doi: 10. 3969 / j. issn. 1674 - 1242. 2024. 01. 008

呼气信号三分类癌症检测模型的设计及评价

岳静文[1]，郝丽俊[1,2]
（1. 上海健康医学院医疗器械学院，上海　201318;
2. 上海理工大学健康科学与工程学院，上海　200093）

【摘要】目的　电子鼻呼气检测技术是一种极具潜力的无创癌症检测技术。然而，现有的检测模型研究大多围绕将某一类疾病患者与健康人的区分展开。为了通过电子鼻呼气检测实现对多种癌症的鉴别，该文基于 KNN 构建 了一种三分类癌症检测模型。**方法**　首先通过电子鼻采集 210 名志愿者的呼气号样本，其中 160 名肺癌或肝癌患者均为医院的确诊患者，50 名健康对照者则为医院的职工或学生。对呼气样本进行预处理得到大小为 210 × 180 的原始特征数据集。然后通过卡方检验完成数据特征初筛，并利用 LDA 优化方法得到训练特征集。接着利用 *K* 值选择学习曲线，训练并得到最优 KNN 三分类癌症检测模型。最后对模型进行多维度评价。**结果**　优化后的 KNN 三分类癌症检测模型可有效区分健康人、肺癌患者和肝癌患者，性能优于其他模型，平均准确度可达到 92.5%。可见，机器学习算法可助力电子鼻呼气检测在癌症检测中的推广应用。

【关键词】呼吸检测；三分类；特征优化；学习曲线；多维度评价
【中图分类号】TP212.3　　**【文献标志码】**A　　**文章编号：**1674-1242（2024）01-0048-06

Design and Evaluation of a Triple Classification Cancer Detection Model for Breath Signals

YUE Jingwen[1], HAO Lijun [1,2]
(1. College of Medical Instruments, Shanghai University of Medicine & Health Sciences, Shanghai 201318, China;
2. School of Health Science and Engneering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】Objective　Electronic nasal breath testing is a highly promising noninvasive cancer detection technique. However, most of the existing studies on detection modeling has are centered on the differentiation of a certain type of disease from healthy individuals. In order to realize the differentiation of multiple cancers through breath testing with an electronic nose, this paper constructs a three-classification cancer detection model based on KNN. **Methods**　Firstly, breath signal samples were first collected from 210 volunteers through an electronic nose, of which 160 patients with lung or liver cancer were diagnosed patients in the hospital and 50 healthy controls were hospital staff or students. The breath samples were preprocessed to obtain a dataset of size 210 × 180. Then, the initial screening of data features is completed by chi-square test and the training feature set is obtained by using LDA optimization method.

收稿日期：2023-10-15。
基金项目：高水平地方高校建设创新人才培养（A1-2601-23-309001）；上海市分子影像学重点实验室建设项目（18DZ2260400）；国家重点研发计划（2020YFA0909000）。
作者简介：岳静文（2001—），女，硕士研究生，从事生物医学工程研究。
通信作者：郝丽俊，女，讲师，电话（Tel.）：13482703742，邮箱（E-mail）：sunnyshm@163.com。

Then, the optimal KNN triple classification cancer detection model is trained and obtained by using the *K*- value to select the learning curves. Finally, the model is evaluated in multiple dimensions. **Results** The optimally designed KNN triple classification cancer detection model can effectively differentiate between healthy people, lung cancer patients and liver cancer patients, with better performance than other models, and the accuracy can reach about 92.5%. It can be seen that machine learning algorithms can help promote the application of electronic nasal breath testing in cancer detection.

【Key words】 Breath Detection; Triple Classification; Feature Optimization; Learning Curve; Multidimensional Evaluation

0　引言

根据世界卫生组织国际癌症研究机构（International Agency for Research on Cancer，IARC）2020 年的研究数据，肺癌和肝癌是死亡率分别排名第一和第三的两种癌症。2020 年全球肺癌患者新增 220 万人，超过发病率排名第三的结直肠癌患者人数（193 万人）近 30 万人 [1]。在中国，肺癌和肝癌是发病率排在前五，死亡率分别排在第一和第二的两大恶性肿瘤。2020 年我国肝癌的新发病例占据全球肝癌新发病例的 45.3%[2]。肺癌和肝癌有一个共同的特征，就是早期症状不明显，容易被忽略。大多数时候，肺癌或肝癌在确诊时已是中晚期，失去了手术切除根治的机会，导致患者的生存率不高，预后也不尽如人意。肺癌和肝癌的不容易诊断、高发病率和高死亡率，决定了其早期诊断和治疗的重要性。

