New Mathematics and Natural Computation
(2026) 1–21

DOI: 10.1142/S1793005726500365

Clustering-Based Hybrid Model for Predicting Symptoms for Colorectal Cancer: A Fuzzy Machine Learning Approach

M. A. Shafi*, Sayooj Aby Jose†,§,¶, M. S. Rusiman‡ and Anuwat Jirawattanapanit§

*Department of Technology and Management, Faculty of Technology Management and Business
Universiti Tun Hussein Onn Malaysia
Malaysia

†School of Mathematics and Statistics
Mahatma Gandhi University
Kottayam, Kerala

‡Department of Mathematic and Statistics
Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia, Malaysia

§Department of Mathematics, Faculty of Education
Phuket Rajabhat University, Phuket, Thailand
¶sayooaby999@gmail.com

Received 14 November 2023
Accepted 13 December 2024
Published 29 January 2025

Colorectal cancer (CRC) is a form of cancer that originates in the colon (large intestine) or rectum, which are components of the digestive system. It generally begins as small, benign growths known as polyps that can form on the inner lining of the colon or rectum. In Malaysia, colorectal cancer ranks among the most prevalent cancers, especially within the Chinese and Malay communities. A study released by the Malaysian National Cancer Registry indicates that colorectal cancer is the second most common cancer type following breast cancer, impacting both genders. The occurrence of colorectal cancer in Malaysia has been consistently increasing, with approximately 15–20% of cancer cases in the nation being colorectal cancer. This type of cancer arises when cells in the body start to multiply excessively, leading to various symptoms. In this research, a novel hybrid fuzzy linear regression with symmetric parameter clustering combined with a support vector machine (FLRWSPCSVM) model is employed to predict the high-risk symptoms associated with the onset of colorectal cancer in Malaysia. The study analyzed secondary data from 180 patients diagnosed with colorectal cancer and treated in a general hospital in Kuala Lumpur, considering twenty-five independent variables with various combinations of variable types. Furthermore, the model included parameters, errors, and explanations, along with two statistical measurement errors. The findings revealed that

¶Corresponding author.

FLRWSPCSVM identified ovarian symptoms and a history of cancer symptoms as high-risk indicators for the development of colorectal cancer, with the lowest mean square error (MSE) recorded at 100.605 and a root mean square error (RMSE) of 10.030.

Keywords: Colorectal cancer; fuzzy machine learning approach; high-risk symptom; statistical errors.

1. Introduction

Statistical and mathematical models play a pivotal role in modern disease analysis, offering valuable insights into disease progression, transmission dynamics, and the impact of interventions. These models have been widely applied to various diseases, including colon cancer, tuberculosis, dengue fever, chickenpox, and COVID-19, to guide healthcare strategies and policy decisions. For instance, mathematical modeling has been used to understand interactions between colon cancer and the immune system[16] and to optimize treatment schedules based on circadian rhythms.[6] Machine learning approaches have also been integrated into survival prediction models for colorectal cancer patients.[2] Similarly, in infectious disease research, models have provided significant insights into tuberculosis epidemiology,[5] dengue fever forecasting,[8] chickenpox dynamics,[9] and the propagation of COVID-19 variants.[15,22] These examples highlight the importance of computational and mathematical tools in addressing real-world biomedical and public health challenges.

Regression analysis is regarded as one of the most powerful and extensively utilized statistical methodologies in quantitative research. It enables researchers to comprehend, model and forecast the relationships among variables. Through the implementation of regression techniques, researchers can investigate how the values of one or more independent variables (predictors) affect the value of a dependent variable (outcome). The adaptability and practical utility of regression analysis renders it an essential instrument across a variety of disciplines, including economics, healthcare, social sciences, business, and education. This journal will examine the significance of regression analysis within quantitative research, its various types, applications, and fundamental concepts, as well as its inherent strengths and limitations.[7]

Regression analysis constitutes a statistical methodology utilized to evaluate the relationships between variables. The fundamental principle underlying this approach is that the dependent variable (Y) is influenced by one or more independent variables (X). Typically, the relationship among these variables is expressed through a mathematical equation that illustrates how variations in the independent variables impact the dependent variable. This analytical technique is a vital instrument in quantitative research, furnishing critical insights into the interconnections between variables while also providing predictive capabilities for prospective outcomes. Its applications are extensive, encompassing diverse fields such healthcare. Thereby assisting researchers and policymakers in making well-informed decisions. Although

regression analysis boasts significant advantages, including its capacity to model intricate relationships and generate predictions, researchers must remain vigilant regarding its limitations, particularly in relation to underlying assumptions and the interpretation of causality. In conclusion, regression analysis serves as a foundational element of quantitative research, facilitating a more profound comprehension of the interactions among various factors and their influence on outcomes.[1,21]

Regression analysis has been performed using a variety of methodologies. Data analyzed using popular approaches such as linear regression, fuzzy regression, and ordinary least squares regression construct the regression function. Nonparametric regression approaches include the use of a regression function that is contained inside a certain group of operations, some of which could have limitless dimensions. How effectively the study's results perform depends on the structure of the data generation process and how it connects to the regression approach used. Since the method of collecting the data is normally great or there is clear data, the technique or the results are acceptable if assumptions are made. If enough evidence is provided, these assumptions may sometimes be tested. Even when assumptions are just slightly broken, regression models for prediction are often useful, albeit they may not perform at their best. In many situations, particularly when determining tiny effects or issues with causality based on observational data.[3,20]

FLRWSPCSVM is a method designed to enhance the precision and reliability of models by categorizing similar parameters or data points. The concept behind statistical process control (SPC) is to group the related parameters according to their attributes, which simplifies the regression model and boosts its effectiveness. In the realm of fuzzy linear regression, symmetric parameter clustering acts as a preparatory step where parameters (such as risk factors, demographic details, or characteristics) are organized into symmetric clusters. These clusters are established based on their influence on the dependent variable and possess comparable properties or relationships with the outcome. By grouping these parameters, the model decreases the complexity of the interactions between variables, reduces noise and enhances the overall consistency of the regression model. For instance, in the analysis of medical data, factors, such as age, weight, and family history, might be grouped together since they collectively affect the probability of a disease outcome. Organizing these interconnected parameters allows the model to concentrate on the most pertinent features and avoid overfitting, leading to more dependable and generalized predictions.[23,24]

