entropy

MDPI

Article

A Novel Method for Colorectal Cancer Screening Based on Circulating Tumor Cells and Machine Learning

Eleana Hatzidaki [1], Aggelos Iliopoulos [1] and Ioannis Papasotiriou [2,*]

[1] Research Genetic Cancer Centre SA (RGCC), 53100 Florina, Greece; hatzidaki.eleana@rgcc-genlab.com (E.H.); iliopoulos.aggelos@rgcc-genlab.com (A.I.)
[2] Research Genetic Cancer Centre International GmbH, 6300 Zug, Switzerland
* Correspondence: office@rgcc-genlab.com

Abstract: Colorectal cancer is one of the most common types of cancer, and it can have a high mortality rate if left untreated or undiagnosed. The fact that CRC becomes symptomatic at advanced stages highlights the importance of early screening. The reference screening method for CRC is colonoscopy, an invasive, time-consuming procedure that requires sedation or anesthesia and is recommended from a certain age and above. The aim of this study was to build a machine learning classifier that can distinguish cancer from non-cancer samples. For this, circulating tumor cells were enumerated using flow cytometry. Their numbers were used as a training set for building an optimized SVM classifier that was subsequently used on a blind set. The SVM classifier's accuracy on the blind samples was found to be 90.0%, sensitivity was 80.0%, specificity was 100.0%, precision was 100.0% and AUC was 0.98. Finally, in order to test the generalizability of our method, we also compared the performances of different classifiers developed by various machine learning models, using over-sampling datasets generated by the SMOTE algorithm. The results showed that SVM achieved the best performances according to the validation accuracy metric. Overall, our results demonstrate that CTCs enumerated by flow cytometry can provide significant information, which can be used in machine learning algorithms to successfully discriminate between healthy and colorectal cancer patients. The clinical significance of this method could be the development of a simple, fast, non-invasive cancer screening tool based on blood CTC enumeration by flow cytometry and machine learning algorithms.

Keywords: colorectal cancer; circulating tumor cells; flow cytometry; machine learning; SVM; SMOTE

Citation: Hatzidaki, E.; Iliopoulos, A.; Papasotiriou, I. A Novel Method for Colorectal Cancer Screening Based on Circulating Tumor Cells and Machine Learning. *Entropy* **2021**, *23*, 1248. https://doi.org/10.3390/e23101248

Academic Editors: Leonidas P. Karakatsanis and Dimitrios S. Monos

Received: 12 August 2021
Accepted: 21 September 2021
Published: 25 September 2021

1. Introduction

Cancer is extremely complex and heterogeneous. It includes various processes (e.g., evading growth suppressors, resisting cell death, replicative immortality), which manifest as cancer's irregular dynamics in multi-level spatio-temporal scales. In particular, at the molecular level, a large number of interacting molecules (proteins, lipids and ions) constitute a complex network, which results in complex intracellular signaling, non-linear reaction kinetics, gene mutations and dysregulations, regulatory circuits, pathway cross-talks and others [1–4]. The processes are non-linear, and the formation of the hierarchies themselves may be discontinuous [5]. In addition, cancer has many (more than 100 distinct types) incarnations. Different categories of cancer exist such as carcinoma, sarcoma, leukemia, lymphoma and myeloma and central nervous system cancers, all depending on where carcinogenesis was initiated (e.g., skin, bone, blood, brain), as well as different types such as bladder, breast, colon and rectal, endometrial, lung, pancreatic, prostate, etc. All these types and/or categories are characterized by unique features and growth dynamics, increasing the already high levels of complexity needed to be confronted by scientists for dealing with carcinogenesis in prevention, early detection, treatment management and screening post-treatment.

Colorectal cancer (CRC) is one of the most common types of cancer, and it can have a high mortality rate if left untreated or undiagnosed. It is estimated that 10% of all annually diagnosed cancers are CRC [6]. Risk factors include both environmental and genetic ones. Apart from age, obesity, lack of exercise, smoking, alcohol consumption and dietary habits are also implicated [7]. Family history can be a contributing factor in 10–20% of CRC cases [8]. Hereditary CRC syndromes include polyps, Lynch syndrome, inflammatory bowel syndrome and type 2 diabetes. Colon cancer survival rates are poor if diagnosed late. Furthermore, the disease becomes symptomatic mainly at advanced stages. This highlights the importance of early screening. At present, guidelines do not recommend screening for CRC at ages lower than 50, unless there is a family history [9]. However, the fact that CRC incidences show an increase in younger age groups has prompted the re-consideration of the guidelines. Colonoscopy is the current reference method for CRC screening. It is an invasive procedure and requires the use of sedation or anesthesia. CRC biomarkers detected in blood could be an attractive alternative. Although a few prognostic biomarkers are known, such as carcinoembryonic antigen (CEA), microsatellite instability (MSI) and BRAF mutation, there are no diagnostic biomarkers available in clinical practice.

Despite the significant advances in the field, colorectal cancer therapy, recurrence and metastasis continue to face difficulties and new challenges due to cancer's inherent complexity. Therefore, detailed biological information (e.g., differences between cancer states and healthy states and/or between cancer subtypes) and the utilization of advanced mathematical methods are of great importance [10–14]. Specifically, in the last two decades there has been an exponential growth of Machine Learning (ML) algorithms utilized for addressing difficult healthcare challenges including complex biological abnormalities such as cancer [15–19]. ML has introduced novel biomarkers for cancer diagnosis, designed novel personalized drugs and delivered potential treatment strategies [20–23]. For achieving these targets, scientists analyze various types of input data [15], such as genomic (SNPs, mutations, microarrays), proteomic (specific protein biomarkers, 2D gel data, mass spectral analyses), clinical (histology, tumor staging, tumor size, age, weight, risk behavior, etc.), high-resolution images (which are involved almost in every cancer diagnosis), demographic, epidemiological or combinations of some of these. The analyses were based on a variety of ML techniques utilized for the development of predictive models. These include Artificial Neural Networks (ANNs), Deep Learning (DL), Bayesian Networks (Bns), Support Vector Machines (SVMs), Decision Trees (Dts) and others [16].

In this direction, [24] developed a deep learning-based method to measure the similarity between CRC tumors and cancer cell lines. The datasets considered contained copy number alterations, gene expression and point mutations. The model learns latent factors that represent clinically relevant patterns and explains the variability of molecular profiles across tumors and cell lines, providing best-matching cell lines to different cancer subtypes. In addition, [25] developed a multi-parameterized ANN to score the risk of colorectal cancer based solely on personal health data from the National Health Interview Survey (NHIS). The ANN was tested per Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) level 2a and level 2b protocols. The result showed that the ANN is comparable to current methods of scoring CRC risk (including those using biomarkers). Another example is given by [26]. In this comparative study, seven ML algorithms were evaluated in combination with six imputation methods for missing data, all trained and cross-tested with the NHIS and PLCO datasets concerning CRC. The optimal model was an ANN with Gaussian expectation-maximization imputation, which can be used as a non-invasive and cost-effective tool to screen the CRC risk in large populations effectively using only personal health data. Finally, a state-of-the-art transfer-learned deep convolutional neural network was developed recently by [27], who proposed a novel patch aggregation strategy for clinic CRC diagnosis using weakly labeled pathological whole-slide image (WSI) patches. This approach yielded promising results, often even better than most of the experienced expert pathologists when tested in diagnos-

ing CRC. For more information on CRC screening, diagnosis and treatment based on other AI applications, the interested reader can refer to a very thorough review by [28].