目前癌症的诊断主要通过细胞学、组织学、影像学和血清学肿瘤标志物等一系列检测技术进行 [3]。这些检测技术较多地依赖医院的专业仪器，检测方法复杂且成本高，甚至会对人体造成一定的伤害，且多数情况下价格高昂，需要配备熟练的技术人员。另外，这些检测技术更多地依赖患者的症状表现，而这些症状通常在癌症晚期才出现，所以无法很好地实现早期肺癌或肝癌的检测。

电子鼻呼气检测是一种新型的模拟嗅觉系统的呼气检测技术，可通过分析传感器对不同呼出气体的响应变化，即呼气信号的变化，来判断机体的生理和病理情况 [4]。相比其他传统的检测技术，该技术具有无创、安全、采样速度快、更少的医学介入和样本容易采集等优点，更适合应用于日常体检。Nakamoto 等 [5] 于 2002 年采用电子鼻收集分析了 25 例肝癌患者、10 例肝硬化患者和 20 名健康人的呼气信号变化，证实了肝癌患者的呼气与肝硬化患者和健康人群的呼气有显著差异。秦涛等 [6] 于 2009 年尝试构建了肝癌的呼气诊断模型，结果显示该模型对肝癌诊断的灵敏性和特异性可分别达到 83.3% 与 91.7%。但该研究采用的是固相微萃取和气相色谱质谱联用技术，硬件平台价格昂贵且不易操作，不适用于肝癌的普及型筛查。Molik-Oeser 等 [7] 于 2011 年应用电子鼻 Cyranose320 采集了 66 例新确诊的非小细胞肺癌患者和 73 名慢性阻塞性肺疾病患者的呼气信号，其检测方法的灵敏性和特异性可分别达到 84% 和 89%。浙江大学的王平团队 [8] 于 2015 年应用主成分分析（Principal Components Analysis，PCA）和支持向量机（Support Vector Machine，SVM）构建电子鼻肺癌检测模型，取得了较好的结果，为国内电子鼻在医疗检测领域的研究奠定了一定的基础。Shlomi 等 [9] 于 2017 年证实了利用电子鼻可以将肺癌与良性肺结节区分开来，特异性和准确率可分别达到 93% 和 87%，但灵敏性过低，仅有 75%。Amer 等 [10] 于 2021 年通过 PCA 和线性判别分析法（Linear Discriminant Analysis，LDA）对电子鼻采集到的呼气信号进行定量分析，并使用受试者操作特性曲线（Receiver Operating Characteristic，ROC）评估诊断性能，证实了电子鼻可有效区分肝硬化患者、肝癌患者及健康对照组。然而，需要注意的是，这是一项小规模的初步研究，结果有待进一步的大规模研究来佐证。此外，以上这些研究的目的在于利用呼气信号区分同一部位的不同疾病，且样本数量较少，需要进一步验证，对于基于呼气信号是否可以有效区分不同的癌症并未进行探讨。

为了证实电子鼻呼气检测的有效性及在多种癌症筛查中的可行性，还需要进行进一步研究。本文将利用电子鼻采集更多的呼气信号样本，尝试基于不同的特征选择和优化算法，从采样点中筛选特征点，利用 K 近邻（K-Nearest Neighbor，KNN）分类算法构建一种可用于区分健康人、肺癌患者和肝癌患者的三分类癌症检测模型。

1　实验方法及材料

1.1　数据的采集

在本研究中，应用商用电子鼻 ILD3000 先后采集 210 名志愿者的呼气信号样本，其中 160 名肺癌或肝癌患者均为医院的确诊患者，50 名健康对照者则为医院的职工或学生。在采集呼气信号前，要求受试者在 3 小时内不能进食，饮水除外。

数据采集的基本流程为：系统预热后，设备先经过两次清洗，进行第一次呼气信号采集，再经过一次清洗，进行第二次呼气信号采集 [11]。

表 1 为整理后的实验样本。3 类人群（健康人、肺癌患者和肝癌患者）的标签分别记作 0、1 和 2。

表 1　实验样本

Tab.1　experimental samples

基本信息	基本类型		
	健康人	肺癌患者	肝癌患者
样本数 / 例	50	91	69
标签	0	1	2

最终，获得一个 $210\times60\times3$ 的数据集。其中，210 表示样本数；60 表示一个样本包括的采样点数；3 表示 3 个传感器维度。进一步将同一样本对应的 3 组传感器的采样点进行组合，便可得到一个大小为 210×180 的二维数组，其中 180 为采样点。该数据集为原始数据集。

