# RECURRENT GASTRIC CANCER PREDICTION USING RANDOMIZED SEARCH CV OPTIMIZER

Vishnu M K
*Computer Science And Engineering*
*Kongu Engineering College*
Erode,India
mkv1722@gmail.com

Vishal Rupak V R
*Computer Science And Engineering*
*Kongu Engineering College*
Erode,India
vishalrupakvr@gmail.com

Vedhapriyaa S
*Computer Science And Engineering*
*Kongu Engineering College*
Erode,India
vedhapriyaa1212@gmail.com

Ms.M.Sangeetha M.E.,
Assistant Professor (Sr.G)
*Computer Science And Engineering*
*Kongu Engineering College*
Erode,India
sangeethcse@gmail.com

Dr.R.Manjuladevi M.E., Ph.D
Associate Professor
*Computer Science And Design*
*Kongu Engineering College*
Erode,India
rmanjuladevi.gem@gmail.com

Ms.C.Sagana M.E.,
Assistant Professor (Sr.G)
*Computer Science And Engineering*
*Kongu Engineering College*
Erode,India
Sagana.c@gmail.com

*Abstract*—***Gastric cancer has the second-highest mortality rate and the fourth-highest incidence of malignant tumors worldwide. Early stomach cancer has a fair prognosis, but the symptoms are unusual and the signs are not immediately apparent. The goal of the project was to use patient clinicopathological data and machine learning techniques to increase the precision of a prediction model for the emergence of gastric cancer. The current study proposes the first four parameters. By comparing the performance of Gradient Boosted Decision Tree, Random Forest, Decision Tree, and Gradient Boosting Machine, Logistic Regression, it was shown that machine learning approaches can be used to estimate the repetition of patients with malignant after such operation. The parameters influencing postsurgical repetition of stomach cancer have been proposed in the present study. In this case, Random Forest is enhanced using the Random Search Cross Validation optimizer to increase the model's performance in predicting cancer. Random Search Cross Validation performs a random search over parameters, sampling each setting from a range of potential parameter values. The dataset contains 2012 patient records as well as 51 characteristics.***

## I. Introduction

In terms of malignant tumors, gastric cancer has the fourth-highest global incidence and the second-highest fatality rate. Each year, there are around 1 million new instances of stomach cancer worldwide, with China accounting for close to 50.0% of these cases. Although the potential side effects are strange and take time to manifest, the treatment for early-stage stomach cancer is favorable. There was no change in the greater incidence of surgical recurrence (40.0-70.0%). The average time it took for stomach cancer to recur after surgery was 20.5-28.0 months, according to reports. Chemotherapeutic and immunotherapeutic, however, can be used in close coordination to treat the illness, decrease necrosis, and attempt to get ready for surgery when gastric cancer recurs after surgery. A number of organizations have recently undergone changes as a result of big data and machine learning. Precision medicine plans, which integrate machine learning with medical and health big data, have brought a future big data health cause within reach. Making predictions based on past data is a task for which the learning algorithm is especially well suited. By collecting complicated dynamic interactions in the data, the computer vision approach can improve predictive performance more than the classic regression model. Algorithms can currently identify a patient's survival from a phase 1 breast cancer. Machine learning was used to more precisely forecast the likelihood of early biochemical recurrence and the level of malignancy of breast lesions. A branch of artificial intelligence known as machine learning has been applied to the evaluation of tumor risk, the detection of lesions, the prognosis prediction, and the prediction of therapy response. Medical problems that were previously thought to be insurmountable may be successfully solved via deep learning. In conclusion, postoperative recurrence is a significant factor influencing the treatment plan of stomach cancer. Investigating the hazard for postoperative treatment failure of stomach cancer is crucial in order to locate the disease and establish whether it has returned or spread. In order to forecast whether people with stomach cancer would develop malignancies again after surgery, we applied an algorithm.

