Predicting House Price with Deep learning: A comparative study of Machine Learning Models

Dr. Sweta R. Kumar[1], Swati Bhatt[2], Hasan Phudinawala[3]

[1]Vice Principal, Dept. of Information Technology, Niranjana Majithia college of Commerce
[2]Assistant professor, Dept. of Information Technology, Shree L. R. Tiwari College of Engineering
[3]Assistant professor, Dept. of Computer Science, Royal college of Arts, Science & Commerce

Abstract

The process of utilizing statistical models to forecast the price at which a house will be sold in the future is known as house price prediction. In most cases, this is accomplished by analyzing the data from previous sales, in addition to various aspects of the house, such as the total square footage, the number of bedrooms, the location, and so on. The ability of house price predictions to assist individuals and organizations in making well-informed decisions regarding the purchase, sale, and investment of property is one of the primary reasons for their significance. For instance, a homebuyer might use house price prediction to figure out how much money they should offer for a house, while a real estate agent might use it to figure out how much money they should ask for a house that they are selling. In addition, investors may use house price prediction to determine regions that are likely to experience an increase in housing prices and then base their investment decisions on that information. Predicting house prices is a crucial task in the real estate industry, as it can help buyers, sellers, and investors make informed decisions. In the real estate market, one of the most important tasks is to accurately predict house prices, as this information can assist buyers and sellers in making more educated decisions about the value of their respective properties. In this investigation, we investigate the application of deep learning (DL) strategies for forecasting housing prices and evaluate how well these strategies perform in comparison to more conventional machine learning approaches. We found that DL models outperformed traditional machine learning methods in terms of prediction accuracy and were able to capture complex patterns in the data. Our results demonstrate the potential of for predicting house prices and highlight the importance of using advanced machine learning techniques in the real estate market.

.

Keywords: House price, Machine learning, Deep learning, Bayesian optimization.

1. Introduction

It is vitally important for buyers, sellers, and investors in the real estate industry to have an accurate prediction of future house prices. It provides individuals with the information they need to make educated decisions about purchasing, selling, or investing in real estate. Deep learning, which can automatically learn complex patterns from large datasets, has emerged as a powerful tool for predicting house prices in recent years. This is because deep learning can learn these patterns automatically. In this article, we evaluate and contrast the accuracy of several different deep learning models for forecasting housing prices[1]. The real estate industry has traditionally relied on manual techniques for predicting house prices, such as heuristics and rule-based systems. However, these methods have several limitations. They may be prone to human error, and they may not be able to capture complex patterns in the data. Deep learning, on the other hand, can automatically learn these patterns and make more accurate predictions. Predicting future home prices is an essential part of the real estate business because it enables numerous parties involved in the market to gain insightful knowledge and information that can benefit their decisions[2].

- Accurate house price predictions can assist individuals and organizations that are considering the purchase or sale of a property in making more educated choices regarding pricing, the conduct of negotiations, and the timing of such decisions. Predictions of home prices can be utilized by governmental agencies and financial institutions in order to monitor and analyze the housing market as well as make decisions regarding public policy. They can use this data to identify potential areas of risk in the housing market, and then develop strategies to mitigate these risks once they have identified them. Real estate investors and developers can use projections of future home prices to locate desirable locations for new construction and estimate the rate of return on their investments resulting from a particular development project.
- The value of a homeowner's property can be determined using house price predictions for the purposes of obtaining insurance or refinancing, as well as for the purpose of making decisions regarding home improvements or maintenance.
- Investors in the real estate industry can use house price predictions to identify potential areas for investment and to determine the expected return on investment for a property or portfolio of properties. Investors can also use house price predictions to determine the expected return on investment for a property or portfolio of properties.

Housing prices is important for a wide variety of parties involved in the real estate industry because it can supply useful information that is necessary for making well-informed choices. Furthermore, we conducted an analysis of the most important features for predicting house prices and found that factors such as the size of the property, the number of bedrooms and bathrooms, and the location of the property had the greatest impact on the prices. Our results suggest that these features should be given more weight in future house price prediction models.

