Journal of Statistical Sciences and Computational Intelligence
e-ISSN: 3092–9458
2(2), 475–487
DOI: https://doi.org/10.64497/jssci.168
https://jssci.org/index.php/pub

Article

Hybrid gradient boosting-based response surface methodology for modeling nonlinear data

Erick Cheruiyot Kirui [1*], Ayubu Anapapa [2], and John Mutuguta [3]

[1] Department of Mathematics and Actuarial Science, Murang'a University of Technology, Kenya; echeruiyotkirui@gmail.com

[2] Department of Mathematics and Actuarial Science, Murang'a University of Technology, Kenya; anapapa@mut.ac.ke

[3] Department of Mathematics and Actuarial Science, Murang'a University of Technology, Kenya; jmutuguta@mut.ac.ke

* **Correspondence:** echeruiyotkirui@gmail.com

Abstract: The paper introduces a hybrid model, combining Gradient Boosting Machines (GBM) and Response Surface Methodology (RSM) to overcome nonlinear data modeling. A four-factor Central Composite Design (CCD) of 30 experimental runs was simulated to determine the strength of the hybrid GB-RSM model. The RSM model was estimated with a second-order polynomial fitted with maximum likelihood estimation, and the GBM model was trained with five-fold cross-validation, a learning rate of 0.05, tree depths of 3-4, and 1500 trees in R. The GB-RSM model was obtained by training GBM using the RSM residuals to model higher-order nonlinearity. RSM was applied to estimate the linear, quadratic, and interaction terms; the residuals were provided to a GBM to model them to capture non-linearities. Preprocessing was done using normalization and K-Nearest Neighbor imputation of the missing data. Root Mean Square error, Mean Absolute error, as well as Coefficient of Determination R^2 were used to measure the model's performance. It was found that the RSM model had moderate accuracy $(\mathrm{RMSE} = 0.4751, \mathrm{MAE} = 0.2775, R^2 = 0.8209)$ and GBM had better predictive accuracy $(\mathrm{RMSE} = 0.1761, \mathrm{MAE} = 0.1178, R^2 = 0.9754)$. The Hybrid GB-RSM model was better than the two, having the best predictive variance $(\mathrm{RMSE} = 0.0896, \mathrm{MAE} = 0.0491, R^2 = 0.9935)$. The residual diagnostics showed that the GB-RSM residuals were well distributed around zero, which means low bias and low variance with high stability. Combining the advantages of RSM and GBM, the hybrid framework provides better predictive and optimization capabilities, which is why it can be utilized in various data-driven decision-making and optimization.

Keywords: data imputation, gradient boosting, response surface methodology, statistical modeling, hybrid model.

Received: 27 November 2025; Revised: 20 December 2025; Accepted: 10 March 2026; Published Online: 05 June 2026

Copyright: ©2025 the Author(s). Published by JSSCI. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Journal Abbreviation: J. Stat. Sci. Comput. Intell.

1. Introduction

Modeling of complex nonlinear relationships between a response variable and predictor variables remains a daunting challenge in the field of statistical modeling. Traditional statistical methods, especially Response

Surface Methodology (RSM), have been used to model relationships between inputs and outputs under the assumption of linearity [1-2]. Complex nonlinear modeling of the relationships among predictors and responses has been a persistent research problem in fields such as agriculture, engineering, finance, and health care. Conventional methods, especially Response Surface Methodology (RSM), have been widely used because of their ability to capture linear, quadratic, and interaction effects; however, when faced with nonlinearities, high-dimensional data, and missing data, their predictive power suffers a blow [3]. RSM is normally applied in experimental design where the objective is to determine the optimal combination of input conditions likely to achieve the optimum output [4-5].

GBM is an ensemble learning model that builds a sequence of decision trees sequentially, where the decision tree is trained to correct the error of its predecessor [6]. Gradient boosting: Early successes, such as [7], originally proposed the algorithm, set the framework in which weak learners are sequentially improved by minimizing prediction errors by optimizing a specified loss function, suitable for both regression and classification problems. The flexibility and strength of gradient boosting have been useful in overcoming the issue of overfitting by using an imputation algorithm [8]. Gradient boosting applies to both regression and classification problems because of the flexibility of the method to use different loss functions [9]. Its application extends to agriculture, medicine, finance, and even engineering. It also has the advantage of handling missing, incomplete, complex, and nonlinear relationships existing in the data and can also perform well with high-dimensional data [10]. The risk of overfitting the data is inherent in Gradient Boosting, so it incorporates a variety of regularization methods, which keep the model generalizable [11]. Gradient Boosting is especially useful due to its ability to capture complex nonlinear interactions that exist in the data since each decision tree will focus on a different aspect of the residuals, and thus the predictive performance of the model will continue to improve [12].

The general objective was to develop a hybrid model integrating gradient boosting and response surface methodology to model nonlinear and high dimensional data while the specific objectives were to develop an integrated Gradient Boosting-based Response Surface Methodology framework for modeling complex nonlinear relationships between multiple predictors and response and to evaluate the statistical properties, predictive performance, and robustness of the hybrid GB-RSM in comparison with traditional Response Surface Methodology and other machine learning approaches.

2. Methodology

This paper uses a hybrid modeling framework that combines Response Surface Methodology with Gradient Boosting Machines (GBM) to overcome the limitations of nonlinear relationship estimation, high-dimensional data, and missingness. The methodology includes some key steps, such as data preprocessing, model development, and testing. Pre-processing was done to correct discrepancies, cause controlled missingness, and normalize predictors. An RSM model that accommodates linear, quadratic, and interaction effects was then fitted using a polynomial regression. The RSM model residuals were then fed to a Gradient Boosting machine that sequentially modeled the unexplained variance. Hyperparameters of the GBM, including learning rate, estimators, and tree depth, were optimized by response-surface tuning procedures.

