网络出版时间:2025-03-17 10:17:05 网络出版地址:https://link.cnki.net/urlid/34.1065.R.20250314.1614.016

基于机器学习的肺结核肺炎患者判别分析研究

常敏丽[1],由淑萍[2],陈晓蝶[1],陈志斐[3],郑彦玲[3]
(新疆医科大学[1]公共卫生学院、[2]护理学院、[3]医学工程技术学院,乌鲁木齐 830017)

摘要 **目的** 探讨机器学习方法在肺结核患者判别中的可行性。**方法** 从某三甲医院获取860例患者的15个观测指标数据。通过深入挖掘分析数据,采用支持向量机、随机森林及神经网络方法对患者所患疾病做判别分析。**结果** 基于支持向量机、随机森林和神经网络建立的肺结核可疑患者判别模型准确率分别为90%、91%和88%。**结论** 3种机器学习方法均可用于肺结核可疑患者的判别分析。相比较而言,随机森林在肺结核患者与肺炎患者的判别上表现更优。

关键词 肺结核;肺炎;支持向量机;随机森林;神经网络
中图分类号 R 183.3
文献标志码 A **文章编号** 1000-1492(2025)03-0507-08
doi:10.19405/j.cnki.issn1000-1492.2025.03.017

早诊早治肺结核能显著提高康复率,但因其症状与肺炎、慢阻肺、流感慢性支气管炎等[1]疾病高度相似,导致当前对肺结核可疑患者(症状相似但未确诊)的诊断需综合考量患者临床症状并通过查体等多种手段来判定其是否罹患肺结核。后疫情时代人力物力分散的情况下,上述诊断方法似有不足之处。

运用机器学习解决问题的案例有很多[2-6],如王成武 等[7]应用支持向量机构建预测模型得出了支持向量机可应用于心脏病辅助诊断的结论。郝金奇 等[8]使用神经网络模型为防治抗结核药物性肝损伤提供了新思路。王冬 等[9]使用优化的随机森林模型对乳腺癌数据集进行分类诊断得到了较好的分类效果。目前关于肺结核可疑患者的判别分析鲜有报道,因此该研究从方法学出发为肺结核可疑患者是否患病提供科学参考。由于患者数据有限,故仅对肺结核与肺炎患者做判别分析。

1 材料与方法

1.1 数据来源 研究选取病例来源于2020至2021年间前往乌鲁木齐市某三甲医院就诊患者数据库中确诊的肺结核与肺炎患者。为确保模型分类效果有效性,在患者数据库中根据严格纳入排除标准进行了筛选,并在符合纳排标准的患者中采用随机抽样方法以平衡肺结核与肺炎患者数据比例,以此获得用于输入模型的高质量数据集,其中肺结核与肺炎患者各430例,男性与女性分别为375例、485例,肺结核与肺炎患者中女性患者分别占总患者57.21%、55.58%,患者年龄为19~97(61.27±14.44)岁。研究数据涉及的指标包括吸烟史、咳嗽、咳痰、气短、胸闷、气喘、气促、胸痛、咯血、发热、下肢水肿、乏力、关节疼痛、心悸,盗汗。肺结核与肺炎患者不同病例特征比较及变量赋值情况分别见表1及表2。

1.2 数据处理

1.2.1 纳入标准 CT影像学检查结果表明患者的肺部影像呈现出了典型的肺结核或肺炎的影像学特征;由两名或更多具有丰富临床经验的医师通过分析患者影像学资料后,结合患者的临床表现和其他相关检查结果,进行详细诊断并最终确认患者患有肺结核或肺炎;患者就诊时提供了详细的主诉信息,例如对症状的详细描述,如咳嗽、咳痰、胸痛、发热等。

1.2.2 排除标准 CT影像学检查结果未显示出典型的肺结核或肺炎影像学特征,或结果不明确,无法做出明确诊断;诊断过程未由至少两名具有丰富临床经验的医师进行,或诊断过程中未结合患者的临床表现和其他相关检查结果;患者未能提供完整且准确的临床症状描述,如咳嗽、咳痰、胸痛、发热等主诉信息;患者存在严重的共存疾病或状况,可能干扰肺结核或肺炎的诊断或治疗,如严重的心血管疾病、肝脏或肾脏功能不全等。

2024-07-18 接收
基金项目:国家自然科学基金项目(编号:72064036、72174175)
作者简介:常敏丽,女,硕士研究生;
郑彦玲,女,教授,博士生导师,通信作者,E-mail:zhengyl_math@sina.cn

表1 肺结核与肺炎患者不同病例特征比较

Tab. 1 Comparison of different case characteristics of patients with tuberculosis and pneumonia

