INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

IJIE

Vol. 17 No. 7 (2025) 41-51
https://publisher.uthm.edu.my/ojs/index.php/ijie

Application of Support Vector Machine and Gaussian Process Regression for Carbon Emission Prediction in Building Construction

Rufaizal Che Mamat[1]*, Azuin Ramli[2], Aminah Bibi Bawamohiddin[3]

[1] *Centre of Green Technology for Sustainable Cities, Department of Civil Engineering, Politeknik Ungku Omar, 31400 Ipoh, Perak, MALAYSIA*
[2] *Research, Innovation and Commercialisation Unit, Politeknik Ungku Omar, 31400 Ipoh, Perak, MALAYSIA*
[3] *Department of Information Technology and Telecommunications, Politeknik Ungku Omar, 31400 Ipoh, Perak, Malaysia, MALAYSIA*

*Corresponding Author: rufaizal.cm@gmail.com
DOI: https://doi.org/10.30880/ijie.2025.17.07.005

Article Info

Received: 21 January 2025
Accepted: 18 July 2025
Available online: 23 December 2025

Keywords

Climate change, carbon emissions, support vector machine, Gaussian process regression, sustainable development

Abstract

In light of the heightened awareness of climate change, the construction industry is under significant pressure to reduce its carbon footprint. This study aims to apply two advanced intelligent methods, Support Vector Machine (SVM) and Gaussian Process Regression (GPR), to predict carbon emissions during the construction stage of building projects. The models were trained and tested using four input parameters: quantity of construction machinery, fuel consumption rate, carbon emission factor per unit of fuel or electricity consumed, and operating hours of the machinery. The performance of the models is compared to determine the most accurate and reliable predictor. The results demonstrate that the GPR model consistently outperforms the SVM model in terms of accuracy and consistency. The proposed GPR model is poised to be a valuable tool for policymakers and organisations in making informed decisions to mitigate carbon emissions.

1. Introduction

The construction industry is currently facing a critical imperative to urgently reduce carbon emissions to mitigate the imminent threats of climate change. The alarming surge in atmospheric carbon dioxide, which has increased from 278 ppm in 1750 to 421 ppm in 2023 [1], is primarily attributed to human activities, particularly construction activities. This significant increase vividly illustrates the industry's substantial contribution to global warming, posing severe risks to infrastructure, ecosystems, and human health. Addressing carbon emissions in the construction sector is paramount due to its considerable impact on climate change. Immediate and effective measures are imperative to curb emissions and ensure sustainable development. This urgency underscores the need for innovative and accurate predictive models to guide the industry's efforts to reduce its carbon footprint.

Considering the tendency to increase carbon emissions in the future, it is imperative to strengthen current environmental laws to minimise these emissions. Understanding and forecasting carbon emissions during construction is pivotal for devising effective policies and strategies. Thus, developing a robust model for predicting carbon emissions is crucial. This model can serve as a valuable tool for organisations to create and enhance their policies and strategies. While traditional methods are often reactive, capitalising on the latest technology for predicting and reducing carbon emissions is essential [2]. With advancements in artificial intelligence, new opportunities are emerging to address this issue more progressively through data-driven approaches.

Recently, soft computing methods have demonstrated great effectiveness in addressing complex issues, particularly in predicting carbon emissions. As an advanced artificial intelligence approach, artificial neural networks (ANNs) have emerged as a highly effective tool for predicting carbon emissions, showcasing superior accuracy compared to traditional statistical methods [3], [4]. The capacity of ANNs to comprehend intricate non-linear connections has resulted in their extensive application in forecasting emissions at different levels of detail, ranging from regional sizes [5] to national and global dimensions [6]. Nevertheless, a recognised drawback of ANNs is their inclination towards overfitting, which occurs when the model becomes too tailored to the training data, compromising its ability to perform well on unseen data [7]. Hence, it is crucial to employ careful regularisation techniques and extensive cross-validation to reduce overfitting and ensure the robustness of ANN models in predicting carbon emissions [8]. This model poses complex and highly intricate challenges for problem-solving. While ANN models have been widely used, Support Vector Machine (SVM) and Gaussian Process Regression (GPR) are more likely to exhibit lower overfitting than ANN when dealing with high-dimensional data [9], [10]. Furthermore, SVM and GPR can give the best results when the number of samples or data is limited [11], [12]. On the other hand, the ANN model requires a large amount of data to minimise the impact of noise during training.

In this study, support vector regression (SVR) and Gaussian process regression (GPR) models were utilised to predict carbon emissions during the building construction phase. The study aimed to assess the ability of these models to process complex and nonlinear data and to determine which model was most accurate at predicting carbon emissions. A total of 100 data sets were available, containing information such as the number of construction plants and machinery, the carbon emission factor per unit of fuel or electricity consumed by these plants and machinery, the number of hours these plants and machinery operate, and the rate at which they consume fuel or electricity. The SVR and GPR models were trained using 70 datasets and then tested on 30 additional datasets. To enhance prediction accuracy and reliability while mitigating overfitting, 5-fold and 10-fold cross-validation were implemented during the training phase. The SVR and GPR models were trained with optimal parameters using Bayesian optimisation to enhance their prediction accuracy. The study revealed that the GPR model outperformed the SVR model. The most effective model was then used to conduct a sensitivity analysis on the parameters, employing the method developed in this study to determine their significance.

