





Review

Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey

Linjing Liu ¹ , Xingjian Chen ¹, Olutomilayo Olayemi Petinrin ¹ , Weitong Zhang ¹, Saifur Rahaman ¹ ,
Zhi-Ri Tang ¹ and Ka-Chun Wong ^{1,2,*} 

¹ Department of Computer Science, City University of Hong Kong, Hong Kong, China; jingliu3-c@my.cityu.edu.hk (L.L.); xingchen3-c@my.cityu.edu.hk (X.C.); opetinrin2-c@my.cityu.edu.hk (O.O.P.); weitzhang6-c@my.cityu.edu.hk (W.Z.); srahaman2-c@my.cityu.edu.hk (S.R.); gerintang@163.com (Z.-R.T.)

² Hong Kong Institute for Data Science, City University of Hong Kong, Hong Kong, China

* Correspondence: kc.w@cityu.edu.hk

Abstract: With the advances of liquid biopsy technology, there is increasing evidence that body fluid such as blood, urine, and saliva could harbor the potential biomarkers associated with tumor origin. Traditional correlation analysis methods are no longer sufficient to capture the high-resolution complex relationships between biomarkers and cancer subtype heterogeneity. To address the challenge, researchers proposed machine learning techniques with liquid biopsy data to explore the essence of tumor origin together. In this survey, we review the machine learning protocols and provide corresponding code demos for the approaches mentioned. We discuss algorithmic principles and frameworks extensively developed to reveal cancer mechanisms and consider the future prospects in biomarker exploration and cancer diagnostics.

Keywords: machine learning; early cancer detection; liquid biopsy



Citation: Liu, L.; Chen, X.; Petinrin, O.O.; Zhang, W.; Rahaman, S.; Tang, Z.-R.; Wong, K.-C. Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey. *Life* **2021**, *11*, 638. <https://doi.org/10.3390/life11070638>

Academic Editor: Alfredo Conti

Received: 7 June 2021

Accepted: 24 June 2021

Published: 30 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

When cells mutate, they could divide uncontrollably and eventually form cancer [1]. According to the World Health Organization, cancer accounts for nearly 10 million deaths in 2020. Unfortunately, this number is estimated to be still climbing in the following decades and will reach 27 million new cases in 2040 [2]. As the second factor of death, cancer accounts for one-sixth of deaths worldwide each year [3]. Therefore, fighting against cancer is a huge challenge for global public health. Early detection, followed by tailored site-specific treatment, plays an important role in the front-line cure of cancer and could reduce the eventual mortality of cancer patients [4–6].

Cancer is associated with mutated genes; and genetic analysis is increasingly applied in cancer diagnosis [7]. The traditional methods for genetic testing on cancer patients are sampling from tumor tissues. However, tumor tissue biopsy is limited by several drawbacks such as invasive acquisition, clinical complications, sample preservation, and tumor heterogeneity [8–10].

Liquid biopsy [7,11], which surmounts the limitation of tissue biopsy, is evaluated as a potential tool for early cancer detection and monitoring [12]. By sampling from blood, stool, urine, saliva, and other fluid samples, liquid biopsy provides a non-invasive and feasible cancer detection service [13–16]. Compared with tissue biopsy, liquid biopsy is also more comprehensive to evaluate tumor heterogeneity since tumor sites can release aberrant signals into body fluid [17,18]. Researchers paid significant attention to the different components from liquid biopsy which are associated with cancers [19–23].

As the possibility or severity of tumor in the body is relevant to the liquid biopsy components, accurate cancer prediction based on the characteristics of these components becomes a significant problem. The application of machine learning protocols has been

widely studied in recent years, proving to be valuable in early cancer detection. Nevertheless, the required knowledge to implement these methods is high, posing an obstacle to researchers who are looking to get started on liquid biopsy analysis and early cancer detection. Therefore, this review not only focuses on the published research of machine learning in early cancer detection but also demonstrates the entire implementation procedure in this effort.

The rest of this review is written in the following sections. Section 2 introduces the procedures of implementing machine learning, including data preprocessing, model selection, model evaluation, and hypothesis testing. Section 3 summarizes the mainly liquid biopsy components associated with cancer. Section 4 is an overview of the most widely used machine learning algorithms and the relevant literature with corresponding datasets. Section 5 is the discussion on this topic. For all machine learning protocols and algorithms, we provided the code demo as a tutorial available at (<https://github.com/ElaineLIU-920/Code-Demo-for-ML-procedures-and-algorithms>, accessed on 10 June 2021).

2. Machine Learning Related Procedures

The data sets of cancer liquid biopsies are large and complex. Therefore, it is difficult to deal with using traditional methods. Machine learning algorithms, as a potential tool, can automatically analyze and identify regularities from data and then predict future data based on the obtained experience. For machine learning, the detection of cancer is regarded as a supervised problem, which is called a classification task. In this section, we focus on the supervised machine learning protocols and some of the preparatory work before implementing these methods. This section is organized according to a typical workflow for supervised machine learning. Firstly, we will discuss some techniques for data preprocessing. Moreover, model evaluation and selection methods which include the performance metrics for supervised learning are also discussed. Next, we will introduce the hypothesis test to indicate the statistical significance.

2.1. Data Preprocessing

Data preprocessing is a fundamental step of the machine learning implementation, which has been stated to have a significant influence on the performance of machine learning models [24,25]. Data preprocessing consists of missing-value solution, normalization, dimension reduction, and feature reconstruction. As the future data is unknown in reality, we suggest that all data preprocessing methods are only applied to training data.

2.1.1. Missing Value

Missing value cannot be avoided in a dataset, which may create an obstacle for predictors. Inappropriately handling strategy will easily result in extracting poor knowledge, and wrongly prediction [26].

The first option to deal with this problem is to delete samples with missing values [27–29], which may result in discarding a large number of samples and increasing bias prediction [30]. Alternatively, the missing value can be filled by the mean, mode, or a random value [25]. Moreover, some model-based methods are also employed to predict the missing value [30]. Model-based methods do neither delete missing-value samples nor fill the value by simple imputation; Instead, it builds a model for the missing feature based on inferences from existing complete data.

Model-based methods consist of two steps: (1) Build a regression or classification model based on complete samples for the feature which is corresponding to the missing values; (2) Predict on the incomplete samples with its existed feature as input, and then the output is an estimate of missing value [31].

2.1.2. Normalization

The main advantage of implementing normalization is that it prevents the predictions of later stages from being dominated by relatively large or small values in the data set.

Besides, normalization is significant to ensure comparability over different samples. In this section, we will introduce three commonly used normalization methods, namely Z-Score standardization, Max-min normalization, and Decimal scaling [32].

- Z-Score standardization. In Formula (1), A is feature (attribute), x_A is the original value of feature A , x'_A is the normalized value; μ is the mean of feature A , and σ is the standard deviation of feature A .

$$x'_A = \frac{x_A - \mu}{\sigma} \quad (1)$$

- Max-Min normalization. Max-Min normalization, also called deviation standardization, is transformed by Formula (2), where min is the minimum of feature A ; max is the maximum of the feature A .

$$x'_A = \frac{x_A - \min}{\max - \min} \quad (2)$$

- Decimal scaling. This method is realized by moving the decimal point position according to the absolute maximum of feature A . In Formula (3), j is the smallest integer such that all x'_A is less than 1, $j = \lceil \log_{10} \max \rceil$. Here, max is the maximum of the feature A .

$$x'_A = \frac{x_A}{10^j} \quad (3)$$

2.1.3. Dimension Reduction

Feature is the observation of samples, which is also synonymous with input variables or attributes. The dimension of the dataset is the number of variables measured on each sample, equal to the number of features. Owing to the development of detection technology, the available samples have increased explosively in terms of dimension. When machine learning algorithms are applied to these high-dimensional data [33,34], dimension curse becomes a crucial issue to resolve, which is especially severe in bioinformatics [35,36].

One of the problems with the high-dimensional dataset is that some algorithms tend to perform poorly on high-dimensional data, as not all features are valuable for prediction. In many cases, a large amount of the features are irrelevant or redundant with the learning task, resulting in overfitting for learning models [32]. In addition, high dimensional data will also increase computation time as well as the memory of storage. Moreover, if the dimension of data is very high, visualization becomes quite difficult.

Feature extraction (also known as feature transformation, feature projection or dimension reduction specifically) and feature selection are two dimension reduction techniques [37] to solve these problems. The choice of feature extraction or feature selection depends on different data types and applications. We will next briefly introduce some typical approaches for dimension reduction.

A. Feature Extraction

Feature extraction method develops a transformation from the original high-dimensional feature space into a new low-dimensional space. The essence of feature extraction reduction is to learn a mapping function $f : X \rightarrow X'$, where X is the original data, and X' is a low-dimensional vector representation after data mapping. Linear mapping and non-linear mapping methods are two main types to implement feature extraction [38]. Linear mapping is mainly represented by principal component analysis (PCA) [39,40], linear discriminant analysis (LDA) [41], and non-negative matrix factorization (NMF) [42], while non-linear mapping is mainly represented by locally linear embedding (LLE) [43] and Isomap [44].

The advantage of feature extraction is that it decreases the dimension of feature through data transformation, which enables obtaining a lower feature space without losing information. However, it is precisely for this reason that the new space is obtained from the linear or non-linear transformation of the original space, causing the inexplicability of the new features.

B. Feature Selection

Different from feature extraction, feature selection directly selects a valuable subset features and removes noisy, redundant, or irrelevant features from the original dataset, which only contains the important information to solve the problem [45–47]. Based on different pathways of combining feature selection strategy with machine learning models, feature selection techniques are categorized into three types: filter method, wrapper method and embedded method [48].

Filter methods, independent of any learning models, assess the importance of features based on the statistical and intrinsic properties of the original dataset. In this setup, importance ranking is adopted as the principal criteria for feature selection. By reserving high-scoring features and removing low-scoring features, a subset with a lower dimension of features is obtained. Many filter-type methods have been studied, including Pearson correlation coefficient [49], F-statistic [50], Chi-squared-statistic [51] and Mutual information [52].

Wrapper methods adopt different search algorithms to generate the subsets of features. Subsequently, a specific subset is evaluated by training and testing the performance of the classification model, which is wrapped in the search algorithm. The whole process works iteratively until the highest learning performance is achieved or the desired number of selected features is obtained. A wide range of search strategies can be used, including Sequential Selection Algorithms, Recursive Feature Elimination, and Meta-heuristic Algorithms (e.g., genetic algorithm) [53,54].

Embedded methods explore the optimal subset of features during the process of constructing a learning model. Similar to the wrapper methods, the embedded methods are specific to the adopted machine learning algorithm. Least absolute shrinkage and selection operator, Elastic net and Ridge regression are three typical regularization algorithms [55,56].

Detail comparison of these three pathways to implement feature selection is discussed in [48,57]. As feature selection merely explores a valuable subset of the original feature, it retains the semantics of the original features, which possesses the advantage of interpretable analysis. However, some information may be lost when employing feature selection methods, as only a subset is reserved, and some of the features will be omitted.

Feature extraction, as well as feature selection, has the ability to improve model performance, computational efficiency, utilization of memory storage, and data visualization. Therefore, both of these two methods are employed as effective dimension reduction techniques, used alone or in combination.

2.1.4. Feature Construction

Feature construction is also known as attribute generation. Different from dimension reduction, in some cases, the features may be insufficient to describe the problem for learning models. Therefore, feature construction is adopted utilized to enrich the data. According to the definition, taken from Motoda and Liu [58], feature construction aims to discover the hidden relationships of original features by constructing new high-level features. Similar to feature selection, the process of constructing feature can also be categorised into three classes: filter methods, wrapper methods, and embedded methods [59,60]. For numerical features, simple algebraic operators such as addition, subtraction, multiplication, and division are often used to compound features.

2.2. Model Evaluation

Model evaluation is the process of assessing the performance of models on the future data [61]. In the straight forward, it aims to evaluate how well the built model by estimating the generalization error on unseen data. A good machine learning model should perform well not only on the training data but also on the future data. Therefore, before implementing a model for production, we should be fairly sure that the performance of the model will not decline when confronted with the new data. For most practical applications, the true performance of the model cannot be calculated as we do not have real future data. Hence,

it is important to use new data for model evaluation to prevent the likelihood of overfitting problems to the training set. Holdout, bootstrap, and cross-validation are most commonly used method for model evaluation [62–64].

2.2.1. Holdout Method

Holdout method is the simplest model evaluation method, which directly splits the dataset into two portions: training set and test set. For example, we randomly choose 2/3 of the whole dataset as the training set and 1/3 as the test set. Firstly, we utilize the training set to fit and build the model. Subsequently, we evaluate the built model on the test set by comparing the predictions of the label and the ground truth. To some extent, the test set represents the new and unseen data in practice. As the estimation result obtained by applying the holdout method once is often not reliable, it necessitates the repeating of splitting and evaluating several times, which is called the repeated holdout method. The average performance evaluation is reported as the final estimation result. We usually utilize about 2/3 to 4/5 of the dataset for training and the rest for testing.

It should be noted that we cannot train and evaluate the model based on the training dataset simultaneously, which is called resubstitution evaluation or resubstitution validation. As resubstitution evaluation would introduce optimistic bias due to overfitting on resubstitution samples, we cannot ascertain whether the model works because it remembers the training data or because it could generalize well on new data.

2.2.2. Cross-Validation

The basic idea of cross-validation is to divide the data into different subsets. In this setup, some of these subsets are used to train the model and the rest are used to test the model until all the samples have been used for testing. k -fold cross-validation strategy is most commonly used in the classification research [65]. With k -fold cross-validation, the dataset is partitioned into k disjoint subsets, the union of which is equivalent to the whole dataset. A single subset from these k disjoint subsets is retained as the test data to evaluate the classifier, and the remaining $k - 1$ subsets are used as training data. This process is then repeated for k times until all subsets are used as the test data exactly once. The performance evaluation results on k test set are averaged as the performance estimation for the classifier.

The step-by-step instruction of k -fold cross-validation is summarized as below. Figure 1 is the diagram of k -fold cross-validation.

- **Step 1:** Randomly split the original dataset into k equal folds.
- **Step 2:** Select one of these folds as test set, and the remaining $k - 1$ folds as training set to build model.
- **Step 3:** Compute generalization performance of the built model on the test set.
- **Step 4:** Repeat step 2 to step 3 for k times until each fold has and only has one chance to act as the test set, and the remaining folds act as the training set.
- **Step 5:** Report the average of generalization performance on all test sets as an estimations of the model performance.

The different values of k , which is usually five, ten or equal to the number of instances in the dataset, determine the different subtypes of cross-validation. Assuming that the dataset includes n samples, if $k = n$, we obtain a special case of the cross-validation, namely, the leave-one-out cross-validation (LOOCV). Obviously, the LOOCV method is not affected by the partition of samples, as there is only one unique way for n samples to be divided into n subsets, each of which contains only one sample. Although the evaluation results of the LOOCV method are often considered to be more accurate, LOOCV method also has unbearable computational overhead when the dataset is relatively large. For example, LOOCV needs to build 1 thousand models if the dataset contains 1 thousand samples. However, 5-fold cross-validation and 10-fold cross-validation only need to build five and ten models, respectively.

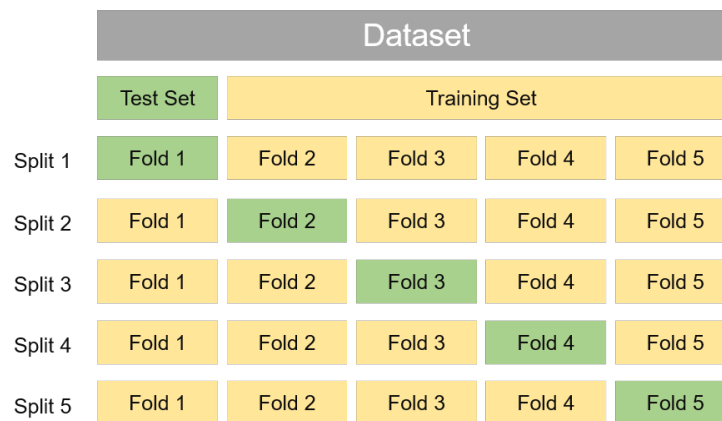


Figure 1. *k*-fold Cross-validation.

