data preprocessing and feature selection. Our specific objectives are to:

- Identify the most significant features affecting real estate prices and optimize their selection and engineering for improved forecasting accuracy.
- Evaluate the effectiveness of ML and DL algorithms in predicting real estate prices using a comprehensive dataset.
- Provide better estimations of property values to reduce investment uncertainty and improve capital investment strategies.

We hypothesize that:

- Advanced feature engineering will significantly improve prediction model performance.
- Advanced ML and DL algorithms will achieve superior accuracy compared to traditional methods.
- Improved forecasting accuracy will lead to better estimations of property values, thereby reducing investment uncertainty and enhancing capital investment strategies.

Ultimately, our goal is to enhance the accuracy and reliability of real estate price forecasts and advance the field of real estate analytics.

Literature Review:

Significant research has been conducted to develop accurate methods for predicting real estate prices. The following literature review outlines key studies and methodologies in this field. Ziweritin et al. [10] addressed three primary approaches: polynomial regression with features raised to various powers, multivariate regression models utilizing multiple features, and linear regression with square feet as a variable. The multivariate regression model using square feet, bedrooms, and bathrooms produced the best results because it considered multiple influential factors, leading to more accurate predictions. Lee et al. [11] studied ML applications, such as linear regression, SVR, KNN, and random forest in real estate price prediction. Linear regression achieved the lowest prediction error at 0.3713 due to its simplicity and effectiveness in capturing the linear relationship between features and prices. Ho et al. [12] used three machine learning techniques: GBM, random forest, and SGD-based SVR. The accuracy and predictive power of these algorithms were highlighted by the researchers when they compared them to forecast property prices. The computational economy of SVM was emphasized, whereas RF and GBM showed better prediction accuracy with fewer mistakes because these algorithms can handle complex patterns and interactions between features. The study came to the conclusion that ML, particularly RF and GBM, has potential for precise property price forecasting. Al Kurdi et al. [13], demonstrated that the decision tree achieved exceptional accuracy, with True Negative Rates (TNR) and True Positive Rates (TPR) exceeding 92%. The decision tree performed well due to its ability to model non-linear relationships and interactions among features effectively.

Xu et al. [14] predicted home prices using deep learning approaches, including CNNs. Some of the features used for prediction were the type of housing, the building area, the location, and macroeconomic indicators such as GDP, property asset, and consumption level. The CNN model's effectiveness was confirmed by the testing results, which showed an accuracy of 98.68% and a mean square error of 0.01055. CNN excelled because it can capture spatial hierarchies and complex patterns in the data. Das et al. [15] integrated ML techniques with a GSNE to collect and combine neighborhood information for price prediction. The integration of neighborhood information improved the model's performance by adding valuable contextual data.

Table 1. Summary of Previous Studies on Real Estate Price Prediction: Algorithms Used, Performance Metric and Best Declared Algorithm/s

Ref.	Algorithms Used	Performance Metric	Best Algorithm Declared
[10]	Linear Regression, MR, PR	RMS	Multivariate regression
[11]	Linear Regression, SVR, KNN, RF Regression	PE	Linear regression
[12]	SGD-based SVR, RF, GBM	EM	RF, GBM
[13]	DT, RF, AdaBoost, NB, Logistic Regression	Accuracy (TPR, TNR)	Decision Tree
[14]	CNN	MSE	CNN
[15]	GSNE + ML	Accuracy	GSNE + ML
[16]	MLP + ARIMA	MAE	MLP+ARIMA
[17]	Regression Models, Hybrid Models, ML Models	MAE	Multi-kernel DL regression
[18]	Light GBM	MAPE	Light GBM
[19]	RF Classification, DT, NB, Linear Regression	Prediction Effectiveness	Linear regression

Nouriani et al. [16] used a hybrid methodology combining DL and time series forecasting approaches. They used a DL model with four hidden layers and an ARIMA model to forecast trends in real estate values. The hybrid approach succeeded because it leveraged the strengths of both deep learning for capturing complex patterns and ARIMA for time series trends. Yousif et al. [17] compiled research using various methods like regression, hybrid models, and ML models, suggesting a new benchmark dataset REPD-3000. The researchers found their multi-

kernel DL regression model to be more efficient than others. The multi-kernel approach allowed the model to capture different types of data patterns effectively. Li et al. [18] used the LightGBM framework, concluding that attaching geodata, logarithm, and apartment brandence resulted in the lowest MAPE. LightGBM performed well because of its ability to handle large datasets with efficiency and accuracy.

Gampala et al. [19] used supervised learning techniques, including DT, NB, linear regression, and RF classification, finding linear regression to be the most accurate for predicting house prices. Linear regression was effective because it provided a straightforward and interpretable model for the relationship between features and house prices. Table 1 integrates findings from several studies, providing an overview of methodologies, algorithms, and performance measures for evaluating real estate price predictions.

Paper Organization:

There is a comprehensive breakdown of real estate price prediction containing ten sections within this study. The significance of real estate price forecasting stressed in the introduction (section 1) with reference to its impact on the involved stakeholders and the economy. Section 2 provides the study's objective that specifically identifies the aims and scope of our research work. Section 3 focuses on related work which includes a summary of the existing literature and studies on real estate price prediction, as well as the technique and tactics used by researchers. The whole methodology, including data collection, preprocessing procedures, and the standard by which the appropriate prediction model was chosen, is described in Section 5. In section 6, there are elaborate explanations of the algorithms used which include Linear Regression, Gradient Boosting, Random Forest, CNN, and KNN.

Section 7 presents the findings and discussion, which analyses and evaluates each algorithm in terms of performance index and evaluates its performance to predict real estate values more accurately. In Section 8, we integrate our research findings with existing literature to highlight how our results align with or extend current knowledge in real estate price prediction. In the conclusion (Section 9), the study's important findings are presented, stressing contributions to the field and offering areas for further research. The paper is supported by a complete list of references (Section 10), proving the study's legitimacy and academic integrity.

