Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Score Based Colorectal Cancer Risk Assessment: A Comprehensive Machine Learning Approach for Heterogeneous Data

Bharathi M P[1*], Dr. Samitha Khaiyum[2], Dr. Veena R[3], Dr. Shivakumar Swamy S[4]
[1*]Research Scholar, VTU-RC, Dayananda Sagar College of Engineering, Bangalore
[2]Professor and Head, Department of MCA - VTU, Dayananda Sagar College of Engineering, Bangalore [3]Consultant Pathologist and Head of Histopathology and Digital Pathology, HCG Cancer Hospital, Bangalore [4]Sr. Consultant Radiologist, Department of Radiology, HCG Cancer Hospital, Bangalore

ARTICLE INFO

Received: 18 Dec 2024

Revised: 17 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

The role of technology in modern healthcare is increasingly critical and transformative. Machine learning has transformed cancer diagnosis and treatment, showing remarkable success in colorectal cancer (CRC) management. Their potential in personalized care is redefining the future of medical practices. Early detection of colon cancer and polyps is crucial to reducing CRC-related mortality and morbidity. However, selecting the most effective early screening method remains a challenge. This prospective study proposes a simple, efficient, and reliable scoring system to assess CRC risk levels (low, medium, high, very high). Heterogeneous parameters such as age, gender, tumor stage, tumor grade, and CEA levels are integrated into the scoring. CNN model is implemented for prediction and pretrained VGG16 model used for obtaining tumour stage value using CT scan images during the first phase. A comprehensive dataset combining image-based and clinical features was created. In the second phase, a random forest model was applied to assess collective risk factors. The developed model aims to assist clinicians in diagnosis, treatment planning, and patient monitoring. Additionally, model can also be used for disease prognosis using a biomarker CEA (Carcinoembryonic Antigen). The random forest algorithm obtained 96% accuracy for the dataset with the heterogeneous parameters.

KEYWORDS: Healthcare, colorectal cancer, deep learning, machine learning, random forest, linear regression, neural network, heterogeneous data

I. INTRODUCTION

Cancer poses a major public health challenge and places a substantial economic burden on healthcare systems worldwide. Colorectal cancer (CRC) is a widely occurring form of cancer globally, standing as the third most diagnosed cancer in men and the second in women across the world [1].

Identifying colon cancer poses challenges because of its potential occurrence across various segments of gastrointestinal tract—the large intestine, small intestine, and rectum. The tumour's propensity to manifest in diverse locations within this complex system complicates early detection and diagnosis strategies. Often its symptoms are difficult to identify in the early stages [2]. This makes it more important for individuals to learn about prevention, management, treatment plans and prognosis.

Initially various machine learning algorithms, specifically deep learning algorithms, were presented to assist medical experts to predict and classify these cancer cells. Image analysis is the most frequently used modality in which deep learning technique is used to discover the intensity to which the cancerous cells have spread to the neighbouring organs [3].

A. PARAMETERS USED IN PROPOSED STUDY

Colorectal cancer is usually detected through routine screening examinations. Proposed research focused on multi-omics data or heterogeneous data to achieve a better understanding of cancer prediction and cancer progression by assessing the risk level.

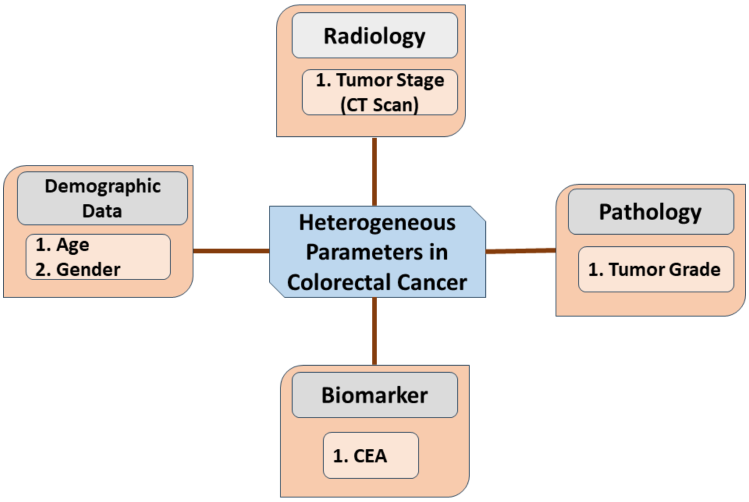


Figure 1: Heterogeneous data referred to in COLON CANCER

The parameters used in the proposed study are depicted in above figure 1. Medical images are one of the important tools to provide proper assistance to medical experts. Various imaging techniques are used to find important features or insights for diagnosis. In [4], the researchers have provided a detailed study about the image modalities. They proved that images of different forms in healthcare are the important prediction factor of colorectal cancer. Also, it is proved that it cannot decide the risk level or the survival time of patients. Various parameters are considered for monitoring the patient's health condition and help with treatment planning.

II. LITERATURE SURVEY

Cancer disease is a significant public health challenge with profound economic implications for healthcare systems globally. According to a recent review, the global cancer burden grew to 19.3 million new cases and resulted in 10 million deaths in the last year [5]. Also, various studies have proven that image data and clinical data when combined into an integrated approach, becomes an essential method for staging and for further prognosis [6]. Various studies were carried out using different image modalities like CT scan images, MRI, PET scan and analysis of histopathology images for the prediction, classification, and treatment plans of colon cancer. A study [7] explores the strengths and limitations of diagnostic tools such as
CT/MRI, endoscopes, genetics, and pathological assessments. It highlights the significant advancements facilitated by deep learning techniques in these areas. The main application of AI in therapeutic recommendations for colon cancer hold promising potential, offering notable advancements in clinical and translational oncology.

