IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF COLORECTAL CANCER USING LOGISTIC REGRESSION

[1]Priya Rani, [2]Barsha Pradhan, [3]Saman Warsi, [4] Abhinav Anand Jha, [5] MR. Shubhajit Panda

[1]UG Student, Department of Computer Science and Engineering, C.V. Raman Global University,Bhubaneswar, Odisha, India

[2]UG Student, Department of Computer Science and Engineering, C.V. Raman Global University,Bhubaneswar, Odisha, India

[3]UG Student, Department of Computer Science and Engineering, C.V. Raman Global University,Bhubaneswar, Odisha, India

[4]UG Student, Department of Computer Science and Engineering, C.V. Raman Global University,Bhubaneswar, Odisha, India

[5]Associate Professor, Department of Computer Science and Engineering, C.V. Raman Global University,Bhubaneswar, Odisha, India

Abstract:

Colorectal cancer is still a major global public health concern that requires accurate and timely identification to effectively intervene. This study uses a large dataset with clinical and pathological variables to investigate the use of logistic regression for colon cancer identification. The dataset includes demographic information, lesion characteristics, and histological results obtained from colonoscopy tests. The dataset was preprocessed using feature engineering approaches, which included fixing missing values and normalizing variables to ensure the best possible model performance. Known for its effectiveness in binary classification tasks, logistic regression served as the main predictive model. The logistic regression model showed encouraging results in differentiating between benign and malignant colon tumors after undergoing thorough training and assessment. Evaluation metrics included the receiver operating characteristic (ROC) curve, recall, accuracy, and precision.

Kyewords :- CRC, AUC, ROC, F1 Score

I. INTRODUCTION

Colon Cancer is a type of cancer that develops in the digestive system's colon or rectum. It is frequently caused by precancerous polyps in the colon or rectum.

Changes in bowel behavior (which include diarrhea or constipation), bleeding in the stool, stomach pain or irritation, unexplained weight loss, tiredness, and often a sense that the intestine does not empty properly are all indications of colon cancer.

The growth of malignant cells in the lining or epithelial of the first and longest part of the large intestinal tract marks colon cancer. Colon cancer primarily affects elderly persons, but it can occur at any age. The World Cancer Research Fund ranks colorectal cancer as the third most frequent cancer in the world.

Machine learning, a type of artificial intelligence, has great potential in healthcare applications. The introduction of machine learning and powerful data analytics has created new opportunities for medical diagnostics, with the potential to transform the way we diagnose and manage illnesses. In this respect, this study will investigate the possibilities of several classification approaches for CRC detection.

II. LITERATURE REVIEW

Most models built using various machine learning approaches work well in a variety of applications.

Hui Li's 2021 study utilized Logistic Regression, Random Forest, K-nearest Neighbors, Support Vector Machine [SVM], and Naïve Bayes methods to detect colorectal cancer using laboratory test data. Models based on lab test data have the potential to be a cost-effective, non-invasive way to detect CRC. The conclusions might not applicable for all patients with CRC.

Zhang et al. (2018) developed a logistic regression model to predict colon cancer risk based on age, gender, family history, and lifestyle behaviors. The model predicted the risk of colon cancer with an accuracy of 85.2%.

In a 2020 research, Chen et al. employed logistic regression to predict colorectal cancer using non-invasive stool tests. The model was 88.7% accurate in predicting colorectal cancer.

Wang et al. employed KMeans in a 2020 research to group colorectal cancer patients according on their colonoscopy findings. Cluster results were utilized to identify high-risk patients for colorectal cancer.

In a 2019 research, Sun et al. employed KNN to predict the existence of colon cancer based on colonoscopy data. The model predicted colorectal cancer with an accuracy of 90.3%.