2. Methodology

Powerful machine learning algorithms, specifically SVR and GPR, were employed to predict carbon emissions during building construction. The main objective was to assess and compare these models for complex and non-linear construction datasets. The methodology began with the selection of four crucial input parameters: the running hours of construction machinery, the carbon emission factor per unit of fuel or energy used, the total number of construction machinery and plants, and the fuel or electricity consumption rate. These attributes directly impact carbon emissions, and the data were obtained from relevant literature. The models were trained using 100 records with complete parameter information. The data were normalised, outliers were removed, and the dataset was divided into training and testing subsets to ensure its quality and consistency. Bayesian optimisation was employed to optimise the model parameters for optimal performance. To minimise overfitting and enhance prediction accuracy, 5-fold and 10-fold cross-validation were utilised to evaluate the robustness of the models.

2.1 Dataset and Pre-processing

In this study, four variables are considered: the number of operating hours of these plants and machinery, the carbon emission factor per unit of fuel or electricity consumption, the total number of construction plants and machinery, and the rate of fuel or electricity consumption, which is determined using the formula for carbon emissions, as stated in Eq. (1). Each data for this input parameter consists of a range of different values and is based on the literature, as presented in Table 1.

$$CE = \sum_{j=1}^{n} M_j \times r_j \times h_j \times F_j \tag{1}$$

In this equation, CE represents carbon emissions, M_j stands for the quantity of the j_{th} piece of construction machinery or plant (where $j = 1,2,3,\ldots n$), n is the total number of building sites, and j is the specific type of building site. F_j represents the carbon emission factor per unit of fuel or power consumption for the j_{th} construction plant and machinery; h_j denotes the number of hours these machines are operational; and r_i, represents the rate of fuel or electricity consumption for these machines.

To develop accurate prediction models, effective data pre-processing is crucial. This includes cleaning the data, normalising it, removing outliers, selecting features, and segmenting the data to ensure that it is clean,

Int. Journal of Integrated Engineering Vol. 17 No. 7 (2025) p. 41-51

relevant, and reliable. Using the Min-Max normalisation technique, each value is transformed to a scale of 0 and 1 based on the minimum and maximum values of each parameter. Additionally, the study employs Boxplot and Interquartile range (IQR) techniques to identify and eliminate outliers that can cause bias and inaccuracy. After pre-processing, the data is divided into training and test sets. In this study, 70% of the data is randomly selected as the training set to develop the model, while the remaining data is used as a test set to evaluate the model's performance.

Table 1 *Range of parameter input value*

Input parameter	Range value
M_j	1-5
r_j (L/h) or (kW/h)	500-1000
h_j	8-18
F_j (kgCO2e/L) or (kgCO2e/kW)	0.5-3

2.2 SVM Model Development

This study utilised support vector machines and Gaussian process regression to analyse the data. The SVM is a learning technology first introduced by Cortes & Vapnik [13]. SVM is a machine-learning technique for problem-solving through classification. The application of SVM to address regression difficulties is known as support vector regression (SVR). The SVR technique effectively addresses complex issues that are not linear and involve multiple variables, but have limited data [14]. Additionally, SVR achieves this by mitigating the risks associated with the problem's structure. Su et al. [15] found that SVR converges quickly. Mamat et al. [16] found that SVR's capacity to discover correlations between input and output data makes it a more efficient solver of multidimensional function estimation issues.

As previously discussed, SVR is a supervised machine-learning algorithm used to predict discrete values. It operates on the same underlying principle as SVM. In contrast to traditional regression models that aim for one-to-one mapping, the SVR seeks the best-fit line [17]. The hyperplane with the most points is considered the best match in the SVR model. The main purpose of SVM is to determine the appropriate decision hyperplane to maximise the distance between two classes of samples on both sides of the hyperplane while maintaining good generalisation ability. What is interesting is that SVR can be applied in various domains, including estimation, and is particularly well-suited for small datasets with nonlinear characteristics. For this reason, SVR is proposed as a computational tool to predict carbon emissions. The following is a linear regression representation of the data set:

$$D=\left\{\left(x^1,y^1\right),\left(x^2,y^2\right),\ldots\left(x^n,y^n\right)\right\}x\in\square^m,y\in\square \tag{2}$$

where x and y represent the linear function's attributes and labels, respectively.

$$f\left(x\right)=\left\langle w\middle|x\right\rangle+b \tag{3}$$

The regression function can be obtained as:

$$\min\Phi\left(w,\xi\right)=\min\left(\frac{1}{2}\left\|w\right\|^2+c\sum_i\xi_i^-+\xi_i^+\right) \tag{4}$$

where W and C stand for the cost and support vector, respectively, to modify the training model. To eliminate errors, the slack variables are ξ_i^- and ξ_i^+. To obtain satisfactory results in SVM, the Loos function minimises the error between the label and the prediction data. In this study, the ε-insensitive loss function was employed, as follows:

$$L_e=\left\{\begin{array}{c}0\,for\left|f\left(x\right)-y\right|\prec\varepsilon\\\left|\left|f\left(x\right)-y\right|-\varepsilon,otherwise\right.\end{array}\right\} \tag{5}$$

