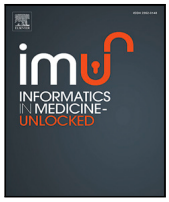
Informatics in Medicine Unlocked 28 (2022) 100824

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Most dominant metabolomic biomarkers identification for lung cancer

Utshab Kumar Ghosh *, Fuad Al Abir, Nahian Rifaat, S.M. Shovan, Abu Sayeed, Md. Al Mehedi Hasan

Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh

ARTICLE INFO

Keywords:
Metabolomic biomarkers
Normal distribution test
Variance test
Student's T-Test
Kruskal–Wallis Test
Recursive Feature Elimination

ABSTRACT

Metabolomic biomarkers play a vital role in the early identification and prediction of cancer. It is possible to save numerous lives if biomarkers are used to assist medical providers in diagnosing their patients faster. Many researchers have been trying to identify the crucial biomarkers in the early diagnosis of diseases. This paper presents several steps divided into two phases for determining the most important metabolomic biomarkers in the blood for lung cancer prediction using Plasma and Serum samples. We used the Shapiro–Wilk Test, Bartlett's Test, Levene's Test, Student's t-Test, and Kruskal–Wallis Test in the first phase to determine the potential biomarkers. Recursive Feature Elimination with Random Forest was used to identify the final most dominant metabolomic biomarker at the second phase. Lastly, we ended with Ridge Classifier and XGBoost Classifier to assess the consistency of our approaches. Despite the declining number of metabolites up to a greater level, our prediction accuracy was 100% and 90.91% for Plasma and Serum samples, respectively which is higher than the state-of-the-art method. Finally, we made some analysis using the most dominant metabolites that can serve as a source of inspiration for our work.

1. Introduction

Lung cancer has been a threat in the field of medicine for a long time. According to recent studies, an estimation of 235,760 new cases will be diagnosed, and 131,880 people will die from lung cancer in 2021 in the US [1]. The same survey results reported 228,820 diagnoses and 135,720 deaths in 2020 [2]. So without any doubt, the number of patients is increasing at an alarming rate every year. Patients with invasive lung and bronchus cancer were identified from the SEER 18-registry database, covering 28% of the US population [3]. Lung cancer patients can be recovered if they are diagnosed early. Life-loss years vary from 6.16 for Stage I cancer to 16.21 for Stage IV [4].

Metabolomics aims to comprehensively analyze wide arrays of metabolites in biological samples [5]. Metabolite measurements bear fundamental regulatory importance to be used as diagnostic markers for biological conditions, including diseases and response to chemical treatment [6]. It is a beneficial field of study in the field of disease detection. In this regard, biomarkers can play a crucial role in disease detection and identification. One of the most common diseases faced with the earth today is cancer. Suppose the metabolites are recognized, either not present or present up to a tolerable amount in healthy cases. In that case, it will have a huge impact on the identification of cancer.

In this paper, our objective is to show the impact of the measure of some specific metabolomic biomarkers in lung cancer diagnosis. We worked with some lung cancer patients using those metabolomic biomarkers present in Plasma and Serum samples of blood of those patients. We analyzed 158 metabolites to find out the most significant metabolomic biomarkers. We classified a person as a normal or a lung cancer patient based on the specific metabolites. We also found the hierarchical differences and relations between the metabolites using the Agglomerative Hierarchical Clustering Technique [7]. Eventually, we evaluated our approaches in terms of accuracy to identify lung cancer patients.

We divided our proposed methodology mainly into two parts—Feature Selection and Classification. At first, in feature selection, for looking into the distribution of each feature in our dataset, we used Shapiro–Wilk Test to check if the features were normally distributed [8,9]. Then, we checked if our dataset maintained homogeneity (or equality) of variances. For that, we used Bartlett's Test for the features with normal distribution and Levene's Test for the features without normal distribution [9–11]. Finally, Student's t-Test [12] for features with Homoscedasticity (or equal variances) and Kruskal–Wallis Test [13] for the features with Heteroscedasticity (or unequal variances) were used and using the test statistics and p-Value, to obtain the most dominant metabolites from a large number of list of those. We obtained the most dominant metabolites in the case of Plasma and Serum

* Corresponding author.
E-mail addresses: kumarutshab@gmail.com (U.K. Ghosh), alabir.fuad@gmail.com (F. Al Abir), nahian.rifaat@gmail.com (N. Rifaat), sm.shovan@gmail.com (S.M. Shovan), abusayeed.cse@gmail.com (A. Sayeed), mehedi_ru@yahoo.com (M.A.M. Hasan).

https://doi.org/10.1016/j.imu.2021.100824
Received 11 September 2021; Received in revised form 6 December 2021; Accepted 14 December 2021
Available online 30 December 2021

samples. We deleted the rest. Then we used Recursive Feature Elimination [14] and obtained the best accuracy with some classifiers with a very few metabolites of Plasma and Serum samples. For classification, we went through some classifiers like Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), Ridge Classifier, XGBoost Classifier, etc. We noted the results of the best classifiers (Ridge Classifier and XGBoost Classifier for Plasma and Serum sample, respectively).

At a glance, the overview of our paper is as follows: in the following section, a literature review about the existing strategies on this problem domain is presented. Then in the Methodology section, we described the datasets, and our proposed architecture is explained elaborately. The Results section provides the details of our experimental results and comparison with the previous works. We conclude the paper with a Conclusion. We also mention some practical applications of our work in the Practical Implications section and the limitations of our work in the Limitations and Future Research section.

2. Literature review

Previous works have taken up the task of identifying metabolomic biomarkers from blood samples that can be used to identify cancer patients. Kumar et al. [12] used Student's t-Test and Kruskal–Wallis Test to identify significant metabolites that attribute to lung cancer and its identification in patients. They used the plasma, and serum samples from the dataset [15] and ranked those metabolites according to the importance score calculated using a Support Vector Machine (SVM) classifier with a radial basis kernel function. Masrur et al. [16,17] used Student's t-Test, Kruskal–Wallis Test, and Mann–Whitney–Wilcoxon Test to distinguish the differentially expressed metabolites obtained from plasma and serum blood samples. The authors of this work used cluster heatmap plots and fold change values to differentiate up and down-regulated metabolites and used Recursive Feature Elimination (RFE) to order and select the metabolites. Finally, a Support Vector Machine (SVM) classifier was used with these metabolites to classify control, and disease subjects in the dataset [15].