**978-1-6654-8035-2/22/$31.00 ©2022 IEEE**

## II. Literature Review

Sung Hoon Noh [1] comprised 1035 patients, 520 of whom were receiving adjuvants of oxaliplatin and capecitabine, and 515 of whom were being monitored. The median follow-up for the trial's population with a treatment goal was 624 months (IQR 54–70). 139 (27% of patients) and 203 (39%) of those in the prophylactic capecitabine plus cisplatin group experienced disease-free mortality events, respectively (stratified relative risk [HR] 058; 95% CI 047-072; p00001). In the adjunct capecitabine + oxaliplatin group, the predicted 5-year disease-free survival was 68% (95% CI 63-73) as compared to 53% (47-58) in the monitoring alone group. In the supplemental capecitabine and oxaliplatin group, 103 patients (20%) had passed away by the clinical cutoff date compared to 141 patients (27%).in observation group.In comparison to the observation group, which received only adjuvant capecitabine and oxaliplatin, the estimated 5-year overall survival was 78% (95% CI 74-82) in the capecitabine and oxaliplatin group. No information on negative incidents was gathered after the initial analysis.

The mean classifier performance values of RCB class prediction (AUC, 0.86) and DSS projection (AUC, 0.92) depending on XGBoost, as well as RFS prediction with logistic regression, show that Tahmassebi, Amirhessam[3], Computer vision with mpMRI consistently performed well. The XGBoost classifier regularly and accurately outperforms the others when contrasted to other classifiers. The most crucial elements in determining RCB class were lesion diameter variations, the general patterns of shrinking, and mean travel times on Dynamic contrast - enhanced, minimal ADC on DWI, and peritoneal edema on T2-weighted imaging. The most crucial variables in forecasting RFS were volume dispersion, plasma concentration flow, median transit time, DCE-MRI lesions size, lowest, max, and mean ADC with DWI. The characteristics included lesion size, volume distribution, mean plasma flow, and maximal ADC with DWI were the features most important for predicting DSS.

Migita K [10], The standard deviation was 3.3 kg/m2, and the mean BMI was 22.5 kg/m2 at the time of surgery. The BMI subgroup revealed that 134 patients (21%) were overweight, 431 (67.6%) were of a normal weight, and 73 (11.4%) were underweight. Patients who were underweight had a 5-year overall survival (OS) rate of 66.6%, patients who were normal weight had an OS rate of 81.3%, and patients who were overweight had an OS rate of 79.9% (P = 0.001). The OS rate was significantly lower in stage I cancer patients who were underweight compared to stage II and stage III cancer patients who were normal weight or overweight. Being underweight was identified in the multivariate analysis as an independent predictor of OS, but not in patients with stage II or stage III illness.

K.Kulig [11], 492 individuals, or 25% of all patients, were overweight in 1992. Increased BMI was associated with greater rates based on inter abscesses and postoperative cardiac complications . However, death rates and other issues remained the same. Patients with a BMI over 25 kg/m2 had a significantly shorter median disease-specific mortality (25.7 months, P = 0.003) than those with a BMI under 25 kg/m2. These differences were caused by the lower proportion of patients with T3 and T4 cancers, metastatic lymph nodes, metastatic disease, and non-curative resections. In a Cox proportional risk analysis, it was discovered that age, level of infiltration, lesion metastasis, distant metastases, and type of residual cancer were all independently predicted factors.

## III. Proposed Work

The tumor regrowth of stomach cancer in patients following surgery can be predicted using machine learning approaches. The model is trained and tested using the clinicopathological dataset for the patient. The mean value from the training set's nearest neighbors is used to impute missing values. Machine learning models can be more accurate by using optimization algorithms.

### 3.1 Data Preprocessing

#### 3.1.1 *Data Reduction*

Reducing the quantity of storage space required is the process of data reduction. Data minimization can lower expenses and improve storage effectiveness. Storage providers frequently use the words raw capacity and effective capacity, which refer to data after reduction, to characterize storage capacity. Twelve factors, including height (cm), gender, age (year), BMI (kg/m2) and weight (kg), operation time (minutes), location, tumor size (cm), chemotherapy, dissection, death, and circumference—were chosen out of the 51 qualities. Included is the target variable Recurrence.

#### *3.1.2 Data Cleaning*

Data cleaning involves locating and removing inaccurate or corrupt data from a current record, spreadsheet, or data warehouse. KNN Imputer, which offers Imputation for filling in missing values using k-Nearest Neighbors, was used to fill in the missing values.