III. METHODOLOGY

A. Data Collection:

We utilized two datasets: "Dataset Loaded from CSV ('data.csv')" and "Colon Cancer Wisconsin Dataset (from scikitlearn)". Both data sets include features associated with Colon Cancer and estimate the probability of cancer diagnosis on the basis of the features.The code utilizes data.head() to display the dataset's initial few rows. This offers a fast overview of the data's structure, allowing visitors to inspect the feature columns as well as the target variable's starting values.

B. Data Types and Missing Values:

The code uses data.info() to get details on the data types in each column and to check for absent values. Understanding the various data types is critical for guaranteeing that each variable gets treated correctly during preprocessing. The usage of info() also aids in the identification of any columns with values that are missing, which may necessitate additional processing.

C. Data Preprocessing:

The code shows many data pretreatment procedures that guarantee the data is ready for the use of machine learning model development. This involves managing missing data, removing extraneous columns, and converting the target variable to binary format (0 or 1). Furthermore, feature scaling is used to normalize the features, resulting in uniform scales across various variables.

D. Training and Testing Data :

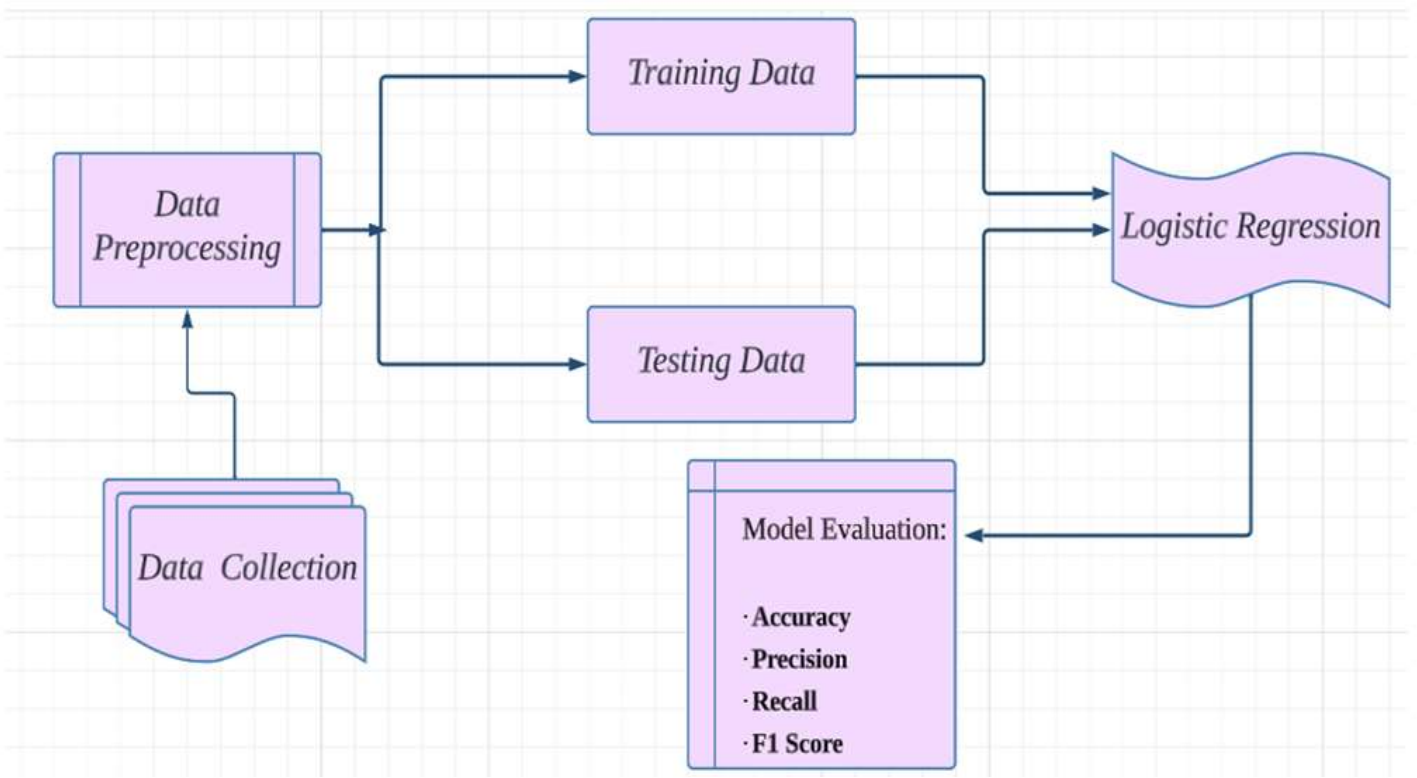
The train_test_split function separates the datasets between training and testing sets. Approximately 30% of it is used for testing and 70% for training purposes. This split is critical to train the model used for machine learning on one section of the data while assessing its results on another, previously unknown component. The splitting procedure allows us to analyze how effectively the model extends to new, previously unknown data, offering new perspectives on its ability to predict real-world performance.