Next, Eq. (4) needs to be solved with the following subjects:

$$y_i - w \times x_i - b < \varepsilon + \xi_w^- w \times x_i + b - y_i \tag{6}$$

$$\varepsilon + \xi_i^- w \times x_i + b - y_i < \varepsilon + \xi_i^- + \xi_i^+ \tag{7}$$

Considering that $\xi_i^+ \geq 0$, the following is a representation of the solution to Eq. (4):

$$\max\left(-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\left\langle x_i, x\right\rangle\right) + \sum_{i=1}^{l}\alpha_i\left(y_i - \varepsilon\right) - \alpha_i^*\left(y_i + \varepsilon\right) \tag{8}$$

where α_i and α_i^{*+} are the Lagrange multipliers subject to the subsequent limitations:

$$\sum_{i}^{l}\left(\alpha_i - \alpha_i^*\right) = 0 \tag{9}$$

$$0 < \alpha_i^*, \alpha_i \leq \square, i = 1,...,l \tag{10}$$

2.3 GPR Model Development

Gaussian process regression is a popular machine-learning technique for predicting and modelling complex system behaviour. The Gaussian process uses the infinite-dimensional Gaussian distribution, which is a generalisation of the multivariate normal distribution. Gaussian processes are useful for statistical modelling, regressing to multiple target values, and analysing higher-dimensional mappings [18]. Gaussian processes are stochastic in the fields of probability theory and mathematical statistics. This multivariate Gaussian distribution has applications in machine learning, signal processing, and other domains.

The GPR is a machine-learning regression algorithm that assumes noise and a Gaussian process prior, with Bayesian inference as the solution. This Bayesian technique offers the added benefit of providing a degree of prediction uncertainty and can handle both linear and nonlinear regression issues [19]. A prior probability distribution over functions in function space can be defined using Gaussian processes. With a covariance matrix and a mean vector, they expand the idea of a Gaussian distribution to the context of functions. GPR is a general approximation of continuous functions in a compact space, regardless of the kernel function's form.

Gaussian processes can predict new data without needing validation, since they consider previous data knowledge and functional correlations [20], [21]. This allows Gaussian process regression models to discover the prediction distribution corresponding to a new test input. Essentially, the Gaussian process is a multivariate Gaussian distribution.

The idea or concept of the Gaussian process is crucial and fundamental to GPR. In simple terms, a Gaussian process is a probability distribution over functions, where each function is essentially a random variable. Similar to fundamental fitting methods, GPR does not require the anticipated order of approximation to be specified [10]. The Gaussian process is fully defined by the covariance function, $k(x, x')$, and the mean function, $m(x)$, which characterise the covariance between any two data points and x's. In a finite dataset with n observations, $D = [x_i, y_i]$time$^{x_{l=1}}$, x_i is the i_{th} instance's input vector, and y_i is the i_{th} instance's observation value. The random variables $f(x_1)$, $f(x_2)$, ... $f(x_n)$ have a joint Gaussian distribution, as illustrated in Eq. (11):

$$f\left(x\right) = GP\left(m\left(x\right), k\left(x, x^{'}\right)\right) \tag{11}$$

In this case, the mathematical expressions for the mean function $m(x)$ and the kernel function $k(x, x')$ are found in Eq. (12) and Eq. (13):

$$m\left(x\right) = E\left[f\left(x\right)\right] \tag{12}$$

$$k\left(x, x'\right) = E\left[\left(f\left(x\right) - m\left(x\right)\right)\left(f\left(x'\right) - m\left(x'\right)\right)\right] \tag{13}$$

Int. Journal of Integrated Engineering Vol. 17 No. 7 (2025) p. 41-51

The noise in the measurements can be considered when developing a general model of the GPR problem, as shown in Eq. (14).

$$y = f\left(x\right) + \xi \tag{14}$$

2.4 Training and Testing of Model

The data on carbon emissions, calculated through Eq. (1), underwent a pre-processing step where they were normalised between 0 and 1. To simulate the SVM and GPR models, the data were randomly divided into training and test sets, with 70% allocated for training and 30% for testing. Four statistical indicators were used to assess the effectiveness of the models: mean absolute percentage error (MAPE), coefficient of determination (R^2), mean absolute error (MAE), and root-mean-squared error (RMSE). In regression analysis, R^2 was utilised to assess the goodness of fit [22]. This method assesses sample predictability. MAE is the mean absolute difference between each sample and the mean, while RMSE is the difference between the measured and projected variables. It determines the unit of mistake size. MAPE computes the average percentage of absolute errors and assesses model reliability. The formulas for calculating these statistical indicators are listed below, as shown in Eq. (15) to Eq. (18):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(CE_C - CE_p\right)^2}{\sum_{i=1}^{n}\left(CE_C - \overline{CE_p}\right)^2} \tag{15}$$

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}\left(CE_C - CE_p\right)^2}{n}} \tag{16}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|\left(CE_C - CE_p\right)\right|}{n} \tag{17}$$

$$MAPE = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}\left|\left(CE_C - CE_p\right)\right|}{CE_C}\right) \tag{18}$$

where CE_C represents the average of calculated carbon emissions, while CE_p represents the predicted carbon emissions and $\overline{CE_p}$ is the average of predicted carbon emissions.