In the field of healthcare, this combined approach can forecast various outcomes, including the probability of certain illnesses (like colorectal cancer) and the intensity of symptoms. By managing ambiguous data, such as unclear symptom explanations or partial medical records, the model can deliver more precise and detailed predictions. Additionally, this hybrid model is exceptionally effective for analyzing intricate, non-linear connections in data with many dimensions. This scenario is often

seen in sectors like finance, where anticipating stock market movements or consumer habits necessitates understanding the interactions among several factors.[14]

Fuzzy Linear Regression with Symmetric Parameter Clustering with Support Vector Machines constitutes a sophisticated and adaptable methodology for forecasting symptoms associated with colorectal cancer. This hybrid model capitalizes on the advantages of fuzzy logic, machine learning, and advanced clustering techniques to establish a resilient framework capable of addressing uncertainty, modeling intricate relationships, and enhancing predictive accuracy. As healthcare datasets grow in complexity, such integrated approaches present a promising avenue for the advancement of effective diagnostic tools and decision support systems within the field of medical practice. Since, the new hybrid FLRWSPCSVM model can accurately predict colorectal cancer high-risk symptoms, many elements of colorectal cancer may be understood.[25,27] Furthermore, the new hybrid FLRWSPCSVM models are anticipated to have the smallest measurement error model. The newly developed hybrid FLRWSPCSVM models predict high-risk colorectal cancer symptoms more precisely and effectively, with lower model error. The development of an accurate prediction model for the analysis and monitoring of high-risk colorectal cancer symptoms will help to reduce the scope of any new issues that may arise. Some important points include the following:

- The FLRWSPCSVM model employs a methodology that involves a hybrid approach to knowl- edge creation, incorporating health sciences and technology assessment.
- This approach aims to predict high-risk symptoms of colorectal cancer and facilitate an innova- tive process for the early detection and management of different stages of colorectal cancer.
- The hybrid FLRWSPCSVM model aligns with the objectives of the World Health Organization's efforts in addressing colorectal cancer.
- The new hybrid FLRWSPCSVM model plays a crucial role in determining suitable indices for evaluating science and technology within the health research system.

Therefore, it will change the classification of cancer's risk level and raise people's awareness of colorectal cancer. Additionally, it will assist the oncology department in lowering the price of medical supplies and equipment.

2. Materials and Methods

The population for the study was patients diagnosed with colorectal cancer at all stages, and a general hospital in Kuala Lumpur, Malaysia provided secondary data that included real colorectal cancer data.[12] Data were collected and recorded by physicians and nurses using cluster sampling. Statistical Package for Social Sciences, Matlab, Weka Explorer, and Microsoft Excel were used for machine learning.

2.1. *Fuzzy linear regression model*

Every discipline may benefit from statistical analysis, particularly linear regression. In a fuzzy regression approach known as fuzzy linear regression, certain model components are represented by fuzzy numbers. Hideo Tanaka studied a method known as the fuzzy linear regression model (FLR) in 1982. The primary goal of estimating values in the study is acquired as fuzzily defined quantities that reflect the fuzziness of the system's architecture. The traditional secret interval and observation mistakes are associated at the same time. The fuzzy model does not need any assumptions.[4,17]

Due to fuzzy parameters, the input and output data are ambiguous. The fuzzy parameters' expression of the system structure's ambiguity serves as the model's justification for data inconsistencies.[18]

Fuzzy output is denoted as $(\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_g)$, where α_i is the center and $(\zeta_0, \zeta_1, \ldots, \zeta_g)$ is a width of the fuzzy triangular diagram. The linear function of fuzzy linear regression is as follows:

$$Y = A_0(\alpha_0, \zeta_0) + A_1(\alpha_1, \zeta_1)x_1 + \cdots + A_g(\alpha_g, \zeta_g)x_g, \tag{1}$$

where $A = [A_0, A_1, \ldots, A_g]$ is a vector of fuzzy coefficients provided in the form of a triangular fuzzy number and $X = [x_1, x_2, \ldots, x_g]$ is a vector of independent variables. By using the provided data and resolving the linear programming issue, FLR's fitting model may be improved. In addition, a linear programming problem may be used to fine-tune the fuzzy parameter[26]:

$$\begin{aligned} &\alpha^t x_i + (1-H)\sum_j c_j|x_{ij}| \geq y_i + (1-H)e_i \\ &-\alpha^t x_i + (1-H)\sum_j c_j|x_{ij}| \geq -y_i + (1-H)e_i. \end{aligned} \tag{2}$$

2.2. *Fuzzy linear regression with symmetric parameter*

Fuzzy Professional researchers often use fuzzy linear regression with symmetric parameters (FLRWSP) for studying confusing phenomena. With symmetric parameters, fuzzy linear regression may reflect a variety of hazy and ambiguous circumstances. In the study, fuzzy linear regression was used by the researcher to assess food quality, particularly that of fried donuts. Model theory is helpful from a scientific and engineering perspective for the conceptual framework and findings that can be immediately implemented in system models utilizing the fuzzy approach and new developments in fuzzy logic.[4] If $\tilde{A}_i (i = 0, 1, \ldots, n)$ is a symmetrical fuzzy number and x_i is a crisp real number, and will be a triangular fuzzy number, $\breve{Y} = (f^c(x), f^s(x))$ when $f^c(x)$ is the mode and $f^s(x)$ is spread of triangular fuzzy number.

