Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

A Comparative Study of Multiple Linear Regression-Clustering-SVM and Fuzzy Linear Regression-Symmetric Parameter Clustering-SVM Hybrid Models in Predicting Colorectal Cancer

Nur Ain Ebas[1,*], Muhammad Ammar Shafi[2], Mohd Saifullah Rusiman[3], Toh Yoke Teng[1], Zeety Md Yusof[1], Nurain Izzati Mohd Yassin[4], Banan Badeel Abdal[5]

[1] Department of Civil Engineering, Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia
[2] Department of Technology Management and Business, Universiti Tun Hussein Onn Malaysia
[3] Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia
[4] Department of Civil Engineering, School of Engineering and Computing, MILA University, Nilai, Negeri Sembilan, Malaysia
[5] College of Administration and Economics, University of Duhok, Iraq

ARTICLE INFO

Article history:
Received 21 January 2025
Received in revised form 2 July 2025
Accepted 23 September 2025
Available online 1 October 2025

Keywords:
colorectal cancer; fuzzy linear regression; multiple linear regression; hybrid model; error metrics

ABSTRACT

Colorectal cancer (CRC) remains a leading cause of mortality worldwide, with early detection being crucial for improving patient outcomes. In order to predict the colorectal cancer, this study compares two hybrid machine learning models, which are Multiple Linear Regression Clustering with Support Vector Machine (MLRCSVM) and Fuzzy Linear Regression with Symmetric Parameter Clustering with Support Vector Machine (FLRWSPCSVM). Secondary data was obtained from a general hospital in Kuala Lumpur. It includes 180 colon cancer patients as respondents, with data collected and recorded by nurses using cluster sampling. The size of the tumor is the dependent variable, while colorectal cancer symptoms and factor are the independent variables. Mean square error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE) are used to evaluate the performance of these models. The results indicate that FLRWSPCSVM outperforms MLRCSVM in terms of accuracy and robustness in handling uncertain or noisy data, highlighting its potential as a powerful tool for early colorectal cancer diagnosis.

1. Introduction

Colorectal cancer (CRC) is the most prevalent type of cancer that affects the digestive tract. It is the second most common cause of cancer-related death worldwide and the third most common cancer in both men and women. Given the poor prognosis of CRC, it is of great importance to make a more accurate prediction of this disease. According to Yadav & Kumar [1], early diagnosis, appropriate treatment strategy, accurate assessment of treatment response, and correct prognosis are essential for a better result. Research conducted by Phillips et al., [2] identifies factors linked to colorectal cancer screening initiation at age 50 among individuals aged 50-75, using descriptive and

* *Corresponding author.*
E-mail address: afiqahjamil@uitm.edu.my

logistic regression methods. According to the study, there is limited widespread screening for average-risk individuals in Malaysia due to the low prevalence of colorectal cancer. Yusoff et al., [3] investigated participation and obstacles to colorectal cancer screening in Malaysia, highlighting that it is the most common cancer in males and the third most common in females.

According to Sawicki et al., [4], among all cancers, lung cancer ranks first across both genders, accounting for approximately 11.6% of total cases. For women, breast cancer shares the same percentage, while prostate cancer is the second most common type of cancer in men at 7.1%. Colorectal cancer (CRC) stands as the third most common cancer (6.1%) and the second leading cause of cancer-related deaths (9.2%). A study by Douaiher et al., [5] suggests that by 2035, fatalities from rectal and colon cancer will rise by 60% and 71.5%, respectively. Hannisdal & Thorsen [6] analyzed prognostic factors such as Dukes' stage, age, and inflammatory markers using regression techniques, emphasizing the importance of these variables in survival outcomes. By examining the correlations between a dependent variable and several independent factors, multiple linear regression (MLR) makes it easier to identify important predictors of colorectal cancer [7]. While hybrid MLR models offer valuable insights, they face challenges, including model assumptions and the need for robust diagnostic procedures to ensure accurate and reliable predictions.

Fuzzy concept capable of managing uncertainty data that is not precise to a specific point value. There are a lot of researchers' studies that involve fuzzy. One of the researchers, Nopour et al., [8], develops a fuzzy logic-based clinical decision support system (FL-based CDSS) to identify CRC patients. Chaira [9] studied the novel intuitionistic fuzzy c-means clustering method using intuitionistic fuzzy set theory effectively clusters that group CT scan brain pictures efficiently and aid in the identification of abnormalities. Ramathilagam et al., [10] proposed a fuzzy probabilistic c-means method to effectively identify cancer subtypes in colon cancer datasets, improving clustering accuracy and enhancing interpretability of the structure. Fuzzy c-means is one of the most popular ongoing areas of research among all types of researchers, including computer science, mathematics, and other areas of engineering, as well as all areas of optimization practices. Nayak et al., [11] conducted a survey on FCM and its applications in more than one decade to show the efficiency and applicability in a mixture of domains. Early CRC detection using computational technologies can significantly improve the overall survival possibility of patients.