2.1. Experimental Design and Data Simulation

This paper used simulated experimental data to identify the strengths of the integrated GB-RSM model using a rotatable four-factor central composite design. Four factors x_1, x_2, x_3, x_4 were studied at 5-levels $(-\alpha, -1, 0, +1, +\alpha)$ where α is the axial distance that facilitates rotatability. The axial distance was $\alpha = 2$, and the number of experimental runs was 30, consisting of center points, axial, and factorial points. The

outcome variable y was generated on a nonlinear combination of factors and their interactions. The K-Nearest Neighbor (KNN) technique was used to impute the missing values (5% of the data), and all variables were normalized before modelling. GBM model was trained on five-fold cross-validation, with a learning rate of 0.05, tree depth of three to four, and 1500 trees in R version 4.4.3. The hybrid GB-RSM model was developed in two steps, where the RSM was first fitted to yield residuals, which were later fitted to the GBM to describe the local nonlinear trends that the surface of polynomials was unable to capture. The resulting hybrid prediction was determined by adding the RSM and GBM components. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination were used to evaluate model performance on an 80%-20% train-test split. To evaluate the bias, variance, and stability, residual diagnostics, Q-Q plots, and predicted-versus-observed plots were performed. All visualizations and simulations, such as 2D contour and 3D surface plots, were performed in R with *caret, rsm, gbm, dplyr, plotly,* and *viridis* packages.

2.2. Response Surface Methodology

The first approach was to apply response surface methodology to model the relationship between the response variable and predictor variables. A second-order polynomial approximation for the response variable y was fitted to capture the linear, quadratic, and interaction terms as shown below:

$$y = f_{RSM}(X) + \varepsilon, \tag{1}$$

where $f_{RSM}(X)$ is a second order polynomial

$$f_{RSM}(X) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \beta_{ii} x_i^2 + \sum_{i<j}^{n} \beta_{ij} x_i x_j + \epsilon, \tag{2}$$

where β_0 is the constant coefficient, β_i the linear coefficient, β_{ii} is the regression coefficient for the quadratic term, and β_{ij} is the coefficient for the interaction term, and ϵ is the error term. These coefficients were estimated using maximum likelihood estimation. Then, the residuals were computed as the difference between the observed and the predicted values as shown below:

$$r_i = y_i - f_{RSM}(X), \tag{3}$$

where r_i is the residual from the i^{th} data point, y_i is the observed value of the response, while $f_{RSM}(X)$ is the predicted value from the RSM model.

2.3. Gradient Boosting

Gradient boosting improves the prediction by fitting the residuals of the RSM model. It sequentially builds the models from weak learners to predict the residuals. For a GB model, the initial step was to fit the first decision tree $f_1(x)$ to the errors (residuals) from the RSM model to minimize the errors. The following steps were followed:

First, the initial decision tree was fitted as follows:

$$F_1(x) = F_0(x) + \lambda f_1(x), \tag{4}$$

where $F_0(x)$ is the initial prediction from the RSM model, $f_1(x)$ is the initial decision tree fitted to the RSM errors and λ is the learning rate which determines the contribution of each decision tree in the model. The residuals at t iterations are given by:

$$r_i^{(t)} = y_i - F_{t-1}(x_i), \tag{5}$$

where y_i is the observed value of the i^{th} observation, $r_i^{(t)}$ is the residual of the i^{th} observation at $t -$

https://doi.org/10.64497/jssci.168

iteration and $F_{t-1}(x_i)$ is the prediction of the i^{th} observation at $t-1$ iterations. The model was sequentially updated by fitting weak learners (decision trees) on the errors of the previous models. The resulting model is given below:

$$F_t(x) = F_{t-1}(x) + \lambda f_t(x), \tag{6}$$

where λ is the learning rate, $f_t(x)$ is the new decision tree at t iteration trained on residuals, and $F_{t-1}(x)$ is the prediction from $t-1$ iteration and $F_t(x)$ is the prediction at iteration t. After T iterations, the final prediction is the sum of the prediction from the RSM model and the subsequent adjustments made by the weak learners.

This study adopted a leaf-based approach over the node-based approach because in a node-based gradient boosting, the correction is applied across the entire decision tree after it has been trained on the residuals in order to make a general correction to all observations in the dataset. The leaf-based gradient boosting, on the hand, applies the correction term at the leaf level of the tree and not the entire tree to update the model for all observations [13-14]. Correction terms are given to each leaf depending on the average residual of the training data points within the leaf. For each learner $f_t(x)$, data is split into K_m leaves, and the update for each K leaf is computed based on the residuals for the data points in the leaf such that:

$$C_{t,k} = argmin \sum_{x \in R_{t,k}} L(y_i, F_{t-1}(x) + c), \tag{7}$$

$C_{t,k}$ is the correction term. This term represents the adjustments made to the prediction for the $k-th$ leaf at $t-iteration$. It is determined by minimizing the loss function that falls within a particular region. $L(y_i, F_{t-1}(x) + c)$ is the loss function, which measures the difference between the observed and the predicted values, and c is the correction applied to the leaf at t −iteration. $R_{t,k}$ is the region or leaf of the decision tree where k represent each node.