1.2　基于卡方检验的特征初筛

卡方检验是一种常用的假设检验方法，它主要用于比较两个或两个以上样本的比率，并分析两个分类变量之间的相关性。卡方值越大，两个分类变量之间的偏离程度越高。当将其用于特征初筛时，可根据式（1）计算特征 f 与类别 b 之间的卡方值。

$$x^2(f,b)=\frac{n\times(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

式中，n 表示数据总数；A 表示属于 b 类且含有特征 f 的数据；B 表示不属于 b 类但含有特征 f 的数据；C 表示属于 b 类但不包含特征 f 的数据；D 表示不属于 b 类也不包含特征 f 的数据 [12]。

Kang 等 [13] 将卡方检验特征选择用于网络异常检测，使分类精度提高了 3%。Thaseen 等 [14] 利用卡方检验特征选择来构建入侵检测模型，使分类准确率有所提高。在本研究中，以基于 KNN 分类器的十折交叉验证的平均准确率为依据，通过卡方检验对原始数据集进行特征初筛，将每个样本的特征数据初步减少至 170 个。

1.3　基于 LDA 的特征优化

LDA 作为一种强大的监督算法，其主要思想是通过考虑类内和类间的散点计算线性判别式 [15]。在本研究中，经过特征初筛，数据的维度依然比较大，为此利用 LDA 进一步优化特征以降低特征维度。

具体步骤如下。

（1）对数据尤其是以不同尺度测量的数据进行归一化预处理。

（2）分别利用式（2）和式（3）计算类内和类间散点矩阵。

$$S_w=\frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{L}(x_i^l-\mu_l)(x_i^l-\mu_l)^{\mathrm{T}} \quad (2)$$

$$S_b=\frac{1}{N}\sum_{l=1}^{L}N_l(\mu_l-\mu)(\mu_l-\mu)^{\mathrm{T}} \quad (3)$$

式中，l 是类标签；x_i^l 是类别 l 的某样本特征向量；N_l 是类别 l 的总样本数；μ_l 是平均向量；L、N 和 μ 分别是类别数、所有类的样本数和全局特征向量平均值。

（3）计算并按照递减顺序排序 S_w^{-1} 和 S_b 矩阵的特征向量与特征值。

（4）定义由对应最大特征值的 S_w^{-1} 和 S_b 矩阵的前 k 个特征向量形成的线性判别式的 $N\times K$ 大小的投影矩阵 $w=(w_1,w_2,\cdots,w_k)$。

（5）使用方程将样本 x 的大部分信息重建到新的 k 维子空间 Y 上。

经过 LDA，本研究最终将特征优化至 2 维，大幅减少了模型构建中的运算量。

1.4　基于 KNN 构建三分类癌症检测模型

KNN 是使用最广泛的惰性学习方法之一。

KNN 算法具有高精度、对离群值不敏感和不使用数据输入假设等优点。它的工作原理是给定一组 n 个训练实例，在给定新的预测样本时，KNN 分类器将识别其对应的 K 个最近的相邻训练实例，然后将最多相邻的类标签分配给新的测试样本 [16]。

在本研究中，基于 KNN 算法对健康人、肺癌患者和肝癌患者进行分类的具体步骤如下。

（1）对优化后的特征数据集进行标准化处理，并将处理后的样本划分为测试集和训练集。

（2）基于十折交叉验证和学习曲线，确定最佳的 K 值，得到三分类癌症检测模型。

（3）遍历测试集，计算每个测试样本与所有训练实例之间的欧式距离，并记录所有距离。

（4）对每个测试样本，排序其与所有训练实例的距离，选出最近的 K 个训练实例，并记录其对应的标签（本研究中的标签分别为健康人、肺癌患者和肝癌患者）。进一步统计 K 个训练实例对应最多的标签类别，并将测试样本划分为该类别。

2　实验结果

2.1　模型的影响因素分析

2.1.1　特征优化前后的性能对比

在本研究中，利用卡方检验法和 LDA 相结合的方法完成特征优化。表 2 为特征优化前后的性能对比。为了获得较为客观的结果，在此取十折交叉验证后的平均准确率进行对比。

