



Data and text mining

CancerEMC: frontline non-invasive cancer screening from circulating protein biomarkers and mutations in cell-free DNA

Saifur Rahaman ¹, Xiangtao Li², Jun Yu³ and Ka-Chun Wong ^{1,*}

¹Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, ²School of Artificial Intelligence, Jilin University, Changchun, Jilin, China and ³Institute of Digestive Diseases and The Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong SAR

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on August 27, 2020; revised on December 19, 2020; editorial decision on January 18, 2021; accepted on January 20, 2021

Abstract

Motivation: The early detection of cancer through accessible blood tests can foster early patient interventions. Although there are developments in cancer detection from cell-free DNA (cfDNA), its accuracy remains speculative. Given its central importance with broad impacts, we aspire to address the challenge.

Method: A bagging Ensemble Meta Classifier (CancerEMC) is proposed for early cancer detection based on circulating protein biomarkers and mutations in cfDNA from blood. CancerEMC is generally designed for both binary cancer detection and multi-class cancer type localization. It can address the class imbalance problem in multi-analyte blood test data based on robust oversampling and adaptive synthesis techniques.

Results: Based on the clinical blood test data, we observe that the proposed CancerEMC has outperformed other algorithms and state-of-the-arts studies (including CancerSEEK) for cancer detection. The results reveal that our proposed method (i.e. CancerEMC) can achieve the best performance result for both binary cancer classification with 99.17% accuracy (AUC=0.999) and localized multiple cancer detection with 74.12% accuracy (AUC=0.938). Addressing the data imbalance issue with oversampling techniques, the accuracy can be increased to 91.50% (AUC=0.992), where the state-of-the-art method can only be estimated at 69.64% (AUC=0.921). Similar results can also be observed on independent and isolated testing data.

Availability: <https://github.com/saifurcubd/Cancer-Detection>

Contact: kc.w@cityu.edu.hk

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Genetic modifications often cause cancers through evolutionary diversity and selections with uncontrolled growth of cells (Colaprico, 2020; Nowell, 1976; Tao *et al.*, 2019). It is one of the leading causes of death for both men and women worldwide (Torre *et al.*, 2015). It has 18.1 million new cases, with 9.6 million deaths in 2018 (Bray *et al.*, 2018; Hassan and De Rosa, 2020). Early cancer detection can inform early medical interventions that reducing the patient mortality rate. It was found that late cancer detection can reduce patient survival rates (Hiom, 2015). Early cancer detection is accessible and can be performed in various ways. Recently, liquid biopsy through blood test is a standard procedure for early cancer detection from molecular biomarkers, genetic variants, and mutations in circulating cell-free DNA (cfDNA).

A molecular biomarker can indicate the disease prognosis in a patient. It can be defined based on a specific protein, a fragment of the protein, DNA mutation, or even an RNA strand (Kumar, 2006; Hanash *et al.*, 2008). In particular, protein biomarkers are specific and sensitive to clinical cancer detection, management, and monitoring (Hüttenhain *et al.*, 2019). Blood plasma is commonly adopted as the protein biomarker source with minimal invasive damages toward patients (Surinova *et al.*, 2011). Recently, many biomarkers are introduced to detect different cancer types at early and late stages. In particular, four protein analytes are discovered in the early detection of ovarian cancer in 2005 (Mor *et al.*, 2005). Since then, it has triggered our interest in cancer protein biomarkers for cancer detection (Stoeva *et al.*, 2006). Later on, many protein biomarkers are proposed and identified from blood test analytes for cancer

detection at the early stages (Cohen et al., 2018; Visintin et al., 2008; Wang et al., 2020; Wong et al., 2019) and for different types of cancers such as colorectal cancer (Karl et al., 2008; Pei et al., 2007), breast cancer (Whitwell et al., 2020; Harbeck et al., 2014), liver cancer (Bertino et al., 2012), lung cancer (Buszewski et al., 2012), pancreatic cancer (Takadate et al., 2013), esophageal cancer (Napier et al., 2014), and gastric cancer (Rugge et al., 2015).

Cell-free DNA (cfDNA) based liquid biopsy is an appealing clinical application for early cancer detection (Shuo Li et al., 2020). Examining cfDNA through blood can give a non-invasive medical test for cancer patient detection (Cristiano et al., 2019). Currently, the blood test is widely used for early cancer detection. The early prostate cancer detection studies based on prostate-specific antigen assessment are still debated (Pinsky et al., 2017). Many other researchers are working on the sequencing of cancer-based somatic variations in circulating cfDNA for early cancer detection (Cohen et al., 2018; Cristiano et al., 2019; Phallen et al., 2017; Razavi et al., 2019), gastric cancers (Kim et al., 2019), colorectal cancer (Osuni et al., 2019), lung cancer (Gandara et al., 2018), breast cancer (Cristina, 2019; O'Leary et al., 2018; Garcia-Murillas et al., 2015), lung cancer in early-stage (Abbosh et al., 2017) and late-stage human malignancies (Bettegowda et al., 2014). ABEMUS (Casiraghi et al., 2020) is developed to detect the somatic single-nucleotide variants in cfDNA for cancer detection and recurrent cancer growth detection from sequencing data (Caravagna et al., 2018). However, cfDNA-based blood tests (liquid biopsies) have limitations in cancer localizations (Bettegowda et al., 2014). Therefore, several studies recommend the direction in which both cfDNA mutations and protein biomarkers should be combined (Cohen et al., 2017).

The somatic variants in cfDNA, circulating tumor DNA (ctDNA) and different protein biomarkers of blood plasma analytes can enhance the cancer detection performance using machine learning tools and techniques at early stages (Wong et al., 2019). Through machine learning, cancer detection can be achieved based on different data types such as pathological data, clinical data, liquid biopsies data, and cfDNA mutations data.

In this article, protein biomarker concentrations and identified mutations in plasma cfDNA/ctDNA data collected from cancer patients and healthy controls (Cohen et al., 2018) are adopted to detect cancers at different stages (i.e. stages I to III) according to the American Joint Commission on Cancer (AJCC) as well as the localization of surgically resectable eight cancer types such as lung, liver, colorectum, ovary, esophagus, stomach, breast, and pancreas. In particular, breast and ovarian cancer are common cancers in women. Ovarian cancer has caused 152 000 deaths annually worldwide (Whitwell et al., 2020).

Researchers are working on cancer detection using machine learning algorithms; for instance, network-based multi-task learning model (Wang et al., 2020), deep-learning (Chen et al., 2019; Wong et al., 2019), and conjunctive Bayesian networks for cancer predictability pathways (Hossenli et al., 2019). Recent research in early cancer detection from mutations in cfDNA and different protein biomarkers data through blood test analytes are CancerA1DE, Deep Learning, Decision tree, Naïve Bayes, Random Forest, and CancerSEEK (Cohen et al., 2018; Wong et al., 2019). CancerA1DE relies on Average One-Dependence Estimators (AODE) (Webb et al., 2005). AODE is a semi-naïve Bayesian machine learning estimator that can classify through aggregating many one-dependence classifications. Deep Learning used deep feed-forward neural networks with one, two, and three hidden layer/s with other default settings in weka tools (Hall et al., 2009). CancerSEEK uses logistic regression for frontline binary cancer classification as Cancer or Normal class and random forest (RF) for cancer type localization. However, the existing CancerA1DE method has achieved satisfactory prediction performance for binary cancer classification. Nevertheless, it does not give satisfactory results for cancer localization to the class imbalance problem. The class imbalance problem for cancer localization detection data is still challenging. The existing methods are limited to insufficient accuracy in cancer detection due to the data uncertainty from cfDNA mutation scores and protein biomarker data from multianalyte blood test. Those methods

are hardly realistic, along with high standard deviations and class imbalance in cancer localization data. Therefore, we need to propose other machine learning techniques that can handle the uncertainties and limitations.