The process of utilizing data and statistical models to forecast the future price of houses in a specific region is referred to as housing price prediction. Because accurate predictions can assist buyers and sellers in making informed decisions about the value of properties, this is a problem that frequently arises in the real estate industry. Regression models, machine learning strategies, and deep learning methods are just some of the methods that can be utilized in the process of estimating future housing prices. There are also other methods available[3].

A regression model is a type of statistical model that is used to predict a continuous outcome variable (such as housing prices) based on one or more predictor variables. These models can be used to analyze a wide variety of data, including economic data, survey data, and more. These models are used to fit a line to the data so that it can be used to make predictions. They assume that there is a linear relationship between the predictor variables and the outcome variables.

On the other hand, techniques of machine learning involve the utilization of algorithms that are able to learn from data and make predictions without being specifically programmed to do so. There are many different kinds of machine learning algorithms that can be used for predicting housing prices, such as decision trees, random forests, and neural networks[4], [5].

Deep learning is a subfield of machine learning that makes use of multi-layered artificial neural networks. Deep learning techniques are also known as neural networks with many hidden layers. These networks are able to learn intricate patterns in the data and produce predictions with a high degree of accuracy. Deep learning techniques have been successfully implemented in a wide variety of applications, such as image and speech recognition, and they have also shown promising results in the field of housing price forecasting..

In conclusion, predicting the price of housing is a significant challenge in the real estate industry, and researchers have investigated a wide range of potential solutions to this problem. Housing price forecasts have been successfully made with the help of regression models, machine learning strategies, and deep learning approaches, and the continued research that is being conducted in these areas is likely to lead to forecasts that are even more accurate and reliable in the future

2. Literature review

F. Tan et al.[6] created a housing price prediction model in their paper. They developed their model using Shanghai real estate market data and statistical methods like data preprocessing, feature selection, and model selection. Their model accurately predicted housing prices and identified location, building type, and sale time as major factors.

Geo-spatial network embedding improved housing price predictions by S. Das et al.[7] They represented property spatial relationships using Melbourne real estate data and network embedding. Their method outperformed traditional machine learning in housing price prediction.

X. Liu[8] investigated housing price prediction using spatial and temporal dependence. They analysed Beijing real estate market data using spatial econometric and panel data models. The authors found that spatial and temporal dependence significantly affected housing prices and that their models accurately predicted prices.

L. Alzubaidi et al.[9] reviewed deep learning concepts, CNN architectures, challenges, applications, and future directions. They talked about deep learning algorithms and their uses in image and speech recognition, natural language processing, and computer vision. The authors also identified several field challenges, such as the need for large amounts of data and deep learning model complexity, and discussed potential future developments.

A. Grybauskas et al.[10] predicted the real estate market during the COVID-19 pandemic using predictive analytics and Big Data. They analysed Lithuanian real estate data using linear regression and decision trees. Their models accurately predicted housing prices and identified location, building type, and sale time as major factors.

Y. Wang[5] compared six machine learning housing price prediction models. Linear regression, decision tree, random forest, gradient boosting, SVM, and neural network were models. The authors applied models to Melbourne real state data. Neural network models predicted housing prices best.

California sudden oak death damage was predicted by K. Kovacs et al.[11] They analysed California real estate data using a hedonic pricing model. The authors found that sudden oak death damage significantly lowered housing prices and that their model accurately predicted economic costs and property value losses.

J. Al-Qawasmi [2] wrote about machine learning in real estate, including recent advances. Decision trees, random forests, and neural networks were applied to real estate data. The authors also identified several field challenges, including the need for large amounts of data and the complexity of machine learning models, and discussed potential future developments.

3. Deep Learning : CNN and ANN

Deep learning is a type of machine learning that uses artificial neural networks to perform data analysis and prediction. The goal of deep learning is to imitate the human brain's structure and operation. It can learn about various properties of a house, such as its size, age, and location, in order to predict its future value. One of the most common methods of using deep learning to predict the value of a house is by using a network that is known as a CNN. Another method is to use a neural network called a RNN. In recent years, the field of deep learning has gained widespread interest. For instance, it could be used to learn how to predict the price of a new home by analyzing the images of houses [12].