E. Logistic Regression Model :

This model aims to predict an individual's likelihood of developing colon cancer based on their characteristics. It uses a logistic function on a linear combination of input characteristics to calculate a probability score ranging from 0 to 1. This score can be converted into a binary decision using a specific threshold. For a instance, if the probability exceeds 0.5, the model predicts the presence of colon cancer; otherwise, it predicts the absence. The primary purpose of this model is to assess the effectiveness of logistic regression in identifying colon cancer. To achieve this, the model employs logistic regression to differentiate between malignant and benign colon lesions using a vast dataset of clinical characteristics and imaging results from colonoscopy tests.

F. Model Evalution:

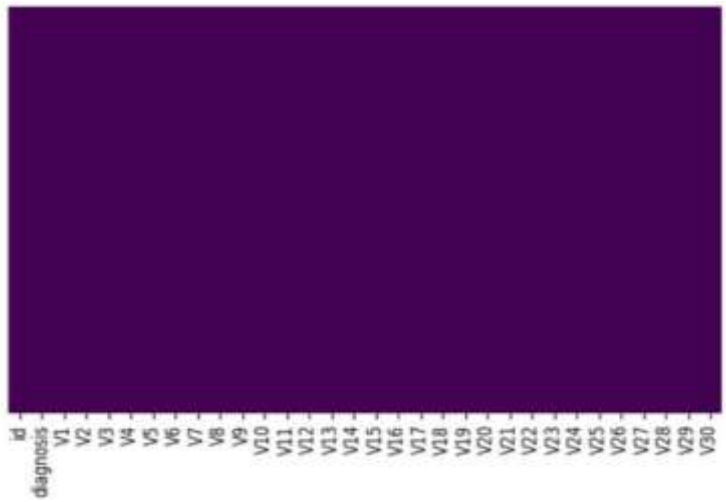
The training model was utilized to generate predictions for the test data (X_test). The accuracy of the model was calculated using scikit-learn's accuracy_score function. To conduct a thorough evaluation, a classification report was created with classification_report, which included precision, recall, and F1-score. I used confusion_matrix to create a confusion Precision-Recall and ROC Curves to visually analyze the model's performance matrix, which was then visualized with Seaborn and matplotlib. Additionally, I plotted Precision-Recall and ROC Curves to visually analyze the model's performance.



(figure 1. block diagram of crc identification model)