2.2.3. Bootstrapping

The bootstrapping method is a re-sampling technique to draw sample data repeatedly with replacement from the original dataset, proposed by Bradley Efron in 1979 [66]. The workflow of bootstrap method is summarized as following:

- **Step 1:** The size of original dataset is *n*. We randomly select one instance from this dataset and then assign it to the *j*_{th} bootstrap dataset. Repeating this process until the size of *j*_{th} bootstrap sample reaches *n*.
- **Step 2:** Fit a model to *j*_{th} bootstrap dataset and compute the performance.
- **Step 3:** Repeat Step 2 and 3 for *b* times. Calculate the model performance as the average over the *b* estimates. If accuracy is the performance metric, then the model performance for bootstrapping is:

$$ACC_{boot} = \frac{1}{b} \sum_{j=1}^b \frac{1}{n} \sum_{i=1}^n (1 - L(\hat{y}_i, y_i)) \tag{4}$$

In 1983, Bradley Efron described the 0.632 Estimate [67] to address the bias of the bootstrap approach aforementioned. The bias in the conventional bootstrap method is owing to the fact that the bootstrap approach only utilize approximately 63.2% of the samples from the whole dataset. For example, we can calculate the probability that a specific sample, from a dataset with size *n*, is not selected as as following:

$$P_N = \left(1 - \frac{1}{n}\right)^n \tag{5}$$

The value of Equation (5) is asymptotically equivalent to $\frac{1}{e} \approx 0.368$ when $n \rightarrow \infty$. Therefore, the probability that the specific sample is chosen as:

$$P_C = 1 - \left(1 - \frac{1}{n}\right)^n \approx 0.632 \tag{6}$$

Subsequently, to adjust the bias that is owing to the sampling strategy, Bradley Efron introduced the 0.632 Estimation method, computed by the Formula (7):

$$ACC_{0.632boot} = \frac{1}{b} \sum_{i=1}^b (0.632ACC_{h,i} + 0.368ACC_{r,i}) \tag{7}$$

where $ACC_{r,j}$ is the resubstitution accuracy, and $ACC_{h,j}$ is the accuracy on out-of-bag samples (samples which are not selected as the bootstrap samples). The 0.632 Bootstrap could address the pessimistic bias, however, an optimistic bias may occur. Therefore, 0.632 + Bootstrap was proposed [68].

$$ACC_{0.632+boot} = \frac{1}{b} \sum_{j=1}^b (\omega ACC_{h,j} + (1 - \omega) ACC_{r,j}) \tag{8}$$

Instead of using a fixed weight $\omega = 0.632$, $0.632 + \text{Bootstrap}$ compute the weight ω as

$$\omega = \frac{0.632}{1 - 0.368R} \tag{9}$$

where R is the relative overfitting rate:

$$R = \frac{(-1)(ACC_{h,j} - ACC_{r,j})}{\gamma - (1 - ACC_{h,j})} \tag{10}$$

where γ is the no-information rate. We can calculate γ through fitting a model to a dataset that contains all possible combinations between features x'_i and target class labels y_i :

$$\gamma = \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n L(y_i, f(x'_{i'})) \tag{11}$$

Additionally, the no-information rate γ could be estimates as:

$$\gamma = \sum_{k=1}^K p_k (1 - q_k) \tag{12}$$

where p_k is the percentage of examples belonging to class k and observed in the dataset, and q_k is the percentage of examples that the classifier predicts to belong class k .

2.2.4. Performance Evaluation Metrics

There are four types of possible outcomes for classification tasks, true positive, true negative, false positive, and false negative. The definition of these four terms is listed in Table 1.

Table 1. Definition of terms.

Term	Definition
True Positive (TP)	The prediction is positive and it is actually positive.
False Positive (FP)	The prediction is positive but it is actually negative.
True Negative (TN)	The prediction is negative and it is actually negative.
False Negative (FN)	The prediction is negative but it is actually positive.

These four outcomes are often listed on the confusion matrix. The following confusion matrix (Table 2) is an illustration for the case of binary classification.

Table 2. Confusion matrix.

Actual \ Predict	Yes	No
	Yes	TP
No	FP	TN

Next, we will introduce some model evaluation metrics.

Accuracy (also known as recognition rate) is defined as the fraction of correct predictions. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Precision (also known as positive predictive value, PPV) is defined as the fraction of correct positive predictions among all of positive predictions.

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

Recall (also known as sensitivity, true positive rate, TPR) is defined as the ratio of true positive predictions with respect to all of the examples that truly belong in positive class.

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

F_β score consider both precision and recall together as an evaluation index. The β parameter allows us to control the trade-off of importance between precision and recall. $\beta < 1$ focuses more on precision while $\beta > 1$ focuses more on recall. When $\beta = 1$, it is called F_1 score.

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (16)$$

Brier score is used to check the goodness of a predicted probability score, whose values range between 0 and 1. For binary classification, the score is given by:

$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2 \quad (17)$$

where p_i is the prediction probability, and the term o_i is equal to 1 if the event occurred and 0 if not.

Receiver Operating Characteristic Curve (ROC Curve) is the plot between the true positive rate and false positive rate. Following (Figure 2) is an example of the ROC curve. The area under the ROC curve (AUC) is to measure how well the classifiers make correct predictions on the different thresholds.

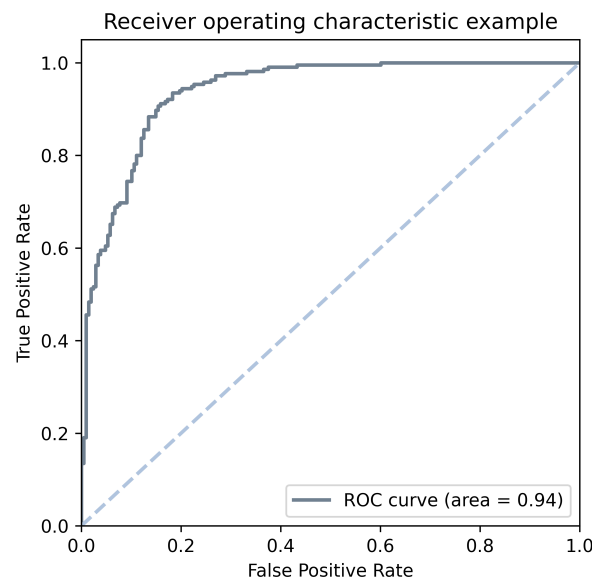


Figure 2. Receiver Operating Characteristic Curve.

2.3. Model Selection

With the development of machine learning, researchers proposed many efficient machine learning algorithms. For each algorithm, there are several hyperparameters that can be tuned to fit different datasets. Using different hyperparameters and algorithms to fit the training data sets results in different candidate models. As we are usually interested in obtaining the best-performing model from these candidate models, we need to find an

approach to evaluate their respective performance in order to rank them. Model selection is the process of selecting the best machine learning model from the candidate models, which are built based on the training dataset. It involves the selection of different types of models (e.g., KNN, SVM, RF, etc.) and the selection of models with different hyperparameters for a certain type (e.g., different kernels for SVM).

As mentioned before, it is essential to evaluate our model with new data to prevent the possibility of overfitting on the training set. However, in order to select the best model, we need to evaluate the candidate models while building the model. In light of the fact that we cannot evaluate the candidate models on the test set. Otherwise, we will obtain a model that performs best on the test set but may not generalize well in practice. To evaluate the model as we build and adjust the model, we create a third subset of the dataset, called the validation set. If we have plenty of data, which may be at least 1000 to infinite, we could straightforwardly create the validation set. To evaluate the model as we build and tune the model, we could randomly split the full dataset into training, validation, and test sets. Then, we would fit candidate models on the training set with a different configuration of hyperparameters and algorithms. Subsequently, we can evaluate the performance of candidate models on the validation set and select the winning model which performs best (model evaluation and selection). With the hyperparameters of the best model, we retrain it using the training + validation set, and the generalization performance of the final model is evaluated on the test set (model evaluation). If the performance on the test set is similar to the performance on the validation set, there is reason to believe that the model will perform well on future data. Finally, we retrain the model on the full dataset (training, validation, and test set) for production use.

However, we rarely have such sufficient datasets in practice. We mainly have two approaches, re-sample methods and analytical methods, to implement model selection for a limited size of the dataset [69].

2.3.1. Re-Sample Methods

For re-sample methods, we expand the sample size by repeating a random re-sampling training set and then compute the average of prediction error as the estimation. In general, we split the training dataset into sub-training and validation sets. Sub-training set is used to fit candidate models for different algorithms and hyperparameters. The validation set is used to evaluate these candidate models and select the best model. Model evaluation does not change, in which test set is still utilized to estimate the performance of the final selected model.

We can adopt the aforementioned methods (holdout, bootstrapping and cross-validation) of model evaluation to split the training dataset again. By far, the most widely used is the cross-validation method, which includes many subtypes. Here, nested cross-validation method [70] will be detail for an example. Up to now, we have two tasks: the first task is to select the best model across candidate algorithms and corresponding hyperparameters; and the second task is to estimate the generalized performance of the best model. The nested cross-validation method includes an inner loop and an outer loop. In the inner loop, the target is to select the best model, whereas, in the outer loop, the target is to estimate the generalization performance of the best model selected by the inner loop. Figure 3 illustrates the procedure of the nested cross-validation. It works as follows:

- **Step 1:** Randomly split the whole dataset into K equal folds (outer loop).
- **Step 2:** Select one of them as the test set, and the remaining $k - 1$ folds as the training set.
- **Step 3:** Randomly split the training set into K' equal sub-folds (inner loop).
- **Step 4:** Select one of the sub-folds as the validation set and the remaining $k' - 1$ folds as the sub-training set. Then we train candidate models under different algorithms and hyperparameters with the sub-training set. Next, we evaluate the performance of candidate models on the current validation set.
- **Step 5:** Repeat step 4 for k' times, so that each sub-fold has and only has one chance to act as the validation set, and the remaining sub-folds act as the sub-training set.

- **Step 6:** We then compute the average performance of candidate models on all validation sets and select the winning model with the best performance.
- **Step 7:** With the hyperparameters of the best model from Step 6, we retrain it with the whole training set and then evaluate the generalization performance of the best model on the current test set.
- **Step 8:** Repeat step 2 to step 6 for k times, so that each fold has and only has one chance to act as the test set, and the remaining folds act as the training set.
- **Step 9:** Report the average of generalization performance on all test sets as an estimate of the model performance.

Lastly, we retrain the best model using the whole dataset for deployment. For brevity, nested CV with K outer folds and K' inner folds is denoted as $K \times K'$ nested CV. Typical values for $K \times K'$ are 10×10 , 10×5 , or 5×5 , etc.

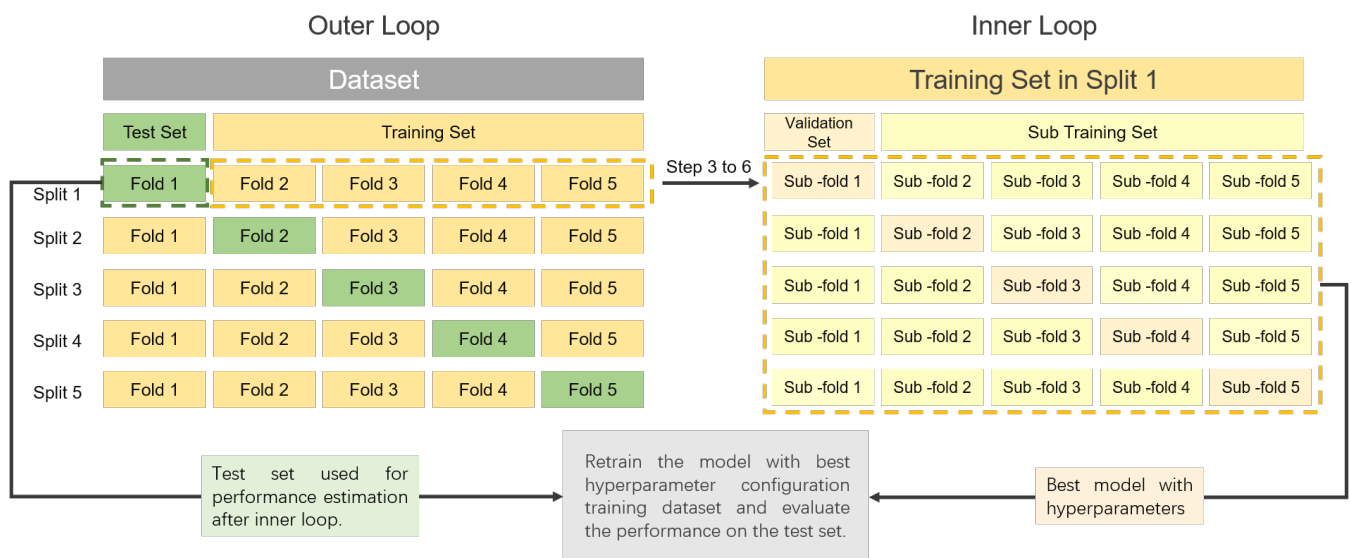


Figure 3. Nested 5×5 Cross-validation.

2.3.2. Analytical Measures

Compared with re-sample methods, the analytical methods not only evaluate model performance but also consider the model complexity. In addition, as analytical methods approximate the test error from the training error, which does not need to repeat several times, it could improve the efficiency of model selection. In this part, three typical used analytical criteria are introduced for model selection.

Akaike Information Criterion (AIC) is a scoring criterion to measure the performance of statistical models, named for the Japanese statistician Hiroji Akaike who proposed AIC in 1973 [71].

$$AIC = -2 \cdot \text{loglike} + 2 \cdot d \tag{18}$$

Formula (18) is a mathematical formulation of AIC, where loglike is the maximized log-likelihood; d is a measure of model complexity, such as the number of parameters for linear models. It is noted that the form of d for nonlinear and complex models differ and should be carefully derived. To use AIC for model selection, we simply choose the model with the smallest AIC over the set of models considered.

Bayesian Information Criterion (BIC), also known as the Schwarz criterion, was derived from Bayesian probability, and inferred by Schwarz [72]. Like AIC, it is applicable for models that are fitted by the maximum of likelihood. If we use the same formalism defined in Formula (18), the generic form of Bayesian Information Criterion is defined as follows:

$$BIC = -2 \cdot \text{loglike} + (\log n) \cdot d \tag{19}$$

It is straightforward to find that BIC is proportional to AIC. Compared with AIC, BIC punishes heavily on models, which possess more parameters and higher complexity. Although it looks similar, the original idea of BIC is not similar to AIC, but obtained from a Bayesian perspective.

Minimum Description Length (MDL) is motivated from an optimal coding viewpoint, proposed by Rissanen [73]. MDL recommends us to select the model from an information theory perspective. If we want to transmit our model and the prediction, a good solution from the view of coding is to encode the message with shortest length. According to Shannon's theorem [74], the length to describe our problem is:

$$Length = -\log Pr(\mathbf{y}|\theta, M, \mathbf{X}) - \log Pr(\theta|M) \quad (20)$$

In Formula (20), M is our model with θ as parameter. $Pr(\mathbf{y}|\theta, M, \mathbf{X})$ is the conditional probability of the model output with attribute \mathbf{X} . The first term of Formula (20) represents the average code length for transmitting the difference between the output of the model and the ground truth, whereas the second term represents the average code length for transmitting the model parameter vector θ .

One advantage of the analytical measure for the model selection approach is that it does not require a validation dataset. It means that all of the data can be used to build the model, and we can score the candidate models directly. However, the analytical measure also has the limitation of the inability to form general statistics across different types of models. For a more detailed discussion about analytical measures, the material can be obtained from [75].

2.3.3. Hyperparameter Tuning

The hyperparameters of machine learning algorithms enable the model to be tailored to different datasets. Therefore, hyperparameter tuning, which refers to the searching of an appropriate hyperparameter configuration, is an important process for the application of machine learning. Grid search, random search, Bayesian optimization, and meta-heuristic algorithms are most commonly used for hyperparameter tuning.

Grid search is an exhaustive search strategy exploring a grid of evenly spaced values. Generally, grid search can find the global optimum value by setting a large search range and a fine grid. It involves generating a uniform grid of hyperparameter configuration across the search space. With this search strategy, we simply build the model for each potential combination of all of the hyperparameters and evaluate the model to select the one which achieves the best results. The downside is that the number of potential hyperparameter combinations to be explored grows exponentially with the number of hyperparameters. It is quite inefficient to try all hyperparameter combinations one by one, which could take days or even weeks, especially on a large dataset.

Different from grid search, random search simply draws some random samples instead of trying all hyperparameter settings. This strategy randomly samples model hyperparameters following a sampling distribution (e.g., uniform) for a number of iterations. For each iteration, we build the model under a hyperparameter combination, which is randomly sampled from the aforementioned distribution. Subsequently, we evaluate each chosen hyperparameter configuration and select the best one. On account of randomness, it is not guaranteed that random search always finds the optimal solution.

Meta-heuristic algorithm is a generic optimization framework that can resolve almost all optimization problems as it is a problem independent. The iterative generation process of meta-heuristic algorithm realizes the robust searching mechanism by balancing exploration (diversification) and exploitation (intensification) under different intelligent concepts. Therefore, it enables the black-box optimization problem of hyperparameter tuning solvable with an optimal or near-optimal solution. Genetic algorithm [76], Particle Swarm optimization [77], Simulated Annealing [78], and Tabu Search [79] are already introduced for hyperparameter tuning.

Bayesian optimization for machine learning parameter tuning was proposed by J. Snoek (2012) [80]. It works under the assumption that the mapping between hyperparameter setting and generalization performance was sampled from a Gaussian process. We first construct the distribution by the observation of hyperparameters and corresponding generalization performance. Subsequently, the acquisition function was adopted to determine the next point of hyperparameters to evaluate the performance and add the observation to update the distribution. We iteratively repeat these two steps until converging to an optimum. In this setup, the information of the previous hyperparameter setting is included to adjust the exploring process.

2.4. Hypothesis Testing

Once we obtain the final model, we usually want to compare our method with the state-of-the-art methods to prove that it beats or performs as well as the advanced method. With model evaluation methods and performance metrics, it seems possible to compare the performance of the different models by first using an evaluation method to measure certain performance metrics of the models and then comparing the value of performance directly. However, the performance of models and the difference between models may be misleading because of the sampling error instead of essential differences. To make the performance of the model statistically significant, we will introduce hypothesis testing in this part.

Hypothesis testing is a statistical inference method to distinguish whether the results are due to sampling error or intrinsic differences. For hypothesis testing, we firstly compute a statistic from the samples and assume that it follows a certain distribution. If the probability of the statistic to obey this distribution is very low, we may reject the hypothesis; If not, we may accept it. For example, if we have a model with an average error rate which is ϵ_0 . In hypothesis testing, we may assume that the error rate is less than ϵ_0 . If the test result is consistent with the hypothesis, we accept the hypothesis; otherwise, we should reject the hypothesis as the error rate has a high probability to be greater than ϵ_0 .