In predicting colorectal cancer (CRC) risk, hybrid fuzzy linear regression models are becoming increasingly important, especially when dealing with uncertain and variable patient data. These models combine fuzzy logic with conventional regression methods to improve predictive accuracy and ease of interpretation. Zeng & Zheng [12] mentioned that by estimating parameters with precise inputs and fuzzy outputs, fuzzy linear regression can minimize least square errors to draw significant conclusions. According to Bisserier et al., [13], a modified version of the fuzzy linear model can encompass observed data, ensuring a comprehensive representation of output evolution, which is vital for assessing CRC risk. In hereditary non-polyposis colorectal cancer patients, fuzzy modeling has been successfully employed to predict CRC risk, uncovering significant associations between genetic mutations, smoking, and CRC risk by Brand et al., [14]. Zhang et al., [15] proved the integration of fuzzy regression with machine learning algorithms enhances diagnostic capabilities, achieving high levels of sensitivity and specificity in detecting CRC. Both models perform well. Hence, this study aimed to compare FLRWSPCSVM and MLRCSVM to predict colorectal cancer.

2. Methodology

2.1 Data Scope

This research utilized secondary data from a general hospital in Kuala Lumpur, involving 180 colon cancer patients as respondents. Nurses collected and documented the data using cluster sampling. The study incorporated continuous data, including dependent and independent variables. The tumor size served as the dependent variable, while independent variables consist of twenty-five variables about the symptoms and factors involved. All the twenty-five variables are explained in Table 1. The tumor sizes investigated ranged from 20mm to 100mm. Doctors conducted face-to-face interviews with patients regarding their colorectal cancer experiences, obtaining immediate responses. The questionnaire consists of twenty-four variables, excluding TNM staging. Colorectal cancer staging focuses on the tumor's size and its spread within the human bowel from a polyp. Medical professionals used tests and scans to determine the cancer stage in patients. Tumor, node, and metastases (TNM) are widely employed in oncology worldwide. According to the American Cancer Society [16], TNM describes the primary tumor's size (T), the presence of cancer cells in lymph nodes (N), and cancer spreading to other body parts (M).

Table 1
A detailed explanation of the variables used

Variables	Variable Name	Variables Category
Y	Tumor size	Dependent variable
x_1	Gender	Factor
x_2	Age	Factor
x_3	Ethnic	Factor
x_4	Family history	Factor
x_5	Small bowel	Symptom
x_6	Weight loss	Symptom
x_7	Diarrhea	Symptom
x_8	Anemia	Symptom
x_9	Blood stool	Symptom
x_{10}	Abdominal pain	Symptom
x_{11}	icd10 (Place where CRC existed by patient)	Symptom
x_{12}	TNM	Symptom
x_{13}	Diabetes Mellitus	Symptom
x_{14}	Crohn's Disease	Symptom
x_{15}	Ulcerative collitis	Symptom
x_{16}	Polyp	Symptom
x_{17}	History of cancer(s)	Factor
x_{18}	Endometrial	Symptom
x_{19}	Gastric	Symptom
x_{20}	Urinary tract	Symptom
x_{21}	Hepatobiliary	Symptom
x_{22}	Ovarian	Symptom
x_{23}	Other cancer	Factor
x_{24}	Intestinal	Symptom
x_{25}	Colorectal	Symptom

2.2 Model Scope

This study employs r-value and the coefficient of determination (r^2) to assess the correlation within a dataset explained by a statistical model. Subsequently, multiple linear regression, as a linear model, must satisfy certain assumptions before analysis can proceed. These assumptions include constant variance, normal distribution, and absence of multicollinearity. Following this, the analysis will focus on identifying significant variables and calculating metric errors. Identifying significant variables is crucial for determining factors and symptoms that provide valuable information for early detection and prediction of colorectal cancer. The MLRCSVM model consists of three stages. The first stage involves with multiple linear regression (MLR) to handle the linear relationships between features. Then, clustering is applied to segment the data into k clusters based on feature similarity, and the last stage is SVM classification to classify the data into CRC and non-CRC categories based on the clusters formed.

Furthermore, various fuzzy linear models will be applied, including fuzzy linear regression (FLR), fuzzy linear regression with symmetric parameter (FLRWSP), fuzzy linear regression with asymmetric parameter (FLRWAP), and fuzzy c-means method. Among these, FLRWSP is expected to perform best based on MSE, RMSE, MAE, and MAPE values. This model will be combined with fuzzy c-means clustering to create an optimal model for predicting colorectal cancer tumor size. Additionally, support vector machines (SVM) will be utilized to enhance tumor size prediction accuracy. These methods aim to minimize errors in MSE, RMSE, MAE, and MAPE values for predicting colorectal cancer tumor size. The model yielding the lowest values across these metrics will be selected as the optimal model for predicting colorectal cancer tumor size. Achieving the best model is crucial for obtaining an approximation that closely predicts tumor size, particularly in the early stages (stages I and II) of colorectal cancer.

