BH-index: A predictive system based on serum biomarkers and ensemble learning for early colorectal cancer diagnosis in mass screening

Antonio Battista[a], Rosa Alessia Battista[b], Federica Battista[c], Gerardo Iovane[d,*], Riccardo Emanuele Landi[e]

[a]*A.O.U. S. Giovanni di Dio e Ruggi d'Aragona, UOC Chir Urg, UOC Laboratorio Analisi, Salerno, Italy*
[b]*IRCCS Ospedale San Raffaele, University Vita Salute S. Raffaele, Milan, Italy*
[c]*IRCCS Foundation Policlinico San Matteo, University of Pavia, Pavia, Italy*
[d]*Department of Computer Science, University of Salerno, Salerno, Italy*
[e]*Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy*

Abstract

Keywords: B-index, Colorectal cancer diagnosis, Artificial intelligence, Simulation, Ensemble learning, Mass screening, Machine learning, Majority voting

1. Introduction

Colorectal cancer (CRC) is one of the most common malignancies among the general population, accounting for 10% of all diagnosed cancers. CRC represents the third most frequent malignant tumor in men while is even only the second in women after breast cancer. In 2020 its estimated incidence has been 19:100.000 worldwide, while the mortality rate represents the 9.4% of all cancer-related deaths [1]. Over the past decades, the scientific world witnessed a slight decrease in CRC's incidence and mortality, supposedly related to both implementations of mass screening programs and subsequent earlier recognition of the disease at initial stages. Hemoglobin and DNA alteration detected in stool samples along with endoscopy is the most diffused and recommended screening test in the general population. However, they are burdened by quite low sensitivity and invasiveness, respectively [2]. For this reason, the eventual development of new minimally invasive, highly sensitive, and specific approaches has a crucial role in aiming for recognition of the disease at the earliest stage possible.

*Corresponding author
Email address: giovane@unisa.it, University of Salerno - Via Giovanni Paolo II, 132 - 84084 Fisciano (SA) (Gerardo Iovane)

Artificial Intelligence introduced relevant perspectives of the solution, especially through supervised machine learning aimed at approximating unknown patterns among relevant data. Significant parameters can efficiently and reliably allow inferences about patients' health status; pattern recognition in colorectal cancer diagnosis is in continuously considerable expansion [3, 4, 5, 6].

In this study, we propose a further improvement of B-index [7], a mathematical tool based on Artificial Neural Networks and extended reality for non-invasive early colorectal cancer diagnosis. We faced the prediction problem of cancer presence and staging classification by combining the outcomes provided by multiple models through an ensemble learning, i.e. the majority voting, approach [8]. We performed a comparative analysis of the performances, as binary and staging predictors, provided by four machine learning models: RF (Random Forest) [9], XGB (XGBoost) [10], SVM (Support Vector Machine) [11], and ANN (Artificial Neural Network) [12].

2. Materials and methods

2.1. Serum biomarkers

Carcinoembryonic Antigen (CEA) [13, 14, 15, 16, 17, 18], Ceruloplasmin [19], Haptoglobin [20, 21, 22], Transferrin [23, 24, 25], Tissue Polypeptide Antigen (TPA) [26, 27], CA 19.9 [28, 29, 30], CA 72.4 [31, 32, 33], C-reactive Protein (CRP) [34, 35, 36], CA 50 [28, 37], C4 Complement [38, 39], CA 125 [40, 41, 42], Alpha-1-Antitrypsin [20, 43], Alpha-2-Macroglobulin [44, 39, 45], Ferritin [24, 46, 47, 48], Retinol Binding Protein (RBP) [49, 50, 51], Alpha-1-Acid Glycoprotein [44, 52], Complement C3 [39, 53].

2.2. Dataset

The dataset was built thanks to the participation, between January 2011 and January 2013, of 345 patients (97 found healthy and 248 affected by colorectal cancer) enrolled at several Surgical Departments in Italy. All blood samples were collected preoperatively by venipuncture, 24 hours after hospital admission to reduce psychological stress-induced modification of tested serum parameters. All patients underwent preoperative evaluation including physical examination, colorectal endoscopy, histopathological analysis on biopsies, contrast total body CT scan. Pathological TNM staging was established on surgical specimens according to American Joint Committee on Cancer staging system after CRC surgical excision. Patients were subsequently divided into 4 subclasses according to disease stage: 74 patients classified as stage I, 61 patients belonging to stage II, 76 patients in stage III while stage IV was made up of 37 patients. Ninety-seven healthy controls were selected regardless of sex and age, ranging between 45 and 80 years old. The dosage of the serum parameters described in Section 2.1 was evaluated in all participants' blood samples. Evaluation of mean, STDEV, and t-Student was assessed in the different subgroups: CRC patients (stage I to IV) and participants in apparent good health.

2.3. Explanatory data analysis

In Figure 1 we show the two principal components of the dataset by employing PCA (Principal Components Analysis) [54], which is an unsupervised approach that permits to reduce the dimensionality of the feature space. The two components highlighted in the graph represent the main components of the observation space, and therefore the dimensions, normalized for having zero mean, with the greatest variance in the dataset.

We notice that the observations accumulate more in a precise point of space and that they expand slightly along the diagonals. The expansion that concerns the negative values of the first main component (PCA 1) presents more negative cases (no CRC presence), while the one that concerns the positive values of the aforementioned main component presents most of the positive cases (CRC presence). The expansion of positive cases highlights some outliers that are close and others very far from the most significant density zone.

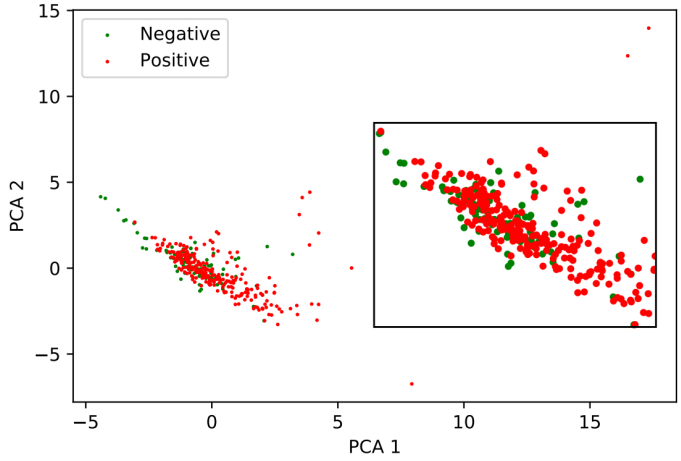


Figure 1: Bidimensional visualization of the original dataset through PCA for the binary problem. Green and red points represent negative and positive patients, respectively.