表 2　特征优化前后的性能对比

Tab.2　Performance comparison before and after feature optimization

	特征优化前	特征优化后
卡方检验法	0.62	0.63
LDA	0.63	0.85

由表 2 可以看到，使用卡方检验法对特征进行选择后，分类效果有少许提升。而使用 LDA 对特征进行进一步优化后，分类器的性能再次提升了 22%。

2.1.2　不同 K 值对模型性能的影响

K 值是 KNN 三分类癌症检测模型中的重要参数。测试样本的类别与 K 值的选择密切相关，改变 K 值，测试样本的类别也会发生改变。图 1 为基于不同 K 值（20 以内）的模型学习曲线。图中，横坐标为不同的 K 值，纵坐标为测试集中 3 类测试样本的准确率。显然，对于本研究设计的三分类癌症检测模型，当 $K \geqslant 7$ 时，其对三类测试样本的鉴别准确率基本稳定在 0.90。考虑到计算量，本研究设置 K=7。

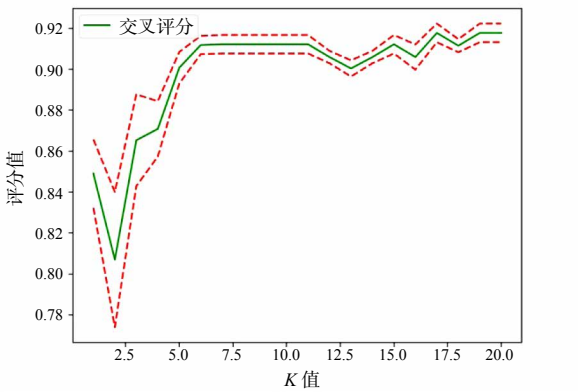

图 1　基于不同 K 值的模型学习曲线

Fig.1　Model learning curves based on different K Values

2.2　模型的多维度评价

在得到最优特征集和最优模型参数后，本研究对模型进行了多维度评价。

2.2.1　混淆矩阵

首先将 120 个训练样本的特征优化集按照 8 ∶ 2 的比例随机划分为训练集和测试集，得到 42 个测试样本，包括 10 个健康人样本、16 个肺癌患者样本和 16 个肝癌患者样本。

在利用训练集构建三分类癌症检测模型后，输入测试集，得到混淆矩阵，图 2 为一次测试的结果。图中，0、1、2 分别代表健康人、肺癌患者和肝癌患者。

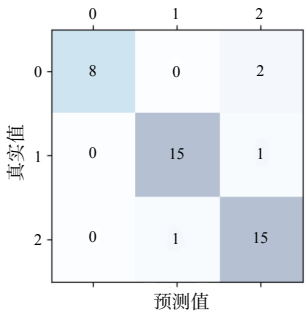

图 2　分类癌症检测模型的混淆矩阵

Fig.2　Confusion of three categorical models

由图 2 可以看到，该模型在鉴别 3 类测试样本时，正确预测值都比较高。对于 10 个健康人样本，

被正确诊断的数量为 8 个，准确率为 0.8；对于 16 个肺癌患者样本，被正确诊断的数量为 15 个，准确率为 0.94；对于 16 个肝癌患者样本，被正确诊断的数量为 15 例，准确率为 0.94。

2.2.2　ROC 曲线

图 3 为三分类癌症检测模型鉴别 3 类测试样本的 ROC 曲线。显然，调参后的 KNN 三分类癌症检测模型对 3 类测试样本呼气信号的区分度都很高，AUC 值均超过 0.95。