After T iterations, the final model is given by:

$$F_T(x) = F_{t-1}(x_i) + \sum_{t=1}^{T} \sum_{k=1}^{K_m} C_{t,k} I\left(x \in R_{t,k}\right), \tag{8}$$

$F_T(x)$ is the final prediction at T iteration, $C_{t,k}$ is the correction term for the $k-th$ leaf at T −iteration, $R_{T,k}$ is the subset of the data points for the $k-th$ leaf at T −iteration and $I\left(x \in R_{t,k}\right)$ is the indicator function that applies adjustments in the regions $R_{t,k}$.

$$I\left(x \in R_{t,k}\right) = \begin{cases} 1, & for\ x \in R_{t,k} \\ 0, & otherwise \end{cases}$$

The integrated Gradient boosting with the RSM model is given by

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\beta}_{ij} x_i x_j + F_T(x), \tag{9}$$

$$F_T(x_i) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\beta}_{ij} x_i x_j + F_{t-1}(x) + \sum_{t=1}^{T} \sum_{k=1}^{K_m} C_{t,k} I\left(x \in R_{t,k}\right), \tag{10}$$

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_i^2 + \sum_{i<j}^{k} \hat{\beta}_{ij} x_i x_j + F_{t-1}(x) + \sum_{t=1}^{T} \sum_{k=1}^{K_m} C_{t,k} I\left(x \in R_{t,k}\right). \tag{11}$$

3. Results and Discussions

https://doi.org/10.64497/jssci.168

The regression analysis performed with the help of a second-order Response Surface Model (RSM) offers a thorough insight into the nonlinear and interactive relationships between the four predictor variables (x_1, x_2, x_3, x_4). This model has both linear, quadratic, and cubic terms of interaction; thus, it can depict both curvature and multi-variable complex interactions that cannot be represented in a pure second-order (quadratic) model.

$$\hat{y} = 0.4564 + 0.1055x_1 + 1.2321x_2 - 0.0783x_3 + 8.394x_4 + 2.2207{x_1}^2 + 2.7594{x_2}^2 - 0.4470{x_3}^2 - 1.6523{x_4}^2 + 3.900x_1x_2 - 1.217x_1x_3 + 0.5177x_1x_4 + 4.1523x_2x_3 - 0.0135x_2x_4 + 5.679x_3x_4. \quad (12)$$

Parameter Estimates with Standard Errors, t-values, and p-values

Term	Estimate	Std. Error	t-value	p-value
(Intercept)	0.4564	0.0825	5.5331	0.0000
x1	0.1055	0.1726	0.6110	0.5414
x2	1.2321	0.1774	6.9447	0.0000
x3	−0.0783	0.1743	−0.4491	0.6535
x4	8.3940	0.1729	48.5547	0.0000
x1:x2	3.9004	0.4470	8.7253	0.0000
x1:x3	−1.2172	0.4449	−2.7361	0.0064
x1:x4	0.5177	0.4363	1.1865	0.2358
x2:x3	4.1523	0.4533	9.1592	0.0000
x2:x4	−0.0135	0.4451	−0.0303	0.9758
x3:x4	5.6795	0.4433	12.8116	0.0000
x1^2	2.2207	0.2963	7.4952	0.0000
x2^2	2.7594	0.3123	8.8349	0.0000
x3^2	−0.4470	0.2978	−1.5010	0.1338
x4^2	−1.6523	0.2966	−5.5698	0.0000

Residual Std. Error = 1.6405 on NA DF | R^2 = 0.7914, Adjusted R^2 = 0.7876 | F-statistic = 208.38 on 14 and NA DF (p < <2e-16)

Figure 1: RSM Output

Error! Reference source not found. shows that the model yielded an R^2 of 0.7914 and adjusted R^2 of 0.788, which means that about 78.8 percent of the total variance in the response variable is accounted for by the predictors in the model. Adjusted R^2 is used to explain the complexity of the model and ensures that most predictors also have significant explanatory power and do not overfit the data. The large F-statistic $208.3, p < 2.2e-16)$ also confirms the statistical significance of the overall regression model. These findings indicate that the second-order model gives a good fit and sufficiently represents the underlying structure of the system, but there is still residual variance that cannot be explained.

These findings show that some predictor variables such as x_2 and x_4 have a strong relationship with the response variable, while x_1 and x_3 exhibited a weak association. The variables x_2 had a coefficient of 1.2321 and a p-value of 0.000, indicating a strong and positive relationship between x_2 and the response. Also, x_4 had a coefficient of 8.390 and a p- value of 0.000, demonstrating a highly significant and positive relationship with the outcome variable. On the other hand, x_1 and x_3 did not show a significant relationship with the response variable. The x_1 variable has a coefficient of 0.1055 and a p-value of 0.5414 that it does not have a significant effect on the response. Also, x_3 has a coefficient of -0.0783 and a p-value of 0.6535, which indicates that the variable is statistically insignificant. These two variables do not have a significant impact on the outcome variable when analyzed individually.

Nevertheless, the terms of interaction provide a more complex situation. Even though x_1 and x_3 do not significantly impact the response, their interaction (x_1: x_3) significantly impacts it, with an estimate of -1.2172 and a p-value of 0.0064. This implies there is a negative relationship between these two predictors; that is, the x_1 impact on the response is reduced when x_3 is high. Conversely, the relationship between x_2 and x_3 (x_2: x_3) significantly impacts the response, and the coefficient of this relationship is 4.1523, with a p-value of 0.0000, which means that the two variables together increase the response. The interaction effects such as x_1: x_2 and x_3: x_4 also reveal how complex the manner in which these predictors interact to determine the result.

