Advances in Applied Mathematics 应用数学进展, 2023, 12(5), 2532-2545
Published Online May 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.125254

Hans汉斯

机器学习预测癌症五年存活率

杨心蕙

青岛大学数学与统计学院，山东 青岛

收稿日期：2023年4月28日；录用日期：2023年5月21日；发布日期：2023年5月31日

摘　要

癌症生存率对于癌症患者的临床治疗具有重要的意义，本篇论文旨在探究出可以准确预测癌症患者五年生存率的机器学习方法。采用的数据特征是TCGA网站上下载的多组学数据。我们探究出mRMR特征选择法和逻辑回归分类器以及SVM分类器的方法组合可以使五年存活率的准确率达到0.85以上，甚至可以超过0.9。由于我们分类时采用的是五折交叉验证，可以表明我们的结果稳健性较高。同时这两种方法组合的结果中AUC值和F1值也比较高，再次证实了这两种方法组合的优势。

关键词

五年生存率，机器学习，多组学

Machine Learning Predicts Five-Year Cancer Survival Rates

Xinhui Yang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Apr. 28th, 2023; accepted: May 21st, 2023; published: May 31st, 2023

Abstract

Cancer survival is of great importance to the clinical management of cancer patients and the aim of this thesis is to explore machine learning methods that can accurately predict the five-year survival rate of cancer patients. The data features used are multi-omics data downloaded from the TCGA website. We explore that the combination of the mRMR feature selection method and the logistic regression classifier and SVM classifier can result in an accuracy of more than 0.85 and even more than 0.9 for the five-year survival rate. Since we use a five-fold cross-validation for our classification, our results are robust. Also the AUC and F1 values are higher in the results of the combination of these two methods, which again confirms the advantages of the combination of these

文章引用: 杨心蕙. 机器学习预测癌症五年存活率[J]. 应用数学进展, 2023, 12(5): 2532-2545.
DOI: 10.12677/aam.2023.125254

two methods.

Keywords

Five-Year Survival Rate, Machine Learning, Multi-Omics

1. 引言

癌症是威胁人类生命的一项重大疾病，癌症导致全世界六分之一的死亡。癌症生存率一直是人类关注的主要健康问题。癌症的治疗在很大程度上取决于对预后的判断，准确预测癌症预后是定制和个性化治疗方案的基础。之前的已经有很多研究针对于不同癌症的患者经过治疗后的生存能力的预测，其中预测患者的五年生存率不只是每位患者关心的重大问题，也是一项重大的公共卫生问题[1]。

为了准确预测癌症患者的五年生存率，研究人员已经提出了很多有效的癌症预后的统计分析方法[2]，例如 Cox 比例风险回归、Kaplan Meier 估计和对数秩检验。这些用于预测的癌症五年生存率方法的主要数据来源主要是临床数据，包括癌症诊断、癌症类型、肿瘤分级等。在最近的研究中，研究者们开发了更多类型的数据用于更好地了解癌症状态以预测五年生存率。这些数据是患者样本的高通量和高维多组学数据。多组学数据包括基因组数据(即全基因组数据、单核苷酸多态性(SNP)数据、拷贝数变异(CNV)数据等)、表达数据(即 mRNA 和 miRNA 数据)、蛋白质组数据和表观遗传数据(即甲基化和其他染色体修饰) [3]。由于多组学数据高通量高数据量的属性，使得传统的统计方法在应用方面面临巨大的挑战。

研究人员开发并应用了其他的方法包括机器学习方法来解决这些问题。其中一些机器学习方法在预测癌症生存率方面取得了令人满意的效果，包括：支持向量机分类器、贝叶斯网络和决策树等。机器学习方法已成为癌症预后研究人员的流行工具[4]。

本篇论文基于三个癌症数据应用不同特征降维方法和机器学习分类器方法，对癌症患者进行二分类，标签分别为“0”，“1”，表示癌症患者生存没有超过 5 年和生存超过 5 年。通过比较分析，在我们所应用的特征降维方法和分类方法中确定了能使分类准确率最高的方法组合，并且在另外的癌症上测试确定的方法的效果。

2. 数据和方法

2.1. 数据收集

此次研究用到的 TCGA 数据均下载于 https://xenabrowser.net/datapages/网站。共下载三个癌症数据分别是乳腺癌(BLCA)、膀胱癌(BRCA)、多发性骨髓癌(LGG)。

由于我们要研究的问题是癌症患者的五年生存率，因此我们需要用到患者的五年生存信息。患者的生存信息来自 xena 网站的生存数据。确定每位患者的五年生存信息只需生存数据中的“OS”和“OS.time”列。我们结合这两列给定生存信息中的癌症患者标签“0”或“1”。给定标签的依据为：OS 表示从随机化开始至(因任何原因)死亡的时间。OS.time 表示患者发生该事件所用的时间。因此如果某位患者对应的 OS 为 0 表示患者仍然存活，若此时其对应的 OS.time 大于 5 年，这说明，该患者生存超过 5 年，此时我

们会给定标签“1”，若该患者对应的 OS.time 小于 5 年，并不能说明该患者存活超过 5 年，这种情况下，我们并不能确定该患者的标签；另外，如果某位患者对应的 OS 为 1 表示该患者已经死亡，若其对应的 OS.time 少于 5 年，说明该患者存活没有超过 5 年，此时我们会给定标签为“0”，若该患者对应的 OS.time 大于 5 年，这说明该患者存活超过五年后死亡，此时我们会给定标签为“1”。

接下来我们需要收集不同癌症患者的多组学特征数据。不同癌症用到的多组学数据不同，见表 1。

Table 1. Multi-omics data for different cancers
表 1. 不同癌症的额多组学数据

癌症类型	多组学数据
BRCA	CNV RNA RPPA
BLCA	CNV RNA miRNA RPPA
LGG	CNV RNA exon

其中，CNV 表示基因水平的拷贝数变异数据、RNA 表示 RNA 上的基因表达数据、miRNA 表示 miRNA 上的基因表达数据、RPPA 表示蛋白质表达数据、exon 表示外显子表达数据。

