Novel Biomarkers for Colorectal Cancer Prediction

Mohamed Ashraf[1], M.M. El-Gayar[2]*, Eman Eldaydamony[3]
Department of Information Technology, Faculty of Computers and Information Mansoura University, Mansoura 35516, Egypt[1, 2, 3]
Department of Computer Science, Arab East Colleges, Riyadh, Saudi Arabia[2]
Department of Information Technology, Faculty of Computers and Information Mansoura University, Mansoura 35516, Egypt[3]

***Abstract*—Most researchers work on solving the important issue of identifying biomarkers linked to a certain disease, like cancer, in order to assist in the disease's diagnosis and treatment. Several research have recently suggested several methods for identifying genes linked to disease. A handful of these methods were created specifically for CRC gene prediction, though. This research presents a novel prediction technique to determine new biomarkers related to CRC that can assist in the diagnosing process. First, we preprocessed four Microarray datasets (GSE4107, GSE8671, GSE9348 and GSE32323) using RMA (Robust Multi-Array Average) method to remove local artifacts and normalize the values. Second, we used the chi-squared test for feature selection to identify some significant features from datasets. Finally, the features were fed to XGBoost (eXtreme Gradient Boosting) to diagnose various test scenarios. The proposed model achieves a high mean accuracy rate and low standard deviation. When compared to other systems, the experiment findings show promise. The predicted biomarkers are validated through a review of the literature.**

***Keywords*—Colorectal cancer (CRC); microarray; biomarkers; gene expression omnibus; feature selection; chi-squared test; XGBoost**

I. Introduction

Particularly with the current state of treatment, cancer is a complex disease [1]. Colorectal cancer is one of the leading causes of cancer-related deaths worldwide. Colorectal cancer is a type of cancer that affects the latter part of the gut intestine within the digestive system [2]. Colorectal cancer is a cancer that occurs in the last 15 centimeters of the colon that meets part of the rectal region, and these two types of cancer together call colorectal cancer [3]. In most cases, colorectal cancer begins as a small mass of non-cancer cells called adenomatous polyp, after a period of time the solids that have formed become cancer masses present in the colon. These masses may be small and accompanied by very few symptoms [4]. There may be no symptoms of colorectal cancer, especially in its early stages. Common symptoms of colorectal cancer:

- Constipation or diarrhea.
- The feeling that the intestines are not completely emptied.
- Blood in the feces.
- Frequent pain caused by gases, bloating or feeling full.
- Weight loss for no reason.
- Persistent fatigue.
- Vomiting and nausea.

Colorectal Cancer is the third most common cancer in men and the second most common in women worldwide [2]. Despite recent advances in surgical and multimodal therapies, the overall survival of advanced CRC patients remains very low. Periodic tests can significantly reduce and prevent the incidence of this disease. If colorectal cancer is detected early enough, it can mostly be treated [5].

Determining the patient's cancer kind and its biomarkers [6] is one of the biggest problems facing researchers. The biomarker is the most crucial element in cancer research since it facilitates therapy and reduces the cost and time required for diagnosis. Thus, one of the researchers' most important tasks is to identify the most relevant biomarker. Because cancer involves dynamic genetic alterations, researchers have worked hard to investigate how to diagnose and forecast the disease.

Next-generation sequencing (NGS) and microarray data [7] are two important sources of potentially helpful molecular patterns. The abundance of gene expression data [8] facilitates the discovery of disease class and cancer-related biomarkers. Microarray technology can be used to study the entire genome, proteome, and transcriptome in various cells and tissues. The ability to analyze vast amounts of data quickly is one of the advantages of microarray technology.

In the context of microarray technology, feature selection techniques [9, 10] fall into three categories: filter, embedding, and wrapper techniques. The filter approach, which is independent of the predictor and does not require the classifier, assesses and ranks the genes in relation to the class label. Gene interactions and correlations are not taken into consideration. In contrast, the wrapper approach depends on whether features are added or removed while evaluating the subgroup features using classification methods. When compared to alternative classification algorithms, the filter approach is faster but less accurate than the wrapper method. In contrast, the wrapper approach is slower computationally but yields more accurate results than the filter method. The embedded approach involves building a particular classifier but uses search methods for subsets of optimal characteristics. The merging of the filter and wrapper techniques is intended to reduce the wrapper method's computational complexity issue.

This proposed model uses the chi-squared test [11] for feature selection to identify some significant features from the datasets for colorectal cancer classification. Then, the features were fed to XGBoost (eXtreme Gradient Boosting) [12] to diagnose different test cases and identify new biomarkers related to colorectal cancer that can help in the early diagnosis

of that disease. Finally, the results were validated using SHAP algorithm to explain the importance of the proposed biomarkers in identifying colorectal cancer and expose the significant biomarkers.

The main contribution of the proposed system is the identification of new biomarkers related to CRC that can assist in the early diagnosis of the disease.

For the reader's convenience, seven sections make up the remainder of this work. The relevant literature, existing flaws, and how the suggested approach gets around them are covered in Section II. In Section III, the materials and methods are explained. Section IV introduces the datasets, assessment measures, and results. The experimental results and novel biomarkers are discussed in Section V. By applying the SHAP algorithm to validate the results, Section VI illustrates the significance of novel biomarkers. Lastly, Section VII concludes and summarizes the plans for future work. Table I lists the used abbreviations in this paper for easy reference of the reader.

TABLE I. THE USED ABBREVIATIONS

CRC	Colorectal cancer	IG	Information Gain
RMA	Robust Multi-array Average	ANN	Artificial Neural Network
XGBoost	eXtreme Gradient Boosting	P-SVM	Penalized Support Vector Machine
SHAP	Shapley Additive Explanations	XAI	Explainable Artificial Intelligence
NGS	Next Generation Sequencing	DEGs	Differentially Expressed Genes
TCGA	The Cancer Genome Atlas	CAD	Colorectal Adenocarcinoma
GEO	Gene Expression Omnibus	GA	Genatic Algorithm
CatBoost	Categorical Boosting	MCC	Moffitt Cancer Center
PCA	Principal component analysis	X^2	Chi-squared
DT	Decision Tree	NB	Naive Bayes
TP	True positive	RF	Random Forest
FP	False positive	AB	Adaboost
LR	Logistic Regression	GBDT	gradient boosting decision tree
LGBM	Light gradient-boosting machine	GB	gradient boosting
SVM	Support Vector Machine	LDA	Linear Discriminant Analysis
FN	False negative	TN	True negative
SE	Sensitivity	SPC	specificity
TPR	True positive rate	FPR	False negative rate
LASSO	Least Absolute Shrinkage and Selection Operator	AUPR	Area under precision-recall
NCBI	National Center for Biotechnology Information	AUC	Area Under the Curve
GSNFS	Gene Sub-Network-based Feature Selection	ACC	Accuracy
mRMR	minimum Redundancy Maximum Relevance	SD	Standard Deviation

II. RELATED WORK

In the field of biology, predicting genes linked to a disease is regarded as an active research topic. Genes linked to these diseases have been found and predicted by many researchers; some of these studies have focused specifically on colorectal cancer. Table II shows a summary of the current studies. For example, Ahmadieh-Yazdi et al. [13] presented an approach to predict disease-related biomarkers using the TCGA dataset [14, 15] and GEO dataset [8]. First, they used LASSO and P-SVM methods as feature selection to identify the most relevant DEGs. DEGs frequently chosen by these techniques were chosen for additional analysis. Second, they applied the Multilayer Perceptron technique in conjunction with the Artificial Neural Network (ANN) method to evaluate the effectiveness of each method's gene selection in distinguishing primary samples from metastatic malignancies.

Maurya et al. [16] proposed a novel framework to identify genes associated with CRC using the TCGA dataset and GEO dataset. They used Boruta as a features selection method to select significant genes. The most relevant genes were then utilized to create an ML-based prognostic classification model with Random Forest classifier.

Li, S et al. [17] proposed a supervised learning framework based on deep learning (DeepCSD) to identify cancer subtypes. They designed a minimalist feed-forward neural network to capture the distinct molecular features in different cancer subtypes.

Al-Rajab et al. [18] proposed a framework that provides a two-step multi-filter hybrid model to select features for the classification of colon cancer was proposed. A mixture of the Information Gain (IG) and Genetic algorithms (GA) is used in the initial stage of feature selection. In order to lower the amount of genes and acquire more relevant genes, the second stage employs the minimum Redundancy Maximum Relevance (mRMR) filter approach. Utilizing appropriate machine learning techniques for further analyzing the data is the final step. With the suggested framework model, it was discovered that the SVM, Decision Tree, Naïve Bayes, and K-Nearest Neighbor classifiers provided accurate and promising results.