The quadratic terms are also extremely important in modeling. All the coefficients of ${x_1}^2$, ${x_2}^2$ and ${x_4}^2$ are significant, and the p-value of each of these is 0.0000. These quadratic terms note that the association between the predictors and the response is not entirely linear, but is curvilinear in nature. The coefficients of ${x_1}^2$ and ${x_2}^2$ are positive, which means that their effect on the response increases with increasing levels of these variables, which is a convex relationship. Conversely, the negative coefficient of ${x_4}^2$ indicates that at increased levels of x_4, the relationship between the x_4 and the response can change to a concave form.

The results demonstrate a number of significant and important relationships between pairs of predictors. The interaction between x_1 and x_2 is positive, with a coefficient of 3.9004 and a p-value of 0.0000, which means that there is a synergistic effect of the interaction between these two variables. This implies that an increase in both x_1 and x_2 will yield a higher response relative to the individual contributions, indicating that the two predictors synergize to improve the system performance. In the same way, the response is positively and significantly affected by the interaction of x_2 and x_3, with a coefficient of 4.1523 and a p-value of 0.0000 indicating that both x_2 and x_3 jointly have an additional impact. This result indicates that interaction terms are also significant to consider in modeling because they can demonstrate combined effects that would be neglected in an analysis of individual predictors at any given moment.

Conversely, the interaction of x_1 and x_3 exhibits a negative impact with a coefficient of -1.2172 and a p-value of 0.0064. It implies that x_1 and x_3 have an inverse relationship; the larger they are, the lesser the effect of the two. It means that there is a negative correlation between x_1 and x_3, as the positive impact of the increase in one variable is weakened by the positive impact of the increase in the other variable. The relationships between x_1 and x_4 and between x_2 and x_4 were, however, revealed as statistically insignificant (p-values > 0.5), and so the two groups of predictors do not exert a significant impact on the response when they interact. This implies that x_1, x_2, and x_4 alone are significant; however, their combinations do not result in any meaningful synergies.

RSM (MLE) Model Performance Metrics

Overall model fit statistics

Metric	Value
Log-Likelihood	−1,889.3080
AIC	3,810.6170
BIC	3,888.8170
R^2	0.7798
Adjusted R^2	0.7764
F-test p-value	0.0000

Figure 2: MLE Model Performance

https://doi.org/10.64497/jssci.168

Error! Reference source not found. provides some of the main performance measurements of a model using Maximum Likelihood Estimation (MLE). The log-likelihood of -1889.3080 gives some idea of how well the model fits the data, with larger values desired to give a better fit. Both the AIC of 3810.6170 and the BIC of 3888.8170 are model comparison measurements, with a lower value indicating a superior model, but they are to be compared to other models. The p-value F is 0.0000, which validates the overall statistical significance of the model, i.e., the regression model is a good predictor. All these metrics indicate good performance of the model, and the balance between the explanatory power and model complexity is good.

3.1. 3D Surface with Contour Plots

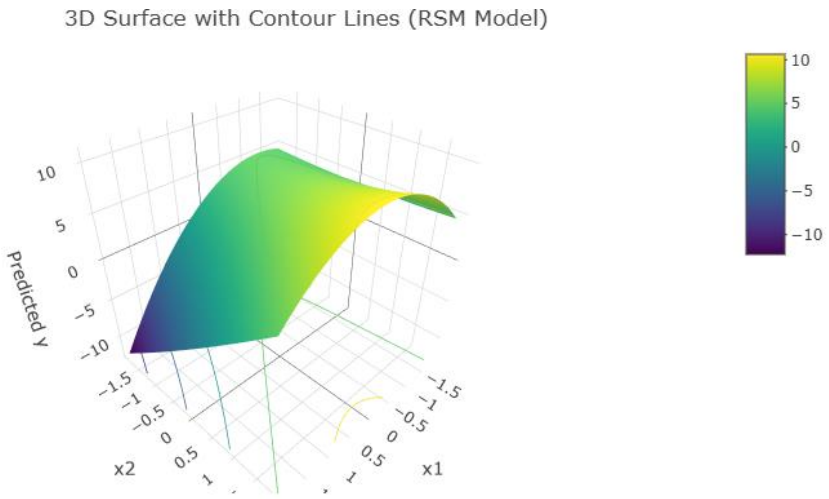


Figure 3: Response Surface

Error! Reference source not found. shows a 3D surface plot with contour lines that show the response surface produced by the fitted Response Surface Methodology (RSM) model of factors x_1 and x_2. The plot reflects a smooth, curved surface, which indicates that the model has successfully captured both linear and quadratic relationships between the independent variables and the predicted response Y. The gradient of colors between purple (low response values) and yellow (high response values) shows areas of the lowest and highest predicted responses, demonstrating the joint effect of changes in the factors on the outcome. The curvature and the form of the contours indicate that the interaction between x_1 and x_2 is nonlinear. The darker lines of contour in some areas suggest a steep response surface, where minimal changes in the level of factors cause significant changes in the response. Generally, the surface suggests that the area with the highest response increase or decrease is around one side of the design space, affirming that the RSM model is able to effectively elicit effects of critical factors and possibly the optimum operating conditions.