3. Results and Discussion

To gain insight into the future sustainability performance of cities, it is crucial to accurately predict carbon emissions in construction. This involves a critical task for environmental management and law. In this study, two experiments were conducted using 100 computed data points to compare the predictive capabilities of SVR and GPR models for carbon emissions. MATLAB R2022b was used to develop and test the GPR and SVR models. In this study, the regression learner tool was selected due to its excellent features for building, training, and optimising regression models, as well as its intuitive interface. The tool supports various hyperparameter optimisation methods, including Bayesian, random, and grid searches, and also allows for parallel processing. The study excluded other methodologies that did not fulfil the required standards.

In this study, the effectiveness of support vector machines and Gaussian process regression was assessed for predicting carbon emissions. The focus was on their accuracy, ability to handle non-linear data, and suitability for complex construction data. To improve prediction accuracy, the hyperparameters of both models were systematically fine-tuned within a predefined range. For the SVM model, the hyperparameters explored were the box constraint, kernel scale, epsilon, and kernel function. Through 60 iterations of Bayesian optimisation, the best hyperparameters were determined for an SVM with a Gaussian kernel: a kernel scale of 0.055635, an epsilon of 0.00102244, and no data standardisation.

For Gaussian process regression, the hyperparameter tuning process included variables such as sigma, the base function, various kernel functions, the kernel scale, and whether to standardise the data. Utilising Bayesian optimisation, the most effective hyperparameters for GPR were pinpointed. The optimal configuration comprised a constant base function, a Matern 5/2 kernel, a sigma of 0.012030, a kernel scale of 0.0202626, and no data standardisation. These refined settings were crucial in accurately identifying the dataset's underlying patterns, thereby enhancing the model's predictive accuracy. The Matern 5/2 kernel, in particular, proved to be highly effective due to its ability to handle non-linearities and capture intricate relationships in the data. This kernel function is known for its flexibility and robustness in modelling complex phenomena, which is crucial when dealing with multifaceted construction data. By eschewing data standardisation, the model retained the original data distribution, which can sometimes enhance the interpretability and performance of the GPR model in specific contexts.

This optimisation process not only highlighted the GPR model's superior capability in dealing with complex, non-linear datasets, but also underscored the importance of meticulously selecting and fine-tuning hyperparameters. The chosen sigma and kernel scale values were critical in defining the smoothness and variance of the predictions, directly impacting the model's ability to generalise from the training data to unseen data. The Bayesian optimisation method provided a systematic approach to navigating the hyperparameter space, ensuring that the selected parameters maximised the model's predictive performance. A summary of the hyperparameter search ranges and the final optimised parameters for both models is presented in Table 2.

Both 5-fold and 10-fold cross-validation methods were utilised to train and evaluate each model, ensuring a thorough assessment of their robustness and generalizability across different data partitions. This approach, which involved using the entire dataset for training and validation, effectively minimised the risk of overfitting. The effectiveness of this methodology is illustrated in Fig. 1(a) and Fig. 1(b), which depict the R^2 and Mean Absolute Percentage Error (MAPE) values for the refined SVM and GPR models during the testing phase, utilising 5-fold cross-validation. Likewise, Fig. 2(a) and Fig. 2(b) display the performance metrics for the optimised GPR and SVM models with 10-fold cross-validation.