2.2. 数据处理

2.2.1. 数据预处理

由于下载的原始数据存在缺失值和非标准化等问题，我们需要对数据进行预处理才能采用其他方法对多组学数据进行降维。

对于直接在网站上下载下来的数据我们首先对该数据进行缺失值处理。因为不同组学数据中存在缺失值的列很多，如果我们直接删除存在缺失值的所有列，会导致一部分重要信息被删除，为了确保我们保留了原始数据的重要信息，而且有效处理确实信息，我们只删除缺失值占该列数据的比例超过 20%的列。在每个癌症下将不同的组学数据分别删除超过给定阈值的特征列后，我们将不同组学数据和生存数据中的样本取交集，这样我们就可以得到所要研究的患者样本的标签和特征。其中样本标签为“0”、“1”分别表示样本生存没有超过 5 年和样本生存超过 5 年，样本特征即为不同癌症下的多组学特征。

我们对确定标签后的样本进行整理，不同癌症的样本及其标签分布见表 2：

Table 2. Distribution of labels for different cancer samples
表 2. 不同癌症样本标签分布

癌症类型	0	1
BRCA	81	200
BLCA	142	38
LGG	98	65

接下来我们继续处理数据中的缺失值，我们采用 k 紧邻插补法对缺失值进行插补。由于多组学数据量巨大，大量的特征中存在很多与预测目标不想管的特征，若我们在进行目标分类时考虑到这些特征的话，会对我们的分类结果造成不良的影响。因此我们要删除与预测目标不相关的特征。我们过滤掉没有信号(零均值)或低方差的特征。采用的方法时方差筛选，即删除方差小于给定阈值的列。由于不用组学数据的原始数据大小不同，取值范围不同，并且特征的个数也不同，不同组学的方差筛选阈值也是不同的。

不同组学数据的阈值见表 3：

Table 3. Variance screening thresholds for different histological data
表 3. 不同组学数据的方差筛选阈值

多组学数据	阈值
CNV	0.001
RNA	0.1
miRNA	0.1
RPPA	0
exon	0.05

对于 RPPA 表达数据，我们只过滤掉没有变化(方差等于零)的特征，因为可用特征因 RPPA 数量少而受到限制。所有癌症类型都使用相同的方差阈值。

下面我们对特征进行标准化，采用的是最大最小值标准化的方法。

数据预处理流程见图 1：

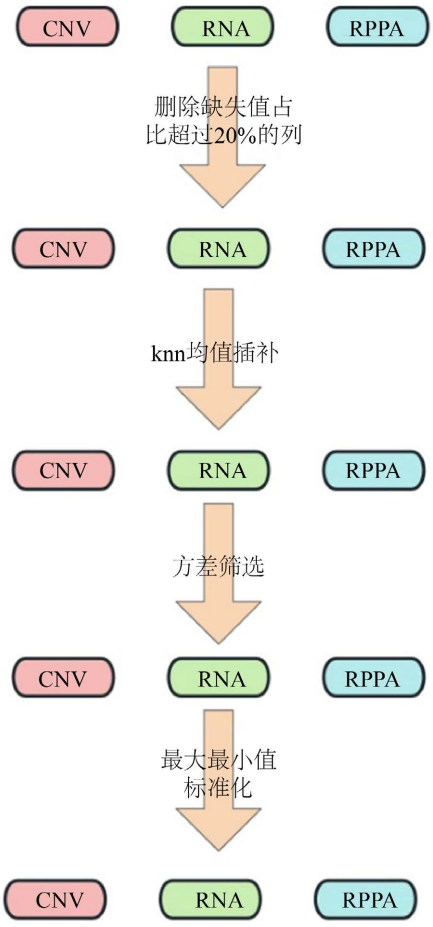

Figure 1. Data pre-processing flow chart
图 1. 数据预处理流程图

2.2.2. 特征降维

我们在筛选出变化大的特征并且对特征数据进行标准化后，对数据进行机器学习降维。我们采用三种降维方式，分别是随机森林特征选择方法，主成分分析法(pca)和最大相关最小冗余特征选择方法(mRMR 特征选择)，在后面的结果部分会展示这三种降维方法结果的优劣。下面分表表述这三种降维方法在不同癌症下应用的细节。

随机森林特征选择方法：

在进行随机森林特征选择时我们将每个癌症的不同组学数据串联起来一起进行特征选择。我们给予 10000 颗决策树平均不纯度衰减的计算来评估特征重要性。在随机森林特征选择方法的结果中，我们为每一个特征计算其重要分数，并按照重要分数由大到小对特征进行排序。重要分数越高，说明该特征对于标签分类越重要。给定一个阈值，筛选出特征重要性大于给定阈值的特征，这些筛选出的特征即为我们要保留的特征。

在以往的研究中，每个癌症的在特征选择后剩余的特征个数为 1000 左右，因此我们要保留的特征也为 1000 左右。当阈值为 0.0002 时，每个超过阈值的特征大约在 1000 左右，因此在三个癌症的阈值相同，都为 0.0002。因为我们在进行随机森林特征选择时是所有癌症放在一起选择的。每个癌症下的特征数据经过特征选择分别剩下的特征个数为：1399、1260、903。

主成分分析法：

在进行主成分分析降维时我们在每个癌症下将多组学数据分开处理，不同组学数据设定相同的方差贡献率 0.95。经过主成分分析法降维后的不同癌症下的不同组学特征数量在表 中展示。

最大相关最小冗余特征选择方法：

在应用 mRMR 特征选择方法时需要事先指定需要筛选出的特征的个数。由于 CNV、RNA、miRNA 的样本量大，RPPA、exon 的样本量小，因此我们对前面几个多组学数据给定的需要筛选出的特征个数为 100，后面两个组学数据的特征个数为 50。我们根据 pca 降维后各组学的特征数量给出 mrmr 特征选择各组学所需的特征数量为 100 或 50。为确保我们给定的特征数量的合理性我们对其进行调参。根据在不同特征个数下分类结果的准确性、AUC 和 F1 的值来确保特征个数为 100 和 50 的合理性。调参结果在结果部分进行展示。

经过上面的处理，我们就可以根据筛选出的特征结合癌症患者对应的标签，对其进行分类。

经过一系列处理的不同癌症组学数据的特征数量见表 4：

Table 4. Number of features for different cancer histology data
表 4. 不同癌症组学数据的特征数量

		原始	缺失值处理	方差筛选	随机森林特征选择	pca 特征选择	Mrmr 特征选择
BRCA	CNV	24776	24776	24776	1399	118	100
	RNA	20530	20530	19004		211	100
	RPPA	281	189	189		90	50
BLCA	CNV	24776	24776	24776	1260	101	100
	RNA	20530	20530	19180		150	100
	miRNA	2210	549	548		109	100
	RPPA	245	195	195		80	50
LGG	CNV	24776	24776	24776	903	58	100
	RNA	20530	20530	17891		120	100
	exon	485577	395708	27801		118	50

