

doi:10.3969/j.issn.2095-1035.2024.01.012

基于同步辐射 X 射线荧光光谱与一维卷积神经网络的癌症筛查方法

魏超杰¹ 李超² 解宏鑫³ 王欣⁴ 李玉锋³ 李玉文¹ 刘杨¹ 王伟^{1*}

(1. 中国农业大学 工学院, 暨全国金属组学创新研究中心, 北京 100083;

2. 安徽医科大学 第二附属医院, 暨全国金属组学创新研究中心, 合肥 230032;

3. 中国科学院高能物理研究所, 中国科学院-香港大学金属组学与健康和环境联合实验室, 中国科学院纳米生物效应与安全性重点实验室, 暨全国金属组学创新研究中心, 北京 100049;

4. 安徽医科大学 基础医学院, 暨全国金属组学创新研究中心, 合肥 230032)

摘要 癌症是全球范围内引起高发病率与高死亡率的疾病之一。现有癌症检测方法耗时、昂贵、专业人员依赖性强, 开发一种无损、快速筛查方法非常重要。在前期工作基础上, 发展了基于同步辐射 X 射线荧光光谱技术(SRXRF)与深度学习技术结合的一种非靶标金属组学方法筛查癌症患者。首先, 分析控制组与癌症组共 269 份血清样本的 SRXRF 谱线, 得到 Ca、Mn、Zn、Ge、Br 在两类人群中具有代表性差异, 可以作为癌症筛查的标志物; 其次, 对于平均光谱进行归一化(Normalization)、迭代自适应加权惩罚最小二乘法(airPLS)、Savitzky-Golay 平滑(SG)、标准正态变换(SNV)的预处理, 并建立偏最小二乘判别分析(PLSDA)、K 近邻法(KNN)、软独立建模分类法(SIMCA)的化学计量学模型, 三种模型对癌症筛查的最优准确率分别为 89.89%、93.26%、90.95%; 最后, 基于像素级光谱, 搭建三种一维卷积神经网络(1DCNN)模型, 三种模型准确率分别为 93.56%、95.24%、93.27%, 相对于化学计量学模型均有所提高, 增加卷积层的数量有助于数据特征提取, 模型准确率提高了 1.68%。将三种模型卷积层提取获得的特征进行 t -分布随机邻域嵌入算法(tSNE)降维可视化, 得到 1DCNN 提取的特征具有显著可分性, SRXRF 结合 1DCNN 模型开发的非靶标金属组学方法在实现癌症的快速筛查方面具有潜力。

关键词 癌症筛查; 血清; X 射线荧光光谱; 一维卷积神经网络; 非靶标金属组学

中图分类号: O657.34 文献标志码: A 文章编号: 2095-1035(2024)01-0104-08

Research on Cancer Screening Method Based on Synchrotron Radiation X-ray Fluorescence Spectroscopy and One-dimensional Convolutional Neural Network

WEI Chaojie¹, LI Chao², XIE Hongxin³, WANG Xin⁴, LI Yu-Feng³,
LI Yuwen¹, LIU Yang¹, WANG Wei^{1*}

收稿日期: 2023-10-10 修回日期: 2023-10-30

基金项目: 国家自然科学基金面上项目(32272410)

作者简介: 魏超杰, 男, 博士研究生, 主要从事信息与图像处理技术研究。E-mail: chaojie0721@163.com

* 通信作者: 王伟, 男, 教授, 主要从事信息与图像处理技术研究。E-mail: playerwxw@cau.edu.cn

引用格式: 魏超杰, 李超, 解宏鑫, 等. 基于同步辐射 X 射线荧光光谱与一维卷积神经网络的癌症筛查方法[J]. 中国无机分析化学, 2024, 14(1): 104-111.

WEI Chaojie, LI Chao, XIE Hongxin, et al. Research on Cancer Screening Method Based on Synchrotron Radiation X-ray Fluorescence Spectroscopy and One-dimensional Convolutional Neural Network[J]. Chinese Journal of Inorganic Analytical Chemistry, 2024, 14(1): 104-111.

(1. College of Engineering & National Consortium for Excellence in Metallomics,
China Agricultural University, Beijing 100083, China;

2. The Second Affiliated Hospital & National Consortium for Excellence in Metallomics,
Anhui Medical University, Hefei, Anhui 230032, China;

3. CAS-HKU Joint Laboratory of Metallomics on Health and Environment & CAS Key Laboratory for Biomedical Effects of
Nanomaterials and Nanosafety & Beijing Metallomics Facility & National Consortium for Excellence in Metallomics, Institute of
High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China;

4. School of Basic Medical Sciences, & National Consortium for Excellence in Metallomics,
Anhui Medical University, Hefei, Anhui 230032, China)