Material and Methods:

The methodology section provides a detailed explanation of the systematic approach and procedures used in this study. It covers the steps taken to estimate real estate values, including data collection, data preprocessing, modeling, and the analysis of results and discussion.

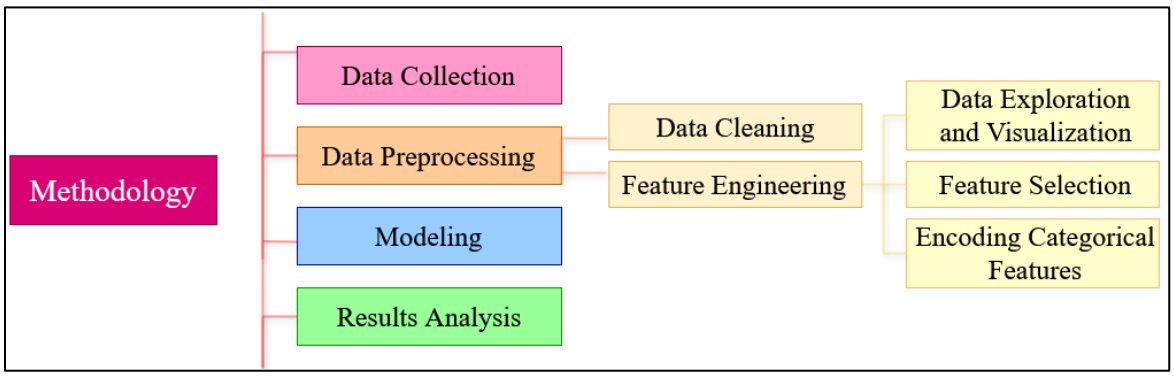

Figure 1. Methodology Used in the Study

Figure 1 depicts the flow of methodology, used to estimate real estate values; divided into four major branches: data collection, data preprocessing, modeling, and results analysis. The data collecting stage is responsible for gathering relevant data from a variety of sources. The data processing stage is divided into two phases: data cleaning and feature engineering. Data cleaning is dealing with missing numbers, deleting duplicates, and fixing errors to assure data

quality. Feature engineering consists of three important activities: data exploration and visualization, feature selection, and category feature encoding. Data exploration and visualization provide information on data distribution and linkages. Feature selection determines the most relevant attributes for the model, which improves its performance. Encoding categorical features converts non-numeric data into a numerical representation appropriate for modeling. Selecting and training suitable machine learning and deep learning algorithms is the next step in the modeling branch after preprocessing. Lastly, the results and discussion section analyses performance of models and discusses the effect of results on prediction of real estate prices. Results analysis is covered in the section 7.

Data Collection:

The dataset used in this study, named "House Prices 2023 Dataset" was sourced from Kaggle and compiled from the Zameen.com website. It contains 168,000 entries that detail various real estate properties in Pakistan. This comprehensive database provides 20 attributes of the property characteristics, geographical aspect, and the execution process. Table 2 highlights all the attributes included in this dataset.

Table 2: Attributes of Real Estate Dataset

S.no	Variables	Data Type
1	property_id	Integer
2	location_id	Integer
3	page_url	String
4	property_type	String
5	price	Integer
6	location	String
7	City	String
8	province_name	String
9	latitude	Real
10	longitude	Real
11	baths	Integer
12	area	String
13	purpose	String
14	bedrooms	Integer
15	date_added	String
16	agency	String
17	agent	String
18	Area Type	String
19	Area Size	String
20	Area Category	String

Data Preprocessing:

The data preprocessing describes the essential steps taken to prepare the raw data for modeling. This involves data cleaning to ensure data quality and feature engineering to enhance the dataset's predictive power. By transforming and refining the data, this stage lays the foundation for effective model training and accurate real estate value estimation.

Data Cleaning:

The quality of the dataset was assessed through detailed documentation provided by its owner. The owner performed extensive data preprocessing and created a new version called "Cleaned_data_for_model" which is used in this research. He also shared the notebook on Kaggle, documenting the entire data cleaning process and how the cleaned version of the dataset was created. The training dataset from Kaggle required no additional cleaning, as the dataset owner had provided multiple versions, including a cleaned version. Consequently, no further cleaning procedures were necessary for the dataset used in the experiments.

Feature Engineering:

Feature engineering involves transforming and manipulating raw data to generate useful features that improve the model's predicted performance. When predicting real estate values with the supplied dataset, the feature engineering process may be split down into three distinct steps:

Data Exploration and Visualization:

In data visualization, scatter plots show outliers and the association between two variables by placing each observed dataset on the cartesian grid. Figure 2 illustrates the relationship between bath and price, showing a positive correlation: as baths (x-axis) increase, so does price (y-axis). It also highlights one outlier with 400 baths.