Case characteristics	Total (number)	Pulmonary tuberculosis [n(%)]	Pneumonia [n(%)]	χ^2 value	*P* value
Gender				0.232	0.680
Male	375	184(49.1)	191(50.9)		
Female	485	246(50.7)	239(49.3)		
Age(years)				5.219	0.076
0 – 34	46	24(52.2)	22(47.8)		
35 – 54	194	83(42.8)	111(57.2)		
55 – 100	620	323(52.1)	297(47.9)		
Nationality				39.596	<0.001
Han	486	205(42.2)	281(57.8)		
Hui	53	27(50.9)	26(49.1)		
Uyghur	280	179(63.9)	101(36.1)		
Kazak	28	15(53.6)	13(46.4)		
Russian	2	0(0)	2(100.0)		
Kirgiz	1	1(100.0)	0(0)		
Man	1	0(0)	1(100.0)		
Mongolian	7	3(42.9)	4(57.1)		
Tujia	1	0(0)	1(100.0)		
Xibo	1	0(0)	1(100.0)		
Smoking history				0.091	0.821
Yes	248	126(50.8)	122(49.2)		
No	612	304(49.7)	308(50.3)		
Cough				248.353	<0.001
Yes	419	94(22.4)	325(77.6)		
No	441	336(76.2)	105(23.8)		
Coughing up phlegm				123.457	<0.001
Yes	306	75(24.5)	231(75.5)		
No	554	355(64.1)	199(35.9)		
Shortness of breath				39.568	<0.001
Yes	216	68(31.5)	148(68.5)		
No	644	362(56.2)	282(43.8)		
Chest tightness				5.467	0.024
Yes	165	69(41.8)	96(58.2)		
No	695	361(51.9)	334(48.1)		
Wheezing				4.524	0.046
Yes	53	19(35.8)	34(64.2)		
No	807	411(50.9)	396(49.1)		
Tachypnea				12.858	<0.001
Yes	102	68(66.6)	34(33.3)		
No	758	362(47.8)	396(52.2)		
Chest pain				12.464	0.001
Yes	92	30(32.6)	62(67.4)		
No	768	400(52.1)	368(47.9)		
Hemoptysis				5.273	0.029
Yes	75	47(62.7)	28(37.3)		
No	785	383(48.8)	402(51.2)		
Fever				359.460	<0.001
Yes	493	109(22.1)	384(77.9)		
No	367	321(87.5)	46(12.5)		
Lower extremity edema				7.550	0.010
Yes	23	18(78.3)	5(21.7)		
No	837	412(49.2)	425(50.8)		
Fatigue				42.914	<0.001
Yes	192	136(70.8)	56(29.2)		
No	668	294(44.0)	374(56.0)		
Arthralgia				140.298	<0.001
Yes	201	174(86.6)	27(13.4)		
No	659	256(38.8)	403(61.2)		
Palpitations				34.233	<0.001
Yes	47	43(91.5)	4(8.5)		
No	813	387(47.6)	426(52.4)		
Night sweats				136.242	<0.001
Yes	131	127(96.9)	4(3.1)		
No	729	303(41.6)	426(58.4)		

表2 变量赋值表

Tab. 2 Variable Assignment Table

Main symptom	Assign value
Gender	Male = 1, Female = 0
Age	0 – 14 = 0, 15 – 34 = 1, 35 – 54 = 2, 55 – 100 = 3
Nationality	Han = 0, Hui = 1, Uyghur = 2, Kazak = 3, Russian = 4, Kirgiz = 5, Man = 6, Mongolian = 7, Tujia = 8, Xibo = 9
Smoking history	Yes = 1, No = 0
Cough	Yes = 1, No = 0
Coughing up phlegm	Yes = 1, No = 0
Shortness of breath	Yes = 1, No = 0
Chest tightness	Yes = 1, No = 0
Wheezing	Yes = 1, No = 0
Tachypnea	Yes = 1, No = 0
Chest pain	Yes = 1, No = 0
Hemoptysis	Yes = 1, No = 0
Fever	Yes = 1, No = 0
Lower extremity edema	Yes = 1, No = 0
Fatigue	Yes = 1, No = 0
Arthralgia	Yes = 1, No = 0
Palpitations	Yes = 1, No = 0
Night sweats	Yes = 1, No = 0

1.3 统计学处理

1.3.1 支持向量机 支持向量机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。它可以在将样本完全分为两类的前提下，使决策边界与支持向量的间隔尽可能的远，以达到最好的分类效果。引入核函数后，可通过核函数将原本线性不可分的样本投影到高维空间转化为线性可分的样本[10]。采用了径向基核函数建立支持向量机模型，使用 R4.1.3 中的“e1071”包进行分析。

1.3.2 随机森林 随机森林模型是由一定数量的决策树形成，树越多，随机森林的鲁棒性就越强[11]。它能够利用多棵树对现有患者数据进行训练并预测

新入患者的疾病类别。当给定一个患者数据后，每棵树都会根据患者主诉得出一个分类结果，随机森林会选择决策树中出现次数最高的结果作为森林的分类结果。建模时单个决策树数量取值为500，分裂节点预选变量个数取值为6。使用R4.1.3软件中“randomForest”包进行分析。