According to Shuwen et al. [19], the early detection of liver metastasis is crucial for improving the prognosis of patients with colorectal adenocarcinoma (CAD), and the utilization of a single biomarker in conjunction with a classification model has greatly enhanced the ability to predict the metastasis of various cancer types. There aren't many reports on CAD, though. Thus, the purpose of this study was to determine the most appropriate classification model for CAD patients with liver metastases and, using that model, investigate the gene's metastatic process. For the purpose of identifying the CAD system with liver metastases, the CatBoost model—which was constructed using 33 feature genes—exhibited the greatest classification performance.

TABLE II. A REVIEW OF A FEW RECENT STUDIES FOR COMPARISON

Study	Year	Analysis	Methodology	Dataset
Ahmadieh-Yazdi et al. [13]	2023	Identifying the disease-related biomarkers based on LASSO and P-SVM	LASSO, P-SVM, ANN	TCGA, GEO
Maurya et al. [16]	2023	Identifying genes linked to colorectal cancer using gene expression	Boruta, RF	TCGA, GEO
Li, S et al. [17]	2022	Predicting cancer subtypes based on DeepCSD	NN, Deep learning	GEO
Al-Rajab et al. [18]	2021	Identifying cancer cells based on two-stage approach	A hybrid of IG and GA, DT, KNN, SVM and NB	GEO
Shuwen et al. [19]	2020	Predicting biomarkers from classifier for liver metastasis of colorectal adenocarcinomas using machine learning models	LR, NN, SVM, RF, GBDT, Catboost.	GEO, MCC
Kozuevanich et al. [20]	2020	Predicting genes related to disease based on GSNFS	Correlation-based, Information gain, GSNFS	GEO
Yanke et al. [21]	2020	Identifying the disease-related biomarkers based on Random Forest and Deep learning	Random walk restart algorithm, RF, Deep learning	TCGA, GEO
Ram et al. [22]	2017	Identifying the disease-related biomarkers based on Random Forest	RF	GEO, Array Express databases

Kozuevanich et al. [20] proposed that the combination of Gene Sub-Network-based Feature Selection (GSNFS) and feature selection is very promising to identify biomarkers associated with CRC because it requires fewer subnetworks to build a classifier and provides a performance comparable to that of a full data set classifier.

Yanke et al. [21] proposed a model that used complex networks, machine learning methods and deep learning technology to look for probable genes linked to colorectal cancer in the following seven types of colorectal cancer data: LUAD, LUSC, UCEC, BRCA, COAD, HNSC, and KIRC. The signed random walk restart algorithm was employed in this suggested model to extract features. The random forest is the machine learning technique employed in this model as a colorectal cancer classifier. This model also makes use of deep learning technologies to look for putative colorectal genes and offer a novel method of diagnosing colorectal cancer.

Ram et al. [22] proposed a model that used the Random Forest algorithm to rank and select the genes needed to properly diagnose and treat cancer. While preserving its accuracy for prediction, the Random Forest method produced extremely tiny gene groups.

In conclusion, previous studies did not explore all the biomarkers associated with colorectal cancer, nor did they achieve the highest percentages in the disease classification process. To overcome the several limitations of the current studies, as mentioned above, we designed a novel prediction system that primarily identifies new biomarkers related to CRC based on microarray dataset that can assist in the diagnosing process and that is considered a very important advantage compared to previous studies. First, we preprocessed four Microarray datasets (GSE4107 [23], GSE8671 [24], GSE9348 [25] and GSE32323 [26]) using RMA (Robust Multi-Array Average) method27 to remove local artifacts and normalize the values. Second, we used the chi-squared test [11] for feature selection to identify some significant features from datasets. Finally, the features were fed to XGBoost (eXtreme Gradient Boosting) [12] to diagnose various test scenarios. The proposed model achieves a high mean accuracy rate and low standard deviation. When compared to other systems, the experiment findings show promise. The predicted biomarkers are validated through a review of the literature.

III. MATERIALS AND METHODS

The main contribution of the proposed system is the identification of new biomarkers related to CRC that can assist in the early diagnosis of the disease. Using the RMA (Robust Multi-Array Average) approach [27], we first preprocessed four Microarray datasets (GSE4107 [23], GSE8671 [24], GSE9348 [25], and GSE32323 [26]) to eliminate local artifacts and normalize the data. Secondly, we selected certain important features from datasets using the chi-squared test [11] for feature selection. XGBoost (eXtreme Gradient Boosting) [12] was then fed the features in order to diagnose several test scenarios. The model that has been suggested has a low standard deviation and a high mean accuracy rate. The experiment results show promise when compared to other systems. Based on a review of the literature, the expected biomarkers are confirmed.

The proposed prediction system is depicted in Fig. 1 with a unique four-step architecture. First, the preprocessing step comprises four steps: background correction, normalization, summarization and log2 transformation for removing local artifacts and noise. Second, the most relevant features are selected using the Chi-squared test [11] as a feature selection. Third, these proposed features are fed to the XGBoost algorithm [12] for identifying various test cases. In conclusion, we assess the proposed system's performance using five metrics, revealing encouraging outcomes in comparison to other methods. The following subsections contain more information on the suggested prediction system.

A. Preprocessing

In the preprocessing step, we prepared and improved the original dataset to feed it to the feature selection algorithm in order to improve our proposed system and obtain accurate results. The microarray datasets were preprocessed using RMA (Robust Multi-Array Average) approach [27] which involves three main steps: Background correction, Normalization, and Summarization. First, the microarray datasets were background corrected to remove local artifacts and noise so measurements aren't so affected by neighboring measurements. Second, we normalized these datasets to remove array effects so measurements from different arrays are comparable. Then, we

summarized the normalized datasets for combining probe intensities across arrays so the final measurement represents gene expression level. Finally, the log2 transformation was applied to the gene expression levels as almost all preprocessing methods return expression levels on log2 scale which is the approximately right scale.

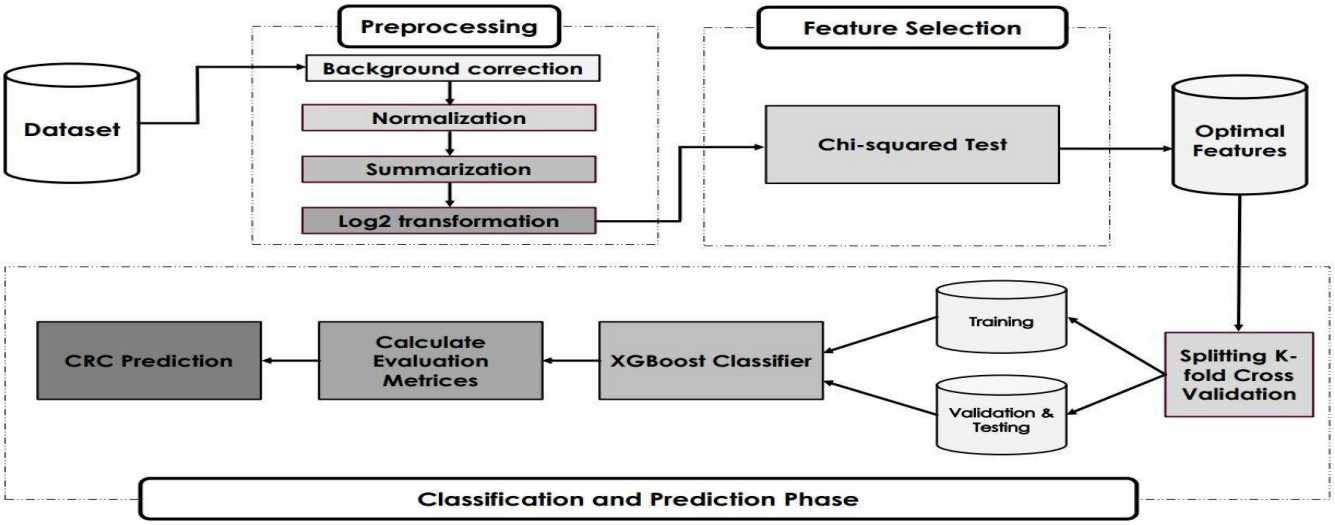

Fig. 1. The proposed prediction system.