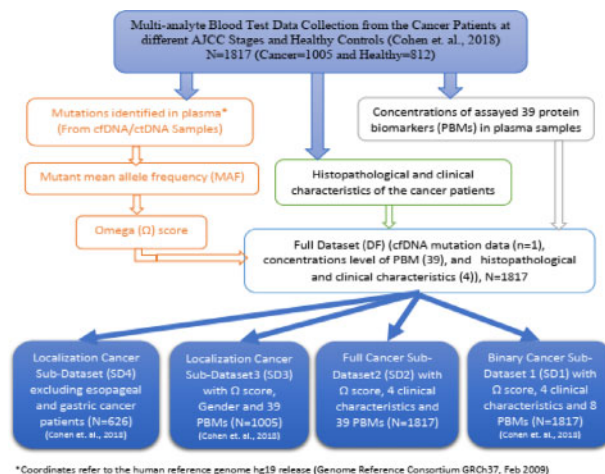
In this article, an ensemble meta classifier (CancerEMC) with average one-dependent estimators (AODE) is proposed for cancer detection. It can be implemented for both binary cancer detection as 'Cancer' or 'Normal' and cancer localization types detection. For CancerEMC, different attribute selection methods are used to select the best protein biomarker attribute set for both binary and localized cancer detection. Moreover, ADaptive SYNthetic (ADASYN) (He et al., 2008) and Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) are found beneficial toward CancerEMC for oversampling of unbalanced cancer localization data that increase the performance accuracy of localized cancer detection.

The remaining of this article is organized as follows: Section 2 briefly presents the methods and materials for cancer detection from cfDNA mutation score and protein biomarker levels in blood test analytes; Section 3 provides the experimental setup, results, performance analysis, and comparison with existing other methods of cancer localization detection through blood test analytes.

2 Materials and methods

2.1 Datasets

In this study, we have collected the multi-analyte blood test data (mutations of cfDNA/ctDNA and the assayed protein biomarker levels in blood plasma samples) and clinical characteristics of cancer detection from Cohen et al. (2018), as shown in Figure 1. It consists of two datasets collected from cancer patients and healthy controls by the blood test (liquid biopsy) as a non-invasive medical test. The first dataset has 1817 patients' blood test sample data, where 1005 cancer patients are diagnosed at the median age of 64 (range 22–93) with different AJCC stages I to III. These cancers are identified at eight different organs such as breast, lung, colorectum, liver, ovary, stomach, pancreas, and esophagus. The others are 812 healthy control individuals at the average age of 55 (28–65) without a cancer history. The mentioned dataset consists of thirty-nine protein biomarker concentrations in plasma samples and omega score calculated from the detected mutations in cfDNA samples of the same patient and their clinical characteristics [age, sex, race (ethnicity) and histopathology]. In particular, Cohen et al. (2018) used nine selected features (Omega score and eight protein biomarkers selected by a straightforward optimization) from the first dataset for binary cancer detection. The second dataset consists of 626 cancer patient data samples with forty-one features (Omega score, thirty-nine protein biomarkers and age) for cancer type localization.



*Coordinates refer to the human reference genome hg19 release (Genome Reference Consortium GRCh37, Feb 2009)

Fig. 1. Methodology overview for early cancer detection

We have processed the multi-analyte blood test dataset into four sub-datasets (shown in Figure 1) to evaluate the proposed CancerEMC method in different aspects. Those four sub-datasets are denoted as subdataset1 (SD1) for binary cancer detection as same as the first dataset of Cohen *et al.* (2018) for binary cancer detection; subdataset2 (SD2) for binary and localized cancer detection with all features except histopathology; subdataset3 (SD3) which consists of only 1005 cancer patient data for localized cancer detection and finally; subdataset4 (SD4) with only 626 samples with forty-one features as same as the second dataset of Cohen *et al.* (2018).

2.2 Features analysis and selection

In total, we have 1817 patient samples with one feature (Omega score) related to the cfDNA mutations, thirty-nine features of protein biomarkers (PBMs) concentrations in blood plasma, and four clinical features (age, sex, ethnicity and histopathology). Omega score denotes the cfDNA mutation score calculated from the mutant allele frequency (MAF) in four wells of unique identifier sequences (UIDs) of cfDNA in plasma samples using the following equation (1) (Cohen *et al.*, 2018).

$$\text{Omegascore}(\Omega) = \sum_{i=1}^W w_i \cdot \ln \frac{p_i^C}{p_i^N} \quad (1)$$

where, W is the number of wells in cfDNA, w_i is the ratio of the number UIDs in i th well and the total number of UIDs ($\text{UIDs}_i/\text{UIDs}$), p_i^C is the P -value of Cancer in MAF distribution of i th well and finally, p_i^N is the P -value of Normal in MAF distribution of i th well.

All clinical features are collected from cancer patients and healthy controls except histopathology. The histopathology feature describes the microscopical characteristics of cancer cells/tissues to identify different cancer types. Therefore, it is not included in the input features of the CancerEMC method for cancer detection. Moreover, the 'sex' feature is used in CancerEMC. It is found less important than the other features for binary cancer detection. However, it has remarkable impacts on cancer localization because some cancer types such as breast cancer and ovary cancer are discriminative in female patients. The ethnicity feature represents the individual's genetic invariant and physical traits that also impact binary cancer detection.

We have conducted statistical analysis with visualization tools such as correlation heatmap, data scattering histogram, cluster heatmap, parallel coordinate plot, and feature histogram for feature analysis. Supplementary Figure S1 visualized the Pearson correlational heatmap of thirty-nine protein biomarker features on SD2, where most of the protein biomarkers exhibit low correlation values. Supplementary Figure S2 visualizes the parallel coordinate plot for all features for both binary and localized cancer detection. It has illustrated that all features, including protein biomarkers, are not equally crucial for both binary and localized cancer detection from multianalyte blood test data. Supplementary Figures S3 and S4 show the correlational cluster heatmap on SD1 for binary cancer detection and SD4 for localized cancer detection. From Supplementary Figures S1, S3 and S4, we observed that most protein biomarkers have low correlational values (-0.2 to less than 0.4).