In this study, we built an ML classifier for discriminating colorectal cancer samples from non-cancer, healthy samples. The ML classifier was based on Support Vector Machines (SVM), which were chosen as an appropriate approach for classifying colorectal cancer and non-cancer/healthy samples. We used SVM since they are among the most commonly applied ML algorithms within the field of cancer research and more generally in computational biology [29–33], exhibiting accurate predictive performance. SVM can be used to overcome classification problems concerning datasets with small sample size, high dimensionality and nonlinearity with good generalization capability. In this direction, the SVM classifier was also compared with other classifiers, developed by methods frequently utilized in ML applications, using larger over-sampling datasets generated by the SMOTE algorithm [34].

As a dataset, we used experimental data derived from circulating tumor cells (CTCs) [35] detected by flow cytometry, which can be promising prognostic biomarkers in CRC [36]. In particular, circulating CTCs are cancer cells that are shed from the tumor and travel in blood circulation. CTCs actively leave the tumor tissue and invade the blood stream using a process known as epithelial-to-mesenchymal transition (EMT). During EMT, cancer cells lose their epithelial characteristics and acquire mesenchymal ones. This allows them to become mobile and migrate from the primary to the metastatic site [37]. Today, there is only one FDA-approved detection technique for CTCs. It relies on EpCam-positive and CD45-negative immunoselection of fixed cells. CTCs are then detected using high-resolution imaging combined with immunocytofluorescent staining [38]. The system therefore detects CTCs by counting cells positive for fluorescent signal co-localization in an image captured by a camera. However, EpCam, being an epithelial marker, limits the ability to evaluate CTCs from tumors that have no EpCam expression, or cancer cells that have undergone EMT [39]. More importantly, today's technologies for CTC determination rely mainly on traditional microscopy imaging and therefore suffer from the same limitations. Well focused images are imperative for image analysis; ideally, images should be viewed under different light sources, phase contrast, bright-field and fluorescence, and finally, there is a limitation to the pixel information a microscope can deliver [40]. On the other hand, flow cytometry is a powerful and sensitive cell analysis technique that detects fluorescent signals as cells pass one by one in front of a light source. If the cytometer is a sorter, cells can also be isolated alive and cultured for downstream analyses. We have developed a method for CTC determination in whole blood using flow cytometry. CTCs were defined as CD45-negative, CD31-negative and pan-cytokeratin-positive cells in peripheral blood cells. It was found that our method of CTC detection by flow cytometry had a sensitivity of 86.2% and specificity of 83.9% [41].

The aim of this study was, firstly, to validate our method for CTC determination, and secondly, to use these data to perform binary classification between colorectal cancer and healthy samples. The clinical significance of this method could be the development of a non-invasive cancer screening tool based on blood CTC enumeration by flow cytometry and ML.

2. Materials and Methods

2.1. Sample Collection

This study was not a clinical trial and did not include any interventions. The study was reviewed and approved by our institutional ethics committee. Informed consent was obtained from all patients. Blood samples from a total of 41 healthy individuals/non-cancer patients and 41 CRC patients were collected in sterilized 50-mL falcon tubes containing 7 mL 0.02 M EDTA as an anti-coagulant. Healthy individuals were identified as healthy/non-cancer by their physicians.

2.2. Sample Preparation

A total of 2 mL of blood was mixed with 2 mL of fetal bovine serum in 15-mL centrifuge tubes to regain the cells' shape. The samples were then centrifuged at 1200 rpm for 10 min at room temperature and the supernatant was discarded. A total of 100 µL of sample was transferred to round-bottom tubes for flow cytometry analysis.

2.3. Antibodies and Staining Procedure

Antibodies used were CD45-PE/Cy7, CD31-RPE and pan CK-PE/Cy5. Samples were fixed and permeabilized using LEUCOPERM according to the manufacturer's instructions. Briefly, first samples were stained with surface antibodies for 20 min (CD45 and CD31, 5 µg/mL each), washed with PBS and then fixed with 100 µL Leucoperm Reagent A, washed with PBS, permeabilized with 100 µL Reagent B and stained intracellularly with 5 µg/mL pan-CK antibodies for 20 min and washed again with PBS. After the last wash, cells were re-suspended in 500 µL PBS and were ready for acquisition in a Beckman Coulter FC500 cytometer.

2.4. Sample Blinding

A total of 31 healthy and 31 cancer samples were known to the investigators and were used for the training and validation of the algorithm. Twenty samples were blinded by using 5-digit codes and were used for prediction (test) analysis.

2.5. Sample Acquisition and FCS Data Analysis

Circulating tumor cells were defined as CD45-negative, CD31-negative and pan-cytokeratin-positive cell populations. Non-hematological cells were gated out using a CD45-negative selection. The endothelial cells were then removed using a CD31-negative gating selection. Tumor cells were identified by pan-CK-positive selection. Unstained samples were used as a negative control for gating. FCS Express software was used for fcs data analysis. Figure 1 shows the gating strategy in FCS Express, where the CD31-negative gate is set as a CD45-negative sub-gate, and the pan-CK-positive gate is set as a CD31-negative sub-gate.

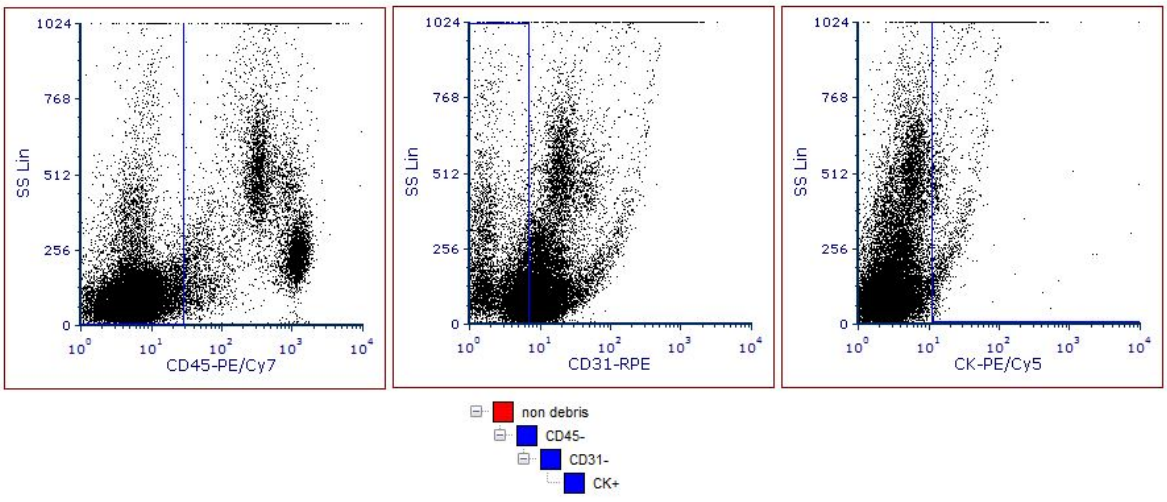


Figure 1. Gating strategy for the identification of CTCs in PBMCs. First on the left plot shows exclusion of CD45-positive cells (hematopoietic); second plot shows exclusion of CD31-positive cells (epithelial); third plot shows selection of pan-CK-positive cells.

2.6. Mathematical Analysis

2.6.1. Two-Sample Kolmogorov–Smirnov Test

The two-sample Kolmogorov–Smirnov test [42] is a nonparametric hypothesis test that evaluates the difference between the Cumulative Distribution Functions (CDFs) of the two samples over the range of x in each dataset. The two-sided test uses the maximum absolute difference between the CDFs of the distributions of the two samples. The test statistic is:

$$D^* = \underset{*}{max}(|\hat{F}_1(x) - \hat{F}_2(x)|), \tag{1}$$

where $\hat{F}_1(x)$ is the proportion of x_1 values less than or equal to x, and $\hat{F}_2(x)$ is the proportion of x_2 values less than or equal to x.