The model of FLRWSP can be written as follows:

$$f^c = a_0 + a_1 x_1 + \cdots + a_n x_n, \tag{3}$$

where

$f^c(x)$ is equation of fuzzy parameter,
$a_0, a_1, \ldots, a_n$ are fuzzy parameters,
$x_1, x_2, \ldots, x_n$ are variable of fuzzy parameter

and membership function Y can be defined as follows:

$$\begin{aligned} Y &= 1 - \frac{-(f^c(x) - y)}{f^s(x)} f^c(x) - f^s(x) \leq y \leq f^c(x), \\ Y &= 1 - \frac{(y - f^c(x))}{f^s(x)} f^c(x) \leq y < f^c(x) + f^s(x). \end{aligned} \tag{4}$$

The target function is defined in the symmetric condition of triangular fuzzy number as

$$\begin{aligned} (1-h)s_0^L + (1-h)\sum_{i=1}^{n}(s_i^L|x_{ji}|) - a_0 - \sum_{i=1}^{n}(a_i x_{ji}) \geq -y_j, \\ (1-h)s_0 + (1-h)\sum_{i=1}^{n}(s_i^L|x_{ji}|) + a_0 + \sum_{i=1}^{n}(a_i x_{ji}) \geq -y_j, \end{aligned} \tag{5}$$

where

s_0^L, s_i^L are spread left of triangular fuzzy number,
s_0, s_i are spread of triangular fuzzy number,
α_0, α_i are modes of triangular fuzzy number,
h is degree of triangular fuzzy number.

2.3. *Fuzzy clustering methods*

A data set may be a member of more than one cluster when using the fuzzy C-means (FCM) clustering technique.[13,19] J.C. Dunn created fuzzy C-means (FCM) clustering in 1973, and J.C. Bezdek enhanced it in 1981. As a result of the algorithm's foundation in fuzzy C-means minimization in the direction of the following objective function or criteria, such as

$$j_m = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m \leq \infty, \tag{6}$$

where m is any real number higher than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith d-dimensional center of the cluster, and $\|*\|$ is any norm indicating the similarity between any measured data and the center. With the updating of membership u_{ij} and cluster centers c_j as described below, fuzzy

partitioning is performed through an iterative optimization of the objective function indicated above:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|X_i - C_j\|}{x_i - C_k} \right)^{\frac{2}{m-1}}}, \quad C_j = \frac{\sum_{i=1}^{N} u_{ij}^m X_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{7}$$

This iteration will end when $\max_{ij} \left\{ |u_{ij}^{k+1} - u_{ij}^{(k)}| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of Jm. In case of the algorithm for minimizing j_m, it can be summarized by the following steps:

(1) Initialize $U = |u_{ij}|$ and matrix, U.
(2) k steps: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m X_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{8}$$

(3) Update U_k, U^{k+1}

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|X_i - C_j\|}{x_i - C_k} \right)^{\frac{2}{m-1}}}. \tag{9}$$

(4) If $\|U^{k+1} - U_k\| < \varepsilon$ afterwards, stop; otherwise, go back to step 2.

2.4. *Support vector machines*

The Support Vector Machines (SVMs) algorithm proposed by Vapnik in 1963 is used to identify subtle patterns in complex data sets. The algorithm performs linear classification and regression analysis to predict the sorting or regression of previously unseen data.[10,11]

Linear SVM machines construct a hyperplane as a decision surface so that the separation margin between positive and negative gives good generalization performance.[25] In support vector learning, three learning machines can be constructed: Polynomial Kernel, Radial Basis Functions Kernal, and Two-Layer Perceptron. Example of kernel functions with the polynomial kernel:

$$K(x, y) = (x^T y + 1)^d. \tag{10}$$

Moreover, when performing linear classification, SVMs can efficiently perform nonlinear sorting using the kernel trick by implicitly mapping their inputs to high-dimensional feature spaces.

3. Multiple Linear Regression Clustering and Support Vector Machines

The hybrid model is defined as a combination of two or more individual models. The proposed approach introduces a hybrid model, termed the Multiple Linear Regression Clustering with Support Vector Machine (MLRCSVM). This model comprises five distinct procedures:

(1) Determine which correlation between Y and X_i has a higher value.
(2) The first stage of the hybrid is the modeling of the multiple linear regression (MLR) clustering, which is a combination of MLR and fuzzy C-means (FCM). The clustering combination between Multiple Linear Regression (MLR) and Fuzzy C-Means (FCM) is based solely on the Y data, as well as on the Y data with independent variables that exhibit higher correlation values. The optimal MLR clustering model is selected based on the smallest values of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).
(3) Identify the MLR clustering and SVM residuals.
(4) The second hybrid stage is making the new hybrid data using the equation in (6).

$$Y = L_t + N_t, \tag{11}$$

where Y represents a new set of data, L_t denotes the residuals from the MLR cluster (linear model) and N_t refers to the residuals from the Support Vector Machine (SVM) model (linear model). The SVM, being a linear model, is capable of minimizing model errors and is less sensitive to outliers, making it a suitable choice for integration into the hybrid model.
(5) Developing a hybrid model using the FLRWSP method

$$\text{ERROR}_{\text{final}} = \frac{(n_1 \times \text{ERROR}_1) + (n_2 \times \text{ERROR}_2)}{n_1 + n_2}, \tag{12}$$

where the numbers for clusters 1 and 2 are, respectively, n_1 and n_2. ERROR_1 is the number error of MLR clustering and ERROR_2 is the number error of the SVM model. Error-values would be the value of MSE and RMSE.

3.1. *A new hybrid fuzzy linear regression with symmetric parameter clustering and support vector machines*

Fuzzy linear clustering regression combining fuzzy linear regression and fuzzy C-means method is proposed. The hybrid combines the fuzzy linear clustering regression model and the support vector machines model. The creation of the hybrid is carried out in five steps as follows and is shown in Fig. 1.

(1) Determine which correlation between Y and X_i has a greater value.