3.2. Model Performance Summary

Table 1: Model Performance Summary

Model	RMSE	MAE	R^2
RSM	0.475	0.277	0.821
GBM	0.176	0.118	0.975
Hybrid GB-RSM	0.090	0.049	0.993

Results in **Error! Reference source not found.** shows that RSM model yielded an

https://doi.org/10.64497/jssci.168

RMSE of 0.475 , MAE of 0.277, and an R^2 of 0.821, which is a good fit to the data. RSM was found to be very effective in capturing the key effect factors and interaction of the factors, with a clear interpretation ability in regression coefficients and surface plots. To evaluate the adequacy and diagnostic validity of the second-order Response Surface Methodology (RSM) model, residual analysis was performed; after fitting the quadratic model, the residuals were computed as differences between the observed and predicted response values as a measure of the unexplained variation, and several diagnostic plots, such as those of the residuals versus fitted values, normal Q-Q plot, and residuals versus predictor plot, were plotted to assess the assumptions and performance of the model. The analysis of the residual values revealed slight curvature patterns, which meant that the quadratic model could not.

The GBM model was much better, with $\mathrm{RMSE} = 0.176$, $\mathrm{MAE} = 0.118$, and $R^2 = 0.975$ Nonlinear dependencies and higher-order interactions among the factors were well captured by the ensemble learning mechanism of GBM, which minimized residual errors sequentially using boosted decision trees. In addition, GBM was found to be very robust to the imputed data, adapting to missingness patterns and noise in the predictors. Although the model had better predictive power, it was limited in transparency over the effects of factors compared to RSM due to its non-parametric, black-box nature.

The hybrid GB-RSM model that combines the parametric format of Response Surface Methodology (RSM) with the nonlinear learning ability of Gradient Boosting Machines (GBM) showed the highest overall predictive performance ($RMSE = 0.089, MAE = 0.049$, $R^2 = 0.993$). In this context, the response surface in the RSM component was able to capture the global and continuous curvature of the response surface due to its use of quadratic terms and was also able to estimate main effects as well as second-order interactions. But since RSM is limited by its fixed quadratic functional form, the residual analysis indicated the existence of other nonlinear patterns in the data. To overcome such shortcomings, GBM was trained on the RSM residuals and enabled it to model local deviations and higher-order nonlinear dependencies. This step-wise integration played a major role in improving predictive power, interpretative power, and strength. The hybrid model preserved the explanatory power of RSM- giving it explicit regression coefficients and response surfaces that can be visualized and the flexibility of GBM to deal with nonlinear and high-dimensional data. This means that the GB-RSM hybrid framework will provide a sensible and interpretable modeling approach, which is ideal in optimization studies of agricultural, environmental, and engineering systems that are described by nonlinear, complex, and multivariate interactions.

3.3. Residual and Diagnostic Analysis

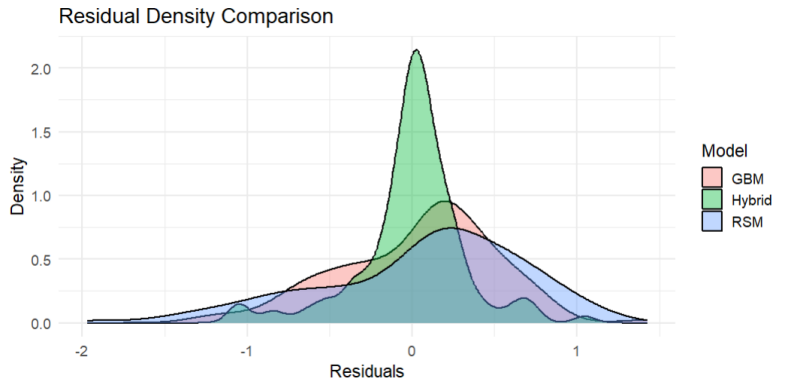


Figure 4: Residual Density Analysis

The comparison plot of the residual density in **Error! Reference source not found.** shows the nature of the distribution of the residuals of the three models: Response Surface Methodology (RSM), Gradient Boosting

https://doi.org/10.64497/jssci.168

Machine (GBM), and hybrid GB-RSM framework. The residuals of the hybrid model have a narrow and sharply peaked distribution around zero, indicating that the majority of the prediction errors are small and evenly distributed. This trend is typical of a high predictive precision model with low bias. Conversely, a broader distribution of the residuals is observed in the RSM model, which shows greater variability and systematic errors of the actual response surface. Residual values of the GBM model are not too far in either direction compared to RSM, although they are more concentrated than the hybrid model. These findings suggest that, although GBM is superior to the pure RSM in terms of nonlinear effects capturing, the hybrid process offers a better balance because the residual dispersion is minimized and the predictions are more likely to cluster around the actual mean. The Q-Q plot indicates that although the data are distributed around the normal line, the tails are far to the right and to the left, which indicates the existence of heavy-tailed behavior and outliers. Collectively, these findings suggest that, despite its reasonable first-order approximation, RSM breaks the normality and homoscedasticity assumptions of regression and hence may not be as reliable in predicting more complex phenomena.