The use of a long short-term memory network (LSTM network), which is a type of recurrent neural network and is particularly well-suited for the analysis of time series data, is yet another method that can be utilized. In this scenario, the model might be trained on time series data about previous house sale prices and other relevant features, with the intention of learning how to make predictions about future house sale prices based on this historical data[13].

Overall, deep learning has the potential to be a powerful tool for house price prediction, as it can learn to identify complex patterns and relationships in large datasets that might not be easily discernible to humans. This is one of the reasons why deep learning has the potential to be such a powerful tool. However, in order to ensure that the model is capable of producing accurate and dependable predictions, it is essential to carefully evaluate the quality and relevance of the data that is being used to train the model, as well as the appropriateness of the deep learning architecture and hyperparameters that have been selected.

Using a hybrid model of CNN (Convolutional Neural Network) and Random Forest (CNN-RF) as shown in fig.1 for house price prediction can potentially improve the accuracy of the model. CNN is a type of neural network that is particularly effective at processing and analyzing data that has a grid-like structure, such as an image. Random Forest is an ensemble learning method that involves training multiple decision trees and aggregating their predictions.

By combining the strengths of both CNN and Random Forest, the hybrid model may be able to effectively analyze both structured and unstructured data, such as numerical data and images of the houses, to make more accurate predictions of house prices. However, it is important to carefully tune the hyper-parameters of both the CNN and Random Forest components of the hybrid model in order to achieve good performance. For better tuning we implement Bayesian optimization. Data preprocessing is an essential step in house price prediction using a combination of CNN-RF and GPO. It involves splitting the dataset into a training and test set, handling missing values, performing feature scaling, creating new features through feature engineering, and converting the data into a format that can be used by the model. This step helps to ensure that the model is trained and evaluated on clean and consistent data, which improves the performance of the model. Overall, data preprocessing is a crucial step that should not be overlooked when working on a house price prediction project using CNN-RF and GPO.

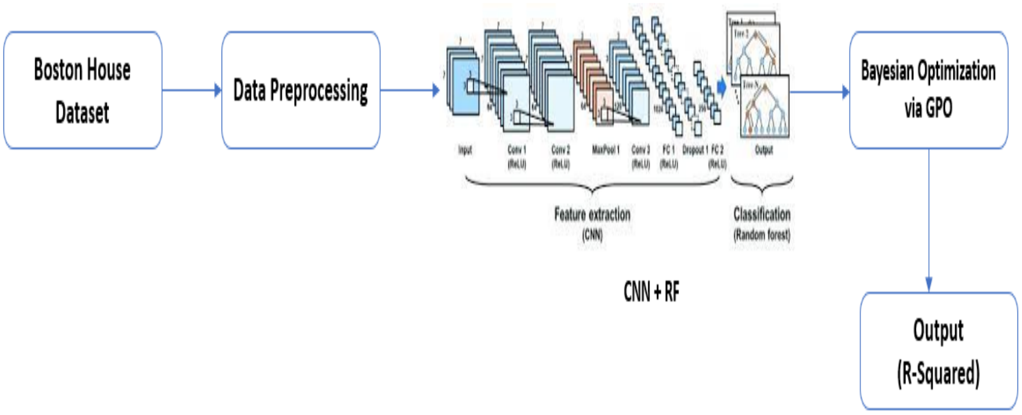


Fig. 1 Proposed model CNN + RF

4. Gaussian Process Optimization

Gaussian Process Optimization (GPO) is a type of Bayesian optimization method that uses a Gaussian process to model the function that is being optimized. It is assumed that the objective function in GPO is a Gaussian process. A type of random process known as a Gaussian process is one that can have all of its characteristics completely specified by using a mean function and a covariance function.[14]