根据上表，我们可以看出，在同一种癌症下随机森林筛选出的特征数量是最多的，主成分分析法次之，mRMR 特征选择方法选出的特征是最多的。但是对于同一种特征降维方法，三种癌症筛选出的特征数量是差不多的。

下面我们就可以利用分类器根据整理完成的数据对患者样本进行分类，根据分类结果评价各种特征选择方法和分类器效果的好坏。

3. 结果

3.1. 机器学习方法结果比较

经过对不同癌症的多组学数据进行预处理、降维后可以将数据输入分类器进行分类。对于每个癌症均采用 9 个分类器进行分类，为了保证结果的稳健性，均采用五折交叉验证。下表中展示了 9 个分类器在五折交叉验证下的三个指标值，分别是准确率(ACC)、AUC 和 F1。

结果见表 5：

Table 5. Classification forecast results
表 5. 分类预测结果

			逻辑回归	随机森林	决策树	XGBoost	LGBM	KNC	Adaboost	svm	朴素贝叶斯
BRCA	RF	ACC	0.75 (+/−0.04)	0.72 (+/−0.03)	0.71 (+/−0.05)	0.73 (+/−0.03)	0.74 (+/−0.05)	0.71 (+/−0.01)	0.75 (+/−0.03)	0.76 (+/−0.03)	0.71 (+/−0.09)
		AUC	0.77 (+/−0.05)	0.73 (+/−0.07)	0.58 (+/−0.05)	0.63 (+/−0.07)	0.69 (+/−0.07)	0.54 (+/−0.10)	0.75 (+/−0.03)	0.78 (+/−0.05)	0.70 (+/−0.09)
		F1	0.73 (+/−0.05)	0.66 (+/−0.04)	0.68 (+/−0.05)	0.67 (+/−0.04)	0.70 (+/−0.07)	0.62 (+/−0.04)	0.72 (+/−0.04)	0.75 (+/−0.03)	0.71 (+/−0.09)
	pca	ACC	0.68 (+/−0.04)	0.70 (+/−0.03)	0.64 (+/−0.05)	0.65 (+/−0.03)	0.69 (+/−0.04)	0.72 (+/−0.01)	0.57 (+/−0.04)	0.70 (+/−0.02)	0.71 (+/−0.01)
		AUC	0.70 (+/−0.03)	0.61 (+/−0.07)	0.53 (+/−0.04)	0.56 (+/−0.09)	0.58 (+/−0.06)	0.56 (+/−0.07)	0.57 (+/−0.04)	0.69 (+/−0.07)	0.66 (+/−0.08)
		F1	0.67 (+/−0.04)	0.60 (+/−0.04)	0.61 (+/−0.03)	0.59 (+/−0.05)	0.61 (+/−0.04)	0.62 (+/−0.02)	0.61 (+/−0.01)	0.62 (+/−0.04)	0.59 (+/−0.01)
	mrmr	ACC	**0.86 (+/−0.05)**	0.79 (+/−0.03)	0.70 (+/−0.09)	0.78 (+/−0.05)	0.79 (+/−0.04)	0.78 (+/−0.05)	**0.85 (+/−0.03)**	**0.87 (+/−0.05)**	0.86 (+/−0.05)
		AUC	0.92 (+/−0.04)	0.81 (+/−0.07)	0.61 (+/−0.11)	0.83 (+/−0.05)	0.83 (+/−0.07)	0.78 (+/−0.08)	0.85 (+/−0.03)	0.92 (+/−0.04)	0.89 (+/−0.03)
		F1	0.86 (+/−0.05)	0.75 (+/−0.05)	0.69 (+/−0.08)	0.76 (+/−0.04)	0.78 (+/−0.04)	0.75 (+/−0.06)	0.81 (+/−0.05)	0.86 (+/−0.06)	0.85 (+/−0.06)
BLCA	RF	ACC	0.77 (+/−0.05)	0.79 (+/−0.02)	0.69 (+/−0.08)	0.77 (+/−0.05)	0.82 (+/−0.02)	0.77 (+/−0.03)	0.64 (+/−0.06)	0.81 (+/−0.02)	0.75 (+/−0.06)
		AUC	0.69 (+/−0.06)	0.75 (+/−0.06)	0.50 (+/−0.06)	0.70 (+/−0.14)	0.71 (+/−0.09)	0.49 (+/−0.08)	0.62 (+/−0.08)	0.73 (+/−0.05)	0.65 (+/−0.09)
		F1	0.72 (+/−0.05)	0.71 (+/−0.04)	0.67 (+/−0.06)	0.71 (+/−0.05)	0.77 (+/−0.03)	0.70 (+/−0.04)	0.73 (+/−0.04)	0.74 (+/−0.02)	0.72 (+/−0.07)