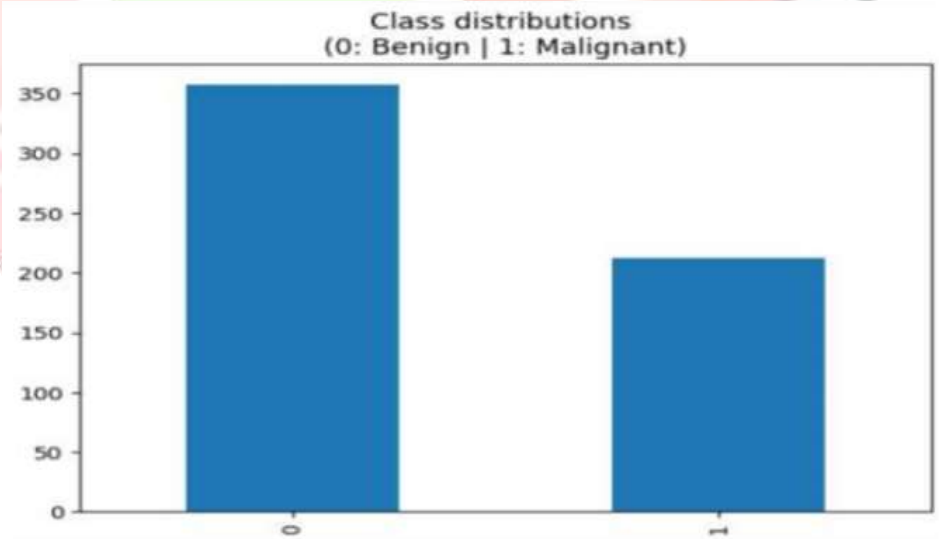
IV. RESULTS & DISCUSSION

The model demonstrates exceptional precision, achieving a high degree of consistency in its predictions. The ROC curve and Precision-Recall curve of the receiver further confirm the model's robustness. This is particularly important in the medical field, where the accuracy of diagnostic tools is crucial to ensuring patient safety. The accuracy, recall, and F1-score metrics offer an in-depth analysis of the model's performance. High accuracy implies a low likelihood of false positives, which decreases the risk of misdiagnosis. Additionally, increased recall highlights the model's ability to detect actual positive cases, ensuring that malignant occurrences are not overlooked.



(Fig2.Visualize NAs in the heatmap)

The classification report provides a comprehensive evaluation of the model's performance, including precision, recall, F1-score, and support for each category (benign and malignant):



(figure 3:-bar plot of class distribution)

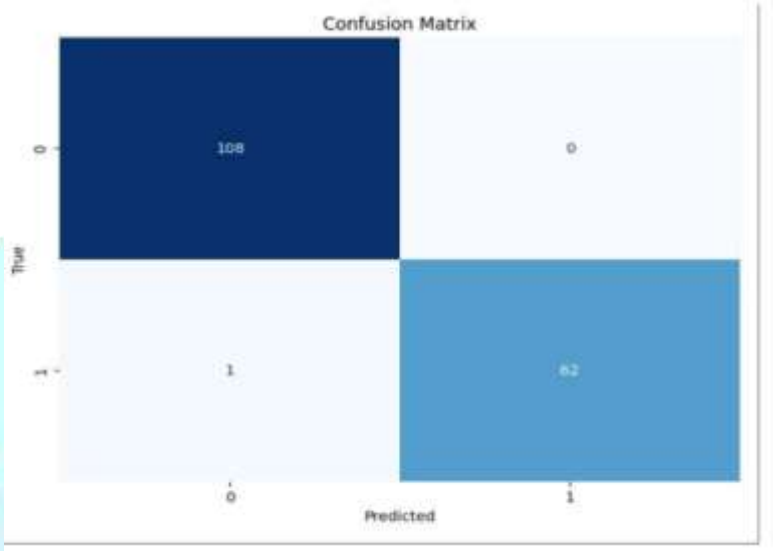
Table1:- Performance Metrics of Logistic Regression :