2.6.2. Wilcoxon Rank Sum Test

The Wilcoxon rank sum test is a nonparametric test for two populations [43]. In particular, this test examines the null hypothesis that two samples are drawn from continuous distributions with equal medians, against the alternative hypothesis that they are not. The test assumes that the two samples are independent. The Wilcoxon rank sum test is equivalent to the Mann–Whitney U-test, which is a nonparametric test for equality of population medians of two independent samples X and Y. Specifically, the Mann–Whitney U-test statistic, U, is the number of times a y precedes an x in an ordered arrangement of the elements in the two independent samples X and Y. It is related to the Wilcoxon rank sum statistic in the following way: If X is a sample of size n_x, then:

$$U = W - \frac{n_x(n_x + 1)}{2} \tag{2}$$

2.6.3. Support Vector Machines

In this paper, in order to solve this binary classification problem, we apply a powerful classifier, the support vector machine (SVM). SVM aims to create a decision boundary between two classes in order to predict the labels from one or more feature vectors [44,45]. This decision boundary is known as the hyperplane. Its orientation is crucial for the best separation of the closest data points from each of the classes. These closest points are called support vectors. In particular, for given a labeled training dataset:

$$(x_1, y_1), \ldots, (x_n, y_n), x_i \in R^d \wedge y_i \in (-1, 1), \tag{3}$$

where x_i is a feature vector representation and y_i is the class label (negative or positive) of a training compound i, and the optimal hyperplane can be defined as:

$$wx^T + b = 0, \tag{4}$$

where w is the weight vector, x is the input feature vector and b is the bias. In the best case scenario, w and b would satisfy the following inequalities for all elements of the training set:

$$wx_i^T + b \geq 1 \; if \; y_i = 1, \tag{5}$$

$$wx_i^T + b \leq -1 \; if \; y_i = -1. \tag{6}$$

Therefore, the objective of training an SVM model is to find the proper w and b so that the hyperplane separates the data and maximizes the margin $1/\|w\|^2$.

However, many binary classification problems do not have a simple hyperplane as a useful separating criterion. For such problems, instead of using a linear SVM classifier, we can alternatively use the kernel method. This method enables us to model higher dimensional, non-linear models, while retaining nearly all the simplicity of an SVM separating hyperplane. Specifically, the kernel method transforms the data into higher dimensional spaces to make the data separable.

In general, a kernel function is defined as:

$$G(x, y) = < f(x), f(y) >, \quad (7)$$

where G is the kernel function, x and y are n dimensional inputs and f is used to map the input from n dimensional to m dimensional space. Finally, the term $< x, y >$ denotes the dot product. This class of functions includes polynomials and Radial Basis Function (RBF). In particular, polynomials (e.g., linear, quadratic, cubic) are defined as:

$$G(x, y) = (1 + x'y)^p, \quad (8)$$

where p is some positive integer, while RBF kernel is defined as:

$$G(x, y) = exp(-\|x - y\|^2). \quad (9)$$

Of course, the choice of kernel function, among other parameters, can greatly influence the performance (e.g., reduce or increase the classification probability error) of an SVM model. One can choose between the available kernels through trials and, depending on the nature of the problem, select the best one. One way to find the optimal kernel in a statistically rigorous fashion is by using cross-validation.

Particularly, cross-validation is a procedure used to avoid under- and overfitting [46]. It is a process in which the dataset is randomly partitioned into a training and a test set. In this paper, we used a k-fold cross validation procedure. In particular, this method splits the data randomly into k equal (or almost equal) parts. Then, the algorithm runs k times, using k-1 of the parts as a training set and the remaining part as a test set. Each time the algorithm runs, a different test set is used, so that over the k runs of the algorithm, all the instances in the dataset are used as a test set. The success of the algorithm is the sum of the correct classification over each of the runs. However, even cross-validation can overestimate the prospective performance of ML methods. Therefore, we also conducted a truly blind test in order to demonstrate the prospective capabilities of our cross-validated model [20].

2.6.4. Comparison between Different Classifiers

One drawback of this study is the relatively small dataset, which can lead to biased models that are not generalizable. Therefore, in order to further test the generalizability of our method, we also compared the performances of many classifiers, in addition to the SVM classifier, developed by methods frequently utilized in ML applications. In particular, we developed optimizable models from classification trees [47], discriminant analysis [48], logistic regression [49], naïve Bayes [50], k-nearest neighbor (kNN) [51] and ensemble methods, including boosted trees, bagged trees (random forest), subspace discriminant, subspace kNN and RUSBoosted trees [50–53]. The hyperparameter search range for the different classifiers was: (a) Classification trees: *Maximum number of splits*: 1–163, *Split criterion*: Gini's diversity index, Maximum deviance reduction. (b) Discriminant Analysis: Linear, Quadratic, Diagonal Linear, Diagonal Quadratic. (c) Naïve Bayes: *Distribution names*: Gaussian, Kernel, *Kernel type*: Gaussian, Box, Epanechnikov, Triangle. (d) SVM: *Multiclass method*: One-vs.-All, One-vs.-One, *Box constraint level*: 0.001–1000, *Kernel scale*: 0.001–1000, *Kernel function*: Gaussian, Linear, Quadratic, Cubic, *Standardize data*: true, false. (e) kNN: *Number of neighbors*: 1–82, *Distance metric*: City block, Chebyshev, Correlation, Cosine, Euclidean, Hamming, Jaccard, Mahalanobis, Minkowski (cubic), Spearman, Distance weight: Equal, Inverse, Squared inverse, *Standardize data*: true, false. (f) Ensemble: *Method*: Bag, GentleBoost, LogitBoost, AdaBoost, RUSBoost, *Number of learners*: 10–500, *Learning rate*: 0.001–1, *Maximum number of splits*: 1–163, *Number of predictors to sample*: 1–2.

In order to perform the comparison, since the dataset is small, we updated the original dataset, generating over-sampling datasets. Therefore, we tested the performance of all the above classifiers using the over-sampling datasets. In order to create the over-sampling datasets, we used a robust method named Synthetic Minority Over-sampling Technique

(SMOTE) [34,54]. This is an over-sampling approach that creates synthetic minority class samples. This technique is widely used and performs better than simple over-sampling. In particular, the SMOTE samples are linear combinations of two similar samples from the minority class (x and x^R) and are defined as:

$$s = x + u * (x^R - x), \tag{10}$$

where u is randomly chosen from $U(0, 1)$ and differs for each SMOTE sample. This guarantees that a SMOTE sample lies on the line joining the two original samples used to generate it [34,54]. For more information on SMOTE and its updates, the interested reader can refer to [55].

2.6.5. Performance Measures for Binary Classifiers

The performance analysis of the model can be measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC). They are all based on true positives (TP, correctly predicted positive (cancer) samples); true negatives (TN, correctly predicted negative (non-cancer/healthy) samples), false positives (FP, normal samples wrongly predicted as being cancer samples) and false negatives (FN, cancer samples wrongly predicted as non-cancer/healthy) [56].

In particular, Accuracy is the percentage of correctly predicted samples, and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \tag{11}$$

and is used for estimating the overall performance of the classifier.

Sensitivity or True Positive Rate (TPR) is the percentage of samples correctly predicted as cancer samples, and is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \times 100\%. \tag{12}$$

The opposite of sensitivity is called False Negative Rate (FNR) or Miss Rate and is equal to FNR = 1 − TPR.

Specificity or True Negative Rate (TNR) is the percentage of samples correctly predicted as non-cancer/healthy samples, and is defined as:

$$Specificity = \frac{TN}{TN + FP} \times 100\%. \tag{13}$$

Precision or Positive Predictive Value (PPV) is the percentage of samples correctly predicted as cancer from all positive predictions, and is defined as:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{14}$$

The opposite of Precision is False Discovery Rate (FDR) equal to FDR = 1 − PPV.