Continued

		ACC	0.72 (+/−0.05)	0.79 (+/−0.02)	0.72 (+/−0.03)	0.78 (+/−0.02)	0.78 (+/−0.04)	0.77 (+/−0.02)	0.44 (+/−0.09)	0.79 (+/−0.01)	0.79 (+/−0.01)
	pca	AUC	0.54 (+/−0.06)	0.51 (+/−0.12)	0.47 (+/−0.04)	0.45 (+/−0.09)	0.50 (+/−0.12)	0.48 (+/−0.06)	0.44 (+/−0.09)	0.49 (+/−0.10)	0.54 (+/−0.07)
		F1	0.69 (+/−0.04)	0.71 (+/−0.04)	0.67 (+/−0.02)	0.69 (+/−0.02)	0.71 (+/−0.06)	0.70 (+/−0.02)	0.67 (+/−0.02)	0.70 (+/−0.02)	0.70 (+/−0.02)
		ACC	**0.93 (+/−0.01)**	0.82 (+/−0.04)	0.74 (+/−0.06)	0.78 (+/−0.04)	0.79 (+/−0.05)	0.83 (+/−0.03)	0.80 (+/−0.07)	**0.93 (+/−0.02)**	0.91 (+/−0.02)
	mrmr	AUC	0.98 (+/−0.01)	0.85 (+/−0.05)	0.62 (+/−0.08)	0.66 (+/−0.07)	0.75 (+/−0.08)	0.84 (+/−0.07)	0.80 (+/−0.07)	0.97 (+/−0.02)	0.96 (+/−0.02)
		F1	0.93 (+/−0.01)	0.75 (+/−0.07)	0.74 (+/−0.06)	0.70 (+/−0.04)	0.75 (+/−0.06)	0.78 (+/−0.04)	0.78 (+/−0.08)	0.93 (+/−0.02)	0.90 (+/−0.03)
		ACC	0.83 (+/−0.03)	0.80 (+/−0.05)	0.71 (+/−0.06)	0.78 (+/−0.07)	0.81 (+/−0.05)	0.77 (+/−0.02)	0.89 (+/−0.04)	0.82 (+/−0.04)	0.75 (+/−0.08)
	RF	AUC	0.93 (+/−0.02)	0.88 (+/−0.03)	0.72 (+/−0.05)	0.88 (+/−0.05)	0.89 (+/−0.04)	0.84 (+/−0.04)	0.89 (+/−0.04)	0.92 (+/−0.02)	0.84 (+/−0.07)
		F1	0.84 (+/−0.03)	0.80 (+/−0.05)	0.70 (+/−0.06)	0.78 (+/−0.07)	0.81 (+/−0.05)	0.77 (+/−0.02)	0.77 (+/−0.05)	0.82 (+/−0.04)	0.75 (+/−0.08)
		ACC	0.77 (+/−0.05)	0.69 (+/−0.06)	0.69 (+/−0.05)	0.73 (+/−0.04)	0.72 (+/−0.05)	0.71 (+/−0.03)	0.83 (+/−0.03)	0.76 (+/−0.05)	0.64 (+/−0.02)
LGG	pca	AUC	0.88 (+/−0.05)	0.78 (+/−0.10)	0.70 (+/−0.06)	0.78 (+/−0.05)	0.82 (+/−0.06)	0.79 (+/−0.04)	0.83 (+/−0.03)	0.86 (+/−0.05)	0.84 (+/−0.06)
		F1	0.77 (+/−0.05)	0.66 (+/−0.06)	0.69 (+/−0.05)	0.73 (+/−0.05)	0.72 (+/−0.05)	0.70 (+/−0.04)	0.76 (+/−0.06)	0.76 (+/−0.05)	0.55 (+/−0.04)
		ACC	**0.93 (+/−0.02)**	0.82 (+/−0.03)	0.76 (+/−0.04)	0.85 (+/−0.06)	0.88 (+/−0.04)	0.88 (+/−0.07)	**0.97 (+/−0.03)**	**0.93 (+/−0.03)**	0.82 (+/−0.04)
	mrmr	AUC	0.98 (+/−0.02)	0.94 (+/−0.03)	0.76 (+/−0.04)	0.94 (+/−0.02)	0.96 (+/−0.02)	0.94 (+/−0.05)	0.97 (+/−0.03)	0.99 (+/−0.01)	0.95 (+/−0.03)
		F1	0.93 (+/−0.02)	0.82 (+/−0.03)	0.76 (+/−0.04)	0.85 (+/−0.06)	0.88 (+/−0.04)	0.88 (+/−0.07)	0.93 (+/−0.05)	0.93 (+/−0.03)	0.82 (+/−0.04)

我们将准确率最高的数值加粗标记出来。由上表我们可以发现，在三种特征选择方法中 mRMR 特征选择法对应的准确率是最高的，并且由 mRMR 特征选择方法筛选出的特征数量是最少的，因此在我们所采用的三种特征选择方法中 mRMR 特征选择法表现是最好的。其次我们还可以看出，在我们所采用的九种分类器中，逻辑回归和 mRMR 分类器在这三个癌症样本中准确率是最高的。并且其 AUC 和 F1 的值也是最高的。因此我们可以得出结论，在我们所采用的三个特征选择方法和九种分类器方法中能够使准确率达到最高并且 AUC 和 F1 的值也表现最好的方法组合为 mRMR 特征选择法和逻辑回归分类器、以及 mRMR 特征选择法和 svm 分类器。

3.2. mRMR 特征选择法调参结果

由于 mRMR 特征选择法有一个重要的参数，即需要筛选出的特征的个数，这个参数是我们需要实现给定的。我们最终结果采用的参数值为 100 或 50，为证明参数选取的合理性，我们该参数进行调参。由

于我们上面已经探讨出能够使分类结果表现最好的两组方法组合，因此在调参过程中我们采用这两个方法组合进行调参结果的比较。

调参结果的折线图见图 2、图 3：

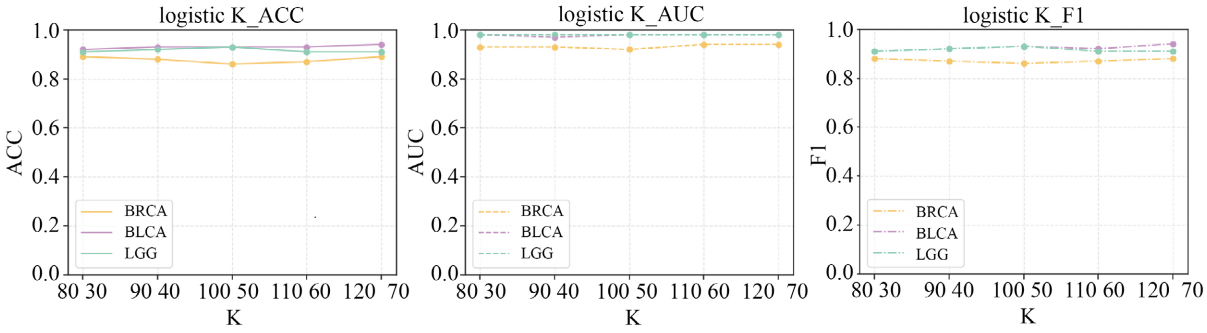

Figure 2. Line graph of the change in the value of each indicator in a logistic regression classifier during the process of co-variance