1.3.3 神经网络 神经网络是通过训练学习输入和输出之间的关系。在训练期间，网络通过反向传播算法来调整权重和偏置，以使网络产生更准确的输出。反向传播算法通过计算输出结果和实际结果之间的误差，并反向传播到网络中的每个神经元来更新权重和偏置[8]。每个神经元接收来自其他神经元的输入，并将这些输入加权总和，并通过激活函数来产生输出。使用R4.1.3中的“neuralnet”包进行分析。

2 结果

2.1 数据集比较 主要使用Excel 2021对原始数据进行预处理，使用RStudio 2022训练模型并评估模型性能。抽取了860例患者数据集中的75%作为训练数据，25%作为测试数据用于预测并计算预测准确率。训练集由317例肺结核患者数据与328例肺炎患者数据构成。测试集由113例肺结核患者数据与102例肺炎患者数据构成。训练集与测试集中肺结核患者与肺炎患者信息比较见表3。

2.2 支持向量机 为了构建一个性能卓越的支持

表3 训练集与测试集中肺结核患者与肺炎患者信息比较

Tab. 3 Comparison of information on patients with tuberculosis and pneumonia in the training and test sets

Case characteristics	Training set					Test set				
	Total	Pulmonary tuberculosis [*n* (%)]	Pneumonia [*n* (%)]	χ^2	*P* value	Total (number)	Pulmonary tuberculosis [*n* (%)]	Pneumonia [*n* (%)]	χ^2	*P* value
Gender				0.004	0.951				1.168	0.280
Male	280	138(49.3)	142(50.7)			95	46(48.4)	49(51.6)		
Female	365	179(49.0)	186(51.0)			120	67(55.8)	53(44.2)		
Age(years)				3.218	0.200				1.998	0.368
0–34	28	14(50.0)	14(50.0)			18	10(55.6)	8(44.4)		
35–54	154	66(42.9)	88(57.1)			40	17(42.5)	23(57.5)		
55–100	463	237(51.3)	226(48.8)			157	86(54.8)	71(45.2)		
Nationality				32.751	<0.001				8.154	0.148
Han	360	147(40.8)	213(59.2)			126	58(46.0)	68(54.0)		
Hui	48	25(52.1)	23(47.9)			5	2(40.0)	3(60.0)		
Uyghur	204	130(63.7)	74(36.3)			76	49(64.5)	27(35.5)		
Kazak	25	13(52.0)	12(48.0)			3	2(66.7)	1(33.3)		
Russian	1	0(0)	1(100.0)			1	0(0)	1(100.0)		
Kirgiz	1	1(100.0)	0(0)			0	0(0)	0(0)		
Man	1	0(0)	1(100.0)			0	0(0)	0(0)		
Mongolian	3	1(33.3)	2(66.7)			4	2(50.0)	2(50.0)		
Tujia	1	0(0)	1(100.0)			0	0(0)	0(0)		
Xibo	1	0(0)	1(100.0)			0	0(0)	0(0)		
Smoking history				0.635	0.425				0.621	0.431
Yes	184	95(51.6)	89(48.4)			64	31(48.4)	33(51.6)		
No	461	222(48.2)	239(51.8)			151	82(54.3)	69(45.7)		
Cough				204.623	<0.001				45.306	<0.001
Yes	317	65(20.5)	252(79.5)			102	29(28.4)	73(71.6)		
No	328	252(76.8)	76(23.2)			113	84(74.3)	29(25.7)		
Coughing up phlegm				103.087	<0.001				21.905	<0.001
Yes	227	50(22.0)	177(78.0)			79	25(31.6)	54(68.4)		
No	418	267(63.9)	151(36.1)			136	88(64.7)	48(35.3)		
Shortness of breath				34.269	<0.001				6.701	0.010
Yes	154	44(28.6)	110(71.4)			62	24(38.7)	38(61.3)		
No	491	273(55.6)	218(44.4)			153	89(58.2)	64(41.8)		
Chest tightness				4.323	0.038				1.126	0.289
Yes	125	51(40.8)	74(59.2)			40	18(45.0)	22(55.0)		
No	520	266(51.2)	254(48.8)			175	95(54.3)	80(45.7)		