Area under the Curve (AUC) is a measure of the model's overall performance. AUC for binary classification [56] is given by:

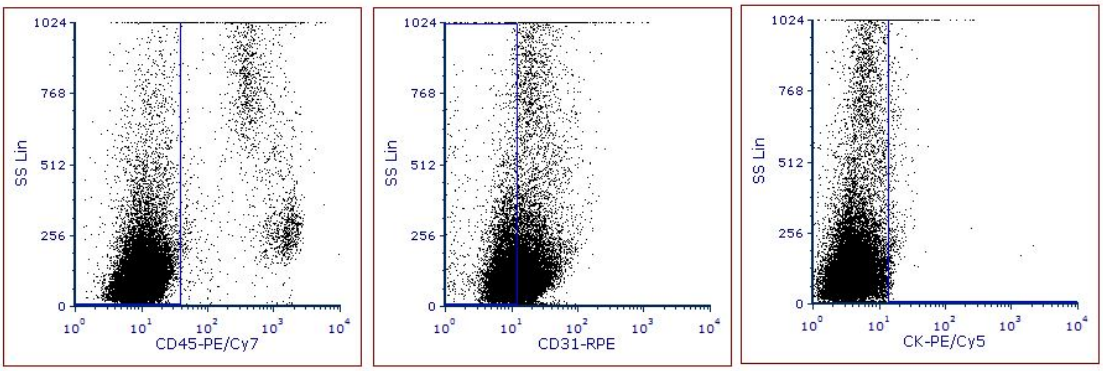
$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{15}$$

The maximum AUC is 1, which corresponds to a perfect classifier, while for a classifier that randomly assigns observations to classes, AUC = 0.5. Larger AUC values indicate better classifier performance. A rough rule of thumb is that the accuracy of tests with AUCs between 0.50 and 0.70 is low; between 0.70 and 0.90, the accuracy is moderate; and it is high for AUCs over 0.90 [57].

AUC is the primary statistic we obtain from a Receiver Operating Characteristics (ROC) curve [58], which plots the tradeoffs between sensitivity and 1-specificity. In particular, ROC graphs are two-dimensional graphs in which sensitivity (TPR) is plotted on the Y-axis and FPR (1-TNR) on the X-axis, for different thresholds of the classifier output. They are useful for organizing classifiers and visualizing their performance. In such a graph, the point (0, 1) represents perfect classification.

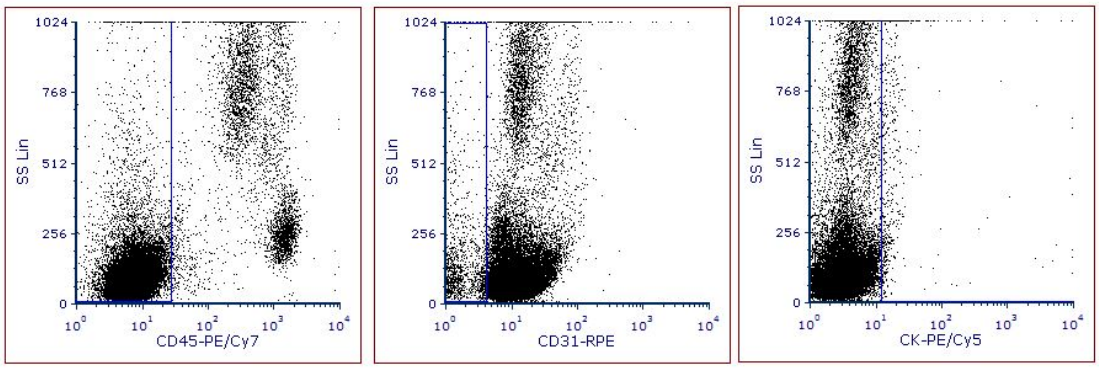
3. Results

After data acquisition, CTCs from the 31 cancer and 31 healthy samples were calculated using FCS Express. Figure 2 shows a healthy sample analysis. No CTCs were found as it is denoted by the column # of Events (number of Events). Figure 3 shows a cancer sample. Five CTCs were found using the same method.



Gate	# of Events	X Geometric Mean	Y Geometric Mean	% of gated cells	% of all cells
None	50000	3.94	n/a	100.00	100.00
CD45-	3228	5.88	316.03	6.46	6.46
CD31-	730	3.10	428.37	1.46	1.46
CK+	0	n/a	n/a	0.00	0.00

Figure 2. Representative analysis of a healthy sample. No cells were found to be CD45-/CD31-/CK+ as denoted in the column # of Events (number of Events).



Gate	# of Events	X Geometric Mean	Y Geometric Mean	X Median	Y Median	% of gated cells	% of all cells
None	50000	3.34	n/a	3.25	87.00	100.00	100.00
CD45-	2150	3.10	254.57	3.25	231.50	4.30	4.30
CD31-	797	1.41	264.22	1.18	251.00	1.59	1.59
CK+	5	38.47	291.29	47.83	274.00	0.01	0.01

Figure 3. Representative analysis of a cancer patient sample. Five cells were found to be CD45-/CD31-/CK+ as denoted in the column # of Events (number of Events).

3.1. Statistical Tests

Before we built the classifier, we tested for differences between the cancer and non-cancer/healthy distributions and their medians. This was achieved using two non-parametric hypothesis tests, namely the two-sample Kolmogorov–Smirnov (KS) test and the Wilcoxon rank sum (WRS) test. Both tests revealed significant statistical differences in terms of distributions and medians. In particular, the KS test rejected the null hypothesis, namely that the data are from the same continuous distribution with a p-value equal to $1.85^{-6} \ll 0.05$. In addition, the WRS test rejected the null hypothesis, namely that the data are samples from continuous distributions with equal medians with a p-value equal to $2.86^{-7} \ll 0.05$. Therefore, the cancer and non-cancer/healthy samples have significant statistical differences, both in terms of their distributions as well as their medians. This information indicates that an efficient classifier can be built based on this dataset. All computations for the statistical tests were performed in MATLAB [59], using the Statistics and Machine Learning Toolbox.

3.2. SVM Classifier

We used a 5-fold cross validation and MATLAB's Bayesian Optimization function *bayesopt* to find the best (optimized) classification SVM model. In particular, the hyperparameter search range included *box constraint level*: 0.001–1000, *kernel_scale*: 0.001–1000 and *kernel_function*: Gaussian, linear, quadratic, cubic. The optimized SVM model consisted of a quadratic kernel function (*scale* = 1, *order* = 3) and *box constraint level* equal to 3.0685. The data were standardized.

The results of the optimized SVM are shown in the confusion matrix (Figure 4). In particular, in this figure the total number of observations in each cell is presented (central panel). The rows correspond to the true class, and the columns correspond to the predicted class. Diagonal and off-diagonal cells correspond to correctly and incorrectly classified observations, respectively. As it can be seen in this panel, considering the cancer samples as positives, the true positives (TP) were found equal to 23, true negatives (TN) = 28, false positives (FP) = 3 and false negatives (FN) = 8. Based on these values, we estimated the performance measures using Equations (11)–(14). In particular, the accuracy of the classifier was found to be $51/62 \times 100\% = 82.3\%$.