We have applied different feature selection (FS) methods to select vital protein biomarker features for the binary and localized cancer detection, such as Random Forest Feature Selection (RFFS) algorithm, Information Gain Ratio (InfoGainRatio), Recursive Feature Elimination (RFE) using logistic regression, random forest, extra tree classifier and Extreme Gradient Boosting tree (XGBoost) (Chen and Guestrin, 2016). To consider the unbiased cancer prediction, we have adopted 10-folds cross-validation (CV) with and without feature selection method and obtained median accuracies, AUCs in ROC space, f-Scores, along with CV error estimations as shown in Supplementary Table S1. We used the Root mean squared error (RMSE) and Mean absolute error (MAE) for error estimation in 10-folds CV (Varma and Simon, 2006; Ambrose and McLachlan, 2002). It is observed that feature selection before the CV gives better results when considering error estimation and other evaluation metrics. In the CancerEMC, the FS process of all feature selection methods used the 10-fold CV and other FS criteria to select significant protein biomarkers for cancer detection. Therefore, we have conducted a feature selection procedure before the CV and resampling techniques. RFFS algorithm, RFE and XGBoost methods are implemented in Python 3.0 sklearn package. The InfoGainRatio features selection method is implemented in Weka 3.4 tools with default parameters. XGBoost is an ensemble boosting machine learning algorithm for both data classification and feature selection. In this article, the RFFS algorithm, XGBoost, RFE and InfoGainRatio selection methods are compared for significant protein biomarker feature selection from thirty-nine protein biomarkers for both binary and localized cancer detection. XGBoost protein biomarker importance bar with their average gains across all splits is visualized in Supplementary Figure S5 for binary cancer detection and Supplementary Figure S6 for localizing cancer detection. Protein biomarkers feature selection using the RFFS algorithm with their importance values are depicted in Supplementary Figure S8 and S9 and for binary and localized cancer detection, respectively. RFFS algorithm used the Gini importance calculated by averaging the decrease in impurity all over the random forest trees for features importance.

Supplementary Figure S8 and S9 illustrate that the fifteen protein biomarker features (From IL-8 to Thrombospondin-2 in Supplementary Fig. S8) significant for binary cancer detection and 19 protein biomarker features (From IL-6 to sFas in Fig. S9) for localized cancer detection, respectively. On the other hand, Cohen *et al.* (2018) selected only eight protein biomarkers (PBMs) with a straightforward optimization techniques in SD1. The scatter histogram of binary cancer detection features with green color for cancer patients and orange color for healthy patients is shown in Supplementary Figure S7. The details are tabulated in Table 1.

Table 1 illustrates that the RFFS algorithm has the best accuracy to select the minimum number of protein biomarker features for binary and localized cancer detection.

After analyzing all PBM, clinical and cfDNA features for cancer detection, we have reached the following observations: (1) Different PBMs have different impacts on cancer detection. (2) PBM features are not suitable to perform cancer detection alone (Bettegowda *et al.*, 2014). (3) The cfDNA mutation score (omega score) and PBMs data can increase cancer detection performance (Cohen *et al.*, 2017). (4) Finally, the clinical features such as sex also have impacts on cancer localization.

Table 1. Protein biomarker (PBM) features selection with maximal classification accuracy based on CancerEMC

Rank	Method	Binary (no. of selected PBM) [accuracy]	Cancer types (no. of selected PBM) [accuracy]
1	RFFS Algorithm	15 [99.17%] (Supplementary Fig. S8)	19 [73.1629 %] (Supplementary Fig. S9)
2	XGBoost	14 [99.1194%] (Supplementary Fig. S5)	25 [72.4042%] (Supplementary Fig. S6)
3	RFE with Logistic Regression	25 [98.8993 %]	25 [70.4473 %]
4	RFE using Random Forest	6 [98.8442 %]	13 [70.2875 %]
5	RFE using Extra Tree	17 [98.90%]	30 [71.5655 %]
6	Info Gain Ratio	39 [99.00%]	34 [72.6837 %]

We have used the additional features: age, ethnicity, and sex with previously mentioned omega scores of cfDNA and selected fifteen significant protein biomarker features from SD2 for binary cancer detection. To localize cancer detection, we selected nineteen protein biomarker features by the RFFS algorithm along with omega score and sex attributes for cost-effective blood test. We have found that these nineteen PBMs are already used to detect previously mentioned cancer types. Therefore, CA 15-3, CA19-9, sHER2, sEGFR2, sErbB2, IL-6, IL-8, Midkine, Prolactin, GDF 15, CD44, Leptin, sFas, and TIMP-2 PBMs are used for breast cancer; AFP and OPN are used for liver cancer; EGFR2, NSE, CEA, Midkine, Thrombospondin-2, GDF-15, HGF, TGF α , and Leptin are used for lung cancer; CA 19-9, HGF, OPN, Thrombospondin-2, sHER2, TGF α , and TIMP-2 are used for pancreas cancer; CA-125, sHER2, and GDF-15 are used for ovary cancer; CEA, IL-6, GDF 15, CD44, TIMP-2, Leptin, and sHER2/sEGFR2/sErbB2 are used for colorectal cancer; and finally CA19-9, sHER2, OPN, HGF, and TGF α are used for Upper GI (upper gastro-intestinal) cancer (Sung and Cho 2008; Borrebaeck, 2017; cancer.gov, 2020; Cao et al., 2012; Filippou et al., 2020; Hassan et al., 2009; Jiang et al. 2019; Kim et al., 2017; Matsumoto et al. 2017; Spanopoulou and Gkretsi, 2020). Moreover, the state-of-the-art methods (CancerSEEK and CancerA1DE) (Cohen et al., 2018; Wong et al., 2019) are compared based on all thirty-nine protein biomarker features for localized cancer detection along with sex attribute. However, those methods cannot give better cancer detection compared to the proposed CancerEMC methods with a smaller number of protein biomarkers. Therefore, the CancerEMC method is more cost-effective than the existing method due to the minimal protein biomarkers needed.

2.3 Oversampling techniques for data imbalance

Data imbalance is one of the most critical problems for data classification. Generally, the number of samples in real dataset is usually not equally distributed to classes, leading to majority class and minority class (also known as class imbalance issue). It can create pitfalls in data science such as model overfitting or underfitting. In particular, the imbalanced data classification problems can bias the classification result in favor of the majority class. In machine learning techniques, class imbalance problems are handled in two approaches (Chawla et al., 2002). The first one is to assign cost functions to training samples (Pazzani et al., 1994). Another one is the original dataset's resampling obtained by either undersampling to mitigate the majority classes or oversampling to reduce the minority classes (Kubat and Matwin, 1997).

In this article, the localized cancer detection data has the data imbalance problem since each cancer type is not equally distributed. In particular, we observe that colorectal cancer is significant with the highest patient count in SD4, as shown in [Supplementary Figure S10](#). It illustrates the red bar for colorectal class as the majority class and blue bars for liver and ovary classes, which are minority classes. The localized cancer detection performance has been biased toward the colorectal majority class. Hence, we need to solve this data imbalance problem through resampling techniques.

Considering such a phenomenon, we apply different undersampling techniques (ClusterCentroids, RandomUnderSampler, NearMiss, InstanceHardnessThreshold, CondensedNearestNeighbour, OneSidedSelection and EditedNearestNeighbours) to minimize the data imbalance issue of SD4. The InstanceHardnessThreshold method gives better results in both ROC space and accuracy than others. However, it has generated small numbers of data instances (308) for the CancerEMC method. Its accuracy is not good enough, as shown in [Supplementary Table S2](#). We apply different oversampling methods (SMOTE, ADASYN, SMOTETomek, SVMsSMOTE and RandomOverSampler) in the same SD4 data. The SMOTE method gives the highest AUC values as shown in [Supplementary Table S2](#). According to the SMOTE reference article, we also applied SMOTE oversampling to the minority classes combined with InstanceHardnessThreshold under-sampling to the majority class in SD4 and obtained the 82.64% median accuracy with the AUC of 0.976 as shown in [Supplementary Table S2](#). From [Supplementary Table S2](#), we observed that the SMOTE oversampling technique is better than

other oversampling/undersampling techniques in accuracy for cancer detection using CancerEMC. Therefore, SMOTE oversampling techniques are employed to the CancerEMC method for addressing the data imbalance problems in cancer localization. In this article, the Adaptive Synthetic Sampling Method (ADASYN) also recommends mitigating the data imbalance.