Let us take the error rate under Student's t -test [81] for example. If we adopt k -fold cross-validation, the model evaluation process provides k error rates as we split the dataset into k folds. We denote these error rates with $\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_k$, then the average error rate and standard deviation are μ (Formula (21)) and σ (Formula (22)).

$$\mu = \frac{1}{k} \sum_{i=1}^k \hat{\epsilon}_i \quad (21)$$

$$\sigma = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (\hat{\epsilon}_i - \mu)^2} \quad (22)$$

We can compute the t -statistic value following Formula (23), which obeys a t -distribution of $k-1$ degree of freedom.

$$t_0 = \frac{\sqrt{k}(\mu - \epsilon_0)}{\sigma} \quad (23)$$

For $H_0 : \mu \leq \epsilon_0$ vs. $H_1 : \mu \geq \epsilon_0$, we fail to reject the hypothesis with level of significance α if $t_0 \in (-\infty, t_{\alpha, k-1})$ or p -value is greater than the α .

2.4.1. Paired t -Test

Paired t -test [82] is a specific Student's t -test, which is used when the two samples are matched or paired. It is a two-sided test for the null hypothesis that two samples have identical average values.

When comparing machine learning models using paired t -test, we firstly evaluate each model on the same k -fold cross-validation split of the dataset and compute a performance score for each split. For example, we use $\epsilon_1^A, \epsilon_2^A, \dots, \epsilon_k^A$, and $\epsilon_1^B, \epsilon_2^B, \dots, \epsilon_k^B$, to denote the test error rate of model A and model B in the same split. Secondly, we calculate the difference of each pair $d_i = \epsilon_i^A - \epsilon_i^B$. If these two models perform the same, the mean of their difference

should be zero. Therefore, we can compute the mean μ and the variance σ^2 of d_1, d_2, \dots, d_k . Next, we compute the t -statistic value by Formula (24).

$$t_0 = \left| \frac{\sqrt{k}\mu}{\sigma} \right| \quad (24)$$

For H_0 : Model A and Model B perform the same vs. H_1 : Model A and Model perform differently, we fail to reject the hypothesis with the level of significance α if t_0 is less than $t_{\alpha/2, k-1}$ or p -value is greater than the α . It means that there is no significant difference between model A and model B.

Valid use of paired t -test is based on the independence of each evaluation. However, in this case, the sub-training sets have overlap with each other, which means they lack independence in the evaluation. If we stick with this hypothesis test, it will lead to overestimating the probability of rejecting the hypothesis.

2.4.2. 5×2 Cross-Validation Paired t -Test

To address this problem, 5×2 cross-validation, which repeats the 2-fold cross validation five times, was adopted as the evaluation method. We randomly scramble the dataset before each 2 fold cross validation to ensure that each observation occurs only once in the training or test dataset. Then, the evaluation results are tested using a paired t -test. For 5×2 cross-validation paired t -test [82], the computation of statistic value is slightly different with paired t -test. If we define $d_i^j (i = 1, 2; j = 1, 2, \dots, 5)$ to denote the difference of test error rate between model A and model B in the i fold of j repetition, then the average performance in each repetition is $\mu_j = (d_1^j + d_2^j)/2$ and the variance is $\sigma_j^2 = (d_1^j - \mu_j)^2 + (d_2^j - \mu_j)^2$. With these denotation, we define the statistic of 5×2 cross-validation paired t -test in Formula (25).

$$t_0 = \left| \frac{\mu_1}{\sqrt{\frac{1}{5} \sum_{j=1}^5 \sigma_j^2}} \right| \quad (25)$$

The statistic obeys a T-distribution of 5 degrees of freedom. Similar with paired t -test, if the statistic t_0 is less than $t_{\alpha/2, 5}$, model A and Model B perform equivalently. Conversely, the performance of the two models is significantly different, and the one with a lower average error rate performs better.

2.4.3. Wilcoxon Signed-Rank Test

Wilcoxon signed-rank test [83] is a non-parametric hypothesis test, proposed by Frank Wilcoxon in 1945. It applies to the case of two related or paired samples to assess whether their populations have the same distribution. In the specific, it checks whether the difference between the paired observations comes from a population with a median of zero. We can utilize Wilcoxon signed-rank test to compare two different models A and B, at the statistic level as following steps. It should be noted that the denotations not mentioned in this section are the same as in paired t -test.

- **Step 1: Build the null hypothesis.** H_0 : the performance distributions of model A and model B are equal, H_1 : the performance distributions of model A and model B are not equal.
- **Step 2: Calculate the difference.** For $i = 1, 2, \dots, k$, calculate the difference of each pair $d_i = \epsilon_i^A - \epsilon_i^B$.
- **Step 3: Rank the difference.** Order the difference according to its absolute value $|d_i|$ from the smallest to largest value. Define $r(|d_i|)$ to denote the rank. The rank of smallest $|d_i|$ is 1; Ties (pairs with equal $|d_i|$) ranks equal to the average of the orders they cross; The differences equal to zero are omitted when ranking. For example,

if we have six difference values ($d_1 = 5, d_2 = 0, d_3 = -2, d_4 = 5, d_5 = 1, d_6 = 6$), the absolute values of them are $|d_1| = 5, |d_3| = 2, |d_4| = 5, |d_5| = 1, |d_6| = 6$. Therefore, the ranks of them are $r(|d_1|) = 3.5, r(|d_3|) = 2, r(|d_4|) = 3.5, r(|d_5|) = 1, r(|d_6|) = 5$.

- **Step 4: Compute the statistic.** W^+ is the sum of the ranks of the originally positive differences. Conversely, W^- is for negative differences. The test statistic of Wilcoxon signed-rank test is $W = \min(W^+, W^-)$. W can be compared to the critical value table for the Wilcoxon signed-ranks test in [84]. Let k' denote the number of pairs included in the ranking. If $W > W_{critical,k'}$, we reject the null hypothesis. For the instance of step 3, the value of W^+ is 13, and the value of W^- is 2. Therefore, W is 2. As $W_{critical,5}$ is 0 and $W > W_{critical,5}$, we reject H_0 and accept H_1 . It means the model with a higher performance score is statistically significant. Moreover, the sum of the positive difference ranks ($W^+ = 13$) is larger than the sum of the negative difference ranks ($W^- = 2$), showing a positive advantage from model A. Consequently, our analysis provides significant evidence that model A performs better than model B in statistical significance.
- **Step 5: Compute the z-score.** If $k' > 30$, we can implement a large sample approximation. For the population of statistic W , the mean is μ_W (Formula (26)) and the standard deviation is S_W (Formula (27)) [84].

$$\mu_W = \frac{k'(k' + 1)}{4} \tag{26}$$

$$S_W = \sqrt{\frac{k'(k' + 1)(2k' + 1)}{24}} \tag{27}$$

Therefore, the z-score is

$$z_0 = \frac{W - \mu_W}{S_W} \tag{28}$$

Then, we compare the obtained z_0 value to the critical value of normal distribution or calculate the p -value. As a general rule, we set the level of risk to be $\alpha = 0.05$. If p -value is less than 0.05 or the absolute value of z_0 is greater than $z_{\alpha/2}$, we will reject the null hypothesis.

2.4.4. McNemar’s Test

For models that are both very large and built for large datasets, it usually takes several days or weeks to train a single model. Therefore, it is impractical or expensive to perform multiple copies of the model. McNemar’s Test [82,85,86] is capable of comparing the models that can be executed only once. It is named for Quinn McNemar, who proposed it in 1947 [85].

To implement McNemar’s Test, we adopt the holdout method to train and test models A and B. For each model, we record the classification results on the test set and tabulate the outcomes on the following Table 3. It is the contingency table which lists the detail of misclassification by model A and B. For example, n_{00} is the number of samples misclassified by both model A and B; n_{01} is the number of samples misclassified by model A but not by model B.

Table 3. Contingency table.

	Model B	
Model A	Misclassification	Correct classification
Misclassification	n_{00}	n_{01}
Correct classification	n_{10}	n_{11}

Similar to before, we have the null hypothesis that the two models have no difference in performance. In other words, the error rates of these two models are the same, which means that $n_{01} = n_{10}$. Consequently, we build the statistic as Formula (29). The statistic follows the χ^2 distribution with 1 degree of freedom. At the significance level of 0.05, the critical value of 1 degree of freedom is 3.841 [87].

$$\chi_0^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{29}$$

Then, we compare the obtained value of χ_0^2 to the critical value ($\chi_{1-\alpha/2}^2$ and $\chi_{\alpha/2}^2$) of chi-square distribution or calculate the p -value. For the classical approach, if $\chi_0^2 < \chi_{1-\alpha/2}^2$ or $\chi_0^2 > \chi_{\alpha/2}^2$, we should reject the null hypothesis, which assumes that model A and B performs equally. For the p -value approach, if p -value $< \alpha$, we should reject the null hypothesis and state the conclusion that the performance of model A and B are significantly different. In this way, there is sufficient evidence at the significance level of $\alpha = 0.05$ to conclude that these two models perform at a different error rate level.

2.4.5. Friedman Test and Post-Hoc Test

If one model performs well on some data sets and poorly on others compared to other models, how can we tell if this model outperforms the others or not? Friedman test and the corresponding post hoc test are employed to explore the answer to this question.

The Friedman test is a non-parametric hypothesis test to compare multiple models on different datasets, which is proposed by Milton Friedman [88,89]. The procedure of Friedman test involves ranking the models for each dataset separately and calculate the Friedman statistic to infer whether these models perform differently. Suppose we compare k models on N datasets and let R_i represents the average rank of i th algorithm on all datasets. For example, if we have three models, A , B , and C , and four datasets, the best performing algorithm ranks 1, the second-best ranks 2 and the last one ranks 3. In the case of ties, the average of the ranks they across are assigned. Table 4 is an example of the ranking results, where $R_1 = 1.125$, $R_2 = 2.250$, $R_3 = 2.625$.

Table 4. Ranks of model comparison.

Rank \ Dataset	Model		
	Model A	Model B	Model C
Dataset 1	1.5	1.5	3
Dataset 2	1	3	2
Dataset 3	1	2.5	2.5
Dataset 4	1	2	3
Average Rank	1.125	2.250	2.625

The null hypothesis states that there is no difference among all models, which means that the average rank R_i should be equivalent. The average and variance of R_i are $(k + 1)/2$ and $(k^2 - 1)/12$, respectively. The variable χ_0^2 , defined by Formula (30), is distributed according to the χ^2 distribution with $k - 1$ degree of freedom when k and N are big enough.

$$\begin{aligned} \chi_0^2 &= \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \sum_{i=1}^k \left(R_i - \frac{k+1}{2} \right)^2 \\ &= \frac{12N}{k \cdot (k+1)} \left(\sum_{i=1}^k R_i^2 - \frac{k(k+1)^2}{4} \right) \end{aligned} \tag{30}$$

However, the above statistic is over conservative and a new statistic F_0 as Formula (31) is adopted.

$$F_0 = \frac{(N - 1)\chi_0^2}{N(k - 1) - \chi_0^2} \tag{31}$$

F_0 follows a F -distribution with $k - 1$ and $(k - 1)(N - 1)$ degrees of freedom. Next, we compare the obtained value of F_0 to the critical value $F_{\alpha}(k - 1, (k - 1)(N - 1))$ or calculate the p -value. For classical approach, if $F_0 < F_{1-\alpha/2}$ or $F_0 > F_{\alpha/2}$, we should reject the null hypothesis. For p -value approach, if p -value $< \alpha$, we should reject the null hypothesis and state the conclusion that there is at least one model has different performance at the significance level of $\alpha = 0.05$.

If the null hypothesis that all models have the same performance is rejected, it indicates that the performance of the models is significantly different, necessitating proceeding with a post hoc test to distinguish which models perform better than others. There are several available pathways to realize post hoc test. If we want to compare all the models with each other, the Nemenyi test [90,91] is a commonly used method. If we only aim to compare all models with a control model, such as comparing your proposed model with the state-of-the-art models, the Bonferroni correction procedure, step-up procedure, or step-down procedure are also appropriate [92–95]. Here, we take Nemenyi test as an example of post hoc test after Friedman test. A more detail description about other procedures, please refer to [96–98].

To implement the Nemenyi test, we first define a critical difference (CD) as Formula (32).

$$CD = q_{\alpha} \sqrt{\frac{k(k + 1)}{6N}} \tag{32}$$

where q_{α} (Table 5) is the critical value based on the Studentized range statistic divided by $\sqrt{2}$.

Table 5. Critical value q_{α} for the two-tailed Nemenyi test [97].

q_{α} \ k	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

If the difference between the average rank of two models exceeds the critical difference, the assumption that the two algorithms perform the same should be rejected with corresponding confidence. For the example in Table 4, $CD_{0.05} = 1.657$ and $CD_{0.10} = 1.414$; $|R_1 - R_2| = 1.125$, $|R_1 - R_3| = 1.500$ and $|R_2 - R_3| = 0.375$. With the significant level of 0.05, there is no difference in all of the models. With the significant level of 0.05, model A performs better than model B , as the difference of their average ranks exceeds 1.414.

3. Liquid Biopsy Components

During the formation and growth of primary tumors, cells undergo active release, necrosis, or apoptosis [99,100]. In these process, various components are released into the liquid, including circulating tumor cells, cell-free DNA, circulating tumor DNA, cell-free RNA, exosomes, and tumor educated platelets(TEPs) [101].

3.1. Circulating Tumor Cells

The presence of circulating Tumor Cells (CTCs) was firstly identified by Ashworth (Australia) in 1869 [102]. When Ashworth performed an autopsy on a metastatic breast cancer patient, cells similar to those from the primary tumor were found in the blood. CTCs are currently defined as the tumor cells that shed or migrate actively into the vessel

from the primary tumor or metastatic sites and then circulate in the bloodstream [103]. The opinion of tumor self-seeding suggests that CTCs can recirculate back, resulting in the possibility of metastases, which is responsible for the majority of deaths associated with cancer [104,105]. As the access to peripheral blood circulation is a prerequisite for distant metastasis of tumors [106], detection of tumor cells in blood will indicate the possibility of distant metastasis of tumors [107].

Although the content of CTCs is extremely rare, CTCs is still a potential alternative to invasive biopsies as an origin of tumor tissue for the detection and monitoring of cancers [108–111]. These circulating tumor cells can be enriched and detected via different technologies that take advantage of their physical and biological properties [112]. The technology to obtain these cells is an evolving field of research and is challenged by the ability to isolate CTCs in a condition that can be utilized for molecular analysis and propagation into CTC derived xenografts [113].

CTC is isolated from peripheral blood, which can avoid invasive and complex biopsy procedures. The culture of tumor cell lines takes a long time and is homogeneous, which cannot accurately reflect the genetic diversity and the changing tumor microenvironment. In contrast, CTCs-derived xenografts can reflect the biological characteristics of cancer more accurately, providing a visual window for studying the dynamic evolution of cancer and allowing monitoring of the longitudinal evolution of tumors at the molecular level.

As a marker of early diagnosis, CTCs also has some limitations. A reasonable and effective enrichment method is the most important and urgent problem to be solved. The main challenge is to obtain a sufficient number of CTCs that are optimally available for further evaluation. Besides, techniques for assessing the molecular characteristics of CTCs are still evolving, and the standard in this effort for clinical practice should be unified.

3.2. Cell-Free DNA and Circulating Tumor DNA

In 1948, Mandel and Metais, researchers from France, firstly found nucleic acids circulating in the human blood [114,115]. Circulating cfDNA refers to the DNA which is released into the blood by necrotic or apoptotic cells, or active release [116,117]. For cancer patients, part of cfDNA comes from tumor cells. This subpopulation of the cfDNA is ctDNA. In 1977, scientists firstly confirmed the presence of ctDNA in the blood of cancer patients [118]. ctDNA is single or double stranded [113] and comes from either living, dying tumor cells or CTCs [119–121]. The majority of cfDNA are released from normal cells. Therefore, ctDNA only occupies a small proportion of the cfDNA [101].

The concentration of cfDNA in the blood could increase owing to certain events such as cancer, autoimmune, smoking, pregnancy, intense exercise and tissue damaging therapies [122–128]. Likewise, the fraction of ctDNA may vary due to various factors [129]. Although ctDNA analysis provides a viable option for the diagnosis of early cancer, existing techniques cannot overcome the difficulties of sensitivity analysis. How to standardize the testing method is still a problem to be solved.

3.3. Cell-Free RNA

In 1993, Lee firstly discovered the miRNA [130], which are intracellular non-coding RNA molecules containing about 22 nucleotides. The miRNAs play an important signaling role by mediating the post-transcriptional silencing in various cellular activity [131]. Circulating or cell-free miRNA (cfRNA) refers to those miRNAs that identified in the biological fluids [131]. The high turnover rate of tumor cells needs the high expression of specific genes, leading to the large amounts generation of cfRNA [132]. Therefore, researchers identified the corresponding alteration in the blood of cancer patients [101,133].

The limitation of miRNA is reflected in the inconsistency in the selection of internal or external reference genes for quantitative detection; miRNAs from different sources, such as plasma, serum, whole blood, and exosomes, have differences in quality and quantity during the separation process; Some studies have small sample sizes, which may lead to

unreliable results. Therefore, the isolation and quantification of miRNA and the methods used for data analysis still need more verification.