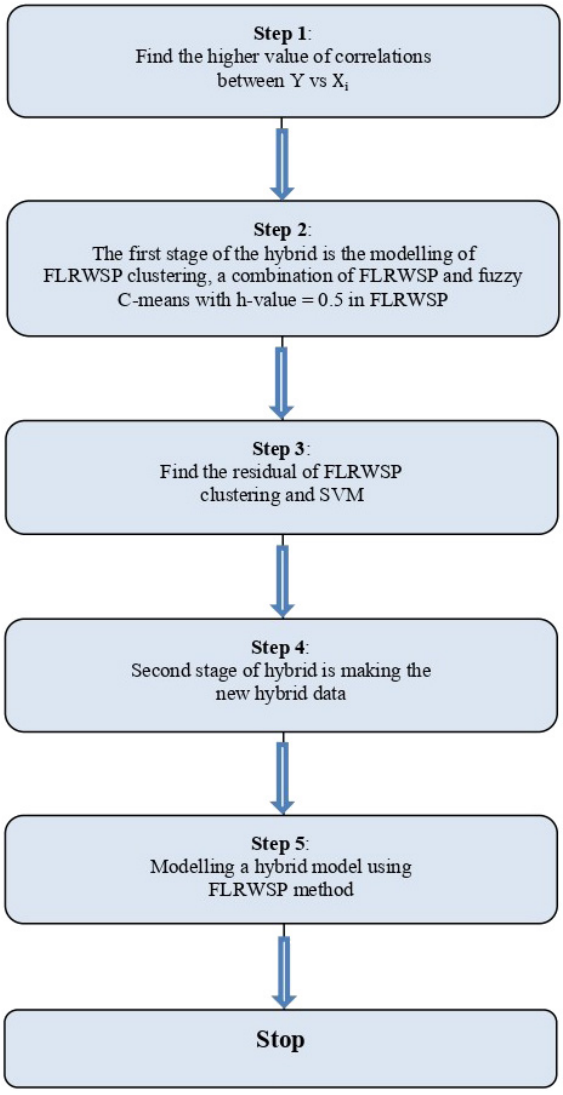

Fig. 1. Framework of hybrid model.

(2) The first stage of the hybrid is the modeling of the FLRWSP cluster, which is a combination of FLRWSP and fuzzy C-means. The combination of clustering between FLRWSP and FCM is based on the Y data alone and the Y data toward several higher values of the independent variables that have higher correlation values. Based on the MSE and RMSE values with the least values, the optimal FLRWSP clustering is chosen. The FLRWSP clustering is optimized by adjusting the h value between 0 and 1 (in 0.1 steps) with the lowest value of statistical error.

(3) Identify the FLRWSP clustering and SVM residuals.
(4) The second hybrid stage is making the new hybrid data using the equation in (6).

$$Y = L_t + N_t. \tag{13}$$

Here, Y is a new set of data, L_t is the FLRWSP cluster residual (nonlinear model), and N_t is the SVM model residual (linear model). SVM is a linear model that can minimize model error and is not too sensitive to outliers, hence it is selected for a hybrid model.
(5) Developing a hybrid model using the FLRWSP method

$$\text{ERROR}_{\text{final}} = \frac{(n_1 \times \text{ERROR}_1) + (n_2 \times \text{ERROR}_2)}{n_1 + n_2}, \tag{14}$$

where the numbers for clusters 1 and 2 are, respectively, n_1 and n_2. ERROR_1 is the number error of MLR clustering and ERROR_2 is the number error of the SVM model. Error-values would be the value of MSE and RMSE.

3.2. *Performance measurement of error*

Rotation estimation is another name for error-based performance assessment or cross-validation. This method is used to assess the outcomes of a statistical study on a separate data set. It is mostly used when predicting outcomes is the aim and when determining how accurate a predictive model will be in actual usage. Standard approaches include MSE and RMSE.[24]

(1) The mean squared error (MSE) calculates the average of the squares of mistakes that correspond to the squared error loss anticipated value. The MSE equation is similar to Eq. (15)

$$\text{MSE} = \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{N}. \tag{15}$$

(2) The Root Mean Square Error (RMSE), which measures how distant the residuals (prediction errors) are from the regression line's data points, is the standard deviation of the residuals. The RMSE equation is the same as Eq. (16).

$$\text{MSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{N}}, \tag{16}$$

where y represents the real data, $\tilde{y}$ represents the predicted data, and N represents the total number of observations.

4. Results

In this study, 180 patients' secondary data were the responses, and there were 25 variables: gender (A1), ethnicity (A2), age (A3), icd10 (A4), TNM Staging (A5),

family history (A6), diabetes mellitus (A7), Crohn's disease (A8), ulcerative colitis (A9), polyp (A10), history of cancer (A11), endometrial (A12), gastric (A13), small bowel (A14), hepatobiliary (A15), urinary tract (A16), ovarian (A17), other cancer (A18), intestinal obstruction (A19), colorectal (A20), weight loss (A21), diarrhea (A22), blood stool (A23), anemia (A24), abdominal (A25). In contrast, the dependent variable for colorectal cancer is the tumor size. Utilizing MSE and RMSE statistical cross-validation methods, it was possible to compare these models. The most accurate model for forecasting colorectal cancer tumor size is the one with the lowest MSE and RMSE. The most accurate model for predicting high-risk symptoms of colorectal cancer is the one with the lowest MSE and RMSE.

4.1. *Multiple linear regression clustering with support vector machines*

In this hybrid model, multiple linear regression clustering (a combination of multiple linear regression and fuzzy c-means clustering) is integrated with a SVM model. This hybrid approach leverages both linear and linear modeling capabilities to predict high-risk symptoms of colorectal cancer. The multiple linear regression clustering and SVM models complement each other, each excelling at capturing distinct data characteristics, allowing for effective modeling of both linear and nonlinear patterns in the data.

4.2. *Cluster* 1 *for MLRCSVM model*

The hybrid model, combining multiple linear regression clustering with the SVM, analyzed a dataset of 85 colorectal cancer patients or respondents, identified as Cluster 1. This model was employed to examine 25 predictor variables and predict high-risk symptoms of colorectal cancer. Upon analysis, 16 predictor variables were found to be statistically significant for colorectal cancer. These significant variables include age, ethnic group, ICD-10 site, TNM staging, family history, Crohn's disease, ulcerative colitis, history of cancer, small bowel, urinary tract, gastric, ovarian, other cancers, intestinal obstruction, weight loss, diarrhea, anemia, blood stool and abdominal. The significance of these variables was determined based on p-values less than 0.05, as detailed in Table 1.