As sufficient progress in the advancements in AI and machine learning has been brought in this digital era, the various multi-modal learning is applied in healthcare problems to extract relevant features from high- dimensional data, data distributions without prior assumption. Nevertheless, much of the existing literature considered a single aspect of combining image analysis with tabular data such as genomics or clinical datasets [8]. They also expressed here that integration of all these data is also necessary which can assist clinical experts in better patient care. We have conducted a survey which questionnaire-based [9] in which various radiologists, pathologists and clinical experts expressed their opinion about the challenges in analyzing the cancer stages which require information like age, tumor size from image data, bio marker, and other clinical data.

Many studies have utilized CNN architecture for image processing [10,11]. The A study in 2020 [12] compared the performance of seven supervised machine learning algorithms in classifying colorectal cancer patients into low, medium, or high-risk categories based on age and family history. The study found that Artificial Neural Networks (ANN) achieved the highest accuracy, with a rate of 75%. They also demonstrated results with family history and without family history. This study uses dataset from the National Health Interview Survey (NHIS), in which people of different ages were interviewed on past and current health status. The researchers' findings underscore the significance of considering multiple parameters in cancer treatment and monitoring. They suggest that factors such as age and family history play crucial roles in determining cancer risk levels, highlighting the importance of incorporating these variables into cancer management strategies. Numerous risk assessment models for colorectal cancer (CRC) have been developed, emphasizing diverse non-biomarker factors [13]. These models incorporate a wide array of demographics, lifestyle choices, and clinical variables to forecast an individual's likelihood of developing the colorectal cancer (CRC). By integrating information such as age, family history, diet, physical activity, and other pertinent factors, these models offer valuable insights into an individual's susceptibility to CRC without using biomarker or image data. A study devised a novel approach using blood counts [14], age, and gender to detect individuals at higher risk of colon cancer via decision trees and cross-validation techniques. It underscored age and gender as pivotal factors in evaluating CRC risk. Among various machine learning models neural network, linear regression and random forest considered as efficient one. Random forest, a fast and easy-to-use algorithm, utilizes CART (Classification and Regression Trees) procedures for classification and regression tasks. It offers a novel approach to modelling by measuring variable importance through permutation, enhancing classification accuracy [15]. A recent systematic review on machine learning models for colorectal cancer risk prediction identified 11 distinct models specifically created to forecast the risk of developing CRC in asymptomatic populations [16]. Among the identified models, three relied solely on genetic markers, while four integrated laboratory and demographic data. One model considered lifestyle and family history, another focused-on comorbidities and medication history. Additionally, one model combined demographics,

lifestyle, and medical history, another model incorporated genetic markers, lifestyle factors, and family history. A study [17] uses a machine learning algorithm (MeScore) interprets CBC reports to identify individuals with a 10 to 20 times higher risk of occult CRC, aiding in targeted screening colonoscopy recommendations. This analysis involved reviewing demographics and complete blood count (CBC) test results extracted from electronic medical records of the Military Health System (MHS). The authors calculated scores for assessing risk for each patient, which was then converted into some values.

II. PROPOSED RESEARCH

A. MOTIVATION AND PRELIMINARIES OF THE WORK

Conducted a questionnaire-based survey to the professionals in the medical field (doctors, radiologists, pathologist.), encompassing diverse healthcare organizations, diagnostic centers, research laboratories, and hospitals [9]. The aim was to gain insights into their perspectives on the role of technology and the challenges associated with diagnosis, treatment, and the usage of different formats of data derived from various reports. More than 85% of respondents conveyed that the diagnosis of medical conditions depends on various factors. We refer to them as heterogeneous data. Especially in the context of cancer like disease, medical professionals such as doctors, radiologists, and pathologists emphasize the complexity involved in determining the stage of the disease and assessing its spread to different organs. Based on discussions with medical experts, it's noted that upon a cancer diagnosis, crucial information such as medication dosage, treatment strategies, the necessity of chemotherapy, and an estimation of survival time can be gleaned by analysing diverse test reports.

B. OBJECTIVE OF THE PROPOSED RESEARCH

The proposed study makes a significant contribution to the field of medicine by enhancing the evaluation of colorectal cancer risk levels. The objective of the research is to assist medical professionals by integrating the heterogeneous parameters for colorectal cancer risk assessment. Study implemented random forest and linear regression and support vector machine models. The study also includes a comparison between both the models.

C. DATASET DESCRIPTION

The dataset utilized for this study is provided by HCG (Health Care Global Enterprises), a renowned cancer treatment center and hospital in India. The dataset comprises of heterogeneous parameters such as age and gender, CT scan images, along with tumor grade, CEA biomarker values obtained from 600 patients. Initially, CT scan images of colorectal cancer from 600 patients were subjected to analysis using a CNN model, which provided a unified insight into tumor stage. In our study, this output from the images is regarded as a standardized homogeneous parameter. Additionally, we collected heterogeneous parameters such as patient's age, gender, and tumor grade from pathologists at HCG. The dataset underwent necessary preprocessing steps to ensure suitability for the proposed machine learning model. Approval for the study was obtained from both the radiologist and pathologist of HCG hospital.