3.4. Exosomes

Exosomes were first discovered in sheep reticulocytes in 1983 and named by Johnstone in 1987 [134]. It refers to the vesicles released by cells, containing an abundance of proteins, genetic information such as DNA and RNA, and other analytes [135]. With a diameter between 30 nm to 100 nm, it can be detected from plasma, saliva, urine, breast milk, hydrothorax, cerebrospinal fluid, semen and other body fluids [136]. Furthermore, it is stable in extreme pH (pH = 1–13) or freeze-thaw [137]. Since playing a key role in tumor growth and metastasis, the complicated impact of exosomes in cancer mechanism needs to be further studied. These concepts support the potential of exosomes and their components to be applied in the detection of cancer [138].

Exosomes have vesicles that enhance the stability of wrapped genetic components. The similarity between circulating exosomal miRNAs and tumor-derived miRNAs enables the former one potentially useful for screening tests for cancer. In addition, other genetic components inside the exosomes will enrich relevant research on tumor genetics. From the preliminary results obtained, the prospects are very promising. However, the technology of acquiring exosomes is still under development, which is also the main reason for limiting exosome-related research.

3.5. Tumor Educated Platelets

Platelets (also termed thrombocytes) are the second most abundant cell types in peripheral blood, existing as circulating anucleated cell fragments. The largest platelets are about 2–3 microns in diameter [139]. More recently, platelets are implicated a central role in the local and systemic responses to tumor growth [140,141]. Confrontation of platelets with tumor cells by transferring tumor-associated biomolecules ('education') is an emerging research field resulting in the term of tumor-educated platelets (TEPs).

4. Machine Learning Algorithms and Clinical Application in Early Cancer Detection based on Liquid Biopsy

Several machine learning algorithms are used to detect cancer based on the characteristics extracted from liquid biopsy. An overview of all relevant papers are listed in the supplementary document (Table: Summary of related publications) with the direct URL of dataset if available. This section discusses and reviews the publications of the most commonly used algorithms for early cancer detection in recent 10 years. As this systematic survey aims to report wide studies related to early cancer detection based on liquid biopsy incorporating machine learning algorithms, over 400 papers were searched using the following keywords: (liquid biopsy OR exosome OR circulating tumor cell OR circulating tumor DNA OR cell free DNA OR microRNA OR tumor educated platelet) AND (cancer OR carcinoma OR adenocarcinoma OR tumor OR malignancy OR malignant disease) AND (svm OR support vector machine). We searched four extensively used machine learning algorithms by replacing the last keyword. For each algorithm, we checked the top 100 relevant publications in recent 10 years according to the following four criteria. Figure 4 is the workflow of select publications.

- The research is about liquid biopsy.
- The research is about cancer detection.
- The research utilized corresponding machine learning method.
- For several models compared, we only consider the model which performs best.

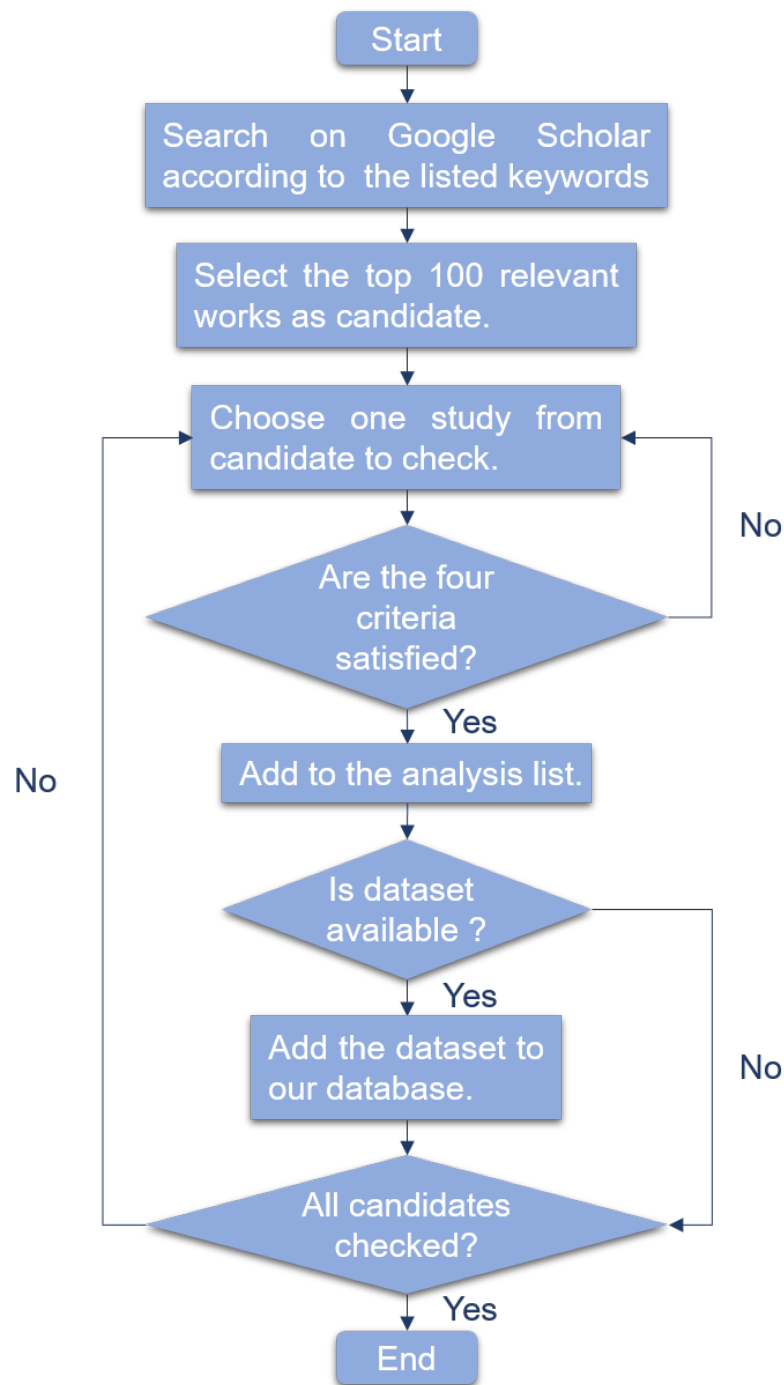


Figure 4. Workflow of search and select publications.

4.1. Traditional Machine Learning Algorithms

For traditional machine learning algorithms, we reviewed linear models, support vector machine and random forest.

4.1.1. Linear Models

Linear models are widely used for supervised learning because of the advantage of implementation simplicity and interpretability. Linear regression, logistic regression and LASSO are some examples of linear models.

A. Principle of Linear Model

Given an input data $\{X, y\}$ for $X = \{x_1, x_2, \dots, x_N\}$. Let $\hat{y} = model(X)$ denote the prediction made by a model for the given input. Coefficients (β) are parameters that define the model by assigning a coefficient to each input, and the bias or intercept is provided by an additional coefficient. The training data is used to estimate the coefficients of the logistic regression algorithm using a learning algorithm known as a maximum-likelihood estimation. The learning algorithm assumes data distribution and produces coefficients that minimize the error of probabilities of model prediction to those in the data.

The logistic regression model can be described with a matrix for the input data X , a vector for the output \hat{y} , and a vector for the coefficients β using linear algebra represented as the Formula (33).

$$\hat{y} = X \cdot \beta \quad (33)$$

Since the above representation is identical to linear regression, which produces real values as outputs instead of class labels, a nonlinear function is used to ensure that the output of the weighted sum is a value between 0 and 1.

Logistic regression uses the logistic function, also known as the sigmoid function, to ensure class labels' prediction. The sigmoid function is an S-shaped curve that maps a real-valued number x into a number between 0 and 1 using Equation (34).

$$f(x) = \frac{1}{1 + e^{-x}} \quad (34)$$

Therefore, for logistic regression, x in Equation (34) is replaced with the weighted sum given in Equation (35) to produce an output between 0 and 1 for two class labels 0 and 1.

$$\hat{y} = \frac{1}{1 + e^{-(X \cdot \beta)}} \quad (35)$$

The output from the model can be interpreted as a probability from a Binomial probability distribution function.

Least Absolute Shrinkage and Selection Operator (LASSO), also known as L1-norm, adds a regularization term which is used to penalize the less important features in a data by making their respective coefficient (β) zero, thereby shrinking their weights to zero. The less important features in Equation (33) having $\beta = 0$ are eliminated, thereby making LASSO useful for feature selection and the creation of simple models. It is beneficial for datasets with high dimensions and high correlation. L1-norm is given by Equation (36)

$$L1_{norm} = \lambda \sum_{m=0}^p \beta_m \quad (36)$$

where λ is the hyperparameter that controls the shrinkage. The bias of the model increases as λ increases while variance increases as λ decreases.

B. The Application of Linear Models in Early Cancer Detection

Linear models have been applied in many ways to detect several types of cancer, either recurrent or metastatic, in different parts of the body. Table 6 is an overview of relevant publications based on linear models.

Table 6. Overview of reference related to Linear Models.

Reference	Method	Dataset Available	URL For Dataset	Cancer Type	Sample Type	Biomarker
[142]	LR	Y	https://www.ncbi.nlm.nih.gov/gds/?term=GSE31682 , accessed on 29 June 2021	Ovarian	Blood	DNA methylation
[143]	LR	N		Oral cancer	Plasma	cfDNA
[144]	LR	N		Oral cancer	Blood	Exosomes
[145]	LASSO	N		Non-Small Cell Lung Cancer	Plasma	cfDNA
[146]	LR	N		Colorectal cancer	Blood	miRNA
[147]	LR, LASSO	Y	http://www.uni-koeln.de/med-fak/clcgp/ , accessed on 29 June 2021	Non-small cell lung carcinoma	Blood	cfDNA
[148]	LASSO	N		Lung cancer	Plasma	Exosomes
[149]	LASSO	Y	https://identifiers.org/ncbi/insdc.sra:SRP302308 , accessed on 29 June 2021	Breast cancer	Blood	cfDNA

Maltoni et al. [150] used a logistic regression model to evaluate the role of altered genes in breast cancer like HER2, PI3KCA for patient prognosis due to the possibility of their correlation with CF-DNA quantity. They collected serum samples from 58 non-relapsed and 21 relapsed patients and analyzed the samples for cfDNA integrity and quantity of all oncogenes. To determine the ability of these genes in predicting a relapse, the logistic regression on a two-marker combination produced an area under curve of 0.627 with a 95% confidence interval. With further clinical validity, the study speculates the potential of cfDNA detected as liquid biopsy in clinical practice.

Gene expression information of original tissues is contained in the nucleosome footprint of cfDNA. This information can be used in the prediction of response to chemotherapy. Yang et al. [149] utilized LASSO to evaluate transcription start site (TSS) regions coverage ability of genes. Based on cfDNA data of 85 healthy individuals and 85 individuals who are breast cancer patients, the coverage at the TSS regions was utilized for the classification of individuals into either having cancer or healthy. The LASSO model was repeated 100 times with a 5-fold cross-validation technique using the R package to prevent bias. A test was implemented using plasma from 30 healthy donors and 60 patients to validate the model independently. The model recorded a significant median AUC of 0.863 for the training cohort and 0.834 for the validation cohort. The model was able to avoid overfitting, as noticed in the recorded AUC. With the analysis, the use of cfDNA nucleosome footprints to predict neoadjuvant chemotherapy was highlighted and verified with the LASSO model. The study will improve personalized decision-making per patients' treatment.

Due to the advancement of lung cancer by the time it is diagnosed, it is the deadliest cancer in the world [151]. El-Khoury et al. [148] used the bootstrap sampling method and LASSO penalization to deduce the suitable combination of protein necessary for predicting outcome to improve early detection and patients' survival. With data comprising 93 healthy donors and 128 lung cancer patients, the level of plasma in 351 proteins was quantified, and the optimal threshold for the biomarker was selected. The validation of the panel was carried out with independent data of 49 healthy donors and 48 patients using logistic regression. With an AUC of 0.999, sensitivity of 0.992, specificity of 0.989, negative predictive value of 0.989 and positive predictive value of 0.992, lung cancer was detected irrespective of the cancer stage, making it possible to detect lung cancer earlier and aiding early treatment.

For early and accurate decisions on treatment strategies, an accurate diagnosis must be made. Therefore, it is vital to distinguish small cell lung cancer (SCLC) from non-small cell lung cancer (NSCLC). Non-small cell lung cancer can be further categorized as squamous cell carcinoma and inter alia adenocarcinoma. Raman et al. [147] collected public data containing 843 samples (small cell lung cancer = 68, squamous cell carcinoma = 351, and inter alia adenocarcinoma = 424) which were filtered based on histology. cfDNA

was extracted was further extracted from plasma. Five classifiers, including random forest, support vector machine, multinomial logistic regression with ridge regularization, multinomial logistic regression with elastic net regularization, and multinomial logistic regression with lasso regularization were evaluated with the data using a leave-one-out cross-validation method. Due to the inability of some classifiers to deal with class imbalance, the authors used a random sampling method to make the number of samples in all classes equal to 68 to make the number of training samples equal to 204. Multinomial logistic regression with ridge regularization, based on iterative one-vs.-all receiver operating curve, had the best performance with a mean area under curve of 0.936. The coefficients of the logistic regression model detected that the prominent regions which differentiate non-small lung cell cancer from small lung cell cancer are located at the chromosome arm, and tumor fraction is a determinant of the prediction probability.

Cucchiara et al. [145], working with the metastatic case of EGFR-positive NSCLC reports the possibility of using the combination of liquid biopsy and radiomics to suggest management of the disease. This can be done by detecting new mutations early. Liquid biopsy is easy to perform, minimally invasive and can be done repeatedly to extract valuable information. cfDNA acquired from plasma of seven metastatic patients was analyzed using digital droplet PCR, and radiomic analysis was also done using computed tomography images. The authors were able to compare the EGFR mutation dynamics in cfDNA with the radiomic features. They used a logistic LASSO regression model to estimate the correlation between the variation in the radiomics features and the EGFR mutation status using a 27-fold Monte Carlo cross-validation method. The model implemented a feature reduction, and maximum likelihood estimation was done for the remaining features. Based on these performance analyses, an early decision can be made for treatment strategy. Although the authors found no significant relationship between the mutational status and tumor volume, there was also no discovered association between the clinical outcomes and the radiomic signatures.

Wei et al. [146] pointed out the need to have less invasive strategies for the early prognosis and detection of colorectal cancer to avoid distant metastasis. The authors extracted extracellular vesicles from plasma samples and used nanoparticle tracking analysis, transmission electron microscopy with western blotting to identify the extracellular vesicles. The samples contained 37 colorectal cancer patients, 22 colorectal adenoma patients and 42 non-cancerous control participants. It was discovered that circulating EV-miR-193a-5p can efficiently distinguish the three classes. Especially with an AUC of 0.752, it can distinguish colorectal cancer patients from the two other classes and with an AUC of 0.759, it can distinguish colorectal cancer from non-cancer. This shows circulating EV-miR-193a-5p can identify colorectal cancer than precancerous lesions. In addition, due to the importance of age factor in colorectal cancer, a logistic regression model was implemented to integrate the age with a cutoff of 55 years and circulating EV-miR-193a-5p. The integration of the age factor increased the area under curve from 0.752 to 0.775 and 0.759 to 0.795 for distinguishing colorectal cancer patients from the two other classes and colorectal cancer from non-cancer, respectively. The integration of the age factor using the model can quickly identify colorectal cancer in high-risk individuals.

Oral cancer, being one of the most frequent cancer in the world, Lin et al. [143] identified the correlation between the progression of oral squamous cell carcinoma and cfDNA. The identification of the biomarkers is essential to improve diagnosis and treatment. Plasma was extracted from 121 oral cancer patients and 50 individuals for control while ensuring that the cfDNA size distribution is similar in oral cancer patients and control donors. Analyses on the dataset revealed that the mean concentration of cfDNA in oral cancer patients was significantly higher than that of the control group. The adjusted odds ratios were determined using binary logistic regression analysis, and a confidence interval of 95% was achieved. With a statistical significance test of $p < 0.05$, the study established the relationship between cfDNA and oral cancer.

Due to the role that serum exosome plays in the development of cancer, Li et al. [144] identified protein content in serum exosome based on 30 samples. The samples included oral cancer patients with lymph node metastasis, oral cancer patients with no lymph node metastasis, and healthy controls. Oral cancer patients have a high rate of lymph node metastasis [152]. A binary logistic regression analysis was carried out to compare the use of four biomarkers (ApoA1, CXCL7, PF4V1, F13A1) and their combinations based on the area under curve. This study deduced that the four biomarkers from serum exosomes could help diagnose oral cancer-lymph node metastasis.

Due to the lack of early detection and resistance to chemotherapy, ovarian cancer is the most lethal cancer in gynecology [153,154]. Li et al. [142] performed a two-stage epigenome-wide association study to identify methylation biomarkers for epithelial ovarian cancer. The authors selected 24 cancer cases, and 24 age-frequency matched control cases for genome-wide methylation profiling, and 206 cancer cases with 205 age-frequency matched control cases. Independent t -test and χ^2 test was used for the continuous and categorical variables, respectively. The correlation between the blood cell counts and the DNA methylation was estimated using Pearson correlation analysis. A logistic regression model was further built for the differentially methylated cpG sites in the validation stage, and it was evaluated based on the receiver operating characteristics curves. With the study, the identified set of blood-derived DNA methylation signatures and its association with epithelial ovarian cancer will serve as a tool for the early detection of ovarian cancer.

Linear models have been successfully applied to different cancer types, including breast cancer, colorectal cancer, oral cancer, lung cancer, etc. Ranging from classification to the selection of important features for further prognosis, the application of linear models as machine learning tools is important.