B. Feature Selection

Feature selection stage seeks to reduce the set of features in microarray datasets by generating a new set of features from the preexisting ones. The majority of the data in the original set should be able to be summarized by these selected ones. This stage aids in lowering computation time, complexity, and model overfitting. Thus, we made an effort to narrow down the microarray datasets to the most important features. Suppose the classification algorithm receives incorrect or irrelevant features as input. In that scenario, it is unable to produce an accurate prediction because the machine learning model's performance depends heavily on the quality of the input data. As a result, we attempted to select the most notable features of most of the datasets. The selected features help us to correctly classify key genes associated with CRC. This is a crucial step in our proposed predictive model. If the features are not chosen correctly, the classification may be invalidated, impacting the predictive model.

There are three types of feature selection [9] in the microarray technology context: filter, wrapper and embedded. In the filter method, the genes are sorted and evaluated according to the class label. Correlation and gene-to-gene interactions are not considered, and it is independent of the predictor without using a learning algorithm (classifier). In contrast, the wrapper method depends on using the learning algorithm (classification algorithm) to add or remove features in order to evaluate the subset features.

This section outlines the feature selection method that produced encouraging results when compared to state-of-the-art methods: The Chi-squared test [11].

- Chi-squared test:

When large sample sizes are available, a statistical hypothesis test called a chi-squared test (also known as a chi-square [11] or X^2 test) is employed in the study of contingency tables. In layman's words, the main purpose of this test is to determine whether the two categorical variables (i.e., the two contingency table dimensions) have no effect on the test statistic. When the test statistic is chi-squared distributed under the null hypothesis, the test is considered valid. To ascertain whether there is a statistically significant discrepancy between the expected and observed frequencies in one or more categories of a contingency table, one can apply Pearson's chi-squared test. A Fisher's exact test is substituted for contingency tables with smaller sample sizes. The observations are categorized into classes that are mutually exclusive in the standard applications of this test. An X^2 frequency distribution is followed by the test statistic generated from the observations if the null hypothesis, which states that there are no differences between the classes in the population, is true. Assessing the observed frequencies' likelihood under the null hypothesis is the aim of the test. When the observations are independent, test statistics happen to follow an X^2 distribution. To verify if a pair of random variables are independent based on observations of each other, X^2 tests are also available. Datasets with categorical features are subjected to the Chi-square test. The desired number of features is determined with the optimal Chi-square scores by calculating the Chi-square between each feature and the target.

In 1900, Pearson published a paper on the X^2 test which is considered to be one of the foundations of modern statistics. In this paper, Pearson investigated a test of goodness of fit.

$$\sum_{i=1}^{k} p_i = 1 \quad (1)$$

$$\sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = n \quad (2)$$

Suppose that n observations in a random sample from a population are classifed into k mutually exclusive classes with respective observed numbers of observations x_i (for i = 1,2,. . . ,k), and a null hypothesis gives the probability p_i that an observation falls into the *i*th class. So we have the expected numbers $m_i = np_i$ for all i.

$$x^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - n \quad (3)$$

Pearson proposed that, under the circumstance of the null hypothesis being correct, as n → ∞ the limiting distribution of the quantity given above is the X^2 distribution as shown in Eq. (3). Pearson dealt first with the case in which the expected numbers *mi* are large enough known numbers in all cells assuming every observation *xi* may be taken as normally distributed, and reached the result that, in the limit as n becomes large, X^2 follows the X^2 distribution with *k* - 1 degrees of freedom.

To reduce the computational time and complexity for the classifier, the Chi-squared test is used to reduce the set of features and concurrently keep key features. We use Chi-squared test as feature selection to choose n features from the training model that have the highest score, based on threshold=5. After applying the Chi-squared test, we obtain 130 features instead of 54,675 features selected by Chi-squared test, as shown in Table III. As seen in Algorithm 1, we use the Chi-squared test to indicate the algorithm for the proposed feature selection after the preprocessing steps of RMA approach.

TABLE III. CHI-SQUARED FEATURE SELECTION AND THEIR NUMBERS

Dataset	# Of features before feature selection	# Of features after feature selection
GSE4107	54675	130
GSE8671	54675	130
GSE9348	54675	130
GSE32323	54675	130

Algorithm 1: The proposed preprocessing and feature Selection

Data: List of genes L_0

Result: The matrix of the most significant features F

Remove local artifacts and noise from the genes in L_0 and update it
Normalize the genes in L_0;
Represent probe intensities across arrays in L_0 as gene expression levels L_1
Transform the gene expression levels L_1 to log2 scale L_2
Initialize microarray matrix of features w
Apply the Chi-squared test for feature selection
Select the features with higher score than threshold=5 in matrix F

C. Classification

The selected features of the Chi-squared test are fed to the XGBoost classifer [12]. This classifier predicts the important genes linked to colorectal cancer (CRC) and diagnoses several test cases. Our tests demonstrate the superiority of XGBoost over state-of-the-art machine learning methods for both regression and classification tasks. XGBoost classifier is a machine learning algorithm that is applied for tabular and structured data. XGBoost is a fast and efficient implementation of gradient-boosted decision trees. XGBoost stands for extreme gradient boosting which implies that it is a large-scale machine learning technique with numerous components. XGBoost is an ensemble learning method. Relying just on the output of a single machine learning model may not always be adequate. Combining the prediction ability of several learners can be done methodically with ensemble learning [28-31]. All of the output from several models is aggregated into a single model as a result. The ensemble's models, often referred to as base learners, may come from different learning algorithms or from the same learning algorithms. The most popular ensemble learning models include bagging, boosting, stack generalization, and expert mixtures. However, two widely regarded methods for ensemble learning [28] are bagging and boosting. While these two methods can be used to a variety of statistical models, decision trees have been the most common application for them. The following is a summary of the main equations included in the XGBoost classifier:

1) Objective function: A regularized objective serves as the objective function in XGBoost and must be optimized throughout the training phase. It is composed of two terms: a regularization term that penalizes complexity to prevent overfitting, and a loss term that calculates the difference between the actual and predicted values.

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} (f_k) \quad (4)$$

Where:

- $l(y_i, \hat{y}_i)$ is the loss function which measures the difference between the actual target value $\hat{y}_i$.
- (f_k) is the regularization term which penalizes the complexity of the model, where represents the k-th tree.
- θ represents the model parameters.

2) Gradient and hessian tree ensemble prediction: Gradient boosting is used in XGBoost to optimize the objective function. The first and second-order derivatives of the objective function with respect to the anticipated scores must be calculated in order to carry out gradient boosting.
Where:

$$g_i = \frac{\partial}{\partial \hat{y}_i} l(y_i, \hat{y}_i) \quad (5)$$

- g_i is the first-order derivative (gradient) of the loss function.

$$h_i = \frac{\partial^2}{\partial {\hat{y}_i}^2} l(y_i, \hat{y}_i) \quad (6)$$

- h_i is the second-order derivative (Hessian) of the loss function.

3) Tree ensemble prediction: The XGBoost model's final prediction is the weighted sum of the predictions made by several trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (7)$$

Where:

- $f_k(x_i)$ is the prediction from the k-th tree.

The XGBoost algorithm [12], which creates decision trees iteratively to minimize the above-defined objective function, is based on these equations. In numerous machine learning

applications, XGBoost provides state-of-the-art performance by optimizing the objective function. As seen in Algorithm 2, we illustrate the algorithm for the suggested classification based on the XGBoost classifier.

Algorithm 2: The proposed classification with XGBoost

Data: $D_{Train} = (x_1, y_1), (x_2, y_2), \dots, (x_N, x_N)$ $D_{Test} = (x_1, y_1), (x_2, y_2), \dots, (x_N, x_N)$
Result: Training model, and Prediction results
Use D_{Train} to train the XGBoost
Initialize the model as $h_0(x)$
For m = 1 → M do
  For i = 1 → N do
    Compute the loss function: $F(y, h_{m-1}(x_i))$
    Compute the residuals: rh,i
  End
  Fit a regression tree m_{th} to the $r_{h,i}$ values to create the terminal regions "tree leaf nodes" $R_{m,j}$, j = 1, 2, … , J; where J is the number of leaf nodes in the tree.
  For j = 1 → J do
    Get the v_m, j;
  End
  Update the weak classifier $h_m(x)$
End
Get the fnal model *H*(*x*)
Use D_{Test} to evaluate the prediction model

For s = 1 → N do
  Process XGBoost prediction model
  Get the predicted label
End
Utilizing the real and expected labels as inputs, compute the evaluation metric

IV. Experimental Results

This part includes the description of the datasets, the specifications for the hardware and software, and evaluation metrics and results. In the results subsection, first, we used the chi-squared test [11] for feature selection to identify some significant features from microarray datasets [7]. Second, the features were fed to XGBoost (eXtreme Gradient Boosting) [12] to diagnose various test scenarios. The proposed model achieves a high mean accuracy rate and low standard deviation. Third, when compared to other state-of-the-art classification algorithms, the experiment findings show promise. Finally, we use two performance measures [32] to provide some figures and tables that support a desired idea.