2.2.1 Multiple Linear Regression

In colorectal cancer (CRC) research, multiple linear regression (MLR) plays as a significant tool for analyzing and forecasting outcomes. This method is applied to examine the relationships between colorectal cancer tumor size and the symptoms and factors involved. Yang et al., [17] have utilized MLR to evaluate prognostic elements influencing post-surgery survival rates in CRC patients, where essential independent factors discovered include Dukes stage and tumor differentiation. The research reported 3- and 5-year survival rates of 63.2% and 60.8%, respectively, underscoring MLR's significance in survival analysis. In a specific study by Shafi et al., [18], MLR was applied to assess mortality rates across various CRC stages.

According to Kunter et al., [19], it is essential to consider and fulfill the assumptions of a regression model before its implementation. Brant [20] outlined the key assumptions for multiple linear regression models. To ensure reliable outcomes in multiple linear regression, several conditions must be met. Firstly, there should be a linear relationship between the independent and dependent variables, with changes in predictors having a direct, proportional effect on the outcome. Secondly, the error variance should remain constant across all predictor levels (homoscedasticity), ensuring consistent predictions. Thirdly, multicollinearity should be avoided, meaning that independent variables should not be highly correlated with each other, as this can obscure their individual impacts. Lastly, the residuals of the model should be normally distributed to guarantee accurate and easily interpretable results.

According to Osborne [21], the multiple linear regression model can be expressed in Eq. (1).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + \varepsilon \quad (1)$$

where Y_i represents the dependent variable, $\beta_0, \beta_1, \beta_2, \beta_j$ are constant values, and $X_{i1} \ldots X_{ij}$ denote unknown parameters or independent variables. The findings revealed that extended fuzzy correlation and regression analysis outperformed traditional MLR in this context.

2.2.2 Fuzzy Linear Regression

In the realm of colorectal cancer (CRC) research, fuzzy linear regression (FLR) has proven to be an invaluable method for predicting outcomes, especially when dealing with imprecise and uncertain data. This technique is particularly useful in clinical environments where patient information may be ambiguous. Tanaka [21] introduced the concept of fuzzy linear regression. His research explored the use of fuzzy linear functions in regression analysis for ambiguous phenomena. Typically, in regression models, discrepancies between observed and estimated values are attributed to measurement errors. Additionally, according to Tanaka [21], these discrepancies were thought to be influenced by the ambiguity of the system's structure.

Consider two sets, X and Y, and a function $f(x, a)$ that maps X to Y. When parameters are represented by fuzzy sets A, the function is termed a fuzzy function, denoted as $f(x, A)$. For a given x, the fuzzy set $Y = f(x, A)$ is mapped from the fuzzy set A and defined as in Eq. (2).

$$f: X \to \xi(y); \quad Y = f(x, A) \quad (2)$$

where $\xi(y)$ is the set of all fuzzy subsets on Y. The fuzzy set Y is defined by the membership function in Eq. (3).

$$\mu_Y(y) = \begin{cases} max_{\{a|y=f(x,a)\}} & \{a|y = f(x,a)\} \neq \emptyset \\ 0 & otherwise \end{cases} \quad (3)$$

where $\mu_A(a)$ is membership function and a is a fuzzy set on the product space of parameters. Eq. (1) was given x and its image was A. The fuzzy parameters were assumed to be limited types of fuzzy sets. Fuzzy parameters were defined by fuzzy sets as illustrated in Figure 1. The fuzzy sets are in Eq. (4) and Eq. (5).

$$\mu_A(a) = \min[\mu_{A_f}(a_f)] \quad (4)$$

where

$$\mu_{A_f}(a_f) = \begin{cases} \frac{1-|\alpha_f - a_f|}{\Im_f}; & \alpha_f - \Im_f \leq \alpha_f \leq \alpha_f + \Im_f \\ 0 & otherwise \end{cases} \quad (5)$$

and $\Im_f > 0$.