D. METHODOLOGY

The below diagram depicts the steps in the proposed framework. The various steps in proposed study follows the inputs from HCG doctors. As per their opinion, colorectal cancer risk level is different in different cases based on age, gender, habits and other health related issues. As the different patient cases are being explained by the doctors, the study received qualitative inputs for building the risk assessment model. For example, how people with different age, gender, health related issues, or/and patients with tumour removal surgery were considered for monitoring to plan the treatment. A patient with tumor stage-III aged below 40 years may have less risk compared to a person above 60 years with stage-III. Also, the impact of tumor grade plays a major role in understanding the risk level and treatment plans.

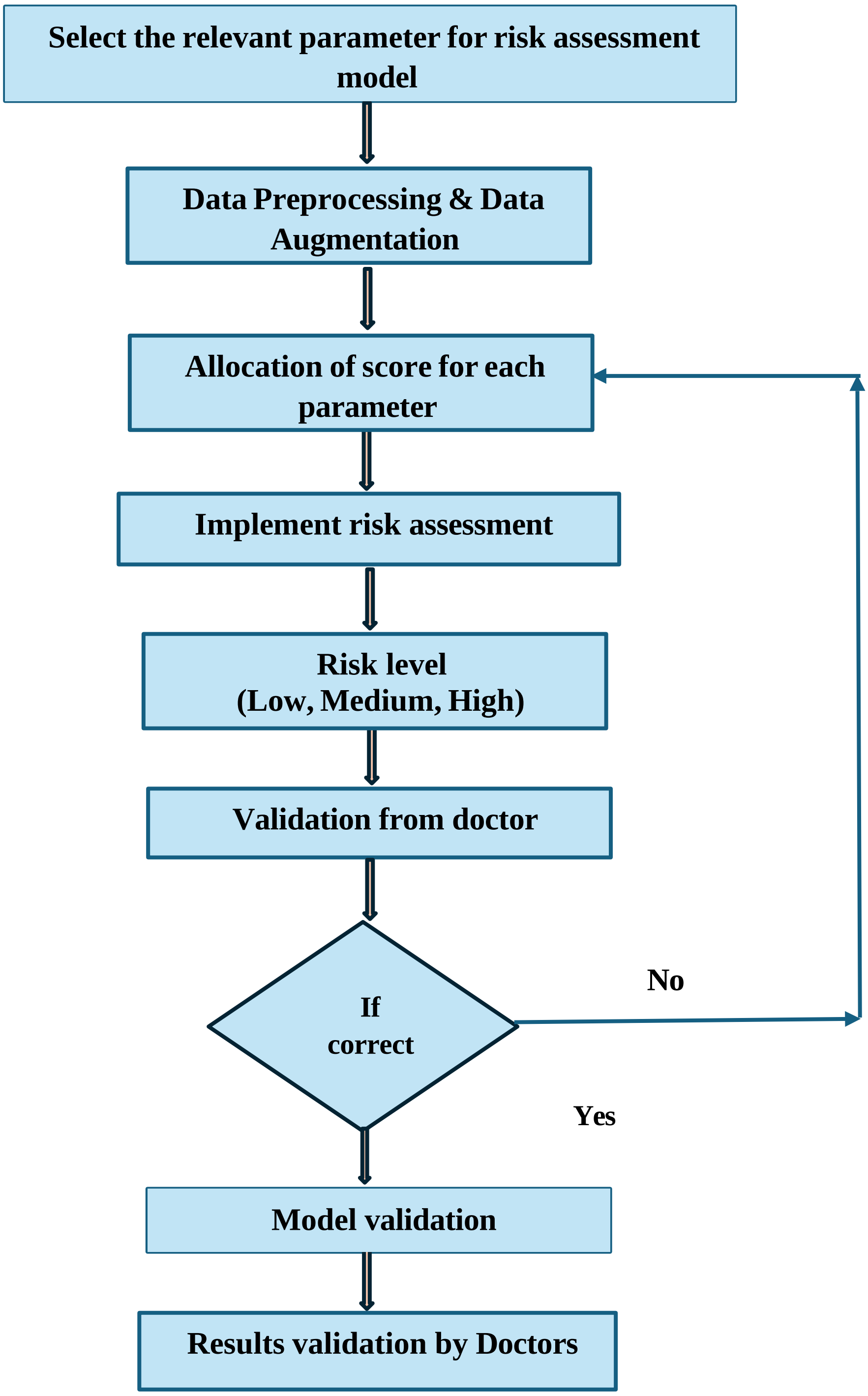

Select the relevant parameter for risk assessment model
Data Preprocessing & Data Augmentation
Allocation of score for each parameter
Implement risk assessment
Risk level (Low, Medium, High)
Validation from doctor
If correct
No
Yes
Model validation
Results validation by Doctors

Figure 2: Steps in the proposed study

The first phase of the study utilizes deep learning architectures CNN and VGG16 models for colorectal cancer prediction and stratification respectively [18, 19]. The proposed study tried to demonstrate the impact of heterogeneous parameters in colorectal cancer diagnosis, treatment, and monitoring. By understanding the impact, the individual score has been assigned to each parameter