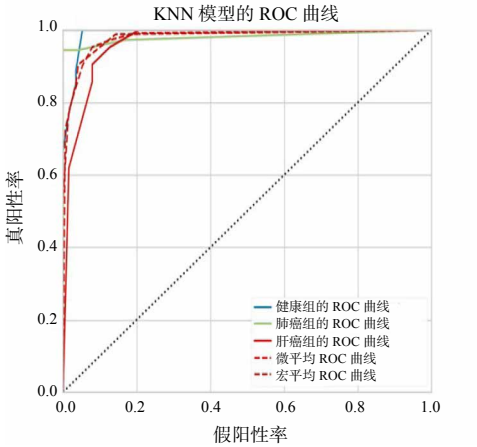

图 3　三分类癌症检测模型的 ROC 曲线

Fig.3　ROC curve of triple cancer detection model

2.2.3　其他性能指标

本研究进一步对三分类癌症检测模型的其他性能指标进行了计算。进行 10 次随机实验后，求得各个性能指标的平均值如表 3 所示。

表 3　三分类癌症检测模型的其他性能指标

Tab.3　Other properties of the triple cancer detection model

类别	平均精准率	召回率	F1 指标
健康人	0.92	0.91	0.91
肺癌患者	0.91	0.92	0.92
肝癌患者	0.90	0.91	0.90

由表 3 可以看出，经过特征选择优化和 KNN 模型调参后，该模型的性能整体表现良好。该模型可以较好地区分 3 类测试样本，平均精准率、召回率和 F1 指标均达到了 0.90 以上。通过对比表中模型对 3 类测试样本的区分性能，发现该模型对肺癌患者的识别能力优于对健康人和肝癌患者的识别能力。

此外，本研究先后利用留一法、随机划分（8 ∶ 2）法和十折交叉验证法对训练特征集进行了多次划分，得到不同的训练集和测试集。变换训练集构建模型，并统计其对不同测试集的预测准确度平均值和方差，完成对模型的泛化性评价。结果显示，在 3 种方法中，改变训练集和测试集，预测的准确度相对比较稳定，10 次测试值的方差不超过 11%，说明该模型的泛化性较好。

2.3　与其他模型的性能对比

本研究进一步将 KNN 模型与 SVM 模型和朴素贝叶斯（Naive Bayes，NB）模型在特征优化前后进行了性能对比。利用十折交叉验证法评价这 3 种模型，得到三者特征优化前后准确率对比的柱状图，如图 4 所示。

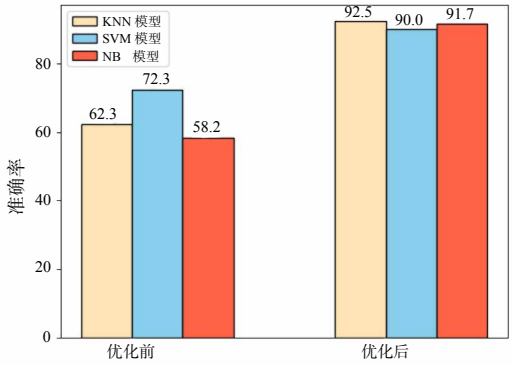

图 4　特征优化前后 3 种模型的准确率对比

Fig.4　Comparison of accuracy of three models before and after feature optimization

由图 5 可以看到，在特征优化前，KNN 模型在鉴别 3 类人群时，其准确率高于 NB 模型，低于 SVM 模型；在特征优化后，KNN 模型的准确率优于其他两种模型。

3　结论

电子鼻呼气检测技术是近年来发展迅猛的一种新型无创癌症检测技术。该技术被越来越多地应用于医学领域，包括肺癌、糖尿病和慢性阻塞性肺疾病等的检测和诊断 [17]。

本研究基于电子鼻采集人体呼气信号，基于 KNN 构建三分类癌症检测模型，探索研究了呼气检测技术对不同癌症鉴别的可能性。实验结果表明，该模型可以高效、准确地完成对健康人、肺癌患者和肝癌患者 3 类人群的鉴别，为未来实现基于呼气的癌症无创筛查技术提供了一定的理论基础。但本研究还存在一些不足之处，在接下来的工作中，我

们计划从以下几个方面开展工作。

（1）KNN 模型性能的进一步提升。理论上，KNN 权值对检测模型的性能有一定影响，但本研究并未对此进行探索。在下一步的工作中，我们将充分考虑 KNN 权值问题，探索一种监督权重分配方法，看是否能够进一步提升模型的检测效果 [18]。

（2）检测模型的进一步探索与改进。例如，可组合应用多种算法构建集成检测模型，从而取长补短，提升检测性能 [19]。

（3）多中心临床数据的采集与模型的外部验证。为了提升并验证模型性能，后续将从多家医院采集更多实验数据进行实验，并采用分层抽样的方式对数据进行划分。

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 71:209-249.

[2] FERLAY J, COLOMBET M, SOERJOMATARAM I, *et al*. Cancer statistics for the year 2020: an overview [J]. **International Journal of Cancer**, 2021, 149(2):778-789.