图 2. 逻辑回归分类器在调参过程中各指标值变化的折线图

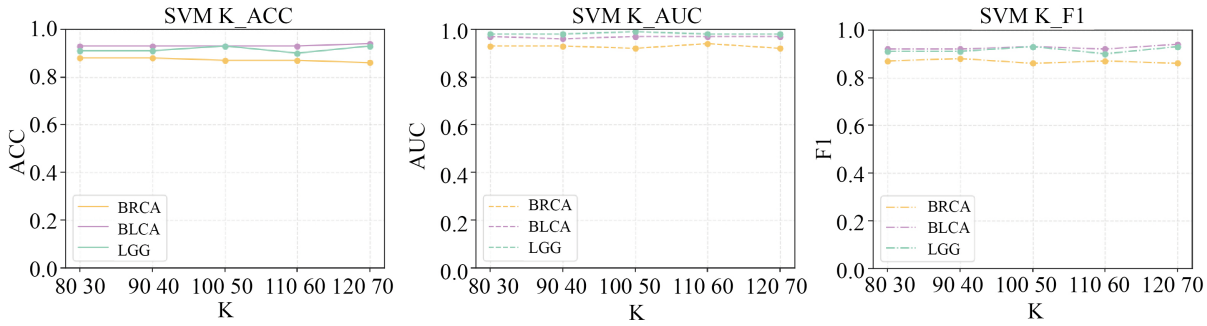

Figure 3. Line graph of the change in the value of each indicator during the tuning process of the SVM classifier

图 3. SVM 分类器在调参过程中各指标值变化的折线图

上面两组图片分别表示分类器为逻辑回归分类器和 SVM 分类器在特征选择方法为 mRMR 的不同特征个数下的结果评价标准的值。在每组图片中由左至右分别是 ACC、AUC、F1。

由上面两图可以看出，当我们给定 mRMR 特征选择方法要选出不同的特征数量时，三种癌症下的 ACC、AUC 和 F1 的值变化并不大，这说明我们选择的方法稳健性很强。结合三种癌症的分类结果，我们考虑整体的分类效果，将 mRMR 特征选择方法需要筛选出的特征数量指定为 100 和 50。

3.3. 单组学数据预测结果

针对表现效果最好的两种组合方法：mRMR 特征选择法、逻辑回归分类法和 mRMR 特征选择方法、SVM 分类器，我们将其应用带三个癌症的不同组学数据上，分别将不同组学数据单独进行分类预测，任意两个或三个组学数据组合进行预测，预测结果见图 4~6。

观察下面三组图，我们可以发现总体来看在每个癌症下多组学数据预测效果要比单组学效果好，但是并不是考虑到所有组学数据的效果是最好的。在单组学进行预测时 CNV 数据集的预测效果时最差的，RNA 数据集的效果是最好的，其次是 exon 数据集和 miRNA 数据集。并且我们可以发现在进行多组学数据预测时，包含 CNV 数据集的多组学数据预测效果要差一些，包含 RNA 数据集的预测效果要好一些。

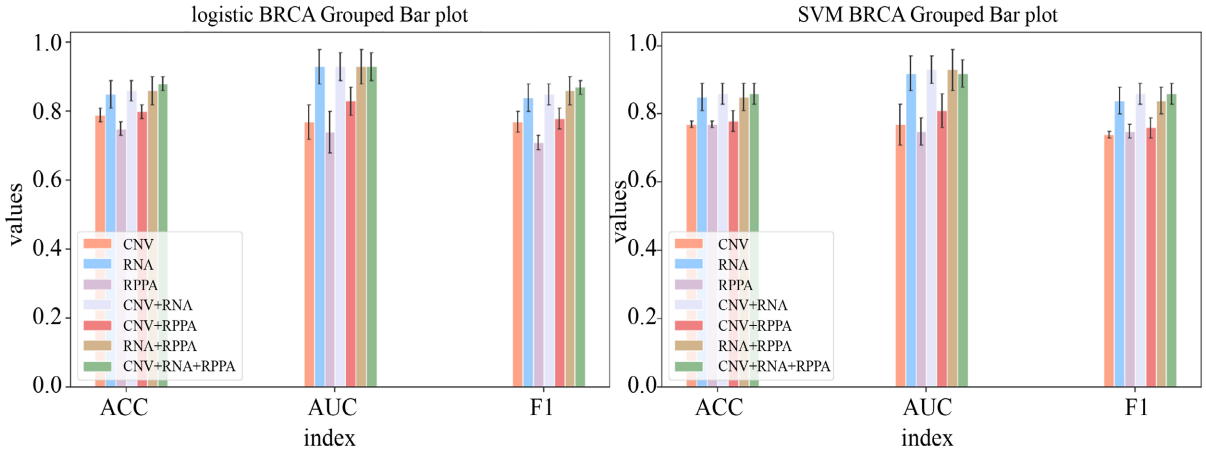

Figure 4. Histogram of results of logistic regression classifier and SVM classifier for each histology data under BRCA
图 4. BRCA 下各组学数据逻辑回归分类器和 SVM 分类器结果柱状图

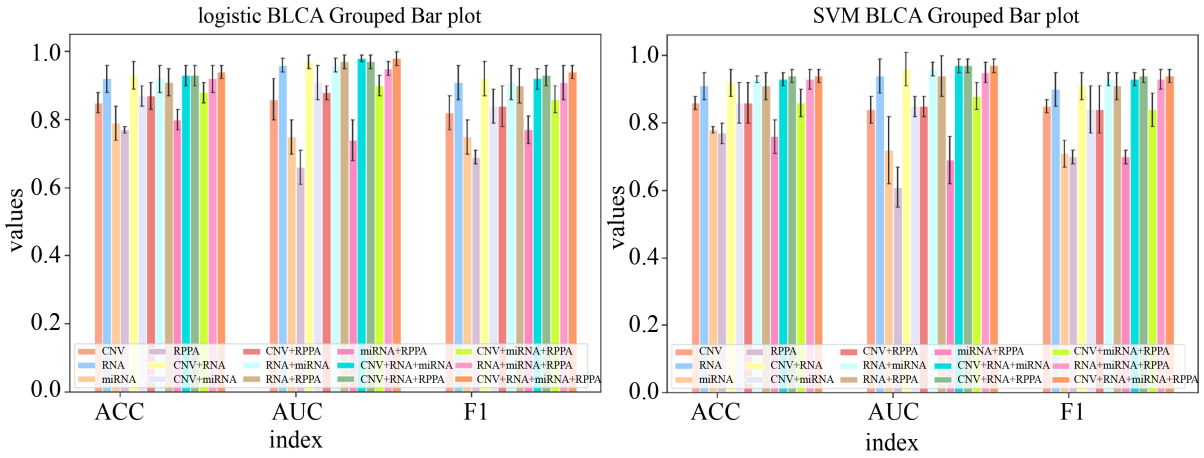

Figure 5. Histogram of results of logistic regression classifier and SVM classifier for each histology data under BLCA
图 5. BLCA 下各组学数据逻辑回归分类器和 SVM 分类器结果柱状图