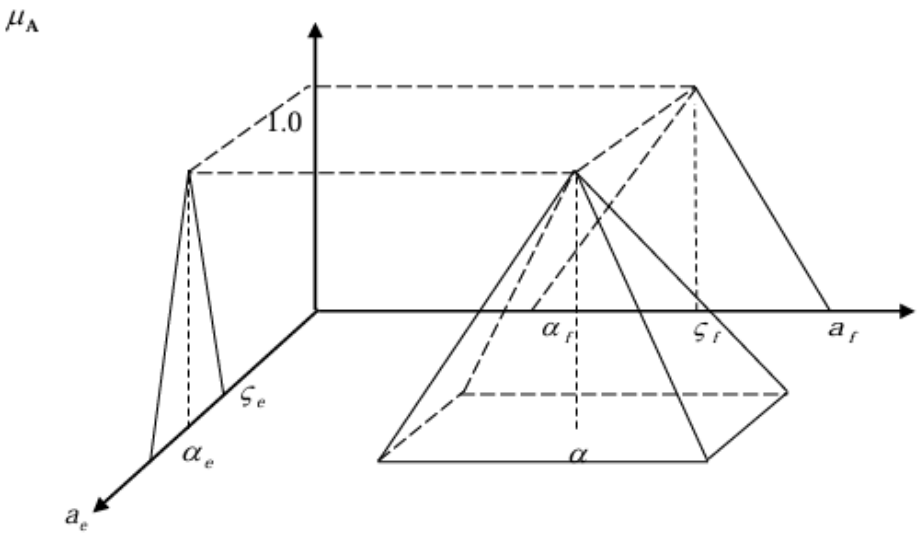

Fig. 1. Fuzzy set of parameters

Researchers have employed FLR to forecast tumor's size in CRC patients, showing enhanced precision compared to conventional linear regression techniques. Zolfaghari [22] revealed that utilizing FLR with fuzzy data produced superior predictions, with the fuzzy linear regression model achieving the lowest metric errors.

2.2.3 MLRCSVM Model

A combined approach utilizing MLR and support vector machines was created to estimate tumor size in CRC, exhibiting improved performance with reduced mean square error (MSE) and root mean square error (RMSE) compared to conventional MLR, according to Shafi et al., [18]. By combining MLR with other approaches like support vector machines, it enhances predictive accuracy, making it an invaluable resource in clinical environments. There are assumptions needed in this hybrid model. The three assumptions apply, such as constant variance, normality, and multicollinearity. Modeling of MLR clustering is a combination of the MLR and FCM methods. SVM model to be chosen to hybrid because SVM is a linear model that is not too sensitive to outliers and can minimize the error of the model. Furthermore, the SVM model used different software to find the residual, and the software used is Weka Explorer software. The calculation of SVM model residual is explained as in Eq. (6).

$$\varepsilon_S = Y_S - \hat{Y}_S \tag{6}$$

where ε_S is residual of SVM, Y_S is observation data of SVM and $\hat{Y}_S$ is the prediction data of SVM. According to Shafi et al., [23], there are five steps involved in making a new hybrid MLRCSVM model: Step 1: Identify the higher value of correlation, r between Y and X_i.

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}} \tag{7}$$

where n is number of samples, X is independent variable, Y is dependent variable. If there are many values of r between Y and X_i, find several Y and X_i with a higher value of r or consider the Y and X_i with r is higher than 0.40.
Step 2: The first stage of hybrid is the modeling of MLR clustering, which is a combination between MLR and fuzzy C-means. The combination of clustering between MLR and FCM is based on Y data alone and Y data toward independent variables, which have a higher correlation value. The best of MLR clustering will be chosen based on the smallest values of MSE and RMSE.
Step 3: Determine the residual of MLR clustering and SVM.
Step 4: The second stage of hybrid is making the new hybrid data using the equation in Eq. (8).

$$Y_{MLRCSVM} = L_M + N_S \tag{8}$$

where $Y_{MLRCSVM}$ is a new dataset, L_M is the residual of MLR clustering (linear model), and N_S is the residual of SVM model (linear model).
Step 5: Modeling a hybrid model using the MLR method and the SVM method, where the final new error is in Eq. (9).

$$ERROR_{final} = \frac{(n_1 \times ERROR_{MLR}) + (n_2 \times ERROR_{SVM})}{n_1 + n_2} \tag{9}$$

where $n_1$and n_2 are the numbers of data for cluster 1 and cluster 2, respectively. $ERROR_{MLR}$ is the number error of MLR clustering, and $ERROR_{SVM}$ is the SVM model's error number. Error values would be the values of MSE, RMSE, MAE, and MAPE.

