RESEARCH ARTICLE | AUGUST 27 2024

The prediction of high-risk symptom for colorectal cancer using a new hybrid of fuzzy statistical machine learning approach

Muhammad Ammar Shafi ✉; Mohd Saifullah Rusiman; Siti Afiqah Muhamad Jamil; Mohd Arif Mohd Zim

AIP Conf. Proc. 3123, 020017 (2024)
https://doi.org/10.1063/5.0225096

The Prediction of high-risk symptom for Colorectal Cancer using a New Hybrid of Fuzzy Statistical Machine Learning Approach

Muhammad Ammar Shafi[1, a)], Mohd Saifullah Rusiman[2, b)], Siti Afiqah Muhamad Jamil[3, c)] and Mohd Arif Mohd Zim[4, d)]

[1]*Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia*
[2]*Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Panchor, Johor, Malaysia*
[3]*College of Computing, Informatics and Media, School of Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*
[4]*KPJ Healthcare Berhad, Damansara Specialist Hospital 2, Damansara, Kuala Lumpur, Malaysia*

[a)] Corresponding author: ammar@uthm.edu.my
[b)]saifulah@uthm.edu.my
[c)]afiqahjamil@uitm.edu.my
[d)]arifmz@kpjdamansara2.com

Abstract. Colorectal cancer (CRC) is a type of cancer that develops in the human colon and rectum. The body's cells proliferating out of control, which is the cause of colorectal cancer, results in these symptoms. Nevertheless, there is still disagreement on the precise signs of a high-risk CRC. The linear regression model struggles with erroneous and ambiguous data. Because the idea of fuzzy set theory can deal with data that does not refer to a precise point value, fuzzy machine learning, a new hybrid linear fuzzy regression with symmetric parameter clustering with a support vector machine model (FLRWSPCSVM), is used in this study to predict the high-risk symptoms causing the development of colorectal cancer in Malaysia (uncertainty data). After analysing secondary data from 180 colorectal cancer patients who underwent treatment in a general hospital, 25 separate symptoms with diverse combinations of variable types were considered in the analysis. Together with the model's parameters, errors, and justifications, two statistical measurement errors were also included. The least values of mean square error (MSE) are 100.605 and root mean square error (RMSE) is 10.030 for FLRWSPCSVM, which were determined to be ovarian and a history of cancer symptoms to be the high-risk symptom for developing colorectal cancer. To monitor and control the high-risk symptoms that can affect colon cancer and lower patient mortality, the hospitality industry could also benefit from this study.

INTRODUCTION

A modelling and analysing several variables as part of a study and establishing a relationship between a dependant variable and one or more independent variables, regression analysis is used as a quantitative research method [1]. One of the common models for data analysis is regression analysis. Its appeal stems from a variety of factors. The statistical equation that describes the relationship between the dependant and independent variables is derived using the analysis. Due in part to its multivariate nature, it has a strong explanatory power. Regression analysis is widely utilised in a variety of domains, including forecasting and prediction, applied sciences, economics, engineering, and computer science [2].

Regression analysis's use in forecasting and prediction is mostly overlapping with machine learning. Regression analysis is also used to examine the relationships between the independent and dependent variables and to comprehend how they connect to one another. Regression analysis can therefore be used to determine if the independent and

The 6th ISM International Statistical Conference (ISM-VI) 2023
AIP Conf. Proc. 3123, 020017-1–020017-8; https://doi.org/10.1063/5.0225096
Published under an exclusive license by AIP Publishing. 978-0-7354-5030-1/$30.00

dependent variables are related causally. Care must be given while choosing the data to be used because this could result in delusions or fictitious associations.

Regression models for prediction can be useful even when assumptions are marginally false, especially when there are minor effects or questions about causality. Outliers are data points that dramatically deviates from other observations, which could indicate experimental error. Regression models are unable to manage real-world situations or data due to the level of uncertainty caused by environmental, measurement, or human factors. Lotfi A. Zadeh developed a model to address the van Gunes phenomenon, such as the fuzzy model.