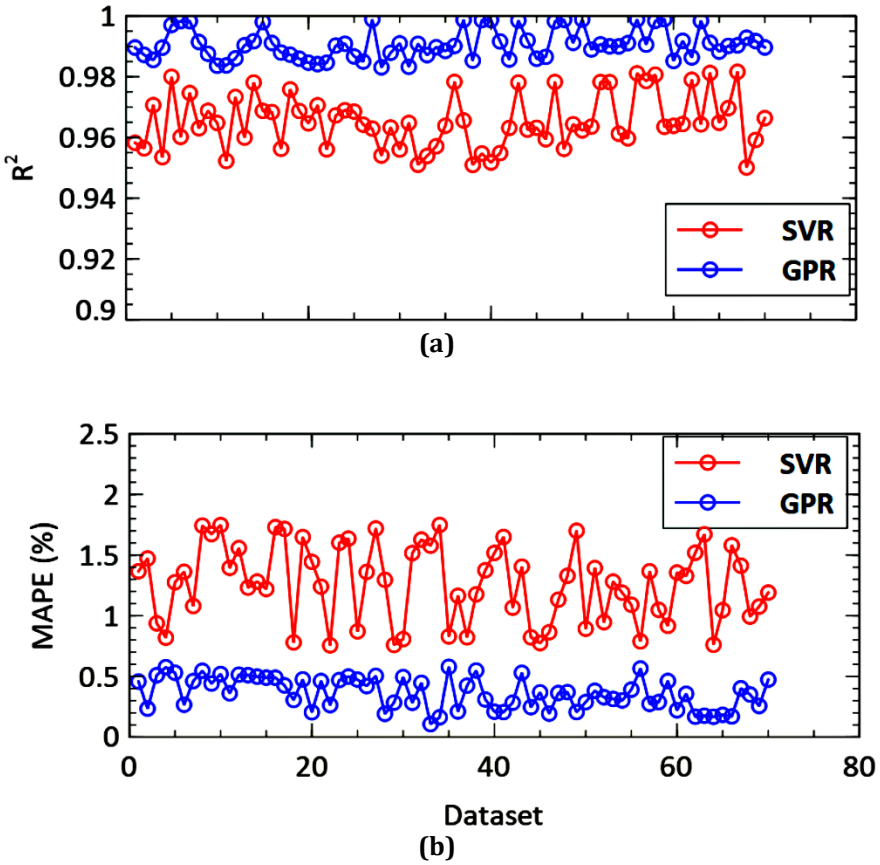

Fig. 1 *Performance index for SVR and GPR models trained with 5-fold cross-validation for (a) R^2; and (b) MAPE*

Int. Journal of Integrated Engineering Vol. 17 No. 7 (2025) p. 41-51

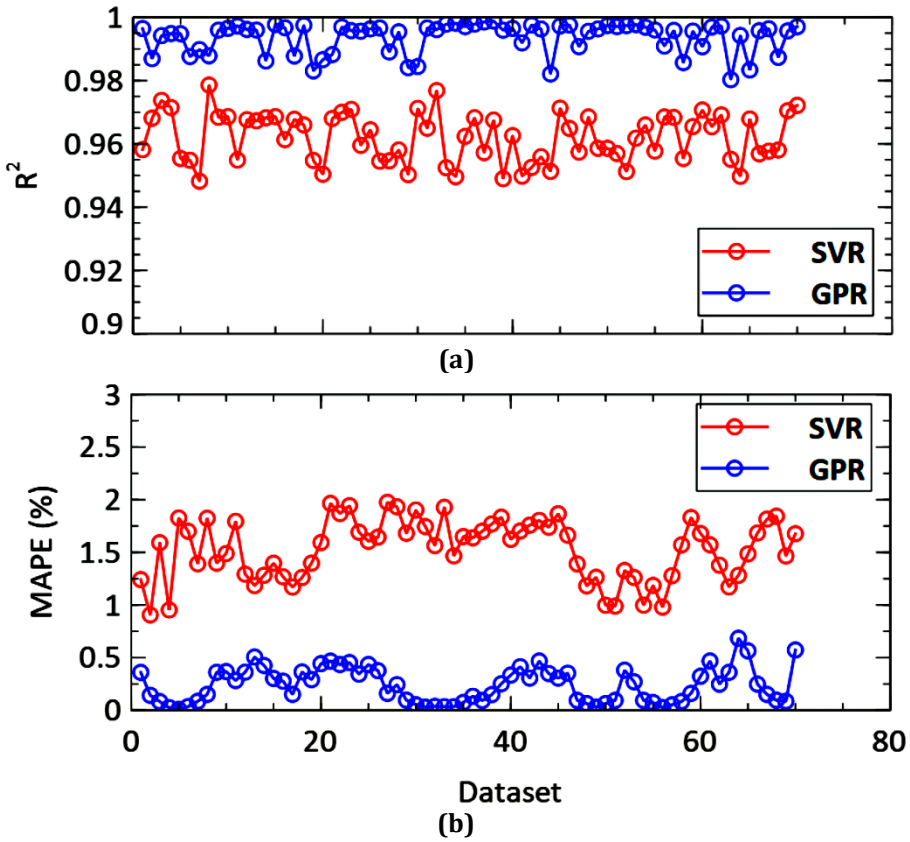

Fig. 2 *Performance index for SVR and GPR models trained with 10-fold cross-validation for (a) R^2; and (b) MAPE*

The comparative analysis reveals minimal differences in the performance metrics between the two cross-validation techniques. This indicates that both models maintain consistent accuracy and reliability across different validation schemes. Such consistency underscores the robustness of the models, ensuring stable and dependable results irrespective of the cross-validation method applied. Upon closer examination, the use of cross-validation serves multiple purposes. Primarily, it allows for the comprehensive utilisation of the dataset, ensuring that each subset of the data is used for both training and validation, thus providing a balanced and exhaustive evaluation of the model's performance. The 5-fold cross-validation divides the data into five subsets, where each subset is used as a validation set while the remaining four are used for training. This process is repeated five times, with each subset being used as the validation set exactly once. Similarly, the 10-fold cross-validation follows the same procedure but divides the data into ten subsets, enhancing the granularity of the assessment.

The results presented in Fig. 1 and Fig. 2 demonstrate that the models exhibit minimal variance in their performance metrics across the two cross-validation techniques. This lack of significant variation suggests that the models are not overfitting to any particular subset of the data, thereby confirming their robustness. The stable R^2 values indicate a high degree of correlation between the predicted and actual values, while the consistent MAPE values reflect the models' precision in prediction across different data splits.