#### *3.1.3 Training and Test dataset*

Using the train test split technique from Scikit, a training set and a test set were created from the dataset in a 4:1 ratio. It is employed to randomly select train and test subsets from arrays or matrices. It is a quick tool that encapsulates next (ShuffleSplit) and input validation(). Splitting (and optionally subsampling) data in a single call using split(X, y) and application allows for quick data entry.

### *3.2 Training Process*

Random forest model and Gradient Boosted decision Tree were trained using RandomForestClassifier, GradientBoostingClassifier from sklearn.ensemble respectively. Linear Regression model was trained using LogisticRegression from sklearn.linear_model. The Decision Tree model was trained using DecisionTreeClassifier from sklearn.tree. The Light Gradient Boosting Machine was trained using the LGBMClassifier from lightGBM. All models were trained using the same training dataset obtained from the train_test_split of scikit-learn.

3.3 *Flow Diagram*

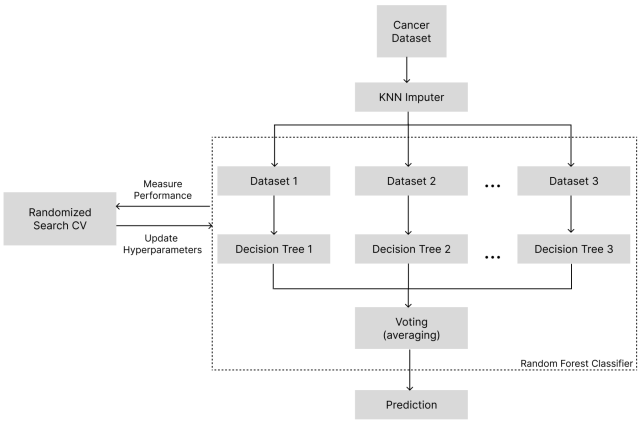


Figure 3.3 Flow Diagram

## IV. Methods and Algorithms

### *4.1 Machine Learning Algorithm*

Logistic Regression, Random forest, Gradient Boosting Classifier, LightGBM and Decision Tree are some of the several machine learning techniques employed.

### *4.2 Optimization*

The Random Search Cross Validation optimizer was used to improve the Random Forest model. The open-source Python machine learning toolkit scikit-learn offers methods for adjusting model hyperparameters. The complete Random Forest model optimization procedure employing Random Search CV may be summed up as follows:

Step 1 : A search space is described as a bounded domain of hyperparameter values.
Step 2 : Sample that domain's points at random.
Step 3 : Create a random search instance and fit it to a model using Scikit-Learn. Indicate the number of cross-validation folds to use and the number of iterations, or the number of possible combinations to try.
Step 4 : Get the best params and best model from the optimizer. Compare the base model and the top random search model to see if random search produced a superior model.

## V. Results And Discussion

### *5.1 Dataset Description*

The proposed work is evaluated on the patient's clinicopathological dataset. It has 2012 patient records in the dataset. We have selected 15 attributes to work with missing values are filled using KNN-Imputer using mean (K = 5).

#### *5.1.1 Sample Dataset*

The data can be found in the BioStudies database, accession number S-EPMC4344235. 2012 patients were included in this study. Gender, age, treatment-related factors, pathological characteristics, and the follow-up period pertaining to survival status were all collected in the retrospective investigations.
https://docs.google.com/spreadsheets/d/1sQGe4QUPyGlxbzeSqRpQTBP38W1bJYQ7I6t4bd-wOIk/edit?usp=sharing