Zheng et al. [18] performed a study on serum samples of lung cancer patients and healthy people implementing RF algorithm, SVM, and PLS-DA algorithms. They identified 15 differential metabolites matching the NIST database, of which five were found most significant ones in differentiating the lung cancer patients. Xie et al. [19] worked with plasma samples of 110 lung cancer patients and 43 healthy individuals to detect lung cancer, especially of early stage. They used Random Forest, SVM, etc., and found 46 most influential metabolic biomarkers from 61, which are present in those plasma samples. They obtained a classification accuracy of 100% with their machine learning implications on that dataset. Ruiying et al. [20] identified 35 metabolites of serum samples which were different between Non-small cell lung cancer (NSCLC) patients and healthy individuals, where 6 metabolites were chosen as the most dominant. Zhang et al. [21] used the mRMR method to rank input features, and the authors combined incremental feature selection with Random Forest to select optimum features for classification. Yuan et al. [22] used the Monte-Carlo Feature Selection method, and then they used the Iterative Feature Selection method with SVM to classify patients. The methods used in these works can be improved by trying out different types of tests that we had done in our work to be discussed in the future sections. Zhang et al. [23] worked on the diagnosis of different stages of lung cancer analyzing plasma metabolites by applying multivariate analysis and logistic regression model.

Moreover, metabolomics study has been widely used in the diagnosis of other diseases too. Shu et al. [24] approached with machine learning-based models on plasma samples and identified powerful biomarker combinations that can predict COVID-19 cases. Biomarkers from plasma samples were analyzed, and the risk of heart failure was assessed by Chirinos et al. [25]. They used a tree-based pipeline optimizer (TPOT) platform to classify the patients and normal individuals. Uchiyama et al. [26] analyzed the serum samples using the Advanced Scan Package of Japan and identified 139 metabolites, of which 16 had a better correlation with colorectal cancer (CRC). Nishiumi et al. [27] analyzed plasma samples using GC-QqQ-MS and observed 8 metabolites to significantly correlate to CRC. Long et al. [28] also made a research on blood samples for global metabolomics profile analysis on CRC and colorectal adenoma polyps using LC-MS/MS.

In the literature, the scope of improvement in the prediction performance while selecting a lesser number of metabolomic biomarkers using statistical and machine learning algorithms is the main motivation behind this research. We aim at early diagnosis of lung cancer more accurately with less number of metabolites than that of Kumar et al. [12], and Masrur et al. [16,17].

Table 1
Subject distribution by control and disease in the datasets.

Sample type	Control	Disease
Plasma	41	41
Serum	41	41

Table 2
Subject distribution by gender in the datasets.

Sample type	Male	Female
Plasma	20	62
Serum	20	62

3. Methodology

3.1. Dataset description

The dataset we used in our study was produced under the study ID: ST000392 by Oliver Fiehn [15]. It was produced by the time-of-flight mass spectrometry GC-TOF-MS technique [29]. All the samples collected were of two types, Plasma and Serum. Both samples contained 82 subjects, for which data of 158 metabolites were given in the dataset. Among the subjects, the number of cancer patients and control patients was 41 each, and 20 of the subjects were male, and the remaining 60 subjects were female. Each of these are summarized in Tables 1 and 2.

All of these blood samples were provided by two institutions, the University of California at Davis Medical Center and the Fred Hutchison Cancer Research Center. They provided the samples from their bio-repositories. The samples were collected using (Ethylenediaminetetraacetic acid) EDTA tubes and were stored at −80 degree Celsius [15].

The raw data of the GC-TOF-MS were processed using the ChromaTOF software (v. 2.32) and to find the peak and mass spectral deconvolution. All the samples were collected with the consent of the individuals in strict adherence to the IRB protocols approved by the Institutional Review Board of each institution with the aim of the usage of the samples restricted to research purposes only. The files containing the result were exported and filtered using the UC Davis Metabolomics BinBase database for consistency.

3.2. Proposed architecture

There is a flow diagram given as Fig. 1. This diagram is a summary of our different approaches with Plasma and Serum samples to make it easier to think and go through with them, which we will go deeper with the details of the approaches in the next sessions.

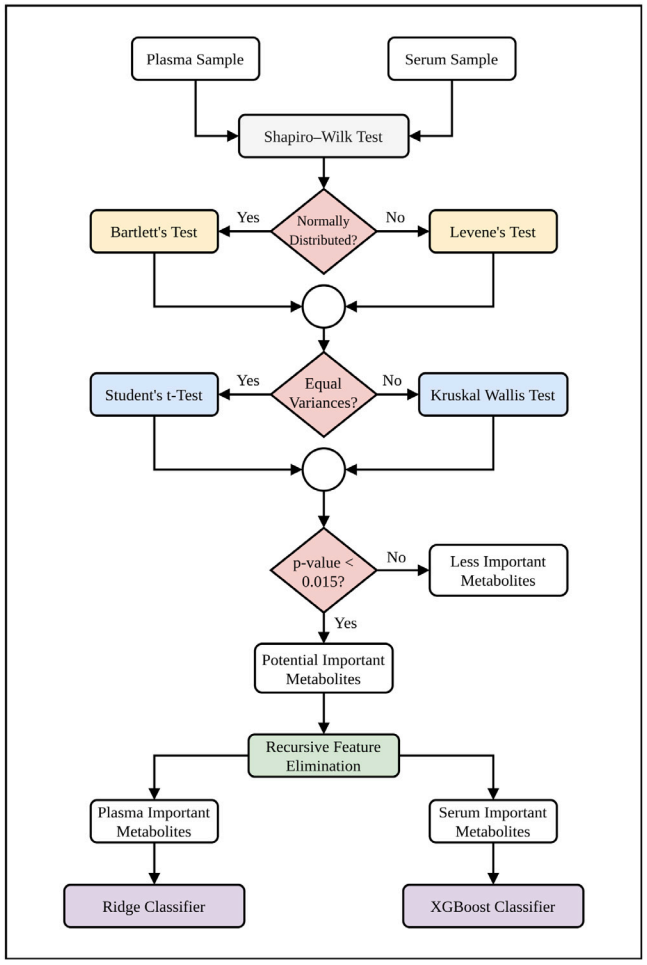


Fig. 1. Flow diagram of our proposed methodology.