Table 3 presents a comparative analysis of the performance of the GPR and SVM models using 5-fold and 10-fold cross-validation techniques. This table showcases the average performance metrics for predicting carbon emissions on the test data. Remarkably, the GPR model consistently surpasses the SVM model in all evaluation metrics across both cross-validation settings. Specifically, under the 5-fold cross-validation regime, GPR achieves an impressive R^2 value of 0.9989, whereas SVM records a slightly lower R^2 value of 0.9875. This substantial difference underscores GPR's superior predictive accuracy and its ability to capture the underlying patterns in the data more effectively.

In the 10-fold cross-validation scenario, GPR not only maintains its superiority but also enhances its performance with an even higher R^2 value of 0.9996 compared to SVM's R^2 value. This consistent outperformance of GPR indicates its robustness and reliability in different cross-validation settings, making it a more dependable model for predicting carbon emissions in complex datasets. The results emphasise the importance of selecting appropriate models and hyperparameters tailored to the specific characteristics of the data. The higher R^2 values achieved by GPR suggest that its kernel functions, particularly the Matern 5/2 kernel, are better suited for capturing the intricate nonlinear relationships present in the construction data. On the other hand, the SVM, while still effective, does not achieve the same level of precision, potentially due to its kernel function and parameter settings. The minimal differences in performance metrics between the 5-fold and 10-fold cross-validation for both

models indicate their stability and reliability. This stability is crucial for practical applications, where consistent performance across different data splits ensures that the model's predictions will be reliable and robust in real-world scenarios.

Table 2 *Optimisation results of SVR and GPR utilising the Bayesian optimisation*

Model	Search range for hyperparameter	Optimisation of hyperparameters
SVR	Box constraint: 0.0001 – 10000 Kernel function: Gaussian, Quadratic, Linear, and Cubic Kernel scale: 0.001 – 10000 Epsilon: 1.22552 e-06 – 1.22552 Standardise data: true and false	Kernel function: Gaussian Kernel scale: 0.055635 Epsilon: 0.00102244 Standardise data: no
GPR	Sigma: 0.0001 – 0.3725 Basic function: Constant, zero, and linear Kernel function: Exponential, Matern 5/2, Rational Quadratic, and Squared Exponential Kernel scale: 0.001 – 10000 Standardise data: true and false	Sigma: 0.012030 Kernel function: Matern 5/2 Basic function: Constant Kernel scale: 0.0202626 Standardise data: no

Table 3 *Bayesian optimisation results for SVR and GPR*

Index	R^2	RMSE	MAPE	MAE
Model	5 – Fold Cross-Validation			
SVR	0.9875	0.0222	1.3426	0.0186
GPR	0.9989	0.0048	0.1925	0.0052
	10 – Fold Cross-Validation			
SVR	0.9888	0.0208	1.2632	0.0174
GPR	0.9996	0.0022	0.0882	0.0024

Moreover, the GPR model shows a substantial reduction in RMSE, MAE, and MAPE values, highlighting its superior prediction accuracy and precision. Although the SVM model is relatively less accurate, these metrics indicate that it remains a valuable tool for estimating carbon emissions. The data demonstrate that the GPR model outperforms the SVM model in predicting carbon emissions.

This study reaffirms the outstanding predictive ability of the GPR model in estimating carbon emissions, even with a relatively small sample size of 100 data points. By opting not to follow the common trend of using complex deep learning models for carbon emissions prediction, this research underscores the potential for accurate predictions without relying on intricate frameworks, especially when working with limited datasets. The success of simpler machine learning models, particularly GPR, in this context challenges the prevailing notion that complex models and extensive datasets are necessary for optimal performance.

The effectiveness of GPR demonstrated in this study suggests that simpler models can indeed provide robust and reliable predictions, thereby offering a viable alternative to more complex deep learning models. This has significant implications for practical applications where large datasets are often unavailable. The findings suggest that using simpler models, such as GPR, can be a practical and efficient solution when data is limited. This research also adds to the ongoing debate about the trade-off between a model's complexity and the amount of data required. It provides a compelling example of how basic machine learning models can effectively address real-world challenges, particularly in fields with constrained data availability. The observed performance of GPR indicates that it can capture the essential patterns in the data with less complexity, making it a suitable choice for many practical applications.

In summary, this study emphasises the value of simpler machine learning models in predictive tasks, particularly when dealing with limited data. The superior performance of GPR in this context illustrates that high accuracy and precision do not necessarily require complex models. These findings advocate for the broader adoption of simpler models, such as GPR, which can deliver reliable predictions with less computational burden and complexity, thereby making them more accessible and practical for a wide range of applications. This research

Int. Journal of Integrated Engineering Vol. 17 No. 7 (2025) p. 41-51

highlights the need to reassess the role of model complexity in predictive tasks, particularly in scenarios where data is a limiting factor.

3.1 Sensitivity Analysis

Sensitivity analysis helps understand the impact of model variables on the model. This study employed the cosine amplitude method for sensitivity analysis, as it is the most commonly used technique in machine learning models [23]. The cosine amplitude sensitivity analytical approach can be expressed as follows:

$$RI = \frac{\sum_{i=1}^{n} V_i P_i}{\sqrt{\sum_{i=1}^{n} V_i^2 \sum_{i=1}^{n} P_i^2}} \tag{19}$$

Eq. (19) calculates the relative importance (RI) of each input variable. In the equation, V_i represents the variable input into the models, P_i represents the predicted output, and n indicates the number of training data points. The results obtained regarding carbon emissions are depicted in Fig. 3. The most significant model variables for CE were M_j and h_j, which had a greater influence compared to r_j and F_j.