Hideo Tanaka invented FLR in 1982 to measure the prediction of unclear phenomena, taking into account input and output relations. Fuzzy logic offers a mathematical foundation for dealing with ambiguity without presumptions, allowing for more accurate data analysis, even if the inaccuracy is not evenly distributed.

Many studies have been conducted on colon cancer, and various models and approaches have been utilised to address issues. The majority of researchers still employ outdated models like descriptive analysis, linear regression analysis, and others [3,4,5]. The outcome of the study's such as applied to determine the high-symptoms related to CRC. Moreover, estimates of tumor size are provisional and approximate. To solve and identify the high-risk symptoms of colorectal cancer, the study must utilise the most up-to-date prediction models, such as fuzzy logic to handle the problem since the CRC data information is imprecise.

Several elements of colorectal cancer may be recognised since the new hybrid FLRWSPCSVM model may accurately predict colorectal cancer high-risk symptoms [6-7]. Also, the smallest measurement of error model is anticipated by the new hybrid FLRWSPCSVM models. In other words, the new hybrid FLRWSPCSVM models are more accurate and effective at predicting the high-risk signs and symptoms of colorectal cancer. The potential new issue will be reduced by developing an accurate prediction model for the investigation and monitoring of colorectal cancer high-risk symptoms.

MATERIALS AND METHODS

Patients with colorectal cancer of all stages made up the study's population, while secondary data on actual colorectal cancer rates came from a general hospital in Kuala Lumpur, Malaysia and the symptoms such as icd10, TNM Staging (TNM), diabetes mellitus (DM), Crohn's disease (CD), ulcerative colitis (UC), polyp (P), endometrial (E), gastric (G), small bowel (SB), hepatobiliary (H), urinary tract (UT), ovarian (O), intestinal obstruction (IO), colorectal (C), weight loss (WL), diarrhoea (D), blood stool (BS), anemia (A), abdominal (AB). Doctors and nurses used cluster sampling to gather and record the data. Machine learning was carried out using the Statistical Package for Social Sciences, Matlab, Weka Explorer, and Microsoft Excel.

Fuzzy Linear Regression Model (Tanaka, 1982)

Any field can benefit from statistical analysis, particularly linear regression. In a fuzzy regression approach known as fuzzy linear regression, some model components are represented by fuzzy numbers. Hideo Tanaka conducted research on the FLRM method 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 errors are associated at the same time. The fuzzy model does not require any assumptions [8,9].

The input and output data that are ambiguous because they are based on fuzzy parameters. The model's explanation for data discrepancies is the system structure's ambiguity as expressed by fuzzy parameters [10].

$Yi = (\alpha_0, \varsigma_0)$, where α_0 is the fuzzy triangle diagram's centre and ς_0 is its width, is used to represent fuzzy output. The following is the fuzzy linear regression's linear function:

$$Y = A_0(\alpha_0, \varsigma_0) + A_1(\alpha_1, \varsigma_1)\, x_1 + \ldots + A_g(\alpha_g, \varsigma_g)\, x_g \quad (1)$$

Where $A=[A0,A1...Ag]$ is a vector of fuzzy coefficients represented as a triangular fuzzy number and $X=[i,c_i]$ is a vector of independent variables. By using the provided data and resolving the linear programming issue, FLR's fitting model can be improved. In addition, a linear programming problem can be used to fine-tune the fuzzy parameter [11].

$$\alpha^t xi + (1-H)\sum_j c_j |x_{ij}| \geq y_i + (1-H)e_i \quad (2)$$

$$-\alpha^t xi + (1-H)\sum_j c_j |x_{ij}| \geq -y_i + (1-H)e_i \quad (3)$$

Fuzzy Linear Regression with Symmetric Parameter (Zolfaghari, 2014)

Apart Professional researchers frequently employ fuzzy linear regression with symmetric parameters (FLRWSP) for studying confusing phenomena. Using symmetric parameters, fuzzy linear regression can reflect a variety of hazy and ambiguous conditions. In the study, fuzzy linear regression was used by the researcher to assess food quality, particularly that of fried donuts. From a scientific and technical standpoint, model theory is beneficial since it offers the conceptual framework and results that can easily be applied in system models using the fuzzy approach and recent developments in fuzzy logic [12, 13,14,15]. The definition of a triangle fuzzy number is *=(f c(x), f 3(x))* when *fc(x)* is the mode and *f3(x)* is the spread of a triangular fuzzy number. If *(i = 0, 1,..., n)* is a symmetrical fuzzy number and x_i is a crisp real integer.