Abstract Cancer is one of the diseases that cause high incidence rate and mortality worldwide. Existing cancer detection methods are time-consuming, expensive, and highly dependent on professionals, making it crucial to develop a non-destructive and rapid screening method. On the basis of previous work, this article developed a non target metabolomics method for screening cancer patients based on the combination of Synchrotron Radiation X-ray Fluorescence Spectroscopy (SRXRF) and deep learning technology. Firstly, by analyzing the SRXRF spectra of 269 serum samples from the control group and the cancer group, it was found that Ca, Mn, Zn, Ge, and Br had representative differences between the two populations and could be used as biomarkers for cancer screening. Secondly, normalization, adaptive iteratively reweighted penalized least squares (airPLS), Savitzky-Golay smoothing (SG), and Standard Normal Variate (SNV) were performed on the average spectrum, and chemometric models of Partial Least Squares Linear Discriminant Analysis (PLSDA), K-Nearest Neighbor (KNN), and Soft Independent Modeling of Class Analogy (SIMCA) were established. The optimal accuracy rates of the three models for cancer screening were 89.89%, 93.26%, and 90.95%, respectively. Finally, based on pixel level spectra, three one-dimensional Convolutional Neural Network (1DCNN) models were constructed, with accuracy rates of 93.56%, 95.24% and 93.27%, respectively. Compared with chemometric models, all models improved significantly. Increasing the number of convolutional layers was conducive to extract data features, and the model accuracy improved by 1.68%. The features extracted from the convolutional layers of three models were visualized using *t*-distributed Stochastic Neighbor Embedding (tSNE) for dimensionality reduction, and it was found that the features extracted by 1DCNN had significant separability. In summary, the nontargeted metallomics method developed by SRXRF combined with 1DCNN model is potential in achieving rapid cancer screening.

Keywords cancer screening; serum; X-ray fluorescence; convolutional neural network; non-targeted metallomics

癌症是引起全球范围内高发病率和高死亡率的公共卫生疾病之一。根据国际癌症研究机构(International agency for research on cancer, IARC)的报告,人类癌症的发病率和死亡率正在迅速增加^[1]。对于晚期癌症患者,经过有效治疗 5 年存活率小于 50%^[2]。若能尽早发现并治疗,能够在很大程度上降低死亡率。而临床上对于癌症患者筛查需要一系列操作,速度较慢,因此有必要开发一种快速癌症筛查的方法。

目前癌症的临床诊断方法有体液生化指标检测、X 射线影像^[3]、计算机断层成像^[4]、核磁共振影像^[5]、超声影像^[6]、内镜检查^[7]等,这些方法存在主

观依赖性强、价格昂贵、局限性大等不足,迫切需要快速、准确、无损的癌症筛查方法。金属组学^[8]是系统研究生命体内自由或络合的全部金属/类金属元素的含量、分布、化学种态及其功能的一门新兴学科。以健康与癌症人群体内血液的金属组作为靶标物,有助于癌症患者的筛查。

微量元素的浓度是身体疾病的另一种信息来源。众所周知,某些元素在调节和决定生物体健康的多种生化过程中发挥着重要作用。CHEN 等^[9]通过使用原子荧光光谱法(AFS)和电感耦合等离子体原子发射光谱法(ICP-AES)测定健康人与心血管患者血液与尿液中的 9 种元素(铬、铁、锰、铝、镉、

铜、锌、镍和硒),发现正常人血液与尿液内所有的元素浓度均高于心血管疾病组,使用决策树建立分类模型最高可达到 97.6% 的准确率。CARTER 等^[10]以微波诱导等离子体原子发射光谱法(MIP-ONS)与电感耦合等离子体质谱法(ICP-MS)测定脚趾甲 22 种元素并结合其他指标作为特征,采用随机森林建立糖尿病预测模型,最优模型的受试者工作特征曲线下面积达到 0.9。LIN 等^[11]采用 ICP-AES 与 ICP-MS 技术测定血清中 39 种微量元素,结合机器学习算法对精神分裂症的识别准确率能够达到 99.21%。因此人体内元素水平可以反应健康状态,用于疾病的筛查。

使用电感耦合等离子体光学发射光谱(ICP-OES)和 ICP-MS 法可以量化元素的含量,具备精确度高、检出限低等优点。但具有破坏性,同时测量浓度差异较大的多种元素时,需要使用多种技术组合。而 SRXRF 具有预处理少、非破坏、宽分析范围[从轻元素(锂、硼)到重金属元素(铀、钷)]、检测速度快、适用样品广等优点。SRXRF 技术通过校正仪器响应和建立标准曲线来预测元素的含量,但定量分析受到多种因素影响。因此本文尝试使用 SRXRF 作为研究手段进行癌症患者的筛查。