The sensitivity analysis using the cosine amplitude method indicates that the parameters M_j (total machinery and construction plant) and hj (operating hours) have greater relative importance in influencing carbon emissions compared to r_j (rate of fuel or electricity usage) and F_j (carbon emission factor per unit of fuel or power usage). This means that changes in the number of machines and operating hours have a more significant impact on carbon emission forecasts compared to changes in the rate of fuel or electricity usage or the carbon emission factor. These results are crucial for understanding the factors that most significantly impact carbon emissions in construction and can inform more informed decisions to mitigate environmental impacts.

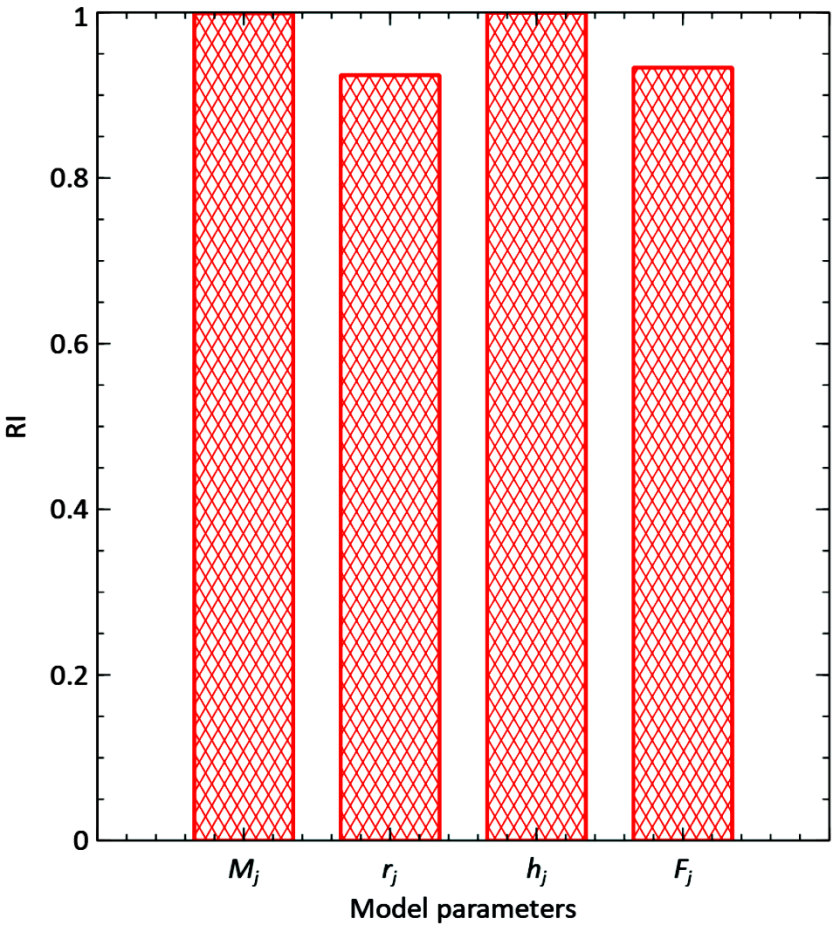

Fig. 3 *Relative importance of model parameters*

4. Conclusions

This study compares support vector machines with regression and Gaussian process regression in predicting carbon emissions. The model used data from computational outcomes and various parameter values. The GPR model consistently shows better prediction accuracy than the SVR model, with lower average MAPE values (0.1925) and higher average R^2 values (0.9989 vs. 0.9875). Sensitivity analysis reveals that M_j and h_j have a greater impact on carbon emissions than r_j and F_j. This research demonstrates the effectiveness of machine learning algorithms, particularly GPR, in accurately predicting carbon emissions during construction. The paper lays the groundwork for improving environmental regulations and makes a significant contribution to the field. To enhance the precision and reliability of carbon emissions predictions, future research could incorporate additional data sources. A more comprehensive and diverse selection of carbon emission data could improve the model's accuracy and applicability across a wider range of civil engineering construction projects.

Acknowledgement

This research was funded by the TVET Applied Research Grant Scheme (Grant No. T-ARGS/2024/BK01/00076) from the Department of Polytechnic and Community College Education, Malaysian Ministry of Higher Education.

Conflict of Interest

The authors declare that they have no conflict of interest regarding the publication of this paper.