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

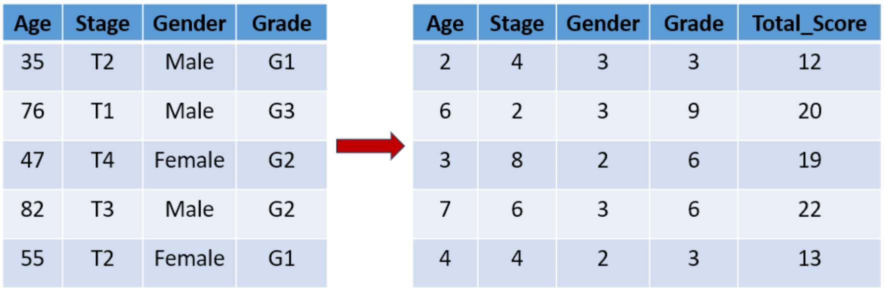
in the study. As the different patient cases are being explained by the doctors, the study received qualitative inputs for building the risk assessment model. Based on the qualitative inputs given by the doctors, they allocated some weightages (scores) for each parameter. This score allocation is to serve the requirement of machine learning model to predict the risk level. The pathologist of the hospital described the tumour grades along with its impact through different cases. Even a few complex cases with family history, biomarker details, patient with multiple issues were also given for understanding the impact of parameters. As a result, based on data availability, doctors suggested using age, gender, tumour stage, tumour grade and CEA biomarker data for the study.

III. DATA PREPROCESSING

A. Data Preparation

Data preparation ensures that machine learning algorithms can effectively process and extract insights from datasets [20]. This process involves preparing the data for the algorithm in a required format, addressing missing or invalid data, which can delay or reduce the algorithm's performance. As the study focuses on heterogeneous parameters collected in different formats from hospital. After collecting all the values were recorded in an excel file. A customized python function has been employed to generate scores for qualitative values into quantitative values.

Table 1: Generated Score values for each parameter and records



Age	Stage	Gender	Grade
35	T2	Male	G1
76	T1	Male	G3
47	T4	Female	G2
82	T3	Male	G2
55	T2	Female	G1

Age	Stage	Gender	Grade	Total_Score
2	4	3	3	12
6	2	3	9	20
3	8	2	6	19
7	6	3	6	22
4	4	2	3	13

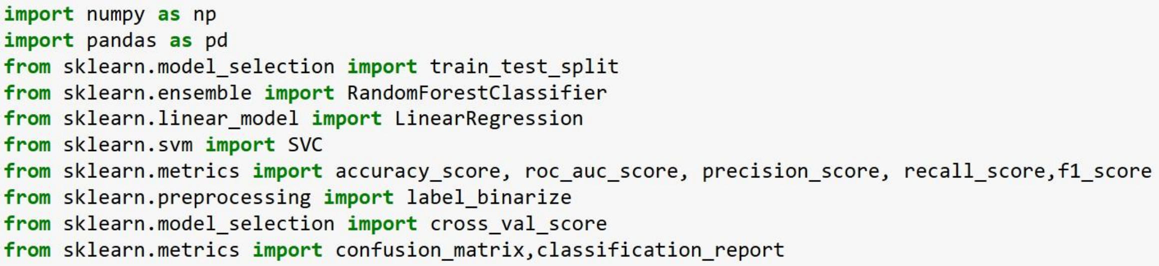
Out of 600 patient records, few were discarded, valid details were missing. Few patient details were empty values as they were not recorded from hospital records.

- Data preprocessing for machine learning involves handling missing values, encoding categorical variables, scaling features, and addressing outliers. Categorical variables are typically encoded as numerical values using techniques like one-hot encoding. Patient gender columns are categorical. These values are encoded with male as 0 and female as 1.
- The dataset is cleaned by removing any missing values (MVs) and replacing them with their mean values. Using fillna(value=-1) for missing values in stage and fillna(value=-0) for missing values in grade columns.
- CEA value was missing for maximum patient records as it is used as prognostic factor. Additional score generated for CEA for the value with 5 for more than 10 value of CEA.
- Data transformation is performed after cleansing the data to remove all missing NaN and null values. The Z-score method is then used to transform the data. This scaling process standardizes the features to values between 0 and 1, accommodating the features' varying magnitudes and types (categorical or non-categorical)

B. Data Augmentation

The heterogeneous data from the hospital were total of 600 which are details of colorectal cancer patients. Few records were discarded as the CT scan image was not found. After filtering accurate records were 514. To increase the size of the dataset, the study employed SMOTE augmentation technique.

Synthetic Minority Over-sampling Technique (SMOTE) is a data augmentation method used to address class imbalance by generating synthetic samples for the minority class. It works by interpolating existing minority class instances to create new, realistic data points. SMOTE helps improve model performance by ensuring balanced class distribution, reducing bias toward majority classes. In this study, SMOTE was applied to generate additional 400 samples based on Age, Gender, Stage, Grade, and Total Score, enhancing dataset diversity for more robust risk assessment in colorectal cancer. Several packages shown in Figure 3 are installed to execute algorithms used in the study's second phase. These libraries offer pre-built modules and functions that streamline the process of data manipulation, model building, evaluation, and deployment.



```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, roc_auc_score, precision_score, recall_score,f1_score
from sklearn.preprocessing import label_binarize
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix,classification_report
```

Figure 3: Packages installed for the model building

III. MODEL IMPLIMENTATION

The below table 2 describes each parameter and the score allocation. Male patients are assigned a higher score of 3, while female patients are assigned a score of 2 due to the increased risk factor associated with colorectal cancer in males. Scores ranging from 1 to 7 were allocated to different age groups: 20-30, 31-40, 41-50, 51-60, 61-70, 71-80, and those above 80 years respectively, reflecting the varying risks associated with age in colon cancer diagnosis. Also, tumour grade values are assigned with some quantitative values assigned by pathologists. The risk levels identified for the proposed model are low, medium, high and very high. The first three risk levels are based on the age, gender, tumour stage and tumour grade. Very high-risk level is for patients with additional biomarker CEA value incorporated with other parameters. This step is mainly utilized by doctors to understand disease prognosis.