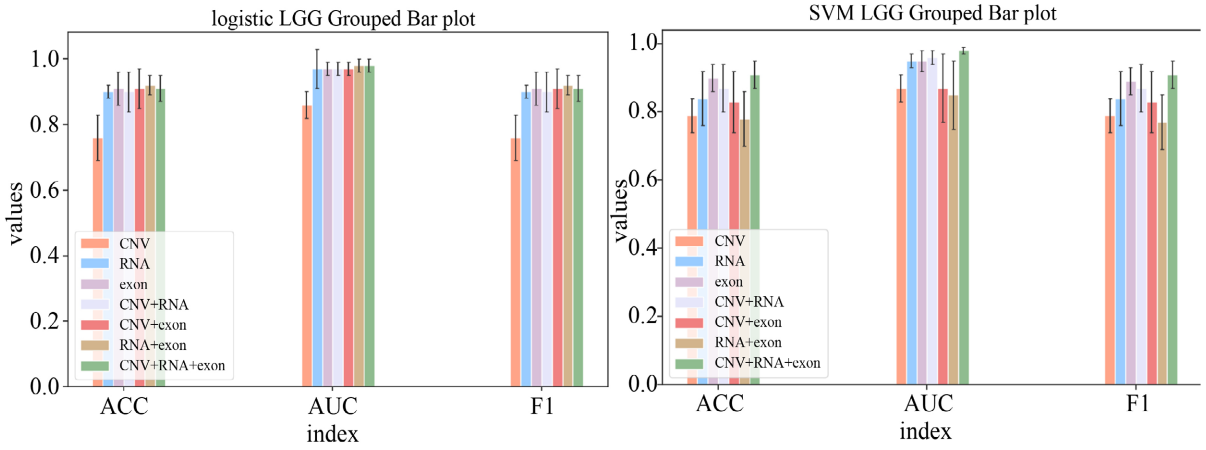

Figure 6. Histogram of results of logistic regression classifier and SVM classifier for each histology data under LGG
图 6. LGG 下各组学数据逻辑回归分类器和 SVM 分类器结果柱状图

4. 方法原理

4.1. 预处理方法

4.1.1. k 近邻插补

k 近邻插补(knearestneighborimputation)是 olgatroyanskaya 提出的一种基于数据局部相似性的插补方法。每个样本的缺失值都是使用训练集中找到的 n_neighbors 最近邻的平均值估算的。如果两个样本都没有丢失的特征很接近，那么这两个样本就是相近的。

k 近邻插补的基本思想是，对于含缺失值的样本，其缺失的数据可参考与它最类似的 k 个样本。具体地说，k 近邻插补将数据集划分为两个集合，一个集合包含所有的完全样本(即不含缺失值的样本)，另外一个集合包含所有的不完全样本(即存在缺失值的样本)。对于每个不完全样本，求其在完全样本集中的 k 近邻，对于缺失值是分类属性的，则插补 k 近邻样本中该属性值的众数；对于缺失值是数值属性的，则插补 k 近邻样本中该属性值的平均数。由于不完全样本的缺失值是根据“相邻”样本求得，因此 k 近邻插补方法不会增加过多的新样本信息[5]。

4.1.2. 方差筛选

方差筛选法又叫做方差过滤法。方差是衡量一个变量的离散程度(即数据偏离平均值程度的大小)。变量的方差越大，我们就可以认为它的偏离程度越大，也就是意味着这个变量对模型的贡献和作用会更加明显，因此要保留方差较大的变量。反之要剔除没有意义的特征[6]。

4.2. 降维方法

4.2.1. 随机森林特征选择法

随机森林算法本质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。特征选择采用随机的方法去分裂每一个节点，然后比较不同情况下产生的误差。能够检测到的内在估计误差、分类能力和相关性决定选择特征的数目。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类[7]。

4.2.2. 主成分分析法

主成分分析法(pca)是一种运用线性代数的知识来进行数据降维的方法，它将多个变量转换为少数几个不相关的综合变量来比较全面地反映整个数据集。这是因为数据集中的原始变量之间存在一定的相关关系，可用较少的综合变量来综合各原始变量之间的信息。这些综合变量称为主成分，各主成分之间彼此不相关，即所代表的的信息不重叠[8]。

4.2.3. mRMR 特征选择方法

mRMR 最初由 peng 提出，被机器学习研究人员用于 DNA microarray data 分类。后来被广泛应用于其他领域，比如：眼动作识别、性别分类或者分析多光谱卫星图像。本篇论文将其应用于多组学特征的筛选。

mRMR 被用来对一组特征在特定分类任务中的重要性进行排序。这种方法可以根据特征与目标的相关性进行排序，同时，特征的冗余度也会受到惩罚。其主要目的是利用相互信息(MI) (用 I 表示)找到一组特征 X 和类别 c 之间的最大依赖性。

由于计算成本高并且样本量少，另一个好的计算方法是确定最大相关性标准。又因为根据最大相关性标准来计算会导致大量的冗余。因此我们必须添加特征冗余的惩罚。结合上面两个思想得到 Mrmr 特

征选择方法的准则。在实践中，可以采用一种贪婪的算法，求得最终筛选出的特征。综上所述，mRMR的原理：在原始特征集合中找到与最终输出结果相关性最大(Max-Relevance)，但是特征彼此之间相关性最小的一组特征(Min-Redundancy) [9]。

4.3. 分类方法

4.3.1. 逻辑回归

逻辑回归[10]也称作 logistic 回归分析，是一种广义的线性回归分析模型，属于机器学习中的监督学习。其推导过程与计算方式类似于回归的过程，但实际上主要是用来解决二分类问题(也可以解决多分类问题)。通过给定的 n 组数据(训练集)来训练模型，并在训练结束后对给定的一组或多组数据(测试集)进行分类。其中每一组数据都是由 p 个指标构成。

我们所要处理的二分类问题。由于分成两类，我们便让其中一类标签为 0，另一类为 1。我们需要一个函数，对于输入的每一组数据，都能映射成 0~1 之间的数。并且如果函数值大于 0.5，就判定属于 1，否则属于 0。而且函数中需要待定参数，通过利用样本训练，使得这个参数能够对训练集中的数据有很准确的预测。