In Figure 2 we show the Box and Whisker plots of the dataset distribution for the binary problem, in which each box represents upper and lower quartiles and outliers are displayed as individual points. Biomarkers that present greater dispersion are Haptoglobin, Transferrin, Tissue Polypeptide Antigen, Alpha-1-Antitrypsin, Alpha-2-Macroglobulin, Ferritin, Alpha-1-Acid Glycoprotein, and Complement C3, both for positive and negative patients. Each parameter of the input space has a significant number of outliers, most of which occur in the case of positive patients. For negative patients the outliers are all relatively close to the whiskers, except for Tissue Polypeptide Antigen which presents two observations very far from the interquartile range, while for positive patients we find occurrences of outliers very far from the relative confidence intervals; for the Tissue Polypeptide Antigen, e.g., we verify a wider diffusion of the observations

beyond the box range. These aspects suggest that the observations related to negative patients are more adequate than those belonging to negative patients, since in the first case, even if the amount of data is limited, the distribution of the markers is defined more densely in the proximity of the whiskers. Thus, the distribution defined by the interquartile ranges and the relative medians is likely to the true distribution of the negatives. The high dispersion of outliers in the case of positive patients shows a high degree of variance in the distribution and, consequently, suggests the need to acquire more observations.

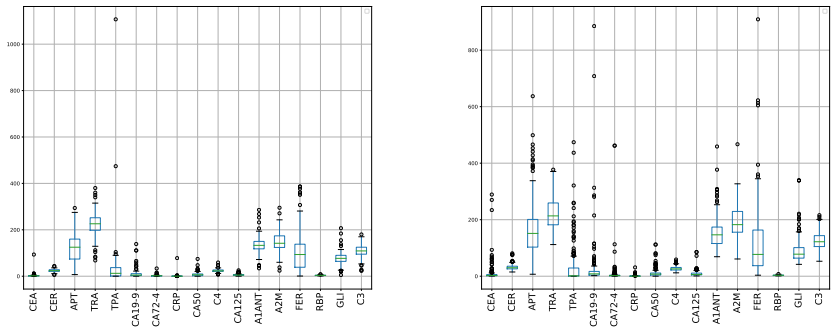


Figure 2: Features distribution related to the negative (n=97) and positive (n=248) patients.

Similar to the binary, Figures 3 and 4 show the Box and Whisker plots of the dataset distribution for the staging problem. Biomarkers' distribution for stage I presents greater dispersion against the other stages, but its outliers occur more in the proximity of the whiskers. Stages II and III have narrower confidence intervals and, except for Ferritin, they involve, similarly to stage I, all outliers in the proximity of the whiskers. Stage IV, on the other hand, while presenting less dispersive confidence intervals, has a large number of outliers far from the interquartile ranges, especially as regards CA19-9, which for stage I, II, and III assumes values lower than 300, but which for the IV it exceeds the limit of 800. Another interesting aspect is that the Ferritin assumes values lower than 400 in the case of stages I and IV, but values higher than the limit of 600 for stages II and III. Thus, the high dispersion of outliers for the positives is most affected by the variance associated with stage IV.

The elimination of the outliers does not represent a convenient choice both for the binary and staging decisions, as it would eliminate useful information. However, as we will see later, it will be possible, through an ensemble learning approach, to reduce the variance introduced by the decisions concerning positive patients without removing outliers.

2.4. Ensemble learning

Ensemble learning is a statistics and machine learning approach that combines multiple learners to obtain better predictive performances than those

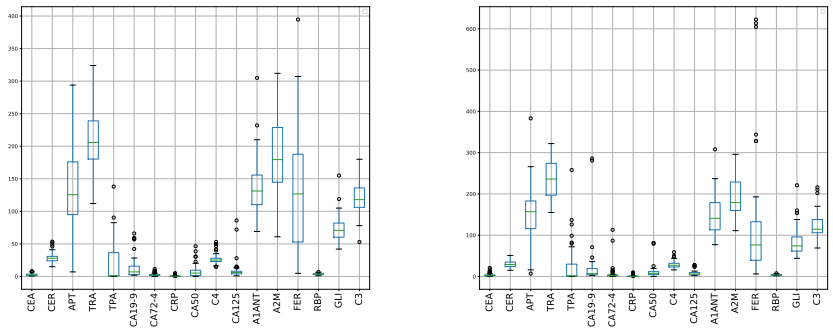


Figure 3: Features distribution related to the positive patients for stages I (n=74), on the left, and II (n=61), on the right.

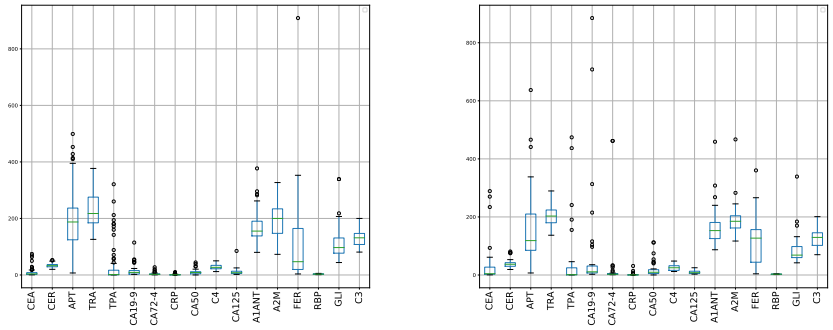


Figure 4: Features distribution related to the positive patients in the training set for stages III (n=76), on the left, and IV (n=37), on the right.

which can be obtained by deploying the combined models individually. Since any classification error is described by the bias-variance trade-off, i.e. the balance between the accuracy and the precision of the classifier when trained on different sets [55], an ensemble model allows to reduce a learner's variance without increasing the bias, i.e. the *bagging*, and a learner's bias without increasing the variance, i.e. the *boosting*.

In general, ensemble learning deals with the training of several weak learners, i.e. which all have a high level of variance, as in the case of bagging, or bias, as in the case of boosting, the combination of which generates a strong learner that provides better performance than those achieved by each of the combined models taken individually. In bagging, the predictions of several weak learners with low bias but high variance are combined to obtain a new model characterized by a lower variance. In boosting, the predictions of several weak learners with a low variance but high variance are combined to obtain a new model characterized by

a lower bias. Another significant ensemble approach is *stacking*, which consists of combining multiple model predictions through a meta-model, i.e. another learner, instead of a deterministic process.

The ensemble approach can be considered as the translation in terms of machine learning of the "unity is strength" concept. Ensemble learning can also be adapted, as we will see in Section 3.2, to combine multiple strong learners to obtain better performances. In the present study, we adopt the majority voting, which is a bagging approach consisting of performing the average of the predictions provided by multiple learners that, in this work, are trained on a common dataset.

2.5. Feature selection

Feature selection was carried out through the evaluation of the significance of the input dimensions after the training phase of Random Forest and XGBoost models.

The importance of a feature is evaluated as the decrease of the impurity related to a given node, weighted by the probability of reaching that same node (the number of samples that reach the node divided by the total number of samples). This metric is computed by the Gini Importance, also called Mean Decrease in Impurity (MDI), which counts the times a feature is involved in splitting a node, weighted by the number of samples it splits [56]. For a binary tree, the importance of the i-th feature on a decision tree is defined as:

$$FI_i = \frac{\sum_{j \in S_{i,j}} NI_j}{\sum_{k \in K} NI_k} \tag{1}$$

where $S_{i,j}$ is the number of times the i-th feature is involved in splitting the j-th node of the tree, K is the total number of the nodes of the tree, and NI_j is the importance of the j-th node, which is defined as:

$$NI_j = w_j C_j - w_{L(j)} C_{L(j)} - w_{R(j)} C_{R(j)} \tag{2}$$

where w_j is the weighted number of samples reaching the j-th node, C_j is the impurity value of the j-th node; $L(j)$ and $R(j)$ are the child nodes, respectively, from left and right split on the j-th node [57]. A high value of the FI_i score indicates a high significance for a given feature.