Table 2: Scoreboard for the Heterogeneous Parameters Used

Parameter	Type	Score (c)	Risk Level (P)
Age			
20>=30	**Number**	1	
31>=40		2	
41>=50		3	

51>=60		4	• Low • Medium • High • Very High
61>=70		5	
71>=80		6	
Above 80		7	
Gender			
Male – 0	Categorical Value (Number)	Male – 3	
Female - 1		Female – 2	
Tumor Stage			
T1	Number	2	
T2		4	
T3		6	
T4		8	
Tumor Grade			
G1	Number	9	
G2		6	
G3		9	
CEA (Carcinoembryonic Antigen)			
	Float	Normal Range: 0-2.9 ng/mL	

The initial step data-splitting approach ensures that the model generalizes well to new, unseen data by learning from diverse subsets of the dataset.

The dataset containing 514 records was split into 70% training (360 records), 15% validation (77 records), and 15% test (77 records) to ensure balanced model training, hyperparameter tuning, and final evaluation. The training set was used to train the Random Forest model, the validation set helped in optimizing hyperparameters and preventing overfitting, and the test set provided an unbiased evaluation of model performance.

Random forest is the main algorithm used in the study because it can handle non-linear relationships and complex interactions between multiple parameters.

```
random forest model = RandomForestClassifier (n_estimators=50, criterion='gini', max_depth=None,
min_samples_split=2, min_samples_leaf=1, max_features='auto', bootstrap=True, random_state=1)
```

Algorithm efficiently handles the risk assessment by leveraging 50 decision trees (n_estimators=50) and the gini criterion minimizes impurity for optimal splits. The first step of training produced 93% test accuracy.

Further to enhance the performance of random forest model, Hyperparameters are selected based on experimentation and cross-validation to find the optimal values for a given dataset. In the second experiment, the grid search method is used to select the best combination of parameters for model training.

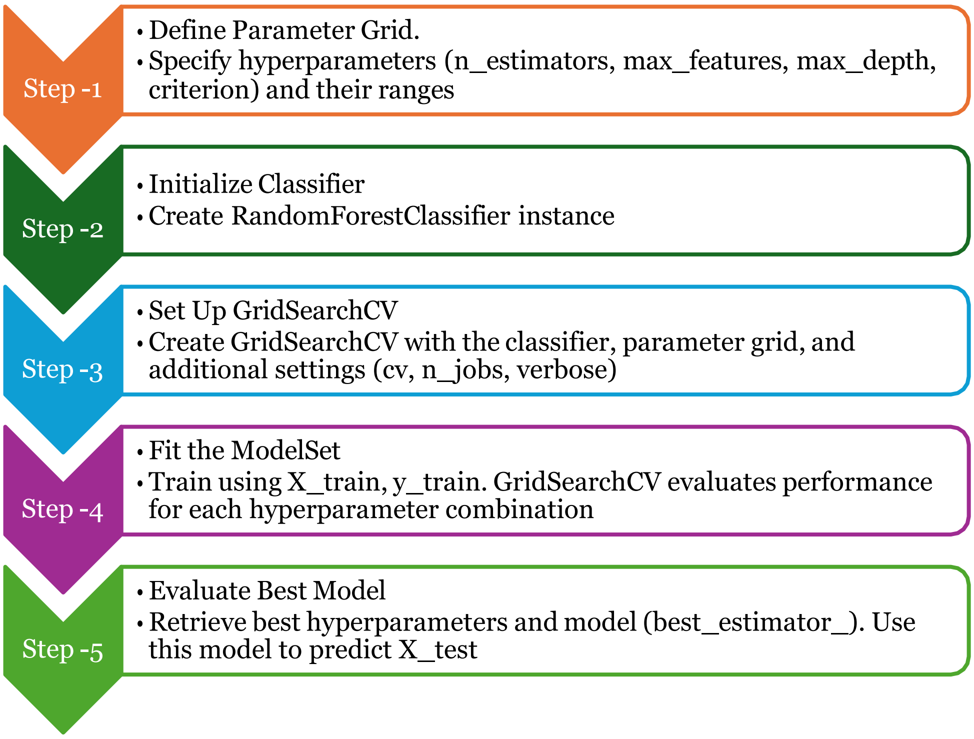
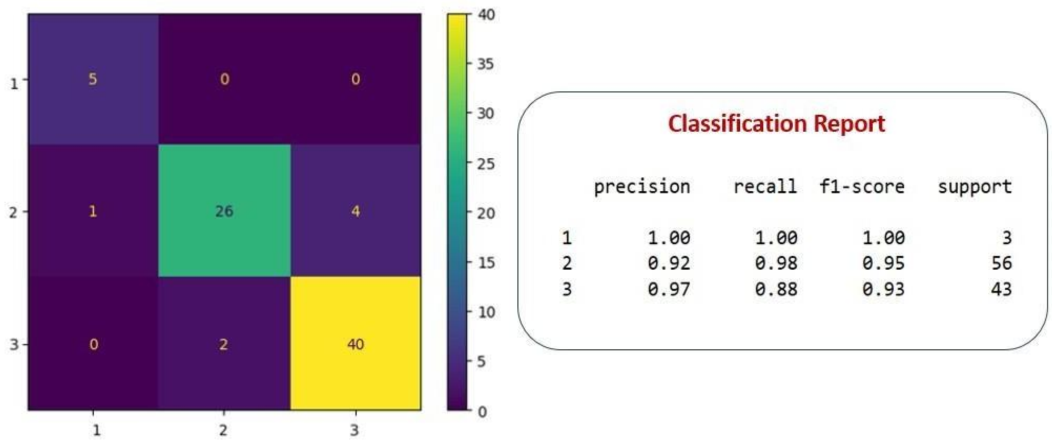