SMOTE is an oversampling method (Chawla et al., 2002) that solves the class imbalance problem in a real dataset. It relies on k-nearest neighbor (KNN) for oversampling in imbalanced datasets. For SD4, we have applied the SMOTE method using the imbalance-learn python package. It 260% oversampled to breast cancer class, 486.05% for liver cancer class, 313.5% for lung cancer class, 375.47% for ovary cancer class, 276.12% for pancreas cancer class and finally 215.5% for upper GI cancer class, resulting in the equal number of patient data instances for each cancer type. The updated instance number of each class of SD4 is shown in [Supplementary Figure S11](#).

Finally, after applying oversampling techniques the patient data have been balanced for localized cancer detection. The balanced data instance statistics after using the oversampling technique to SD4 are shown in [Supplementary Figure S11](#). We have adopted the balanced data of oversampling techniques to the proposed CancerEMC method and observed that the SMOTE techniques could mitigate the data imbalance problem. The effects of oversampling will be discussed in the results section.

2.4 CancerEMC

Ensemble classifier is a fundamental approach that can use more than one base classifier learning algorithm for training to achieve robust classification results (Zhang and Ma, 2012). It is an attractive approach that can enhance the weak learner classification performance by increasing the classification accuracy. Ensemble learning aims at finding a set of learning algorithm models, providing better performance than individual learning algorithms. It is a supervised algorithm that combines the other supervised learning algorithms to ensemble different training models to achieve an overall excellent classification performance. It has many ensemble techniques, such as Boosting, Bagging, Stacking, Grading, and Voting (LeDell, 2015).

For performance comparisons, several machine learning approaches are selected for cancer detection such as support vector machine (SVM), naïve Bayes, random forest, decision tree, DNTB, Bayesian network, the multi-objective evolutionary fuzzy classifier (MEOFuzzyC), neural network, knowledge-based classifier, deep learning, KNN, average one dependence estimators (AODE), and logistic regression. Although some approaches have achieved satisfactory performance in binary cancer classification, those are still not sufficient for cancer type localization due to the data imbalance problem. Decision tree, logistic regression, naïve Bayes, random forest, and deep learning have already been employed in previous research with the same dataset (Cohen et al., 2018; Wong et al., 2019). The remaining methods are implemented using machine learning tools (Weka). However, those methods and the existing methods are limited to insufficient accuracy in localized cancer detection due to the data uncertainty from cfDNA mutation scores and protein biomarker data in the blood test. Those methods are hardly realistic due to the very low correlation coefficient, high variances, and class imbalance in cancer localization data. Therefore, this cancer detection system needs other machine learning techniques to handle the uncertainties and mentioned limitations. In this study, we proposed to adopt the combination of individual machine learners through an ensemble method for cancer detection, such as cancer ensemble meta classifier (CancerEMC). CancerEMC employed bootstrap aggregating (Bagging) with average one-dependence estimators (AODE) as base learner. The bagging method combines the bootstrapped replica of the base training model learner. AODE learner is a variant of the naïve Bayes model. A detailed description of the Bagging and AODE method can be found in [Supplementary Document](#). The overview of CancerEMC method is shown in [Figure 2](#).

The CancerEMC method framework consists of five essential components in a sequential manner: (i) Collecting multianalyte

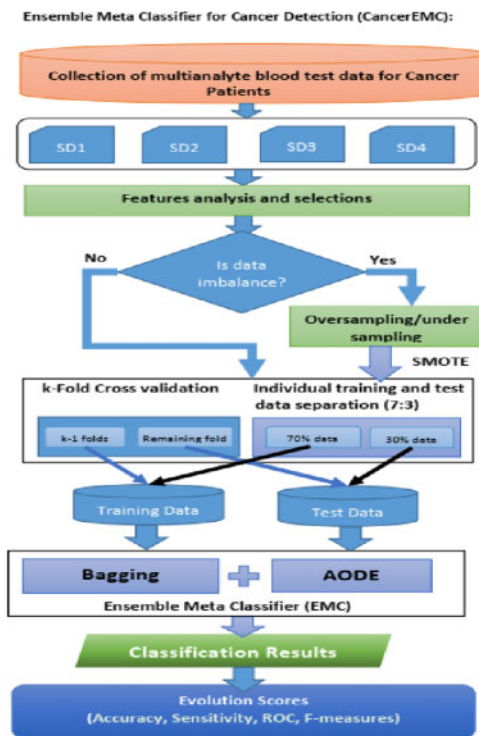


Fig. 2. Overview of CancerEMC for cancer detection

blood test data. (ii) Features analysis and selection. (iii) Over-sampling/under-sampling for imbalanced data. (iv) Two independent evaluation procedures to evaluate the model. Firstly, dataset stratification for nested cross-validation (NCV). Secondly, random training and testing data separation for isolated testing evaluation. (v) Ensemble meta classifier (EMC) methodology. The detailed description of CancerEMC is described in the following steps:

Step1: Collect the multianalyte blood test data with cfDNA mutations and protein biomarker (PBM) concentration levels from cancer patients and healthy control patients.

Step2: Organize the collected data into four sub-datasets as SD1 to SD4, where SD1 is for binary cancer detection (cancer or normal) only while SD2 is adopted for binary cancer detection and cancer type localization. In contrast, SD3 and SD4 are created only for additional localized cancer detection performance benchmarks as previously mentioned.

Step3: Select the significant protein biomarker (PBM) features by using an effective feature selection method. Here, the RFFS algorithm is used to select the significant PBM features illustrated in Section 2.2.

Step4: Check for the data imbalance problem. If any data imbalance is found, then oversampling and/or under-sampling techniques apply to mitigate data imbalance; otherwise, skip to the next step 5. Here, the SMOTE oversampling technique is used for the data imbalance problem of synthetic resamples to remove data imbalance problems. The SMOTE method is described in Section 2.3.

Step5: Partition the full dataset to training data and test data. For k-fold nested cross-validations, the full dataset is divided into k folds, where the kth fold is used as the validation data. The remaining k-1 folds are used as the training data with rotations for k iterations. We proposed another inner loop within the training data with 5-folds cross-validation to select each model parameter for each of the k iterations. The k-fold nested cross-validation (NCV) is used to validate the performance of CancerEMC under

different data folds. We used the 10-fold NCV (with the 5-folds inner-loop cross-validation) for cancer detection to validate the CancerEMC method. Secondly, we randomly separated the full dataset into 70% training data and 30% independent test data. Specifically, 70% of data are used to train each model, and the remaining 30% of data are reserved for testing the trained models.

Step6: Apply the Ensemble Meta Classifier (EMC) framework and follow the two aforementioned evaluation procedures for the cancer detection performance benchmark. To build the EMC, we employ the bagging learning method with AODE as a base learner. We used Weka 3.8.4 API in Java to implement the CancerEMC meta classifier with the base classifier AODE. Number of iterations is set 50–200, bagPercent is set to 100, batchSize is set to 100, and random seed is set to 1.