2.6. Basic and starting solution: B-index

The deployed predictive model was an Artificial Neural Network, developed with simultaneous evaluation of all the serum proteins and antigens tested [26, 58, 59, 32], to distinguish healthy controls from colorectal cancer patients. The classes 0 (negative outcome) and 1 (positive result) were the two possible output values; a positive response identified a patient who probably incurred one of the four TNM stages (I, II, III, and IV). The input dimensions were all the serum levels of the above 17 critical biomarkers.

To better evaluate the results achieved by the predictive model, new sets of samples were generated by performing an extended reality simulation whose deployed algorithm was characterized by the following properties.

i) Data are generated in the function of the TNM classes to avoid admixtures. In particular, each parameter varies in the range $[0, 99]$ of real numbers.

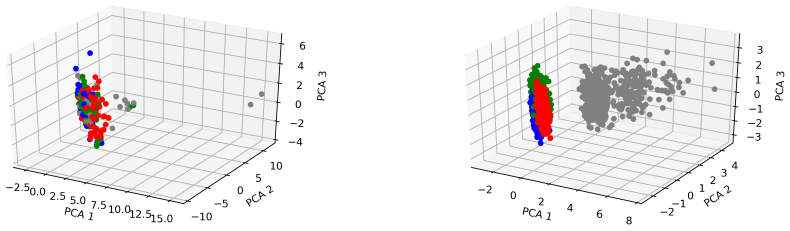
ii) Input parameters are fuzzified through the analysis of the entire distribution of the original samples; variation ranges divided into the five classes: i) *low*, i.e. range $[0, 20[$; ii) *below average*, i.e. range $[20, 40[$; iii) *average*, i.e. range $[40, 60[$; iv) *above average*, i.e. range $[60, 80[$; v) *high*, i.e. range $[80, 99]$.

iii) Each original profile of parameters is remapped into a set of integers on the above fuzzy set, i.e., into the range $[1, 5]$ of integers.

iv) All possible configurations of profiles are generated according to the related TNM class. The transferrin, ferritin, C3, and C4 dimensions are not involved in the recombination process but remapped with the original class values.

v) After 400,000 cases for each TNM class are generated, the fuzzified parameters' tuples are converted into real values using random inter-class generation. For example, in case a parameter has a score of 4, a random number in the $[60, 80]$ range is generated, e.g., 68.75.

The advantage of the fuzzification process also reduces the computational cost of the extended reality procedure. Bounds on the generation of extended data were set to ensure a fully representative fidelity of the original dataset. The quantity of generated extended data was two million samples. In Figure 5, we show the visualization of the real and extended data through PCA (Principal Components Analysis) related to the positive CRC patients.



a) Real PCA components. b) Extended PCA components.

Figure 5: Visualization of real (a) and extended (b) PCA components related to positive CRC patients. Blue, green, red, and gray points represent I, II, III, and IV TNM stages, respectively.

Samples for model training, validation, and testing were chosen by considering data related to all the possible TNM outcomes. The data structure was characterized by the association between serum biomarkers' values and the TNM stage outcome. Among all the 345 participants, 51 were chosen for the

validation set, while the authors generated two million samples (500k for each TNM class) for the test set through an extended reality procedure. The goal of testing the model on extended data was to obtain a larger dataset than the original one to simulate tests on a huge amount of participants. The tests on extended data permitted to study of the differentiation properties which are not visible in low statistics, with the advantage of having the same general properties as the original dataset. The generated data were computed according to the TNM classes, the fuzzification of diagnostic parameters, and the remapping of correspondent values for each profile in the original dataset. Given the TNM, all possible combinations of the 17 input dimensions were created through conditioned stochastic generation. Model's performances were the following:

- Accuracy on real data (n=51): 80.08%;
- Accuracy on extended data (n=2M): 91.11%;
- False positives on real data (n=51): 33.30%;
- False positives on extended data (n=2M): 47.20%;
- False negatives on real data (n=51): 10.20%;
- False negatives on extended data (n=2M): 3.00%;
- Sensitivity on real data (n=51): 89.80%;
- Sensitivity on extended data (n=2M): 99.97%;
- Specificity on real data (n=51): 66.70%;
- Specificity on extended data (n=2M): 76.42%.

The study obtained good results, but it faced the prediction of cancer presence only. As we will see, the improved solution enforced the previous model and also faced the TNM staging prediction issue by taking advance of the very same extended reality simulation.

2.7. Proposed improving solution: BH-index

Based on the previous study, we have built an improvement system whose purpose is to diagnose the presence of colorectal cancer and the related TNM stage. The proposed system is divided into the following chores:

- Binary predictor: it provides the prediction regarding the probability of absence/presence of cancer;
- Staging predictor: it provides, based on the binary predictor outcome, the prediction regarding the cancer degree of staging (I, II, III, IV).

As shown in Figure 6, the binary predictor considers the 17 biomarkers' values of a given patient and outputs one of the two following states: 1 in case of significant cancer presence probability and 0 in the opposite case. The staging predictor takes as input the same biomarkers acquired by the binary predictor and outputs one of the following states: 1, 2, 3, and 4 in case of the significant relative probability of incurring, respectively, in TNM stages I, II, III, and IV. The activation of the computation related to the staging predictor is asserted only when the binary predictor returns a positive outcome for a given input of biomarkers.

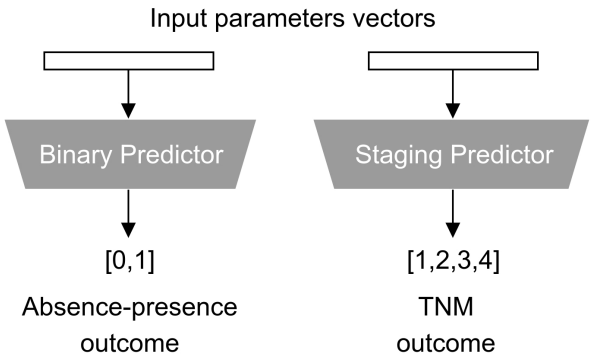


Figure 6: Representation of the binary and staging predictors.

We performed a comparative analysis of the performances, as binary and staging predictors, related to RF (Random Forest), XGB (XGBoost), SVM, and ANN models. The choice of training a set of models of different types also allows to perform the ensemble, i.e., as described in Section 2.3, to combine the relative predictions to obtain greater performances.