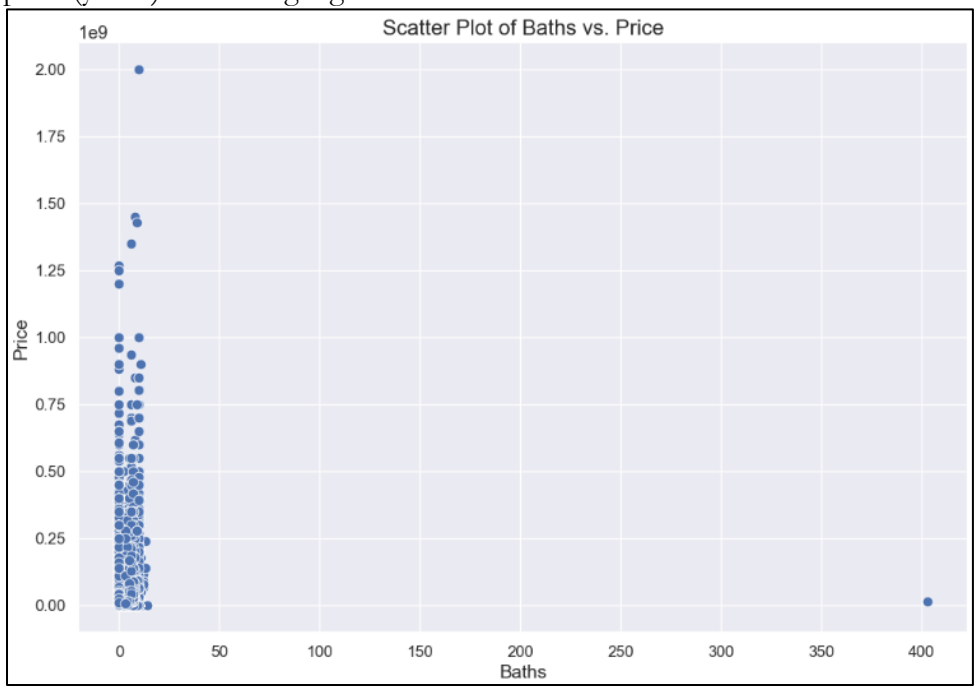

Figure 2. Scatter plot of Baths vs. Price

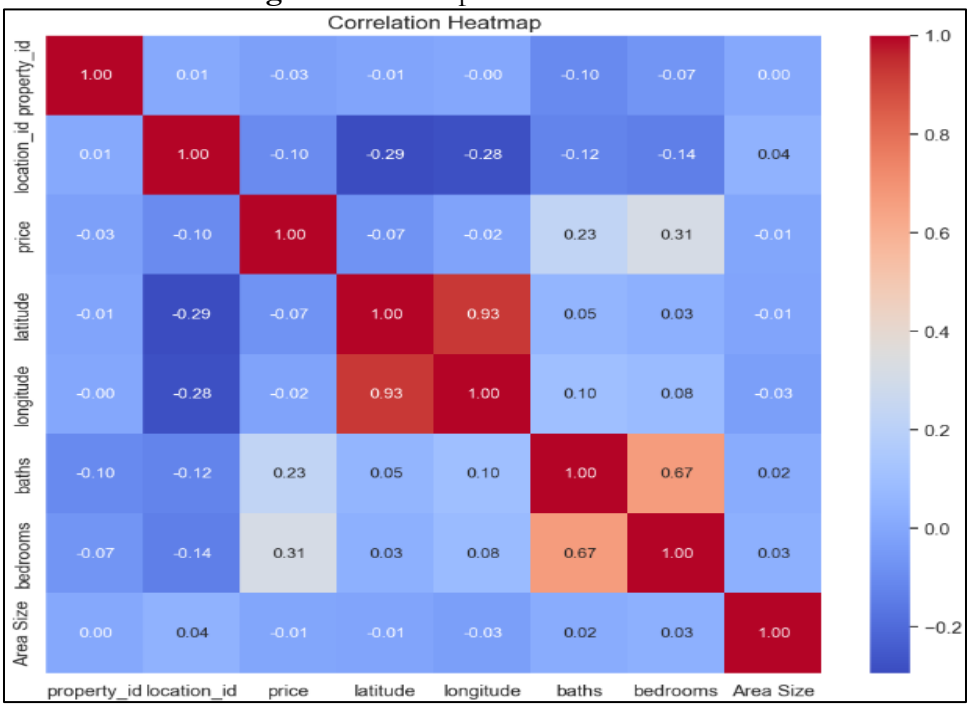

Figure 3. Correlation Heatmap

In figure 3, a heatmap illustrates the influence of pricing factors. It highlights that the number of bedrooms significantly impacts price, with a coefficient of 0.31. Relationships between other factors are also visible; for instance, increasing bathrooms correlates positively with bedrooms by 0.67. Figure 4 utilizes a count plot to depict the frequency of each property type (flat, house, penthouse, farmhouse, lower section, upper portion, and room). It shows that houses appear over 100,000 times, while flats are around 40,000 times, each represented by independent bars. Figure 5 illustrates price distributions for property types. The x-axis shows property types, and the y-axis shows prices. Vertical bars represent price ranges, with the tallest bar indicating more flats available than other types.