Step7: Generate the evaluation matrices from the classification results of Step6 for cancer detection. We generated two independent evaluation metrics respectively.

Step8: Generate and visualize the accuracy (ACC), sensitivity (SN), Area Under Curve (AUC), F-measures, micro-average (PR_{micro}), and macro-average precision (PR_{macro}) for cancer detection results from multianalyte blood test using the below equations (2) to (7) for two evaluation procedures as mentioned in step 5.

Step8: Compute the evaluation scores of the CancerEMC method

2.5 Performance evaluation

For the performance evaluations of the proposed CancerEMC along with other methods, we have computed the evaluation metrics that are Accuracy (ACC), Sensitivity (SN), AUC and F-measures.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = (TP + TN)/n \quad (2)$$

$$SN = TP/(TP + FN) \quad (3)$$

$$PR = TP/(TP + FP) \quad (4)$$

$$F - measure = 2 \times \frac{SN \times PR}{SN + PR} \quad (5)$$

$$PR_{micro} = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c} \quad (6)$$

$$PR_{macro} = \frac{\sum_c PR_c}{C} \quad (7)$$

where PR_{micro} is micro-average precision, PR_{macro} is macro-average precision, c is the class index and C is the total number of classes.

For the benchmark, we have used the k-fold cross-validation (CV) test with $k=10$ and 7:3 training-test data splitting ratios for CancerEMC along with other methods. We also used the receiver-operating characteristic (ROC) curves to compute for AUC and visualize the proposed CancerEMC method's comparisons with other existing and typical machine learning classifiers. The micro-average AUC and macro average AUC are also computed to evaluate different cancer types. Besides, the convex hull of the ROC curve, nested CV and 30% independent test data before resampling are also used to validate CancerEMC method.

3 Results and discussion

In this cancer detection study, CancerEMC is proposed to detect cancer from multianalyte blood test data that include the omega score and different protein biomarker concentrations with clinical history. To build this CancerEMC method, we followed the step by step procedure of the CancerEMC method (Figure 2) as mentioned in Section 2.4. We employed the RFFS algorithm to select the vital protein biomarker features and the bagging ensemble meta classifier with AODE as the internal base learner. We have used the python machine learning packages along with Weka 3.8.4 (Hall et al., 2009) tools and API to build the CancerEMC method. RFFS

algorithm is implemented using the sklearn python package. Bagging ensemble meta classifier and its base learner AODE are implemented using weka tools and API. Oversampling methods to solve the data imbalance problem are implemented using the imbalance-learn python package.

To generate the cancer detection result, we have used the available multianalyte blood test dataset from previous studies, as described in Section 2.1. After that, we reorganize the collected data into four sub datasets, SD1 to SD4. SD1 is the same data as the first dataset of Cohen *et al.* for binary cancer detection; SD2 is full blood test data with all feature attributes for both binary and localized cancer detection; SD3 is for cancer localization with 1005 cancer patients at the 99% specificity level. Finally, SD4 is for localized cancer detection with 626 cancer patient data (Figure 1). We have analyzed and found the significant protein biomarkers for CancerEMC by applying the RFFS algorithm through recursive feature elimination with different classification learners and other methods (Section 2.2). Different oversampling techniques are used to produce a new balanced dataset for solving the data imbalance problems. SMOTE and ADASYN oversampling techniques are applied to find out balanced data for cancer detection. Finally, the bagging ensemble meta classifier with the AODE base learner method is implemented for cancer detection.

Overfitting is one of the most common and concerning problems in machine learning classification and prediction (Claeskens *et al.*, 2008). It occurs to a model that closely fits for limited data and not reasonably fit for future unseen data. It can affect data classification and prediction results. Data overfitting can be prevented using cross-validation (CV), ensemble, resampling, removing irrelevant features, model regularization, and early stopping through loss function. In this study, overfitting can occur due to data imbalance, redundant features, and insufficient data. Hence, we need to handle the data overfitting problem for the proposed CancerEMC method. Therefore, feature selection methods, different resampling techniques, 10-folds nested CV for parameter regularization, and random independent test data separation are used to handle the CancerEMC method overfitting and validation. Independent test data are randomly separated through splitting the original dataset into a training dataset and an independent test dataset (i.e. not used to train the model). We discuss and present the proposed CancerEMC method results for the sub-datasets SD1–SD4, respectively.

For SD1, it has 1817 patient instances with twelve features and a binary class label for cancer detection from multianalyte blood test data. We have applied the proposed CancerEMC method for binary cancer detection as ‘Cancer’ or ‘Normal’. We have also used the clinical characteristics (age, sex, and ethnicity) and nine features mentioned in previous studies to enhance binary cancer detection accuracy. Finally, we observe that CancerEMC can achieve the classification accuracy of 97.91% and the AUC of the ROC curve is 0.9979, as shown in Figure 3, Table 2, and Supplementary Table S5. Subsequently, we have also applied supervised machine learning classifiers with the Weka default parameter values on SD1 for comparisons with CancerEMC results in Table 2 and Figure 3. We also applied 70% data for training and 30% data for testing on SD1, as shown in Table 2. Again, CancerEMC with SMOTE has applied to SD1 and obtained the accuracy value 97.51%, which slightly decreased as shown in Supplementary Table S5.

For SD2, it also has 1817 data instances of multianalyte blood test data with omega score, protein biomarkers, and clinical characteristics. We have applied the RFFS algorithm through recursive feature elimination (described in Section 2.2) to select the significant protein biomarkers for binary classification with omega score of cfDNA and clinical characteristics such as age, sex, and ethnicity. Based on 10-fold CV, CancerEMC obtained the highest accuracy value $ACC = 99.17\%$ with $AUC = 0.999$. CancerEMC is also tested on 30% independent test data randomly separated from SD2 for binary cancer detection with $ACC = 99.08\%$ and $AUC = 0.999$ (shown in Table 2). Again, CancerEMC with SMOTE was applied to SD2 and obtained the accuracy value of 99.05% as shown in Supplementary Table S5.

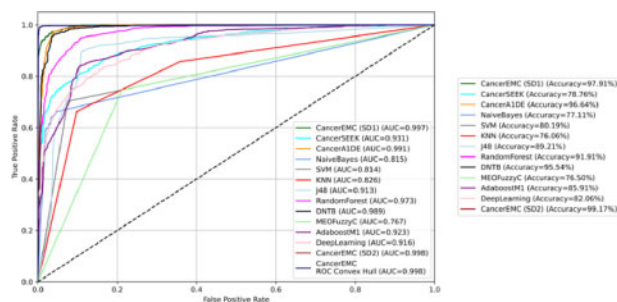


Fig. 3. ROC curve comparisons for binary cancer detection

For binary cancer detection, SD1 and SD2 sub datasets are used to the proposed CancerEMC method and existing and other machine learning methods to generate and evaluate the binary cancer detection from the multianalyte blood test. We observe that the proposed CancerEMC method performed better than CancerAIDE, CancerSEEK methods, and other supervised learning methods in terms of both the 10-folds CV method and 30% independent test data for binary cancer detection. It achieved a better accuracy value of 99.174% than existing methods (Cohen *et al.*, 2018; Wong *et al.*, 2019) for binary cancer classification.