3.4. Residual Diagnostics of the RSM Model

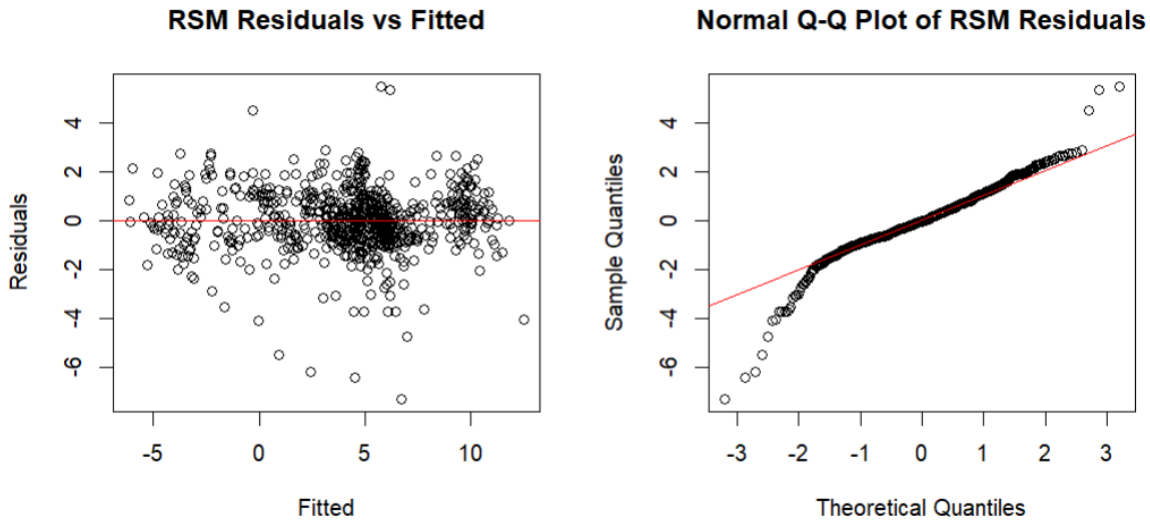


Figure 5: RSM Residuals

Error! Reference source not found. shows the diagnostic plots of the Response Surface Methodology (RSM) model offer understanding on the suitability of the fitted quadratic surface, as well as whether the regression assumptions employed are being met. To determine whether the residuals have systematic patterns which may indicate model misspecification, nonlinearity or heteroscedasticity, the residuals versus fitted values plot (left panel) is utilized. The residuals, in this case, seem to be randomly distributed around the horizontal zero line, and it is a sign that the RSM model is not strongly biased in its structure. The fitted values, however, increase the variability of residuals in a slight degree, which is exhibited by a mild funnel-shaped dispersion. This non-homogeneous variance indicates that there is heteroscedasticity, which means that the model is more likely to generate high prediction error when the fitted response is large. The plot also shows some individual clusters and outliers, which can be due to the effects of some of the experimental runs with extreme factor combinations or interactions that are not well represented by the quadratic model.

The right panel of the normal Q-Q plot also tests the normality assumption of the residues, which is a major requirement for valid statistical inference in RSM. Ideally, when there is a normal distribution in the residuals, the points would be close to the 45-degree reference line. The middle part of the data in this plot compares well

with the theoretical curve, implying that most of the residuals are normally distributed. There are, however, observable deviations at both tails, with the residuals moving downwards (on the left) and upwards (on the right). Such departures imply heavy tails and the existence of a handful of extreme residuals which do not follow normal behavior. The common occurrence of such deviations in the RSM models is when the true response surface possesses a greater nonlinearity or interaction than can be represented by a second-order polynomial.

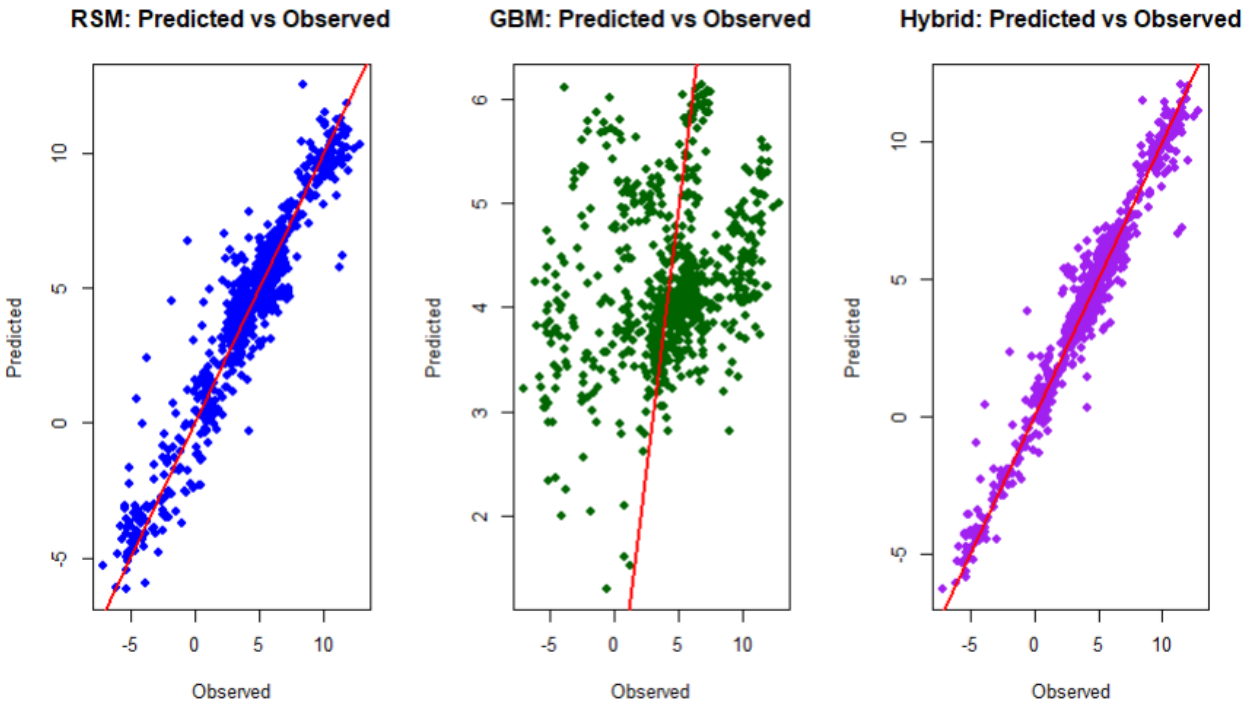


Figure 6: Predicted vs Observed Values for RSM, GBM and Hybrid GB-RSM Models

Error! Reference source not found. is a comparative analysis of three models: Response Surface Methodology (RSM), Gradient Boosting Machine (GBM), and a Hybrid model, using the observed vs. predicted scatter graphs. The primary goal of these kinds of plots is to determine the fit between the model predictions and the real-life data. In a perfect model, all scatter points would fall on the line, and all predictions would be 100 percent accurate and reliable, but this is not the case; this means that the red diagonal line in each scatter plot indicates that the predictions of the model are not 100 percent accurate and reliable. When a point is not on the line, it is an indication of an error in prediction by the model.