3.3. Feature selection

We performed the Feature Selection Process in two different phases. At first, several tests were compiled to obtain the potential metabolomic biomarkers. Since there are things like normal distribution, equal or unequal variances in the dataset features, so we had to perform those tests. In the second phase, we performed Recursive Feature Elimination to filter out and reduce the number of most dominant metabolites.

3.3.1. Shapiro–Wilk Test

From the Monte Carlo simulation, it has been observed that Shapiro–Wilk Test is the best for the normal distribution test for a given significance. The next candidates in this testing field are Anderson–Darling, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling Tests [30].

First, let us assume a null hypothesis, $\boldsymbol{H_0}$ that all of the n samples of every feature are normally distributed. We can reject this hypothesis if we get a p-value less than or equal to 0.05. p-Value = 0.05 was taken as a standard value for most of the test phases. But for the phase of Student's t-Test and Kruskal–Wallis Test, p-Value was set at 0.015 on the trial and error basis in order to get an optimized number of most dominant metabolomic biomarkers.

Now it is time to test the statistical significance. If $\boldsymbol{X_1 < X_2 < \cdots < X_n}$ is an ordered sample of size n to be tested for non-normality, $\bar{X}$ is the sample mean, $X_{(i)}$ is the ith order statistic and a_i are constants generated from the covariances, variances and means of the ith [i=1, 2, …, n] sample: Shapiro–Wilk Test Statistic,

$$W_{shapiro} = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{1}$$

Thus, the test statistic and the corresponding p-value were calculated with the features of our dataset. Now, two cases happened based on that p-value:

if p-Value > 0.05, then the feature is normally distributed.
if p-Value <= 0.05, then the feature is not normally distributed.

We, later on, used Bartlett's test and Levene's test to check if their variance was equal. Bartlett's test is better for the features with normal distribution, and Levene's test is better for the rest [31]. That is why we had to check the distribution of the features with the Shapiro–Wilk Test first for obtaining the best output.

3.3.2. Bartlett's Test

We used Bartlett's Test on the features with normal distributions only. Let us assume a null hypothesis, $\boldsymbol{H_0}$ that all of the n samples have equal variance. More mathematically, if the variance is σ_i for samples $\boldsymbol{i = 1, 2, \ldots, n}$:

$$\boldsymbol{H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2} \tag{2}$$

We can reject this hypothesis and stand with alternate hypothesis, $\boldsymbol{H_a}$ if we get any pair of variances not equal. More mathematically, if σ_i and σ_j are a pair of variance of ith and jth $\boldsymbol{[i, j = 1, 2, \ldots, n]}$ samples:

$$\boldsymbol{H_a : \sigma_i^2 \neq \sigma_j^2} \tag{3}$$

Now it is time for test statistics. If there are $\boldsymbol{N}$ samples, where $\boldsymbol{n_i}$ is the number of samples for the ith feature and $\boldsymbol{S_i^2}$ is the variance of a sample $\boldsymbol{n_i}$ features from the ith population $\boldsymbol{[i = 1, 2, \ldots, k]}$, where $\boldsymbol{k}$ is the total number of features: Barlette Test Statistic,

$$\boldsymbol{W_{bartlett} = 2.3026\frac{q}{c}} \tag{4}$$

Here,

$$\boldsymbol{q = (N - k)log(S_p^2) - \sum_{i=1}^{k}(n_i - 1)log(S_i^2)}$$
$$\boldsymbol{c = 1 + \frac{1}{3(k-1)}[\sum_{i=1}^{k}(\frac{1}{N_i - 1}) - \frac{1}{N-k}]}$$
Pooled Variance, $\boldsymbol{S_p^2 = \frac{(k_i - 1)S_i^2}{N-k}}$

After the calculation of the test statistic as well as p-value from each of the features, two cases happened based on that p-value:

p-Value > 0.05 denotes Homoscedasticity of a feature.
p-Value <= 0.05 denotes Heteroscedasticity of a feature.

3.3.3. Levene's Test

The null hypothesis, $\boldsymbol{H_0}$, and the alternate hypothesis, $\boldsymbol{H_a}$ are the same as mentioned in Bartlett's Test section. So, we jump directly to test statistics. We used Levene's test on the features without normal distributions only. Now, if a variable $\boldsymbol{X}$ with a sample of size $\boldsymbol{N}$ is divided into $\boldsymbol{k}$ subgroups, where $\boldsymbol{N_i}$ is the sample size of the ith subgroup:

Levene Test Statistic,

$$\boldsymbol{W_{levene} = \frac{(N - k)\sum_{i=1}^{k}(\bar{Z}_{i\cdot} - \bar{Z}_{\cdot\cdot})^2}{(k - 1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i\cdot})^2}} \tag{5}$$

where, $\bar{Z}_{i\cdot}$ are the group means of the Z_{ij} and $\bar{Z}_{\cdot\cdot}$ is the overall mean of the Z_{ij}. Z_{ij} has definitions of:

$$\boldsymbol{Z_{ij} = |X_{ij} - X_i|} \tag{6}$$

Here, X_{ij} is the value of measured variable for jth case from the ith group. X_i can be any one of 3 different definitions, which are $\bar{X}_i$, $\widetilde{X_i}$ and $\bar{X}_i^{10\%}$. They are called mean, median, and 10% trimmed mean respectively of the ith subgroup. Just like Bartlett's Test, two cases happened on the basis of the p-value:

p-Value > 0.05 denotes Homoscedasticity of a feature.
p-Value <= 0.05 denotes Heteroscedasticity of a feature.