The first experimentation, as described in Section 3.1, was conducted by training the aforementioned models in a balanced way, i.e. with the same distribution of positives and negatives in both training and validation sets. In this way, we allow the comparison of the binary predictor with the related works and the evaluation of the most significant biomarkers. The second experimentation, as described in Section 3.2, was conducted by training the same chosen models as above in a positive-oriented unbalanced way, i.e. with a different ratio of positives and negatives in the training and validation sets, and, subsequently, by combining, through an ensemble voting mechanism, the predictions of the models which provided us the better performances. In this way, we strengthen the binary predictor performances by minimizing false negatives and reducing the noise in the prediction to increase the overall success factor.

In both the experiments, the dataset was the same available for the prodromic study, i.e. the one described in Section 2.2. Training and validation processes were based on the real samples, while the test was performed based on the samples generated by the extended reality approach. Regarding the ex-

perimentation conducted with balanced sets, 276, 69, and 2 million samples were allocated, respectively, to the training, validation, and test sets; Regarding the experimentation conducted with positive-oriented unbalanced sets, 294, 51, and 2 million samples were allocated to the training, validation, and test sets. Regarding accuracy metrics, the validation score indicates the degree of models' generalization on the set of real data and its accuracy in predicting the patient's status. The test score suggests the model's behavior, during the simulation phase, on configurations of highly variant input parameters. These two metrics provide an overall evaluation of the models' behavior when predicting outcomes for both real and simulated patients.

2.8. Performance metrics

Models performance is evaluated regarding the validation and the test sets, respectively, for real and extended data. We evaluate the models employing the accuracy metric for the validation and test sets. As for the best performing models, we take into consideration also the sensitivity and specificity metrics for the binary predictor, while for the staging predictor we evaluate the confusion matrix.

Considering a binary decision problem in which the two possible outcomes are positive/negative, we provide the following definitions: a true positive (TP) is an outcome where the model correctly predicts the positive class; a true negative (TN) is an outcome where the model correctly predicts the negative class; a false positive (FP) is an outcome where the model incorrectly predicts the positive class; a false negative (FN) is an outcome where the model incorrectly predicts the negative class. The expressions of the accuracy, sensitivity, and specificity scores of a model are:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}.$$

Accuracy represents the number of correctly predicted data points out of all the data points, i.e. the number of correct predictions out of the total number of predictions performed within a test; sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive); this can also be represented in the form of a True Positive Rate (TPR). Specificity is defined as the proportion of the actual negatives which got predicted as the negative (or true negative); this proportion could also be called a True Negative Rate (TNR).

Regarding the staging problem, we use the confusion matrix, which is, as shown in Table 1, a tabular way of visualizing the performance of a multiclass prediction model. Each entry in a confusion matrix denotes the number

of predictions made by the model where it classifies the classes correctly or incorrectly. In the following sections, we will employ its normalized version, which consists of dividing each element by the total number of data points whose actual class is indexed by the row.

	1	**2**	**3**	**4**
1	$N_{1\|1}$	$N_{2\|1}$	$N_{3\|1}$	$N_{4\|1}$
2	$N_{1\|2}$	$N_{2\|2}$	$N_{3\|2}$	$N_{4\|2}$
3	$N_{1\|3}$	$N_{2\|3}$	$N_{3\|3}$	$N_{4\|3}$
4	$N_{1\|4}$	$N_{2\|4}$	$N_{3\|4}$	$N_{4\|4}$

Table 1: Example of a confusion matrix for a multi-class decision problem in which the four possible outcomes are stages I, II, III, and IV, respectively enumerated as 1, 2, 3, and 4. The generic element $N_{C_i|C_j}$ denotes the number of times the model predicts class C_i when the actual class is C_j.

3. Results and discussion

3.1. Results with balanced sets

In this Section, we discuss the results achieved with the training based on balanced datasets, identifying the biomarkers that contribute most significantly to the decision. In Sections 3.1.1 and 3.1.2, we show the results achieved for binary and staging predictors, respectively.

3.1.1. Binary predictor results

As can be seen in Table 2, the binary predictors deployed with RF and XGB generalized the validation better than the test set, while SVM and ANN models generalized the test better than the validation set. By performing feature selection with the approach described in Section 2.4 for the binary problem concerning RF and XGB, we encountered, as shown in Figure 7, that the most significant markers are Ceruloplasmin, which is highly significant for both the algorithms, and Alpha-2-Macroglobulin, which is more significant for RF than for XGB. The other three highly significant markers are Transferrin, CA 72.4, and C-reactive Protein; the former and the second are more decisive for XGB than for RF, while the latter is not relevant for RF but highly decisive for XGB. For both the algorithms, the least significant markers are CA-50 and A-1-Antitrypsin; for this reason, we have carried out a new training of the models by resizing the feature space in this sense.

Training based on the results of the features selection (all features except CA 50 and A-1-Antitrypsin) show, similarly to the case of training based on all features, that RF and XGB generalized the validation better than the test set, while SVM and ANN generalized the test better than the validation set. We note, however, a slight improvement in the accuracy of XGB and ANN on the validation set, while we find, on the same set, a worsening of RF and SVM. Compared to the previous study, we obtained that XGB improves the binary

Binary predictor (all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	276	69	2M	84.05	57.06
XGB	276	69	2M	90.00	71.92
SVM	276	69	2M	76.81	99.55
ANN	276	69	2M	76.81	86.89
Binary predictor (all features except CA 50 and A-1-Antitrypsin)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	276	69	2M	82.60	57.96
XGB	276	69	2M	91.30	71.70
SVM	276	69	2M	72.46	95.27
ANN	276	69	2M	79.71	91.31

Table 2: Results for the binary problem with balanced sets. Train with 199 positives and 78 negatives; test with 51 positives and 19 negatives.

prediction of 11.22% on real data. The details related to the performance are the following:

- Accuracy on real data (n=69): 91.30%;
- Accuracy on extended data (n=2M): 71.70%;
- False positives on real data (n=69): 26.31%;
- False positives on extended data (n=2M): 9.12%;
- False negatives on real data (n=69): 0.02%;
- False negatives on extended data (n=2M): 33.09%;
- Sensitivity on real data (n=69): 90.00%;
- Sensitivity on extended data (n=2M): 66.90%;
- Specificity on real data (n=69): 93.33%;
- Specificity on extended data (n=2M): 90.87%.

A comparison with the previous study reveals that false positives and false negatives have, respectively, decreased by 38.08% and increased by 30.09% on extended data; sensitivity decreased by 33.07%, and specificity increased by 14.45% on extended data. Regarding real data, false positives and false negatives have decreased by 6.99% and 10.18%; sensitivity and specificity have increased by 0.2% and 26.63%.