2.2.2 FLRWSPCSVM Model

There is no assumption needed in this hybrid model. Modeling of FLRWSP clustering is a combination of FLRWSP and FCM method. The combination between FLRWSP and the FCM method is based on the higher value of r. Shafi et al., [23] introduced a new hybrid FLRWSPCSVM model to predict colorectal cancers. According to Shafi et al., [23], there are five steps involved in creating a hybrid model shown below:
Step 1: Identify the correlation with the greater value Y versus X_i as shown in Eq. (7).
Step 2: The modeling of FLRWSP clustering, which combines FLRWSP and fuzzy C-means with $h = 0.5$ in FLRWSP, is the first stage of the hybrid. The lowest MSE and RMSE values among five higher correlation variables will be chosen using the FCM method.
Step 3: Find the FLRWSP clustering and SVM residual.
Step 4: the second stage of hybrid is making the new hybrid data using the equation in Eq. (10).

$$Y_{FLRWSPCSVM} = L_F + N_S \tag{10}$$

where $Y_{FLRWSPCSVM}$ is a new dataset, L_F is the residual of FLRWSP clustering (non-linear model), and N_S is the residual of SVM model (linear model). SVM model to be chosen to hybrid because SVM is a linear model that is not too sensitive to outliers and can minimize the error of the model.
Step 5: Modeling a hybrid using the FLRWSP method, the final error can be measured as in Eq. (11).

$$ERROR_{FLRWSPCSVM} = \frac{(n_1 \times ERROR_1) + (n_2 \times ERROR_2) + \cdots + (n_j \times ERROR_j)}{n_1 + n_2 + \cdots + n_j} \tag{11}$$

where n_1, n_2, n_j are the number of data for cluster 1, cluster 2 until cluster j respectively. $ERROR_1$, $ERROR_2, ERROR_j$ is the error of FLRWSPCSVM model cluster 1, cluster 2 until cluster j respectively. Error values would be the values of MSE, RMSE, MAE and MAPE. This new hybrid method was introduced by Shafi, et al., [23].

2.3 Cross Validation Technique

Evaluating predictive models relies heavily on cross-validation techniques, which employ various metrics to assess performance and generalizability. These metrics include MSE, RMSE, MAE and MAPE. According to Tarekegn et al., [24] MSE measures the average squared difference between predicted and actual values, giving more weight to larger errors, while RMSE provides error in the same units as the output, enhancing interpretability.

MSE is represented by Eq. (12), with lower values indicating better predictive performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

RMSE, shown in Eq. (13), is more easily understood due to the same units as the target variable. Lower RMSE values signify better performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

MAE, as shown in Eq. (14), offers a straightforward interpretation by calculating the average of absolute differences. It demonstrates greater resilience to outliers compared to MSE and RMSE. Superior model performance is indicated by lower MAE values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

MAPE, depicted in Eq. (15), is beneficial for assessing error magnitude relative to actual values. Enhanced predictive accuracy is reflected by a reduced MAPE value. This metric is particularly valuable when comparing models across datasets with different scales.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{15}$$

Although these metrics are commonly employed, they have certain limitations. MSE and RMSE can be sensitive to outliers, while MAPE may be challenging to interpret when actual values approach zero. Consequently, selecting the appropriate metric is essential for accurate model evaluation.

3. Results

3.1 Evaluation of MLR and SVM Hybrid Model Performance for Predictive Analysis

Table 2 shows independent variable, x_3 is chosen for clustering because it reached the smallest error value compared to the other values with MSE = 23.476.

Table 2
Correlation between Y and X_i using MLR

Correlation	Correlation value	MSE value
$Y - x_5$	0.942	119.194
$Y - x_4$	0.914	119.193
$Y - x_3$	0.000	23.476
$Y - x_2$	0.820	119.188
$Y - x_1$	0.225	119.018

The estimate MLRC model equation as in Eq. (16).

$$\hat{Y} = 0.969 + 0.972x_3 \tag{16}$$

Table 3
Summary of the MLRC

Methods	Value
MSE	19.144
RMSE	4.375

There are 1000 data in the residual of SVM. The maximum value is 35.38 while the minimum value is -47.38. The residual of SVM model is shown in Figure 2.

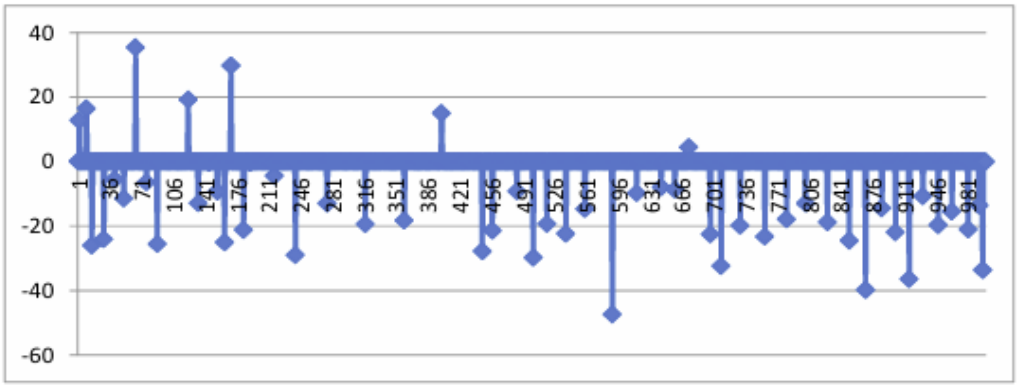

Fig. 2. Residual of SVM model

Table 4 shows the summary of validation techniques using MSE, RMSE, MAE and MAPE for the MLRCSVM model.