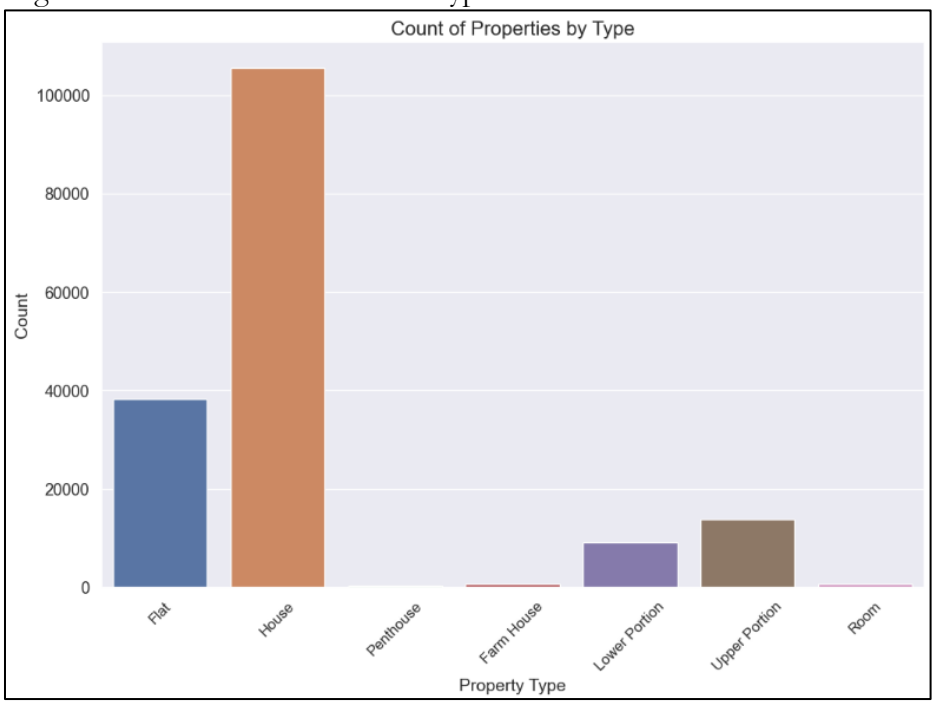

Figure 4. Count Plot

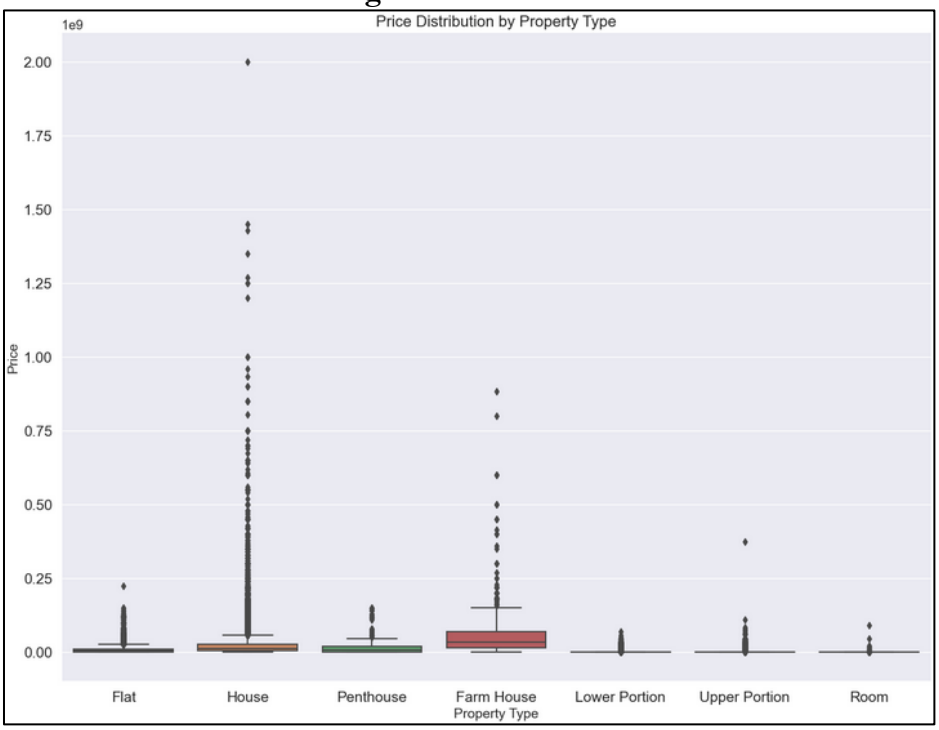

Figure 5. Box Plot of Price Distribution by Bedrooms

Feature Selection:

Following the above exploratory data analysis, we proceeded on a careful feature selection process to determine the most relevant elements determining property pricing. Feature selection criteria were based on both domain knowledge and empirical analysis. Key factors influencing feature selection include:

- **Relevance to Price Prediction:** Features with a strong theoretical or empirical relationship to property prices were prioritized. For example, attributes such as the number of bedrooms, bathrooms, and property area were selected due to their established impact on real estate values.
- **Correlation Analysis:** Features were evaluated for their correlation with the target variable (price). High correlation coefficients, such as those observed between the number of bedrooms and price, indicated their importance.
- **Feature Importance Scores:** Algorithms like RF and GB were used to assess feature importance. Features with high importance scores were retained for modeling.

Using above feature selection criteria, we found a subset of factors with a high potential to influence housing prices. Among these variables, bathrooms, area, and bedrooms were found to be intuitively associated to property price. This revised feature set served as the foundation for further modelling efforts, ensuring that our predictive models were based on strong, data-driven insights acquired from exploratory analyses.

Several datasets can complement the "House Prices 2023 Dataset" to enhance our analysis:

- **Economic Indicators:** Interest rates, inflation rates and GDP growth datasets can help put into perspective the macroeconomic fundamentals that accompany real estate prices.
- **Demographic Data:** Quantitative data such as population growth, mean household income, and employment rates can depict demand side forces affecting real estate prices.
- **Geospatial Data:** Additional aspects such as features in the particular neighborhoods, crime density, and quality of schools can enhance location-wise price estimates.
- **Historical Sales Data:** Other historical datasets concerning the sales prices and trends can be incorporated in the model to increase predictive power by capturing long-term trends.

Encoding Categorical Features:

Feature engineering plays a crucial role in machine learning, with encoding targeted properties being a key component. This involves transforming the property type feature into a format that enables learning algorithms to function effectively. To achieve this process, one of the methods called one hot encoding is used to encode categorical data so as to have new features where every category has its own feature that indicates whether or not that category exists. It is the encoding technique in which all the categorical data was converted, and each category in a categorical feature was represented by its binary variable. The categories “apartment”, “house”, and “condo” were encoded into binary variables in the property type feature where 1 indicates the existence of the category and 0 indicates the absence of such a category. This transformation enables the algorithms to differentiate as well as learn from categorical features in addition to numerical characteristics.

Algorithms Used:

In our research study, we adopted a combined approach to address a complex problem, utilizing multiple ML and DL techniques. This integration improved the experimental process and returned precise results. These are the five algorithms performed which involve linear regression, gradient boosting, random forest, CNN, and KNN. The study ensures the chosen algorithms are suitable for the dataset through the following steps:

- **Exploratory Data Analysis (EDA):** To analyze the distributions, correlations, and outliers for the selected dataset, we performed EDA. This analysis helped in the selection of algorithms that are more in line with the data structure.
- **Feature Engineering:** Some features were deliberately chosen to be engineered because they were needed to give the algorithms necessary data, relevant for real estate prices.
- **Algorithm Comparison:** We used 7-9 algorithms like Gradient Boosting, KNN, CNN etc. We drew a comparison between them by evaluating them under accuracy, RMSE, R-square, and MSE. Out of these, we chose top 5 algorithms based on their performance.