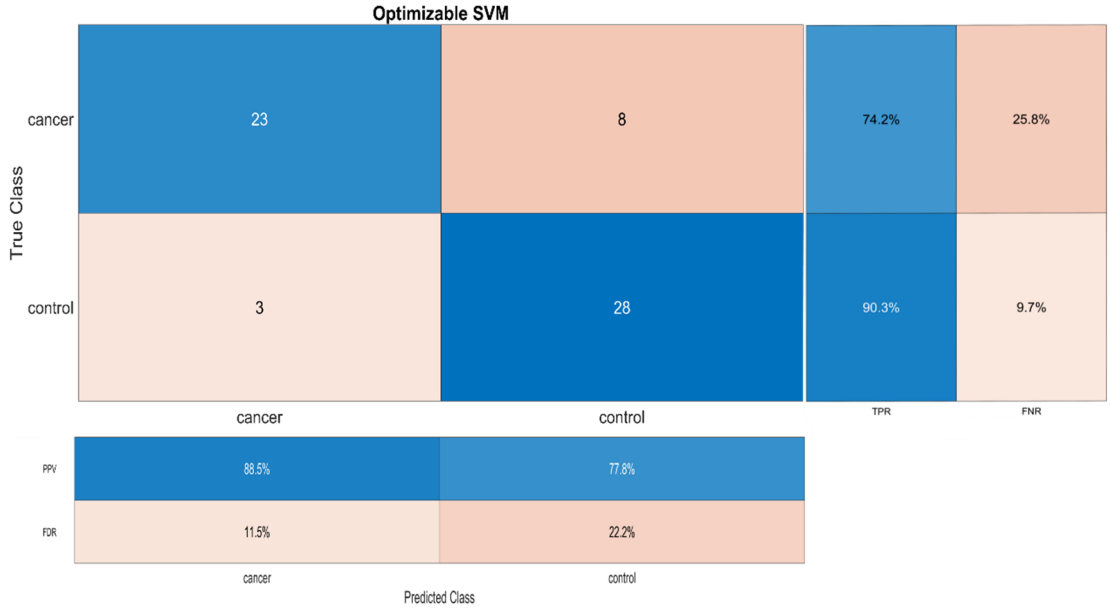


Figure 4. Confusion matrix for the optimized SVM classifier. In the big panel, the rows correspond to the true class, and the columns correspond to the predicted class. Diagonal and off-diagonal cells correspond to correctly and incorrectly classified observations, respectively. The sensitivity (TPR) is shown in the right panel, first column and miss rate (FNR) in the right panel, second column. Additionally, the precision (PPV) is shown in bottom panel, first row and false discovery rate (FDR) in the bottom panel, second row.

In addition, in the right panel, the row summary displays the percentages of correctly and incorrectly classified observations for each true class. This panel shows that the sensitivity (TPR) is equal to $23/31 \times 100\% = 74.2\%$ and the miss rate (FNR) is equal to $8/31 \times 100\% = 25.8\%$. This means that 23 samples were correctly classified as cancer samples and eight samples were falsely classified as non-cancer/healthy (false negatives) out of 31 cancer samples. Similarly, the specificity (TNR) is $28/31 \times 100\% = 90.3\%$, while $3/31 \times 100\% = 9.7\%$ were falsely classified as cancer samples.

Finally, the bottom panel displays a summary of the percentages of correctly and incorrectly classified observations for each predicted class. Specifically, this panel shows the results concerning the precision (PPV) and False Discovery Rate (FDR) of the optimized SVM model. As it is shown, PPV is equal to $23/26 \times 100\% = 88.5\%$ for the cancer samples and $28/36 \times 100\% = 77.8\%$ for the non-cancer/healthy samples. The FDR is $100\% - 88.5\% = 11.5\%$ for the cancer samples and $100\% - 77.8\% = 22.2\%$ for the non-cancer/healthy samples, respectively.

In Figure 5, the ROC curve for the optimized SVM is shown. In the same figure, the AUC, the optimal point for the current classifier (orange dot) and the ROC curve for a random classifier (diagonal red dotted line) are also shown. The random classifier identifies an equal amount of positives and negatives correctly. Therefore, the AUC for a random classifier is 0.5. Any classifier that appears in the lower right triangle performs worse than random guessing. As it can be seen, in Figure 5, the AUC of the optimized classifier is $0.85 >> 0.5$, indicating a moderate-to-high accuracy classifier [57].

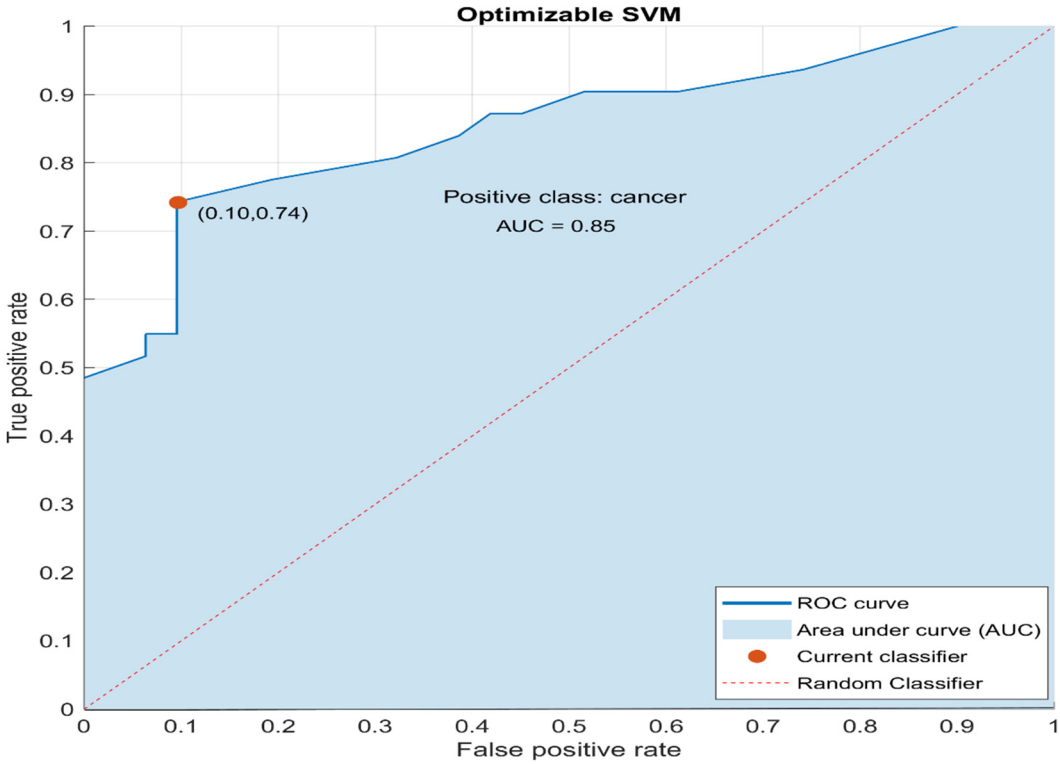


Figure 5. Receiver operating characteristic (ROC) curve for the optimized SVM classifier. The AUC is equal to 0.85. The optimal operation point for the current classifier is also shown (orange dot). The ROC curve of a random classifier is also shown (dotted red line).

In addition, as it is shown, the optimal point (the point that will result in the lowest number of overall errors: FN + FP) for the classifier is found for TPR = 0.74 and FPR = 0.10, near the Y-axis. Classifiers appearing on the left-hand side of an ROC graph are rather "conservative", namely they make positive classifications only with strong evidence, making few false positive errors [58].

3.3. Blind Set

In order to further test the performance of the optimized SVM classifier, we examined its performance in a totally blind set. As mentioned, this set includes 10 cancer and 10 non-cancer/healthy samples. The results are summarized in Figure 6 and reveal that TP = 8, TN = 10, FP = 0 and FN = 2. Therefore, the accuracy in the blind set was found to be $18/20 \times 100\% = 90.0\%$. Moreover, the sensitivity (TPR) was found equal to 80.0%, the miss rate equal to 20.0%, the specificity equal to 100% and the precision equal to 100.0%. Finally, the AUC for the blind set was found equal to 0.98. All computations were performed in MATLAB [59] using the Statistics and Machine Learning Toolbox.