[3] National Cancer Institute. Cancer trends progress report [R/OL]. (2021-01-07)[2024-03-01]. http://progressreport.cancer.gov./introduction.

[4] 张敏，黄钢 . 基于电子鼻的肺癌无创检测研究 [J]. **现代仪器与医疗**，2021（6）：1-5，13.
ZHANG Min, HUANG Gang. Non-invasive detection of lung cancer based on electronic nose[J]. **Modern Instruments**, 2021(6):1-5,13.

[5] NAKAMOTO T, KANNO T. A preliminary study of diagnosis of hepatoma using an artificial olfactory system [J]. **Cancer Detection and Prevention**, 2002, 26(1):32-36.

[6] 秦涛，刘虎，高署，等 . 肝癌患者呼气中挥发性标志物的筛选与定量分析 [J]. **安徽医科大学学报**，2009，44（1）：4-8.
QIN Tao, LIU Hu, GAO Shu, *et al*. Screening and quantitative analysis of volatile markers in the breath of hepatocellular carcinoma patients[J]. **Acta Universitatis Medicinalis Anhui**, 2009, 44(1): 4-8.

[7] MOLIK-OESER C, TAHANOVICH S, SCHROEDER O, *et al*. Discriminating NSCLC from COPD using patterns derived from an electronic nose [J]. **European Respiratory Journal**, 2011, 38(Suppl 55):2792.

[8] WANG P, CHEN D, CHEN Y, *et al*. A novel electronic nose system based on PCA-SVM for early diagnosis of lung cancer [J]. **Journal of Pharmaceutical and Biomedical Analysis**, 2015 (102):379-385.

[9] SHLOMI D, ABUD M, LIRAN O, *et al*. Detection of lung cancer and EGFR mutation by electronic nose system[J]. **Journal of Thoracic Oncology**, 2017, 12(10): 1544-1551.

[10] AMER R E, DIAD M S, IBRAHIM H M, *et al*. Exhaled breath analysis using electronic nose for liver cirrhosis and hepatocellular carcinoma: a pilot study [J]. **Journal of Breath Research**, 2021, 15(4):046007.

[11] 熊征斯，黄钢，郝丽俊 . 基于机器学习的肝癌无创检测 [J]. **北京生物医学工程**，2020，39（1）：74-79.
XIONG Zhengsi, HUANG Gang, HAO Lijun. Non-invasive detection of liver cancer based on machine learning[J]. **Beijing Biomedical Engineering**, 2020, 39(1):74-79.

[12] 田豆 . 深度强化学习分类预测模型及其在脑卒中发病风险预测应用研究 [D]. 太原：太原理工大学，2021.
TIAN Dou. Deep reinforcement learning classification prediction model and its application to stroke incidence risk prediction research[D]. Taiyuan: Taiyuan University of Technology, 2021.

[13] KANG K. Decision tree techniques with feature reduction for network anomaly detection[J]. **Journal of the Korea Institute of Information Security and Cryptology**, 2019, 29(4):795-805.

[14] THASEEN I S, KUMAR C A, AHMAD A, *et al*. Integrated instruction detection model using Chi-Squqre feature selection and ensemble of classifiers[J]. **Arabian Journal for Science and Engineering**, 2019, 44(4):3357-3368.

[15] NKENGFACK L C, TCHIOTSOP D, ATANGANA R, *et al*. A comparison study of polynomial-based PCA, KPCA, LDA and GDA feature extraction methods for epileptic and eye states EEG signals detection using kernel machines[J]. **Informatics in Medicine Unlocked**, 2021(26):1-16.

[16] JIANG Y, ZHOU Z. Editing training data for KNN classifiers with neural network ensemble[C]//International Symposium on Neural Networks, Dalian, China, 2004.

[17] SMULKO J, CHLUDZINSKI T, MAJCHRZAK T, *et al*. Analysis of exhaled breath for dengue disease detection by low-cost electronic nose system [J]. **Measurement**, 2022, 190:110733.

[18] HOMAEINEZHAD M R, ATYABI S A, TAVAKKOLI E, *et al*. ECG arrhythmia recognition via a neuro-SVM-KNN hybrid classifier with virtual QRS image-based geometrical features[J]. **Expert Systems with Applications**, 2012(39):2047-2058.

[19] SHITOLE S. Respiratory diseases detection using machine learning[J]. **International Journal for Research in Applied Science and Engineering Technology**, 2021:2698-2701.