A. Datasets Description

Gene Expression Omnibus (GEO) [8] database served as the primary, accessible, and all-inclusive source for the gene expression raw data used in this paper when microarray data was deposited. The current study used four microarray data sets, as shown in Table IV, as follows:

- GSE4107 [23]: 22 samples and 54675 genes were included in this dataset. The data were split into two groups, each comprising 10 normal and 12 tumor samples.
- GSE8671 [24]: 64 samples and 54675 genes were included in this dataset. The data were split into two groups, each comprising 32 normal and 32 tumor samples.
- GSE9348 [25]: 82 samples and 54675 genes were included in this dataset. The data were split into two groups, each comprising 12 normal and 70 tumor samples.
- GSE32323 [26]: 34 samples and 54675 genes were included in this dataset. The data were split into two groups, each comprising 17 normal and 17 tumor samples.

TABLE IV. Description of the Datasets' Gene Expression

Dataset	Classification Type	# of Samples
GSE4107 [23]	normal	10
	tumor	12
GSE8671 [24]	normal	32
	tumor	32
GSE9348 [25]	normal	12
	tumor	70
GSE32323 [26]	normal	17
	tumor	17

B. Evaluation Metrics

This proposed work used five metrics [32] for measuring the performance of our proposed system, including accuracy (ACC), standard deviation (SD), precision, recall, and F1-Score, which are defined using Eq. (8) *to* Eq. (12).

$$ACC = \frac{TN+TP}{TN+FP+TP+FN} \quad (8)$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad (9)$$

- x_i being the result of the i-th measurement and $\bar{x}$ being the arithmetic mean of the n results considered.

$$Precision = \frac{TP}{FP+TP} \quad (10)$$

$$Recall = SEN = TPR = \frac{TP}{FN+TP} \quad (11)$$

$$F1 - score = \frac{TP}{TP+0.5(FN+FP)} \quad (12)$$

A true positive (TP) gene is one that has been accurately predicted to be a CRC gene, and this must be made clear. Reliability of genes correctly predicted as non-CRC genes is known as true negative (TN) rate. Erroneously predicted genes as colorectal cancer (CRC) genes are known as false positives (FP). False negative (FN) is also the rate of genes that are misclassified as not being CRC genes. The percentage of accurate results over all results based on TP and TN is known as the accuracy rate, or ACC. It evaluates how accurate the proposed system is. The precision can be defined as the ratio of correctly predicted results to the total number of wrong and accurate predictions, with "results" denoting the positive genes. The rate of correctly predicted results over all correctly

predicted results is known as the SEN, recall, or TPR, where "results" refers to the negative genes.

C. Results

All of the experimental results from this investigation, together with pertinent analysis, are provided in this subsection. The comparison of feature selection, the comparison of classification techniques, and the comparison with other prediction systems comprised the three parts of the experimental results.

1) Feature selection comparison: In the feature selection stage, we attempted to select the most notable features of most of the datasets. The selected features help us to correctly classify key genes associated with CRC. This is a crucial step in our proposed predictive model. If the features are not chosen correctly, the classification may be invalidated, impacting the predictive model. We employed the Chi-squared test to identify the important features from microarray datasets in order to construct our prediction system. The features from two state-of-the-art features (PCA and LASSO) are compared with the suggested features to validate them. When compared to other systems, the experiment findings show promise. We preformed the experiments based on microarray dataset using the XGBoost classifier with 5-fold and 10-fold cross-validation technique [33]. We evaluated the results using two performance metric [32]: **ACC** and **SD**.

Microarray dataset: Table V shows the performance comparison of the proposed features based on chi-squared test and features from state-of-the-art feature selection techniques: PCA and LASSO. For 5-fold cross-validation, the proposed features achieved the following: ACC equals 93.0% and SD equals 0.13. For 10-fold cross-validation, the proposed features achieved the following: ACC equals 91.8% and SD equals 0.14. The proposed features based on chi-squared test achieve a high mean accuracy and low standard deviation as shown in Table V.

TABLE V. THE PROPOSED FEATURES' EFFECTIVENESS WAS ASSESSED USING THE CHI-SQUARED TEST AND COMPARED TO TWO OTHER METHODS, NAMELY PCA AND LASSO WITH 5- AND 10-FOLD CROSS-VALIDATION USING THE MICROARRAY DATASET

Metric	K-fold	ACC (%)	SD
PCA	5 10	86.0 90.0	0.26 0.2
LASSO	5 10	82.0 80.0	0.15 0.22
Chi-squared test (proposed)	5 10	**93.0** **92.0**	**0.13** **0.14**

As shown in Fig. 2, the proposed features based on the chi-squared test achieve a high mean accuracy (ACC).

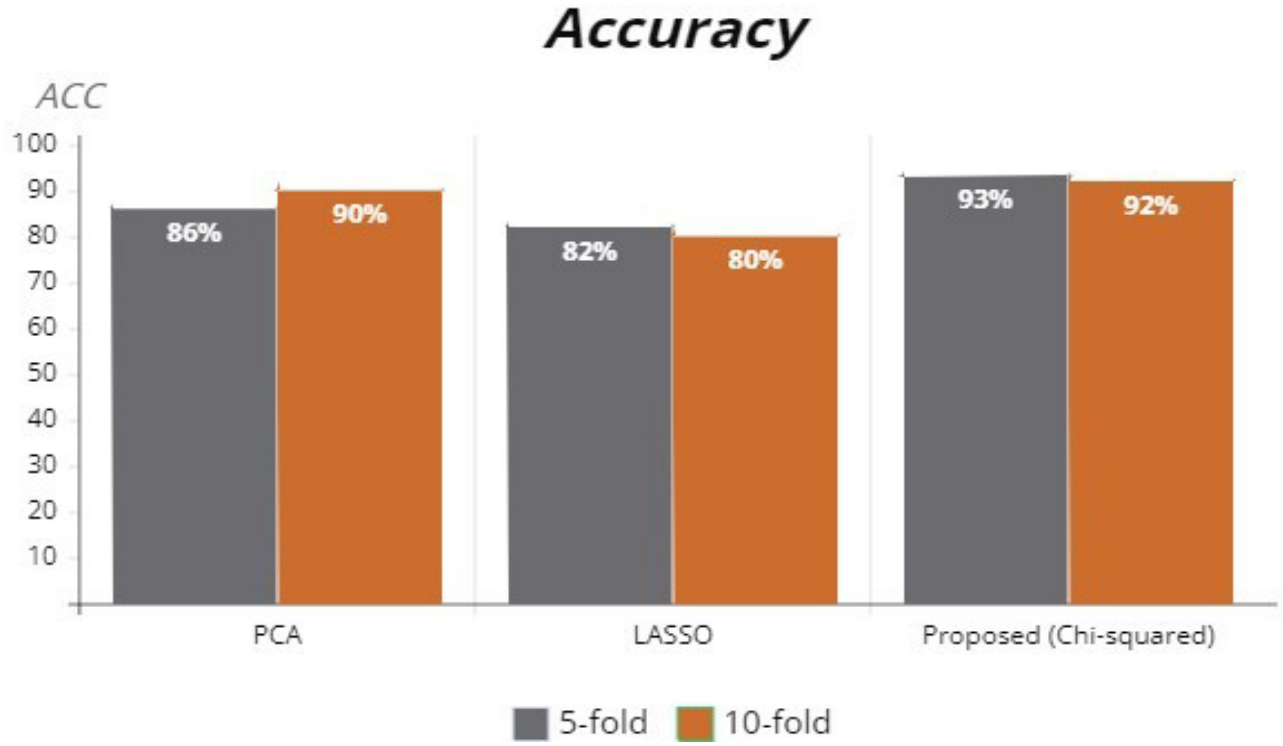

Fig. 2. The comparison between the proposed feature selection method and some state-of-the-art methods is based on accuracy.

2) Classification algorithm comparison: The selected features of the Chi-squared test are fed to the XGBoost classifier. This classifier predicts the important genes linked to colorectal cancer (CRC) and diagnoses several test cases. Our tests demonstrate the superiority of XGBoost over state-of-the-art machine learning methods for both regression and classification tasks. The results from three state-of-the-art classifiers (RF, LGBM, and LR) are compared with the XGBoost classifier to validate them. When compared to other systems, the experiment findings show promise. We

preformed the experiments based on four microarray datasets using the XGBoost classifier with 5-Fold and 10-fold cross-validation technique. We evaluated the results using five performance [32] metric: ACC, precision, recall, F1-Score and SD.