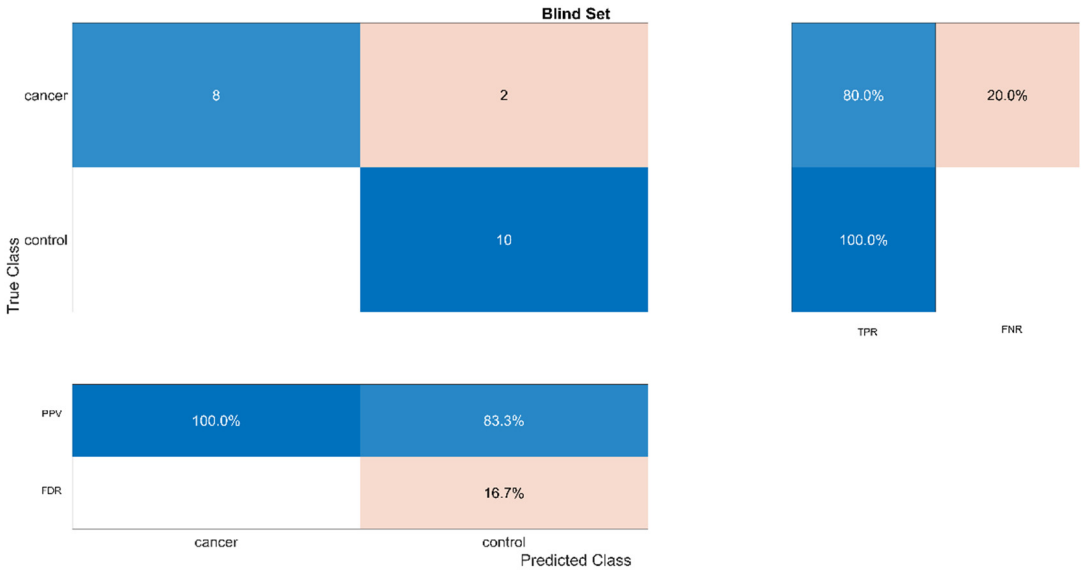


Figure 6. Similar to Figure 5. Confusion matrix of the optimized classifier for the blind set.

Overall, the results demonstrate that the SVM classifier, based on CTCs enumerated by flow cytometry, can successfully discriminate between healthy and colorectal cancer patients with high values of performance measures.

3.4. Comparison between Different Classifiers

In order to further test the generality of our results, we compared different classifiers developed by models frequently utilized in ML applications. All computations were performed in MATLAB [59] using the Statistics and Machine Learning Toolbox. The classifiers were developed on SMOTE-generated over-sampling datasets, using the MATLAB package *smote* [60]. This function synthesizes new observations based on existing (input) data and a K-nearest neighbor approach.

In addition, for the generation of the over-sampling datasets, both the training as well as the blind sets were taken into consideration. In particular, we generated six new datasets by varying both the amount of over-sampling (N) as well as the number of considered nearest neighbors (K). We considered N = 1 (2-fold observations), N = 3 (3-fold observations) and N = 10 (10-fold observations), and K = 5, 10, 20, 30. Therefore, since the initial set (including both the training and blind sets) consists of 41 colon cancer samples and 41 healthy samples, the resulting datasets contained: D1(N = 1, K = 5) = 164 samples, D2(N = 1, K = 10) = 164 samples, D3(N = 3, K = 10) = 328 samples, D4(N = 3, K = 20) = 328 samples, D5(N = 10, K = 20) = 902 samples and D6(N = 10, K = 30) = 902 samples.

The results of the performances of the classifiers are shown in Tables 1 and 2. Specifically, in Table 1 we show the validation accuracy of the optimized models (the optimization of the models was performed using a 5-fold cross validation and MATLAB's Bayesian Optimization function *bayesopt*). As it can be seen, all models achieved high validation accuracies, above 84%, while the differences are not so big, even for the linear benchmark

method of logistic regression. This is expected, since we only have one feature as an input. The highest validation accuracies (%) were found for the classifiers based on SVM (D1—86.0, D2—89.0, D3—89.6, D4—87.2), for on ensemble classifiers (D1—86.0, D4—87.2, D6—88.2) and for on classification trees (D1—86.0, D5—87.6).

Table 1. Validation accuracy of the optimized models, for datasets generated using SMOTE technique, using various parameters such as N = 1, 3, 10 and K = 5, 10, 20, 30.

	D1	**D2**	**D3**	**D4**	**D5**	**D6**
	N = 1	**N = 1**	**N = 3**	**N = 3**	**N = 10**	**N = 10**
	K = 5	**K = 10**	**K = 10**	**K = 20**	**K = 20**	**K = 30**
Trees	86.0	88.4	88.4	85.1	87.6	88.1
Discriminant	84.1	87.2	85.7	86.0	86.7	88.0
Logistic Regression	84.1	86.0	86.0	86.6	87.1	87.8
Naïve Bayes	84.1	86.6	85.7	86.0	86.9	88.1
SVM	86.0	89.0	89.6	87.2	87.3	88.0
KNN	85.4	87.8	89.6	85.7	84.4	87.9
Ensemble	86.0	88.4	88.4	87.2	86.9	88.2

Table 2. Estimated Area Under Curve (AUC) of the optimized models, for datasets generated using SMOTE technique, using various parameters such as N = 1, 3, 10 and K = 5, 10, 20, 30.

	D1	**D2**	**D3**	**D4**	**D5**	**D6**
	N = 1	**N = 1**	**N = 3**	**N = 3**	**N = 10**	**N = 10**
	K = 5	**K = 10**	**K = 10**	**K = 20**	**K = 20**	**K = 30**
Trees	0.89	0.88	0.92	0.88	0.94	0.86
Discriminant	0.89	0.88	0.91	0.92	0.94	0.93
Logistic Regression	0.89	0.88	0.91	0.92	0.94	0.95
Naïve Bayes	0.88	0.88	0.89	0.92	0.94	0.94
SVM	0.84	0.89	0.88	0.89	0.94	0.95
KNN	0.89	0.88	0.91	0.92	0.94	0.92
Ensemble	0.89	0.89	0.92	0.92	0.94	0.94

We also estimated the AUC for each of the optimized models and the results are presented in Table 2. As it can be seen, all models achieved high values of AUC ($\geq$0.84). In this case, the highest values were achieved by ensemble methods and in particular by Gentle Adaptive Boosting (GentleBoost) [61] for four datasets (D1—0.89, D2—0.89, D4—0.92, D5—0.94), and the Bootstrap Aggregation and Random Forest (Bag) [62] for one dataset (D3—0.92). Logistic regression also yielded the highest AUC for four datasets (D1—0.89, D4—0.92, D5—0.94, D6—0.95), whereas the other classifiers attained the highest performances for fewer datasets.

Taking into account the results concerning validation accuracies as well as AUC, it can be concluded that, even though all ML classifiers yielded high performances, SVM performed better according to the validation accuracy metric, while ensemble methods performed better according to the AUC metric. However, compared to AUC, accuracy is simpler and easier to interpret, while it is mostly used for evaluating supervised binary classifiers with balanced classes, taking into account both true positive as well as true negative predictions. Therefore, based on accuracy results, in this study, we chose SVM for developing efficient and robust ML classifiers.

4. Discussion

In the present study, we developed an SVM classifier for performing binary classification between colorectal cancer and non-cancer/healthy samples. The main feature used for the classification is the number of CTCs from cancer and non-cancer/healthy samples, as obtained from flow cytometry. In this study, 31 colorectal cancer and 31 non-cancer/healthy samples were used for the development of the SVM classifier. In addition, the SVM classifier was tested in a blind test set, which included 10 cancer samples and 10 non-cancer/healthy samples. Finally, in order to further test the efficiency and generalizability of the proposed method, we generated various over-sampling datasets by applying the SMOTE algorithm and used these datasets in order to develop and compare various ML classifiers.