续表

Case characteristics	Training set					Test set				
	Total	Pulmonary tuberculosis [*n* (%)]	Pneumonia [*n* (%)]	χ^2	*P* value	Total (number)	Pulmonary tuberculosis [*n* (%)]	Pneumonia [*n* (%)]	χ^2	*P* value
Wheezing				5.328	0.021				0.033	0.855
Yes	41	13(31.7)	28(68.3)			12	6(50.0)	6(50.0)		
No	604	304(50.3)	300(49.7)			203	107(52.7)	96(47.3)		
Tachypnea				21.211	<0.001				8.785	0.003
Yes	25	1(4.0)	24(96.0)			11	1(9.1)	10(90.9)		
No	620	316(51.0)	304(49.0)			204	112(54.9)	92(45.1)		
Chest pain				11.139	0.001				1.789	1.181
Yes	67	20(29.9)	47(70.1)			25	10(40.0)	15(60.0)		
No	578	297(51.4)	281(48.6)			190	103(54.2)	87(45.8)		
Hemoptysis				2.834	0.092				2.690	0.101
Yes	55	33(60.0)	22(40.0)			20	14(70.0)	6(30.0)		
No	590	284(48.1)	306(51.9)			195	99(50.8)	96(49.2)		
Fever				271.539	<0.001				87.795	<0.001
Yes	373	80(21.4)	293(78.6)			120	29(24.2)	91(75.8)		
No	272	237(87.1)	35(12.9)			95	84(88.4)	11(11.6)		
Lower extremity edema				2.843	0.092				4.972	0.026
Yes	14	10(71.4)	4(28.6)			9	8(88.9)	1(11.1)		
No	631	307(48.7)	324(51.3)			206	105(51.0)	101(49.0)		
Fatigue				24.207	<0.001				20.091	<0.001
Yes	139	94(67.6)	45(32.4)			53	42(79.2)	11(20.8)		
No	506	223(44.1)	283(55.9)			162	71(43.8)	91(56.2)		
Arthralgia				98.861	<0.001				42.197	<0.001
Yes	156	131(84.0)	25(16.0)			45	43(95.6)	2(4.4)		
No	489	186(38.0)	303(62.0)			170	70(41.2)	100(58.8)		
Palpitations				27.581	<0.001				6.838	0.009
Yes	36	33(91.7)	3(8.3)			11	10(90.9)	1(9.1)		
No	609	284(46.6)	325(53.4)			204	103(50.5)	101(49.5)		
Night sweats				92.855	<0.001				43.004	<0.001
Yes	92	88(95.7)	4(4.3)			39	39(100.0)	0(0.0)		
No	553	229(41.4)	324(58.6)			176	74(42.0)	102(58.0)		

向量机模型，研究中采用了网格搜索(Grid Search)法寻找最优模型参数。通过遍历预设的参数组合来寻找模型性能的最佳点。首先为松弛变量和核函数系数设置了不同的数值范围，并生成了一个参数网格。使用该网格对支持向量机模型进行训练和评估，以确定哪个参数组合能够提供最佳的分类性能。为了更准确地评估模型的泛化能力，研究采用了十折交叉验证(10-fold Cross Validation)策略。将数据集分成10等份，每次训练时使用9等份作为训练集，剩余1等份作为验证集。重复进行10次，每次验证集不同，最终性能指标取10次验证的均值。这种方法可以有效地减少过拟合风险，并提供模型在未知数据上的性能估计。经过一系列的训练和验证，发现当松弛变量设置为4，核函数系数为0.135时，基于径向基核函数的支持向量机模型达到了最优的性能。十折交叉验证的寻优过程见图1。

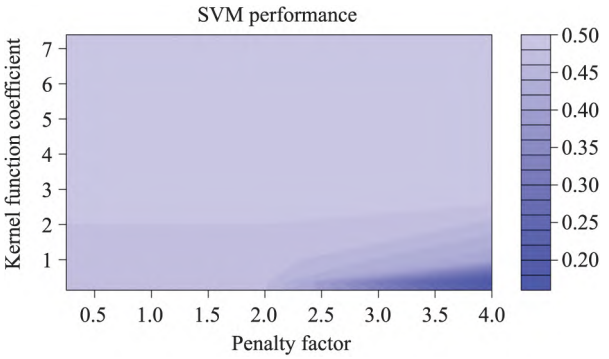

图1 支持向量机分类器寻优过程

Fig. 1 Optimisation process of support vector machine classifier

递归特征消除法可以通过递归减少自变量集合大小选择最重要的自变量。使用递归特征消除法分别列出了从1个自变量到15个自变量每个自变量子集大小的准确性和Kappa系数以及它们的标准

差。可以得出随着自变量数量的增加，模型的准确性和 Kappa 系数通常会提高，这表明更多的自变量有助于提高模型分类性能。自变量子集中最重要的 5 个自变量为发热、咳嗽、盗汗、关节疼痛和咳痰，它们对支持向量机模型分类的贡献最大。准确性和 Kappa 系数的标准差提供了模型性能稳定性的度量。较低的标准差意味着模型性能更加稳定。

2.3 随机森林 在随机森林模型中，研究采用袋外数据(out-of-bag，OOB)来评估模型的性能。袋外数据是指在构建决策树的过程中未被选中用来构建任何树的数据样本。通过这些未参与模型训练的样本，能够得到模型对未知数据预测能力的估计。黑线展示了整体模型的错误率随着决策树棵数的变化情况。而红线和绿线分别代表了肺炎和肺结核预测误差的动态变化。随着决策树棵数的增加模型的性能在初期有了显著的提升，见图 2。这可能是由于更多的决策树能够捕捉到数据中的复杂模式和噪声。然而，当决策树的数量达到一定程度后，模型的性能改善逐渐放缓，并开始趋于稳定。这可能是因为数据中的可分性已经被模型充分利用，进一步增加决策树数量对于提升性能的帮助有限。