3.1.2. Staging predictor results

Table 3 shows the staging problem results by training the same four models used for the binary predictor. It can be noticed that, without feature selection, the real data were better generalized by RF and XGB and that, by deleting CA 50 and A-1-Antitrypsin, the model that reached the best accuracy score

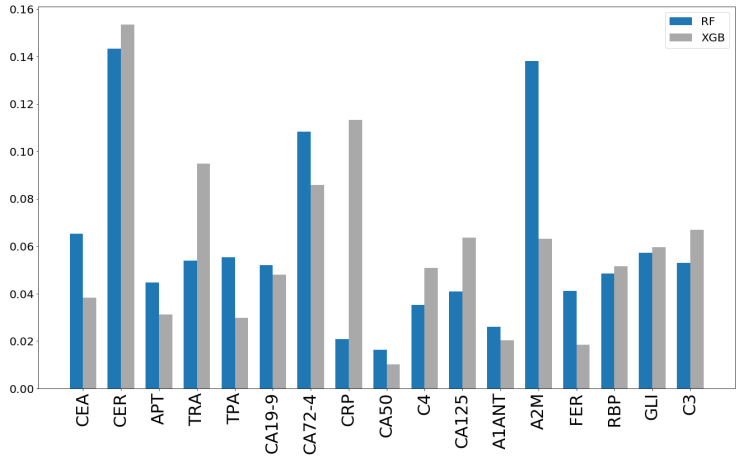


Figure 7: Feature importance related to RF and XGB for the binary problem.

Staging predictor (all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	197	51	2M	58.82	76.16
XGB	197	51	2M	66.66	42.32
SVM	197	51	2M	47.05	37.61
ANN	197	51	2M	45.09	30.78
Staging predictor (all features except CA 50 and A-1-Antitrypsin)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	197	51	2M	62.74	45.50
XGB	197	51	2M	66.66	16.57
SVM	197	51	2M	47.05	37.61
ANN	197	51	2M	56.86	38.66

Table 3: Results for the staging problem with balanced sets. Training with 57, 53, 59, and 28 samples for stages I, II, III, and IV, respectively; test with 17, 8, 17, and 9 samples for stages I, II, III, and IV, respectively.

on real data was XGB. By training the models separately, we deduce that the maximum achieved accuracy score is 66.66% on real data.

Unlike the binary, for the staging predictor, we verify that all the biomarkers are significant for both RF and XGB. We did not encounter further improvements on real data, except for a reduction of the noise in the prediction due to a general lowering of accuracy related to extended data (especially for RF and XGB). In Figure 8 we notice that the spectrum of the significance of the features generally presents high values, and therefore it is possible to state that for the staging problem the markers identified in the present study represent a necessary basis for the decision, suggesting the identification of further features or the acquisition of more data. In particular, we note that for the staging problem Carcinoembryonic Antigen and Alpha-1-Acid Glycoprotein are of particular importance; the first of these two biomarkers is highly and averagely significant for RF and XGB, respectively, while the second is relevant for both algorithms.

The other three highly significant markers are Ceruloplasmin, C-reactive Protein, and Retinol Binding Protein; the former is decisive both for XGB and RF, while the second and the latter are more relevant for XGB. For both the algorithms, the least significant markers are Transferrin and Complement C3. Thus, we can conclude that the best staging predictor is the XGB model characterized by the confusion matrices shown in Table 4.

	Real data			
	1	**2**	**3**	**4**
1	0.70	0.12	0.18	0.00
2	0.12	0.62	0.00	0.26
3	0.00	0.18	0.82	0.00
4	0.34	0.00	0.33	0.33

	Extended data			
	1	**2**	**3**	**4**
1	0.12	0.86	0.01	0.01
2	0.79	0.08	0.12	0.01
3	0.76	0.08	0.06	0.10
4	0.00	0.00	0.60	0.40

Table 4: Confusion matrices of the model trained with XGB for the staging predictor with balanced sets on real and extended data. Row and column indices represent, respectively, the true expected and the predicted outcomes of the model.

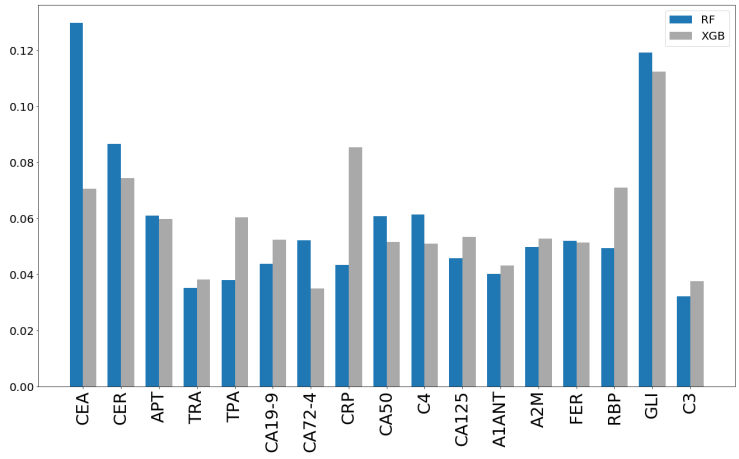


Figure 8: Feature importance related to RF and XGB for the staging problem.

The performance related to the staging predictor on real data shows that stage I is confused to an acceptable level with II and IV, while stage II is confused with I and IV. The best score is obtained for stage III, which is confused only with II. The worst performances are obtained with stage IV, which is often confused with I, II, and IV. As for the performance concerning the extended data, the staging predictor confuses almost all the occurrences of stages II and III with those of stage I, while stage IV is confused only with III.

3.2. Results with positive-oriented unbalanced sets and majority voting

In this Section, we discuss the results achieved with the training based on positive-oriented unbalanced datasets by performing dimensionality reduction through PCA and combining the three best models through a majority voting

approach. In Sections 3.2.1 and 3.2.2, we show the results achieved for binary and staging predictors, respectively.

3.2.1. Binary predictor results

As shown in Table 5, the binary predictors deployed with RF, XGB, and ANN generalized the validation better than the test set, while the SVM model generalized suitably both of them. By performing dimensionality reduction through PCA, we have reduced the input domain to 3 main components, but we did not encounter further improvements. Compared to the previous study, we obtained that XGB improves the binary prediction of 10.11% on real data. Furthermore, importance weights analysis suggested that all the features, i.e., the 17 input biomarkers, are statistically relevant to perform the outcome's prediction.

Binary predictor (all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	294	51	2M	84.31	57.88
XGB	294	51	2M	90.19	64.21
SVM	294	51	2M	86.27	96.25
ANN	294	51	2M	86.27	73.45
Binary predictor (PCA dimensions, all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	294	51	2M	74.50	80.00
XGB	294	51	2M	74.50	88.30
SVM	294	51	2M	76.47	88.90
ANN	294	51	2M	74.50	67.82

Table 5: Results for the binary problem with positive-oriented unbalanced sets. Train with 215 positives and 80 negatives; test with 35 positives and 17 negatives.

Further experimentation was performed by adopting ensemble learning, i.e. combining ANN, SVM, and XGB models without dimensionality reduction. This combined model provided the binary decision according to the mean value of the outcomes predicted by the above three models. Results show 98.03% and 84.63% accuracy scores, respectively, for real and extended data. The details related to the performance are the following:

- Accuracy on real data (n=51): 98.03%;
- Accuracy on extended data (n=2M): 84.63%;
- False positives on real data (n=51): 7.69%;
- False positives on extended data (n=2M): 10.50%;
- False negatives on real data (n=51): 0.00%;
- False negatives on extended data (n=2M): 16.57%;
- Sensitivity on real data (n=51): 100.00%;

- Sensitivity on extended data (n=2M): 57.44%;
- Specificity on real data (n=51): 92.30%;
- Specificity on extended data (n=2M): 96.94%.