3.3.4. Student's t-Test

In our work, we used Student's t-Test [32] on each metabolite that had a p-value > 0.05 and demonstrated equal variance after applying Bartlette's Test and Levene's Test on the metabolites. The Student's t-Test is a parametric test method. In the Student's t-Test, if a random sample $X_{11}, X_{12}, X_{13}, \ldots, X_{1n1}$ follows a normal distribution with mean μ_1 and variance σ_1^2 and another random sample $X_{21}, X_{22}, X_{23}, \ldots, X_{2n2}$ follows a normal distribution with mean μ_2 and variance σ_2^2, then the null hypothesis, H_0 and the alternate hypothesis, H_a are to be tested is, $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$. The test statistic (for $\sigma_1^2 = \sigma_2^2$) is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7}$$

Here,

$$\bar{X}_1 = \sum_i^{n_1} \frac{X_{1i}}{n_1};\ \bar{X}_2 = \sum_i^{n_2} \frac{X_{2i}}{n_2}$$
$$s_1^2 = \frac{1}{n_1 - 1}\sum(X_{1i} - \bar{X}_1)^2;\ s_2^2 = \frac{1}{n_2 - 1}\sum(X_{2i} - \bar{X}_2)^2;$$
$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

Similarly, for $\sigma_1^2 \neq \sigma_2^2$, the test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \tag{8}$$

Here, the arithmetic means of the sample 1 and sample 2 are $\bar{X}_1$ and $\bar{X}_2$ respectively and s_1^2, s_2^2 are their respective variances. The test statistics correspond to equal variance and unequal variance, respectively. Then, the p-value is calculated with respect to the derived t value with $n_1 + n_2 - 2$ degrees of freedom [12].

If X is a metabolomics data matrix that contains both types of samples (cancer and control), then for the ith metabolite X_{ij}, $[j = 1, 2, \ldots, n_1$ is a sample for type-1 (e.g. cancer) with sample size $n_1]$ and X_{ik}, $[k = 1, 2, \ldots, n_2$, is the sample for type-2 (e.g. control) with sample size $n_2]$; we assume H_0, "the ith metabolite is not differentially expressed between cancer vs control group". In this study, H_0 is rejected if p-value < 0.015.

3.3.5. Kruskal–Wallis Test

Kruskal and Wallis [13] proposed a non-parametric test that is used on the data that do not satisfy the property of equal variance. We used this non-parametric test on the metabolites that did not satisfy the condition of p-value > 0.05 upon using Levene's Test and Bartlett's Test on these metabolites. The test statistic for Kruskal–Wallis for k groups each of size n_i is defined by: Kruskal–Wallis Test Statistic,

$$W_{Kruskal} = \frac{1}{s^2}\left[\left[\sum_{i=1}^{k} \frac{R_i}{n_i} - N\frac{(N+1)^2}{4}\right]\right] \tag{9}$$

Here,

N: the total number of observations
R_i: sum of the ranks for the ith sample
$s^2 = \frac{1}{N-1}\left[\sum_{i,j} R_{ij}^2 - N\frac{(N+1)^2}{4}\right]$

3.3.6. Recursive feature elimination

After we had a set of filtered metabolites obtained from all the gradual tests, we decided to reduce the number of those metabolites up to some more extent. So we used Recursive Feature Elimination [14] to obtain a small feature subset, removing several features at a time. This technique follows the following steps:

Step 1. Train the classifier
Step 2. Compute the ranking criterion for all features
Step 3. Remove the feature with smallest ranking criterion

Here, we chose Random Forest Classifier (RF) [33] to train. For any given tree in an RF, there is a subset of the learning set not used by it during training because each tree was grown only on a bootstrap sample. These subsets, called out-of-bag (OOB), can be used to give unbiased measures of prediction error [34]. Each feature is shuffled, and in the shuffled dataset, an OOB estimation of the prediction error is calculated. Also, the irrelevant features do not change the prediction error when altered in this way, opposite to the very relevant ones.

Again, we used cross-validation in this technique (RFECV) with a step size of 5 to find out all probable best combinations of most dominant metabolites of the Plasma sample, numbered from 1 to n_P. We did the same thing with the Serum sample for the metabolites numbered from 1 to n_S. [n_P and n_S are the number of filtered biomarkers of Plasma and Serum samples, respectively, obtained through all the tests mentioned in Section 3.3].

3.4. Classification

We chose two different supervised machine learning algorithms to differentiate normal persons and patients. Ridge Classifier [35,36] and XGBoost Classifier [37] for the Plasma and Serum samples, respectively.

3.4.1. Ridge Classifier

We used Ridge Classifier in our Plasma samples. Ridge Classifier uses the Ridge regression model to create a classifier. Our dataset consists of binary classes. So the classifier was used in the following ways:

The target variable is converted into +1 or −1 based on the class to which it belongs. Then, the Ridge regression model is built to predict our target variable. The loss function: L_0 = Mean Squared Error + l2 penalty.

Then, the Ridge regression's prediction value was calculated based on decision_function. If the value is greater than 0, it is predicted as Disease class. Otherwise, it is predicted as the Control class. So, using the Ridge classifier with cross-validation [38] of 10 splits in our Plasma Sample, we got predictions from our sample. The accuracy measurements are mentioned in Table 9.