In the next subsections, we will discuss each algorithm and how it fits into our research study.

Linear Regression:

Linear regression is used to establish a linear relationship between one or more parameters. Simple regression and MLR are the two types of linear regression. When two variables have a linear connection, a straightforward linear regression model is used. However, when there is a linear relationship between two or more independent variables, an MLR model is used, and when there is a polynomial relationship between the variables, a polynomial regression model is used [20].

Gradient Boosting:

Gradient boosting is a learning method that uses basic predictors to construct a set of diverse strong learners that can perform better than any single learner in the set. The most used base learner is the fixed size decision tree [21]. Like boosting, gradient boosting is also known to perform well in regression problems [22]. It is very flexible in terms of the specific application needs and handles different loss functions as well [23].

Random Forest:

Random forest is a combination of classification or regression trees and has been widely applied, for example to SDM [24]. It is an ML algorithm derived from decision trees and bagging [25]. It is an excellent tool for the exploration of high-dimensionality datasets [26].

Convolutional Neural Networks:

Neural networks consist of units, akin to neurons, connected by adjustable strengths in a learning process. Each unit integrates information from its synapses to determine its activation status, influencing whether its response is linear or nonlinear [27]. CNN is increasingly utilized for prediction, grouping, pattern recognition, and classification across various domains, surpassing traditional regression and statistical models [28]. It mimics the brain's operation with numerous simple processors interconnected by weighted connections. Nodes, akin to neurons, process information received locally or through these connections, influencing their individual outputs [29]. Outputs are generated by output units, with hidden units positioned between input and output layers [30].

K-Nearest Neighbor:

K-nearest neighbor (KNN) is supervised machine learning algorithm mainly applied to predict classes of hitherto unknown data points [31]. It offers a better solution to the problem of approximation which initially was introduced in statistics in the form of nonparametric discrimination [32]. It has been applied widely and successfully to many fields such as text classification, pattern recognition and image processing [33].

Results and Discussions:

To assess the performance of the model once it had been trained, we utilized metrics such as Root Mean Square Error (RMSE), Mean Square Error (MSE), R-squared, and accuracy. These metrics help to determine the adequacy of the model's prediction of the events that occurred in a given period of time.

For gradient boosting, we employed decision trees where multiple trees or stages were built in a way that each new tree worked to reduce errors made by the previous trees. To increase the efficiency of the gradient boosting method, the hyperparameters like learning rate, number of trees, and other parameters were optimized using grid search and random search methods. Parameters like learning rate, tree depth and tree number were tuned to get the best set of values for minimizing the generalization errors. To prevent overfitting, where the ML model memorizes the training data and fails to generalize to new data, we incorporated early stopping. Table 3 explains the details of the performance of five models that have been applied in this current research. The performance measures used are MSE, RMSE, R-squared, and accuracy. These metrics give an indication of how far off each of the models are from predicting the real estate prices to perfection, where the smaller values of MSE and RMSE are preferred, higher values of R-squared for better explanation and higher values of accuracy percentage for more precise model. These metrics provide an indication of how accurately the models predict real estate prices. Lower values of MSE and RMSE indicate better performance, as they reflect smaller differences between predicted and actual prices. Higher R-squared values are preferred as they signify a better fit of the model to the data, explaining more of the variance. Additionally, higher accuracy percentages denote a more precise model in terms of correct predictions.

Table 3. Models Performance Metrics

Model	MSE	RMSE	R-Squared	Accuracy
Linear Regression	2.5e+4	15.81	0.78	78%
Random Forest	2.2e+4	14.83	0.80	80%
Gradient Boosting	2.4e+4	15.49	0.79	79%
CNN	6.0e+4	24.49	0.60	60%
KNN	1.9e+4	13.79	0.85	86%

The correlation matrix between actual and predicted prices is displayed in the form of a heatmap in Figure 6. The heatmap illustrates the correlation between actual and predicted prices, with darker shades indicating a stronger correlation. This visualization is useful when trying to understand how well each model performs in terms of capturing association between variables and identifying aspects such as accuracy of predictions and potential bias that may exist in the models.