A comparison with the previous study reveals that false positives and false negatives have, respectively, decreased by 36.7% and 13.57% on extended data; sensitivity decreased by 42.53% and specificity increased by 20.52% on extended data. Regarding real data, false positives and false negatives have decreased by 25.61% and 10.20%; sensitivity and specificity have increased by 10.20% and 25.60%.

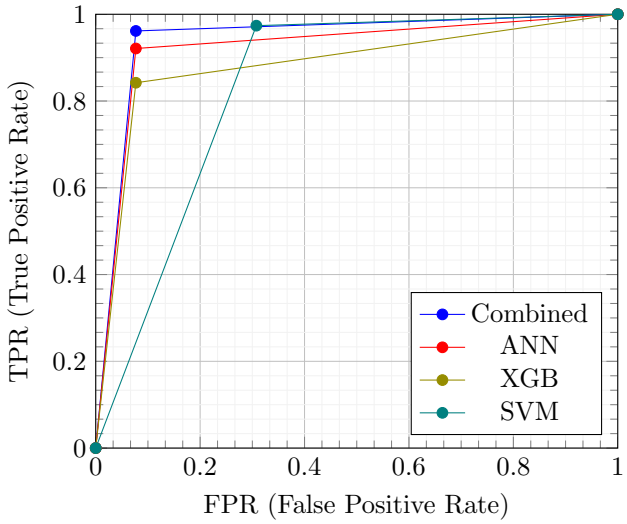


Figure 9: ROC curves of ANN, XGB, SVM and the combined model for the binary problem with positive-oriented unbalanced sets.

The combined model results in an improvement since higher accuracy on real data is promising for the diagnosis's correctness. Furthermore, the decreasing accuracy of extended data indicated the simulated evaluation's relevance as a measure of the prediction's noise. From the plot in Figure 9, it is possible to analyze the trend of the ROC (Receiver Operating Characteristic) curves, which highlight the relationships, on real data, between TPR (True Positive Rate) and FPR (False Positive Rate) for the three single models and the combined one. It can be seen that the combined model inherits the FPR score from ANN and XGB, as well as the TPR score from the SVM, resulting in a better generalizing model. Compared to the SVM-only, the combined model provides much greater accuracy on real data and improves the B-index performance by 16% against the previous work.

3.2.2. Staging predictor results

Table 6 shows the staging problem results by training the same four models used for the binary predictor. It can be noticed that, without dimensionality reduction, the real data were better generalized by RF and ANN and that, by adopting PCA dimensionality reduction, the model that reached the best accuracy score on real data was RF. By training the models separately, we deduce that the maximum achieved accuracy score is 60% on real data.

Staging predictor (all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc.** (%)	**Test acc.** (%)
RF	223	25	2M	44.00	49.15
XGB	223	25	2M	56.00	31.76
SVM	223	25	2M	32.00	24.02
ANN	223	25	2M	44.00	56.41
Staging predictor (PCA dimensions, all features)					
Model	**Train (n)**	**Validation (n)**	**Test (n)**	**Validation acc. (%)**	**Test acc. (%)**
RF	223	25	2M	60.00	47.37
XGB	223	25	2M	56.00	37.58
SVM	223	25	2M	44.00	54.51
ANN	223	25	2M	48.00	74.46

Table 6: Results for the staging problem with positive-oriented unbalanced sets. Training with 71, 52, 69, and 31 samples for stages I, II, III, and IV, respectively; test with 8, 5, 9, and 3 samples for stages I, II, III, and IV, respectively.

Even for the staging predictor, the importance weights analysis verified that all the input features are significant, but, unlike the binary predictor, the reduction of dimensionality via PCA introduced a slight improvement. Figure 5 in Section 2.2 shows the visualization of the three main components extracted from the real and the extended datasets composed of patients affected by cancer in association with the TNM stages.

We note the high complexity of classification due to the significant density of points present in the plot associated with negative values of component 2. It is also noticeable that the data related to a type IV TNM are more distinguishable than the others, at least as regards the positive values of the second PCA component. The presence of outliers related to the IV stage determines, in the extended dataset, a well-defined decision area isolated from the high-density section associated with the other TNM stages. This aspect suggests that predicting a type IV stage may be more reliable than the different TNM configurations. This observation, however, represents only a hypothesis since the amount of real data available for the present study was limited; in fact, the acquisition of further data in the field could reveal, in the future, different confidence intervals between the TNM points.

The combination of the three best models, i.e., RF, XGB, and ANN with dimensionality reduction, was also performed for the staging by using the same ensemble voting mechanism used for the binary predictor. The results show that performances achieved 60% and 64.35% accuracy scores, respectively, on real and extended data. Thus, the best staging predictor was the combined model characterized by the confusion matrices shown in Table 7.

	Real data			
	1	**2**	**3**	**4**
1	0.75	0.25	0.00	0.00
2	0.00	0.75	0.25	0.00
3	0.00	0.60	0.40	0.00
4	0.50	0.25	0.00	0.25

	Extended data			
	1	**2**	**3**	**4**
1	0.96	0.04	0.00	0.00
2	0.96	0.04	0.00	0.00
3	0.00	0.25	0.75	0.00
4	0.00	0.00	0.18	0.82

Table 7: Confusion matrices of the combined model for the staging predictor with positive-oriented unbalanced sets on real and extended data. Row and column indices represent, respectively, the true expected and the predicted outcomes of the model.

Performances regarding the extended reality test resulted better against the validation. This aspect means that the combined model generalizes better extended than the real data. We also notice that, regarding real data, the least confusing stages are I and II, while III and IV are mostly confused, respectively, with II and I stages. Regarding extended data, the combined model often confuses III with the II stage. These results suggest the need to acquire more data for the training or introduce a certain degree of determinism into the model. Further medical knowledge can be beneficial to define confidence intervals for the staging decision.

3.3. Discussion

3.4. Comparison to the related works

4. Conclusions and future works

References

[1] Globocan 2020 accessed on december 17, 2020.
URL `https://gco.iarc.fr/today/home`

[2] J. S. Lin, M. A. Piper, L. A. Perdue, C. M. Rutter, E. M. Webber, E. O'Connor, N. Smith, E. P. Whitlock, Screening for colorectal cancer: Updated evidence report and systematic review for the us preventive services task force, JAMA 315 (23) (2016) 2576–2594.

[3] L. Bottaci, P. J. Drew, J. E. Hartley, M. B. Hadfield, R. Farouk, P. W. Lee, I. M. Macintyre, G. S. Duthie, J. R. Monson, Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions, The Lancet 350 (9076) (1997) 469–472.

[4] H. Goyal, R. Mann, Z. Gandhi, A. Perisetti, A. Ali, K. Aman Ali, N. Sharma, S. Saligram, B. Tharian, S. Inamdar, Scope of artificial intelligence in screening and diagnosis of colorectal cancer, Journal of Clinical Medicine 9 (10) (2020) 3313.