Table 2 presents a comparative evaluation of the model performance for the MLRCSVM (cluster 1) model. The recorded values for the MLRCSVM (cluster 1) model were 64.453 for MSE and 8.028 for RMSE. All identified significant variables have a notable impact on the progression of colorectal cancer. The estimated model for the symptoms and factors associated with colorectal cancer, derived from the hybrid multiple linear regression clustering combined with the SVM (Cluster 1).

$\breve{Y} = 64.228 + -0.573\text{age} + 4.321$ ethnic group $+ 5.836$ icd10 $+ 2.024$ TNM Staging $+ -9.670$ family history $+3.130$ Crohn's disease $+3.127$ ulcerative colitis $+4.924$ history of cancer $+ -6.784$ small bowel $+3.414$ urinary tract $+13.736$ ovarian

Table 1. The parameter of the MLRCSVM model (cluster 1).

Independent variables	Beta (β)	Sig. value
(Constant)	64.228	∗0.000
Gender (A1)	0.515	0.722
Age at Diagnosis (years) (A2)	−0.573	∗0.000
Ethnic Group (A3)	4.321	∗0.000
ICD 10 Site (A4)	5.836	∗0.000
TNM Staging (A5)	2.024	∗0.000
Family History (A6)	−9.670	∗0.000
Diabetes Mellitus (A7)	−0.577	0.732
Crohn's Disease (A8)	3.130	∗0.047
Ulcerative colitis (A9)	3.127	∗0.037
Polyp (A10)	−2.535	0.093
History of cancer (A11)	4.924	∗0.002
Endometrial (A12)	−0.372	0.805
Gastric (A13)	−1.966	0.218
Small bowel (A14)	−6.784	∗0.000
Hepatobiliary (A15)	−1.081	0.476
Urinary tract (A16)	3.414	∗0.033
Ovarian (A17)	13.736	∗0.000
Other cancer (A18)	4.193	∗0.009
Intestinal Obstruction (A19)	−6.583	∗0.000
Colorectal (A20)	2.785	0.061
Weight loss (A21)	4.677	∗0.002
Diarrhoe (A22)	5.555	∗0.001
Anemia (A23)	−4.315	∗0.036
Blood stool (A24)	−1.765	0.268
Abdominal (A25)	−2.69	0.341

Note: *significant at 0.5.

Table 2. Summary of the MLRCSVM model (cluster 1).

Methods	Value
MSE	64.453
RMSE	8.028

+4.193 other cancer + − 6.583 intestinal obstruction +4.677 weight loss +5.555 diarrhoe + − 4.315 anemia.

4.3. *Cluster 2 for MLRCSVM model*

The hybrid model, combining multiple linear regression clustering and SVM techniques, was applied to a dataset of 95 colorectal cancer patients or respondents (Cluster 2). This model was used to analyze 25 predictor variables and predict high-risk symptoms of colorectal cancer. Following the analysis, only one predictor

Table 3. The parameter of the MLRCSVM model (cluster 2).

Independent variables	Beta (β)	Sig. value
(Constant)	85.701	∗0.001
Gender (A1)	2.639	0.651
Age at Diagnosis (years) (A2)	−0.266	0.278
Ethnic Group (A3)	3.820	0.445
ICD 10 Site (A4)	−2.811	0.612
TNM Staging (A5)	−0.291	0.857
Family History (A6)	−27.602	∗0.006
Diabetes Mellitus (A7)	−3.357	0.605
Crohn's Disease (A8)	8.956	0.176
Ulcerative colitis (A9)	−1.361	0.830
Polyp (A10)	0.051	0.994
History of cancer (A11)	12.222	0.082
Endometrial (A12)	−3.208	0.619
Gastric (A13)	−13.351	0.054
Small bowel (A14)	2.221	0.720
Hepatobiliary (A15)	2.256	0.741
Urinary tract (A16)	0.218	0.974
Ovarian (A17)	13.288	0.058
Other cancer (A18)	5.744	0.331
Intestinal Obstruction (A19)	2.377	0.725
Colorectal (A20)	−2.419	0.700
Weight loss (A21)	3.677	0.545
Diarrhoe (A22)	−4.404	0.446
Anemia (A23)	−6.876	0.376
Blood stool (A24)	−4.867	0.399
Abdominal (A25)	−3.782	0.608

Note: *significant at 0.5.

Table 4. Summary of the MLRCSVM model (cluster 2).

Methods	Value
MSE	150.090
RMSE	12.251

variables were found to be statistically significant: family history. The significance of these variables was determined based on p-values less than 0.05, as shown in Table 3.

Table 4 presents a comparative evaluation of the model performance for the MLRCSVM (cluster 2) model. The recorded values for the MLRCSVM (cluster 2) model were 150.090 for MSE and 12.251 for RMSE. All significant variables exert a considerable influence on the effects of colorectal cancer. The estimated model for the symptoms and factors associated with colorectal cancer, derived from the hybrid multiple linear regression clustering and SVM approach (Cluster 2).

$$\breve{Y} = 85.701 + -27.602 \text{ family history}.$$

Table 5. Measurement error for cluster 1.

Methods	Value
MSE	101.232
RMSE	10.061

4.4. *A new hybrid FLRWSPCSVM*

A new hybrid FLR model was created utilizing SVM clustering and two types of measurement error: RMSE and MSE, each of which have a fuzzy parameter. Based on Table 5, MSE and RMSE error measurements can be used to analyzed and to obtain the values which were 101.232 and 10.061. The most accurate colorectal cancer tumor size prediction model has the lowest error value.

4.5. *Cluster* 1 *for a hybrid model*

The 85 data were utilized as responders in Cluster 1 of the hybrid FLR using clustering and SVM model. The parameters of model cluster 1 and the residual in Fig. 2 are as follows.

The fuzzy mean value of tumor size (mm) can be explained by ovarian with the highest fuzzy parameter = 18.864 as in Table 6.