3.4.2. XGBoost Classifier

We used XGBoost Classifier for our best approach, which is a scalable tree boosting system invented by Tianqi Chen and Carlos Guestrin [37]. It was started off as a research project as a part of the Distributed Machine Learning Community (DMLC) group. This algorithm falls in the category of boosting techniques among the ensemble techniques and has been regarded as a great performer in various cases where notion stems for the construction of additive models. Let, $b_k(x)$ be a function that is addressed as a base learner. The additive model is thus the sum of base learners:

$$f(x) = \sum_{k=1}^{M} b_k(x) \tag{10}$$

for $k = 1, 2, \ldots, M$ where M is the number of base learners. The minimization of the risk $L = (f(x, y))$ for the base learners of the previous equation can be written as,

$$\begin{aligned} b(x) &= argmin_b \sum_D L(f_{k-1}(x) + b(x), y) \\ &= argmin_b \sum_D [b(x)g(x, y) + \frac{1}{2}b^2(x)h(x, y)] \end{aligned} \tag{11}$$

where $D = (x, y)$ is a dataset and

$$g(x, y) = \frac{\partial L(f_{k-1}(x), y)}{\partial};\ h(x, y) = \frac{\partial L(f_{k-1}(x), y)}{\partial f^2} \tag{12}$$

The additive model of Eq. (10) is thus updated iteratively with the boosting as

$$f_k(x) = f_{k-1}(x) + b(x) \tag{13}$$

In our work, we used a tree-boosting XGBoost algorithm. The tree model can be written as

$$f(x) = \sum_{j=1}^{T} w_j I[x \epsilon R_j] \tag{14}$$

where w_j is the constant fit in region R_j and I being the set of indices of input x assigned to the jth leaf for $j = 1, 2, \ldots, T$, where T is the number of leaves of a tree. To grow a tree, it need to learn the constant w_j and the regions R_j from data. For the optimized leaf weight w^*, Eq. (14) is submitted into the second equation of (11) to yield

$$\begin{aligned} w^* &= argmin_w \sum_D \sum_{j=1}^{T} [g(x,y) w_j + \frac{1}{2} h(x,y) w_j^2] \\ &= argmin_w \sum_{j=1}^{T} [G w_j + \frac{1}{2} H w_j^2] \end{aligned} \tag{15}$$

where,

$$G = \sum_D g(x,y)$$
$$H = \sum_D h(x,y)$$

For a fixed structure, the optimized leaf weights w^* can be determined as $w^* = \frac{G}{H}$

Finding the optimized leaf weights of w^* is equivalent to learning the leaf weights. We have to find the split, which maximizes again, which is the loss reduction of a split. The gain for a fixed structure is derived by the substitution into the previous equation by the equation before it.

$$w^* = -\frac{l}{2} \sum_{k=1}^{M} \frac{G^2}{H} \tag{16}$$

The binary splits l(left and right) are determined by maximizing the gain A according to Eq. (16) given by

$$A = \frac{1}{2} [\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G^2}{H}] \tag{17}$$

where the subscripts L and R denote the left and right branches of a tree, respectively.

Now, using this XGBoost classifier in the Serum sample, we got predictions from our dataset this time. Cross-validation with 10 splits was used to obtain a more accurate result. The accuracy measurements for all of them in the Serum sample are also mentioned in Table 9.

4. Results

4.1. Biomarker selection

4.1.1. Phase 1: Potential important biomarkers

We employed our methodology on the dataset prepared by Oliver Fiehn [15], which is described in Section 3.1. From Shapiro–Wilk Test, as mentioned in Section 3.3.1 on our Plasma sample, only 19 features (20 features for Serum sample) had given a result of p-value > 0.05, which declines the state of 100% normal distribution of our dataset. Now we should check the homogeneity of variances. Due to Heteroscedasticity, the type I error rate could be affected in our dataset prediction, resulting in false positives.

So, we proceeded with the Plasma sample for testing equality of variances of those 19 features with Bartlett's Test as mentioned in Section 3.3.2. Alternatively, we used Levene's Test from Section 3.3.3 for the rest of the features (139) as they were not normally distributed. In the case of the serum sample, it was 20 features for Bartlett and 138 for Levene (see Table 3).

There were 19 features with a normal distribution, as we had mentioned earlier. All of them showed a result with a p-value under 0.05. So, we obtained 19 features with the equality of variances and 0 of the opposite case.

Table 3
Characteristic count of the normally distributed features obtained from Shapiro–Wilk Test.

Sample type	# normally distributed	# not normally distributed
Plasma	19	139
Serum	20	138

Table 4
Characteristic count of the homoscedastic and heteroscedastic features obtained from Levene's Test and Bartlett's Test.

Sample type	# homoscedastic	# heteroscedastic
Plasma	138	20
Serum	141	17

Table 5
Potential plasma biomarkers and the p-Values obtained from phase 1.

Metabolites	p-Value
Asparagine	7.772173e−04
Benzoic acid	1.573528e−03
Tryptophan	8.134184e−03
Uric acid	4.475238e−03
Alpha-ketoglutarat	8.283066e−03
Citrulline	2.255381e−03
Glutamine	1.008927e−03
Hypoxanthine	1.008319e−02
Malic acid	1.492488e−03
Methionine sulfoxide	2.922638e−03
Nornicotine	1.416206e−02
Octadecanol	1.307513e−02
3-phosphoglycerate	4.305257e−06
5-methoxytryptamine	3.768184e−06
Adenosine-5-monophosphate	1.172319e−09
Aspartic acid	9.288817e−06
Lactic acid	2.299930e−05
Maltose	1.725126e−04
Maltotriose	9.523541e−03
N-methylalanine	1.212036e−02
Phenol	5.512039e−06
Phosphoethanolamine	1.511919e−03
Pyrophosphate	1.044920e−07
Pyruvic acid	2.825701e−04
Taurine	6.989644e−07

After Bartlett's Test and Levene's Test on our Plasma sample, we got 138 features with Homoscedasticity (features named parametric). So we used Student's t-Test from Section 3.3.4 for finding out the test statistics and p-value. The remaining 20 features were found with Heteroscedasticity (features named non-parametric), and so they needed Kruskal–Wallis t-Test as mentioned in Section 3.3.5. This ratio was slightly different for the serum sample as recorded in Table 4.

Going through all these steps, we obtained some important metabolites based on the p-value of the features, using Student's t-Test and Kruskal–Wallis Test. We marked a metabolite as a potential for which a p-Value < 0.015 was found. There were 26 such metabolites in the plasma sample. It was 16 in the case of the Serum sample. The important metabolites along with their test results are given in Tables 5 and 6.

4.1.2. Phase 2: Final important biomarkers

Lastly, we performed Recursive Feature Elimination from Section 3.3.6 and observed that only 19 metabolites from the Plasma sample were enough to provide us the best accuracy. It was only 7 at the time of the Serum sample. The most dominant metabolites are shown in Table 7.