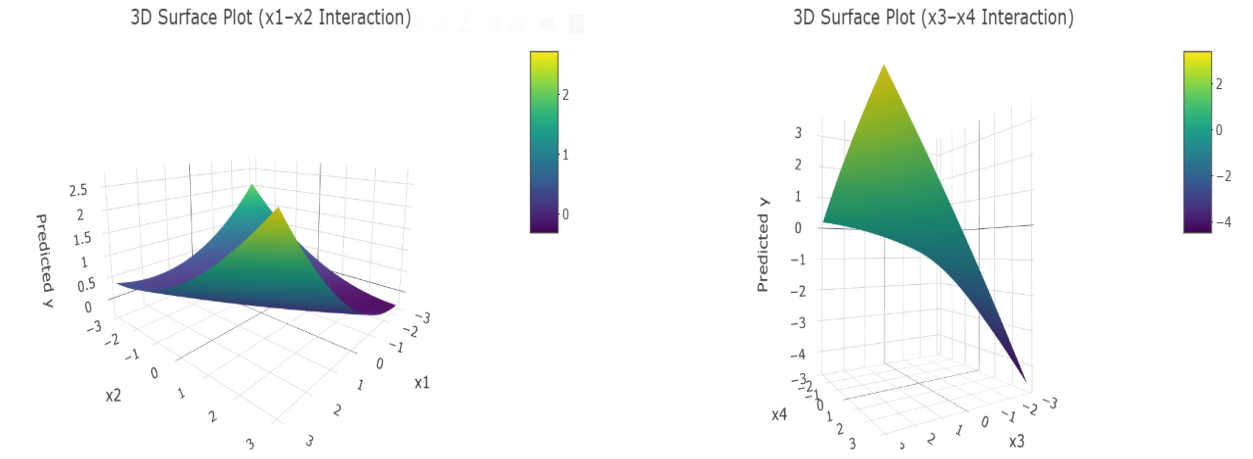


Figure 7: RSM Contour Plots

The two 3D response surfaces in **Error! Reference source not found.** show different patterns of interaction in the fitted response surface model. The initial surface (the relationship between $x_1$and x_2) shows a strong curvature, which is a typical characteristic of a strong second-order effect. This curvature shows that the effect of these factors is not additive but synergistic, i.e., the overall effect of $x_1$and x_2 on the response variable is even larger than the sum of the effects of the two factors. The surface increases evenly at the extremes of the two factors, indicating a nonlinear dependence in which the predicted response is growing with the deviation of the two variables out of the central levels. This is statistically significant quadratic and interaction terms, which lead to model curvature and a stationary region due to an optimal or critical point in the experimental domain.

Conversely, the surface by x_3 and x_4, has a regular gradient structure, which indicates that the behavior of the response has first-order behavioral influence with a minimal nonlinear deviation. The slope on the surface shows opposite direction effects in that an increase in x_3 leads to increased predicted response levels and a corresponding increase in x_4 translates to weakening of the response, indicating opposite main effects and a negative interaction in the system being modeled. Statistically, the association between x_3 and x_4 is almost null, and that linear coefficients are the leading factor in explaining the response pattern. These contrasting surfaces explain the statistical heterogeneity of factor interactions, meaning that the response function is statistically heterogeneous with sensitivities that vary spatially as a result of interaction between nonlinear curvature and directional linearity. The $x_1{:}x_2$ pair interaction shows a strong non-linear variability, whereas the $x_3{:}x_4$ interaction shows a comparatively nonlinear effect exhibiting a dependent structure, hence suggesting that the response surface is hierarchically driven by the $x_1{:}x_2$ pair. The fact that x_1 and x_2 define a region with localized curvature and a global linearity implies that the model should be optimized on the region where x_1 and x_2 are most sensitive to changes and predictive uncertainty is high.

4. Conclusion and Recommendation

Gradient Boosting-Based Response Surface Methodology (GB-RSM) model yielded better results, which was a significant enhancement over the traditional Response Surface Methodology (RSM) and gradient boosting model. The GB-RSM model was robust in dealing with high-dimensional, nonlinear data with missing values by integrating the interpretability of RSM, which models linear, quadratic, and interaction effects, and the adaptive learning ability of Gradient Boosting (GBM). Higher-order interactions and nonlinearities that could not be explained by RSM alone could be captured by the model, which made it a more accurate and robust prediction. The diagnostics of the residual showed that the GB-RSM model was stable with low bias and variance, and thus highly predictive. This is a hybrid method that takes advantage of the strengths of both methods and is especially applicable in complex uses where high accuracy, flexibility, and interpretability are needed.