Step -1
• Define Parameter Grid.
• Specify hyperparameters (n_estimators, max_features, max_depth, criterion) and their ranges
Step -2
• Initialize Classifier
• Create RandomForestClassifier instance
Step -3
• Set Up GridSearchCV
• Create GridSearchCV with the classifier, parameter grid, and additional settings (cv, n_jobs, verbose)
Step -4
• Fit the ModelSet
• Train using X_train, y_train. GridSearchCV evaluates performance for each hyperparameter combination
Step -5
• Evaluate Best Model
• Retrieve best hyperparameters and model (best_estimator_). Use this model to predict X_test

Figure 4: Steps in hyperparameter tuning using GridSearch CV

The above figure 4 depicts the steps followed in a grid search technique. After hyperparameter tuning, the model obtained an accuracy of 96%. Cross-validation technique is used for model evaluation. Evaluation of the generalizability of models is crucial, especially for medical diagnostic models. To achieve this, a two-step method was adopted for model evaluation. Firstly, a standard 5-fold and 10-fold cross-validation (CV) technique was employed to ensure unbiased learning and establish a reliable model. In 5-fold, it divides the dataset into five equally sized folds, training the model on four folds and testing the remaining fold iteratively.

B. RESULTS & DISCUSSIONS

The research outcome establishes a meaningful relationship between the heterogeneous parameters with integrated score, shedding light on potential risk assessment indicators for cancer treatment and prognosis. A previous study [20] demonstrated risk assessment using gender, age and CBC values. We implemented random forest for risk assessment. To rigorously evaluate the model's performance, we employed additional data solely for testing purposes.

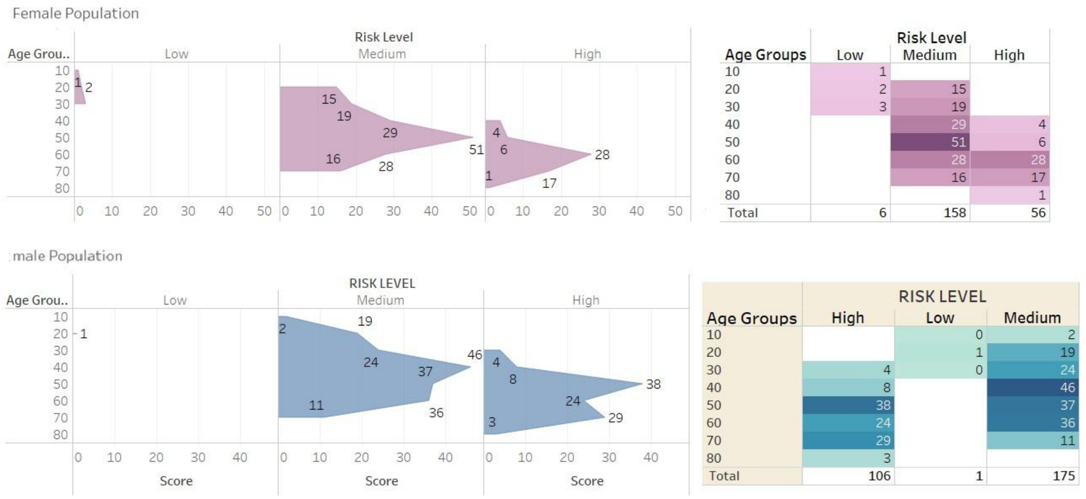


```
        precision    recall  f1-score   support

1          1.00       1.00      1.00         3
2          0.92       0.98      0.95        56
3          0.97       0.88      0.93        43
```

Figure 3: Result from Random Forest

In the proposed model, each parameter influencing risk has an assigned score based on its impact and three defined ranges for low, medium, and high-risk levels. The low and high-risk levels are determined by the parameter's lower and upper bounds, while the medium-risk level is based on the midpoint of these bounds. As a parameter is incorporated, the model compares the patient's score against these ranges during each iteration. The model processes all relevant features, continuously updating the risk score. Ultimately, the model predicts the patient's overall risk level by aggregating the individual scores and comparing them to the established risk ranges for a comprehensive assessment.