使用 importance() 函数评估随机森林模型分类时各自变量的贡献，MeanDecreaseAccuracy 和 MeanDecreaseGini 是随机森林模型中用于评估自变量重要性的两个指标。前者反映了自变量在模型中增加时平均准确率的下降，数值越大表示自变量越重要。

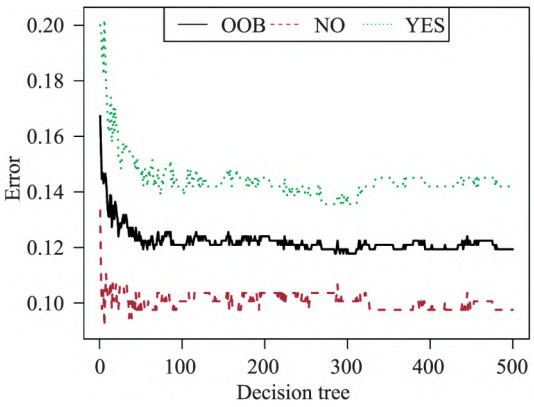

图 2 随机森林袋外数据误差

Fig. 2 Random forest out-of-bag data errors

后者反映了自变量在模型中增加时基尼指数的下降，基尼指数用于衡量数据的纯度，数值越大表示自变量越重要。因此可以得出发热、盗汗、咳嗽、关节疼痛和心悸对于随机森林模型分类的贡献最大，见表 4。

2.4 神经网络 对于神经网络模型，研究中构建了具有两个隐藏层(第一个层有 4 个神经元，第二个层有 3 个神经元)的非线性神经网络模型，用于对训练集进行训练。其中输出层使用了非线性激活函数，可以引入激活函数的跳跃性，有助于减少训练过程中的梯度消失问题，加快学习速度。同时也可以使神经网络拥有更强的表达能力，从而在训练数据上获得更好的拟合效果。以便模型能够捕捉和模拟

表 4 各模型自变量贡献

Tab. 4 Contribution of the independent variables of each model

Variable	SVM				RF		NNM			
	Accuracy	Kappa coefficient	Accuracy standard deviation	Kappacoefficient standard deviation	Mean decrease accuracy	Mean decrease gini	Hidden layer1 neuron1	Hidden layer1 neuron2	Hidden layer1 neuron3	Hidden layer1 neuron4
Smoking history	0.821	0.642	0.055	0.112	8.632	5.864	1.256	3.025	0.427	11.252
Cough	0.821	0.642	0.055	0.112	36.945	48.920	0.433	-13.799	2.585	31.077
Coughing up phlegm	0.839	0.676	0.032	0.063	14.362	13.271	-2.071	-19.751	-1.634	8.472
Shortness of breath	0.862	0.723	0.039	0.077	8.708	7.159	-4.716	15.852	-0.997	-12.465
Chest tightness	0.871	0.742	0.040	0.080	4.411	5.139	10.490	-3.913	-1.057	-47.618
Wheezing	0.877	0.754	0.042	0.083	7.469	3.531	-707.716	-8.198	-3.559	-9.488
Tachypnea	0.880	0.761	0.044	0.090	7.830	1.768	-696.732	-54.926	-1.197	-116.570
Chest pain	0.884	0.767	0.044	0.089	2.316	3.939	-11.089	-4.117	-1.671	-112.934
Hemoptysis	0.890	0.780	0.043	0.086	8.528	5.061	-5.355	19.613	-2.633	-45.857
Fever	0.893	0.786	0.041	0.082	76.789	99.069	-8.542	2.871	3.577	-3.550
Lower extremity edema	0.891	0.783	0.038	0.077	0.040	0.918	8.413	-31.243	-0.651	7.198
Fatigue	0.890	0.779	0.037	0.073	8.141	7.963	-1.989	-6.662	-1.093	6.731
Arthralgia	0.885	0.770	0.048	0.097	34.722	18.364	9.392	33.716	-0.728	-1 000.704
Palpitations	0.884	0.767	0.046	0.091	20.332	4.883	7.269	13.109	-1.573	-178.704
Night sweats	0.882	0.764	0.049	0.098	61.704	30.373	753.792	29.326	-0.599	256.940

SVM: support vector machine; RF: random forest; NNM: neural network model.