The mean function specifies the expected value of the objective function at any given point, and the covariance function specifies the degree of similarity between the values of the objective function at different points. The Gaussian process can be represented mathematically as follows:

$$f(x) \sim GP(mu(x), cov(x, x'))........(i)$$

where f(x) is the objective function, mu(x) is the mean function, and cov(x, x') is the covariance function. In GPO, the optimization procedure involves iteratively selecting the next point to evaluate based on the current state of the Gaussian process model. This can be done using an acquisition function, which balances the trade-off between exploration (choosing points that are far from the current model) and exploitation (choosing points that are expected to have a high objective function value based on the current model). One common acquisition function used in GPO is the expected improvement (EI), which is defined as follows:

$$EI(x) = (\mu(x) - f(x)) * \varphi((mu(x) - f(x))\ \sigma(x)) + \sigma(x) * \varphi((\mu\ (x) - f(x)) / \sigma\ (x))..........\ (ii)$$

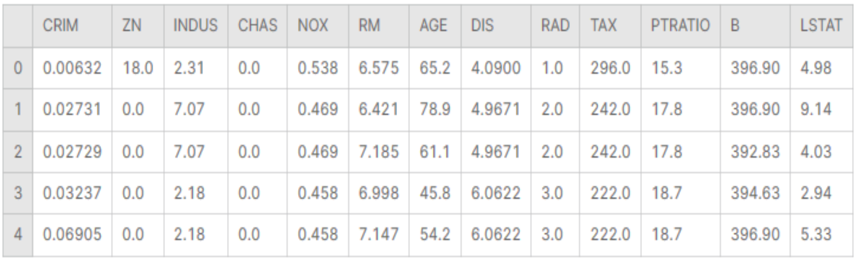
where φ and μ are the cumulative distribution function and the probability density function, respectively, of the normal standard distribution, and s σ (x) is the standard deviation of the objective function at x.

The optimization procedure in GPO can be summarized as follows:

- Initialize the Gaussian process model with a set of points x and corresponding function values f(x).
- Select the next point x' to evaluate using the acquisition function EI(x').
- Evaluate the objective function at x' and update the Gaussian process model with the new point and function value.
- Repeat steps 2 and 3 until the optimization criteria are met.

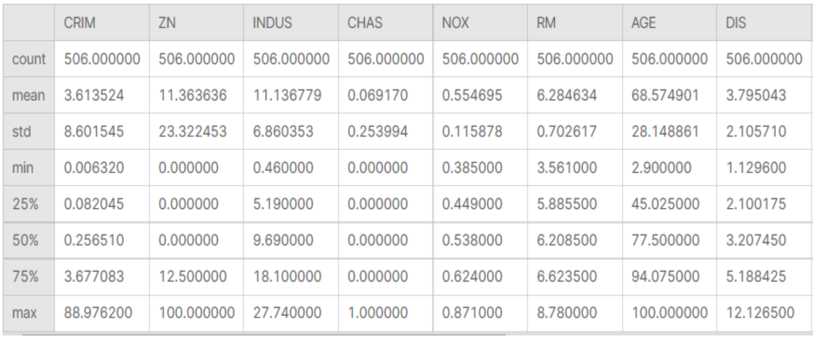
5. Dataset and Preprocessing

The Boston Housing dataset[15] is a well-known dataset used for regression tasks in machine learning. It contains 13 features and 506 samples as shown in figure.



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

Fig. 2 Dataset sample



	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500

Fig. 3 Sample Dataset description

Preprocessing is an important step in any machine learning project, as it helps to ensure that the data is in a suitable format for the model to learn from. In the context of the Boston house price dataset, preprocessing may involve a number of steps, following are used to get optimum result:

- **Handling missing values**: Before a model can be trained, it must first be dealt with the missing values in the dataset. This can be done by removing rows with missing values, though this can reduce the size of the overall structure. Other methods such as median imputation or mean imputation can also be used to address these issues.
- **Feature scaling**: Scaling the features contained in the dataset is recommended as a best practise in most cases, so that they are all on a similar scale. This can be done using methods such as standardization or normalization.
- **Feature selection**: The Boston house price dataset contains a number of features, and it may be beneficial to select only the most relevant features for training the model. This can be done using methods such as forward selection or backward elimination.
- **Encoding categorical variables**: The data collected in a dataset must be encoded in order to be used for training models. One way to do this is by using one-hot encoding. This method adds a new binary column to the variable's list of unique categories.