基于 SRXRF 的平均光谱偏最小二乘判别分析(PLSDA)与支持向量机(SVM)的机器学习算法,预处理方法单一,样本数量有限^[12]。为了从大数据集中挖掘数据特征,此次研究增加了血清样本数量,以同一个样本不同位置的光谱作为输入开发卷积神经网络,以适应大数据发展时代下的癌症筛查方法研究。

受生物视觉感知机制启发的卷积神经网络(Convolutional Neural Network, CNN)被证明可以从原始数据中提取有效特征进行分类或者回归的有效手段^[13]。二维卷积神经网络在图像处理^[14]和目标检测^[15]方面取得一定的研究成果。但一维卷积神经网络(One dimensional-convolutional neural network, 1DCNN)可以处理序列数据,且具有较少的参数量,较小的过拟合风险,成功用于近红外光谱^[16]、拉曼光谱^[17]和其他一维信号^[18]等研究。因此,构建 1DCNN 模型,从 SRXRF 数据中提取深层特征,建立癌症患者的筛查模型具有可行性。

在本研究中,将 SRXRF 和 1DCNN 结合展示非靶标金属组学在血清样本中筛选癌症患者中的应用。通过分析光谱差异,研究控制组与癌症组的标志物。建立预处理与化学计量学和 1DCNN 的癌症分类模型。并以特异性,灵敏性验证模型性能。本

研究有助于推动 SRXRF 技术在癌症患者筛查方法的应用研究。

1 实验方法

1.1 样品准备

本研究的血清样本采集实验在安徽医科大学第二附属医院进行,于上午 6:00—8:00 采集志愿者空腹状态下的静脉血液,保存于离心管内,使用 3 000 r/min 离心机离心 10 min,取上层血清。无各类疾病的健康志愿者与病理证实的肺癌与胃癌志愿者提供血清样本,两者在年龄、性别上无明显差异。健康志愿者的血清样本称为控制组,癌症志愿者为癌症组,共收集 269 份血清样本,包含 161 份控制组与 108 份癌症组样本。所有志愿者均签署了知情同意书,本工作获得了安徽医科大学伦理委员会同意。

1.2 数据获取

在北京同步辐射装置 4W1B 束线上进行 SRXRF 数据采集,设备运行的能量为 2.5 GeV,束流强度为 150~250 mA^[19]。将 60 μ L 血清滴在滤纸上,风干后固定在移动平移台上,采用 Si(Li)固态探测器检测 X 射线荧光强度。在每个血清样本的不同位置进行 22 次连续点扫描,共收集 5 918(269 $\times$ 22)条荧光光谱。计算每个样品的 22 条光谱的平均值称为平均光谱,共 269 条用于化学计量学的模型建立,5 918 条荧光光谱称为像素级光谱,用于 1DCNN 模型建立。

1.3 光谱预处理

采用归一化处理、Savitzky-Golay 平滑(SG)、迭代自适应加权惩罚最小二乘法(Adaptive iteratively reweighted penalized least squares, airPLS)、标准正态变化(Standard normal variate, SNV)作为光谱预处理手段。归一化处理可以消除志愿者体内微量元素量级差异对 SRXRF 的影响。采用 15 点 2 次线性 SG 平滑对曲线进行去噪,减少仪器与环境对光谱曲线的影响。目前效果最好的基线校正方法是 airPLS,采用基于误差的迭代加权策略,基于上一次循环拟合的基线和原始信号之间的差异来纠正谱线。SNV 能够消除光程差异、散射和样品稀释等引起的误差,使数据具有可比性。

1.4 化学计量学算法

采用偏最小二乘判别(Partial least squares linear discriminant analysis, PLSDA)、K 近邻算法(K-nearest neighbor, KNN)、簇类独立软模式分类(Soft independent modeling of class analogy, SIMCA)对

癌症组进行筛选。PLSDA 是一种常用的多变量校准方法,它通过将光谱数据投影到一个新的空间中搜索一系列与响应变量高度相关的潜在变量,本研究采用五折交叉验证法优化潜变量的数量^[20]。KNN 是一种非常经典的分类算法,未知样本的类别由其最接近的 K 个临近值来代表^[21]。采用欧式距离作为度量,采用五折交叉验证选择最优邻域值 K。SIMCA 是以主成分分析为基础的定性分析方法^[22],通过将数据投影到最优主成分空间中实现

分析和分类的目的。

1.5 深度学习算法

CNN 是在前馈神经网络的基础上通过增加卷积操作而发展起来的。SRXRF 能够反应血清中微量元素的信息,因此可以将连续的 SRXRF 数据作为一维方向的像素点,使 CNN 应用于 SRXRF 分析具有可行性。本文参考 LeNet 模型,将二维卷积神经网络变形,以连续通道的 SRXRF 光谱作为输入,共提出三种 1DCNN 框架(图 1)。