The FLRWSP model can be expressed as follows:

$$f^s(x) = s_0 + s_1x_1 + \ldots + s_nx_n \quad (4)$$

$$f^c(x) = a_0 + a_1x_1 + \ldots + a_nx_n \quad (5)$$

In the case of a triangular fuzzy number, the goal function is defined as follows:

$$(1-h)s_0^L + (1-h)\sum_{i=1}^n (s_i^L |x_{ji}|) - a_0 - \sum_{i=1}^n (a_i x_{ji}) \geq -y_j \quad (6)$$

$$(1-h)s_0 + (1-h)\sum_{i=1}^n (k_i s_i^L |x_{ji}|) + a_0 + \sum_{i=1}^n (a_i x_{ji}) \geq -y_j \quad (7)$$

Fuzzy Clustering Method

A data set can be a member of more than one cluster when using the fuzzy C-means (FCM) clustering technique. This technique was created by Dunn (1973) and enhanced by Bezdek (1981). As a result of the algorithm's foundation in fuzzy C- means minimization in the direction of the following objective function or criterion, such as:

$$j_m = \sum_{i=1}^N \sum_{j=1}^c u_{ij}^m ||x_i - c_j||^2, 1 \leq m \leq \infty \quad (8)$$

The total number of items is *N*, and the total number of clusters is *C*. The index *r (r=1,..,C)* corresponds to cluster number *r*, while the index *q (q=1..., N)* corresponds to object number *q*. The algorithm for minimising *J* in the case of Euclidean distance can be summed up as follows.

1) Choose cluster centres *'c'* at random. Select the fuzziness exponent, *z* > 1, then the termination tolerance between 0 and 1.

2) By calculating the weighted average for each group and the Euclidean distance, the distance, d_{qr}, is updated for the provided *qr*.

3) Revise the membership values as,

$$u_{qr} = \frac{1}{\Sigma_{k=1}^{c}\left(\frac{d_{qr}}{d_{qk}}\right)^{\frac{2}{z-1}}}, \text{For } z = 1 \quad (9)$$

4) Determine the objective or criterion *J* and iterate until the objective function is minimised. If *k* = 1, 2,.., or, the iteration repeated; otherwise, step 2 was repeated.

Support Vector Machines Model

Vapnik's 1963 Support Vector Machines (SVM) technique is used to find subtle patterns in large, complicated data sets. To forecast the sorting or regression of previously unobserved data, the system does linear classification and regression analysis [16, 17, 18].

The separation margin between positive and negative provides strong generalisation performance, therefore linear SVM machines build a hyperplane as a decision surface [9, 19, 20]. Three learning machines can be built for support vector learning: the polynomial kernel, the radial basis functions kernel, and the two-layer perceptron. An illustration of a polynomial kernel function:

$$K(x,y) = (x^T y + 1)^d \quad (10)$$

Furthermore, by implicitly translating their inputs to high-dimensional feature spaces, SVMs can effectively execute nonlinear sorting when doing linear classification utilising the kernel-method.

A New Hybrid FLRWSPCSVM

Fuzzy linear clustering regression, which combines FLR and FCM. The hybrid model combines the SVM model and the FLRC model. There are five steps involved in creating the hybrid:

1) Determine the correlation between *Y* and X_i that is higher.

2) The modelling of the FLRWSP cluster, which combines FLRWSP with fuzzy C-means, is the first stage of the hybrid. Based on the *Y* data alone and the *Y* data towards multiple higher values of the independent variables that have higher correlation values, FLRWSP and FCM combine clustering. Based on the MSE and RMSE values with the least values, the best FLRWSP clustering is chosen. The FLRWSP clustering is optimised by changing the *h* value between 0 and 1 (in 0.1 steps) with the least amount of statistical error [21].