$\breve{Y} = 1.022 + (4.179, 0)\text{A1} + (-0.319, 0.511)\ \text{A2} + (9.591, 0)\ \text{A3} + (10.551, 0)\ \text{A4} + (3.997, 0)\ \text{A5} + (-2.273, 0)\ \text{A6} + (1.795, 0)\ \text{A7} + (2.556, 0)\ \text{A8} + (-1.574, 0)\ \text{A9} + (-4.216, 0)\ \text{A10} + (8.979, 0)\ \text{A11} + (-1.438, 0)\ \text{A12} + (0.471, 0)\ \text{A13} + (-3.065, 0)\ \text{A14} + (-2.274, 0)\ \text{A15} + (11.171, 0)\ A16 + (18.864, 0)\ \text{A17} + (7.864, 0)\ \text{A18} + (-7.089, 0)\ \text{A19} + (2.574, 0)\ \text{A20} + (3.205, 0)\ \text{A21} + (10.464, 0)\ \text{A22} + (1.167, 0)\ \text{A23} + (-4.212, 0)\ \text{A24} + (-1.222, 0)\ \text{A25}.$

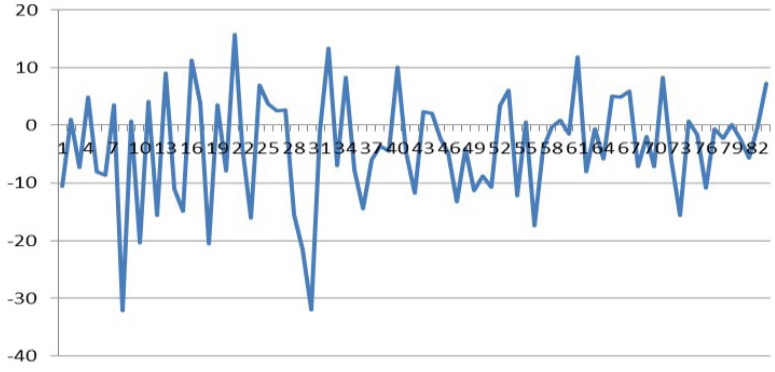

Fig. 2. The FLRWSPCSVM model's residual (cluster 1).

Table 6. The parameter of the FLRWSPCSVM model (cluster 1).

Fuzzy parameter center	α_i	Width c_i
(Constant)	1.022	0
Gender (A1)	4.179	0
Age at Diagnosis (years) (A2)	−0.319	0.511
Ethnic Group (A3)	9.591	0
ICD 10 Site (A4)	10.551	0
TNM Staging (A5)	3.997	0
$\text{Family}_{\text{History}}$ (A6)	−2.273	0
Diabetes Mellitus(A7)	1.795	0
Crohn's Disease (A8)	2.556	0
Ulcerative colitis (A9)	−1.574	0
Polyp (A10)	−4.216	0
History of cancer (A11)	8.979	0
Endometrial (A12)	−1.438	0
Gastric (A13)	0.471	0
Small bowel (A14)	−3.065	0
Hepatobiliary (A15)	−2.274	0
Urinary tract (A16)	11.171	0
Ovarian (A17)	18.864	0
Other cancer (A18)	7.864	0
Intestinal Obstruction (A19)	−7.089	0
Colorectal (A20)	2.574	0
$\text{Weight}_{\text{loss}}$ (A21)	3.205	0
Diarrhoe (A22)	10.464	0
$\text{Blood}_{\text{stool}}$ (A23)	1.167	0
Anemia (A24)	−4.212	0
Abdominal (A25)	−1.222	0

Table 7. Measurement error for cluster 2.

Methods	Value
MSE	100.069
RMSE	10.003

4.6. *Cluster* 2 *for a hybrid model*

The 95 data were utilized as responders in Cluster 2 of the hybrid FLR using clustering and SVM model. According to Table 7, the MSE value is 100.069 and the RMSE value is 10.003. The parameters of model cluster 2 and residual in Fig. 3 are as follows.

The fuzzy mean value of tumor size (mm) can be explained by the history of cancer, with the highest fuzzy parameter = 11.853 as in Table 8.

$\breve{Y} = 1.669 + (3.416, 0)\ \text{A1} + (0.017, 0.834)\ \text{A2} + (4.713, 0)\ \text{A3} + (5.522, 0)\ \text{A4} + (3.158, 0)\ \text{A5} + (-13.984, 0)\ \text{A6} + (7.230, 0)\ \text{A7} + (11.073, 0)\ \text{A8} + (4.094, 0)\ \text{A9} +$

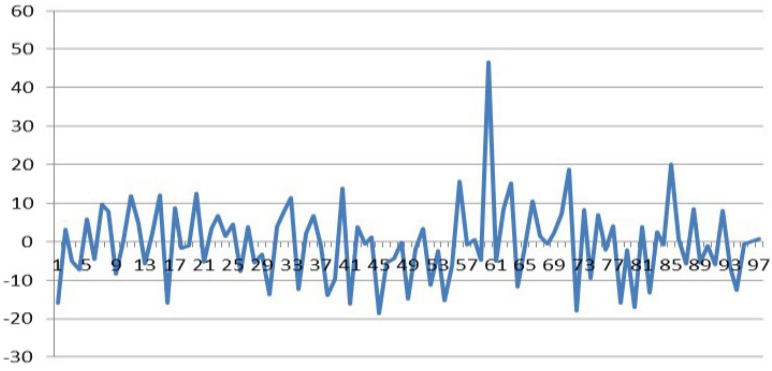

Fig. 3. The FLRWSPCSVM model's residual (cluster 2).

Table 8. The FLRWSPCSVM model's parameter (cluster 2).