GSE4107 Microarray dataset: Table VI, shows the performance comparison of the proposed classifier based on XGBoost and classifiers from state-of-the-art techniques: LGBM, LR and RF. For 5-fold cross-validation, the proposed classifier achieved the following: ACC equals 93.0%, Precision equals 93.0%, Recall equals 100.0%, F1-score equals 96.0%, and Mean SD equals 0.085. For 10-fold cross-validation, the proposed classifier achieved the following: ACC equals 95.0%, Precision equals 90.3%, Recall equals 100.0%, F1-score equals 90.0%, and Mean SD equals 0.187. The proposed classifier based on XGBoost classifier achieved a high mean accuracy rate and low standard deviation as shown in Table VI.

As shown in Fig. 3, the proposed classifier based on XGBoost classifier achieved a high mean accuracy rate (ACC).

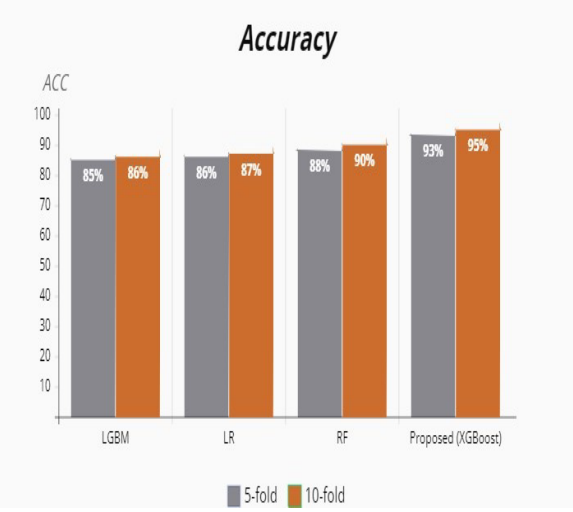

Fig. 3. The comparison between the proposed classifier and some state-of-the-art classifiers is based on accuracy using the GSE4107 dataset.

TABLE VI. USING 5-FOLD AND 10-FOLD CROSS-VALIDATION METHODS BASED ON THE 4107 DATASET, THE PROPOSED SYSTEM'S PERFORMANCE WAS EVALUATED IN COMPARISON TO STATE-OF-THE-ART CLASSIFIERS

Metric	K-fold	ACC (%)	Precision (%)	Recall (%)	F1-score (%)	SD
LGBM	5	85.4	87.2	89.5	82.4	0.180
	10	86.4	84.3	82.9	79.5	0.226
LR	5	86.2	89.6	86.3	87.4	0.155
	10	87.1	86.3	84.9	80.5	0.201
RF	5	88.9	90.2	94.5	91.4	0.120
	10	90.1	87.3	87.9	86.5	0.195
XGBoost	5	**93.0**	**93.0**	**100.0**	**96.0**	**0.085**
	10	**95.0**	**90.3**	**100.0**	**90.0**	**0.187**

GSE8671 Microarray dataset: Table VII, shows the performance comparison of the proposed classifier based on XGBoost and classifiers from state-of-the-art techniques: LGBM, LR and RF. For 5-fold cross-validation, the proposed classifier achieved the following: ACC equals 97.77%, Precision equals 96.0%, Recall equals 100.0%, F1-score equals 97.77%, and Mean SD equals 0.040. For 10-fold cross-validation, the proposed classifier achieved the following: ACC equals 98.0%, Precision equals 98.0%, Recall equals 100.0%, F1-score equals 98.88%, and Mean SD equals 0.037. The proposed classifier based on XGBoost classifier achieved a high mean accuracy rate and low standard deviation as shown in Table VII. As shown in Fig. 4, the proposed classifier based on XGBoost classifier achieved a high mean accuracy rate (ACC).

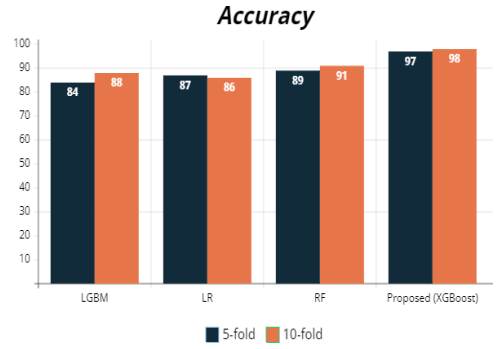

Fig. 4. The comparison between proposed classifier and some state-of-the-art classifiers based on Accuracy using GSE8671 dataset.

GSE9348 Microarray dataset: Table VIII, shows the performance comparison of the proposed classifier based on XGBoost and classifiers from state-of-the-art techniques: LGBM, LR and RF. For 5-fold cross-validation, the proposed classifier achieved the following: ACC equals 98.18%, Precision equals 100.0%, Recall equals 98.18%, F1-score equals 99.05%, and Mean SD equals 0.025. For 10-fold cross-validation, the proposed classifier achieved the following: ACC equals 91.0%, Precision equals 91.0%, Recall equals 100.0%, F1-score equals 95.05%, and Mean SD equals 0.057. The proposed classifier based on XGBoost classifier achieved a high mean accuracy rate and low standard deviation as shown in Table VIII. As shown in Fig. 5, the proposed classifier based on XGBoost classifier achieved a high mean accuracy rate (ACC).

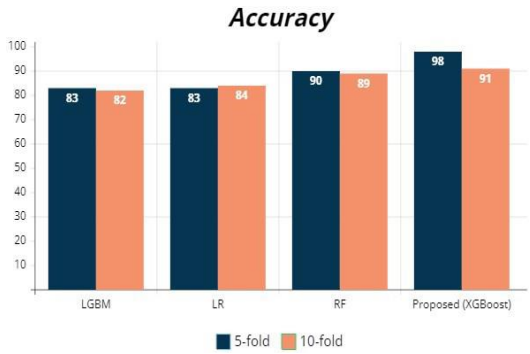

Fig. 5. The comparison between proposed classifier and some state-of-the-art classifiers based on accuracy using GSE9348 dataset.

TABLE VII. USING 5-FOLD AND 10-FOLD CROSS-VALIDATION METHODS BASED ON THE 8671 DATASET, THE PROPOSED SYSTEM'S PERFORMANCE WAS EVALUATED IN COMPARISON TO STATE-OF-THE-ART CLASSIFIERS.

Metric	K-fold	ACC (%)	Precision (%)	Recall (%)	F1-score (%)	SD
LGBM	5 10	84.2 88.5	85.4 85.6	85.5 81.4	84.1 76.4	0.190 0.236
LR	5 10	87.3 86.2	88.4 84.1	84.8 85.2	86.1 78.5	0.164 0.220
RF	5 10	89.4 91.4	91.1 89.5	93.4 88.4	90.2 88.2	0.130 0.188
XGBoost	5 10	**97.77** **98.0**	**96.0** **98.0**	**100.0** **100.0**	**97.77** **98.88**	**0.040** **0.037**

TABLE VIII. USING 5-FOLD AND 10-FOLD CROSS-VALIDATION METHODS BASED ON THE 9348 DATASET, THE PROPOSED SYSTEM'S PERFORMANCE WAS EVALUATED IN COMPARISON TO STATE-OF-THE-ART CLASSIFIERS.

Metric	K-fold	ACC (%)	Precision (%)	Recall (%)	F1-score (%)	SD
LGBM	5 10	83.8 82.1	84.1 81.2	84.2 80.2	81.2 76.8	0.240 0.311
LR	5 10	83.7 84.0	86.1 82.2	84.4 81.3	85.6 81.1	0.188 0.222
RF	5 10	90.1 89.5	92.3 89.3	95.4 88.9	92.7 87.8	0.116 0.187
XGBoost	5 10	**98.18** **91.0**	**100.0** **91.0**	**98.18** **100.0**	**99.05** **95.05**	**0.025** **0.057**

GSE32323 Microarray dataset: Table IX, shows the performance comparison of the proposed classifier based on XGBoost and classifiers from state-of-the-art techniques: LGBM, LR and RF. For 5-fold cross-validation, the proposed classifier achieved the following: ACC equals 96.0%, Precision equals 95.0%, Recall equals 100.0%, F1-score equals 97.14%, and Mean SD equals 0.059. For 10-fold cross-validation, the proposed classifier achieved the following: ACC equals 96.0%, Precision equals 100.0%, Recall equals 95.0%, F1-score equals 96.0%, and Mean SD equals 0.087. The proposed classifier based on XGBoost classifier achieved a high mean accuracy rate and low standard deviation as shown in Table IX.