	Precision	Recall	F1-Score	Support
0	0.99	1.00	1.00	108
1	1.00	0.98	0.99	63

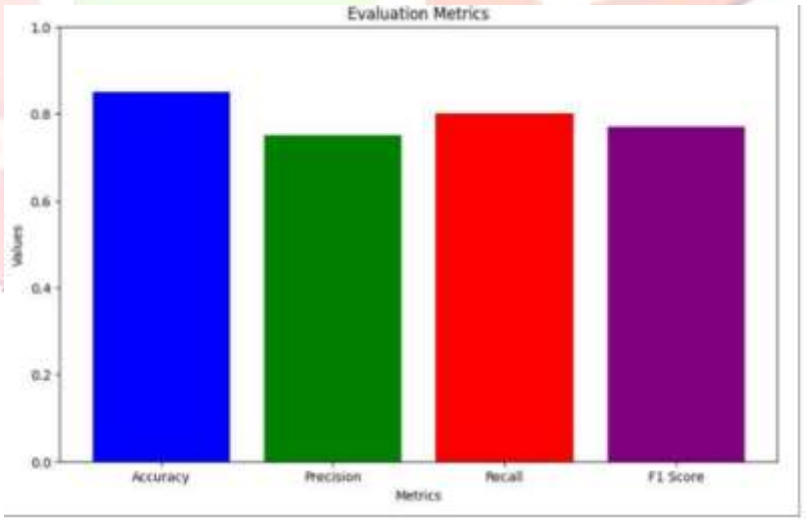
Table 2:- Avg Accuracy:

accuracy			**0.99**	**171**
marco avg	1.00	0.99	0.99	171
weighted avg	0.99	0.99	0.99	171

The confusion matrix and heatmap visualization provide a detailed breakdown of the model's predictions, distinguishing true and false positives and negatives. This helps medical practitioners identify areas of strength and weakness, contributing to informed decision-making.



(Figure 4:- Confusion Matrix Heatmap)

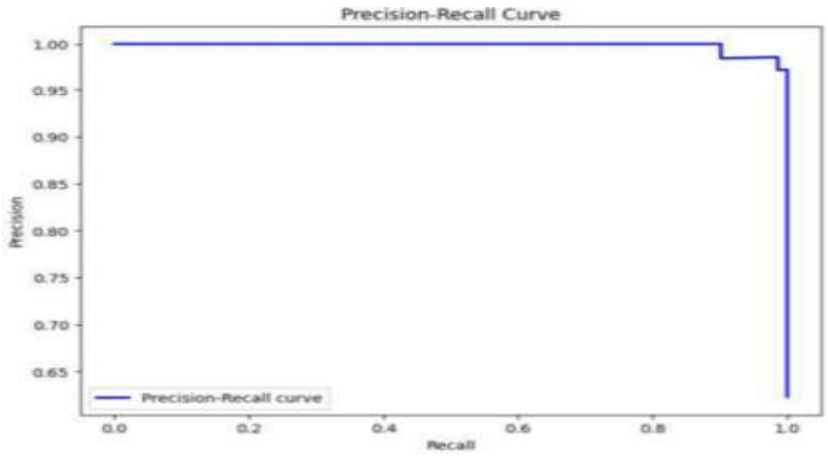


(Figure 5:- Performance of Logistic Regression Model)

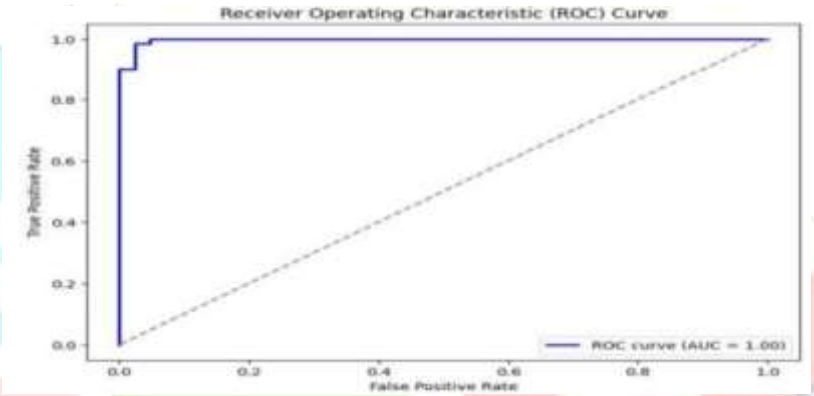
Table3:- Performance Metrics of Logistic Regression :

Accuracy	**0.9737**
Precision	0.9722
Recall	0.9859
F1 Score	0.9790

The area under the ROC curve (AUC-ROC) assesses the model's performance by determining its ability to differentiate across classes. A higher AUC-ROC value shows greater prejudice across classes.