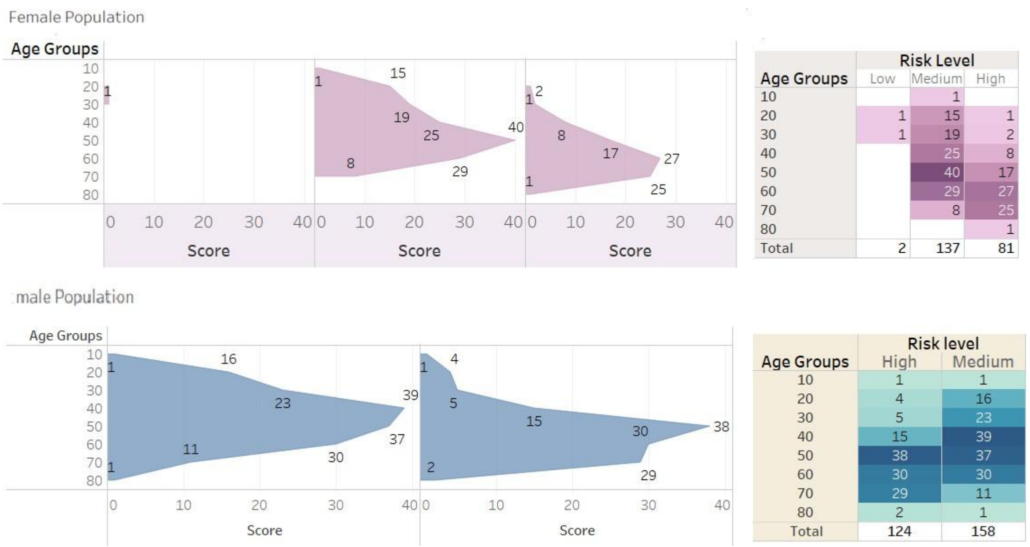
Age Groups	Low	Risk Level Medium	High
10	1		
20	2	15	
30	3	19	
40		29	4
50		51	6
60		28	28
70		16	17
80			1
Total	6	158	56

Age Groups	RISK LEVEL High	Low	Medium
10		0	2
20		1	19
30	4	0	24
40	8		46
50	38		37
60	24		36
70	29		11
80	3		
Total	106	1	175

Figure 4: Risk Level Prediction from Stage, Age, and Gender

The above graph depicts the predictions of risk level when the model was trained by integrating the parameters of tumor stage, age, and gender of the patients. It represents the score versus risk level for each age group and gender. In the female population, there are more patients with medium-risk levels than those with high-risk levels. The count of low-risk cases results from the impact of age and the tumor stage. Similar observations from the male populations show a count of one in a low-risk level because of the age parameter.

Research Article



Female Population

Age Groups	Risk Level		
	Low	Medium	High
10		1	
20	1	15	1
30	1	19	2
40		25	8
50		40	17
60		29	27
70		8	25
80			1
Total	2	137	81

male Population

Age Groups	Risk level	
	High	Medium
10	1	1
20	4	16
30	5	23
40	15	39
50	38	37
60	30	30
70	29	11
80	2	1
Total	124	158

Figure 5: Risk Level Prediction from Stage, Age, Gender and Grade

The above graph depicts the change in the number of predictions at each risk level as the parameter tumor grade has been integrated. The model is trained with updated parameters and the relevant score. The result shows the reduced count of low and medium risk levels in female populations, and the low-risk level count is zero in male populations. This indicates that patients with low risk when added with a score of tumor grade, resulted in a high-risk level due to the increased score.

The proposed model also used for Colorectal Cancer prognosis using a biomarker CEA (Carcinoembryonic antigen). Elevated CEA levels (above 5.0 ng/mL) can indicate the presence of colon cancer. The patient records with CEA greater than 10 will add a score of 5 to the combined score value. Ihe very high-risk level assessed by proposed model based on the combined score greater than or equal to 24. The model validated at different stages by integrating heterogeneous parameters. The varied risk levels are the result from model based on the combined score. Various studies introduced efficient tools and techniques for colorectal cancer prediction. Most of the studies concentrated on image analysis. With the increase in cancer cases oncologists are trying to find technological assistance to analyze the risk level of cancer patients in order to increase the survival time.

A study [21] carried out a comparison between nine supervised and unsupervised machine learning algorithms for colorectal cancer prediction and risk analysis based on a list of important dietary data. A study [22] demonstrated logistic and random forest models for colorectal cancer risk assessment using heterogeneous data like age, gender and hemoglobin concentrations. This study obtained the result in three screening rounds with varied values of two preceding hemoglobin concentrations. Both the algorithms produced 95% accuracy for 219257 patients. This study has proved the efficiency of Random Forest algorithm in colorectal cancer risk assessment. The proposed study introduced a risk assessment technique including image and CEA which plays a major role in medical diagnosis and treatment.

V. CONCLUSION

The study introduces a score-based hybrid machine learning model for colorectal cancer risk assessment, leveraging heterogeneous parameters such as age, gender, tumor stage, and grade. This method not only stratifies individuals into low, medium, and high-risk categories but also holds promise for prognosis when combined with the carcinoembryonic antigen (CEA) biomarker. By incorporating CEA values, we can further refine risk prediction, identifying individuals at a very high risk of colorectal cancer. It is proved that random forest algorithm obtained a better result compared to other previous works. This study concludes that cancer predictive multimodal approaches are capable of better patient care and patient's data management. In this regard, integration of heterogeneous data plays a major role. Also, the use of deep learning techniques is the best choice for getting good results using CT image data.