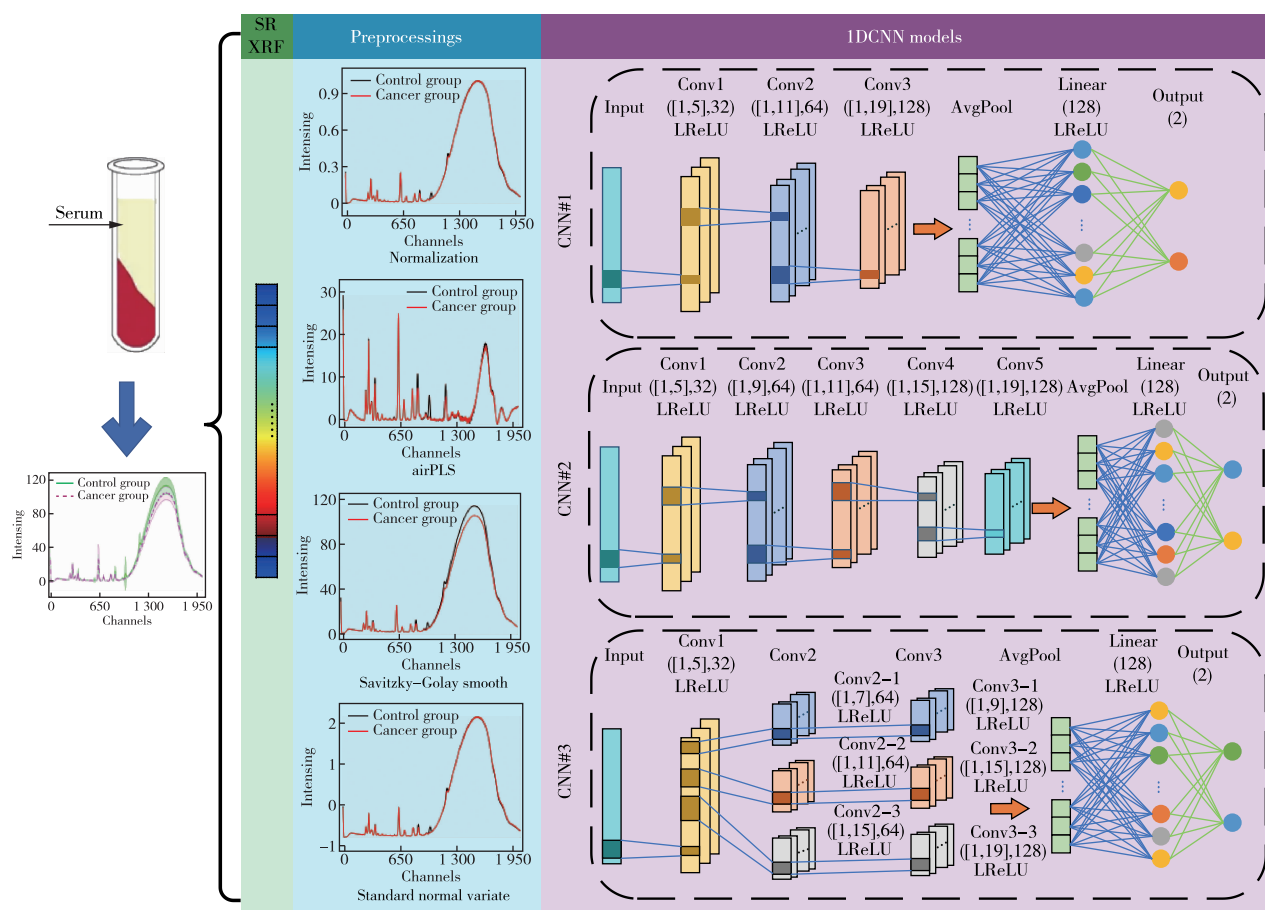


图 1 一维卷积神经网络模型

Figure 1 A one-dimensional convolutional neural network model.

第一个 1DCNN 模型是由三个卷积层和一个全连接层组成的浅 CNN 模型(CNN #1)。三层卷积的卷积尺度大小分别为 1×5 、 1×11 、 1×19 ,卷积核为 32、64、128,步幅为 2、5、9。第二个模型是由五个卷积层和一个全连接层组成深 CNN 模型(CNN #2)。CNN #2 与 CNN #1 第一层卷积相同,第二到五层卷积的卷积尺度大小分别为 1×7 、 1×9 、 1×11 、 1×13 ,卷积核为 64、64、128、128,步幅为 3、3、5、7。第三个是宽 CNN 模型(CNN #3),CNN #3 与前两个模型的第一层卷积相同,而 Conv3、Conv2 与