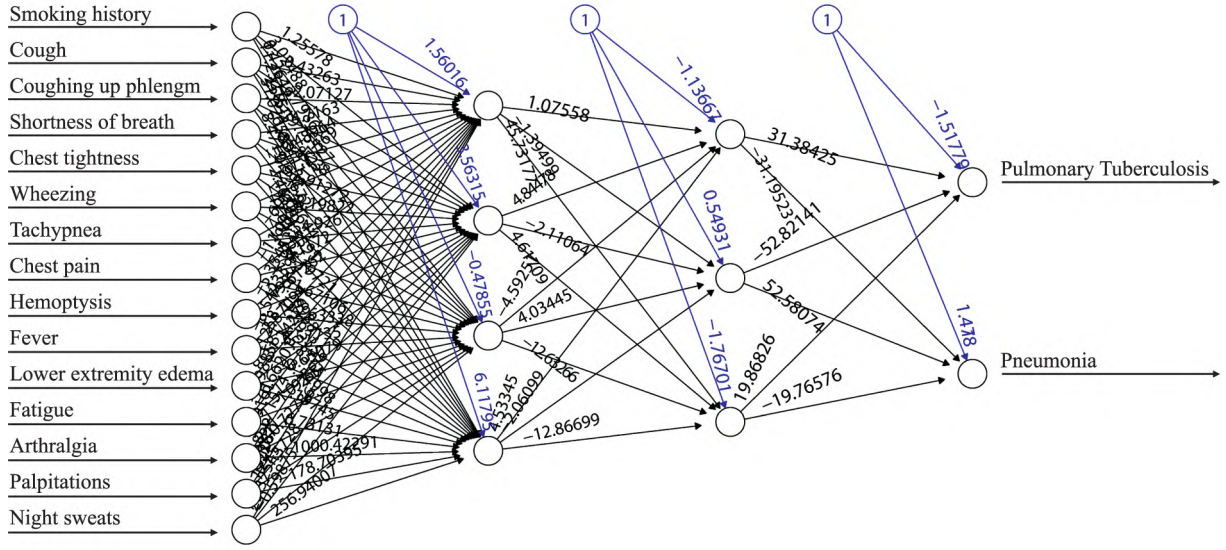

图 3 神经网络结构图

Fig. 3 Neural network diagram

数据中的复杂关系。黑线清晰地展示了每一层与其相关权重之间的直接联系。与此同时，蓝色线则揭示了在拟合过程中每一步所添加的误差项，见图 3。这些误差项共同构成了一个能够表征线性模型误差范围的区间，为模型的精确性和可靠性提供了重要保障。

神经网络的训练通常通过反向传播算法实现，反向传播算法是一种基于梯度下降的优化方法，通过计算输出误差对各层节点权重的梯度，并对权重进行更新，从而不断优化神经网络的性能。表 4 展示了从输入层到第 1 个隐藏层的各个神经元的连接权重，从中可以得出盗汗和咳嗽的权重最大，即对神经网络模型而言，进行判别分析时盗汗和咳嗽的贡献最大。

2.5 性能对比 使用 3 种机器学习方法搭建的肺结核可疑患者判别分析模型性能对比结果详见表 5，基于随机森林的肺结核可疑患者分类模型准确率为 91%，灵敏度为 87%，特异度为 96%，整体优于支持向量机和神经网络。结合不同分类器下肺结核可疑患者判别分类模型 ROC 曲线可得：相比较于支持向量机和神经网络模型，基于随机森林建立的模型在判别肺结核患者与肺炎患者上更具优势，见图 4。

3 讨论

肺结核是中国重点防治的传染病之一，对其进行早期诊断对预防及治疗工作至关重要。当前肺结核诊断主要依赖于临床方法，但在后疫情时代，有限的人力和物力资源使得这种方法在效率和成本上存在挑战。因此基于文献研究[8, 12-13]选用支持向量

表 5 肺结核可疑患者判别分析性能对比

Tab. 5 Comparison of performance of discriminant analysis of patients with suspected tuberculosis

Model	Accuracy(95% *CI*)	Sensitivity	Specificity	AUC
SVM	0.90(0.849 – 0.935)	0.88	0.92	0.941
RF	0.91(0.865 – 0.946)	0.87	0.96	0.945
NNM	0.88(0.833 – 0.923)	0.83	0.94	0.937

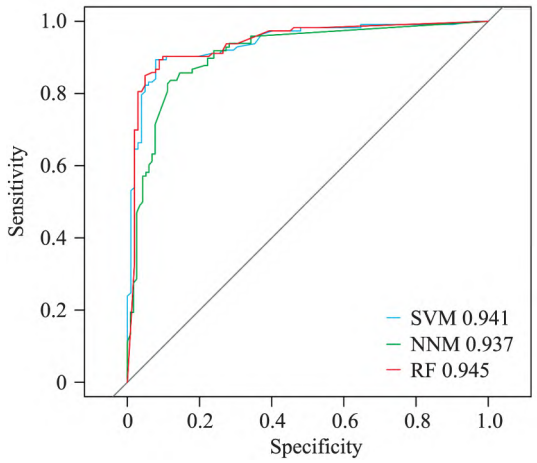

图 4 肺结核可疑患者分类模型 ROC 曲线

Fig. 4 ROC curve of the classification model for suspected tuberculosis patients

机、随机森林和神经网络方法构建了肺结核可疑患者的判别分类模型。研究结果显示，上述3种方法均能通过症状有效区分肺结核患者和肺炎患者，判别精确率均超过87%，其中基于随机森林的模型表现最为突出。通过ROC曲线分析，证实了这3个模型均具有较好的效果和稳健性。