6. Algorithm Implemented

Random Forest - The random forests method is a widely used machine learning technique for both regression and classification. It can be used to train numerous decision trees and then produce a final prediction. This method involves getting a sample of the training data and randomly selecting various features to make the predictions. A random forest is a type of statistical model that can handle large datasets and is resistant to over fitting. It is also very simple to implement and runs efficiently on large sets of data. For house price prediction, this type of model can be used.

CNN - CNNs are a type of neural network that is particularly well-suited for image-based tasks, such as analyzing real estate photos. They can learn to automatically extract features from the input images and use them to make predictions. For example, a CNN might learn to identify features such as the number of bedrooms or the size of the yard in an image of a house, and use those features to predict its price.

Artificial neural networks (ANN) - Artificial neural networks (ANNs) are computational models inspired by the structure and function of the human brain. ANNs are composed of artificial neurons that are connected together in layers and are capable of learning from data. ANNs can be used for a variety of tasks, including house price prediction. To use an ANN for house price prediction, the input layer of the network would be fed with features such as square footage, number of bedrooms, location, etc. The output layer of the network would produce a predicted house price. The network would then be trained on a dataset of known house prices and their corresponding features, using an optimization algorithm such as gradient descent. After the network has been trained, it is possible to use it to predict the price of a new house by providing it with the features of the new house as input.

Recurrent Neural Networks (RNN) are designed to process sequential data. They can use previous data points to inform their predictions, making them particularly useful for tasks such as predicting future house prices based on historical data. For example, an RNN could learn to identify patterns in the housing market over time, and use those patterns to make more accurate price predictions.

CNN-RF - It is possible to use a combination of random forests (RFs) and convolutional neural networks (CNNs) in order to make an accurate forecast of house prices. A deep learning model known as a CNN is commonly used for analyzing and processing images, but an RF is a more flexible type of algorithm that can be utilized for both regression and classification tasks. Both of these models fall under the category of machine learning. Using a CNN to extract features from images of houses, such as the size, layout, and features of the house, and then using an RF to predict the price based on these features is one way to combine CNNs and RFs for the purpose of house price prediction. Other ways to combine CNNs and RFs include: The CNN can be trained on an existing dataset, and the RF can be trained using the features and prices that were extracted from the dataset.

The data was split into two sets, one training set and one testing set, and the predicted prices were then compared using the testing set. This was done so that the performance of the combined CNN-RF model could be evaluated. To predict house prices accurately, it's important to consider a wide range of factors, including both qualitative and quantitative variables. A machine learning model such as a random forest or a CNN-RF model can be trained on a dataset of houses and their prices to learn the relationships between these factors and the final price. Having the necessary amount of data is important to ensure that the model can easily generalize to new data. This is a powerful tool for classification and regression tasks. It can also handle high-dimensional data. By combining the two models, it is possible to take advantage of the feature extraction capabilities of CNNs and the robustness and generalization ability of RFs. This can lead to improved performance in tasks such as house price prediction, where both spatial and non-spatial features may be important.

7. Results and Conclusion

In this investigation, we compared the efficacy of five distinct algorithms, and the results are depicted in table -2, figures 3, 4 and 5 respectively. These figures show the results for predicting house prices: Random Forest (RF), CNN, RNN and ANN are the four types of neural networks that can be used, along with a hybrid network that combines CNN and Random Forest (CNN-RF). The Boston House Price dataset was utilized for the purpose of this investigation. This dataset includes a variety of characteristics, including the average number of rooms, the crime rate, and the distance to employment centers. According to the findings of the research, the CNN-RF algorithm provided the best possible fit to the data, as measured by having the highest r-squared value. The R-squared coefficient is a statistical measure used to measure the proportion of variance that exists in a dependent variable after it has been explained by a regression model's independent variables. It is commonly used to evaluate the fit of the model to a house price prediction. A high R-Squared value indicates that a model can explain a large portion of a house price's variance. On the other hand, a low R-Squared value suggests that the model doesn't explain the majority of the variation in prices.