Conv1 串联,Conv2 与 Conv3 内各含有三个卷积层,并保持并行处理数据,其中 Conv2-1 与 Conv3-1 串联,并保持小的卷积尺寸,Conv2-2 和 Conv3-2 保持中等卷积尺寸,Conv2-3 与 Conv3-3 保持最大卷积尺寸,具体参数如图 1 所示。三种模型的卷积层后采用 BatchNorm 进行归一化和 Leaky ReLU 函数进行激活,激活函数的表达式如式(1)所示:

$$f(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (1)$$

α 称为 Leaky 系数,在此处取 0.1。特征提取后

采用自适应最大池化和全连接层,输出有两个节点,分别对应控制组与癌症组。

模型的训练过程,将 SRXRF 光谱(1×2025)输入到 3 个 1D-CNN 模型中,将标签 0、1 转换为 One-hot 形式,三种模型的输出层为 2 个神经元,使用“Sigmoid”激活函数判断分类概率。设置 Batch size 的大小为 512,初始学习率为 1×10^{-3} ,迭代次数 Epoch 为 600,模型损失函数为 BCEloss,优化器为 Adam(Adaptive moment estimation)优化器,为了训练过程中更好的收敛,Adam 动量项权重衰减系数 β_1 为 0.9,学习率衰减系数为 0.999。采用 t -分布随机邻域嵌入算法(t -distributed stochastic neighbor embedding, tSNE)进行卷积神经网络模型的特征可视化。

1.6 模型评价

在建立模型前,样本被随机分为校正集(2/3)和预测集(1/3),通过计算校正集和预测集的正确分类率(Correct classification rate, CCR)来评价模型,CCR 的计算公式如式(2)表示:

$$CCR/\% = \frac{N_1}{N_2} \times 100\% \quad (2)$$

其中, N_1 和 N_2 分别为校正集和预测集正确的数量和总数。进一步,为了评估模型的鲁棒性和可靠性,从混淆矩阵计算得到特异性与灵敏度,分别使用式(3)或(4)来评估每种类型样本的分类精度。

$$\text{特异性}/\% = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$\text{灵敏度}/\% = \frac{TN}{TN+FP} \times 100\% \quad (4)$$

式中 TP 、 TN 、 FN 、 FP 分别代表真阳性(True positive)、真阴性(True negative)、假阴性(False negative)、假阳性(False positive)。

在 Windows 系统上基于 Matlab 和 Classification toolbox 5.4 搭建化学计量学模型。基于 PyCharm 开发环境与 Pytorch 框架搭建 1DCNN 模型,采用 NVIDIA GeForce RTX 1650 4GB GPU 进行模型训练。

2 结果与讨论

2.1 光谱分析

图 2 显示了两类志愿者血清数据在 0~2 025 通道内的 SRXRF 平均光谱及标准差。光谱曲线基本走势与 HE 等^[12]的研究一致。在本实验中,数据采集条件一致,血清样本体积相同,通过对光谱曲线峰面积的拟合,可反应血清中金属元素的差异,其中

371、588、862、992、1 189 通道附近的峰面积拟合为 Ca、Mn、Zn、Ge、Br。呈现出的规律为控制组中对应的通道强度高于癌症组,即控制组体内上述元素含量高于癌症组。CALLEJÓN-LEBLIC 等^[23]的研究提出 V/Mn 与 V/Zn 的比值作为肺癌中金属标志物,HE 等^[13]提出 Ca 和 Zn 作为癌症标志物,PIACENTI 等^[24]的研究指出癌症组织中的 Ca、Cu、Zn 低于正常组织,上述文献与本研究 SRXRF 的 Ca、Mn、Zn 差异情况一致。

Ge 是一种能在外部诱发癌症的一种微量元素,在肿瘤患者体内作为重要标志物^[25]。曾昭华等^[26]的研究表明,土壤中 Br 与食道癌、宫颈癌死亡率呈负相关,与肝癌、鼻咽癌死亡率成正相关。以上元素对于癌症的筛查具有显著意义,对于提取特征通道,简化模型,识别癌症标志物具有重要意义。

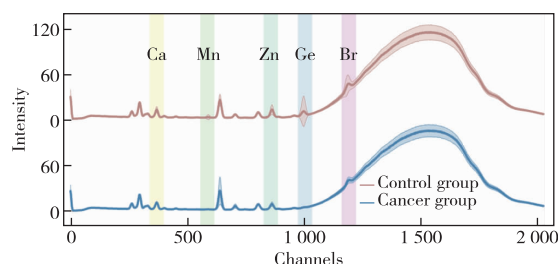


图 2 控制组与癌症组 SRXRF 均值曲线

Figure 2 The averaged SRXRF spectra of serum samples in the control group and cancer group.