在随机森林模型中，袋外数据的预测率为12.4%，表明模型能够有效地区分两类疾病。该模型将35例肺结核患者误判为肺炎患者，将19例肺炎患者误判为肺结核患者，而有282例肺结核患者和309例肺炎患者的判别结果与实际情况相符，显示其判别精度高达91.63%。这一精度与王冬等[9]利用人工鱼群优化随机森林模型对乳腺癌数据集进行分类诊断的结果相当，均显示出优秀的分类效果，证明了分类模型可作为一种有效的辅助诊断工具。此外，Xin et al[13]的研究也表明，随机森林分类器能从多个维度预测残疾老年人的抑郁状态，提高早期抑郁症状的识别能力。

在探究病理特征对模型分类性能的影响时从多个维度和模型综合分析了各自变量的重要性。首先通过递归特征消除法发现增加自变量数量能够有效提升模型分类性能，其中发热、咳嗽、盗汗、关节疼痛和咳痰被支持向量机模型识别为最重要的特征。基于随机森林模型得到了从不同角度揭示自变量重要性的指标，发热、盗汗、咳嗽、关节疼痛和心悸的高值表明这些特征对维持模型准确率至关重要，证实了上述特征在随机森林模型中的核心地位。神经网络模型训练过程依赖于反向传播算法通过调整连接权重来优化网络性能，从神经网络的第一个隐藏层连接权重可以看出，盗汗和咳嗽的权重最大，这表明在神经网络模型中，这两个特征对于分类决策的贡献最为显著。上述3个模型均强调了发热、咳嗽、盗汗等自变量的重要性。这些自变量不仅在临床实践中被广泛用作疾病诊断依据，也在机器学习模型中表现出了良好的分类效果。

上述模型也存在着一定局限性。首先研究对象主要是50至75岁及以上的老年人，主诉可能存在理解偏倚。其次在模型训练时并未考虑老年人常见并发症对肺结核诊断和治疗的影响。最后样本量较小，可能影响从样本到总体推断的准确性。因此，未来的研究需要继续扩大样本规模，严格筛选研究对象，并持续优化输入症状[14]，同时考虑更多分类预测模型，增加与其他分类方法的比较，以期建立更加优化和完善的肺结核可疑患者判别分类模型。该模型有望在节省医疗资源的同时，提高肺结核患者的早期诊断率，进而促进早期治疗。

综上所述，基于机器学习的分类模型在肺结核与肺炎的判别中展示了良好的性能。可以为肺结核可疑患者的早期诊断提供较好的方法学参考。

参考文献

[1] Abdul Hamid M F, Selvarajah S B, Nuratiqah N, et al. Two cases of cryptogenic organizing pneumonia masquerading as tuberculosis (TB) in a TB endemic area [J]. Respirol Case Rep, 2021, 10(1): e0883. doi: 10.1002/rcr2.883.

[2] Choong C S, Ab Nasir A F, Abdul Majeed A P P, et al. Investigation of features for classification RFID reading between two RFID reader in various support vector machine kernel function [M]// Lecture Notes in Electrical Engineering. Singapore: Springer Singapore, 2021: 127-39. doi: 10.1007/978-981-33-4597-3_13.

[3] P. Lakhani, B. Sundaram, 王 臣. 胸部X线影像的深度学习：应用卷积神经网络进行肺结核自动分类 [J]. 国际医学放射学杂志, 2017, 40(5): 601-2. doi: 10.19300/j.2017.r0822.

[3] P. Lakhani, B. Sundaram, Wang C. Deep learning of chest x-ray images: application of convolutional neural networks for automatic classification of tuberculosis [J]. Int J Med Radiol, 2017, 40(5): 601-2. doi: 10.19300/j.2017.r0822.

[4] Sahoo A, Samantaray S, Ghose D K. Multilayer perceptron and support vector machine trained with grey wolf optimiser for predicting floods in Barak river, India [J]. J Earth Syst Sci, 2022, 131(2): 85. doi: 10.1007/s12040-022-01815-2.

[5] Yücelbaş Ş, Yücelbaş C. Autism spectrum disorder detection using sequential minimal optimization-support vector machine hybrid classifier according to history of jaundice and family autism in children [J]. Concurr Comput Pract Exp, 2022, 34(1): e6498. doi: 10.1002/cpe.6498.

[6] Zhu Y, Yin S, Zheng J, et al. O-glycosylation site prediction for Homo sapiens by combining properties and sequence features with support vector machine [J]. J Bioinform Comput Biol, 2022, 20(1): 2150029. doi: 10.1142/S0219720021500293.

[7] 王成武，郭志恒，晏峻峰. 改进的支持向量机在心脏病预测中的研究 [J]. 计算机技术与发展, 2022, 32(3): 175-9. doi: 10.3969/j.issn.1673-629X.2022.03.029.