#### *5.1.2 Correlation Analysis*

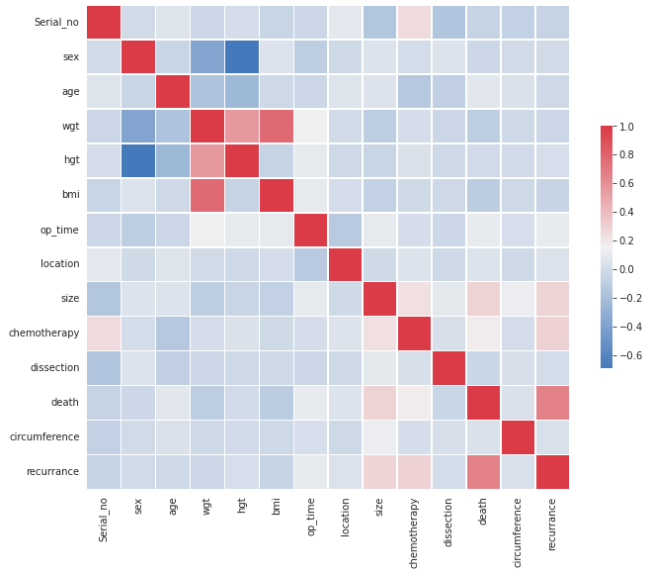


Figure 5.1.2 Correlogram

### *5.2 Test Set Outcomes*

The accuracy of the six algorithm models in the test group was highest for the Random Forest with Random Search CV technique (0.906), followed by the forest algorithm, the Light Gradient Boosting Machine, and the Gradient Boosting Classifier. DecisionTree (1.0), Gradient-Boosting (0.914), Random Forest with Random Search CV (0.902), forest (0.895), gbm (0.88), and Logistic (0.828) have the highest AUC values among the six methods. Random Forest with Random Search CV has the best precision rate (0.7963), followed by Random Forest (0.7959). The Random Forest with Logistic Regression model has the best specificity rate (0.970), followed by the Random Forest and Random Search CV (0.9670). Light GBM has the best sensitivity rate (0.6286), then Random Forest with Random Search CV (0.6143).

### *5.3 Parameter For Evaluation*

The performance is calculated in terms of following measures

1. Accuracy
2. Sensitivity
3. Specificity
4. Precision

Accuracy :

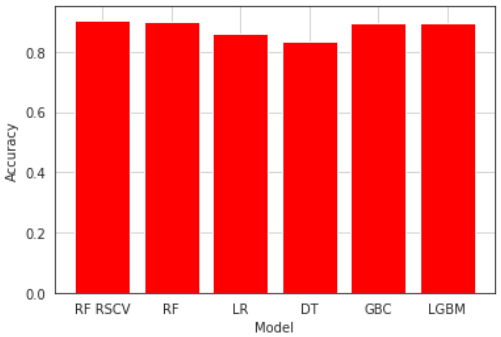


Figure 5.3.1. Accuracy Bar graph

Sensitivity :

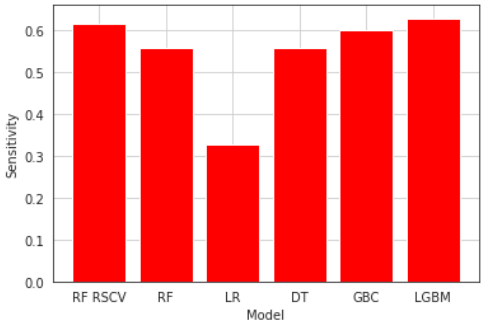


Figure 5.3.2. Sensitivity Bar graph

Specificity :

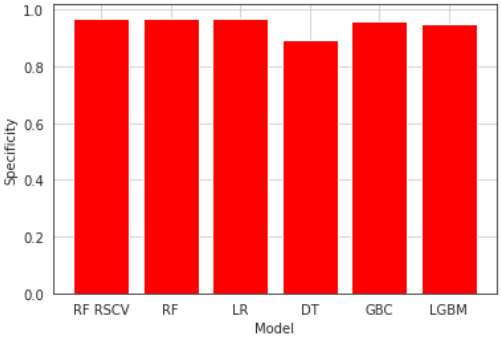


Figure 5.3.3. Specificity Bar graph

Precision:

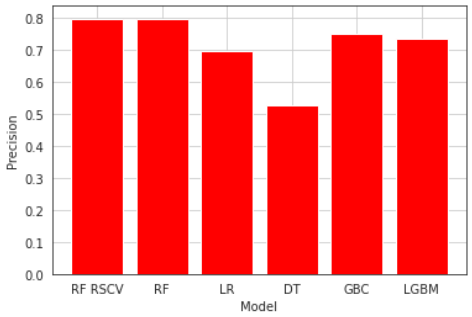


Figure 5.3.4. Precision Bar graph

*5.4 AUC-ROC (Area under the receiver operating characteristic curve)*

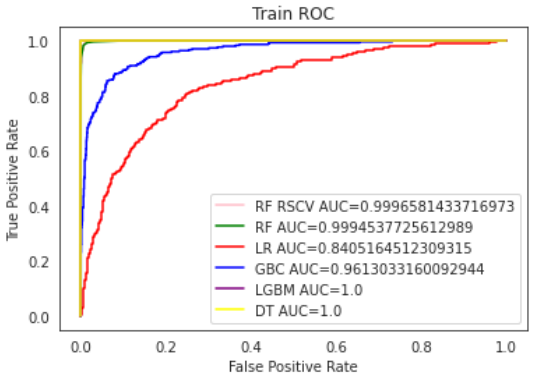


Figure 5.4.1.Train ROC

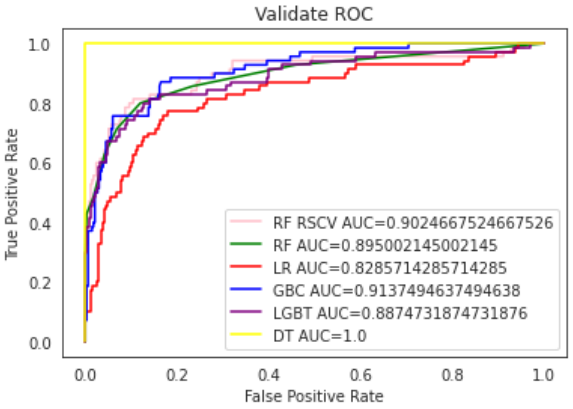


Figure 5.4.2 Validate-ROC

*5.4 Model Comparison*

| S.No | Model | Accuracy | Specificity | Precision | Sensitivity |
|---|---|---|---|---|---|
| 1 | RF RSCV | 90.57 | 96.70 | 79.63 | 61.43 |
| 2 | RF | 89.83 | 97.00 | 79.59 | 55.71 |
| 3 | LR | 85.86 | 97.00 | 69.7 | 32.86 |
| 4 | LGBM | 89.58 | 95.20 | 73.33 | 62.86 |
| 5 | GBC | 89.58 | 95.80 | 75.00 | 60.00 |
| 6 | DT | 83.62 | 89.49 | 52.70 | 55.71 |

TABLE 5.4.1 MODEL COMPARISONS

## Conclusion

Consequently, machine learning can forecast if stomach cancer patients would relapse following surgery. Additionally, BMI, operation duration, weight, and age were the first four factors determining the reappearance of stomach cancer following surgery. The best algorithms for predicting recurrent gastric cancer in cases of stomach cancer were discovered to be GNB, XGBoost, and Random Forest. To further increase the precision of the predictions made by the machine learning models, optimization approaches can be utilized.

In the coming years, more precise machine learning studies with more instances are required in order to identify the most accurate algorithm and maybe make individualized therapies applicable.

## References


[1] Sung Hoon Noh, Sook Ryun Park, Han-Kwang Yang, Hyun Cheol Chung, Ik-Joo Chung, Sang-Woon Kim, Hyung-Ho Kim, Jin-Hyuk Choi, Hoon-Kyo Kim, Wansik Yu, Jong Inn Lee, Dong Bok Shin, Jiafu Ji, Jen-Shi Chen, Yunni Lim, Stella Ha, Yung-Jue Bang, 5-year follow-up of an open-label, randomized phase 3 trial", The Lancet Oncology, Volume 15, Issue 12, 2014, Pages 1389-1396

[2] Dhar D, K, Kubota H, Tachibana M, Kotoh T, Tabara H, Masunaga R, Kohno H, Nagasue N: "Body Mass Index Determines the Success of Lymph Node Dissection and Predicts the Outcome of Gastric Carcinoma Patients". Oncology 2000;59:18-23. doi: 10.1159/000012131.

[3] Marrelli D, De Stefano A, de Manzoni G, Morgagni P, Di Leo A, Roviello F. Prediction of recurrence after radical surgery for gastric cancer: a scoring system obtained from a prospective multicenter study. Ann Surg. 2005 Feb;241(2):247-55. doi: 10.1097/01.sla.0000152019.14741.97. PMID: 15650634; PMCID: PMC1356909.