2.2 化学计量学模型

PLSDA、KNN 和 SIMCA 分类准确率及最优的关键参数列于表 1,参数 LV_s/K 表示 PLSDA 与 SIMCA 的最优潜变量数与 KNN 的最近邻数。在 PLSDA 模型中,预处理后的准确率相对于原始数据均有一定程度的改善,airPLS 基线校正后的准确率最高,校正集、预测集和交叉验证集分别达到了 90.76%、89.76%、88.67%。在 KNN 模型中,原始数据的预测集准确率达到 92.13%,仅有 SG 平滑预处理改善了准确率,而其他预处理准确率低于原始数据预测结果,KNN 的分类结果整体上均优于 PLSDA 模型。在 SIMCA 模型中,归一化预处理取得最优分类结果,校正集、预测集与交叉验证集分别为 91.93%、90.95%和 90.45%。

根据模型预测结果,预处理能够改善模型准确率,但错误地使用预处理方式会削弱有效信息甚至引入新的噪声。模型的选择比预处理方式的选择对模型的分类影响更大,本次数据中 KNN 的模型准确率优于其他模型。

表 1 基于化学计量学的分类性能
Table 1 Classification performance based on chemometrics

Model methods	Pre-processings	LVs/K	Calibration set accuracy/%	Prediction set accuracy/%	Cross-validation set accuracy/%
PLSDA	None	5	86.11	85.39	84.27
	Normalization	4	88.89	87.64	87.27
	airPLS	6	91.11	89.89	88.67
	SG	8	89.44	88.76	88.24
	SNV	5	90.00	89.76	88.34
KNN	None	4	93.33	92.13	91.34
	Normalization	7	91.67	89.89	90.54
	airPLS	6	92.22	89.89	91.56
	SG	4	94.44	93.26	93.87
	SNV	6	92.22	91.01	90.87
SIMCA	None	6	85.78	85.39	86.34
	Normalization	7	91.11	90.95	90.45
	airPLS	7	87.78	87.64	86.55
	SG	8	88.89	85.39	87.45
	SNV	5	88.89	86.52	87.37

为了进一步验证模型的预测能力,预处理与三种模型的最优组合 airPLS-PLSDA、SG-KNN、Normalization-SIMCA 的预测集的混淆矩阵如表 2 所示。三个模型对于癌症组的灵敏性均大于

91.67%,而特异性均大于 84.91%,对于癌症的精确度最高仅为 89.47%,表明易将控制组样本误识别为癌症组样本。采用一维卷积神经网络进一步提高准确率。

表 2 化学计量学模型的混淆矩阵
Table 2 Confusion matrix of chemometrics models

Actual class	Predict class					
	PLSDA		KNN		SIMCA	
	Control group	Cancer group	Control group	Cancer group	Control group	Cancer group
Control group	46	7	49	4	45	8
Control group	2	34	2	34	3	33
Specificity/%	94.44	86.79	94.44	92.45	91.67	84.91
Sensitivity/%	86.79	94.44	92.45	94.44	84.91	91.67
Accuracy/%	95.83	82.92	96.08	89.47	93.75	80.49
Total accuracy/%	89.89		93.26		90.95	

2.3 卷积神经网络模型

为了增加模型的泛化能力增加鲁棒性,以 5 918 条 SRXRF 谱线作为网络的输入。三种 1DCNN 的预测结果如表 3 所示。在模型中,归一化预处理对于 CNN#1 的准确率有所改善,其他预处理相对于原始数据会造成模型的准确率下降,这跟 ZHANG 等^[27]相关研究中表述的一致,虽然预处理可以改善原始数据,但会导致原始信息的丢失,甚至不恰当预处理,反而会引入新的噪声,造成模型准确率下降。

CNN#2 相对于 CNN#1 的准确率有所提高,是卷积层在提取特征方面发挥作用,CNN#1 卷积层数较少,提取特征有限,并不能充分发挥模型优势,而 CNN#2 增加了卷积层的层数,预测集的准确率提高了 1.68%。但在提高准确率的同时,导致

模型参数增加,训练模型花费的时间更多,CNN#1 迭代一个周期仅需 0.29 s,即仅需 180 s 可完成模型的训练,而 CNN#2 迭代一次需要消耗 0.56 s,时间是 CNN#1 的 1.93 倍。CNN#3 相对 CNN#1 而言具有相似的准确率,虽然具备更宽的卷积尺度,但并没有明显提升,原因可能是 CNN#1 中 Conv2 和 Conv3 具备较宽的卷积尺寸,模型的特征提取能力与 CNN#3 相同,均能够提取原始光谱中绝大部分有效特征。

卷积神经网络的卷积层具备提取特征的能力,将三种模型提取特征后的输出(128 个特征)进行 tSNE 降维处理,保留前三个维度,可视化后的结果如图 3 所示,发现控制组和癌症组绝大部分能够很好地分为两类,但是依然有重叠的部分存在,表示神经网络提取的特征能够有效筛查癌症患者。