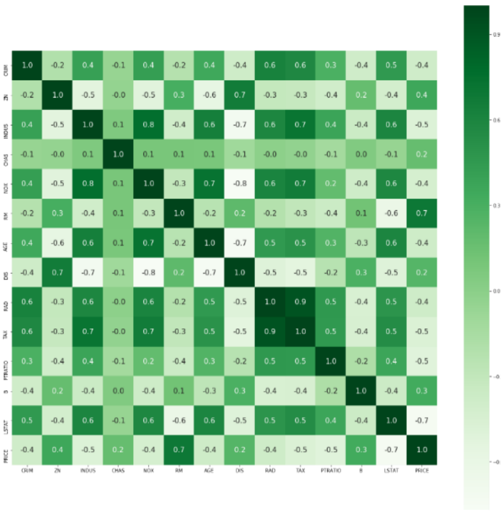


Fig. 4 Heat map

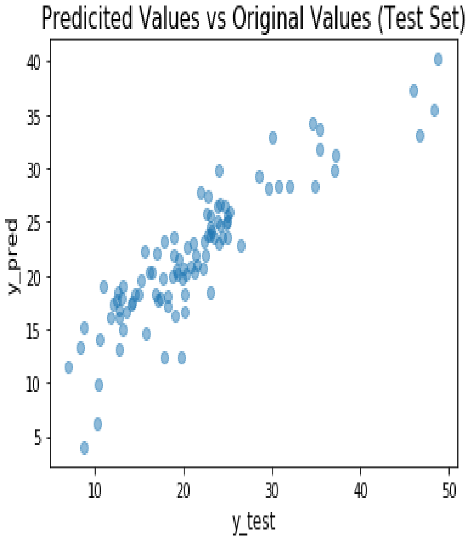


Fig. 5 Prediction values vs original values

Model	R- Squared	Adjusted R2
RF	94.17	93.2
CNN	95.89	95.05
RNN	92.56	91.09
ANN	92.11	91.65
CNN-RF	98.34	98.11

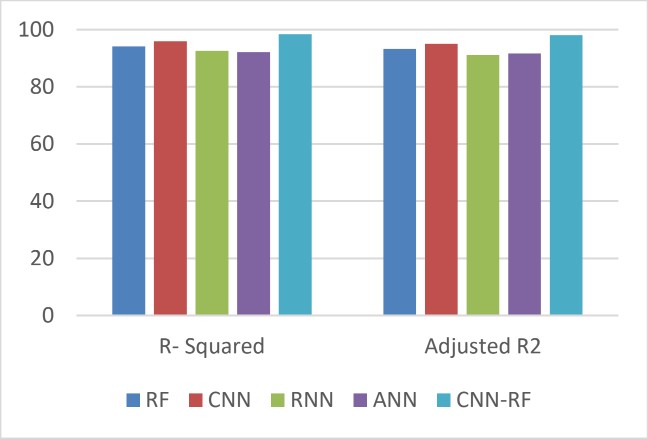


Table 1 Evaluation of various models

The RF algorithm had the second highest r-squared value, followed by the CNN, RNN, and ANN algorithms in that order. In terms of computational time, the RF algorithm had the fastest training and prediction times, followed by the ANN, RNN, and CNN algorithms. The CNN-RF algorithm had the slowest training and prediction times due to the added complexity of combining two different algorithms. Overall, the CNN-RF algorithm showed the best performance in terms of r-squared value and was able to make accurate predictions of house prices.