3) Find the FLRWSP clustering and SVM residual.

4) Using the Eq. 8, the second hybrid step creates the new hybrid data.

$$\mathrm{Y} = \mathrm{Lt} + \mathrm{Nt} \quad (11)$$

Here, L_t is the residual of the FLRWSP cluster (nonlinear model), and N_t is the residual of the SVM model. *Y* is a new data set (linear model). The SVM model is chosen for a hybrid model because it is a linear model that can minimise model error and is not overly sensitive to outliers.

5) Using the FLRWSP approach to model a hybrid

$$\mathrm{Error}final = (n1 \text{ x ERROR1}) + (n2 \text{ x ERROR2}) / n1 + n2 \quad (12)$$

where the numbers for clusters 1 and 2 are *n1* and *n2*, respectively. The MLR clustering number error is ERROR1, and the SVM model number error is ERROR2. The MSE and RMSE numbers would be considered error values.

In a new approach hybrid of FLRWSPCSVM, procedure 1-5 must be fulfilled to get the new approach hybrid of FLRWSPCSVM and it summarized in a flow chart as shown in Fig. 1. The model will be obtained in stop process [22, 23].

.

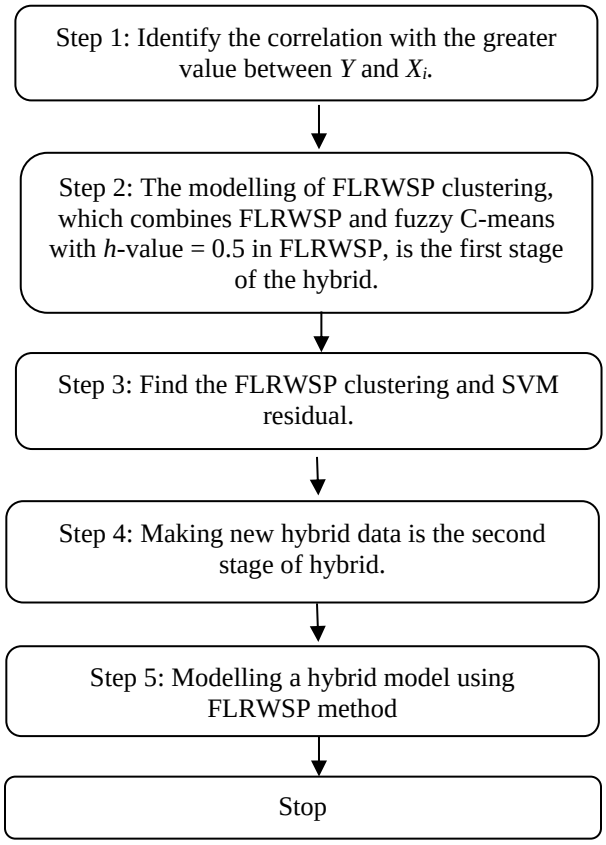

FIGURE 1. Framework of hybrid model

RESULTS

In this study, 180 patients' secondary data were the responses, and there were 19 variables. In contrast, the dependent variable for colorectal cancer is the tumor size. Using MSE and RMSE statistical cross-validation techniques, it was possible to compare these models such as fuzzy linear regression (FLR), fuzzy linear regression with symmetric parameter (FLRWSP), fuzzy linear regression by clustering and a new hybrid FLRWSPCSVM. The most accurate model for forecasting colorectal cancer tumor size is the one with the lowest MSE and RMSE.

A New Hybrid of Fuzzy Modeling

A new hybrid FLR model was created utilizing SVM clustering and two types of measurement error: MSE and RMSE, each of which have a fuzzy parameter. The value of MSE and RMSE may be calculated from the error measurement of MSE and RMSE using the sum of error FLR from clustering and error SVM model. The most accurate colorectal cancer tumor size prediction model has the lowest error value. The new hybrid model divided into two clusters to determine the error for each cluster by MSE and RMSE. Moreover, every cluster will have a parameter of their own follow by the weightage of variables.