4.1.2. Support Vector Machine

Support Vector Machine (SVM) [155] is a supervised learning method for solving data mining problems, first proposed by Cortes and Vapnik in 1995. It aims to build a decision boundary, which is known as the hyperplane, to separate different classes. The positive samples and negative samples each have the closest point to the hyperplane. SVM distinguishes different classes by maximizing the distance between these two points to the hyperplane.

A. Principle of SVM

If the data instances are $\{x_i, y_i\}$ for $i = 1, 2, 3, \dots, N$, where $x_i \in R^d$ and $y_i \in \{1, -1\}$. The two classes in the training data can be separated by a hyperplane $H: \mathbf{w}^T \cdot \mathbf{x} + b = 0$. Furthermore, there are two hyperplanes $H_1: \mathbf{w}^T \cdot \mathbf{x} + b = 1$ and $H_2: \mathbf{w}^T \cdot \mathbf{x} + b = -1$ parallel to H . The positive and negative samples, which are closest to H , just fall on H_1 and H_2 , respectively. Such samples are support vectors. Margin is defined as the distance between H_1 and H_2 in Formula (37).

$$\text{Margin} = \frac{2}{\|\mathbf{w}\|} \quad (37)$$

SVM aims to learn an optimal separating hyperplane H to maximize the margin (minimize \mathbf{w}), while keeping all the points correctly classified. This problem can be summarized as Formula (38).

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{\|\mathbf{w}\|^2}{2} \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i \end{aligned} \quad (38)$$

For non-separable data, slack variable ζ_i is defined to allow data samples to violate the margin or even misclassified. In Formula (39), C is the penalty parameter.

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1, \\ & \xi_i \geq 0, \quad \forall i \end{aligned} \quad (39)$$

When the true model of the dataset is nonlinear, we can map the input data $\mathbf{x} \in R^d$ into a new high dimensional space $\mathbf{z} \in R^d$ employing a nonlinear mapping $\mathbf{z} = \Phi(\mathbf{x})$. After mapping, the problem can be summarized as Formula (40).

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \cdot \Phi(\mathbf{x}_i) + b) \geq 1, \\ & \xi_i \geq 0, \quad \forall i \end{aligned} \quad (40)$$

To solve this problem, we need to rewrite the primal problem into its dual form.

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \\ & \sum_{i=1}^N \alpha_i y_i = 0, \quad \forall i \end{aligned} \quad (41)$$

In Formula (41), α_i is the Lagrange Multiplier. The SVM dual problem contains the inner product of $\Phi(\mathbf{x}_i)$, which is the high-dimensional feature vector. To simplify the calculation, kernel function is defined to replace the inner product as Formula (42).

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \quad (42)$$

B. The Application of SVM in Early Cancer Detection

As a traditional and popular machine learning method, SVM was widely used for early cancer detection. An overview of relevant reference to SVM is provided in Table 7.

Patrick et al. [156] reported a work of glioblastoma detection utilizing SVM with radial basis kernel. In this study, 1158 miRNAs collected from blood were analyzed. They applied SVM and filter based feature selection method to determine a suitable subset of miRNA biomarkers and achieved their best result based on 180 miRNAs with an accuracy of 81%, specificity of 79%, and sensitivity of 83%. Additionally, 52 miRNAs were significantly distinguished by unpaired Student's *t*-test. On this basis, miR-128 and miR-342-3p stand out significantly with a *p*-value of 0.025 under correcting for multiple testing by Benjamini-Hochberg adjustment. This work revealed the possibility of miR-128, miR-342-3p and other important miRNA as biomarkers to detect glioblastom based on the analyses of 20 patients and 20 healthy individuals. It is also an instance of the effectiveness of SVM on a small sample dataset with high dimensions.

Table 7. Overview of reference related to Support Vector Machine.

Reference	Method	Dataset Available	URL For Dataset	Cancer Type	Sample Type	Biomarker
[156]	SVM	N		Glioblastoma	Blood	miRNA
[141]	SVM	Y	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68086 , accessed on 29 June 2021	6 Cancers	Blood	TEP-RNA
[157]	PSO + SVM	Y	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89843 , accessed on 29 June 2021	Non-Small-Cell Lung Cancer	Blood	TEP-RNA
[158]	SVM vs. PCA vs. LDA	N		Oral cancer	Saliva	Exosomes
[159]	PSO + SVM	Y	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107868 , accessed on 29 June 2021	2 Cancers	Blood	TEP-RNA
[160]	SVM	N		prostate-cancer	Blood	Extracellular vesicles
[161]	SVM vs. RF vs. LASSO	Y	https://bigd.big.ac.cn/search/?dbId=&q=PRJCA001138 , accessed on 29 June 2021	3 Cancers	Urine	cfDNA
[162]	SVM	N		Esophageal cancer	Plasma	cfDNA
[163]	PSO + SVM	N		Sarcoma	Blood	TEP-RNA
[164]	SVM	N		Lung Cancer	Serum	miRNA
[165]	SVM + SFLA vs. RF vs. KNN vs. GPC vs. GNB vs. GBM vs. SVM vs. LASSO vs. Elastic Net	Y	https://www.nature.com/articles/s41467-020-18965-w#data-availability , accessed on 29 June 2021	7 Cancers	Plasma	cfDNA

In 2015, Thomas Wurdinger's team from the Netherlands published a study in Cancer Cell showing that mRNA from tumor-educated platelets (TEPS) is potential for diagnosis of various cancers and differentiation of cancer types [141]. This is the first time that the term of tumor-educated platelet proposed. They identified that 1453 mRNAs increased and 793 mRNAs decreased in TEPs compared with healthy platelets. Further analysis indicated that the increased TEP mRNAs were involved in biological processes such as vesicle-mediated transport and the cytoskeletal protein binding while the decreased mRNAs were involved in RNA processing and splicing. A pan-cancer classification based on SVM was implemented, distinguishing 228 patients in 6 cancers from 55 healthy individuals with 96% accuracy. Additionally, TEP mRNA profiles are also demonstrated to be effective in distinguishing the specific tumor type. Besides, they found that the platelet samples of patients possess distinct therapy-guiding markers confirmed in matching tumor tissue. In their further study [157], this team combined particle-swarm optimization (PSO) and SVM to detect non-small-cell lung cancer based on TEPs. PSO was utilized to identify the optimal biomarker panels from large amounts liquid biosources and to tune the parameter of SVM. They termed this pipeline PSO-enhanced thromboSeq. In 2019, they reevaluated the publicly available dataset in [157] and further validate the performance on a new platelet-RNA-sequencing dataset from a healthy donor (HD) and lower-grade glioma (LGG) samples [159]. In this manuscript, the authors not only provided a new dataset but also disclosed the code and state the operation of the code step by step. Heinhuis et al. [163] generalized the pipeline of PSO-enhanced thromboSeq to identify the biomarker for sarcoma on a dataset with 160 samples, achieving a diagnostic accuracy of 87% and AUC of 0.93.

Cario et al. [166] diagnosed oral cancer based on the Fourier-transform infrared (FTIR) spectra of salivary exosomes. The dataset is the whole saliva samples collected from 21 oral cancer patients and 13 healthy individuals. By analyzing the absorbance spectra, they found a number of differences between normal and cancer samples, including changes in

the conformations of proteins, lipids and nucleic acids. Based on these findings, this work adopted the spectra absorbance bands between the 900 cm^{-1} and 3700 cm^{-1} , the ratios and the area under the absorbance spectrum of different three certain band as the input features of classifiers. Principal component analysis–linear discriminant analysis (PCA–LDA) and SVM are included as the discrimination models. In terms of accuracy, SVM achieved a training accuracy of 100% and a cross-validation accuracy of 89%. PCA–LDA showed an accuracy of 95%.

Sunkara et al. [160] presented a centrifugal device for isolation of extracellular vesicles (EVs) from whole-blood. SVM was utilized to analyze the 8 biomarkers to detect 43 prostate-cancer patients from 30 healthy individuals. HSP90 achieved the highest sensitivity (86%), accuracy (88%), specificity (90%), and AUC (0.92) of all the test markers.

Guangzhe et al. [161] applied SVM to detect urothelial carcinoma (UC) from 65 patients with urothelial carcinoma, 58 with kidney cancer, 45 with prostate cancer, and 95 normal individuals by analyzing copy number alterations of urinary cfDNA. In this work, the random forest was first utilized to select the top 50 features. After feature selection, RF, SVM and LASSO were compared and SVM with linear kernel outperformed the other two models. The authors defined UCdetector as a combination of the 50 CNA features selected by the RF and the SVM classifier with linear kernel. UCdetector achieved the AUC of 0.959 under 10 repeats of random splitting on this dataset. Further validation on an independent dataset comprising 24 normal samples and 28 UC patients was implemented. The UCdetector distinguishes UC with an AUC of 0.888. To test the clinical sensitivity of selected 50 CNA features, the authors applied UCdetector on the 410 patients from TCGA and 90 patients from Chinese UTUC WGS data. UCdetector could accurately identify the upper tract urothelial cancers at the AUC of 0.996. Furthermore, the concordance performance of urinary cfDNA was reported to be more sensitive than the urinary sediment. This recent work recognized the top 50 important CNA features from 5000 original features and achieved satisfying performance on different datasets, even on tissue samples from TCGA. For further comment, it demonstrates the power of feature selection based on RF and the identity capacity of SVM.

Shicai et al. [162] combined SALP-seq and SVM as a pipeline to discover new cfDNA-based biomarkers for esophageal cancer. They studied the reads density of all promoters and found high reads density in normal samples and extremely low-density cancer samples on 49 genes. Of these, 34 genes are newly discovered biomarkers. The author further validated the relationship between esophageal cancer and these biomarkers on a dataset with 163 esophageal cancer samples and 11 normal samples. Moreover, 88 important regions associated with esophageal cancer were screened out from the whole genome and 54 of these, located in distal intergenic and proximal regulatory regions, were inferred to be potential diagnostic and prognostic markers for cancer. Additionally, 37 mutated genes, unique in pre-operation patients, were also discovered from a large amount of mutations in thousands of genes in pre- and post-operated esophageal cancer samples and normal samples. In this work, 103 epigenetic markers and 37 genetic markers were discovered for esophageal cancer. Finally, SVM was adopted to detect cancer samples based on 88 cancer-associated regions and achieved a high AUC of 1.0.

Zhang et al. [164] designed a DNA molecular computation platform involving SVM to analyze miRNA profiles from serum samples. They validate the performance based on clinical serum samples from 8 healthy individuals and 14 lung cancer patients with an accuracy of 86.4%.

In our recently published work [165], we proposed an Adaptive Support Vector Machine (ASVM) method by combining Shuffled Frog Leaping Algorithm and SVM for pancreatic and subsequent tumor origin analysis. The proposed method was firstly validated on a cell-free DNA dataset with 423 sample records. We observed an improvement of AUC from 0.832 for SVM to 0.938 for ASVM. The proposed ASVM was competitive or outperformed the other six machine learning models on both the original dataset and additional two datasets.

4.1.3. Random Forest

Random Forest (RF) is an ensemble machine learning approach consisting of randomly selected decision tree subsets for classification and regression. Leo Breiman introduced a random forest algorithm using bootstrapping in the random tree selection method in the early 2000s [167]. It was an enormous improvement in classification and regression machine learning accuracy. It uses the bag of random tree classifications to the ensemble and evaluates the overall classification for the given training and test data set.

A. Principle of Random Forest

The basic principle of the RF algorithm is the bootstrapping aggregation of randomly selected decision trees from given data observations. According to the Breiman [167] RF algorithm, it deals with classification and regression tasks using the random forest to learn. For the general RF regression estimation, Let X is the random input vector, where $X \in R$. We need to predict the response Y using the following Equation (43).

$$m(X) = E[Y|X = x] \quad (43)$$

Now training sample $D_n = ((X_1, Y_1), \dots, (X_N, Y_N))$ for independent input and goal data pairs of D_n dataset that construct estimate with random tree $T m_n : T \rightarrow R$ for m Function (43). Now RF consists of M numbers of random regression trees. The predicted estimation value (m_n) for the j_{th} tree input x is defined as:

$$m_n(x; \Theta_j, D_n) = \sum_{i \in D_n(\Theta_j)} \frac{\mathbb{1}_{X_i \in A_n(x; \Theta_j, D_n)} Y_i}{Z_n(x; \Theta_j, D_n)} \quad (44)$$

where $D_n(\Theta_j)$ is the set of input data points for each tree and $A_n(x; \Theta_j, D_n)$ is the data elements of each input observation, $Z_n(x; \Theta_j, D_n)$ is preselected data for input tree construction from $A_n(x; \Theta_j, D_n)$. Now final random forest estimation is:

$$m_{M, n}(x; \Theta_1, \dots, \Theta_M, D_n) = \frac{1}{M} \sum_{j=1}^M m_n(x; \Theta_j, D_n) \quad (45)$$

For RF supervised classification, it can classify both binary and multi-class datasets [168]. Let input vector and Y is a random class vector with class value 0,1. Now we can predict label Y from input X and D_n dataset. Therefore, RF binary classifier can obtain from the random classification trees as:

$$m_{M, n}(x; \Theta_1, \dots, \Theta_M, D_n) = \begin{cases} 1 & \text{if } \frac{1}{M} \sum_{j=1}^M m_n(x; \Theta_j, D_n) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (46)$$

B. The application of Random Forest in Early Cancer Detection

In recent years, several studies employed RF for early cancer detection from different liquid biopsy data. An overview of relevant references is provided in Table 8.

Song CX et al. [169] applied the RF algorithm to predict lung cancer, pancreatic cancer, and HCC using cfDNA 5-Hydroxy-methyl-cytosine (5hmC) mark in blood plasmas. This study collected the whole-genome cfDNA 5hmC signatures from 49 patients with seven cancer types and eight healthy individuals for their sequence analysis using the 5hmC library. After sequence analysis, copy number variation (CNV) has estimated using PopSV 1.0.0 R package. The RF algorithm and Gaussian Mclust model applied using gene bodies and DhMRs for cancer type prediction with different cancer stages from forty HCC, pancreatic lung cancer patients, and healthy samples. RF algorithm achieved the highest accuracy, 87.5% and 92%, for two feature sets, gene bodies, and DhMRs, while Mclust prediction accuracies are 82.5% and 90%.

Table 8. Overview of reference related to Random Forest.

Reference	Method	Dataset Available	URL For Dataset	Cancer Type	Sample Type	Biomarker
[169]	RF and Mclust	Y	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81314 , accessed on 29 June 2021	7 cancer	Blood	cfDNA
[170]	LR and RF	Y	https://science.sciencemag.org/highwire/filestream/704651/field_highwire_adjunct_files/1/aar3247_Cohen_SM_Tables-S1-S11.xlsx , accessed on 29 June 2021	8 cancer	Blood	cfDNA and protein Biomarkers
[171]	RF	Y	https://github.com/bergerm1/GenomeDerivedDiagnosis , accessed on 29 June 2021	22 Cancers	Plasma	cfDNA
[172]	RF	Y	https://doi.org/10.5281/zenodo.3715312 , accessed on 29 June 2021	intracranial tumors	Plasma	cfDNA
[173]	RF	N		Lung cancer	Serum	miRNA
[174]	RF	Y	https://www.ebi.ac.uk/pride/archive?keyword=PXD018301 , accessed on 29 June 2021	5 cancer	Plasma	EVP
[175]	RF	N		hepatocellular carcinoma	Blood	cfDNA
[176]	RF	N		gastrointestinal cancers	Plasma	cfDNA

Cohen et al. [170] designed CancerSEEK method for early cancer detection using circulating protein biomarkers and mutation in cfDNA from multi-analyte blood test results consist of 1817 blood plasma samples with 1005 cancer patients with eight different type of cancer such as colorectum, liver, ovary, esophagus, pancreas, stomach, breast and lung cancer, and 812 healthy individuals. CancerSEEK method is usually applied for both binary and localize cancer detection from the mentioned blood test. For binary cancer detection, logistic regression (LR) classifier with 10-folds cross-validation involved using omega cfDNA score and eight protein biomarkers. CancerSEEK employed the random forest (RF) classifier with 10-folds cross-validation using omega cfDNA score, 39 protein biomarkers, and patient gender for cancer type localization. CancerSEEK achieved 70% average sensitivity for eight cancer types with 99% specificity, including the sensitivity levels of five cancer types from 69% to 98%.

Later, Nassri et al. [172] applied the binomial RF classifier for gliomas cancer detection with other types of cancer using cfDNA methylation profile from plasma samples. They achieved the highest sensitivity with an AUC value of 0.990.

Penson et al. [171] used the RF machine learning classifier for cancer type detection on tissue biopsies and then validated on two plasma ctDNA datasets. They achieved 73.8% accuracy with 5-folds cross-validation for 22 cancer types, including the highest accuracy of 95%, 87%, and 85% for uveal melanoma, glioma, and colorectal cancer, respectively. It also obtained 75% accuracy from plasma ctDNA genome analysis.

Wang et al. [176] used the RF model for gastrointestinal cancer detection using plasma cfDNA data. The gastrointestinal cancers include the gall bladder, stomach, esophagus, colon, bile duct, pancreas, liver, and rectum cancers. This study also analyzed the cfDNA profile of hepatocellular carcinoma, colorectal cancer, pancreatic cancer patients, and healthy individuals. It obtained the AUC of 0.960, 0.890, 0.910, respectively, using the RF model with 10-folds cross-validation.