Table 4
MSE, RMSE, MAE and MAPE value of the MLRCSVM model

Methods	Value
MSE	1.842
RMSE	1.357
MAE	0.813
MAPE	1.853

3.2 Optimization of Fuzzy Linear Regression with Symmetric Parameters (FLRWSP) for Predictive Accuracy

Zolfarhani et al., [22] introduced the FLRWSP model utilizing FLR. The study incorporated five predictor variables and one dependent variable, focusing on prediction data. Matlab software was employed to determine MSE and RMSE values, followed by optimizing the degree of fitting triangular fuzzy "h". The optimal prediction model for simulation data was identified by the h value yielding the lowest MSE and RMSE. Tables 5 present the analysis results of MSE and RMSE values by degree of fitting while Table 6 presents the analysis results of fuzzy parameter focusing on, $h = 0.8$.

Table 5
MSE and RMSE value by degree of fitting (h value)

h	MSE values	RMSE values
0.0	43094.705	207.593
0.1	109137.966	330.360
0.2	24675.711	157.083
0.3	1025.103	32.017
0.4	475.419	21.804
0.5	183.311	13.539
0.6	196.127	14.005
0.7	166.221	12.926
0.8	164.566	12.823
0.9	187.289	13.685
1.0	60905537.54	7804.199

Table 6
Fuzzy parameter, $h = 0.8$, Zolfarhani et al., [22]

Fuzzy parameter	Center, α_i	Width, c_i
Constant	5.9202	0
A_1	0.305	2.733
A_2	-0.015	0.054
A_3	-0.011	0
A_4	0.356	0
A_5	0.545	0.171

The fuzzy parameter outcomes revealed that $h = 0.8$ provided the best prediction in simulation data, using the smallest possible degree of fitting triangular fuzzy. α_i is a center of fuzzy parameter and c_i denotes the fuzziness of parameter (width). Table 6 displays the fuzzy parameter values for the five predictor variables. The fuzzy mean value of tumor size (mm) can be explained by A_5 with the highest fuzzy parameter= 0.545. The vagueness of simulation data can be explained by A_1, A_2 and A_5. Moreover, the fact that A_2 and A_3 were negative depends on the correlations between A_2 and A_3.

The estimated parameter of fuzzy linear regression with symmeric parameter model for simulation data is as in Eq. (17).

$$\hat{Y} = 5.920 + (0.305, 2.733)A_1 + (-0.015,0.054)A_2 + (-0.011,0)A_3 + (0.356,0)A_4 + (0.545,0.171)A_5 \quad (17)$$

The FLRWSP model, combining fuzzy linear regression with symmetric parameter and fuzzy c-means, was selected due to its lower error values compared to other fuzzy linear regression models in this study, with an MSE of 274.007 and RMSE of 16.553. Fuzzy c-means clustering does not require preceding assumptions before analysis. The study analyzed 1000 observations of raw data using Microsoft Excel and Matlab software.

Table 7
MSE value for independent variables chosen toward dependent variables

	Cluster 1	Cluster 2	MSE value
$Y - A_5(h$ =0.5)	20.417 (496 data)	28.521 (504 data)	24.500
$Y - A_4(h$ =0.1)	21.057 (515 data)	27.801 (485 data)	24.327
$Y - A_2(h$ =0.0)	19.484 (501 data)	29.969 (499 data)	24.716
$Y - A_1(h$ =0.2)	20.540 (529 data)	27.315 (471 data)	23.731
$Y - A_3(h$ =0.3)	20.128 (476 rows of data)	28.817 (524 rows of data)	24.681

Table 7 shows the comparison MSE value between Y and A_5, A_4, A_3, A_2, and A_1. Based on the Table 7, $Y - A_1(h$ =0.2) is chosen as the best clustering with the smallest value of error MSE value 23.731.

3.3 Modelling a hybrid data using FLRWSP method and SVM method

In FLRWSPCSVM models, procedure 1-5 mentioned in Section 2.2.2 needs to be fulfilled in order to get better results. The summary of the model shown in Table 8 proves that FLRWSPCSVM is the best model for simulation data prediction with the lowest MSE and RMSE values of 1.783 and 1.335 compared to other models such as MLR, FLR, FLRWSP, FLRWAP, and SVM.

Table 8
MSE, RMSE, MAE and MAPE value of the FLRWSPCSVM model

Methods	Value
MSE	1.783
RMSE	1.335
MAE	0.764
MAPE	1.594

Table 9 summarizes the performance metrics including MSE, RMSE, MAE, and MAPE for several predictive models based on simulation data.