TABLE IX. USING 5-FOLD AND 10-FOLD CROSS-VALIDATION METHODS BASED ON THE 9348 DATASET, THE PROPOSED SYSTEM'S PERFORMANCE WAS EVALUATED IN COMPARISON TO STATE-OF-THE-ART CLASSIFIERS.

Metric	K-fold	ACC (%)	Precision (%)	Recall (%)	F1-score (%)	SD
LGBM	5 10	89.6 88.8	89.4 85.4	88.7 86.6	86.6 81.4	0.160 0.216
LR	5 10	87.7 89.4	86.5 87.5	87.7 85.5	88.9 84.2	0.176 0.180
RF	5 10	92.8 91.5	92.5 91.5	95.5 93.9	96.5 94.5	0.105 0.125
XGBoost	5 10	**96.0** **96.0**	**95.0** **100.0**	**100.0** **95.0**	**97.14** **96.0**	**0.059** **0.087**

As shown in Fig. 6, the proposed classifer based on XGBoost classifer achieved a high mean accuracy rate (ACC).

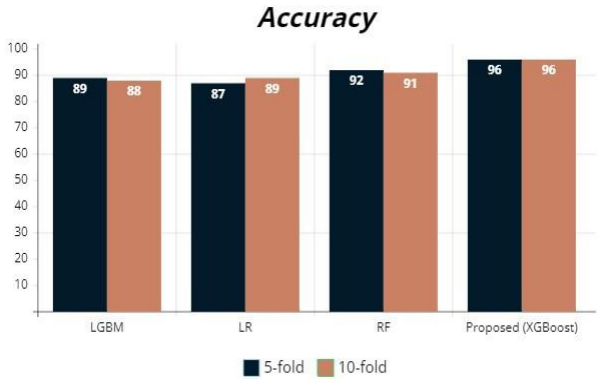

Fig. 6. The comparison between proposed classifier and some state-of-the-art classifiers based on Accuracy using GSE32323 dataset.

3) Comparison with other prediction systems: To verify how well the proposed system performs when using the XGBoost classification algorithm and the chi-squared feature selection technique. We evaluated how well the proposed system performed in comparison to state-of-the-art systems: Ahmadieh-Yazdi et al. [13], Maurya et al. [16], Li, S et al. [17], and Al-Rajab et al. [18].

The proposed system is compared to state-of-the-art systems in Table X. This comparison is based on feature selection, classification method, and performance evaluation using 10-fold cross-validation.

As shown in Fig. **7**, the proposed system achieved the highest mean accuracy rate (ACC), F1-score and Recall compared with state-of-the-art systems.

TABLE X. THE PROPOSED SYSTEM, BASED ON THE MICROARRAY DATASET, WAS COMPARED WITH THE PERFORMANCE METRICS, FEATURE SELECTION TECHNIQUES, AND CLASSIFICATION STRATEGIES EMPLOYED IN STATE-OF-THE-ART SYSTEMS

System	ACC (%)	F1-score (%)	Recall (%)	Classification Method	Feature Selection Method
Ahmadieh-Yazdi et al.[13]	90.0	80.9	84.6	ANN	LASSO
Maurya et al.[16]	99.0	80.2	83.6	RF	Boruta
Li, S et al.[17]	90.0	78.7	82.5	Deep learning	NN
Al-Rajab et al.[18]	DT=93.75,K-NN=93.75,NB=87.5,SVM=81.25	83.0	81.4	DT,K-NN,NB,SVM	a hybrid of IG and GA
Proposed System	**98.18**	**99.05**	**98.18**	**XGBoost**	**Chi-squared**

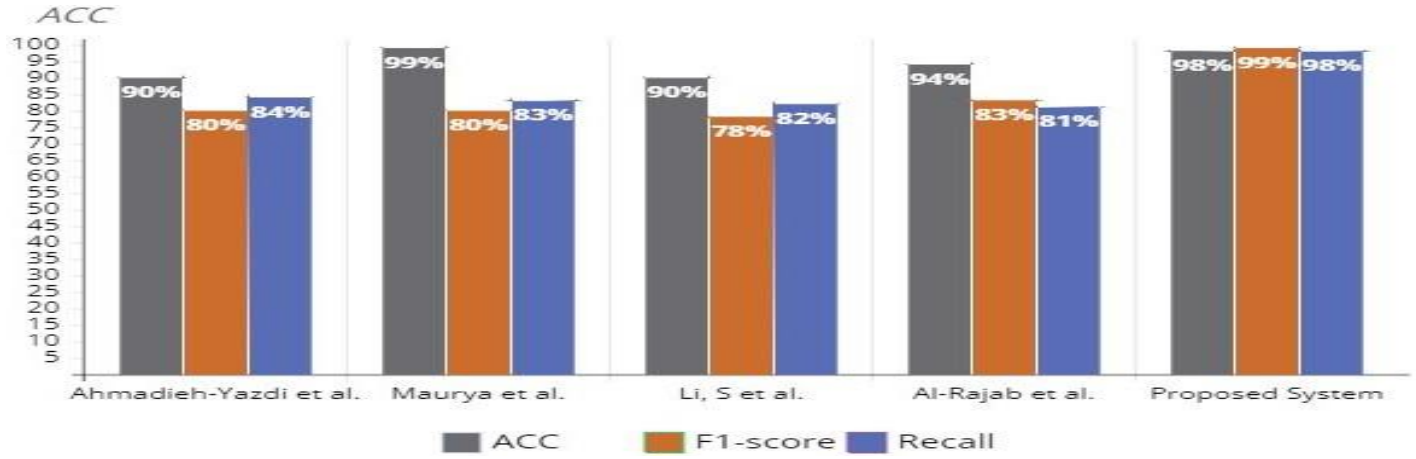

Fig. 7. The comparison between proposed system and some state-of-the-art systems.

V. DISCUSSION

Cancer is a complicated disease, especially in the treatment process so far. Colorectal cancer is one of the leading causes of cancer-related deaths worldwide. Colorectal Cancer is the third most common cancer in men and the second most common in women worldwide. Despite recent advances in surgical and multimodal therapies, the overall survival of advanced CRC patients remains very low. Periodic tests can significantly reduce and prevent the incidence of this disease. If colorectal cancer is detected early enough, it can mostly be treated. The biomarker is the most important component of cancer researches because it helps with the treatment process and saves cost and time in the diagnosis process. Therefore, determining the relevant biomarker is an essential task for the researchers. Researchers have made great efforts to explore an accurate diagnosis and prediction of cancer because it contains genetic changes that are dynamic. As a result, we developed our proposed prediction system to find key genes associated with colorectal cancer that can assist in the early diagnosis of the disease.

Using the RMA (Robust Multi-Array Average) approach, we first preprocessed four Microarray datasets (GSE4107, GSE8671, GSE9348, and GSE32323) to eliminate local artifacts and normalize the data. We employed 5-fold and 10-fold cross-validation methods to evaluate the proposed method. Secondly, we selected certain important features from datasets using the chi-squared test for feature selection. XGBoost (eXtreme Gradient Boosting) was then fed the features in order to diagnose several test scenarios. To verify how well the proposed system performs when using the XGBoost classification algorithm and the chi-squared feature selection technique. We evaluated how well the proposed system performed in comparison to state-of-the-art systems: Ahmadieh-Yazdi et al. [13], Maurya et al. [16], Li, S et al. [17], and Al-Rajab et al. [18]. The model that has been suggested has a low standard deviation and a high mean accuracy rate. The experiment results show promise when compared to other systems.

The results are evaluated using five performance metric: ACC, precision, recall, F1-Score and SD. For 5-fold cross-validation, the GSE4107 dataset achieved the following: ACC equals 93.0%, Precision equals 93.0%, Recall equals 100.0%, F1-score equals 96.0%, and Mean SD equals 0.085. For 10-fold cross-validation, the GSE8671 dataset achieved the following: ACC equals 98.0%, Precision equals 98.0%, Recall equals 100.0%, F1-score equals 98.88%, and Mean SD equals 0.037. For 5-fold cross-validation, the GSE9348 dataset achieved the following: ACC equals 98.18%, Precision equals 100.0%, Recall equals 98.18%, F1-score equals 99.05%, and Mean SD equals 0.025. For 5-fold cross-validation, the GSE32323 dataset achieved the following: ACC equals 96.0%, Precision equals 95.0%, Recall equals 100.0%, F1-score equals 97.14%, and Mean SD equals 0.059.