4.3.2. 随机森林分类器

随机森林[11]是非常具有代表性的 Bagging (装袋法)集成算法，它的所有基评估器都是决策树，分类树组成的森林就叫做随机森林分类器。随机森林就是指利用多棵树对样本进行训练并预测的一种分类器。随机森林是利用多个决策树对样本进行训练、分类并预测地一种算法，主要应用于回归和分类场景。在对数据进行分类地同时，还可以给出各个变量地重要性评分，评估各个变量在分类中所起地作用。分类时，每棵树都投票并且返回得票最多的类。

4.3.3. 决策树

决策树[12]是一种树型结构，其中每个内部节结点表示在一个属性上的测试，每一个分支代表一个测试输出，每个叶结点代表一种类别。决策树学习是以实例为基础的归纳学习。决策树学习采用的是自顶向下的递归方法，其基本思想是以信息熵为度量构造一棵熵值下降最快的树。到叶子节点的处的熵值为零，此时每个叶结点中的实例都属于同一类。决策树的划分依据之一是信息增益。决策树的指标可以选择基尼指数或者是交叉熵。直白来讲就是决策树是一颗树，要分类的样本从树根进入，在树的每个节点通过对样本的某种属性的判断选择不同的路径逐步下降到底,得出其所属类别。

4.3.4. XGBoost

XGBoost [13]并没有提出一种新的机器学习算法，而是基于现有的树打包算法在系统及算法层面(主要是系统层面)进行了改进；系统层面改进点包括：带权重的分位点分割算法、稀疏特征的处理、缓存、特征并行、基于 rabit 的分布式训练等，这些改进使得 xgboost 无论是在单机训练还是分布式训练上的耗时都比 pGBRT/scikit-learn/spark MLLib/R gbm 等有至少几倍的提升；算法层面的改进主要包括：L1/L2 正则化、目标函数二阶导的应用等；boosting 基本思想是叠加多个弱分类器的结果组合成一个强分类器，叠加方法是各个基分类器的结果做加法，在生成下一个基分类器的时候，目标是拟合历史分类器结果之和与 label 之间的残差，预测的时候是将每个基分类器结果相加；每个基分类器都是弱分类器，目前 xgboost 主要支持的基分类器有 CART 回归树、线性分类器。

4.3.5. LGBM

GBDT (Gradient Boosting Decision Tree)是机器学习中一个长盛不衰的模型，其主要思想是利用弱分

类器(决策树)迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT 不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务；在各种数据挖掘竞赛中也是致命武器，据统计 Kaggle 上的比赛有一半以上的冠军方案都是基于 GBDT。而 LightGBM (Light Gradient Boosting Machine)是一个实现 GBDT 算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点。常用的机器学习算法，例如神经网络等算法，都可以以 mini-batch 的方式训练，训练数据的大小不会受到内存限制。而 GBDT 在每一次迭代的时候，都需要遍历整个训练数据多次。如果把整个训练数据装进内存则会限制训练数据的大小；如果不装进内存，反复地读写训练数据又会消耗非常大的时间。尤其面对工业级海量的数据，普通的 GBDT 算法是不能满足其需求的。LightGBM 提出的主要原因就是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于实践[14]。

4.3.6. K 近邻分类器

最近邻分类器是在最小距离分类的基础上进行扩展的，将训练集中的每一个样本作为判断依据，寻找距离待分类样本中最近的训练集中的样本，以此依据进行分类。存在问题是噪声(图像噪声是指存在于图像数据中的不必要的或多余的干扰信息。)我们从上面的图中可以看出，绿色的区域内还有一个黄色区域，此时这个区域应该是绿色，因为黄色的点很有可能是噪声，但我们使用最近邻算法就有可能出现上面的问题。

与最近邻分类器类不同的是，k 近邻分类器是几个测试的样本共同抉择属于哪一个样本。如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中 k 通常是不大于 20 的整数。KNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别[15]。

4.3.7. Adaboost

Adaboost 算法[16]通过将多个弱分类器进行依次的顺序结合，层层递进，层层筛选，最终成为一个强分类器，结构如下图所示。Adaboost 对弱分类器进行迭代训练，每一个阶段训练好的弱分类器都将参与下一轮弱分类器的迭代，使得整个分类器更为严谨。因此，每次迭代都会生成一个新的弱分类器，假设有 n 次迭代，那么就有 n 个弱分类器，并且第 n 次迭代有 $n-1$ 个弱分类器参与进来。每一轮迭代完成的弱分类器中的各个参数都不会改变，只有当前训练的弱分类器，才会学习参数，并且当前学习的弱分类器，有机会纠正前面迭代伦次没有分类正确的数据，最终分类结果是否可靠，需要看训练的所有弱分类器的综合结果。从结构图中我们可以看出，在分类器中有两种数据权重，分别是数据权重和弱分类器权重。其中，弱分类器权重由弱分类器使用数据权重寻找当前分类精度较高的决策点，弱分类器权重越高，说明训练效果越好，则在整个分类器的最终决策中，能够起到更大的作用。

Adaboost 算法中有两种权重，一种是数据的权重，另一种是弱分类器的权重。其中，数据的权重主要用于弱分类器寻找其分类误差最小的决策点，找到之后用这个最小误差计算出该弱分类器的权重，分类器权重越大说明该弱分类器在最终决策时拥有更大的发言权。Adaboost 会对分错的数据加大权重，由于权重增大影响，那么下一轮弱分类器就有更大的把握将当前轮没有正确分类的数据分队，如果下一轮还是没有分对，那么这一点的数据权重将继续增加，这样一轮一轮的持续下去，后面的分类器将会更加注意这个数据的分类，这样将其分对的概率也就越高。

4.3.8. SVM

支持向量机(SVM) [17]是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界

是对学习样本求解的最大边距超平面，可以将问题化为一个求解凸二次规划的问题。与逻辑回归和神经网络相比，支持向量机，在学习复杂的非线性方程时提供了一种更为清晰，更加强大的方式。

具体来说就是在线性可分时，在原空间寻找两类样本的最优分类超平面。在线性不可分时，加入松弛变量并通过使用非线性映射将低维度输入空间的样本映射到高维度空间使其变为线性可分，这样就可以在该特征空间中寻找最优分类超平面。