The GB-RSM hybrid model is recommended to be used in the areas of agriculture, engineering, environmental sciences, and healthcare, where nonlinear relationships, high-dimensional data, and missing values are inherent, due to its better performance in these areas. The model is particularly useful in optimization problems where it is important to know how variables interact, and where both model interpretability and predictive power are of interest. Further studies should be conducted on the optimization of hyperparameters, including the learning rate, count of trees, and the depth of the tree, to enhance the effectiveness and performance of the model. Lastly, the incorporation of other machine learning methods, including neural networks or deep learning, might be considered to work with even more complex data and enhance the performance of a model in a particular application.

Author Contribution: Conceptualization, E. C. Kirui; Methodology, E. C. Kirui; Formal Analysis, E. C. Kirui; Investigation, E. C. Kirui; Data Curation, E. C. Kirui; Writing – Original Draft, E. C. Kirui; Writing – Review & Editing, E. C. Kirui, A. Anapapa and J. Mutuguta; Visualization, E. C. Kirui, A. Anapapa and J. Mutuguta; Supervision, A. Anapapa and J. Mutuguta.

Funding Statement: This research did not receive any funding.

Data availability: The dataset utilized in this research can be provided by the author upon reasonable request.

Acknowledgements: The author would like to thank the Department of Mathematics and Actuarial Science of Murang'a University of Technology, Kenya, for their support and guidance during the entire research process. I would like to extend my special gratitude to Dr. Ayubu Anapapa and Dr. John Mutuguta, who guided and mentored me greatly during the study. Also, special thanks to my lovely wife and daughter, Flossy and Natasha, for their moral support.

Conflict of Interests: Authors declared no conflict of interest.

References

[1] H. Y. Chen and C. Chen, "A study of the response surface methodology model with regression analysis in three fields of engineering," *Applied System Innovation*, vol. 8, no. 4, Art. no. 99, Aug. 2025, doi: 10.3390/asi8040099.

[2] M. Reji and R. Kumar, "Response surface methodology (RSM): An overview to analyze multivariate data," *Indian Journal of Medical Research*, vol. 9, no. 2, pp. 245–252, 2022, doi: 10.18231/j.ijmr.2022.042.

[3] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar, and L. A. Escaleira, "Response surface methodology (RSM) as a tool for optimization in analytical chemistry," *Talanta*, vol. 76, no. 5, pp. 965–977, Sep. 2008, doi: 10.1016/j.talanta.2008.05.019.

[4] Y. Peng, U. Khaled, A. A. A. A. Al-Rashed, R. Meer, M. Goodarzi, and M. M. Sarafraz, "Potential application of response surface methodology (RSM) for the prediction and optimization of thermal conductivity of aqueous CuO (II) nanofluid: A statistical approach and experimental validation," *Physica A: Statistical Mechanics and its Applications*, vol. 554, Art. no. 124353, Sep. 2020, doi: 10.1016/j.physa.2020.124353.

[5] R. K. Dave, T. V. Ramana Rao, and A. S. Nandane, "Improvement of post-harvest quality of pear fruit with optimized composite edible coating formulations," *Journal of Food Science and Technology*, vol. 54, no. 12, pp. 3917–3927, Nov. 2017, doi: 10.1007/s13197-017-2850-y.

[6] A. Ali, R. Jayaraman, E. Azar, and M. Maalouf, "A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework," *Building and Environment*, vol. 252, Art. no. 111268, Mar. 2024, doi: 10.1016/j.buildenv.2024.111268.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[8] T. Domingues, T. Brandão, and J. C. Ferreira, "Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey," *Agriculture*, vol. 12, no. 9, Art. no. 1350, Sep. 2022, doi: 10.3390/agriculture12091350.

[9] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset analysis and feature characteristics to predict rice production based on eXtreme Gradient Boosting," *Journal of Computing Theories and Applications*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.

[10] P. Mahesh and R. Soundrapandiyan, "Yield prediction for crops by gradient-based algorithms," *PLoS One*, vol. 19, no. 8, Art. no. e0291928, Aug. 2024, doi: 10.1371/journal.pone.0291928.

[11] G. S. M. Khamis, Z. M. S. Mohammed, S. M. Alanazi, A. F. A. Mahmoud, F. A. Abdalla, and S. A. Bkheet, "Prediction of myocardial infarction complications using gradient boosting," *Engineering, Technology and Applied Science Research*, vol. 14, no. 6, pp. 18550–18556, Dec. 2024, doi: 10.48084/etasr.9076.

[12] F. Huber, A. Yushchenko, B. Stratmann, and V. Steinhage, "Extreme Gradient Boosting for yield estimation compared with deep learning approaches," *Computers and Electronics in Agriculture*, vol. 202, Art. no. 107346, Nov. 2022, doi: 10.1016/j.compag.2022.107346.

[13] D. Zhao *et al*., "Mapping mangrove leaf area index (LAI) by combining remote sensing images with PROSAIL-D and XGBoost methods," *Remote Sensing in Ecology and Conservation*, vol. 9, no. 3, pp. 370–389, Jun. 2023, doi: 10.1002/rse2.315.

[14] Y. Zhang, Q. Chang, Y. Chen, Y. Liu, D. Jiang, and Z. Zhang, "Hyperspectral estimation of chlorophyll content in apple tree leaf based on feature band selection and the CatBoost model," *Agronomy*, vol. 13, no. 8, Art. no. 2075, Aug. 2023, doi: 10.3390/agronomy13082075.

Disclaimer/Publisher's Note: The views, opinions, and content expressed in all articles are solely those of the respective author(s) and contributor(s) and do not necessarily reflect those of the JSSCI, its editors, or the publisher. JSSCI and its editorial team assume no responsibility for any harm or damage resulting from the use of information, methods, or products mentioned in the publication.