Lastly, the proposed model predicted novel CRC biomarkers that are not present in the databases [6] using the proposed prediction system. The literature review is used to verify these biomarkers. The Biomarkers that were extracted: VIP, CYR61, ADAMTS1, SLC51A, GREM1, PLN, MSI2, FOS, ADH1B, ETNK1, MEP1B, NR1H4, SYNP02, OGN, FOSB, UGT2A3, RGS1 and SERPINF1. We found out that

none of these genes had been linked to colorectal cancer (CRC) based on the literature review [6].

Table XI, show the smallest gene set for different datasets selected by XGBoost. These genes are known as colorectal cancer biomarkers and the function of each of these genes is presented in the mentioned table. The function and annotation of each selected gene were extracted from the NCBI database.

TABLE XI. THE LIST OF BIOMARKERS AND THEIR MAIN FUNCTIONS

Probe ID	Gene Symbol	Function
* "206577_*at*"	**VIP**	Encodes the vasoactive intestinal peptide (VIP) protein. VIP is a neuropeptide that has a wide range of physiological effects and is distributed widely throughout the body.
* "210764_*s_at*"	**CYR61**	Encodes a protein referred to as CCN1 (cysteine-rich protein 61),or cysteine-rich angiogenic inducer 61.
* "222162_*s_at*"	**ADAMTS1**	Encode the ADAMTS1 protein, an enzyme that is engaged in a number of biological functions, such as: Extracellular Matrix Remodeling, Angiogenesis Regulation, Cell Migration and Proliferation, and Tissue Homeostasis and Development.
* "228230_*at*"	**SLC51A**	Encodes the organic solute transporter alpha (OSTα) protein, which is a component of the heteromeric transporter complex that is involved in the transfer of bile acids.
* "218468_*s_at*"	**GREM1**	Encodes gremlin-1, a protein that belongs to the family of bone morphogenetic protein (BMP) antagonists known as DAN (differential screening-selected gene abnormal in neuroblastoma).
* "204939_*s_at*"	**PLN**	Phospholamban is essential for controlling the activity of the sarcoplasmic reticulum calcium ATPase (SERCA), a vital ion transporter.
* "1552364_*s_at*"	**MSI2**	Encodes the Musashi-2 (MSI2) protein, an RNA-binding protein belonging to the Musashi family.
* "209189_*at*"	**FOS**	Encodes the Fos protein, a member of the Fos transcription factor family.
* "209613_*s_at*"	**ADH1B**	Encodes the enzyme alcohol dehydrogenase 1B, which is essential to the liver's ethanol metabolism–the kind of alcohol present in alcoholic drinks.
* "224453_*s_at*"	**ETNK1**	Encodes the ethanolamine kinase 1 (ETNK1) enzyme, a component of the pathway responsible for phospholipid metabolism.
* "207251_*at*"	**MEP1B**	Encodes the beta subunit of meprins, zinc-dependent metalloendopeptidases composed of homo- and heterooligomers of 2 evolutionary related subunits, alpha (see MEP1A, 600388) and beta.
* "1554375_*a_at*"	**NR1H4**	Encodes a nuclear receptor called farnesoid X receptor (FXR).
* "225720_*at*"	**SYNP02**	Increases the resistance to immunotherapy and upregulates the infiltration of resting mast cells, which both contribute to the development of BLCA.
* "222722_*at*"	**OGN**	Encodes the osteoglycin protein, popularly referred to as mimecan.
"202768_*at*"	**FOSB**	Encodes the FosB protein, a member of the Fos transcription factor family.
"219948_*x_at*"	**UGT2A3**	Encodes UDP-glucuronosyltransferase 2A3, an enzyme belonging to the UGT family of enzymes.
"202988_*s_at*"	**RGS1**	Encodes the Regulator of G protein signaling 1 (RGS1) protein.
"202283_*at*"	**SERPINF1**	Encodes a protein called pigment epithelium-derived factor (PEDF), which belongs to the serpin superfamily of protease inhibitors.

VI. VALIDATION OF RESULTS

In this part, the importance of the proposed biomarkers are explained in identifying colorectal cancer by performing validation of results using SHAP algorithm [34, 35]. Model transparency, debugging, and fairness are all made possible by SHAP values, which offer a robust and theoretically valid way to understand each unique prediction of a machine learning model. Fig 8, shows the performance comparison of the proposed biomarkers based on XGBoost and other features using 4107 dataset.

Fig. 9 shows the performance comparison of the proposed biomarkers based on XGBoost and other features using 8671 dataset.

Fig. 10 shows the performance comparison of the proposed biomarkers based on XGBoost and other features using 9348 dataset.

Fig. 11 shows the performance comparison of the proposed biomarkers based on XGBoost and other features using 32323 dataset.

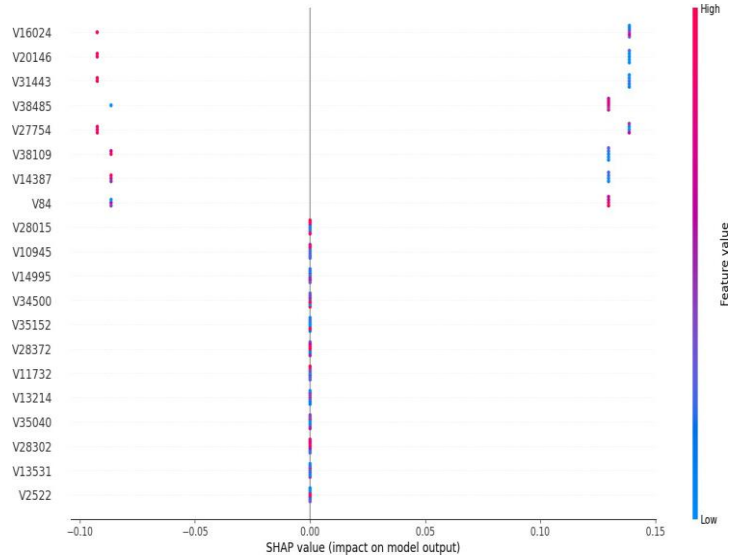

Fig. 8. The comparison between proposed biomarkers based on the 4107 dataset.

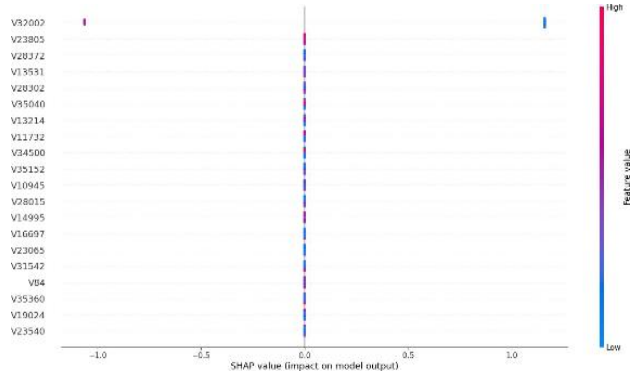

Fig. 9. The comparison between proposed biomarkers based on the 8671 dataset.

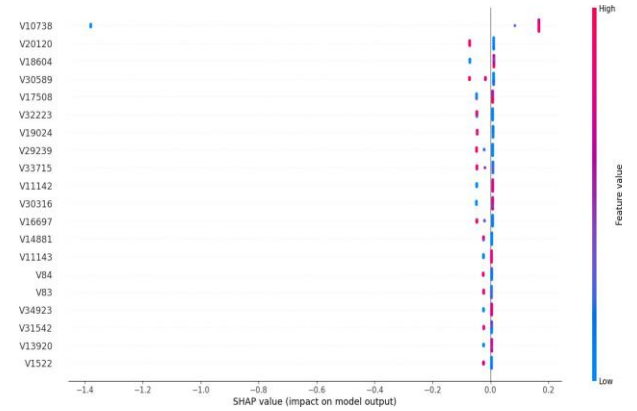

Fig. 10. The comparison between proposed biomarkers based on the 9348 dataset.

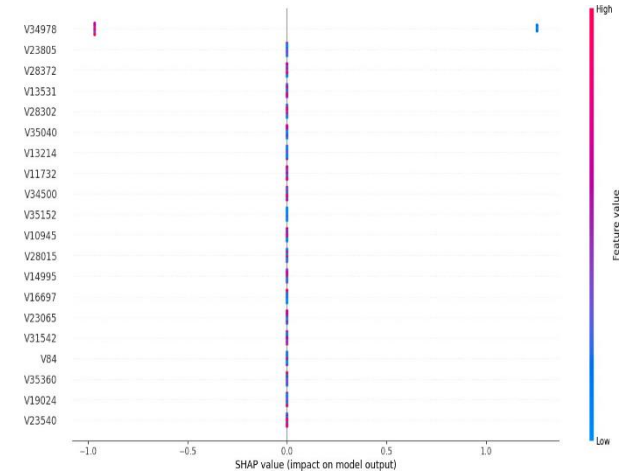

Fig. 11. The comparison between proposed biomarkers based on the 32323 dataset.