Zhang et al. [173] employed the RF algorithm for feature selection and classification of early-stage lung cancer using circulating miRNA from the liquid biopsy with SMOTE oversampling technique. They achieved the highest 96.60% accuracy value (AUC = 0.996) with a maximum of 13 miRNA features. RF identified the top five circulating miRNA features for early lung cancer detection.

Peng et al. [177] applied the RF prediction model for early-stage pancreatic cancer detection of diabetic patients using blood-based plasma biomarkers. The RF model has identified the best biomarkers for early-stage pancreatic cancer patients considering the AUC measure using the leave-one-out cross-validation technique and obtained AUC values of 0.850 and 0.810 with and without the CA19-9 biomarker.

Hoshino et al. [174] employed the RF classifier to identify the biomarkers from extracellular vesicles and particle (EVP) for cancer detection. The research shows that EVP proteins are able to serve as biomarkers for early cancer detection and tumor origin detection. This study used 426 human EVP profile samples for cancer detection and achieved over 90% sensitivity and 88% specificity on both training set and test set.

4.2. Deep Learning

In cancer detection, traditional machine learning algorithms usually rely heavily on the representation of the selected information [178]. However, in most cases, it is difficult to give an effective feature set. In addition, manually designing features requires a lot of manpower and time in complex tasks. Therefore, deep learning came into being. When training the model, deep learning utilizes high-level features to represent low-level features, that is, to build complex concepts by combining simple concepts [179]. Since our survey focuses on commonly used algorithms based on the characteristics extracted from liquid biopsy, and the extracted features are substantially tabular data (i.e., a sample by feature matrix); therefore, here we just discuss the basic deep learning model without introducing the spatial-aware or time-aware blocks in computer vision or natural language processing. A classic case of deep learning is the multilayer perceptron (MLP, also named as a neural network (NN)).

A. Principle of MLP

A multilayer perceptron is a function that maps a set of input values to output values, and this function consists of many simpler functions [180]. It can be considered that each function gives a new representation of the input. Generally, an MLP consists of three different blocks, which are the input layer, hidden layer, and output layer. A 3-layer MLP architecture can be seen in Figure 5. Herein, the input layer accepts the features, that is, the experiment results from liquid biopsy. Hidden layers are between the input and output layers. Each hidden node in the hidden layer is a perceptron (with its own set of weights). Hidden layer can extract a feature pattern from the previous layer and model more complex functions [181]. Hidden layer is also called a fully-connected layer or dense layer. Output layer outputs the final prediction results (e.g., the binary description of sick or healthy).

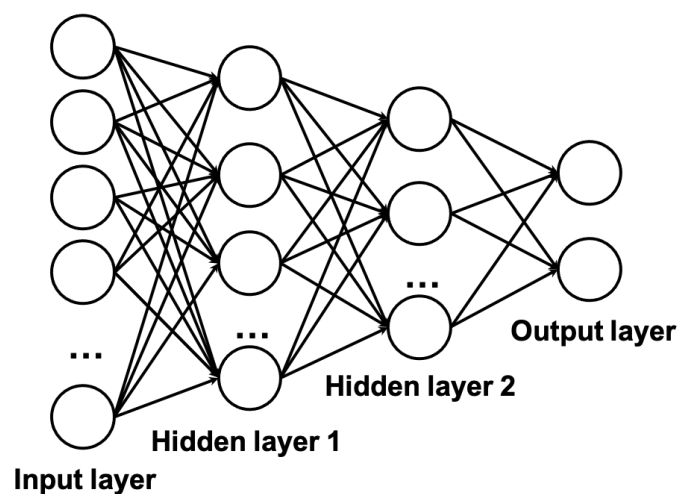


Figure 5. A simple 3-layer MLP architecture.

Formally, for one layer:

$$\mathbf{h} = f(\mathbf{W}^T \mathbf{x})$$

where W is the weight matrix (one column for each node). \mathbf{x} is the input from the previous layer, \mathbf{h} is the output to the next layer. $f(a)$ is the activation function that is applied to each dimension to get the output. In most cases, Rectifier Linear Unit (ReLU) is utilized as the activate function in hidden layers since it is faster and easier to train with [182]. ReLU is an activation function defined as the positive part of its argument:

$$f(x) = x^+ = \max(0, x)$$

where x is the input to a node. ReLU can obtain sparse representation since most nodes will output zero. The activation functions for the output layer can be Softmax for both binary and multi-class classification. Softmax function is defined as:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

where z is the input vector. The term on the bottom of the formula is the normalization term which ensures that all the output values of the function will be sum to 1, thus constituting a valid probability distribution. When training, we can utilize Cross-Entropy Loss Function to optimize the neural network, which is formulated as:

$$L = \frac{1}{N} \sum_i L_i = \frac{1}{N} \sum_i \sum_{c=1}^M -y_{ic} \log(p_{ic})$$

where M is the number of class; y_{ic} is the indicator variable (0 or 1), if the category is the same as the category of sample i , it is 1, otherwise it is 0; p_{ic} is the predicted probability that the observation sample i belongs to the category c .

B. The Application of MLP in Early Cancer Detection

In the past several years, the utilization of neural networks in cancer detection can be summarized into two categories: feature engineering and classification. For feature engineering, a neural network is usually performed to remove the input's noise and extract the most representative features that can best describe the subjects' attributes. This step can also be called feature extraction or dimension reduction [183]. Regarding the neural network for classification, the architecture of the neural networks used in cancer detection varied in depth (shallow and deep architecture), loss function and other parameters [184]. An overview of relevant references is provided in Table 9.

Table 9. Overview of reference related to Deep Learning.

Reference	Method	Dataset Available	URL For Dataset	Cancer Type	Sample Type	Biomarker
[185]	ANN	N		Lung cancer	Blood	Others
[186]	CNN	N		CTCs Detection	Blood	CTCs
[187]	CNN	N		Lung cancer	Blood	cfDNA
[188]	AODE, deep learning, decision tree, naive Bayes	Y	https://science.sciencemag.org/highwire/filestream/704651/field_highwire_adjunct_files/1/aar3247_Cohen_SM_Tables-S1-S11.xlsx , accessed on 29 June 2021	8 Cancers	Blood	Multianalyte

In 2014, Krzysztof et al. [185] introduced Artificial Neural Networks (ANNs) to early lung cancer detection. The dataset, provided by Diagnostic and Monitoring of Tuberculosis and Illness of Lungs Ward in Kuyavia and Pomerania Centre of the Pulmonology

(Bydgoszcz, Poland), includes 193 patients involving 48 features (i.e., blood test results, age, sex, etc.). The training set, validation set and test set are randomly splitted with 97, 48, and 48 samples, respectively. Different ANNs are trained and analysed to achieve the best performance. The optimal architecture was composed of 3-layers MPL (48 input neurons, 9 neurons in hidden layer, 2 output neurons) with learning rate 0.1, epochs 17, linear function for hidden layer, and tangent function for output layer. The obtained classification accuracy is 97.91% and AUC is 0.9983. However, as the dataset is limited, we can not ascertain the high performance is on account of model generalization or the certain dataset splitting.

In 2016, Yunxiang et al. [186] developed a deep 6-layers Convolutional Neural Network (CNN) to detect circulating tumor cells from blood results. A training methodology utilizing k-means clustering was adopted to explore the most representative samples to build the classification boundary. The filter parameters, bias terms, and weights were automatically optimized by back-propagation under 0.1 learning rate setting. The experiment results show that the proposed CNN is superior to SVM on F-score. To validate the effectiveness of proposed training strategy, a comparison experiment was implemented, which indicates that the F-score of CNN increased from 91.2% to 97% with the training strategy. For SVM, the performance only reached 75.4% without the training method and increased to 78.4% after adopting it.

In 2018, Kothen-Hill et al. [186] proposed a CNN-based framework, named Kittyhawk, to distinguish the true cancer mutations from sequencing artifacts even in ultra-low variant allele frequencies (VAFs) at the level of 10^{-4} . Kittyhawk is an 8-layer CNN with a fully-connected output layer, learning rate 0.1, momentum 0.9, and minibatch size 256. Kittyhawk initially proposed the read representation which combines the aligned genomic context, the quality scores, and the complete read sequence. The proposed method was first examined on 201,730 reads in the validation set, achieving an average performance of F1-score 0.961. Subsequently, the generalization capability was demonstrated on the independent lung cancer case with 0.92 F1-score reported.

In 2019, Ka-Chun Wong et al. [188] collected blood test records from 1817 patients to build three deep learning models to detect cancers as the front-line detector in a binary manner (i.e., cancer or normal). Since their datasets have standard and well-crafted input features, they directly adopted the deep feedforward neural networks with one hidden layer, two hidden layers, and three hidden layers (namely, DeepLearning1, DeepLearning2, and DeepLearning3, respectively) for model construction. The remaining training setting follows the default settings of WEKA. However, the performance of the deep learning methods cannot scale to full performance once the specificity level is relaxed.

5. Discussion

From the perspective of machine learning, we find out that even simple machine learning algorithms such as linear models can lead to a high-quality performance for liquid biopsy-based diagnosis for several common cancer types. However, there is no perfect model that performs the best on all datasets. Besides, the performance of machine learning models is diverse under different hyperparameter settings. To ensure the stability, we recommend Bayesian optimization for hyperparameter tuning after considering performance and runtime. With a hyperparameter optimization strategy, the machine learning model is adaptive to different datasets.

In addition, among all the machine learning models, the most popular and widely used are conventional algorithms. This is partly due to the barriers between biology and computer science; it is also partly due to the dataset size limitation. In the current data amount context, the traditional machine learning model such as linear models, support vector machine and random forest are still dominant in early cancer detection for their training speed and robustness on small dataset. We hope that the all-sided review of machine learning procedures and corresponding code demos presented in this survey can act as a reference guide. Definitely, advanced machine learning algorithms could also

be applied for exploring latent biomarkers and the complicated relationship in order to further improve the performance. However, model generalization and complexity have to be balanced in a fair manner.

As limited with the sample size and the interpretability of deep learning models, deep learning was not popular in liquid biopsy cancer detection. From the related studies in the past several years, we can observe that, with the increased data amount from the liquid biopsy, deep learning methods are likely to outperform conventional machine learning methods. However, there are also concerns. The first concern is that deep learning is vulnerable to overfitting. Therefore, regularization, dropout, and early stopping are utilized to prevent neural networks from overfitting. Besides, the birth of batch normalization improves the model baselines and speeds up all structures [189]. Due to the variance shift conflict between dropout and batch normalization, these two methods are not recommended to be adopted simultaneously at bottlenecks except for high-dimensional data. Another concern is the black-box nature of deep learning [190]. Since the hidden layers between input and output layers are complex, it is difficult to extract the most important features and match them with the biological explanation. The explainable framework design is vital to introduce machine learning models into clinical application [191]. In general, the technique for explaining predictions can be categorized into backpropagation-based methods and perturbation-based methods [192]. The recent successes of explainable framework [191–195] do shed light on its promising ability. Therefore, we are still optimistic with its development in cancer detection in the future.

From the perspective of liquid biopsy components, we find out that machine learning is extensively used for single-omics analysis. However, a single type of circulating biomarker seldom fully reveals the essence of tumor occurrence. Therefore, multi-omics detection is another promising direction for early cancer detection and treatment monitoring. The exploration competence of machine learning can enable the capability to figure out the complex causal relationships between different molecular measurements. Therefore, the integration of machine learning methods and multi-omics, including genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics, provides unprecedented opportunities to understand the underlying mechanism of tumor occurrence and early detection.

6. Conclusions

In this survey, we have presented an overview of machine learning protocols and the applications of different machine learning algorithms in the context of early cancer detection based on liquid biopsy. Additionally, we provided code demos for the aforementioned approaches in each procedure of machine learning. Based on the survey of over 400 papers, we have identified that the early cancer detection based on liquid biopsy has been tackled by different machine learning algorithms, which have been applied to multiple cancer types (e.g., pancreatic cancer, hepatocellular carcinoma, breast cancer, oral cancer, etc.) for a wide variety of component (e.g., circulating tumor cells (CTCs), cell-free DNA (cfDNA), circulating tumor DNA (ctDNA), cell-free RNA (cfRNA), exosomes, and Tumor Educated Platelets (TEPs)).

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/life11070638/s1>.

Author Contributions: L.L.: Project administration, Implementation of the code demo, Writing—most of original draft and review; X.C.: Writing—original draft of Deep learning part and review & editing; O.O.P.: Writing—original draft of Linear Models part; W.Z.: Writing—review & editing; S.R.: Writing—original draft of Random Forest part; Z.-R.T.: Writing—review & editing; K.-C.W.: Writing—review & editing and Funding acquisition. All authors have read and agreed to the published version of the manuscript.

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 is also supported by the National Natural Science Foundation of China under Grant No. 32000464.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

- Nahid, A.A.; Kong, Y. Involvement of machine learning for breast cancer image classification: A survey. *Comput. Math. Methods Med.* **2017**, *2017*, 3781951. [[CrossRef](#)]
- Wild, C.; Weiderpass, E.; Stewart, B. *World Cancer Report: Cancer Research for Cancer Prevention*; IARC Press: Lyon, France, 2020; pp. 181–188.
- Society, A. Global cancer facts and figures 4th edition. *Am. Cancer Soc.* **2018**, *1*, 1–73.
- Cree, I.A.; Uttley, L.; Woods, H.B.; Kikuchi, H.; Reiman, A.; Harnan, S.; Whiteman, B.L.; Philips, S.T.; Messenger, M.; Cox, A.; et al. The evidence base for circulating tumour DNA blood-based biomarkers for the early detection of cancer: A systematic mapping review. *BMC Cancer* **2017**, *17*, 697. [[CrossRef](#)]
- Chen, X.; Gole, J.; Gore, A.; He, Q.; Lu, M.; Min, J.; Yuan, Z.; Yang, X.; Jiang, Y.; Zhang, T.; et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **2020**, *11*, 1–10.
- WHO. *Guide to Cancer Early Diagnosis*; WHO: Geneva, Switzerland, 2017.
- Crowley, E.; Di Nicolantonio, F.; Loupakis, F.; Bardelli, A. Liquid biopsy: Monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* **2013**, *10*, 472. [[CrossRef](#)] [[PubMed](#)]
- Shinozaki, M.; O'Day, S.J.; Kitago, M.; Amersi, F.; Kuo, C.; Kim, J.; Wang, H.J.; Hoon, D.S. Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. *Clin. Cancer Res.* **2007**, *13*, 2068–2074. [[CrossRef](#)] [[PubMed](#)]
- Zhou, J.; Shi, Y.H.; Fan, J. Circulating cell-free nucleic acids: Promising biomarkers of hepatocellular carcinoma. In *Seminars in Oncology*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 39, pp. 440–448.
- Cohn, S.L.; Pearson, A.D.; London, W.B.; Monclair, T.; Ambros, P.F.; Brodeur, G.M.; Faldum, A.; Hero, B.; Iehara, T.; Machin, D.; et al. The International Neuroblastoma Risk Group (INRG) classification system: An INRG task force report. *J. Clin. Oncol.* **2009**, *27*, 289. [[CrossRef](#)] [[PubMed](#)]
- Diaz Jr, L.A.; Bardelli, A. Liquid biopsies: Genotyping circulating tumor DNA. *J. Clin. Oncol.* **2014**, *32*, 579. [[CrossRef](#)] [[PubMed](#)]
- Cai, X.; Janku, F.; Zhan, Q.; Fan, J.B. Accessing genetic information with liquid biopsies. *Trends Genet.* **2015**, *31*, 564–575. [[CrossRef](#)]
- The, L.O. Liquid cancer biopsy: The future of cancer detection? *Lancet Oncol.* **2016**, *17*, 123.
- Molina-Vila, M.A.; Mayo-de Las-Casas, C.; Gimenez-Capitan, A.; Jordana-Ariza, N.; Garzón, M.; Balada, A.; Villatoro, S.; Teixido, C.; Garcia-Pelaez, B.; Aguado, C.; et al. Liquid biopsy in non-small cell lung cancer. *Front. Med.* **2016**, *3*, 69. [[CrossRef](#)]
- Strotman, L.N.; Millner, L.M.; Valdes, R.; Linder, M.W. Liquid biopsies in oncology and the current regulatory landscape. *Mol. Diagn. Ther.* **2016**, *20*, 429–436. [[CrossRef](#)] [[PubMed](#)]
- Zhang, W.; Chen, X.; Wong, K.C. Noninvasive early diagnosis of intestinal diseases based on artificial intelligence in genomics and microbiome. *J. Gastroenterol. Hepatol.* **2021**, *36*, 823–831. [[CrossRef](#)] [[PubMed](#)]
- Chen, M.; Zhao, H. Next-generation sequencing in liquid biopsy: Cancer screening and early detection. *Hum. Genom.* **2019**, *13*, 34. [[CrossRef](#)]
- Peeters, M.; Price, T.; Boedigheimer, M.; Kim, T.W.; Ruff, P.; Gibbs, P.; Thomas, A.; Demonty, G.; Hool, K.; Ang, A. Evaluation of emergent mutations in circulating cell-free DNA and clinical outcomes in patients with metastatic colorectal cancer treated with panitumumab in the ASPECCT study. *Clin. Cancer Res.* **2019**, *25*, 1216–1225. [[CrossRef](#)]
- Cescon, D.W.; Bratman, S.V.; Chan, S.M.; Siu, L.L. Circulating tumor DNA and liquid biopsy in oncology. *Nat. Cancer* **2020**, *1*, 276–290. [[CrossRef](#)]
- Di Meo, A.; Bartlett, J.; Cheng, Y.; Pasic, M.D.; Yousef, G.M. Liquid biopsy: A step forward towards precision medicine in urologic malignancies. *Mol. Cancer* **2017**, *16*, 80. [[CrossRef](#)]
- Heitzer, E.; Perakis, S.; Geigl, J.B.; Speicher, M.R. The potential of liquid biopsies for the early detection of cancer. *NPJ Precis. Oncol.* **2017**, *1*, 1–9. [[CrossRef](#)]
- Ilie, M.; Hofman, V.; Long, E.; Bordone, O.; Selva, E.; Washetine, K.; Marquette, C.H.; Hofman, P. Current challenges for detection of circulating tumor cells and cell-free circulating nucleic acids, and their characterization in non-small cell lung carcinoma patients. What is the best blood substrate for personalized medicine? *Ann. Transl. Med.* **2014**, *2*, 107.
- Montani, F.; Marzi, M.J.; Dezi, F.; Dama, E.; Carletti, R.M.; Bonizzi, G.; Bertolotti, R.; Bellomi, M.; Rampinelli, C.; Maisonneuve, P.; et al. miR-Test: A blood test for lung cancer early detection. *JNCI J. Natl. Cancer Inst.* **2015**, *107*. [[CrossRef](#)]