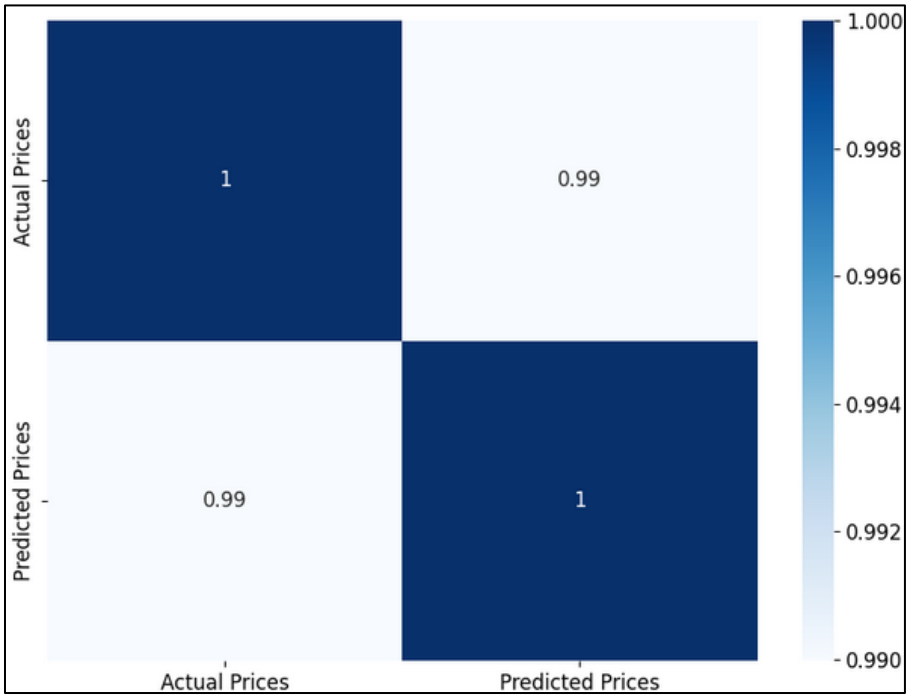

Figure 6. Heatmap of Correlation Matrix

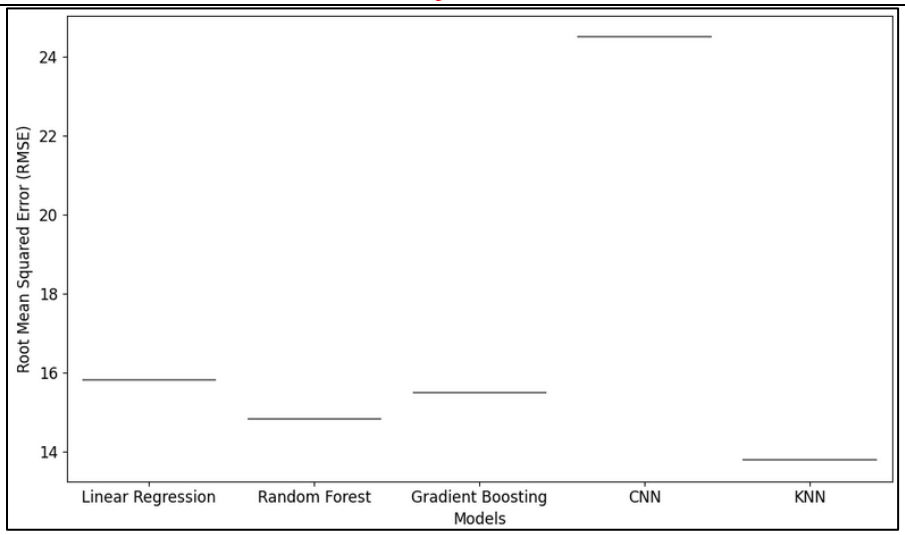

Figure 7. Plot of Models Performance

Figure 7 shows violin plot which summarizes the performance of the five models in terms of RMSE. Each line of the violin plot corresponds to the distribution of RMSE values, where the width of the violin indicates variability. This plot gives a measure of the ability of each model with respect to the prediction errors.

In Figure 8, the bar chart displays the R-squared values of the four models used. R-squared quantifies the extent to which the variation in the dependent variable (real estate prices) can be explained by the variation in the independent variables (features used in the models). A model with a higher R-squared value is better, as it means it can predict a larger percentage of the price variance.

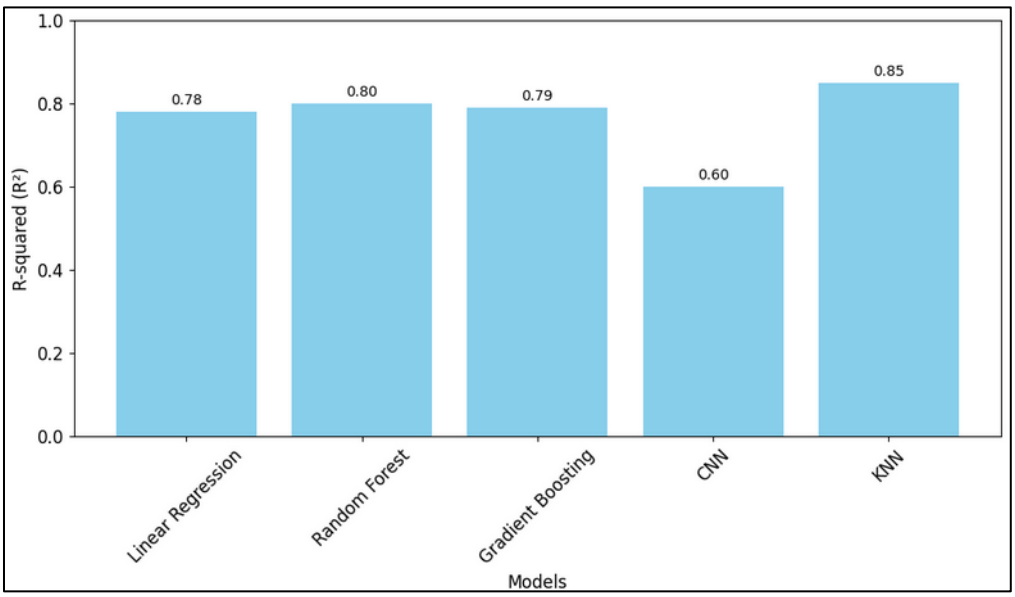

Figure 8. R-squared Values of Models

When comparing the models used for real estate price prediction, the KNN model proved to be the best, achieving the smallest RMSE of 13.79 and the highest coefficient of determination (R-squared) of 0.85. It showed strong performance when tested against real estate prices against linear regression, gradient boosting, random forest, and CNN. Our study faced some of the challenges that include:

- **Complex Feature Relationships:** One of the problems was to handle interactions between different attributes that include location, size of property, and facilities. To explain these interactions, one had to incorporate rather complex models to trace them.

- **Scalability of Models:** One of the major difficulties was to use models that could be easily scaled with the size of the dataset. There was also the consideration of always having to ensure that the models could take large data and still perform well and with correct results.
- **Hardware Resources:** The dataset was a large one and hence it was not easy to perform effective deep learning algorithms since it needs a lot of computation power to run. These models took hours to compile and made it clear that strong hardware resources were needed to process the data in the models.