The results of this study revealed the efficiency of the developed SVM classifier both on the training set as well as on the blind set. In particular, for the training set, the performance measures of the SVM classifier were found to be: accuracy equal to 82.3%, sensitivity (TPR) equal to 74.2%, miss rate (FNR) equal to 25.8%, specificity (TNR) equal to 90.3%, precision (PPV) equal to 88.5% and AUC equal to 0.85. For the blind set, the performance measures of the SVM classifier were found to be: accuracy equal to 90.0%, sensitivity (TPR) equal to 80.0%, miss rate (FNR) equal to 20.0%, specificity (TNR) equal to 100.0%, precision (PPV) equal to 100.0% and AUC equal to 0.98.

One drawback of this study was the relatively small dataset, which can result in misclassifications, while the estimators may produce unstable and biased models, which can fail to generalize efficiently. However, the analysis of over-sampling SMOTE-generated datasets revealed that ML classifiers can also be effective for much bigger (up to 10-fold) datasets. In particular, the estimation of the performance measures of the optimized classifiers showed that all classifiers exhibited very good performances, yielding values above 0.84 for validation accuracy and above 0.84 for AUC. Additionally, SVM performed better according to the validation accuracy metric, while ensemble methods performed better according to the AUC metric. Considering accuracy as a more relevant metric for this supervised binary with balanced classes study, SVM was the selected method.

Therefore, as the results of this study demonstrate, the drawback of the small dataset size is surpassed by the dataset quality, namely the careful feature selection (e.g., CTCs), which provides significant information for the development of effective ML classifiers. In particular, our results indicate that flow cytometry, using the gating strategy described, can be a valuable tool for CTC enumeration with high sensitivity and specificity. In addition to the accuracy of the method, other advantages are also present. Additional markers can also be studied. Immunophenotyping CTCs, that is, the determination of the expression of markers related to steaminess or metastasis, could provide useful clinical information that can aid in cancer prognosis and/or treatment decisions. Additionally, using flow cytometry and sorting, CTCs can be isolated alive and cultured for downstream applications.

Overall, the results show that CTCs enumerated by flow cytometry can provide significant information, which when "fed" into ML algorithms can successfully discriminate between non-cancer/healthy and colorectal cancer patient subjects. Even though the results seem promising, more experiments have to take place in order to obtain larger datasets, while the exploitation of more sophisticated classification techniques is needed to verify and extend the results of this study. ML algorithms are not static products, and can continue to change and improve even once deployed, as new training data become available. However, these issues will be addressed in following studies. In conclusion, the results of this study are promising towards the development of a simple, fast and non-invasive screening method for cancer, using CTC enumeration by flow cytometry from blood samples and machine learning.

Author Contributions: Conceptualization, E.H. and I.P.; methodology, E.H., A.I. and I.P.; software, A.I., E.H. and I.P.; writing—original draft preparation, E.H. and A.I.; writing—review and editing, I.P; supervision, I.P.; project administration, I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are not publicly available due to containing personal information that could compromise participant privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
2. Butcher, E.C.; Berg, E.; Kunkel, E.J. Systems biology in drug discovery. *Nat. Biotechnol.* **2004**, *22*, 1253–1259. [CrossRef]
3. Hornberg, J.J.; Bruggeman, F.; Westerhoff, H.V.; Lankelma, J. Cancer: A Systems Biology disease. *Biosystems* **2006**, *83*, 81–90. [CrossRef]
4. Grizzi, F.; Chiriva-Internati, M. Cancer: Looking for simplicity and finding complexity. *Cancer Cell Int.* **2006**, *6*, 4. [CrossRef] [PubMed]
5. Moore, N.M.; Kuhn, N.Z.; Hanlon, S.E.; Lee, J.S.H.; Nagahara, L.A. De-convoluting cancer's complexity: Using a 'physical sciences lens' to provide a different (clearer) perspective of cancer. *Phys. Biol.* **2011**, *8*, 010302. [CrossRef] [PubMed]
6. Bray, F.; Me, J.F.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
7. Dekker, E.; Tanis, P.J.; Vleugels, J.L.A.; Kasi, P.M.; Wallace, M.B. Colorectal cancer. *Lancet* **2019**, *394*, 1467–1480. [CrossRef]
8. Henrikson, N.B.; Webber, E.M.; Goddard, K.A.; Scrol, A.; Piper, M.; Williams, M.S.; Zallen, D.T.; Calonge, N.; Ganiats, T.G.; Msc, A.C.J.J.; et al. Family history and the natural history of colorectal cancer: Systematic review. *Genet. Med.* **2015**, *17*, 702–712. [CrossRef]
9. Qaseem, A.; Crandall, C.J.; Mustafa, R.A.; Hicks, L.A.; Wilt, T.J. Clinical Guidelines Committee of the American College of Physicians. Screening for Colorectal Cancer in Asymptomatic Average-Risk Adults: A Guidance Statement from the American College of Physicians. *Ann. Intern. Med.* **2019**, *171*, 643–654. [CrossRef]
10. Gentles, A.J.; Gallahan, D. Systems Biology: Confronting the Complexity of Cancer. *Cancer Res.* **2011**, *71*, 5961–5964. [CrossRef]
11. Biemar, F.; Foti, M. Global progress against cancer—Challenges and opportunities. *Cancer Biol. Med.* **2013**, *10*, 183–186.
12. Cagan, R.; Meyer, P. Rethinking cancer: Current challenges and opportunities in cancer research. *Dis. Model. Mech.* **2017**, *10*, 349–352. [CrossRef]
13. Iliopoulos, A.; Beis, G.; Apostolou, P.; Papasotiriou, I. Complex Networks, Gene Expression and Cancer Complexity: A Brief Review of Methodology and Applications. *Curr. Bioinform.* **2020**, *15*, 629–655. [CrossRef]
14. Karakatsanis, L.P.; Pavlos, E.G.; Tsoulouhas, G.; Stamokostas, G.L.; Mosbruger, T.; Duke, J.L.; Pavlos, G.P.; Monos, D.S. Spatial constrains and information content of sub-genomic regions of the human genome. *iScience* **2021**, *24*, 102048. [CrossRef]
15. Cruz, J.A.; Wishart, D.S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Inform.* **2006**, *2*, 59–77. [CrossRef]
16. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef] [PubMed]
17. Munir, K.; Elahi, H.; Ayub, A.; Frezza, F.; Rizzi, A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. *Cancers* **2019**, *11*, 1235. [CrossRef] [PubMed]
18. Apostolou, P.; Iliopoulos, A.C.; Parsonidis, P.; Papasotiriou, I. Gene expression profiling as a potential predictor between normal and cancer samples in gastrointestinal carcinoma. *Oncotarget* **2019**, *10*, 3328–3338. [CrossRef] [PubMed]
19. Iqbal, M.J.; Javed, Z.; Sadia, H.; Qureshi, I.A.; Irshad, A.; Ahmed, R.; Malik, K.; Raza, S.; Abbas, A.; Pezzani, R.; et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell Int.* **2021**, *21*, 1–11. [CrossRef]
20. Menden, M.P.; Iorio, F.; Garnett, M.; McDermott, U.; Benes, C.H.; Ballester, P.J.; Saez-Rodriguez, J. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE* **2013**, *8*, e61318. [CrossRef]
21. Bashiri, A.; Ghazisaeedi, M.; Safdari, R.; Shahmoradi, L.; Ehtesham, H. Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iran. J. Public Health* **2017**, *46*, 165–172.
22. De Silva, D.; Ranasinghe, W.; Bandaragoda, T.; Adikari, A.; Mills, N.; Iddamalgoda, L.; Alahakoon, D.; Lawrentschuk, N.; Persad, R.; Osipov, E.; et al. Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. *PLoS ONE* **2018**, *13*, e0205855. [CrossRef]
23. Levine, A.B.; Schlosser, C.; Grewal, J.; Coope, R.; Jones, S.; Yip, S. Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. *Trends Cancer* **2019**, *5*, 157–169. [CrossRef] [PubMed]
24. Ronen, J.; Hayat, S.; Akalin, A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* **2019**, *2*, e201900517. [CrossRef] [PubMed]
25. Nartowt, B.J.; Hart, G.R.; Roffman, D.A.; Llor, X.; Ali, I.; Muhammad, W.; Liang, Y.; Deng, J. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLoS ONE* **2019**, *14*, e0221421. [CrossRef] [PubMed]
26. Nartowt, B.J.; Hart, G.R.; Muhammad, W.; Liang, Y.; Stark, G.F.; Deng, J. Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. *Front. Big Data* **2020**, *3*, 6. [CrossRef] [PubMed]