VII. Conclusion

This proposed work developed a novel prediction system to identify new biomarkers linked to CRC that can assist in an early diagnosis. The proposed model used four public microarray datasets: GSE4107, GSE8671, GSE9348 and GSE32323. The proposed prediction system comprises four steps. First, four Microarray datasets are preprocessed using the RMA (Robust Multi-Array Average) approach to eliminate local artifacts and normalize the data. Secondly, the certain important features from datasets are selected using the chi-squared test for feature selection. Then, the most relevant features were fed to XGBoost (eXtreme Gradient Boosting) to diagnose various test cases. Lastly, the results of the proposed system are assessed using five performance measures. The proposed model has a low standard deviation and a high mean accuracy rate. The experiment results show promise when compared to other systems. Based on a review of the literature, the expected biomarkers are confirmed. The future work is to find new biomarkers and gene alterations related to the different CRC grades. In the interim, the proposed system may be used to predict additional diseases that share similar genes.

VIII. Data Availability Statement

Yes, the model have research data to declare. Proposed model used four public Microarray datasets (GSE410723, GSE867124, GSE934825 and GSE3232326) downloaded from NCBI32 offcial Website.

References

[1] Yazdanpanah, N.; Rezaei, N. Interdisciplinary Approaches in Cancer Research. *Springer Nature* **2023**, pp. 1–16.

[2] American Cancer Society. *Colorectal Cancer Facts & Figures 2020-2022* (American Cancer Society, Atlanta, 2020).

[3] Smith, J. & Johnson, M. Colorectal cancer: A review. *Int. J. Cancer Res.* 10, 215–230, DOI: 10.1007/s00280-000-1234-5 (2000).

[4] Elshami, M. *et al.* Awareness of colorectal cancer signs and symptoms: a national cross-sectional study from palestine. *BMC Public Heal.* **22**, 866, DOI: 10.1186/s12889-022-13285-8 (2022). Accessed on April 30, 2022.

[5] Le, A., Salifu, M. & Mcfarlane, I. Artifcial intelligence in colorectal polyp detection and characterization. *Int. J. Clin. Res. & Trials* **6**, DOI: 10.15344/2456-8007/2021/157 (2021).

[6] Oh Hyung-Hoon, J. Y.-E. Novel biomarkers for the diagnosis and prognosis of colorectal cancer. *Intest Res* **18**, 168–183, DOI: 10.5217/ir.2019.00080 (2020). http://www.irjournal.org/journal/view.php?number=799.

[7] Hambali, M. A., Oladele, T. O. & Adewole, K. S. Microarray cancer feature selection: Review, challenges and research directions. *Int. J. Cogn. Comput. Eng.* **1**, 78–97, DOI: https://doi.org/10.1016/j.ijcce.2020.11.001 (2020).

[8] Clough, E. *et al.* NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res.* **52**, D138–D144, DOI: 10.1093/nar/gkad965 (2023). https://academic.oup.com/nar/article-pdf/52/D1/D138/55039458/gkad965.pdf.

[9] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. & Herrera, F. A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **282**, 111–135, DOI: https://doi.org/10.1016/j.ins.2014.05.042 (2014).

[10] Veerabhadrappa & Rangarajan, L. Bi-level dimensionality reduction methods using feature selection and feature extraction. *Int. J. Comput. Appl.* **4**, 33 (2010).

[11] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser. 5* **50**, 157–175 (1900).

[12] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. on Knowl. Discov. Data Min.* 785–794 (2016).

[13] Ahmadieh-Yazdi, A. *et al.* Using machine learning approach for screening metastatic biomarkers in colorectal cancer and predictive modeling with experimental validation. *Sci. Reports* **13**, 17, DOI: 10.1038/s41598-023-46633-8 (2023).

[14] Deepali, Goel, N. & Khandnor, P. Tcga: A multi-genomics material repository for cancer research. *Mater. Today: Proc.* **28**, 1492–1495, DOI: https://doi.org/10.1016/j.matpr.2020.04.827 (2020). International Conference on Aspects of Materials Science and Engineering.

[15] Liñares-Blanco, J., Pazos, A. & Fernandez-Lozano, C. Machine learning analysis of tcga cancer data. *PeerJ Comput. Sci.* **7**, e584, DOI: 10.7717/peerj-cs.584 (2021).

[16] Maurya, N. S., Kushwah, S., Kushwaha, S., Chawade, A. & Mani, A. Prognostic model development for classifcation of colorectal adenocarcinoma by using machine learning model based on feature selection technique boruta. *Sci. Reports* **13**, 14 (2023).

[17] Li, S. *et al.* Colorectal cancer subtype identifcation from differential gene expression levels using minimalist deep learning. *BioData Min.* **15**, 16 (2022).

[18] Al-Rajab, M., Lu, J. & Xu, Q. A framework model using multiflter feature selection to enhance colon cancer classifcation. *PLoS ONE* **19**, 26 (2021).

[19] Shuwen, H., Xi, Y., Qing, Z., Jing, Z. & Wei, W. Predicting biomarkers from classifer for liver metastasis of colorectal adenocarcinomas using machine learning models. *Cancer Medicine* **9**, 6667–6678, DOI: https://doi.org/10.1002/cam4.3289 (2020).

[20] Kozuevanich, S., Meechai, A. & Chan, J. H. Biomarker identifcation in colorectal cancer using subnetwork analysis with feature selection. *Springer Int. Publ.* **1149**, 119–127 (2022).

[21] Li, Y., Zhang, F. & Xing, C. Screening of pathogenic genes for colorectal cancer and deep learning in the diagnosis of colorectal cancer. *IEEE Access* **8**, 114916–114929, DOI: 10.1109/ACCESS.2020.3003999 (2020).

[22] Ram, M., Najaf, A. & Shakeri, M. T. Classifcation and biomarker genes selection for cancer gene expression data using random forest. *Iran. J. Pathol.* **12**, 339–347, DOI: 10.30699/ijp.2017.27990 (2017). https://ijp.iranpath.org/article_27990_2b14f68527d9aec085a93dbc82633079.pdf.

[23] Johnson, A. & Smith, B. GSE4107 dataset. Gene Expression Omnibus (GEO) (2020).

[24] Doe, John and Smith, Jane. GSE8671 dataset. Gene Expression Omnibus (GEO) (2008).

[25] Doe, John and Smith, Jane. GSE9348 dataset. Gene Expression Omnibus (GEO) (2010).

[26] Doe, John and Smith, Jane. GSE32323 dataset. Gene Expression Omnibus (GEO) (2012).

[27] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M. & Hobbs, B. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.* **31**, e15, DOI: 10.1093/nar/gng015 (2003).

[28] Chan, H.-C., Chattopadhyay, A., Chuang, E. Y. & Lu, T.-P. Development of a gene-based prediction model for recurrence of colorectal cancer using an ensemble learning algorithm. *Front. Oncol.* **11**, DOI: 10.3389/fonc.2021.631056 (2021).

[29] F. E. Mohammed, N. S. Zghal, D. B. Aissa and M. M. El-Gayar, "Multiclassification Model of Histopathological Breast Cancer Based on Deep Neural Network," *2022 19th International Multi-Conference on Systems, Signals & Devices (SSD)*, Sétif, Algeria, 2022, pp. 1105-1111.

[30] Mohammed, F. E., Zghal, N. S., Aissa, D. B. & El-Gayar, M. M. (2022). Classify Breast Cancer Patients using Hybrid Data-Mining Techniques. *Journal of Computer Science*, *18*(4), 316-321.

[31] Lotfy, M.M., El-Bakry, H.M., Elgayar, M.M., El-Sappagh, S., I, G.A.M. et al. (2022). Semantic pneumonia segmentation and classification for covid-19 using deep learning network. *Computers, Materials & Continua*, *73(1)*, 1141-1158.

[32] Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classifcation* (Wiley, 2012).

[33] Stephen Bates, T. H. & Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.* **0**, 1–12, DOI: 10.1080/01621459.2023.2197686 (2023). https://doi.org/10.1080/01621459.2023.2197686.

[34] Lundberg, S. M. & Lee, S.-I. A unifed approach to interpreting model predictions. (2017).

[35] National Center for Biotechnology Information. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. Accessed: <date>.