Acknowledgments

We gratefully acknowledge HCG (Healthcare Global Enterprises Ltd) providing the dataset with the necessary support and guidance. Also thankful to Dr. Shivakumar Swamy S, Sr. Consultant Radiologist, Dr. Veena R, Consultant Pathologist and Head of Histopathology and Digital Pathology, HCG Cancer Hospital for the valuable inputs and guidance.

References

[1] Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Cancer J. Clin. 68, 394–424 (2021).

[2] Koncina, E., Haan, S., Rauh, S. & Letellier, E. Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. Cancers 12, 2–319 (2020).

[3] Sharma, A.K.; Nandal, A.; Dhaka, A.; Dixit, R. Medical Image Classification Techniques and Analysis Using Deep Learning Networks: A Review. In Health Informatics: A Computational Perspective in Healthcare; Springer: Singapore, 2021

[4] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017

[5] S.Krithika, S. Sivananthan,S.Serrangevi, S.Surendhar "Detecting the Colorectal Cancer by Deep Learning, TIJER - International Research Journal, ISSN 2349-9249, April 2023, volume 10, issue 4.

[6] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," CA: A Cancer Journal for Clinicians, vol. 66, no. 1, pp. 7–30, 2019

[7] Chaoran Yu, Ernest Johann Helwig, The role of AI technology in prediction, diagnosis, and treatment of colorectal cancer", Artificial Intelligence Review (2022) 55:323–343, Springer, Published online: 4 July 2021

[8] Xiao Tan, Andrew T. Su, Hamideh Hajiabadi, Minh Tran, and Quan Nguyen. 2021. Applying Machine Learning for Integration of Multi-Modal Genomics Data and Imaging Data to Quantify Heterogeneity in Tumour Tissues. In Artificial Neural Networks, Hugh Cartwright (Ed.). Springer US, New York, NY, 209– 228

[9] Bharathi Ramesh, Dr. Samitha Khaiyum "Impact of Heterogeneous Data Integration in Healthcare: A Questionnaire based Survey" scopus, DOI:10.5373/JARDCS/V12SP7/20202421 pages: 2805-2814

[10] Rajani S, Veena M.N, "Ayurvedic Plants Identification based on Machine Learning and Deep Learning Technologies", 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 978-1-6654-5635-7/22/$31.00 ©2022 IEEE.

[11] Rajani S, Veena M.N, "Medicinal plants segmentation using Thresholding and Edge based Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,
Volume-8,Issue-6S4, April 2019

[12] Bradley J, Nartowt, Gregory R Hart, Wazir Muhammad, Ying Liang, Gigi F Stark Jun Deng "Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification" Frontiers in Big data

|www.fronters.org, volume-3, Article 6 Doi: 10.3389/fdata.2020.00006.

[13] Usher-Smith, J. A., Walter, F. M., Emery, J. D., Win, A. K., and Griffin, S. J. (2016). Risk prediction models for colorectal cancer: a systematic review. Cancer Prev. Res. 9, 13–26. doi: 10.1158/1940-6207.CAPR-15- 0274

[14] Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc. 2016; 23:879–890. doi:10.1093/jamia/ocv195.

[15] Kinar Y, Akiva P, Choman E, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS ONE. 2017;12:e0171759. doi: 10.1371/journal.pone.0171759

[16] Qarmiche, N., Chrifi Alaoui, M., Otmani, N., El Fakir, S., Tachfouti, N., Bourkhime, H., Omari, M., El Rhazi, K. and Chaoui, N. Machine Learning for Colorectal Cancer Risk Prediction: Systematic Review.
DOI: 10.5220/0010738100003101

[17] Kinar, Y.; Akiva, P.; Choman, E.; Kariv, R.; Shalev, V.; Levin, B.; Narod, S.A.; Goshen, R. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS ONE 2017, 12, e0171759

[18] Bharathi Ramesh and Dr. Samitha Khaiyum, Dr. Shivakumar Swamy S "Predictive Analysis of Colorectal Cancer via CT scans using Convolutional Neural Networks", International journal of Membrane Science and technology, 2023, vol:10,No:3, pp3378- 3387

[19] Bharathi Ramesh and Dr. Samitha Khaiyum, Dr. Shivakumar Swamy S "Stratifying Colorectal Cancer stages through CT Scan images using Convolutional Neural Networks", International journal on recent and innovation trends in computing and communication, ISSN: 2321-8169, Volume:11, Issue:11 2023

[20] Emmanuel MT (2021) A survey on missing data in machine learning. Journal of Big Data, no. 140. https://doi.org/10.1186/s40537-021-00516-9

[21] Abdul Rahman, H., Ottom, M.A. & Dinov, I.D. Machine learning-based colorectal cancer prediction using global dietary data. BMC Cancer 23, 144 (2023). https://doi.org/10.1186/s12885-023-10587-x

[22] Duco T. Mülder, Rosita van den Puttelaar, Reinier G.S. Meester, James F. O'Mahony, Iris Lansdorp- Vogelaar, "Development and validation of colorectal cancer risk prediction tools: A comparison of models", International Journal of Medical Informatics, Volume 178, 2023, 105194, ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2023.105194.