Fig. 2 shows the line plot of all the accuracies obtained against the different number of metabolites using the Recursive Feature Elimination Cross-Validation (RFECV) with Random Forest Classifier. Here it is noticeable that the result is the best for Plasma when the number

Table 6
Potential serum biomarkers and the *p*-Values obtained from phase 1.

Metabolites	p-Value
Cholesterol	0.004834
Threonine	0.011263
Uric aci	0.011401
Inosine	0.014031
Lactic acid	0.001083
N-methylalanine	0.002099
Phenylalanine	0.008690
Aspartic acid	0.000002
Deoxypentitol	0.002537
Glutamic acid	0.005964
Malic acid	0.006954
Phenol	0.000283
Taurine	0.000533

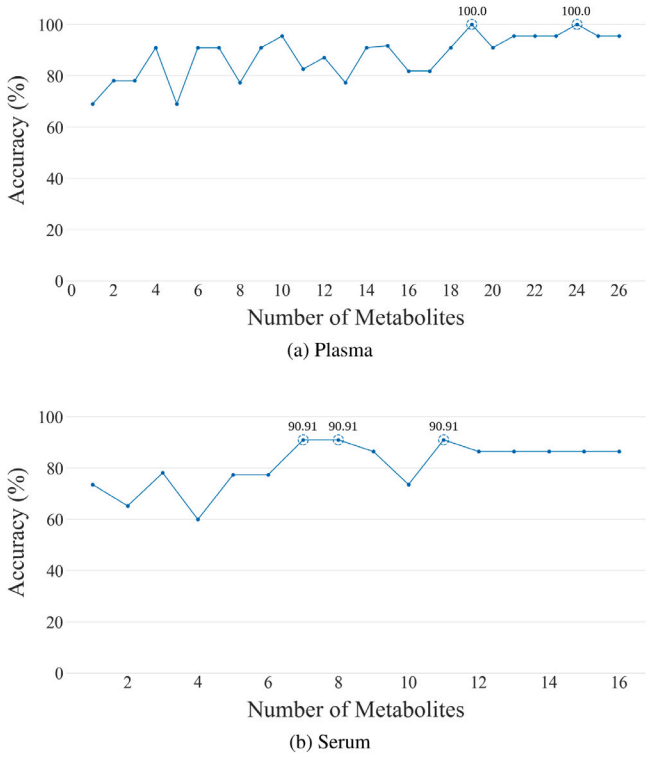


Fig. 2. Accuracy plot of Plasma and Serum samples for different number of biomarkers obtained from RFECV. (a) is of Plasma samples and (b) shows the same of Serum samples. From the plots, we can see the difference of the accuracy varying according to the number of biomarkers we chose. The best performances are annotated of 100% for the Plasma samples with 19 biomarkers and 90.91% with 7 biomarkers for the Serum samples.

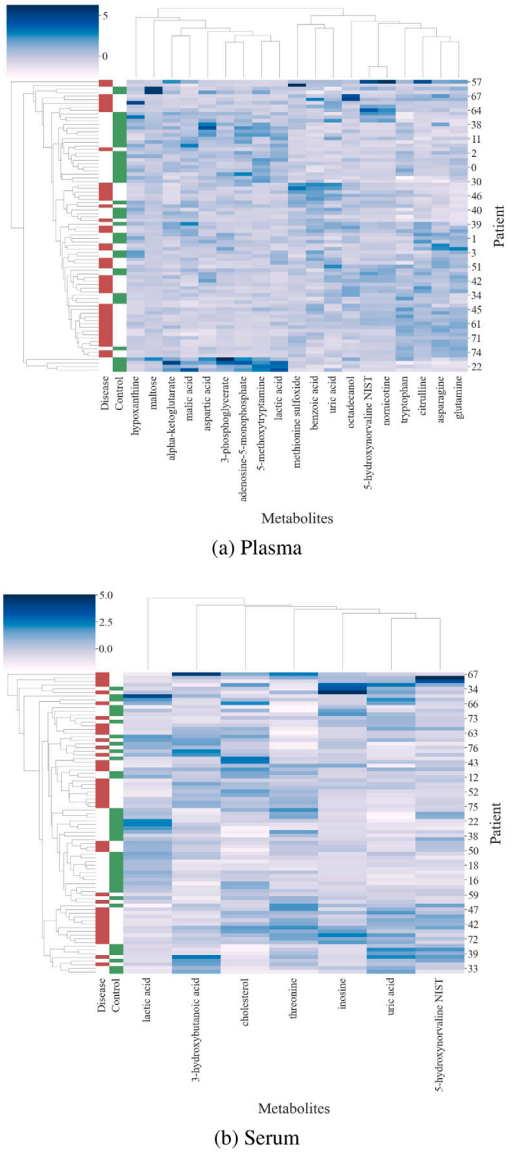


Fig. 3. Clustermaps of final and most dominant biomarkers from Plasma and Serum samples. (a) shows the most important metabolites of Plasma samples and (b) shows the same of Serum samples. The deeper the color, the more dominant it is for the diagnosis of lung cancer. The dendrites shows the cluster relationship among the metabolites. The red spots indicates Disease label and greens are for Normal ones.

of metabolites is 19 or 24. Similarly, the best result from the Serum sample is obtained for 7, 8, and 11 metabolites only.

The final and most dominant metabolomic biomarkers responsible for the prediction are mentioned earlier in Table 7. In our approach, we decreased the number of biomarkers from the work of Kumar et al. [12], and Masrur et al. [17]. A comparison of the approaches to find the least number of important biomarkers is given in Table 8.

4.2. Validity of final important features

4.2.1. Clustermap

Fig. 3 denotes the cluster heatmap of our approach to show the validity of the final important metabolites (19 from plasma and 7 from serum samples) to predict Lung Cancer. The red and green regions of the heatmap indicates the Disease and Control Class respectively. On the other hand, the tiles of the heatmaps show dominance in the case of prediction. The branch-like objects are called dendrites, which formed clusters to show closeness to one another. The two branches under the same hierarchy form a cluster, which denotes the least difference of all other metabolites.