This paper's limitations present opportunities for future expansions which may include investigation on hardware advancements, new optimized algorithms or cloud-based solutions on how to overcome them and boost the accuracy and effectiveness or real estate price forecasts. The study can be extended in these directions for improving the reliability, effectiveness, and generalizability of real estate pricing models in various and rapidly changing environments.

Integrating Our Research Findings with Existing Literature:

The understanding of our study directly impacts real estate price predicting, where KNN stands out as the most efficient of examined algorithms. The results show that the proposed KNN model outperforms the other models by possessing the lowest MSE, alongside the highest accuracy. This outcome deviates from prior studies, wherein RF and GB are credited for high accuracy prediction [12][19]. For example, Ho et al. [12] have discussed how good Random Forest performs, Researcher [19] described the efficiency of Gradient Boosting. However, in our study, KNN showed a better accuracy than these complex models, which indicates that the performance of KNN is particularly best suited for the nature of our dataset. This goes against the established knowledge saying that models with high degrees of complexity yield improved efficiency due to their capacity to handle intricate and nonlinear data patterns as demonstrated by KNN in this case.

In addition, our research fills the gap in previous research where most research work on ensemble methods and deep learning techniques. Therefore, it is advised that future studies should reevaluate traditional algorithms with extended modifications. Ideas for future work include improving feature extraction, optimizing both traditional ML approaches and the hybrid approach, and conducting more experiments using different datasets to verify and extend on the results presented in this work.

Conclusion:

Based on the analysis carried out in the present study, it can be seen that the performance of a large variety of ML algorithms is quite satisfactory especially regarding real estate price prediction. In order to effectively and purposefully undertake this goal of making the existing and the emergent models of real estate valuation more precise and accurate, we initiated a carefully and strategically designed process. To find trends in the data sets, feature extraction and identification of patterns from the previous data was done. We implemented five algorithms including Linear Regression, Gradient Boosting, Random Forest, CNN, and KNN. From the analysis, it was found that KNN outperformed other models in terms of accuracy and stability while working with large datasets. Our study highlights the importance of accurately approximating real estate prices to improve decision-making when buying or investing. We demonstrate how traditional machine learning techniques can address issues like data heterogeneity and market volatility. The aim is to enhance stability, productivity, and transparency in the real estate sector, both nationally and internationally, by providing stakeholders with improved predictive tools. This contributes to sustainable advancement and development in the real estate industry.

Author's Contribution: All authors contributed equally.

Conflict of Interest: There exists no conflict of interest for publishing this manuscript in IJIST.

Project Details: This project completed in June 2023, and resulted in a web-based application for real estate purposes.

References:

[1] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, "House price prediction using hedonic pricing model and machine learning techniques," Concurr. Comput. Pract. Exp., vol. 34, no. 27, p. e7342, Dec. 2022, doi: 10.1002/CPE.7342.

[2] H. Vu Minh, T. Nguyen Hoang, D. Le Doan Minh, and N. Nguyen Minh, "The relevance of factors affecting real estate investment decisions for post pandemic time," Int. J. Bus. Glob., vol. 1, no. 1, p. 1, 2022, doi: 10.1504/IJBG.2022.10056378.

[3] D. Mahmoudinia and S. M. Mostolizadeh, "(A)symmetric interaction between house prices, stock market and exchange rates using linear and nonlinear approach: the case of Iran," Int. J. Hous. Mark. Anal., vol. 16, no. 4, pp. 648–671, Aug. 2023, doi: 10.1108/IJHMA-01-2022-0008/FULL/XML.

[4] D. Broxterman and T. Zhou, "Information Frictions in Real Estate Markets: Recent Evidence and Issues," J. Real Estate Financ. Econ. 2022 662, vol. 66, no. 2, pp. 203–298, Aug. 2022, doi: 10.1007/S11146-022-09918-9.

[5] J. Phangestu and S. Tinggi Manajemen PPM Jakarta, "The Effects of Resource Management to Firm's Performance Through Exploratory and Exploitative Innovations (An Empirical Research on Real Estate Developer Firms in Indonesia)," pp. 128–133, Aug. 2020, doi: 10.2991/AEBMR.K.200812.023.

[6] K. W. D. Hoang, "Machine learning methods in finance: Recent applications and prospects," Eur. Financ. Manag., vol. 29, no. 9, pp. 1657–1701, 2023, [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/eufm.12408

[7] S. H. Zulkarnain, A. S. Nawi, M. A. Esquivias, and A. Husin, "Determinants of housing prices: evidence from East Coast Malaysia," Int. J. Hous. Mark. Anal., vol. ahead-of-print, no. ahead-of-print, 2024, doi: 10.1108/IJHMA-10-2023-0139/FULL/XML.

[8] N. Baptista, J. F. Januario, and C. O. Cruz, "Social and Financial Sustainability of Real Estate Investment: Evaluating Public Perceptions towards Blockchain Technology," Sustain. 2023, Vol. 15, Page 12288, vol. 15, no. 16, p. 12288, Aug. 2023, doi: 10.3390/SU151612288.

[9] D. Broxterman, D. Gatzlaff, M. Letdin, G. S. Sirmans, and T. Zhou, "Introduction to Special Issue: Topics Related to Real Estate Market Efficiency," J. Real Estate Financ. Econ., vol. 66, no. 2, pp. 197–202, Feb. 2023, doi: 10.1007/S11146-022-09928-7/METRICS.

[10] S. Ziweritin, C. Chimezie Ukegbu, T. A. Oyeniran, and I. O. Ulu, "A Recommendation Engine to Estimate Housing Values in Real Estate Property Market," Int. J. Sci. Res. Comput. Sci. Eng., vol. 9, no. 1, pp. 1–7, 2021, doi: 10.26438/ijsrcse/v9i1.17.