Fuzzy Parameter	Center α_i	Width c_i
(Constant)	1.669	0
Gender (A1)	3.416	0
Age at Diagnosis (years) (A2)	0.017	0.834
Ethnic Group (A3)	4.713	0
ICD 10 Site (A4)	5.522	0
TNM Staging (A5)	3.158	0
$\text{Family}_{\text{History}}$ (A6)	−13.984	0
Diabetes Mellitus(A7)	7.230	0
Crohn's Disease (A8)	11.073	0
Ulcerative colitis (A9)	4.094	0
Polyp (A10)	2.792	0
History of cancer (A11)	11.853	0
Endometrial (A12)	4.824	0
Gastric (A13)	−3.694	0
Small bowel (A14)	−3.433	0
Hepatobiliary (A15)	10.677	0
Urinary tract (A16)	5.272	0
Ovarian (A17)	4.936	0
Other cancer (A18)	0.712	0
Intestinal Obstruction (A19)	1.675	0
Colorectal (A20)	−1.232	0
$\text{Weight}_{\text{loss}}$ (A21)	4.492	0
Diarrhoe (A22)	3.808	0
$\text{Blood}_{\text{stool (A23)}}$	−10.831	0
Anemia (A24)	−1.859	0
Abdominal (A25)	−4.707	0

$(2.792, 0)$ A10 + $(11.853, 0)$ A11 + $(4.824, 0)$ A12 + $(-3.694, 0)$ A13 + $(-3.433, 0)$ A14 + $(10.677, 0)$ A15 + $(5.272, 0)$ A16 + $(4.936, 0)$ A17 + $(0.712, 0)$ A18 + $(1.675, 0)$ A19 + $(-1.232, 0)$ A20 + $(4.492, 0)$ A21 + $(3.808, 0)$ A22 + $(-10.831, 0)$ A23 + $(-1.859, 0)$ A24 + $(-4.707, 0)$ A25.

Table 9. The final measurement error.

Methods	Value
MSE	100.605
RMSE	10.030

Table 10. Result for statistical error measurement.

Model of linear regression	MSE	RMSE
MLRCSVM	109.561	10.471
A new hybrid FLRWSPCSVM model	100.605	10.030

The final MSE and RMSE values of the model are shown in (Table 9). After analysis, the residual FLRWSP cluster model and the residual SVM model were hybridized (Table 9), which shows the MSE and RMSE values for this new hybrid model.

5. Comparison of the Clustering Models

A comparative analysis was conducted among several regression and machine learning models, including Multiple Linear Regression with Clustering and Support Vector Machine (MLRCSVM) and Fuzzy Linear Regression with Symmetric Parameter Clustering and Support Vector Machine (FLRWSPCSVM). The results of the comparison, based on MSE and RMSE metrics, are presented in Table 10. It presents a comparative evaluation of the model performance for the MLRCSVM and FLRWSPCSVM models. The recorded values for the MLRCSVM model were 109.561 for MSE and 10.471 for RMSE. For the FLRWSPCSVM model, the statistical error metrics were 100.605 for MSE and 10.030 for RMSE. These values were the lowest rather than MLRCSVM model.

The results demonstrate that the FLRWSPCSVM model yields the lowest values for MSE and RMSE indicating superior predictive accuracy compared to the other model. These findings highlight the enhanced precision of the FLRWSPCSVM model and suggest its promising potential for application across various fields. The proposed Fuzzy Linear Regression with Symmetric Parameter Clustering and Support Vector Machine (FLRWSPCSVM) model outperforms the other models, exhibiting the lowest values for Mean Square Error (MSE) and RMSE. A total of 25 determinants were utilized to predict the high-risk symptoms of colorectal cancer patients at a general hospital in Kuala Lumpur. The analysis incorporated two error metrics MSE and RMSE to compare the performance of Multiple Linear Regression with Clustering and SVM (MLRCSVM) and FLRWSPCSVM. The results indicate that the FLRWSPCSVM model consistently provides the most accurate predictions, with the smallest error values and highlights the significance of 25 determinants in predicting high-risk symptoms of colorectal cancer.

Table 11. MSE and RMSE values in each model.

Model of linear regression	MSE	RMSE
MLRCSVM	109.561	10.471
A new hybrid FLRWSPCSVM model	100.605	10.030

6. Conclusion

The hybrid technique was hybrid with the novel fuzzy machine learning approach, FLRWSP, which combines clustering and the SVM model. To address an ill-defined phenomenon, a new model was developed. Based on two measurement error models, the most informative model applicable to various study domains was found to be a combination of fuzzy linear regression (FLR) clustering and SVM. The model's prediction accuracy, characterized by minimized error values was assessed using metrics such as MSE and RMSE. An overview of the models is presented in Table 11.

The FLRWSPCSVM model is demonstrated to be the most effective for predicting high-risk symptoms of colorectal cancer, as it yields the lowest MSE and RMSE values compared to other models. Furthermore, ovarian cancer and a history of cancer are identified as two high-risk symptoms significantly influencing the development of colorectal cancer. Given its superior predictive performance, the FLRWSPCSVM model is recommended for use in general hospitals to address and manage high-risk symptoms associated with colorectal cancer.

This study provides high-risk symptoms prediction recommendations for future research in Malaysian general hospitals, especially in Kuala Lumpur and globally in Oncology Departments. The FLRWSPCSVM model developed for predicting high-risk symptoms of colorectal cancer based on patient symptoms can help reduce mortality and improve Oncology services. Future research should focus on diverse populations, larger sample sizes, and the use of advanced methods like multiple linear regression and fuzzy linear regression. Investigating additional symptoms and integrating the FLRWSPCSVM model with neural networks could enhance prediction accuracy. Addressing colorectal cancer staging and expanding this research to other hospitals in Malaysia and globally can significantly improve colorectal cancer analysis and patient outcomes.

Acknowledgment

This research was supported by the Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2024/STG06/UTHM/02/2).