4.2.2. Feature importance plot

Fig. 4 shows the cluster Feature Importance plot of our approach to simplify the importance rate of the same metabolites (19 and 7 from plasma and serum samples respectively), most useful to predict Lung Cancer. The importance rate is plotted under a range of 0.0 to 1.0. Random Forest classifier was used to rate out the metabolites.

4.2.3. Comparative analysis

Our different approaches for different samples give us the result of accuracy to predict the patients with lung cancer as given in Table 9.

Table 7
Selected most dominant biomarkers from Plasma and Serum datasets.

Sample type	Metabolites			
Plasma	Asparagine	Benzoic acid	Tryptophan	Uric acid
	5-hydroxynorvaline NIST	Alpha-ketoglutarate	Citrulline	Glutamine
	Hypoxanthine	Malic acid	Methionine sulfoxide	Nornicotine
	Octadecanol	3-phosphoglycerate	5-methoxytryptamine	Adenosine-5-monophosphate
	Aspartic acid	Lactic acid	Maltose	
Serum	Cholesterol	Threonine	Uric acid	3-hydroxybutanoic acid
	5-hydroxynorvaline NIST	Inosine	Lactic acid	

Table 8
Number of most dominant biomarker comparison with the existing methods.

Method	# Plasma	# Serum
Masrur et al. [17]	28	13
Kumar et al. [12]	27	13
Our approach	**19**	**7**

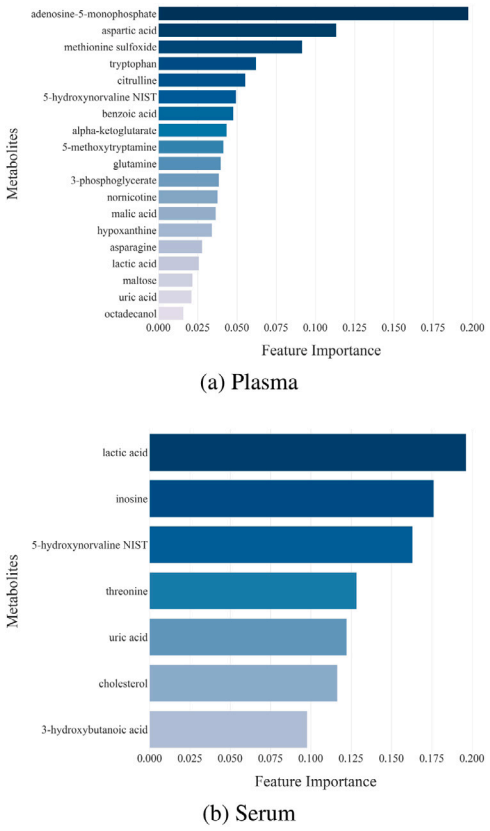


Fig. 4. Feature importance plots of metabolites. (a) shows the relative dominance of the final and most important biomarkers for the early diagnosis of lung cancer using Plasma samples and (b) shows the same using Serum samples.

The prediction accuracy of Masrur et al. [17] was 87.5% for the plasma sample and 83.33% for the serum sample. However, we obtained a better accuracy of 100% and 90.91% for the plasma and serum samples, respectively. We tried Classifiers like Decision Tree, Random Forest, SVM, Ridge, XGBoost, and so on for both the samples. At last, we kept the Ridge Classifier for Plasma and XGBoost Classifier for Serum sample for providing the best results than all other classifiers. The results are given in Table 9.

Table 9
Prediction performance comparison with the state-of-the-art method.

Method		Plasma	Serum
Masrur et al. [17]		87.50	83.33
Our approach	Ridge classifier	**100.0**	75.40
	XGBoost classifier	81.52	**90.91**
	Decision tree classifier	78.64	83.69
	Random forest classifier	83.29	83.27
	Support vector machine	85.43	77.76
	GaussianNB	80.95	78.68

5. Conclusions

Cancer continues to be one of the most common deadly diseases in the world. Every year, large numbers of people lose their lives to different subtypes of cancer. Previous research found that metabolomic biomarkers identified from differentially expressed metabolites in lung cancer patients can have a huge impact on the field of medicine as early and cost-effective measures of identifying cancer patients are crucial to save lives. In this work, we improved the existing method for identifying metabolomic biomarkers with high accuracy. Using the Ridge Classifier, we achieved 100% accuracy for the Plasma sample. For the Serum sample, we obtained 90.91% accuracy using XGBoost Classifier. However, We understand that no approach can be superior to another in every aspect, but it can be better in some aspects. Therefore, it is clear from the results that our approach is more superior compared to the previous ones. Our methodology is efficient both in terms of the predictive accuracy and the reduced number of metabolites for Serum and Plasma samples. The number of most dominant metabolites was only 19 and 7 in the case of Plasma and Serum samples, respectively.

6. Practical implications

The proposed methodology and the results from our study can be impactful in the field of bioinformatics and healthcare to get warned about a patient being affected with lung cancer. If we can be more and more accurate in the early diagnosis of cancer, it will save many lives. Machine learning implications to extract more and more insights, using the information of plasma and serum samples can be a blessing if we keep researching more. Not only cancer, but we can also predict other diseases with the information we are hoping to do as our future work.

7. Limitations and future research

We intend to contribute more to the field of metabolomics and thus plan to work with metabolomic biomarkers for lung cancer and other subtypes of cancer such as kidney, throat, pancreatic cancer, and other diseases. The methods we used in this paper can be applied to other cancer subtypes for future research. It is to be noted that the dataset used in this work is not as big as the datasets used in some other works. Thus, this work is to be extended to larger datasets. Further, research centering around the discussed matters in this section can greatly improve and contribute to the field of metabolomics, disease prediction, and so on. Moreover, we hope to extract people's data from other geographic regions to make the prediction more and more

generalized. Other information like age, past medical reports, bad habits, and routines can impact the diagnosis, and we would like to work on those too. However, we believe our approach can be a base procedure for accurately identifying metabolomic biomarkers needed to identify cancer patients accurately and for future research in the fields.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] American Cancer Society. Cancer facts & figures 2021. Atlanta: American Cancer Society; 2021, p. 17.