Table 9
Summary of MSE, RMSE, MAE and MAPE value for all models (simulation data)

Models	MSE	RMSE	MAE	MAPE
FLRWSP	183.311	13.539	10.851	25.093
FLRWAP	79738.326	282.379	276.898	552.996
SVM	24.137	4.912	1.049	2.009
MLRCSVM	1.842	1.357	0.814	1.853
FLRWSPSVM	1.783	1.335	0.764	1.594

Table 9 indicates that the hybrid FLRWSP clustering with SVM model demonstrates the lowest MSE (1.783), suggesting it best fits the simulation data. In contrast, the FLRWAP model exhibits the highest MSE (79738.326), indicating a poor fit relative to other models, which is also reflected in the RMSE results. The hybrid FLRWSP clustering with SVM model also shows the smallest MAE (0.764), further confirming its predictive accuracy. Conversely, the FLRWAP model displays the largest MAE (276.898), implying substantial errors in predicting actual values. Additionally, the hybrid FLRWSP clustering with SVM model has the lowest MAPE (1.594), signifying the least percentage deviation from actual data, while the FLRWAP model shows the highest MAPE (552.996), indicating significant predictive deviations. Across all performance metrics (MSE, RMSE, MAE, and MAPE), the hybrid FLRWSP clustering with SVM model surpasses other models, consistently showing the smallest error values. This performance suggests it is the most precise and dependable model for predicting the given simulation data. In stark contrast, the FLRWAP model consistently underperforms, exhibiting the largest errors across all metrics. Both models perform well, but hybrid FLRWSP clustering with SVM appears to have a slight edge in terms of overall accuracy and error minimization.

4. Conclusions

The study indicates that the FLRWSPCSVM model surpasses MLRCSVM in colorectal cancer prediction, particularly when dealing with uncertain and noisy data. While MLRCSVM performed adequately with datasets exhibiting clear linear relationships, it encountered difficulties with the noise and uncertainty typical of medical data. In contrast, FLRWSPCSVM excelled under these conditions, leveraging fuzzy logic and symmetric parameter clustering to manage imprecision effectively. However, FLRWSPCSVM's enhanced performance came at the cost of increased computational complexity. The results suggest that FLRWSPCSVM could provide improved diagnostic accuracy for CRC screening, especially in scenarios with variable data quality. Its resilience in noisy environments makes it well-suited for practical applications. This investigation has significant implications for both medical and machine learning fields. In medicine, the findings could lead to more precise and dependable CRC screening tools, potentially enhancing patient outcomes. From an artificial intelligence perspective, the study underscores the efficacy of hybrid models in addressing complex, real-world prediction challenges.

Acknowledgement

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Multidisciplinary Research Grant (MDR) vot (Q703).

References

[1] Yadav, A., & Kumar, A. (2023). "Artificial intelligence in rectal cancer: What is the future*?" Artificial Intelligence in Cancer*. no. 4(2) (2023): 11-22. https://doi.org/10.35713/aic.v4.i2.11

[2] Phillips, K. L., Smith, L. M., Ahn, S., Ory, M. G. and Hochhalter, A. K., "Correlates of initiating colorectal cancer screening beginning at Age 50" *Journal Community Health*, no.19 (2013): 23-30. https://digitalcommons.memphis.edu/health-systems-mgmt-policy-div-facpubs/19

[3] Yusoff, H. M., Norwati, D., Norhayati, M. N. and Amry, A. R. "Participants and Barriers to Colorectal Cancer Screening in Malaysia" *Asian Pacific Journal Cancer Prevention*, no. 13 (2012):, 3983-3987. https://doi.org/10.7314/APJCP.2012.13.8.3983

[4] Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedźwiedzka, E., Arłukowicz, T., Przybyłowicz, K. E. "A Review of Colorectal Cancer in Terms of Epidemiology, Risk Factors, Development, Symptoms and Diagnosis" *Cancers*. No. 13(9)(2021):2025. https://doi.org/10.3390/cancers13092025