27. Wang, K.S.; Yu, G.; Xu, C.; Meng, X.H.; Zhou, J.; Zheng, C.; Deng, Z.; Shang, L.; Liu, R.; Su, S.; et al. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. *BMC Med.* **2021**, *19*, 76. [CrossRef]
28. Mitsala, A.; Tsalikidis, C.; Pitiakoudis, M.; Simopoulos, C.; Tsaroucha, A. Artificial Intelligence in Colorectal Cancer Screening, Diagnosis and Treatment. A New Era. *Curr. Oncol.* **2021**, *28*, 1581–1607. [CrossRef] [PubMed]
29. Chu, F.; Wang, L. Applications of support vector machines to cancer classification with microarray data. *Int. J. Neural Syst.* **2005**, *15*, 475–484. [CrossRef]
30. Zhang, B.; Liang, X.; Gao, H.; Ye, L.; Wang, Y. Models of logistic regression analysis, support vector machine, and back-propagation neural network based on serum tumor markers in colorectal cancer diagnosis. *Genet. Mol. Res.* **2016**, *15*. [CrossRef]
31. Aziz, M.; Hussein, M.A.; Gabere, M.N. Filtered selection coupled with support vector machines generate a functionally relevant prediction model for colorectal cancer. *OncoTargets Ther.* **2016**, *9*, 3313–3325. [CrossRef] [PubMed]
32. Gao, L.; Ye, M.; Wu, C. Cancer Classification Based on Support Vector Machine Optimized by Particle Swarm Optimization and Artificial Bee Colony. *Molecules* **2017**, *22*, 2086. [CrossRef] [PubMed]
33. Huang, S.; Cai, N.; Pacheco, P.P.; Narandes, S.; Wang, Y.; Xu, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51. [CrossRef]
34. Chawla, N.V.; Bowyer, K.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
35. Wang, W.-C.; Zhang, X.-F.; Peng, J.; Li, X.-F.; Wang, A.-L.; Bie, Y.-Q.; Shi, L.-H.; Lin, M.-B. Survival Mechanisms and Influence Factors of Circulating Tumor Cells. *BioMed Res. Int.* **2018**, *2018*, 6304701. [CrossRef]
36. Veyrune, L.; Naumann, D.; Christou, N. Circulating Tumour Cells as Prognostic Biomarkers in Colorectal Cancer: A Systematic Review. *Int. J. Mol. Sci.* **2021**, *22*, 3437. [CrossRef] [PubMed]
37. Ribatti, D.; Tamma, R.; Annese, T. Epithelial-Mesenchymal Transition in Cancer: A Historical Overview. *Transl. Oncol.* **2020**, *13*, 100773. [CrossRef]
38. Cabel, L.; Proudhon, C.; Gortais, H.; Loirat, D.; Coussy, F.; Pierga, J.-Y.; Bidard, F.-C. Circulating tumor cells: Clinical validity and utility. *Int. J. Clin. Oncol.* **2017**, *22*, 421–430. [CrossRef]
39. Gorges, T.M.; Tinhofer, I.; Drosch, M.; Röse, L.; Zollner, T.M.; Krahn, T.; von Ahsen, O. Circulating tumour cells escape from EpCAM-based detection due to epithelial-to-mesenchymal transition. *BMC Cancer* **2012**, *16*, 178. [CrossRef]
40. Agarwal, A.; Balic, M.; El-Ashry, D.; Cote, R.J. Circulating Tumor Cells: Strategies for Capture, Analyses, and Propagation. *Cancer J.* **2018**, *24*, 70–77. [CrossRef]
41. Papasotiriou, I.; Chatziioannou, M.; Pessiou, K.; Retsas, I.; Dafouli, G.; Kyriazopoulou, A.; Toloudi, M.; Kaliara, I.; Vlachou, I.; Kourtidou, E.; et al. Detection of Circulating Tumor Cells in Patients with Breast, Prostate, Pancreatic, Colon and Melanoma Cancer: A Blinded Comparative Study Using Healthy Donors. *J. Cancer Ther.* **2015**, *6*, 543–553. [CrossRef]
42. Marsaglia, G.; Tsang, W.W.; Wang, J. Evaluating Kolmogorov's Distribution. *J. Stat. Softw.* **2003**, *8*, 1–4. [CrossRef]
43. Whitley, E.; Ball, J. Statistics review 6: Nonparametric methods. *Crit. Care* **2002**, *6*, 509–513. [CrossRef]
44. Vapnik, V. Pattern recognition using generalized portrait method. *Autom. Remote Control* **1963**, *24*, 774–780.
45. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
46. Mitchell, T. *Machine Learning*; McGraw Hill: New York, NY, USA, 1997.
47. Krzywinski, M.; Altman, N. Classification and regression trees. *Nat. Meth.* **2017**, *14*, 757–758. [CrossRef]
48. Hardle, W.; Simar, L. *Applied Multivariate Statistical Analysis*; Springer: Berlin, Germany, 2015.
49. LaValley, M.P. Logistic Regression. *Circulation* **2008**, *117*, 2395–2399. [CrossRef]
50. Hastie, T.; Tibshirani, R.; Friendman, J. *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*; Springer: Berlin, Germany, 2013.
51. Abu Alfeilat, H.A.; Hassanat, A.; Lasassmeh, O.; Altarawneh, A.S.A.; Alhasanat, M.B.; Salman, H.S.E.; Prasath, S. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data* **2019**, *7*, 221–248. [CrossRef] [PubMed]
52. Opitz, D.; Maclin, R. Popular Ensemble Methods: An empirical Study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198. [CrossRef]
53. Yang, P.; Yang, Y.H.; Zhou, B.B.; Zomaya, A.Y. A Review of Ensemble Methods in Bioinformatics. *Curr. Bioinform.* **2010**, *5*, 296–308. [CrossRef]
54. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106. [CrossRef] [PubMed]
55. Fernández, A.; García, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
56. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
57. Streiner, D.L.; Cairney, J. What's Under the ROC? An Introduction to Receiver Operating Characteristics Curves. *Can. J. Psychiatry* **2007**, *52*, 121–128. [CrossRef] [PubMed]
58. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
59. MATLAB. *Statistics and Machine Learning Toolbox*; The MathWorks, Inc.: Natick, MA, USA, 2021.
60. Larsen, B.S. Synthetic Minority Over-Sampling Technique (SMOTE). 2021. Available online: https://github.com/dkbsl/matlab_smote/releases/tag/1.0 (accessed on 1 September 2021).

61. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
62. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]