[5] M. C. Hornbrook, R. Goshen, E. Choman, M. O'Keeffe-Rosetti, Y. Kinar, E. G. Liles, K. C. Rust, Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data, Digestive diseases and sciences 62 (10) (2017) 2719–2727.

[6] N. Dimitriou, O. Arandjelović, D. J. Harrison, P. D. Caie, A principled machine learning framework improves accuracy of stage ii colorectal cancer prognosis, npj Digital Medicine 1 (1) (2018) 1–9.

[7] A. Battista, R. A. Battista, F. Battista, L. Cinquanta, G. Iovane, M. Corbisieri, A. Suozzo, Development of a new mathematical tool for early colorectal cancer diagnosis and its possible use in mass screening, Journal of Interdisciplinary Mathematics 22 (6) (2019) 811–835.

[8] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science (2020) 1–18.

[9] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.

[10] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[11] S. Suthaharan, Support vector machine, in: Machine Learning Models and Algorithms for Big Data Classification, Springer, 2016, pp. 207–235.

[12] G. P. Zhang, Neural networks for classification: a survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 30 (4) (2000) 451–462.

[13] L. Molnar, P. Rahoty, E. Bauer, P. Ronay, I. Besznyak, S. Otto, Correlation between the results of carcinoembryonal antigen (cea) test and the clinical stage of colorectal carcinoma., Acta chirurgica Hungarica 27 (1) (1986) 27–34.

[14] O. Lunde, O. Havig, Clinical significance of carcinoembryonic antigen (cea) in patients with adenocarcinoma in colon and rectum., Acta Chirurgica Scandinavica 148 (2) (1982) 189–193.

[15] M. Rosandić-Pilas, N. Hadzić, A. Stavljenic, M. Juricić, M. Scukanec-Spoljar, Relationship between tissue and serum concentrations of carcinoembryonic antigen (cea) in gastric and colonic carcinomas., Acta medica Austriaca 17 (5) (1990) 89–93.

[16] T. Nakagoe, T. Sawai, H. Ayabe, T. Nakazaki, H. Ishikaw, K. Hatano, K. Kajiwara, K. Miyashita, T. Matsuo, T. Nogawa, et al., Prognostic value of carcinoembryonic antigen (cea) in tumor tissue of patients with colorectal cancer., Anticancer research 21 (4B) (2001) 3031–3036.

[17] H. Luo, K. Shen, B. Li, R. Li, Z. Wang, Z. Xie, Clinical significance and diagnostic value of serum nse, cea, ca19-9, ca125 and ca242 levels in colorectal cancer, Oncology Letters 20 (1) (2020) 742–750.

[18] C. Giessen-Jung, D. Nagel, M. Glas, F. Spelsberg, U. Lau-Werner, D. P. Modest, C. Schulz, V. Heinemann, D. D. Gioia, P. Stieber, Preoperative serum markers for individual patient prognosis in stage i-iii colon cancer, Tumor Biology 36 (10) (2015) 7897–7906.

[19] S. B. Nayak, V. R. Bhat, D. Upadhyay, S. L. Udupa, Copper and ceruloplasmin status in serum of prostate and colon cancer patients, Indian journal of physiology and pharmacology 47 (1) (2003) 108–110.

[20] C. Walker, B. N. Gray, Acute-phase reactant proteins and carcinoembryonic antigen in cancer of the colon and rectum, Cancer 52 (1) (1983) 150–154.

[21] S.-Y. Park, S.-J. Yoon, Y.-T. Jeong, J.-M. Kim, J.-Y. Kim, B. Bernert, T. Ullman, S. H. Itzkowitz, J.-H. Kim, S.-i. Hakomori, N-glycosylation status of β-haptoglobin in sera of patients with colon cancer, chronic inflammatory diseases and normal subjects, International journal of cancer 126 (1) (2010) 142–155.

[22] S. Ghuman, M. V. Hemelrijck, H. Garmo, L. Holmberg, H. Malmström, M. Lambe, N. Hammar, G. Walldius, I. Jungner, W. Wulaningsih, Serum inflammatory markers and colorectal cancer risk and survival, British Journal of Cancer 116 (10) (2017) 1358–1365.

[23] J. Sheng, S.-r. Li, Z.-t. Wu, C.-h. Xia, X. Wu, J. Chen, J. Rao, Transferrin dipstick as a potential novel test for colon cancer screening: a comparative study with immuno fecal occult blood test, Cancer Epidemiology and Prevention Biomarkers 18 (8) (2009) 2182–2185.

[24] M. Prutki, M. Poljak-Blazi, M. Jakopovic, D. Tomas, I. Stipancic, N. Zarkovic, Altered iron metabolism, transferrin receptor 1 and ferritin in patients with colon cancer, Cancer letters 238 (2) (2006) 188–196.

[25] L. J. Herrinton, G. D. Friedman, D. Baer, J. V. Selby, Transferrin saturation and risk of cancer, American journal of epidemiology 142 (7) (1995) 692–698.

[26] H. Putzki, A. Student, M. Jablonski, H. Heymann, Comparison of the tumor markers cea, tpa, and ca 19-9 in colorectal carcinoma, Cancer 59 (2) (1987) 223–226.

[27] F. J. van der Sluis, Z. Zhan, C. J. Verberne, A. C. M. Kobold, T. Wiggers, G. H. de Bock, Predictive performance of tpa testing for recurrent disease during follow-up after curative intent surgery for colorectal carcinoma, Clinical Chemistry and Laboratory Medicine (CCLM) 55 (2).

[28] J. Alvarez, J. Marin, J. Jover, R. Fernandez, J. Fradejas, M. Moreno, Sensitivity of monoclonal antibodies to carcinoembryonic antigen, tissue

polypeptide antigen, alpha-fetoprotein, carbohydrate antigen 50, and carbohydrate antigen 19-9 in the diagnosis of colorectal adenocarcinoma, Diseases of the colon & rectum 38 (5) (1995) 535–542.

[29] T. Nozoe, T. Rikimaru, E. Mori, T. Okuyama, I. Takahashi, Increase in both cea and ca19-9 in sera is an independent prognostic indicator in colorectal carcinoma, Journal of surgical oncology 94 (2) (2006) 132–137.

[30] P. Kuusela, H. Jalanko, P. Roberts, P. Sipponen, J. Mecklin, R. Pitkänen, O. Mäkelä, Comparison of ca 19-9 and carcinoembryonic antigen (cea) levels in the serum of patients with colorectal diseases, British journal of cancer 49 (2) (1984) 135.

[31] G. Lindmark, U. Kressner, R. Bergström, B. Glimelius, Limited clinical significance of the serum tumour marker ca 72-4 in colorectal cancer., Anticancer research 16 (2) (1996) 895–898.

[32] V. Yasasever, Z. Sengün, N. Saydan, H. Onat, N. Dalay, Serum values of ca72. 4 in patients with gastrointestinal system tumors comparison with cea and ca 19.9., European journal of gynaecological oncology 13 (5) (1992) 403–408.

[33] Y. Gao, J. Wang, Y. Zhou, S. Sheng, S. Y. Qian, X. Huo, Evaluation of serum cea, ca19-9, ca72-4, ca125 and ferritin as diagnostic markers and factors of clinical parameters for colorectal cancer, Scientific Reports 8 (1).