Author Contribution

The authors confirm their contribution to the paper as follows: ***Study conception and design:*** *Rufaizal Che Mamat;* ***Data collection:*** *Aminah Bibi Bawamohiddin, Azuin Ramli;* ***Analysis and interpretation of results:*** *Rufaizal Che Mamat, Azuin Ramli;* ***Draft manuscript preparation:*** *Rufaizal Che Mamat, Azuin Ramli, Aminah Bibi Bawamohiddin. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Lan, X., Tans, P., & Thoning, K. W. (2024). Trends in globally-averaged CO2 determined from NOAA Global Monitoring Laboratory measurements. Global Monitoring Laboratory. https://doi.org/10.15138/9N0H-ZH07

[2] Gao, Y., Wang, J., & Xu, X. (2024). Machine learning in construction and demolition waste management: Progress, challenges, and future directions. Automation in Construction. https://doi.org/10.1016/j.autcon.2024.105380

[3] Mutascu, M. (2022). CO2 emissions in the USA: New insights based on ANN approach. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-022-20615-1

[4] Che Mamat, R., Ramli, A., Che Omar, M. B. H., Samad, A. M., & Sulaiman, S. A. (2021). Application of machine learning for predicting ground surface settlement beneath road embankments. International Journal of Nonlinear Analysis and Applications. https://doi.org/10.22075/ijnaa.2021.5548

[5] Jin, H. (2021). Prediction of direct carbon emissions of Chinese provinces using artificial neural networks. PLoS One. https://doi.org/10.1371/journal.pone.0236685

[6] Jena, P. R., Managi, S., & Majhi, B. (2021). Forecasting the CO2 Emissions at the Global Level: A Multilayer Artificial Neural Network Modelling. Energies. https://doi.org/10.3390/en14196336

[7] Aksu, İ. Ö., & Demirdelen, T. (2022). The new prediction methodology for co2 emission to ensure energy sustainability with the hybrid artificial neural network approach. Sustainability. https://doi.org/10.3390/su142315595

[8] Mamat, R. C., Ramli, A., Yazid, M. R. M., Kasa, A., Razali, S. F. M., & Bastam, M. N. (2022). Slope stability prediction of road embankment using artificial neural network combined with genetic algorithm. Jurnal Kejuruteraan. https://doi.org/10.17576/jkukm-2022-34(1)-16

[9] Mamat, R. C., & Ramli, A. (2024). A simple solution for estimating the smear effect permeability ratio using finite element method. Journal of Rehabilitation in Civil Engineering. https://doi.org/10.22075/jrce.2023.28956.1751

[10] Mamat, R. C., & Ramli, A. (2023). Evolutionary polynomial regression for predicting the unconfined compressive strength of lime-stabilized. Suranaree Journal of Science and Technology, 30, 010212.

[11] Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. Reliability Engineering & System Safety. https://doi.org/10.1016/j.ress.2023.109126

[12] Da Veiga, S., & Marrel, A. (2020). Gaussian process regression with linear inequality constraints. Reliability Engineering & System Safety. https://doi.org/10.1016/j.ress.2019.106732

[13] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning. https://doi.org/10.1007/BF00994018

[14] Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., & Khoshnevisan, B. (2014). Potential of radial basis function based support vector regression for global solar radiation prediction. Renewable and Sustainable Energy Reviews. https://doi.org/10.1016/j.rser.2014.07.108

[15] Su, H., Li, X., Yang, B., & Wen, Z. (2018). Wavelet support vector machine-based prediction model of dam deformation. Mechanical Systems and Signal Processing. https://doi.org/10.1016/j.ymssp.2018.03.022

[16] Mamat, R. C., Kasa, A., Razali, S. F. M., Samad, A. M., Ramli, A., & Yazid, M. R. M. (2019). Application of artificial intelligence in predicting ground settlement on earth slope. AIP Conference Proceedings. https://doi.org/10.1063/1.5121094

[17] Tran, N. K., Kühle, L. C., & Klau, G. W. (2024). A critical review of multi-output support vector regression. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2023.12.007

[18] Galeazzi, A., de Fusco, F., Prifti, K., Gallo, F., Biegler, L., & Manenti, F. (2024). Predicting the performance of an industrial furnace using Gaussian process and linear regression: A comparison. Computers & Chemical Engineering. https://doi.org/10.1016/j.compchemeng.2023.108513

[19] Rasmussen, C. E. (2004). Gaussian processes in machine learning. In Bousquet, O., von Luxburg, U., & Rätsch, G. (Eds.), Advanced Lectures on Machine Learning. Springer. https://doi.org/10.1007/978-3-540-28650-9_4

[20] Pohlmann, S., Mashayekh, A., Stroebl, F., Karnehm, D., Kuder, M., Neve, A., & Weyh, T. (2024). State-of-Health prediction of lithium-ion batteries based on a low dimensional Gaussian Process Regression. Journal of Energy Storage. https://doi.org/10.1016/j.est.2024.111649

[21] Mamat, R. C., Kasa, A., & Razali, S. F. M. (2019). The applications and future perspectives of adaptive neuro-fuzzy inference system in road embankment stability. Journal of Engineering Science and Technology Review. https://doi.org/10.25103/jestr.125.09

[22] Mamat, R. C., Ramli, A., Khahro, S. H., & Yusoff, N. I. M. (2022). Numerical simulation and field measurement validation of road embankment on soft ground improved by prefabricated vertical drains: A comparative study. Applied Sciences. https://doi.org/10.3390/app12168097

[23] Lawal, A. I., & Kwon, S. (2023). Development of mathematically motivated hybrid soft computing models for improved predictions of ultimate bearing capacity of shallow foundations. Journal of Rock Mechanics and Geotechnical Engineering. https://doi.org/10.1016/j.jrmge.2022.04.005