[11] S. H. Lee, J. H. Kim, and J. H. Huh, "Land Price Forecasting Research by Macro and Micro Factors and Real Estate Market Utilization Plan Research by Landscape Factors: Big Data Analysis Approach," Symmetry 2021, Vol. 13, Page 616, vol. 13, no. 4, p. 616, Apr. 2021, doi: 10.3390/SYM13040616.

[12] W. K. O. Ho, B. S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," J. Prop. Res., vol. 38, no. 1, pp. 48–70, Jan. 2021, doi: 10.1080/09599916.2020.1832558.

[13] B. Al Kurdi, H. Raza, S. Muneer, M. B. Alvi, N. Abid, and M. T. Alshurideh, "Estate Price Predictor for Multan City Townships Using Marching Learning," Int. Conf. Cyber Resilience, ICCR 2022, 2022, doi: 10.1109/ICCR56254.2022.9996072.

[14] X. Xu and Y. Zhang, "Residential housing price index forecasting via neural networks," Neural Comput. Appl. 2022 3417, vol. 34, no. 17, pp. 14763–14776, May 2022, doi: 10.1007/S00521-022-07309-Y.

[15] S. S. S. Das, M. E. Ali, Y. F. Li, Y. Bin Kang, and T. Sellis, "Boosting house price predictions using geo-spatial network embedding," Data Min. Knowl. Discov., vol. 35, no. 6, pp. 2221–2250, Nov. 2021, doi: 10.1007/S10618-021-00789-X/METRICS.

[16] A. Nouriani and L. Lemke, "Vision-based housing price estimation using interior, exterior & satellite images," Intell. Syst. with Appl., vol. 14, p. 200081, May 2022, doi: 10.1016/J.ISWA.2022.200081.

[17] A. Yousif, S. Baraheem, S. S. Vaddi, V. S. Patel, J. Shen, and T. V. Nguyen, "Real estate pricing prediction via textual and visual features," Mach. Vis. Appl., vol. 34, no. 6, pp. 1–13, Nov. 2023, doi: 10.1007/S00138-023-01464-5/METRICS.

[18] T. Li, T. Akiyama, and L. Wei, "Constructing a highly accurate price prediction model in real estate investment using LightGBM," Proc. - 4th Int. Conf. Multimed. Inf. Process. Retrieval, MIPR 2021, pp. 273–276, 2021, doi: 10.1109/MIPR51284.2021.00051.

[19] V. Gampala, N. Y. Sai, and T. N. Sai Bhavya, "Real-Estate Price Prediction System using Machine Learning," Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022, pp. 533–538, 2022, doi: 10.1109/ICAAIC53929.2022.9793177.

[20] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," J. Appl. Sci. Technol. Trends, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/JASTT1457.

[21] H. Tyralis and G. Papacharalampous, "Boosting algorithms in energy research: a systematic review," Neural Comput. Appl. 2021 3321, vol. 33, no. 21, pp. 14101–14117, Apr. 2021, doi: 10.1007/S00521-021-05995-8.

[22] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artif. Intell. Rev., vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5/METRICS.

[23] M. Ihme, W. T. Chung, and A. A. Mishra, "Combustion machine learning: Principles, progress and prospects," Prog. Energy Combust. Sci., vol. 91, p. 101010, Jul. 2022, doi: 10.1016/J.PECS.2022.101010.

[24] G. G.-A. Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, "Modelling species presence-only data with random forests", [Online]. Available: https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/ecog.05615

[25] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, "Random Forest Spatial Interpolation," Remote Sens. 2020, Vol. 12, Page 1687, vol. 12, no. 10, p. 1687, May 2020, doi: 10.3390/RS12101687.

[26] D. Borup, B. J. Christensen, N. S. Mühlbach, and M. S. Nielsen, "Targeting predictors in random forest regression," Int. J. Forecast., vol. 39, no. 2, pp. 841–868, Apr. 2023, doi: 10.1016/J.IJFORECAST.2022.02.010.

[27] H. C. et Al., "A Comparative Study between the Parameter-Optimized Pacejka Model and Artificial Neural Network Model for Tire Force Estimation," J. Auto-vehicle Saf. Assoc., vol. 13, no. 4, pp. 33–38, 2021.

[28] N. Mohamed, M. Bajaj, S. K. Almazrouei, F. Jurado, A. Oubelaid, and S. Kamel, "Artificial Intelligence (AI) and Machine Learning (ML)-based Information Security in Electric Vehicles: A Review," Proc. - 2023 IEEE 5th Glob. Power, Energy Commun. Conf. GPECOM 2023, pp. 108–113, 2023, doi: 10.1109/GPECOM58364.2023.10175817.

[29] A. Alcañiz, D. Grzebyk, H. Ziar, and O. Isabella, "Trends and gaps in photovoltaic power forecasting with machine learning," Energy Reports, vol. 9, pp. 447–471, Dec. 2023, doi: 10.1016/J.EGYR.2022.11.208.

[30] T. Ma, J. Mou, H. Yan, and Y. Cao, "A new class of Hopfield neural network with double memristive synapses and its DSP implementation," Eur. Phys. J. Plus 2022 13710, vol. 137, no. 10, pp. 1–19, Oct. 2022, doi: 10.1140/EPJP/S13360-022-03353-8.

[31] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Sci. Reports 2022 121, vol. 12, no. 1, pp. 1–11, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[32] S. Zhang, "Challenges in KNN Classification," IEEE Trans. Knowl. Data Eng., Oct. 2021, doi: 10.1109/TKDE.2021.3049250.

[33] S. Zhang and J. Li, "KNN Classification With One-Step Computation," IEEE Trans. Knowl. Data Eng., vol. 35, no. 3, pp. 2711–2723, Mar. 2023, doi: 10.1109/TKDE.2021.3119140.

Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.