[34] C. Kersten, J. Louhimo, A. Ålgars, A. Lahdesmaki, M. Cvancerova, K. Stenstedt, C. Haglund, U. Gunnarsson, Increased c-reactive protein implies a poorer stage-specific prognosis in colon cancer, Acta oncologica 52 (8) (2013) 1691–1698.

[35] S.-J. Myung, Colon tumor and inflammation: is c-reactive protein possible colon tumor marker?, The Korean Journal of Gastroenterology 51 (4) (2008) 265–268.

[36] K. H. Allin, B. G. Nordestgaard, Elevated c-reactive protein in the diagnosis, prognosis, and cause of cancer, Critical Reviews in Clinical Laboratory Sciences 48 (4) (2011) 155–170.

[37] B. E. Persson, E. Ståhle, L. Påhlman, B. Glimelius, O. Nilsson, L. Lindholm, B. Norrgård-Pedersen, J. Holmgren, A clinical study of ca-50 as a tumour marker for monitoring of colorectal cancer, Medical oncology and tumor pharmacotherapy 5 (3) (1988) 165.

[38] K. Konishi, H. Yamamoto, K. Mimori, I. Takemasa, T. Mizushima, M. Ikeda, M. Sekimoto, N. Matsuura, T. Takao, Y. Doki, et al., Expression of c4. 4a at the invasive front is a novel prognostic marker for disease recurrence of colorectal cancer, Cancer science 101 (10) (2010) 2269–2277.

[39] A. Mangano, L. Messina, S. Birgillito, F. Stivala, A. Bernardini, Complelent and its fractions (c3-c4) pattern in subjects with msoplasia, Journal of immunopharmacology 6 (3) (1984) 147–162.

[40] A. Quentmeier, P. Möller, V. Schwarz, U. Abel, P. Schlag, Carcinoembryonic antigen, ca 19-9, and ca 125 in normal and carcinomatous human colorectal tissue, Cancer 60 (9) (1987) 2261–2266.

[41] R. Bast Jr, F.-J. Xu, Y.-H. Yu, S. Barnhill, Z. Zhang, G. Mills, Ca 125: the past and the future, The International journal of biological markers 13 (4) (1998) 179–187.

[42] J. Bukowski, S. Góźdź, J. Słuszniak, W. Korejba, A. Zieliński, Ca 19-9 and ca 125 antigens in the sera of patients with cancer of the large intestine in relation to its clinical progress, Wiadomosci lekarskie (Warsaw, Poland: 1960) 42 (1) (1989) 30–34.

[43] H. Jaberie, S. V. Hosseini, F. Naghibalhossaini, Evaluation of alpha 1-antitrypsin for the early diagnosis of colorectal cancer, Pathology & Oncology Research 26 (2) (2019) 1165–1173.

[44] P. Durdey, N. Williams, D. Brown, Serum carcinoembryonic antigen and acute phase reactant proteins in the pre-operative detection of fixation of colorectal tumours, British journal of surgery 71 (11) (1984) 881–884.

[45] M. Sunderic, A. Sediva, D. Robajac, G. Miljus, P. Gemeiner, O. Nedic, J. Katrlik, Lectin-based protein microarray analysis of differences in serum alpha-2-macroglobulin glycosylation between patients with colorectal cancer and persons without cancer, Biotechnology and Applied Biochemistry 63 (4) (2015) 457–464.

[46] F. Li, T. Kishida, M. Kobayashi, Serum iron and ferritin levels in patients with colorectal cancer in relation to the size, site, and disease stage of cancer, Journal of gastroenterology 34 (2) (1999) 195–199.

[47] T. Kishida, J. Sato, S. Fujimori, S. Minami, S. Yamakado, Y. Tamagawa, F. Taguchi, Y. Yoshida, M. Kobayashi, Clinical significance of serum iron and ferritin in patients with colorectal cancer, Journal of gastroenterology 29 (1) (1994) 19–23.

[48] R. L. Nelson, Iron and colorectal cancer risk: human studies, Nutrition reviews 59 (5) (2001) 140–148.

[49] G. Milano, E. H. Cooper, J. C. Goligher, G. R. Giles, A. M. Neville, Serum prealbumin, retinol-binding protein, transferrin, and albumin levels in patients with large bowel cancer, Journal of the National Cancer Institute 61 (3) (1978) 687–691.

[50] P. Charpiot, R. Calaf, J. Di-Costanzo, J. Romette, M. Rotily, J. Durbec, D. Garcon, Vitamin a, vitamin e, retinol binding protein (rbp), and prealbumin in digestive cancers., International journal for vitamin and nutrition research. Internationale Zeitschrift fur Vitamin-und Ernahrungsforschung. Journal international de vitaminologie et de nutrition 59 (4) (1989) 323–328.

[51] M. V. Abola, C. L. Thompson, Z. Chen, A. Chak, N. A. Berger, J. P. Kirwan, L. Li, Serum levels of retinol-binding protein 4 ((rbp4)) and risk of colon adenoma, Endocrine-Related Cancer 22 (2) (2015) L1–L4.

[52] A. T. Kopylov, A. A. Stepanov, K. A. Malsagova, D. Soni, N. E. Kushlinsky, D. V. Enikeev, N. V. Potoldykova, A. V. Lisitsa, A. L. Kaysheva, Revelation of proteomic indicators for colorectal cancer in initial stages of development, Molecules 25 (3) (2020) 619.

[53] A. Archimandritis, G. Theodoropoulos, M. Tryphonos, A. Germenis, M. Tjivras, A. Kalos, A. Fertakis, Serum protein markers (hp, gc, c3) in patients with colon cancer, Human heredity 43 (1) (1993) 66–68.

[54] R. Bro, A. K. Smilde, Principal component analysis, Analytical Methods 6 (9) (2014) 2812–2831.

[55] R. Polikar, Ensemble learning, in: Ensemble machine learning, Springer, 2012, pp. 1–34.

[56] J. L. Grabmeier, L. A. Lambe, Decision trees for binary classification variables grow equally with the gini impurity measure and pearson's chi-square test, International journal of business intelligence and data mining 2 (2) (2007) 213–226.

[57] Scikit-learn github repository accessed on february 3, 2021.
URL `https://github.com/scikit-learn/scikit-learn/blob/0abd95f742efea826df82458458fcbc0f9dafcb2/sklearn/tree/_tree.pyx#L1056`

[58] C.-C. Chen, S.-H. Yang, J.-K. Lin, T.-C. Lin, W.-S. Chen, J.-K. Jiang, H.-S. Wang, S.-C. Chang, Is it reasonable to add preoperative serum level of cea and ca19-9 to staging for colorectal cancer?, Journal of surgical research 124 (2) (2005) 169–174.

[59] Y. Van der Schouw, A. Verbeek, T. Wobbes, M. Segers, C. Thomas, Comparison of four serum tumour markers in the diagnosis of colorectal carcinoma, British journal of cancer 66 (1) (1992) 148.