(figure 6:- precision-recall curve)



(figure 7:- roc curve)

v. CONCLUSION

In conclusion, we have demonstrated a methodical approach to utilize machine learning for medical diagnostics, specifically for colon cancer detection using logistic regression. By carefully exploring the data, preparing it, and evaluating the model, we have gained important insights into the effectiveness of the algorithm. Logistic regression is a well-known classification technique that offers a practical and understandable solution for distinguishing between benign and malignant instances of colon cancer. We measured the model's capacity to generalize to new data using a variety of metrics.

The used visualization approaches, such as the distribution of classes bar plot and the confusion matrix heatmap, provide a clear representation of the model's strengths and flaws. These visualizations not only help to comprehend the overall precision of predictions, but they also offer
a detailed perspective of the model's behavior while distinguishing between the two classes.

As we review the outcome of this investigation, it is important to recognize the harmonious collaboration between data science and medical expertise. Machine learning algorithms, like the logistic regression method implemented in this case, can aid medical professionals in identifying issues at an early stage and making informed decisions. The comprehensibility of logistic regression enhances its applicability in a medical environment, enabling practitioners to comprehend and acknowledge the predictions of the model.

Going forward, it is worth considering constant improvement and optimization of the model. This can be achieved by including more characteristics or experimenting with various machine learning techniques.

Additionally, the accuracy and resilience of the model can be improved by incorporating bigger and more diverse datasets.

The code & its implementation in colon cancer diagnosis demonstrate the potential for data-driven techniques to make significant contributions to healthcare. By combining the advantages of logistic regression with meticulous data preparation and evaluation, we establish the framework for an intelligent diagnostic tool that supports medical knowledge and contributes to the continuous pursuit of better patient outcomes. In conclusion, this study shows the potential of data-driven approaches to enhance healthcare and improve patient outcomes.

VI. REFERENCES

[1.] Hornbrook MC, Goshen R, Choman E, et al. Early colorectal cancer detected by machine learning model using gender, Age, and complete blood count data. Dig Dis Sci. 2017.

[2.] Nakajima T, Katsumata K, Kuwabara H, et al. Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls. Int J Mol Sci. 2018.

[3.]Fabian P, Gaël V, Alexandre G, et al. Scikit-Learn: machine learning in python. J Mach Learn Res. 2011.

[4.]Long NP, Park S, Anh NH, et al. High-Throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer. Int J Mol Sci. 2019.

[5.] Bénard F, Barkun AN, Martel M, et al. Systematic review of colorectal cancer screening guidelines for average-risk adults: summarizing the current global recommendations. World J Gastroenterol. 2018

[6.] American Cancer Society. (2021). Colorectal Cancer Facts & Figures 2020 2022. Retrieved from https://www.cancer.org/

[7.] Harrell, F. E. (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer.

[8.] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[9.] Seaborn Development Team. (2021). Seaborn: statistical data visualization. Retrieved from https://seaborn.pydata.org/ 27

[10.] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

[11.] McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56).

[12.] National Cancer Institute. (2021). Cancer Data Access System (CDAS). Retrieved from https://cdas.cancer.gov/

[13.]Python Software Foundation. (2021). Python Language Reference, version 3.8. Retrieved from https://www.python.org/

[14.] Sklearn Developers. (2021). Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/

[15.] The Pandas Development Team. (2021). pandas: Powerful data structures for data analysis, version 1.3. Retrieved from https://pandas.pydata.org/