Again, CancerEMC has been applied to SD2 for localized cancer type detection under the 10-fold CV. It obtained the classification accuracy $ACC = 83.4893\%$ and $AUC = 0.980$ with median sensitivity = 83.50% as shown in Supplementary Table S5. CancerEMC method was also tested on the 30% independent test data of SD2 for localized cancer detection and obtained the median accuracy of 86.9725% with median $AUC = 0.984$. CancerEMC with SMOTE has been applied to SD2 for localized cancer detection and obtained the accuracy value 95.977%, which has been dramatically increased, as shown in Supplementary Table S5 with a massive number of instances due to the 812 healthy controls.

For SD3, it has only 1005 cancer patient instances of multianalyte blood test data with omega score, sex, and protein biomarkers for localized cancer detection at the 99% specificity level. We adopted the proposed CancerEMC method to SD3 for cancer localization at the 99% specificity level under 10-fold CV. This method has given an accuracy value $ACC = 74.289\%$ (Shown in Supplementary Table S5) that is also better than the previous studies (i.e. CancerSEEK and CancerAIDE). After handling the data imbalance problem, the proposed CancerEMC method with SMOTE has increased the median accuracy to 93.98% using the 10-folds CV shown in Supplementary Table S5.

Finally, for the SD4 sub-dataset with the same features as used in (Wong *et al.*, 2019) at the 99% specificity level, we applied different supervised machine learning methods [i.e. BayesNet, Logistic, SVM, K-NN, AdaboostM1, Random Forest, J48, Decision Tree with Naïve Bayes (DTNB), MultiObjective EvaluataryFuzzyClassifier] with the proposed CancerEMC method. All mentioned methods are implemented in Weka 3.8.4 API tools (Hall *et al.*, 2009) with the same features and default Weka parameter values. We have used SD4 without oversampling and PBM features selection in all methods and obtained the ACC and AUC in Table 3 as visualized in Supplementary Figure S12. For CancerEMC, localized cancer detection results from SD4 are presented in three steps: (i) Apply SD4 without any change (the same as previous studies). (ii) Apply SD4 with protein biomarker features selection. (iii) Apply SD4 with both protein biomarker feature selection and oversampling techniques.

In the first step, we have applied only ensemble meta classifier (EMC) of CancerEMC without feature selection and oversampling by considering 10-fold CV and 30% independent test data. We obtained the median cancer detection accuracy under 10-folds CV as 72.8435% and 66.4894% for 30% independent test data. The ROC curves with AUCs from 10-fold CV for all localized cancer types are shown in Supplementary Figure S13.

In the second step, the RFFS algorithm was applied to the CancerEMC method to select the significant protein biomarkers

Table 2. Numerical comparisons for binary cancer detection

Methods	10-Fold cross-validation (CV)		30% of independent test data	
	Accuracy	AUC	Accuracy	AUC
CancerSEEK (Cohen <i>et al.</i> , 2018)	77.71%	0.930	86.6055 %	0.947
CancerA1DE (Wong <i>et al.</i> , 2019)	96.64%	0.991	96.1468 %	0.994
Deep Learning (Wong <i>et al.</i> , 2019)	82.05%	0.916	83.35 %	0.919
Naïve Bayes (Wong <i>et al.</i> , 2019)	77.10%	0.889	78.3486 %	0.931
SVM	80.18%	0.814	77.4312 %	0.756
k-NN	76.82%	0.762	80.5505 %	0.805
AdaboostM1	94.55%	0.982	88.6239 %	0.949
Random Forest	91.90%	0.913	92.844 %	0.980
J48 (Wong <i>et al.</i> , 2019)	89.21%	0.913	88.2569 %	0.902
DTNB	95.54%	0.990	95.9633 %	0.996
MultiObj-Evilu.FuzzyC	76.49%	0.767	77.6147 %	0.776
CancerEMC (for SD1)	97.91%	0.9989	97.9817 %	0.998
CancerEMC (for SD2)	99.17%	0.999	99.0826 %	0.999

The proposed method is highlighted in bold.

Table 3. Comparisons of median accuracy and AUC for cancer localization on SD4

Methods	10-Folds cross Validation (After oversampling)		30% independent test data separation before Oversampling/under sampling	
	Accuracy	AUC	Accuracy	AUC
CancerSEEK (Cohen <i>et al.</i> , 2018)	62.32%	0.91	Not used	
CancerA1DE (Wong <i>et al.</i> , 2019)	69.64%	0.921	Not used	
DeepLearning (Wong <i>et al.</i> , 2019)	63.73%	0.873	Not used	
NaïveBayes (Wong <i>et al.</i> , 2019)	46.48%	0.794	59.1133 %	0.872
Logistic	62.93%	0.853	65.0246 %	0.890
SVM	55.59%	0.766	62.3974 %	0.876
k-NN	49.36%	0.666	55.8292 %	0.742
AdaboostM1	69.96%	0.882	77.6683 %	0.963
RandomForest	67.57%	0.919	51.2315 %	0.727
J48	56.38%	0.734	74.5484 %	0.950
DTNB	65.17%	0.898	30.5419 %	0.595
MultiObj.Evilu.FuzzyC	41.05%	0.518	72.9064 %	0.955
CancerEMC	91.4966%	0.992	79.6388%	0.966

The proposed method is highlighted in bold.

features (described in Section 2.2) for the cost-effective blood test in cancer detection. According to the descending feature importance order, 19 protein biomarkers with omega score and sex are selected for the CancerEMC method (Supplementary Figure S9). After applying CancerEMC, we obtained the median localized cancer detection accuracy as increased to 74.1214% with AUC = 0.938 in 10-fold CV (the highest accuracy) before oversampling with only 19 protein biomarkers. We also applied the CancerEMC method with 70% training data and 30% independent test data from SD4 and obtained the median accuracy value of 67.5532% for localized cancer type detection. The ROC curves with different cancer types with average micro and macro ROC from the 10-fold CV are shown in Supplementary Figure S16 and Supplementary Table S5. It illustrates that the localized cancer detection on SD4 using the proposed CancerEMC method gives the AUC values with micro-average AUC=0.960 and macro-average AUC=0.951.

In the third step, we used two different oversampling techniques (ADSYN and SMOTE) to mitigate the data imbalance problem (describe in Section 2.3) in SD4 with the selected 19 protein biomarkers, omega score, and sex for localized cancer type detection.

For ADSYN, SD4 has the data imbalance issue with six minority classes and one majority class. Hence, it has to apply six oversampling runs for the six minority classes and obtained a balanced

dataset (1799 data instances). Ensemble meta classifier of the proposed CancerEMC was applied with 10-fold CV on the newly balanced dataset and obtained the accuracy ACC=90.7727%. The ROC curve with the AUC of each cancer type (i.e. average micro and macro ROC) are shown in Supplementary Figure S14. The individual ROC curves with AUC values and micro and macro average ROC curves are shown in Supplementary Figure S15.

For CancerEMC with SMOTE oversampling technique, the SMOTE method has been applied to handle data imbalance in SD4 with the selected 19 protein biomarkers, omega score, and sex for localized cancer type detection. It has oversampled 215% to 486.05% for six minority classes and generated 1764 data instances with an equal number of instances for each cancer type (i.e. 252). CancerEMC method is applied to the newly generated balanced dataset with 10-folds CV and obtained the median accuracy of 91.4966% with AUC=0.992, as shown in Table 3 and Supplementary Table S5. The ROC curves with AUCs of cancer types along with micro and macro average ROC curves are shown in Supplementary Figures S17 and S18.