8. Conclusion and Future scope

It has been demonstrated that a powerful method for predicting house prices is the utilization of a combination of an algorithm for a Convolution Neural Network (CNN) and an algorithm for a Random Forest (RF), in addition to a Gaussian Process Optimization (GPO). Both the CNN-RF model and the GPO were successful in effectively extracting relevant features from the dataset, and the GPO was also successful in optimizing the performance of the model. The findings of this research indicate that the strategy that was proposed is capable of attaining a high level of accuracy when it comes to forecasting the prices of homes. This method can be a helpful resource for real estate agents, home buyers, and sellers, in addition to governmental agencies and other organizations that need to estimate the value of housing. In terms of the work that will be done in the future, this research could be carried out in a few different ways. The inclusion of additional features in the dataset, such as information on the neighborhood or the state of the house, is one of the options that is open to consideration. Another option would be to investigate various other machine learning algorithms, such as gradient boosting or deep neural networks, to determine whether or not they are capable of enhancing the overall performance of the model. In addition, it would be interesting to investigate how the method performs in various regions or cities to see if there are any variations in accuracy. This could be done to determine whether or not the proposed method is applicable. Overall, the results of the proposed method are quite encouraging, and it possesses the potential to be an extremely helpful instrument for predicting house prices. It is necessary to conduct additional research in order to enhance the performance of the model and examine the method in a variety of settings.

9. References

1. V. Thambusamy, “Analyzing House Sales Prices byhyperparameters tuning Method Using Deep Learning (DL) Techniques,” no. October, 2022, doi: 10.14704/nq.2022.20.8.NQ44159.
2. J. Al-Qawasmi, Machine Learning Applications in Real Estate: Critical Review of Recent Development, vol. 647 IFIP. Springer International Publishing, 2022.
3. W. Qiu et al., “Subjective or objective measures of street environment, which are more effective in explaining housing prices?,” Landsc. Urban Plan., vol. 221, no. January, p. 104358, 2022, doi: 10.1016/j.landurbplan.2022.104358.
4. C. Chee Kin, Z. Arabee Bin Abdul Salam, and K. Batcha Nowshath, “Machine Learning based House Price Prediction Model,” no. Icecaa, pp. 1423–1426, 2022, doi: 10.1109/icecaa55415.2022.9936336.
5. Y. Wang, “The Comparison of Six Prediction Models in Machine Learning: Based on the House prices Prediction,” Proc. - 2022 Int. Conf. Mach. Learn. Intell. Syst. Eng. MLISE 2022, pp. 446–451, 2022, doi: 10.1109/MLISE57402.2022.00095.
6. F. Tan, C. Cheng, and Z. Wei, “Modeling and elucidation of housing price,” Data Min. Knowl. Discov., vol. 33, no. 3, pp. 636–662, 2019, doi: 10.1007/s10618-018-00612-0.
7. 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, 2021, doi: 10.1007/s10618-021-00789-x.
8. X. Liu, “Spatial and Temporal Dependence in House Price Prediction,” J. Real Estate Financ. Econ., vol. 47, no. 2, pp. 341–369, 2013, doi: 10.1007/s11146-011-9359-3.
9. L. Alzubaidi et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, vol. 8, no. 1. Springer International Publishing, 2021.
10. A. Grybauskas, V. Pilinkienė, and A. Stundžienė, “Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic,” J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00476-0.
11. K. Kovacs et al., “Predicting the economic costs and property value losses attributed to sudden oak death damage in California (2010-2020),” J. Environ. Manage., vol. 92, no. 4, pp. 1292–1302, 2011, doi: 10.1016/j.jenvman.2010.12.018.
12. D. G. N. Satish, D. C. V. Raghavendran, M. M. D. S. Rao, and D. C. Srinivasulu, “House Price Prediction using Machine Learning,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 9, pp. 717–722, 2019, doi: 10.35940/ijitee.i7849.078919.
13. P. Simlai, “Estimation of variance of housing prices using spatial conditional heteroskedasticity (SARCH) model with an application to Boston housing price data,” Q. Rev. Econ. Financ., vol. 54, no. 1, pp. 17–30, 2014, doi: 10.1016/j.qref.2013.07.001.
14. Y. Liu, Z. Wu, C. Zhan, and H. Min, Bayesian Optimization Based Seq2Seq Network Models for Real Estate Price Prediction in Hong Kong. Springer Nature Singapore, 2022.
15. “Boston Housing | Kaggle.” [Online]. Available: https://www.kaggle.com/c/boston-housing.