表 3 卷积神经网络模型性能

Table 3 Performance of convolutional neural network models

Model methods	Pre-processings	Calibration set accuracy/%	Prediction set accuracy/%	Cross-validation set accuracy/%
CNN # 1	None	93.56	92.78	92.34
	Normalization	94.34	93.56	92.46
	airPLS	93.56	92.34	92.16
	SG smooth	92.78	91.56	92.10
	SNV	93.34	92.25	92.35
CNN # 2	None	96.56	95.24	95.55
	Normalization	95.33	94.12	94.33
	airPLS	94.42	93.78	93.34
	SG smooth	93.28	92.56	92.67
	SNV	93.57	92.35	91.37
CNN # 3	None	94.37	93.27	93.56
	Normalization	93.27	92.01	91.34
	airPLS	94.28	93.17	93.23
	SG smooth	92.18	91.34	91.45
	SNV	92.54	91.34	91.54

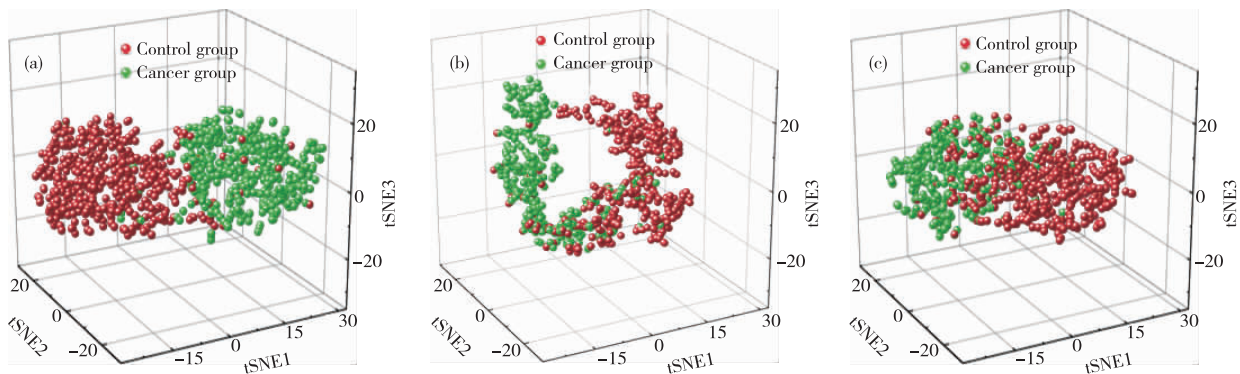


图 3 1DCNN 提取特征 tSNE(a)CNN # 1;(b)CNN # 2;(c)CNN # 3

Figure 3 tSNE diagrams of 1DCNN feature extraction(a)CNN # 1;(b)CNN # 2;(c)CNN # 3.

3 结论

针对控制组与癌症组志愿者的血清样本,采用 SRXRF 对血清样本进行数据采集,获得 5 918 个 SRXRF 数据。对原始谱线分析发现了控制组与癌症组的有效标志元素 Ca、Mn、Zn、Ge、Br。采用预处理与化学计量学结合的方法,建立基于平均光谱的分类模型,模型准确率为 93.6%。采用 1DCNN 进行分类,模型准确率为 95.24%,并可视化提取的特征,提取的特征对筛查癌症患者具有较好的可分性。本研究有利于推动 SRXRF 结合深度学习开发非靶标金属组学方法,实现癌症患者的快速筛查。

参考文献

[1] SUNG H, FERLAY J, SIEGEL R L, et al. Global cancer statistics 2020; GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians,

2021,71(3):209-249.

[2] SITARZ R, SKIERUCHA M, MIELKO J, et al. Gastric cancer:epidemiology,prevention,classification, and treatment[J]. Cancer Management and Research, 2018,10:239-248.

[3] LU L,SUN M Z, LU Q Y, et al. High energy X-ray radiation sensitive scintillating materials for medical imaging, cancer diagnosis and therapy [J]. Nano Energy, 2021, 79: 105437. DOI: 10. 1016/j. nanoen. 2020.105437.

[4] SONAVANE S K,PINSKY P,WATTS J J, et al. The relationship of cancer characteristics and patient outcome with time to lung cancer diagnosis after an abnormal screening CT[J]. European Radiology,2017, 27(12):5113-5118.

[5] TIROTTA I, DICHiarANTE V, PIGLIACELLI C, et al.¹⁹ F magnetic resonance imaging (MRI): from design of materials to clinical applications[J]. Chemical Reviews,2015,115(2):1106-1129.