24. Zhang, S.; Zhang, C.; Yang, Q. Data preparation for data mining. *Appl. Artif. Intell.* **2003**, *17*, 375–381. [[CrossRef](#)]
25. Huang, J.; Li, Y.F.; Xie, M. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Inf. Softw. Technol.* **2015**, *67*, 108–127. [[CrossRef](#)]
26. García, S.; Ramírez-Gallego, S.; Luengo, J.; Benítez, J.M.; Herrera, F. Big data preprocessing: Methods and prospects. *Big Data Anal.* **2016**, *1*, 9. [[CrossRef](#)]
27. Pendharkar, P.C.; Subramanian, G.H.; Rodger, J.A. A probabilistic model for predicting software development effort. *IEEE Trans. Softw. Eng.* **2005**, *31*, 615–624. [[CrossRef](#)]
28. Kosti, M.V.; Mittas, N.; Angelis, L. Alternative methods using similarities in software effort estimation. In Proceedings of the 8th International Conference on Predictive Models in Software Engineering, Lund, Sweden, 21–22 September 2012; pp. 59–68.
29. Rodríguez, D.; Sicilia, M.; García, E.; Harrison, R. Empirical findings on team size and productivity in software development. *J. Syst. Softw.* **2012**, *85*, 562–570. [[CrossRef](#)]
30. Myrtveit, I.; Stensrud, E.; Olsson, U.H. Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Trans. Softw. Eng.* **2001**, *27*, 999–1013. [[CrossRef](#)]
31. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Data preprocessing for supervised learning. *Int. J. Comput. Sci.* **2006**, *1*, 111–117.
32. Patro, S.; Sahu, K.K. Normalization: A preprocessing stage. *arXiv* **2015**, arXiv:1503.06462.
33. Wu, E.Q.; Hu, D.; Deng, P.Y.; Tang, Z.; Cao, Y.; Zhang, W.M.; Zhu, L.M.; Ren, H. Nonparametric bayesian prior inducing deep network for automatic detection of cognitive status. *IEEE Trans. Cybern.* **2020**. [[CrossRef](#)] [[PubMed](#)]
34. Wu, E.Q.; Lin, C.T.; Zhu, L.M.; Tang, Z.; Jie, Y.W.; Zhou, G.R. Fatigue Detection of Pilots' Brain Through Brains Cognitive Map and Multilayer Latent Incremental Learning Model. *IEEE Trans. Cybern.* **2021**. [[CrossRef](#)]
35. 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](#)]
36. Chen, C.C.; Schwender, H.; Keith, J.; Nunkesser, R.; Mengersen, K.; Macrossan, P. Methods for identifying SNP interactions: A review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 1580–1591. [[CrossRef](#)]
37. Khalid, S.; Khalil, T.; Nasreen, S. A survey of feature selection and feature extraction techniques in machine learning. In Proceedings of the IEEE 2014 Science and Information Conference, Warsaw, Poland, 24–26 September 2014; pp. 372–378.
38. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. CSUR* **2017**, *50*, 1–45. [[CrossRef](#)]
39. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [[CrossRef](#)]
40. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417. [[CrossRef](#)]
41. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [[CrossRef](#)]
42. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [[CrossRef](#)]
43. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [[CrossRef](#)]
44. Bernstein, M.; De Silva, V.; Langford, J.C.; Tenenbaum, J.B. *Graph Approximations to Geodesics on Embedded Manifolds*; Technical Report; Citeseer: Princeton, NJ, USA, 2000.
45. Zhou, P.; Hu, X.; Li, P.; Wu, X. Online feature selection for high-dimensional class-imbalanced data. *Knowl. Based Syst.* **2017**, *136*, 187–199. [[CrossRef](#)]
46. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: Berlin/Heidelberg, Germany, 2015.
47. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 856–863.
48. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [[CrossRef](#)]
49. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
50. Weir, B.S.; Hill, W.G. Estimating F-statistics. *Annu. Rev. Genet.* **2002**, *36*, 721–750. [[CrossRef](#)]
51. Liu, H.; Setiono, R. Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, Herndon, VA, USA, 5–8 November 1995; pp. 388–391.
52. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [[CrossRef](#)] [[PubMed](#)]
53. Reunanen, J. Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.* **2003**, *3*, 1371–1382.
54. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [[CrossRef](#)]
55. Kim, S.j.; Koh, K.; Lustig, M.; Boyd, S.; Gorinevsky, D. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res.* **2007**, *8*, 1519–1555.
56. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [[CrossRef](#)]
57. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [[CrossRef](#)]
58. Motoda, H.; Liu, H. Feature selection, extraction and construction. *Commun. IICM* **2002**, *5*, 2.
59. Neshatian, K.; Zhang, M.; Andraea, P. A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming. *IEEE Trans. Evol. Comput.* **2012**, *16*, 645–661. [[CrossRef](#)]

60. Mahanipour, A.; Nezamabadi-pour, H.; Nikpour, B. Using fuzzy-rough set feature selection for feature construction based on genetic programming. In Proceedings of the 2018 3rd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), Bam, Iran, 6–8 March 2018; pp. 1–6.
61. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.12808.
62. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [[CrossRef](#)]
63. Braga-Neto, U.M.; Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **2004**, *20*, 374–380. [[CrossRef](#)]
64. James, G.M. Variance and bias for general loss functions. *Mach. Learn.* **2003**, *51*, 115–135. [[CrossRef](#)]
65. Moreno-Torres, J.G.; Sáez, J.A.; Herrera, F. Study on the impact of partition-induced dataset shift on k -fold cross-validation. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1304–1312. [[CrossRef](#)]
66. Efron, B. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Rev.* **1979**, *21*, 460–480. [[CrossRef](#)]
67. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316–331. [[CrossRef](#)]
68. Efron, B.; Tibshirani, R. Improvements on cross-validation: The 632+ bootstrap method. *J. Am. Stat. Assoc.* **1997**, *92*, 548–560.
69. Hélié, S. An introduction to model selection: Tools and algorithms. *Tutor. Quant. Methods Psychol.* **2006**, *2*, 1–10. [[CrossRef](#)]
70. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 1–8. [[CrossRef](#)]
71. Akaike, H. Information theory and an extension of maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, AS, USA, 2–8 September 1973; pp. 267–281.
72. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]
73. Rissanen, J. A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **1983**, *11*, 416–431. [[CrossRef](#)]
74. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
75. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
76. Holland, J.H.; Reitman, J.S. Cognitive systems based on adaptive algorithms. In *Pattern-Directed Inference Systems*; Elsevier: Amsterdam, The Netherlands, 1978; pp. 313–329.
77. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
78. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [[CrossRef](#)]
79. Glover, F.; Laguna, M. Tabu search. In *Handbook of Combinatorial Optimization*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 2093–2229.
80. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. *arXiv* **2012**, arXiv:1206.2944.
81. Student. The probable error of a mean. *Biometrika* **1908**, *6*, 1–25. [[CrossRef](#)]
82. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [[CrossRef](#)]
83. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics* **1945**, *1*, 80–83. [[CrossRef](#)]
84. Corder, G.W.; Foreman, D.I. *Nonparametric Statistics for Non-Statisticians*; Wiley: Hoboken, NJ, USA, 2011.
85. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)]
86. Everitt, B.S. *The Analysis of Contingency Tables*; CRC Press: Boca Raton, FL, USA, 1992.
87. Wilson, E.B.; Hilferty, M.M. The distribution of chi-square. *Proc. Natl. Acad. Sci. USA* **1931**, *17*, 684. [[CrossRef](#)]
88. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701; reprinted in *J. Am. Stat. Assoc.* **1939**, *34*, 109. [[CrossRef](#)]
89. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **1940**, *11*, 86–92. [[CrossRef](#)]
90. Nemenyi, P. Distribution-free multiple comparisons (doctoral dissertation, princeton university, 1963). *Diss. Abstr. Int.* **1963**, *25*, 1233.
91. Hollander, M.; Wolfe, D.A.; Chicken, E. *Nonparametric Statistical Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 751.
92. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
93. Dunn, O.J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64. [[CrossRef](#)]
94. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **1988**, *75*, 383–386. [[CrossRef](#)]
95. Gibbons, J.D.; Chakraborti, S. *Nonparametric Statistical Inference*; CRC Press: Hoboken, NJ, USA, 2020.
96. Shaffer, J.P. Multiple hypothesis testing. *Annu. Rev. Psychol.* **1995**, *46*, 561–584. [[CrossRef](#)]
97. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
98. Garcia, S.; Herrera, F. An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.

99. Gasch, C.; Bauernhofer, T.; Pichler, M.; Langer-Freitag, S.; Reeh, M.; Seifert, A.M.; Mauermann, O.; Izbicki, J.R.; Pantel, K.; Riethdorf, S. Heterogeneity of epidermal growth factor receptor status and mutations of KRAS/PIK3CA in circulating tumor cells of patients with colorectal cancer. *Clin. Chem.* **2013**, *59*, 252–260. [[CrossRef](#)] [[PubMed](#)]
100. Jahr, S.; Hentze, H.; Englisch, S.; Hardt, D.; Fackelmayer, F.O.; Hesch, R.D.; Knippers, R. DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **2001**, *61*, 1659–1665. [[PubMed](#)]
101. Alimirzaie, S.; Bagherzadeh, M.; Akbari, M.R. Liquid biopsy in breast cancer: A comprehensive review. *Clin. Genet.* **2019**, *95*, 643–660. [[CrossRef](#)]
102. Ashworth, T. A case of cancer in which cells similar to those in the tumours were seen in the blood after death. *Aust. Med. J.* **1869**, *14*, 146.
103. Imamura, T.; Komatsu, S.; Ichikawa, D.; Kawaguchi, T.; Miyamae, M.; Okajima, W.; Ohashi, T.; Arita, T.; Konishi, H.; Shiozaki, A.; et al. Liquid biopsy in patients with pancreatic cancer: Circulating tumor cells and cell-free nucleic acids. *World J. Gastroenterol.* **2016**, *22*, 5627. [[CrossRef](#)] [[PubMed](#)]
104. Kim, M.Y.; Oskarsson, T.; Acharyya, S.; Nguyen, D.X.; Zhang, X.H.F.; Norton, L.; Massagué, J. Tumor self-seeding by circulating cancer cells. *Cell* **2009**, *139*, 1315–1326. [[CrossRef](#)]
105. Rossi, G.; Mu, Z.; Rademaker, A.W.; Austin, L.K.; Strickland, K.S.; Costa, R.L.B.; Nagy, R.J.; Zagonel, V.; Taxter, T.J.; Behdad, A.; et al. Cell-free DNA and circulating tumor cells: Comprehensive liquid biopsy analysis in advanced breast cancer. *Clin. Cancer Res.* **2018**, *24*, 560–568. [[CrossRef](#)]
106. Hayes, D.F.; Cristofanilli, M.; Budd, G.T.; Ellis, M.J.; Stopeck, A.; Miller, M.C.; Matera, J.; Allard, W.J.; Doyle, G.V.; Terstappen, L.W. Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival. *Clin. Cancer Res.* **2006**, *12*, 4218–4224. [[CrossRef](#)]
107. Peitzsch, C.; Tyutyunnykova, A.; Pantel, K.; Dubrovskaya, A. Cancer stem cells: The root of tumor recurrence and metastases. In *Seminars in Cancer Biology*; Elsevier: Amsterdam, The Netherlands, 2017; Volume 44, pp. 10–24.
108. Pantel, K.; Alix-Panabières, C. Circulating tumour cells in cancer patients: Challenges and perspectives. *Trends Mol. Med.* **2010**, *16*, 398–406. [[CrossRef](#)] [[PubMed](#)]
109. Mocellin, S.; Hoon, D.; Ambrosi, A.; Nitti, D.; Rossi, C.R. The prognostic value of circulating tumor cells in patients with melanoma: A systematic review and meta-analysis. *Clin. Cancer Res.* **2006**, *12*, 4605–4613. [[CrossRef](#)]
110. Mehlen, P.; Puisieux, A. Metastasis: A question of life or death. *Nat. Rev. Cancer* **2006**, *6*, 449–458. [[CrossRef](#)]
111. Nagrath, S.; Sequist, L.V.; Maheswaran, S.; Bell, D.W.; Irimia, D.; Ullkus, L.; Smith, M.R.; Kwak, E.L.; Digumarthy, S.; Muzikansky, A.; et al. Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **2007**, *450*, 1235–1239. [[CrossRef](#)] [[PubMed](#)]
112. Alix-Panabières, C.; Pantel, K. Circulating tumor cells: Liquid biopsy of cancer. *Clin. Chem.* **2013**, *59*, 110–118. [[CrossRef](#)]
113. Mamdani, H.; Ahmed, S.; Armstrong, S.; Mok, T.; Jalal, S.I. Blood-based tumor biomarkers in lung cancer for detection and treatment. *Transl. Lung Cancer Res.* **2017**, *6*, 648. [[CrossRef](#)] [[PubMed](#)]
114. Buscaill, L.; Bournet, B.; Cordelier, P. Role of oncogenic KRAS in the diagnosis, prognosis and treatment of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 1–16. [[CrossRef](#)]
115. Mandel, P. Les acides nucléiques du plasma sanguin chez 1 homme. *CR Seances Soc. Biol. Fil.* **1948**, *142*, 241–243.
116. Spellman, P.T.; Gray, J.W. Detecting cancer by monitoring circulating tumor DNA. *Nat. Med.* **2014**, *20*, 474–475. [[CrossRef](#)]
117. Vendrell, J.A.; Mau-Them, F.T.; Béganton, B.; Godreuil, S.; Coopman, P.; Solassol, J. Circulating cell free tumor dna detection as a routine tool for lung cancer patient management. *Int. J. Mol. Sci.* **2017**, *18*, 264. [[CrossRef](#)]
118. Leon, S.; Shapiro, B.; Sklaroff, D.; Yaros, M. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.* **1977**, *37*, 646–650. [[PubMed](#)]
119. Anker, P.; Mulcahy, H.; Chen, X.Q.; Stroun, M. Detection of circulating tumour DNA in the blood (plasma/serum) of cancer patients. *Cancer Metastasis Rev.* **1999**, *18*, 65–73. [[CrossRef](#)] [[PubMed](#)]
120. Stroun, M.; Lyautey, J.; Lederrey, C.; Olson-Sand, A.; Anker, P. About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release. *Clin. Chim. Acta* **2001**, *313*, 139–142. [[CrossRef](#)]
121. van der Vaart, M.; Pretorius, P.J. The origin of circulating free DNA. *Clin. Chem.* **2007**, *53*, 2215. [[CrossRef](#)]
122. Breitbach, S.; Tug, S.; Simon, P. Circulating cell-free DNA. *Sport. Med.* **2012**, *42*, 565–586. [[CrossRef](#)] [[PubMed](#)]
123. Devos, T.; Tetzner, R.; Model, F.; Weiss, G.; Schuster, M.; Distler, J.; Steiger, K.V.; Grutzmann, R.; Pilarsky, C.; Habermann, J.K.; et al. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clin. Chem.* **2009**, *55*, 1337–1346. [[CrossRef](#)]
124. Bulicheva, N.; Fidelina, O.; Mkrtumova, N.; Neverova, M.; Bogush, A.; Bogush, M.; Roginko, O.; Veiko, N. Effect of cell-free DNA of patients with cardiomyopathy and rDNA on the frequency of contraction of electrically paced neonatal rat ventricular myocytes in culture. *Ann. N. Y. Acad. Sci.* **2008**, *1137*, 273. [[CrossRef](#)] [[PubMed](#)]
125. Hu, W.; Yang, Y.; Zhang, L.; Yin, J.; Huang, J.; Huang, L.; Gu, H.; Jiang, G.; Fang, J. Post surgery circulating free tumor DNA is a predictive biomarker for relapse of lung cancer. *Cancer Med.* **2017**, *6*, 962–974. [[CrossRef](#)] [[PubMed](#)]
126. Lee, Y.J.; Yoon, K.A.; Han, J.Y.; Kim, H.T.; Yun, T.; Lee, G.K.; Kim, H.Y.; Lee, J.S. Circulating cell-free DNA in plasma of never smokers with advanced lung adenocarcinoma receiving gefitinib or standard chemotherapy as first-line therapy. *Clin. Cancer Res.* **2011**, *17*, 5179–5187. [[CrossRef](#)]