Cluster 1 for a hybrid model

85 data were utilised as respondents in Cluster 1 of the hybrid FLR using clustering and SVM model. Model Cluster 1's MSE, RMSE value, and parameters are as follows:

TABLE 1. Measurement error of the model cluster 1.

Methods	Value
MSE	162.138
RMSE	12.733

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

$$\begin{aligned}\hat{Y} = 1.314 + (13.003, 0)\ \text{icd10} + (3.916, 0)\ \text{TNM} + (-0.654, 0)\ \text{DM} + (4.24, 0)\ \text{CD} + (0.626, 0)\ \text{UC} \\ + (-3.334, 0)\ \text{P} + (1.103, 0)\ \text{E} + (-2.493, 0)\ \text{G} + (-4.815, 0)\ \text{SB} + (-4.184, 0)\ \text{H} \\ + (8.754, 0)\ \text{UT} + (18.864, 0)\ \text{O} + (-12.827, 0)\ \text{IO} + (3.948, 0)\ \text{C} + (4.139, 0)\ \text{WL} \\ + (7.541, 0)\ \text{D} + (-2.614, 0)\ \text{BS} + (-6.109, 0)\ \text{A} + (0.733, 0)\ \text{AB}.\end{aligned}$$

Cluster 2 for a hybrid model

95 data were used as respondents in Cluster 2 of the hybrid FLR using clustering and SVM model. Model Cluster 2's MSE, RMSE value, and parameters are as follows:

TABLE 2. Measurement error of the model cluster 2.

Methods	Value
MSE	215.140
RMSE	14.667

The fuzzy mean value of tumor size (mm) can be explained by the diabetes mellitus, with the highest fuzzy parameter = 11.157

$\hat{Y}$ = 2.067 + (5.235, 0) icd10 + (3.110, 0) TNM + (7.089, 0) DM + (11.157, 0) CD + (3.984, 0) UC + (2.636, 0) P + (4.606, 0) E + (-3.658, 0) G + (-3.392, 0) SB + (10.439, 0) H + (5.594, 0) UT + (5.541, 0) O + (1.975, 0) IO + (-1.599, 0) C + (4.741, 0) WL + (3.504, 0) D + (-10.959, 0) BS + (-1.644, 0) A + (-4.583, 0) AB.

The final MSE and RMSE values of the model are shown in Table 3 After analysis, the residual FLRWSP cluster model and the residual SVM model were hybridized Table 3, which shows the MSE and RMSE value for this new hybrid model.

TABLE 3. The final measurement error of the model.

Methods	Value
MSE	100.605
RMSE	10.030

TABLE 4. Measurement error in each model.

Regression models	MSE	RMSE
FLR	277.952	16.671
FLRWSP	275.071	16.585
FLR by clustering	201.506	14.195
A new hybrid FLRWSPCSVM model	100.605	10.030

CONCLUSION

The hybrid method was used in conjunction with the novel fuzzy machine learning strategy of FLRWSP using clustering and SVM model. In order to depict an ill-defined phenomenon, a novel model was created. Based on two measurement error models, the most informative model, applicable to any research area, is a combination of FLR by clustering and SVM. The best model with the minimum value of measurement errors, according to the analysis of FLR, FLRWSP, linear fuzzy regression by clustering, and a novel hybrid FLRWSP model by clustering and SVM, is MSE and RMSE. The models' executive summary is displayed in Table 4.

It is shown that machine FLRWSPCSVM models is the best model for foretelling the high-risk indicators of colorectal cancer. This is so because a FLRWSPCSVM has the smallest MSE and RMSE values when compared to other models. Moreover, ovarian cancer and a diabetes mellitus are two symptoms that have a significant impact on the development of colorectal cancer. FLRWSPCSVM Model, which is the best predicted model to identify the high-risk symptom for colorectal cancer, should be used by general hospitals to address the issue.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2021/STG06/UTHM/03/1).