We independently tested the CancerEMC method with 70% training data and 30% independent test data separation before oversampling/under-sampling (not used in the training process) in the same dataset. We obtained the accuracy ACC = 92.4386% with AUC = 0.992 (as shown in Table 3). The sensitivity of cancer type

detection is ranged from 71% to 100%. We observe that the performance of CancerEMC with the SMOTE method is sufficiently increased over the state-of-the-art methods, CancerA1DE and CancerSEEK.

Figure S17, Figure S18, Table 3 and Supplementary Table S5 illustrate that the proposed CancerEMC method has outperformed existing cancer detection methods: CancerA1DE (Wong et al., 2019), CancerSEEK (Cohen et al., 2018), and other standard machine learning classifiers while even not considering the class imbalance issue through SMOTE method. For this cancer tissue localization, the omega score and sex features are used along with PBM features for their relevance to cancer tissue localization. The 'sex' feature is essential because some cancer types such as breast cancer and ovary cancer are prevalent in female patients. We have employed the CancerEMC method to SD4 without the sex feature for breast cancer, resulting in the AUC value of 0.990. After including the sex feature, the AUC value is increased to 0.995.

Finally, we compared the localized cancer detection results of the proposed CancerEMC method with other existing CancerSEEK (Cohen et al., 2018) and CancerA1DE (Wong et al., 2019) method with and without considering the data imbalance problem in SD4. Supplementary Table S5 represents the ACCs of all sub-datasets (SD2, SD3 and SD4) for localized cancer detection for the benchmark comparison among CancerSEEK, CancerA1DE, and the proposed CancerEMC with/without oversampling SMOTE.

For the CancerEMC performance benchmark, we have adopted extensive evaluation procedures, resulting in several possible views on performance differences among different methods (Varma and Simon, 2006; Ambroise and McLachlan, 2002). (i) Original datasets are subsided into four sub-datasets to evaluate the method in different dataset representation. (ii) The results under the 10-folds cross-validation are shown in the second and third columns of Table 3. (iii) The results on the 30% independent test data (before oversampling/under sampling) are shown in the fourth and fifth columns of Table 3. (iv) The results with 10-folds nested cross-validation (5-folds inner loop for parameters regularization) are shown in the second and third columns of Supplementary Table S3. (v) The results on 30% independent test data (after oversampling/under sampling) to avoid test-train data contamination by observing synthetic data effect are shown in the fourth and fifth columns of Supplementary Table S3. (vi) Finally, the results of the 10-folds cross-validation with the same oversampling data are shown in the sixth and seventh columns of Supplementary Table S3. From Tables 3, Supplementary Table S5 and Supplementary Table S3, we found that the proposed CancerEMC method outperformed the other common machine learning algorithm and state-of-the-art studies for cancer detection from multianalyte blood test data.

Supplementary Table S5 illustrates that the SD1 and SD2 with 1005:812 binary cancer and normal class distribution do not incur any serious data imbalance problem. Hence, the oversampling has slightly decreased the binary cancer detection. Therefore, we have included a decision block for deciding on resampling after feature selection or not. According to this block, If a cancer detection dataset with any data imbalance issue, then oversampling and/or under-sampling will be applied. We have also applied the feature selection (FS) in 10-folds CV after oversampling/undersampling for unbiased cancer detection and obtained the median accuracy and AUC for all sub-datasets as shown in Supplementary Table S4.

From Tables 2, 3, Supplementary Tables S3, S4 and S5, we have found that the proposed CancerEMC method outperforms the other machine learning algorithms and state-of-the-art studies (including CancerSEEK published in Science, 2018) in aspects of validation for cancer detection from multi-analyte blood test data.

4. Conclusion

In this study, we have proposed CancerEMC for cancer detection from blood test. It has used ensemble learning with feature selection methods and oversampling techniques to mitigate the data imbalance problem. It has achieved better performance than the state-of-the-art studies for early cancer detection. We believe that such

intelligent methods can contribute to cancer research advances with broad impacts. In the future, we plan to apply the ensemble of knowledge and data-driven learning approaches for cancer detection to enhance the frontline cancer screening performance.

Acknowledgement

The two anonymous reviewers are thanked for their time and efforts, improving numerous aspects of the current study.

Funding

The work described in this paper was substantially supported by the grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11200218], one grant from the Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426], and the funding from Hong Kong Institute for Data Science (HKIDS) at City University of Hong Kong. The work described in this paper was partially supported by two grants from City University of Hong Kong (CityU 11202219, CityU 11203520). This research was substantially sponsored by the research project (Grant No. 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong.

Conflict of Interest: none declared.

References

- Abbosh, C. et al.; The TRACERx consortium. (2017) Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, **545**, 446–451.
- Ambroise, C. and McLachlan, G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, **99**, 6562–6566.
- Bertino, G. et al. (2012) Hepatocellular carcinoma serum markers. In: Fojo, A.T. (ed.) *Seminars in Oncology*, Vol. 39, Elsevier, Amsterdam, pp. 410–433.
- Bettgowda, C. et al. (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.*, **6**, 224ra24–224ra24.
- Borrebaeck, C.A. (2017) Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat. Rev. Cancer*, **17**, 199–204.
- Bray, F. et al. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries CA. *Cancer J. Clin.*, **68**, 394–424.
- Buszewski, B. et al. (2012) Identification of volatile lung cancer markers by gas chromatography-mass spectrometry: comparison with discrimination by canines. *Anal. Bioanal. Chem.*, **404**, 141–146.
- cancer.gov (2020) National Cancer Institute at the National Institutes of Health. <https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-list> (July 2020, date last accessed).
- Cao, D.X. et al. (2012) Osteopontin as potential biomarker and therapeutic target in gastric and liver cancers. *World J. Gastroenterol.*, **18**, 3923–3930.
- Caravagna, G. et al. (2018) Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods*, **15**, 707–714.
- Casiraghi, N. et al. (2020) ABEMUS: platform-specific and data-informed detection of somatic SNVs in cfDNA. *Bioinformatics*, **36**, 2665–2674.
- Chawla, N.V. et al. (2002) Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.
- Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13–17, 2016. ACM, San Francisco, CA, USA. pp. 785–794.
- Chen, R. et al. (2019) Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics*, **36**, 1476–1483.
- Claeskens, G. et al. (2008) *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.
- Cohen, J.D. et al. (2017) Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proc. Natl. Acad. Sci. USA*, **114**, 10202–10207.
- Cohen, J.D. et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, **359**, 926–930.
- Cohen, A.A., Javed, et al. (2017) Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic

- cancers. *Proc. Natl. Acad. Sci. USA*, **114**, 10202–10207. DOI: 10.1073/pnas.1704961114pmid:28874546
- Colaprico, A. *et al.* (2020) Interpreting pathways to discover cancer driver genes with Moonlight. *Nat Commun.*, **11**, 69.
- Cristiano, S. *et al.* (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, **570**, 385–389.
- Filippou, P.S. *et al.* (2020) Midkine (MDK) growth factor: a key player in cancer progression and a promising therapeutic target. *Oncogene*, **39**, 2040–2054.
- Gandara, D.R. *et al.* (2018) Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat. Med.*, **24**, 1441–1448.
- Garcia-Murillas, I. *et al.* (2015) Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.*, **7**, 302ra133–302ra133.
- Hall, M. *et al.* (2009) The Weka data mining software: an update. *ACM SIGKDD Explore Newsl.*, **11**, 10–18.
- Hanash, S.M. *et al.* (2008) Mining the plasma proteome for cancer biomarkers. *Nature*, **452**, 571–579.
- Haibo, H., Yang, B., Garcia, E.A. and Shu, L. (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Natural Net-works, Hong Kong*, 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.
- Harbeck, N. *et al.* (2014) Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow. *Cancer Treat. Rev.*, **40**, 434–444.
- Hassan, M.I. *et al.* (2009) Prolactin inducible protein in cancer, fertility, and immunoregulation: structure, function, and its clinical implications. *Cell. Mol. Life Sci.*, **66**, 447–459.
- Hassan, E.M. and De Rosa, M.C. (2020) Recent advances in cancer early detection and diagnosis: role of nucleic acid-based APA sensors. *Trends Anal. Chem.*, **92**, 9764–9771.
- He, H. *et al.* (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Net-works, Hong Kong*, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.
- Hu, S.C. (2015) Diagnosing cancer earlier: reviewing the evidence for improving cancer survival. *Br. J. Cancer*, **112**, S1–S5.
- Hossain, M.S. *et al.* (2018) A belief rule-based expert system to assess suspicion of an acute coronary syndrome (ACS) under uncertainty. *Soft Comput.*, **22**, 7571–7586.
- Hosseini, S.Z. *et al.* (2019) Estimating the predictability of cancer evolution. *Bioinformatics*, **35**, i389–i397, <https://doi.org/10.1093/bioinformatics/btz332>.
- Jiang, Y.M. *et al.* (2019) Serum thrombospondin-2 is a candidate diagnosis biomarker for early non-small-cell lung cancer. *Biosci. Rep.*, **39**, BSR20190476.
- Karl, J. *et al.* (2008) Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. *Clin. Gastroenterol. Hepatol.*, **6**, 1122–1128.
- Kim, J. *et al.* (2017) Detection of early pancreatic ductal adenocarcinoma with thrombospondin-2 and CA19-9 blood markers. *Sci. Transl. Med.*, **9**, eaah5583.
- Kim, Y. *et al.* (2019) Monitoring circulating tumor DNA by analyzing personalized cancer-specific rearrangements to detect recurrence in gastric cancer. *Exp. Mol. Med.*, **51**, 1–10.
- Kubat, M. and Matwin, S. (1997) Addressing the curse of imbalanced training sets: one sided selection. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, Nashville, Tennessee. pp. 179–186.
- Kumar, S. *et al.* (2006) Biomarkers in cancer screening, research, and detection: present and future: a review. *Biomarkers*, **11**, 385–405.
- LeDell, E. (2015) Scalable Ensemble Learning and Computationally Efficient Variance Estimation. Doctoral Dissertation, The University of California, Berkeley, USA.
- Li, S. *et al.* (2020) Sensitive detection of tumor mutations from blood and its application to immunotherapy prognosis. <https://www.medrxiv.org/content/10.1101/2019.12.31.19016253v1>.
- Matsumoto, K. *et al.* (2017) Hepatocyte growth factor/MET in cancer progression and biomarker discovery. *Cancer Sci.*, **108**, 296–307.
- Mor, G. *et al.* (2005) Serum protein markers for early detection of ovarian cancer. *Proc. Natl. Acad. Sci. USA*, **102**, 7677–7682.
- Napier, K.J. *et al.* (2014) Esophageal cancer: a review of epidemiology, pathogenesis, staging workup, and treatment modalities. *World J. Gastrointest. Oncol.*, **6**, 112.
- Nowell, P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23–28.
- O'Leary, B. *et al.* (2018) Early circulating tumor DNA dynamics and clonal selection with palbociclib and fulvestrant for breast cancer. *Nat. Commun.*, **9**, 896.
- Osumi, H. *et al.* (2019) Early change in circulating tumor DNA as a potential predictor of response to chemotherapy in patients with metastatic colorectal cancer. *Sci. Rep.*, **9**, 17358.
- Pazzani, M. *et al.* (1994) Reducing misclassification costs. In: *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA.
- Pei, H. *et al.* (2007) Proteome analysis and tissue microarray for profiling protein markers associated with lymph node metastasis in colorectal cancer. *J. Proteome Res.*, **6**, 2495–2501.
- Phallen, J. *et al.* (2017) Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.*, **9**, eaan2415.
- Pinsky, P.F. *et al.* (2017) Prostate cancer screening – a perspective on the current state of the evidence. *N. Engl. J. Med.*, **376**, 1285–1289.
- Razavi, P. *et al.* (2019) High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.*, **25**, 1928–1937.
- Rugge, M. *et al.* (2015) Epidemiology of gastric cancer. In: Strong, V.E. (ed.) *Gastric Cancer*, Springer, Berlin, pp. 23–34.
- Spanopoulou, A. and Gkretsi, V. (2020) Growth differentiation factor 15 (GDF15) in cancer cell metastasis: from the cells to the patients. *Clin. Exp. Metastasis*, **37**, 451–464.
- Stoeva, S.I. *et al.* (2006) Multiplexed detection of protein cancer markers with biobarcode nanoparticle probes. *J. Am. Chem. Soc.*, **128**, 8378–8379.
- Sung, H.J. and Cho, J.Y. (2008) Biomarkers for the lung cancer diagnosis and their advances in proteomics. *BMB reports*, **41**, 615–625.
- Surinova, S. *et al.* (2011) On the development of plasma protein biomarkers. *J. Proteome*, **10**, 5–16.
- Takadate, T. *et al.* (2013) Novel prognostic protein markers of resectable pancreatic cancer identified by coupled shotgun and targeted proteomics using formalin-fixed paraffin embedded tissues. *Int. J. Cancer*, **132**, 1368–1382.
- Tao, Y. *et al.* (2019) Improving personalized prediction of cancer prognoses with clonal evolution models. <https://www.biorxiv.org/content/10.1101/761510v1.abstract>.
- Torre, L.A. *et al.* (2015) Global cancer statistics, 2012. *CA Cancer J. Clin.*, **65**, 87–108.
- Varma, S. and Simon, R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91.
- Visintin, I. *et al.* (2008) Diagnostic markers for early detection of ovarian cancer. *Clin. Cancer Res.*, **14**, 1065–1072.
- Wang, Z. *et al.* (2020) Network-based multi-task learning models for biomarker selection and cancer outcome prediction. *Bioinformatics*, **36**, 1814–1822.
- Webb, G.I. *et al.* (2005) Not so naive Bayes: aggregating one-dependence estimators. *Mach. Learn.*, **58**, 5–24.
- Whitwell, H.J. *et al.* (2020) Improved early detection of ovarian cancer using longitudinal multimarker models. *Br. J. Cancer*, **122**, 847–856.
- Wong, K.-C. *et al.* (2019) Early cancer detection from multianalyte blood test results. *iScience*, **15**, 332–341.
- Zhang, C. and Ma, Y. (2012) *Ensemble Machine Learning: Methods and Applications*. Springer, Berlin.