127. Tug, S.; Helmig, S.; Menke, J.; Zahn, D.; Kubiak, T.; Schwarting, A.; Simon, P. Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cell. Immunol.* **2014**, *292*, 32–39. [[CrossRef](#)]
128. Chaudhuri, A.A.; Binkley, M.S.; Osmundson, E.C.; Alizadeh, A.A.; Diehn, M. Predicting radiotherapy responses and treatment outcomes through analysis of circulating tumor DNA. In *Seminars in Radiation Oncology*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 25, pp. 305–312.
129. Haber, D.A.; Velculescu, V.E. Blood-based analyses of cancer: Circulating tumor cells and circulating tumor DNA. *Cancer Discov.* **2014**, *4*, 650–661. [[CrossRef](#)]
130. Lee, R.C.; Feinbaum, R.L.; Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **1993**, *75*, 843–854. [[CrossRef](#)]
131. Hou, J.; Meng, F.; Chan, L.W.; Cho, W.; Wong, S. Circulating plasma MicroRNAs as diagnostic markers for NSCLC. *Front. Genet.* **2016**, *7*, 193. [[CrossRef](#)]
132. Jansson, M.D.; Lund, A.H. MicroRNA and cancer. *Mol. Oncol.* **2012**, *6*, 590–610. [[CrossRef](#)]
133. Trejo-Becerril, C.; Pérez-Cárdenas, E.; Taja-Chayeb, L.; Anker, P.; Herrera-Goepfert, R.; Medina-Velázquez, L.A.; Hidalgo-Miranda, A.; Pérez-Montiel, D.; Chávez-Blanco, A.; Cruz-Velázquez, J.; et al. Cancer progression mediated by horizontal gene transfer in an in vivo model. *PLoS ONE* **2012**, *7*, e52754. [[CrossRef](#)] [[PubMed](#)]
134. Johnstone, R.M.; Adam, M.; Hammond, J.; Orr, L.; Turbide, C. Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes). *J. Biol. Chem.* **1987**, *262*, 9412–9420. [[CrossRef](#)]
135. Sheridan, C. Exosome cancer diagnostic reaches market. *Nat. Biotechnol.* **2016**, *34*, 359–360. [[CrossRef](#)]
136. Rodríguez, M.; Silva, J.; López-Alfonso, A.; López-Muñiz, M.B.; Peña, C.; Domínguez, G.; García, J.M.; López-González, A.; Méndez, M.; Provencio, M.; et al. Different exosome cargo from plasma/bronchoalveolar lavage in non-small-cell lung cancer. *Genes Chromosom. Cancer* **2014**, *53*, 713–724. [[CrossRef](#)]
137. Taverna, S.; Giallombardo, M.; Gil-Bazo, I.; Carreca, A.P.; Castiglia, M.; Chacártegui, J.; Araujo, A.; Alessandro, R.; Pauwels, P.; Peeters, M.; et al. Exosomes isolation and characterization in serum is feasible in non-small cell lung cancer patients: Critical analysis of evidence and potential role in clinical practice. *Oncotarget* **2016**, *7*, 28748. [[CrossRef](#)] [[PubMed](#)]
138. Kahlert, C.; Kalluri, R. Exosomes in tumor microenvironment influence cancer progression and metastasis. *J. Mol. Med.* **2013**, *91*, 431–437. [[CrossRef](#)] [[PubMed](#)]
139. Paulus, J.M. *Platelet Size in Man*; Elsevier: Amsterdam, The Netherlands, 1975.
140. Nilsson, R.J.A.; Balaj, L.; Hulleman, E.; Van Rijn, S.; Pegtel, D.M.; Walraven, M.; Widmark, A.; Gerritsen, W.R.; Verheul, H.M.; Vandertop, W.P.; et al. Blood platelets contain tumor-derived RNA biomarkers. *Blood J. Am. Soc. Hematol.* **2011**, *118*, 3680–3683. [[CrossRef](#)]
141. Best, M.G.; Sol, N.; Kooi, I.; Tannous, J.; Westerman, B.A.; Rustenburg, F.; Schellen, P.; Verschueren, H.; Post, E.; Koster, J.; et al. RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* **2015**, *28*, 666–676. [[CrossRef](#)]
142. Li, L.; Zheng, H.; Huang, Y.; Huang, C.; Zhang, S.; Tian, J.; Li, P.; Sood, A.K.; Zhang, W.; Chen, K. DNA methylation signatures and coagulation factors in the peripheral blood leucocytes of epithelial ovarian cancer. *Carcinogenesis* **2017**, *38*, 797–805. [[CrossRef](#)]
143. Lin, L.H.; Chang, K.W.; Kao, S.Y.; Cheng, H.W.; Liu, C.J. Increased plasma circulating cell-free DNA could be a potential marker for oral cancer. *Int. J. Mol. Sci.* **2018**, *19*, 3303. [[CrossRef](#)] [[PubMed](#)]
144. Li, C.; Zhou, Y.; Liu, J.; Su, X.; Qin, H.; Huang, S.; Huang, X.; Zhou, N. Potential markers from serum-purified exosomes for detecting oral squamous cell carcinoma metastasis. *Cancer Epidemiol. Prev. Biomark.* **2019**, *28*, 1668–1681. [[CrossRef](#)]
145. Cucchiara, F.; Del Re, M.; Valleggi, S.; Romei, C.; Petrini, I.; Lucchesi, M.; Crucitta, S.; Rofi, E.; De Liperi, A.; Chella, A.; et al. Integrating liquid biopsy and radiomics to monitor clonal heterogeneity of EGFR-positive Non-small Cell Lung Cancer. *Front. Oncol.* **2020**, *10*, 593831. [[CrossRef](#)]
146. Wei, R.; Chen, L.; Qin, D.; Guo, Q.; Zhu, S.; Li, P.; Min, L.; Zhang, S. Liquid biopsy of extracellular vesicle-derived miR-193a-5p in colorectal cancer and discovery of its tumor-suppressor functions. *Front. Oncol.* **2020**, *10*, 1372. [[CrossRef](#)] [[PubMed](#)]
147. Raman, L.; Van der Linden, M.; Van der Eecken, K.; Vermaelen, K.; Demedts, I.; Surmont, V.; Himpe, U.; Dedeurwaerdere, F.; Ferdinande, L.; Lievens, Y.; et al. Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small cell lung carcinoma. *Genome Med.* **2020**, *12*, 1–12. [[CrossRef](#)] [[PubMed](#)]
148. El-Khoury, V.; Schritz, A.; Kim, S.Y.; Lesur, A.; Sertamo, K.; Bernardin, F.; Petritis, K.; Pirrotte, P.; Selinsky, C.; Whiteaker, J.R.; et al. Identification of a Blood-Based Protein Biomarker Panel for Lung Cancer Detection. *Cancers* **2020**, *12*, 1629. [[CrossRef](#)]
149. Yang, X.; Cai, G.X.; Han, B.W.; Guo, Z.W.; Wu, Y.S.; Lyu, X.; Huang, L.M.; Zhang, Y.B.; Li, X.; Ye, G.L.; et al. Association between the nucleosome footprint of plasma DNA and neoadjuvant chemotherapy response for breast cancer. *NPJ Breast Cancer* **2021**, *7*, 1–12. [[CrossRef](#)]
150. Maltoni, R.; Casadio, V.; Ravaioli, S.; Foca, F.; Tumedei, M.M.; Salvi, S.; Martignano, F.; Calistri, D.; Rocca, A.; Schirone, A.; et al. Cell-free DNA detected by “liquid biopsy” as a potential prognostic biomarker in early breast cancer. *Oncotarget* **2017**, *8*, 16642. [[CrossRef](#)] [[PubMed](#)]
151. Bray, F.; Ferlay, J.; 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](#)]
152. Rivera, C. Essentials of oral cancer. *Int. J. Clin. Exp. Pathol.* **2015**, *8*, 11884. [[PubMed](#)]
153. Jayson, G.C.; Kohn, E.C.; Kitchener, H.C.; Ledermann, J.A. Ovarian cancer. *Lancet* **2014**, *384*, 1376–1388. [[CrossRef](#)]

154. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2016. *CA Cancer J. Clin.* **2016**, *66*, 7–30. [[CrossRef](#)] [[PubMed](#)]
155. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
156. Roth, P.; Wischhusen, J.; Happold, C.; Chandran, P.A.; Hofer, S.; Eisele, G.; Weller, M.; Keller, A. A specific miRNA signature in the peripheral blood of glioblastoma patients. *J. Neurochem.* **2011**, *118*, 449–457. [[CrossRef](#)] [[PubMed](#)]
157. Best, M.G.; Sol, N.; GJG, S.; Vancura, A.; Muller, M.; Niemeijer, A.L.N.; Fejes, A.V.; Fat, L.A.T.K.; Huis, A.E.; Leurs, C.; et al. Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. *Cancer Cell* **2017**, *32*, 238–252. [[CrossRef](#)] [[PubMed](#)]
158. Zlotogorski-Hurvitz, A.; Dekel, B.Z.; Malonek, D.; Yahalom, R.; Vered, M. FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer. *J. Cancer Res. Clin. Oncol.* **2019**, *145*, 685–694. [[CrossRef](#)]
159. Best, M.G.; GJG, S.; Sol, N.; Wurdinger, T. RNA sequencing and swarm intelligence-enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA. *Nat. Protoc.* **2019**, *14*, 1206–1234. [[CrossRef](#)] [[PubMed](#)]
160. Sunkara, V.; Kim, C.J.; Park, J.; Woo, H.K.; Kim, D.; Ha, H.K.; Kim, M.H.; Son, Y.; Kim, J.R.; Cho, Y.K. Fully automated, label-free isolation of extracellular vesicles from whole blood for cancer diagnosis and monitoring. *Theranostics* **2019**, *9*, 1851. [[CrossRef](#)]
161. Ge, G.; Peng, D.; Guan, B.; Zhou, Y.; Gong, Y.; Shi, Y.; Hao, X.; Xu, Z.; Qi, J.; Lu, H.; et al. Urothelial carcinoma detection based on copy number profiles of urinary cell-free DNA by shallow whole-genome sequencing. *Clin. Chem.* **2020**, *66*, 188–198. [[CrossRef](#)]
162. Liu, S.; Wu, J.; Xia, Q.; Liu, H.; Li, W.; Xia, X.; Wang, J. Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1891–1903. [[CrossRef](#)]
163. Heinhuis, K.M.; In't Veld, S.G.; Dwarshuis, G.; Van Den Broek, D.; Sol, N.; Best, M.G.; Coevorden, F.v.; Haas, R.L.; Beijnen, J.H.; van Houdt, W.J.; et al. RNA-sequencing of tumor-educated platelets, a novel biomarker for blood-based sarcoma diagnostics. *Cancers* **2020**, *12*, 1372. [[CrossRef](#)]
164. Zhang, C.; Zhao, Y.; Xu, X.; Xu, R.; Li, H.; Teng, X.; Du, Y.; Miao, Y.; Lin, H.C.; Han, D. Cancer diagnosis with DNA molecular computation. *Nat. Nanotechnol.* **2020**, *15*, 709–715. [[CrossRef](#)]
165. Liu, L.; Chen, X.; Wong, K.C. Early cancer detection from genome-wide cell-free DNA fragmentation via shuffled frog leaping algorithm and support vector machine. *Bioinformatics* **2021**. [[CrossRef](#)]
166. Cario, C.L.; Witte, J.S. Orchid: A novel management, annotation and machine learning framework for analyzing cancer mutations. *Bioinformatics* **2018**, *34*, 936–942. [[CrossRef](#)] [[PubMed](#)]
167. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
168. Diaz-Uriarte, R.; De Andres, S.A. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **2006**, *7*, 1–13. [[CrossRef](#)] [[PubMed](#)]
169. Song, C.X.; Yin, S.; Ma, L.; Wheeler, A.; Chen, Y.; Zhang, Y.; Liu, B.; Xiong, J.; Zhang, W.; Hu, J.; et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* **2017**, *27*, 1231–1242. [[CrossRef](#)]
170. Cohen, J.D.; Li, L.; Wang, Y.; Thoburn, C.; Afsari, B.; Danilova, L.; Douville, C.; Javed, A.A.; Wong, F.; Mattox, A.; et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **2018**, *359*, 926–930. [[CrossRef](#)] [[PubMed](#)]
171. Penson, A.; Camacho, N.; Zheng, Y.; Varghese, A.M.; Al-Ahmadie, H.; Razavi, P.; Chandarlapaty, S.; Vallejo, C.E.; Vakiani, E.; Gilewski, T.; et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol.* **2020**, *6*, 84–91. [[CrossRef](#)]
172. Nassiri, F.; Chakravarthy, A.; Feng, S.; Shen, S.Y.; Nejad, R.; Zuccato, J.A.; Voisin, M.R.; Patil, V.; Horbinski, C.; Aldape, K.; et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat. Med.* **2020**, *26*, 1044–1047. [[CrossRef](#)]
173. Zhang, Y.H.; Jin, M.; Li, J.; Kong, X. Identifying circulating miRNA biomarkers for early diagnosis and monitoring of lung cancer. *Biochim. Biophys. Acta BBA Mol. Basis Dis.* **2020**, *1866*, 165847. [[CrossRef](#)]
174. Hoshino, A.; Kim, H.S.; Bojmar, L.; Gyan, K.E.; Cioffi, M.; Hernandez, J.; Zambirinis, C.P.; Rodrigues, G.; Molina, H.; Heissel, S.; et al. Extracellular vesicle and particle biomarkers define multiple human cancers. *Cell* **2020**, *182*, 1044–1061. [[CrossRef](#)] [[PubMed](#)]
175. Sprang, M.; Paret, C.; Faber, J. CpG-Islands as Markers for Liquid Biopsies of Cancer Patients. *Cells* **2020**, *9*, 1820. [[CrossRef](#)] [[PubMed](#)]
176. Wang, Y.; Zheng, J.; Li, Z.; Jiang, R.; Peng, J.; Sun, J.; Yang, G.; Yang, X.R.; Huang, A.; Wang, Y.; et al. Development of a novel liquid biopsy test to diagnose and locate gastrointestinal cancers. *J. Clin. Oncol.* **2020**, *38*, 1557. [[CrossRef](#)]
177. Peng, H.; Pan, S.; Yan, Y.; Brand, R.E.; Petersen, G.M.; Chari, S.T.; Lai, L.A.; Eng, J.K.; Brentnall, T.A.; Chen, R. Systemic proteome alterations linked to early stage pancreatic cancer in diabetic patients. *Cancers* **2020**, *12*, 1534. [[CrossRef](#)] [[PubMed](#)]
178. Zhong, G.; Wang, L.N.; Ling, X.; Dong, J. An overview on data representation learning: From traditional feature learning to recent deep learning. *J. Financ. Data Sci.* **2016**, *2*, 265–278. [[CrossRef](#)]
179. Ravi, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; Yang, G.Z. Deep learning for health informatics. *IEEE J. Biomed. Health Informat.* **2016**, *21*, 4–21. [[CrossRef](#)]
180. Ramchoun, H.; Idrissi, M.A.J.; Ghanou, Y.; Ettaouil, M. Multilayer Perceptron: Architecture Optimization and Training. *IJIMAI* **2016**, *4*, 26–30. [[CrossRef](#)]
181. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)]

182. Agarap, A.F. Deep learning using rectified linear units (relu). *arXiv* **2018**, arXiv:1803.08375.
183. Mookiah, M.R.K.; Acharya, U.R.; Ng, E. Data mining technique for breast cancer detection in thermograms using hybrid feature extraction strategy. *Quant. Infrared Thermogr. J.* **2012**, *9*, 151–165. [[CrossRef](#)]
184. Daoud, M.; Mayo, M. A survey of neural network-based cancer prediction models from microarray data. *Artif. Intell. Med.* **2019**, *97*, 204–214. [[CrossRef](#)]
185. Goryński, K.; Safian, I.; Grądzki, W.; Marszał, M.P.; Krysiński, J.; Goryński, S.; Bitner, A.; Romaszko, J.; Buciński, A. Artificial neural networks approach to early lung cancer detection. *Cent. Eur. J. Med.* **2014**, *9*, 632–641. [[CrossRef](#)]
186. Mao, Y.; Yin, Z.; Schober, J. A deep convolutional neural network trained on representative samples for circulating tumor cell detection. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–6.
187. Kothén-Hill, S.T.; Zviran, A.; Schulman, R.C.; Deochand, S.; Gaiti, F.; Maloney, D.; Huang, K.Y.; Liao, W.; Robine, N.; Omans, N.D.; et al. Deep Learning Mutation Prediction Enables Early Stage Lung Cancer Detection in Liquid Biopsy. 2018. Available online: <https://openreview.net/forum?id=H1DkN7ZCZ> (accessed on 10 June 2020).
188. Wong, K.C.; Chen, J.; Zhang, J.; Lin, J.; Yan, S.; Zhang, S.; Li, X.; Liang, C.; Peng, C.; Lin, Q.; et al. Early Cancer Detection from Multianalyte Blood Test Results. *IScience* **2019**, *15*, 332–341. [[CrossRef](#)] [[PubMed](#)]
189. Li, X.; Chen, S.; Hu, X.; Yang, J. Understanding the disharmony between dropout and batch normalization by variance shift. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2682–2690.
190. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **2019**, *7*, 154096–154113. [[CrossRef](#)]
191. Lauritsen, S.M.; Kristensen, M.; Olsen, M.V.; Larsen, M.S.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **2020**, *11*, 1–11. [[CrossRef](#)] [[PubMed](#)]
192. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv* **2017**, arXiv:1711.06104.
193. Lundberg, S.; Lee, S.I. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
194. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv* **2016**, arXiv:1605.01713.
195. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.