This research also was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through GPPS (Q561).

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 (Vot Q424).

REFERENCES

1. J. Lewis, and T. Siaw-Liaw, "Using information management systems and processes to support shared care for colorectal cancer survivors," 2017 IEEE International Symposium on Technology and Society (ISTAS), Sydney, NSW, Australia, 2017, pp. 1-5.
2. A. Agresti, *An Introduction to Categorical Data Analysis* (John Wiley & Sons, Inc, New York, 1996).
3. S. N. Salahudin, H. S. Ramli, M. N. R. Alwi, M. S. Abdullah, and N. A. Rani, Int. J. Recent Technol. Eng. **8**(2S), 643-651 (2019).
4. S. Nizam Salahudin, H. Suhaila Ramli, M. Helmy Abd Razak, M. Safizal Abdullah, and A. Masum, Estudios de Economia Aplicada J. **39**, 1-9 (2021).
5. Y. Dasril, G. K. Wen, B. Nazarudin, N. S. Salahudin, Int. J. Electr. Comput. Eng. **12**(5), 5182–5190 (2022).
6. M. A. Shafi, M. S. Rusiman, and S. N. Syuhada Abdullah, J. Math. Stat. **9**(1), 36–40 (2021).
7. A. H. Diyana Izyan, M. S. Rusiman, C. H. Norziha, M. A. Shafi, O. G. Alma, and S. Suhartono, "A time series analysis for sales of chicken based food product," in *Proceedings of Sciemathic 2020*, AIP Conference Proceedings 2355, 060002 (2021).
8. H. Tanaka, S. Uejima and K. Asai. IEEE Transactions on Systems, Man and Cybernetics. **12**, 903-907 (1982).
9. H. Tanaka, Fuzzy Set Syst. **24**, 363-375 (1987).
10. L. A. Zadeh, Infor. Control. **8**, 338-358 (1965).
11. Y. Ni, "Fuzzy correlation and regression analysis," Ph.D. thesis, University of Oklahoma Graduate College, 2005.
12. Z. S. Zolfaghari, M. Mohebbi and M. Najariyan, Appl. Soft Comput. **22**, 417-423 (2014).
13. P. Pin-Feng, L. Chih-Sheng, Int. J. Manag. Sci. **33**, 497-505 (2005).
14. B. Debasish, P. Srimanta, C. P. Dipak, J. Neural Inf. Process. **10**, 203-224 (2007).
15. B. M. Rhonda, D. D. Jones, H. T. Lynch, R. E. Brand, P. Watson, R. Ashwathnayaran and H. K. Roy, World J Gastroenterol. **12**, 4485-4491 (2006).
16. I. Hyodo, H. Suzuki and K. Takahashi, "Colorectal Cancer Working Group Report," in *30th Asia-Pacific Cancer Conference*. Jpn. J. Clin. Oncol. **40**, 38-43 (2010).
17. S. D. Eduardo, H. Deneo-Pellegrini and A. L. Ronco, Asian Pacific J. Cancer Prev. **12**, 753-759 (2011).
18. Y. H. Mohamed, N. Daud, N. M. Noor and A. A. Rahim. Asian Pacific J. Cancer Prev. **13**, 3983-3987 (2012).
19. K. U. Jayaram, O. Dewey and L. Zhao, Med. Image Anal. **18**, 752-771 (2014).
20. K. Rajeswari and V. Vaithiyanathan, Int. J. Comput. Sci. Netw. **11**, 126-130 (2011).
21. R. Nishihara, K. Wu and P. Lochhead, N. Engl. J. Med. **369**, 1095-1105 (2013).

22. L. Xiaofeng, M. Lin, Z. Sheng, & M. Joseph, “Using fuzzy cmeans and fuzzy integrals for machinery fault diagnosis,” in *International Conference on Condition Monitoring*, Cambridge, England, 2005, pp. 1-9.
23. S. Elizabeth and L. Sujathan, Appl. Math. Sci. **7**, 6297-6307 (2013).