- [6] VAN SLOUN R J G, SOLOMON O, BRUCE M, et al. Super-resolution ultrasound localization microscopy through deep learning [J]. *IEEE Transactions on Medical Imaging*, 2021, 40(3): 829-839.
- [7] KURUMI H, YAMAGUCHI N, ISOMOTO H. Appropriate interval of surveillance endoscopy for early detection of gastric cancer [J]. *Digestive Endoscopy*, 2023, 35(5): 603-605.
- [8] BENETTI F, BREGOLI L, OLIVATO I, et al. Effects of metal (loid)-based nanomaterials on essential element homeostasis: the central role of nanometallomics for nanotoxicology [J]. *Metallomics*, 2014, 6(4): 729-747.
- [9] CHEN H, TAN C, LIN Z, et al. A feasibility study of diagnosing cardiovascular diseases based on blood/urine element analysis and consensus models [J]. *Computers in Biology and Medicine*, 2013, 43(7): 865-869.
- [10] CARTER J A, LONG C S, SMITH B P, et al. Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes [J]. *Expert Systems with Applications*, 2019, 115: 245-255.
- [11] LIN T, LIU T B, LIN Y C, et al. Comparative study on serum levels of macro and trace elements in schizophrenia based on supervised learning methods [J]. *Journal of Trace Elements in Medicine and Biology*, 2017, 43: 202-208.
- [12] HE L N, LU Y, LI C, et al. Non-targeted metallomics through synchrotron radiation X-ray fluorescence with machine learning for cancer screening using blood samples [J]. *Talanta*, 2022, 245: 123486. DOI: 10. 1016/j. talanta. 2022. 123486.
- [13] KIRANYAZ S, AVCI O, ABDELJABER O, et al. 1D convolutional neural networks and applications: a survey [J]. *Mechanical Systems and Signal Processing*, 2021, 151: 107398. DOI: 10. 1016/j. ymssp. 2020. 107398.
- [14] KHAN A, SOHAIL A, ZAHOORA U, et al. A survey of the recent architectures of deep convolutional neural networks [J]. *Artificial Intelligence Review*, 2020, 53(8): 5455-5516.
- [15] QU L Q, HE S F, ZHANG J W, et al. RGBD salient object detection via deep fusion [J]. *IEEE Transactions on Image Processing*, 2017, 26(5): 2274-2285.
- [16] GAO J Y, ZHAO L G, LI J, et al. Aflatoxin rapid detection based on hyperspectral with 1D-convolution neural network in the pixel level [J]. *Food Chemistry*, 2021, 360: 129968. DOI: 10. 1016/j. foodchem. 2021. 129968.
- [17] HAMED MOZAFFARI M, TAY L. Overfitting one-dimensional convolutional neural networks for Raman spectra identification [J]. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2022, 272: 120961. DOI: 10. 1016/j. saa. 2022. 120961.
- [18] PARK H I, CHO T J, CHOI I G, et al. Object classification system using temperature variation of smart finger device via machine learning [J]. *Sensors and Actuators A-Physical*, 2023, 356. DOI: 10. 1016/j. sna. 2023. 114338.
- [19] XIE H X, TIAN X, HE L A, et al. Spatial metallomics reveals preferable accumulation of methylated selenium in a single seed of the hyperaccumulator cardamine violifolia [J]. *Journal of Agricultural and Food Chemistry*, 2023, 71(5): 2658-2665.
- [20] BALLABIO D, CONSONNI V. Classification tools in chemistry. part 1: linear models. PLS-DA [J]. *Analytical Methods*, 2013, 5(16): 3790-3798.
- [21] COVER T M. Nearest neighbor pattern classification [J]. *IEEE Transactions on Information Theory*, 1967, 13(1): 21-27.
- [22] VITALE R, COCCHI M, BIANCOLILLO A, et al. Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial [J]. *Analytica Chimica Acta*, 2023, 1270: 341304. DOI: 10. 1016/j. aca. 2023. 341304.
- [23] CALLEJÓN-LEBLIC B, GÓMEZ-ARIZA J L, PEREIRA-VEGA A, et al. Metal dyshomeostasis based biomarkers of lung cancer using human biofluids [J]. *METALLOMICS*, 2018, 10(10): 1444-1451.
- [24] PIACENTI DA SILVA M, ZUCCHI O L A D, RIBEIRO-SILVA A, et al. Discriminant analysis of trace elements in normal, benign and malignant breast tissues measured by total reflection X-ray fluorescence [J]. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2009, 64(6): 587-592.
- [25] MENCHIKOV L G, POPOV A V. Physiological activity of trace element germanium including anticancer properties [J]. *Biomedicines*, 2023, 11(6). DOI: 10. 3390/biomedicines11061535.
- [26] 曾昭华, 曾雪萍. 中国癌症与土壤环境中溴元素的关系 [J]. *四川地质学报*, 2003(3): 170-173.
ZENG Zhaohua, ZENG Xueping. The relationship between cancer and bromine in soil environment in China [J]. *Acta Geologica Sichuan*, 2003(3): 170-173.
- [27] ZHANG X L, LIN T, XU J F, et al. Deep spectra: an end-to-end deep learning approach for quantitative spectral analysis [J]. *Analytica Chimica Acta*, 2019, 1058: 48-57.