[4] Alarcon-Ruiz, C. A., Heredia, P. & Taype-Rondan, A. "Association of waiting and consultation time with patient satisfaction: secondary-data analysis of a national survey in Peruvian ambulatory care facilities". BMC Health Serv. Res. 19, 439. https://doi.org/10.1186/s12913-019-4288-6 (2019)

[5] J. Kulig, M. Sierzega, P. Kolodziejczyk, J. Dadan, M. Drews, M. Fraczek, A. Jeziorski, M. Krawczyk, T. Starzynska, G. Wallner, "Implications of overweight in gastric cancer: A multicenter study in a Western patient population", European Journal of Surgical Oncology (EJSO), Volume 36, Issue 10, 2010, Pages 969-976, ISSN 0748-7983, https://doi.org/10.1016/j.ejso.2010.07.007. (2010).

[6] Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J. & Samleerat, T. HIV-1 CRF01_AE "coreceptor usage prediction using kernel methods based on logistic model trees". Comput. Biol. Med. 42, 885–889 (2012).

[7] Alarcon-Ruiz, C. A., Heredia, P. & Taype-Rondan, A. "Association of waiting and consultation time with patient satisfaction: secondary-data analysis of a national survey in Peruvian ambulatory care facilities". BMC Health Serv. Res. 19, 439 (2019).

[8] Hsu-Huan Chou, Chia-Jung Kuo, Jun-Te Hsu, Tsung-Hsing Chen, Chun-Jun Lin, Jeng-Hwei Tseng, Ta-Sen Yeh, Tsann-Long Hwang, Yi-Yin Jan,Clinicopathologic study of node-negative advanced gastric cancer and analysis of factors predicting its recurrence and prognosis,The American Journal of Surgery,Volume 205, Issue 6,2013,Pages 623-630,ISSN 0002-9610,

[9] Migita, K., Takayama, T., Matsumoto, S. et al. "Impact of being underweight on the long-term outcomes of patients with gastric cancer".Gastric Cancer 19, 735–743 (2016).

[10] Lo, SS., Wu, CW., Chen, JH. et al. "Surgical Results of Early Gastric Cancer and Proposing a Treatment Strategy". Ann Surg Oncol 14, 340–347 (2007).

[11] Tokunaga, M., Hiki, N., Fukunaga, T. et al. "Better 5-Year Survival Rate Following Curative Gastrectomy in Overweight Patients". Ann Surg Oncol 16, 3245–3251 (2009).

[12] Eom, B. W. et al. "Survival nomogram for curatively respected Korean gastric cancer patients: multicenter retrospective analysis with external validation". PLoS ONE 10, e0119671 (2015)

[13] Wu, B., Wu, D., Wang, M. & Wang, G. "Recurrence in patients following curative resection of early gastric carcinoma". J. Surg. Oncol. 98, 411–414(2008).

[14] Bickenbach, K.A., Denton, B., Gonen, M. et al. "Impact of Obesity on Perioperative Complications and Long-term Survival of Patients with Gastric Cancer". Ann Surg Oncol 20, 780–787 (2013).

[15] Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J. & Samleerat, T. HIV-1 CRF01_AE "coreceptor usage prediction using kernel methods based on logistic model trees". Comput. Biol. Med. 42, 885–889 (2012).

[16] Kruhlikava, I., Kirkegård, J., Mortensen, F. V. & Kjær, D. W. "Impact of body mass index on complications and survival after surgery for esophageal and gastro-esophageal-junction cancer". Scand. J. Surg. 106, 305–310 (2017).

[17] Matsuoka T, Yashiro M. Biomarkers of gastric cancer: Current topics and future perspective. World J Gastroenterol. 2018 Jul 14;24(26):2818-2832. doi: 10.3748/wjg.v24.i26.2818. PMID: 30018477; PMCID: PMC6048430.

[18] Cuocolo, R., Caruso, M., Perillo, T., Ugga, L. & Petretta, M. "Machine learning in oncology: a clinical appraisal". Cancer Lett. 481, 55–62 (2020).

[19] Wenjuan Zhang, Mengjie Fang, Di Dong, Xiaoxiao Wang, Xiaoai Ke, Liwen Zhang, Chaoen Hu, Lingyun Guo, Xiaoying Guan, Junlin Zhou, Xiuhong Shan, Jie Tian,

[20] Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer, Radiotherapy and Oncology,Volume 145,2020,Pages 13-20,ISSN 0167-8140,