支持向量机的分类方法，是在这组分布中找出一个超平面作为决策边界，使模型在数据上的分类误差尽量接近于小，尤其是在未知数据集上的分类误差(泛化误差)尽量小。边际很小的情况，是一种模型在训练集上表现很好，却在测试集上表现糟糕的情况，所以会“过拟合”。所以我们在找寻决策边界的时候，希望边际越大越好。支持向量机，就是通过找出边际最大的决策边界，来对数据进行分类的分类器。也因此，支持向量分类器又叫做最大边际分类器。这个过程在二维平面中看起来十分简单，但将上述过程使用数学表达出来，就不是一件简单的事情了。在几何中，超平面是一个空间的子空间，它是维度比所在空间小一维的空间。如果数据空间本身是三维的，则其超平面是二维平面，而如果数据空间本身是二维的，则其超平面是一维的直线。在二分类问题中，如果一个超平面能够将数据划分为两个集合，其中每个集合中包含单独的一个类别，我们就说这个超平面是数据的“决策边界”。和逻辑回归中的过程一样，SVM 也是通过最小化损失函数来求解一个用于后续模型使用的重要信息：决策边界。

4.3.9. 朴素贝叶斯分类器

朴素贝叶斯分类器[18]是一系列以假设特征之间强(朴素)独立下运用贝叶斯定理为基础的简单概率分类器。该分类器模型会给问题实例分配用特征值表示的类标签，类标签取自有限集合。它不是训练这种分类器的单一算法，而是一系列基于相同原理的算法：所有朴素贝叶斯分类器都假定样本每个特征与其他特征都不相关。

4.4. 分类算法的性能评估方法

k 折交叉验证方法用于评估所检查的数据挖掘模型的性能并比较分类模型的结果。交叉验证是一种重采样技术，用于评估未见数据样本中的数据挖掘技术。在此方法中，机器学习模型被训练和测试 k 次此外，为了比较分类模型的性能，使用了准确度、特异性、灵敏度、kappa 和曲线下面积(AUC)等评估指标的平均值。本论文采用准确率(ACC)，ROC 曲线下的面积(AUC)和 F1 分数[19]：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{F1}=\frac{2\times\mathrm{Pre}\times\mathrm{Recall}}{\mathrm{Pre}+\mathrm{Recall}}$$

5. 结论

本篇论文中我们在三种癌症(BRCA、BLCA、LGG)下结合多组学数据进行特征降维并根据降维后的特征对样本进行五年生存率的分类，每个机器学习分类方法均采用五折交叉验证。通过对比不同特征选择方法和分类方法组合的 ACC、AUC 和 F1 的值，我们发现 Mrmr 特征选择方法和与逻辑回归分类器和 SVM 分类器组合起来，对五年生存率的预测效果是最好的。在五折交叉验证下准确率可以达到 0.85 以上，并且其 AUC 和 F1 表现也很好。同时我们还可以发现，结合不同组学数据分类的结果是不同的，但是总

体来说，结合不同组学数据的分类准确率要比单组学数据好。

我们探索的方法组合在很大程度上准确预测了不同癌症患者的五年生存率，这一结果对于癌症患者的临床治疗有很重要的指导意义。但是该研究同样存在某些不足之处，例如在进行患者的五年生存率预测是，我们只是考虑了患者的组学特征，并没有考虑临床特征例如患者的年龄，性别，癌症分期等；此外我们考虑的多组学数据并不全面。在接下来的研究中我们会继续搜集癌症患者五年生存率的相关数据，并且继续探究新的分类方法，不断提高预测的准确率。

参考文献

[1] Zheng, H., Zhang, G., Zhang, L., *et al*. (2020) Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis. *Frontiers in Oncology*, **10**, Article No. 68. https://doi.org/10.3389/fonc.2020.00068

[2] Kourou, K., Exarchos, T.P., Exarchos, K.P., *et al*. (2015) Machine Learning Applications in Cancer Prognosis and Prediction. *Computational and Structural Biotechnology Journal*, **13**, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

[3] Vasaikar, S.V., Straub, P., Wang, J., *et al*. (2018) LinkedOmics: Analyzing Multi-Omics Data within and across 32 Cancer Types. *Nucleic Acids Research*, **46**, D956-D963. https://doi.org/10.1093/nar/gkx1090

[4] Cruz, J.A. and Wishart, D.S. (2006) Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, **2**. https://doi.org/10.1177/117693510600200030

[5] Altınçay, H. (2011) Improving the *κ*-Nearest Neighbour Rule: Using Geometrical Neighbourhoods and Manifold-Based Metrics. *Expert Systems*, **28**, 391-406. https://doi.org/10.1111/j.1468-0394.2010.00549.x

[6] Schonlau, M. and Welch, W.J. (2006) Screening the Input Variables to a Computer Model via Analysis of Variance and Visualization. In: Dean, A. and Lewis, S., Eds., *Screening*: *Methods for Experimentation in Industry*, *Drug Discovery*, *and Genetics*, Springer, Berlin, 308-327. https://doi.org/10.1007/0-387-28014-6_14

[7] Hasan, M.A.M., Nasser, M., Ahmad, S., *et al*. (2016) Feature Selection for Intrusion Detection Using Random Forest. *Journal of Information Security*, **7**, 129-140. https://doi.org/10.4236/jis.2016.73009

[8] Martinez, A.M. and Kak, A.C. (2001) Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 228-233. https://doi.org/10.1109/34.908974

[9] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[10] Menard, S. (2002) Applied Logistic Regression Analysis. Sage, London. https://doi.org/10.4135/9781412983433

[11] Pal, M. (2005) Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing*, **26**, 217-222. https://doi.org/10.1080/01431160412331269698

[12] Safavian, S.R. and Landgrebe, D. (1991) A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, **21**, 660-674. https://doi.org/10.1109/21.97458

[13] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[14] Aziz, R.M., Baluch, M.F., Patel, S., *et al*. (2022) LGBM: A Machine Learning Approach for Ethereum Fraud Detection. *International Journal of Information Technology*, **14**, 3321-3331. https://doi.org/10.1007/s41870-022-00864-6

[15] Cunningham, P. and Delany, S.J. (2021) *κ*-Nearest Neighbour Classifiers—A Tutorial. *ACM Computing Surveys* (*CSUR*), **54**, 1-25. https://doi.org/10.1145/3459665

[16] Hastie, T., Rosset, S., Zhu, J., *et al*. (2009) Multi-Class Adaboost. *Statistics and Its Interface*, **2**, 349-360. https://doi.org/10.4310/SII.2009.v2.n3.a8

[17] Joachims, T. (1998) Making Large-Scale SVM Learning Practical. Technical Report.

[18] Murphy, K.P. (2006) Naive Bayes Classifiers. University of British Columbia.

[19] Ibrahim, W.H.A. (2020) Performance Evaluation of Classification Algorithms.