[5] Douaiher J., Ravipati A., Grams B., Chowdhury S., Alatise O., and Are C. "Colorectal cancer-global burden, trends, and geographical variations". *Journal of surgical oncology*, no. 115 (2017):619–630. https://doi.org/10.1002/jso.24578
[6] Hannisdal, E. and Thorsen, G. "Regression analyses of prognostic factors in colorectal cancer" *Journal of Surgical Oncology*, no. 37 (1988): 109 – 112. https://doi.org/10.1002/JSO.2930370209
[7] Grégoire, G. "Multiple Linear Regression". *Eas Publications Series*, no. 66 (2014): 45 – 72. https://doi.org/10.1051/EAS/1466005
[8] Nopour, R., Shanbehzadeh, M., & Kazemi-Arpanahi, H. "Developing a clinical decision support system based on the fuzzy logic and decision tree to predict colorectal cancer." *Medical Journal of the Islamic Republic of Iran*, 35 (2021): 44 - 44. https://doi.org/10.47176/mjiri.35.44.
[9] Chaira, T. "A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images." *Applied Soft Computing*, 11 (2011): 1711-1717. https://doi.org/10.1016/j.asoc.2010.05.005
[10] Ramathilagam, S., Kannan, S., and Devi, R."Effective Fuzzy Possibilistic C-Means: An Analyzing Cancer Medical Database", *Soft Computing,* vol. 21 (2017): 2835–2845. https://doi.org/10.1145/2818869.2818870.
[11] Nayak, J., Naik, B., & Behera, H. "Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014" ,(2015): 133-149. https://doi.org/10.1007/978-81-322-2208-8_14.
[12] Zhang, B., Liang, X. L., Gao, H. Y., Ye, L. S. and Wang, Y. G. "Models of logistic regression analysis, support vector machine, and back-propagation neural network based on serum tumor markers in colorectal cancer diagnosis." *Genetics and Molecular Research,* no. 15(2016). https://doi.org/10.4238/GMR.15028643
[13] Bisserier, A., Boukezzoula, R., Galichet, S. (2010). Linear Fuzzy Regression Using Trapezoidal Fuzzy Intervals. In: Bouchon-Meunier, B., Magdalena, L., Ojeda-Aciego, M., Verdegay, JL., Yager, R.R. (eds) Foundations of Reasoning under Uncertainty. Studies in Fuzziness and Soft Computing, vol 249. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10728-3_1
[14] Brand, R. M., Jones, D. D., Lynch, H. T., Brand, R. E., Watson, P., Ashwathnayaran, R., and Roy, H. K. "Risk of colon cancer in hereditary non-polyposis colorectal cancer patients as predicted by fuzzy modeling: Influence of smoking." *World journal of gastroenterology,* 12(28) (2006): 4485–4491. https://doi.org/10.3748/wjg.v12.i28.4485
[15] Zeng, W. and Zheng, X. "Fuzzy Linear Regression Model." *2008 International Symposium on Information Science and Engineering* no. 1 (2008): 183-187.https://doi.org/10.1109/ISISE.2008.143
[16] American Cancer Society. (2012). How is colorectal cancer staged? American Cancer Society.
[17] Yang, Z., Wang, J., Wang, L., Dong, W., Huang, Y, Qin, J., Zhan, W. (2003). Multivariate regression analysis of prognostic factors in colorectal cancer. The Chinese-german Journal of Clinical Oncology, http://dx.doi.org/10.1007/BF02842287
[18] Shafi, M., Rusiman, M. S., Ismail, S. and Kamardan, M.G. (2019). A hybrid of Multiple Linear Regression Clustering model with support vector machine for colorectal cancer tumor size prediction. International Journal of Advanced Computer Science and Applications. 10. 323-328. http://dx.doi.org/10.14569/IJACSA.2019.0100439
[19] Kutner, M.H. and Nachtsheim, C. and Neter, J. 4th Ed. Applied linear regression models (McGraw-Hill/Irwin,2004)
[20] Brand, Rhonda M., David D. Jones, Henry T. Lynch, Randall E. Brand, Patrice Watson, Ramesh Ashwathnayaran, and Hemant K. Roy. Risk of colon cancer in hereditary non-polyposis colorectal cancer patients as predicted by fuzzy modeling: Influence of smoking. World journal of gastroenterology: WJG 12, no. 28 (2006): 4485.
[21] Osborne, Jason W., and Elaine Waters. "Four assumptions of multiple regression that researchers should always test." Practical assessment, research, and evaluation 8, no. 1 (2019): 2.
[22] Zolfaghari, Z. S., Mohebbi, M. and Najariyan, M. (2014). Application of Fuzzy Linear Regression Method for Sensory Evaluation of Fried Donut. Applied Soft Computing, (22), 417-423. https://doi.org/10.1016/j.asoc.2014.03.010
[23] Shafi, M., Rusiman, M. S., Muhamad Jamil, S. A. and Mohd Zim, M. A. "The prediction of high-risk symptom for colorectal cancer using a new hybrid of fuzzy statistical machine learning approach. "*AIP Conference Proceeding*. 3123 (1)(2024): 020017. https://doi.org/10.1063/5.0225096
[24] Tarekegn, A. N., Michalak, K. and Giacobini, M. "Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study Using Multi-Label Datasets." SN COMPUT. SCI. no. 1, 263 (2020). https://doi.org/10.1007/s42979-020-00283-z