Appendix A

CRC	Colorectal cancer
SVM	Support vector machine
FCM	Fuzzy c-method
FLR	Fuzzy linear regression model
MLRCSVM	Multiple linear regression clustering with support vector machine
FLRWSP	Fuzzy linear regression with symmetric parameter
FLRWSPCSVM	Fuzzy linear regression with symmetric parameter clustering with support vector machine
MSE	Mean square error
RMSE	Root mean square error

ORCID

M. A. Shafi https://orcid.org/0000-0002-5299-1374
Sayooj Aby Jose https://orcid.org/0000-0003-4437-1623
M. S. Rusiman https://orcid.org/0000-0001-8255-9884
Anuwat Jirawattanapanit https://orcid.org/0000-0002-6319-0214

References

1. A. Agresti, *An Introduction to Categorical Data Analysis* (John Wiley & Sons, Inc, New York, 1996).
2. L. Buk Cardoso *et al.*, Machine learning for predicting survival of colorectal cancer patients, *Science Reports* **13** (2023) 8874.
3. Y. Dasril, G. K. Wen, B. Nazarudin and N. S. Salahudin, New approach on global optimization problems based on meta-heuristic algorithm and quasi-Newton method, *International Journal of Electrical and Computer Engineering* **12**(5) (2022) 5182–5190.
4. D. I. Amir Hamzah, M. Saifullah Rusiman, N. Che Him, M. Ammar Shafi, O. Gurunlu Alma and S. Suhartono, A time series analysis for sales of chicken-based food product, *AIP Conference Proceedings* **2355** (2021) 060002.
5. K. B. Hamna Mariyam, S. A. Jose, A. Jirawattanapanit and K. Mathew, A comprehensive study on tuberculosis prediction models: Integrating machine learning into epidemiological analysis, *Journal of Theoretical Biology* **597** (2025) 1–11.
6. J. Hesse, T. Müller and A. Relógio, An integrative mathematical model for timing treatment toxicity and Zeitgeber impact in colorectal cancer cells, *NPJ Systems Biology and Applications* **9**(27) (2023) 1–9.
7. L. John and S. W. Teng, Using information management systems and processes to support shared care for colorectal cancer survivors, *IEEE International Symposium on Technology in Society Proceedings* **17**(3) (2017) 1–8.
8. S. A. Jose, K. Mathew, K. B. Hamna Mariyam, A. Jirawattanapanit and W. Anurak, Analyzing and forecasting dengue fever incidence in Thailand: A comprehensive study for public health preparedness, *International Journal of Biomathematics* **17**(3) (2024) 1–15.

9. S. A. Jose, Z. Yaagoub, D. Joseph, R. Ramachandran and A. Jirawattanapanit, Computational dynamics of a fractional order model of chickenpox spread in Phuket province, *Biomedical Signal Processing and Control* **91** (2024) 1–12.
10. S. Kusumadewi, L. Rosita and E. G. Wahyuni, Fuzzy linear regression based on a hybrid of fuzzy C-means and the fuzzy inference system for predicting serum iron levels in patients with chronic kidney disease, *Expert Systems with Applications* **227** (2023) 120–314.
11. M. H. Mashinchi, M. A. Orgun and M. Mashinchi, Solving fuzzy linear regression with hybrid optimization, in *Neural Information Processing*, Lecture Notes in Computer Science, Vol. 5864 (Springer, 2009), pp. 336–343.
12. Ministry of Health Malaysia, *National Strategic Plan for Colorectal Cancer* (*NSPCRC*) *2021–2025* (Ministry of Health Malaysia, Putrajaya, Malaysia, 2021), pp. 1–68.
13. Y. Ni, Fuzzy correlation and regression analysis, Ph.D. thesis, University of Oklahoma Graduate College (2005).
14. J. Nowakov'a and M. Pokorn'y, Fuzzy linear regression analysis, *IFAC Proceedings Volumes* **6** (2013) 245–249.
15. B. I. Omede, S. A. Jose, A. Jirawattanapanit and T. Park, Impact of imperfect vaccination and re-infection on the dynamics of delta and omicron COVID-19 variants in the USA, *Modeling Earth Systems and Environment* **10** (2024) 1–20.
16. E. Raeisi, M. Yavuz and M. Khosravifarsani, Mathematical modeling of interactions between colon cancer and immune system with a deep learning algorithm, *European Physical Journal Plus* **139**(345) (2024) 1–15.
17. H. Tanaka, S. Uejima and K. Asai, Linear regression analysis with the fuzzy model, *IEEE Transactions on Systems*, *Man and Cybernetics* **12** (1982) 903–907.
18. H. Tanaka, Fuzzy data analysis by possibilities linear models, *Fuzzy Sets and Systems* **24** (1987) 363–375.
19. S. Theodoridis and K. Koutroumbas, Clustering algorithms III: Schemes based on function optimization, in *Pattern Recognition*, 4th edn. (Academic Press, 2009), pp. 701–763.
20. S. N. Salahudin, H. S. Ramli, M. H. A. Razak, M. S. Abdullah and A. Masum, Determinants of career success: A case study of male teachers in secondary schools, *Economic and Environmental Studies* **39**(4) (2021).
21. S. N. Salahudin, H. S. Ramli, M. N. R. Alwi, M. S. Abdullah and N. A. Rani, Employee engagement and turnover intention among Islamic bankers in Brunei Darussalam, *International Journal of Recent Technology and Engineering* **8** (2019) 643–651.
22. E. Shamil, S. A. Jose, H. S. Panigoro, A. Jirawattanapanit, B. I. Omede and Z. Yaagoub, Understanding COVID-19 propagation: A comprehensive mathematical model with caputo fractional derivatives for thailand, *Frontiers in Applied Mathematics and Statistics* **10** (2024) 1–18.
23. L. Shoapei, Fuzzy machine learning methods, in *Fuzzy-AI Model and Big Data Exploration* (Springer, 2022), pp. 117–172.
24. M. A. Shafi, M. S. Rusiman, K. Jacob and A. N. Musa, A new intelligent modelling two-stage hybrid fuzzy prediction approach using computation software, *Journal of Intelligent and Fuzzy Systems* **45** (2023) 11013–11019.
25. M. A. Shafi, M. S. Rusiman, S. Ismail and M. G. Kamardan, 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**(4) (2019) 323–328.

26. L. A. Zadeh, Fuzzy sets, *Information and Control* **8** (1965) 338–358.
27. A. S. Zakaria, M. A. Shafi, M. A. M. Zim and S. N. A. M. Razali, The use of fuzzy linear regression modeling to predict high-risk symptoms of lung cancer in Malaysia, *International Journal of Advanced Computer Science and Applications* **14** (2023) 586–593.