[2] American Cancer Society. Cancer facts & figures 2020. Atlanta: American Cancer Society; 2020, p. 17.

[3] Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The effect of advances in lung-cancer treatment on population mortality. N Engl J Med 2020;383(7):640–9.

[4] Mar J, Arrospide A, Iruretagoiena ML, Clèries R, Paredes A, Elejoste I, et al. Changes in lung cancer survival by TNM stage in the basque country from 2003 to 2014 according to period of diagnosis. Cancer Epidemiol 2020;65:101668.

[5] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26(1):51–78.

[6] Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. Metabolite profiling: from diagnostics to systems biology. Nature Rev Mol Cell Biol 2004;5(9):763–9.

[7] Davidson I, Ravi S. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: European conference on principles of data mining and knowledge discovery. Springer; 2005, p. 59–70.

[8] Royston P. Approximating the Shapiro-Wilk W-test for non-normality. Stat Comput 1992;2(3):117–9.

[9] de Gois G, de Oliveira-Júnior JF, da Silva Junior CA, Sobral BS, de Bodas Terassi PM, Junior AHSL. Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de Janeiro—Brazil. Theor Appl Climatol 2020;141(3):1573–91.

[10] Arsham H, Lovric M. Bartlett's test. 2011.

[11] Brown MB, Forsythe AB. Robust tests for the equality of variances. J Amer Statist Assoc 1974;69(346):364–7.

[12] Kumar N, Shahjaman M, Mollah MNH, Islam SS, Hoque MA. Serum and plasma metabolomic biomarkers for lung cancer. Bioinformation 2017;13(6):202.

[13] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Amer Statist Assoc 1952;47(260):583–621.

[14] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1):389–422.

[15] Miyamoto S, Taylor SL, Barupal DK, Taguchi A, Wohlgemuth G, Wikoff WR, et al. Systemic metabolomic changes in blood samples of lung cancer patients identified by gas chromatography time-of-flight mass spectrometry. Metabolites 2015;5(2):192–210.

[16] Masrur T, Hasan MAM. Identification of metabolomic biomarker using multiple statistical techniques and recursive feature elimination. In: 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2). IEEE; 2019, p. 1–4.

[17] Masrur T, Hasan MAM, Mondal MNI. Metabolomic biomarker identification for lung cancer by combining multiple statistical approaches. In: 2019 international conference on electrical, computer and communication engineering (ECCE). IEEE; 2019, p. 1–6.

[18] Zheng Y, He Z, Kong Y, Huang X, Zhu W, Liu Z, et al. Combined metabolomics with transcriptomics reveals important serum biomarkers correlated with lung cancer proliferation through a calcium signaling pathway. J Proteome Res 2021.

[19] Xie Y, Meng W-Y, Li R-Z, Wang Y-W, Qian X, Chan C, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl Oncol 2021;14(1):100907.

[20] Ruiying C, Zeyun L, Yongliang Y, Zijia Z, Ji Z, Xin T, et al. A comprehensive analysis of metabolomics and transcriptomics in non-small cell lung cancer. PLoS One 2020;15(5):e0232272.

[21] Zhang Y-H, Jin M, Li J, Kong X. Identifying circulating mirna biomarkers for early diagnosis and monitoring of lung cancer. Biochimica Et Biophysica Acta (BBA) Mol Basis Disease 2020;1866(10):165847.

[22] Yuan F, Lu L, Zou Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica Et Biophysica Acta (BBA) Mol Basis Disease 2020;1866(8):165822.

[23] Zhang L, Zheng J, Ahmed R, Huang G, Reid J, Mandal R, et al. A high-performing plasma metabolite panel for early-stage lung cancer detection. Cancers 2020;12(3):622.

[24] Shu T, Ning W, Wu D, Xu J, Han Q, Huang M, et al. Plasma proteomics identify biomarkers and pathogenesis of COVID-19. Immunity 2020;53(5):1108–22.

[25] Chirinos JA, Orlenko A, Zhao L, Basso MD, Cvijic ME, Li Z, et al. Multiple plasma biomarkers for risk stratification in patients with heart failure and preserved ejection fraction. J Am Coll Cardiol 2020;75(11):1281–95.

[26] Uchiyama K, Yagi N, Mizushima K, Higashimura Y, Hirai Y, Okayama T, et al. Serum metabolomics analysis for early detection of colorectal cancer. J Gastroenterol 2017;52(6):677–94.

[27] Nishiumi S, Kobayashi T, Kawana S, Unno Y, Sakai T, Okamoto K, et al. Investigations in the possibility of early detection of colorectal cancer by gas chromatography/triple-quadrupole mass spectrometry. Oncotarget 2017;8(10):17115.

[28] Long Y, Sanchez-Espiridion B, Lin M, White L, Mishra L, Raju GS, et al. Global and targeted serum metabolic profiling of colorectal cancer progression. Cancer 2017;123(20):4066–74.

[29] Gohlke RS. Time-of-flight mass spectrometry and gas-liquid partition chromatography. Anal Chem 1959;31(4):535–41.

[30] Razali NM, Wah YB, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. J Stat Model Anal 2011;2(1):21–33.

[31] Gorbunova AA, Lemeshko BY. Application of parametric homogeneity of variances tests under violation of classical assumption. In: Proceedings, 2nd stochastic modeling techniques and data analysis international conference; 2012. pp. 5–8.

[32] Student. The probable error of a mean. Biometrika 1908;6(1):1–25, URL http://www.jstor.org/stable/2331554.

[33] Breiman L. Random forests. Mach Learn 2001;45(1):5–32, URL https://doi.org/10.1023/A:1010933404324.

[34] Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr Intell Lab Syst 2006;83(2):83–90.

[35] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55–67.

[36] Ridgeclassifier, linear model, sklearn. 2007, Accessed: 12-05-2021, URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html,

[37] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm sigkdd international conference on knowledge discovery and data mining; 2016. pp. 785–94.

[38] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the fourteenth international joint conference on artificial intelligence, IJCAI 95, MontrÉAl QuÉBec, Canada, August 20-25 1995, 2 Volumes. Morgan Kaufmann; 1995, p. 1137–45.