[7] Wang C W, Guo Z H, Yan J F. Application of improved support vector machine in heart disease prediction [J]. Comput Technol Dev, 2022, 32(3): 175-9. doi: 10.3969/j.issn.1673-629X.2022.03.029.

[8] 郝金奇，高鹏飞，余艳琴，等. 基于人工神经网络模型分析抗结核药物性肝损伤危险因素 [J]. 中国疗养医学, 2024, 33

(5): 36 – 9. doi: 10.13517/j.cnki.ccm.2024.05.007.

[8] Hao J Q, Gao P F, Yu Y Q, et al. Analysis of risk factors for liver injury caused by anti-tuberculosis drugs based on an artificial neural network model [J]. Chin J Conval Med, 2024, 33(5): 36 – 9. doi: 10.13517/j.cnki.ccm.2024.05.007.

[9] 王 冬，曲 媛，刘玉航，等. 基于优化随机森林算法的乳腺癌分类诊断 [J]. 计算机工程与设计，2022，43(3): 706 – 12. doi: 10.16208/j.issn1000 – 7024.2022.03.015.

[9] Wang D, Qu Y, Liu Y H, et al. Classification diagnosis of breast cancer based on optimized random forest algorithm [J]. Comput Eng Des, 2022, 43(3): 706 – 12. doi: 10.16208/j.issn1000 – 7024.2022.03.015.

[10] 储有兵，费胜巍，范 晞. 基于 WT-SVD-SVM 和 WT-SVD-KNN 的运动想象脑电信号特征提取及分类 [J]. 东华大学学报(自然科学版)，2019，45(6): 881 – 7. doi: 10.3969/j.issn.1671 – 0444.2019.06.012.

[10] Chu Y B, Fei S W, Fan X. Feature extraction and classification of motor imagery electroencephalogram signals based on WT-SVD-SVM and WT-SVD-KNN [J]. J Donghua Univ Nat Sci, 2019, 45(6): 881 – 7. doi: 10.3969/j.issn.1671 – 0444.2019.06.012.

[11] 郭 睿，王忆勤，燕海霞，等. 基于血液动力学原理的中医脉搏波特征提取与识别 [J]. 中西医结合学报，2010，8(8): 742 – 6. doi: 10.3736/icim20100802.

[11] Guo R, Wang Y Q, Yan H X, et al. Feature extraction and recognition of traditional Chinese medicine pulse based on hemodynamic principles [J]. J Chin Integr Med, 2010, 8(8): 742 – 6. doi: 10.3736/icim20100802.

[12] Salem A M, Yakoot M S, Mahmoud O. A novel machine learning model for autonomous analysis and diagnosis of well integrity failures in artificial-lift production systems [J]. Adv Geo-Energy Res, 2022, 6(2): 123 – 42. doi: 10.46690/ager.2022.02.05.

[13] Xin Y, Ren X. Predicting depression among rural and urban disabled elderly in China using a random forest classifier [J]. BMC Psychiatry, 2022, 22(1): 118. doi: 10.1186/s12888 – 022 – 03742 – 4.

[14] 顾天宇，严壮志，蒋皆恢. 基于支持向量机的中风病中医证候分类 [J]. 中医药信息，2021，38(9): 1 – 3. doi: 10.19656/j.cnki.1002 – 2406.20210901.

[14] Gu T Y, Yan Z Z, Jiang J H. Classification of TCM syndrome patterns of stroke based on SVM [J]. Inf Tradit Chin Med, 2021, 38(9): 1 – 3. doi: 10.19656/j.cnki.1002 – 2406.20210901.

Discriminant analysis of pulmonary tuberculosis patients and pneumonia patients based on machine learning

Chang Minli[1], You Shuping[2], Chen Xiaodie[1], Chen Zhifei[3], Zheng Yanling[3]

([1]*College of Public Health*, [2]*College of Nursing*, [3]*College of Medical Engineering and Technology*, *Xinjiang Medical University*, *Urumqi* 830017)

Abstract ***Objective*** To explore the feasibility of machine learning methods in the discrimination of tuberculosis patients. ***Methods*** The data of 15 observation indicators of 860 patients were obtained from a tertiary hospital. Through in-depth mining and analysis of the data, support vector machine, random forest and neural network model methods were used to discriminate the diseases of patients. ***Results*** The accuracies of the TB suspected patient discrimination models based on support vector machine, random forest and neural network were 90%, 91% and 88%, respectively. ***Conclusion*** All three machine learning methods can be used for discriminative analysis of suspected tuberculosis patients. In comparison, random forest performs better in discriminating patients with tuberculosis from those with pneumonia.

Key words pulmonary tuberculosis; pneumonia; support vector machine; random forest; neural network model

Fund programs National Natural Science Foundation of China (Nos. 72064036, 72174175)

Corresponding author Zheng Yanling, E-mail: zhengyl_math@sina.cn