面向结核分枝杆菌抗微生物药物耐药性预测的边缘人工智能

Nishat Tasnim*, Faisal Ahmed ‡‡ , Rokan Uddin Faruqui*

* 吉大港大学计算机科学与工程系，孟加拉国 电子邮箱：tasnim.cse@std.cu.ac.bd、rokan@cu.ac.bd

† 罗维拉-维尔吉利大学计算机工程与数学系，西班牙

‡ NVISION Systems and Technologies SL（INTERA Group），西班牙巴塞罗那 电子邮箱：faisal.ahmed@intera-group.com

arXiv:2611.29143v1 [cs.LG] 22 Nov 2026

摘要——**结核分枝杆菌（MTB）的抗微生物药物耐药性（AMR）对全球结核病防控构成重大挑战，因为耐多药菌株限制了可用的治疗方案。基于基因组数据快速、准确地预测药物耐药性，有助于及时制定治疗决策并采取公共卫生干预措施。然而，许多现有的基因组抗微生物药物耐药性预测计算方法依赖计算密集型流程，因此难以部署在资源受限的边缘设备上。在中低收入国家的医院环境中，这一限制尤为突出，因为当地的计算基础设施和高性能计算资源可能有限。为应对这一挑战，我们提出了一种轻量级的新型机器学习（ML）方法，适用于在边缘设备上运行。该方法结合了从全基因组序列中提取的4-mer频率特征与随机森林分类器，以预测结核分枝杆菌的抗生素耐药性。在独立测试集上，该机器学习方法对20种抗生素的受试者工作特征曲线下面积平均值达到0.82。**

索引词—**抗微生物药物耐药性；结核分枝杆菌；机器学习；随机森林；k-mer；全基因组测序；边缘人工智能。**

I. 引言

结核病（TB）仍是全球传染病致死的主要原因之一。世界卫生组织估计，每年新增病例达数百万例，死亡人数超过百万 [1]。这种疾病由细菌结核分枝杆菌（*MTB*）引起，感染者咳嗽或打喷嚏时释放的空气传播飞沫会传播病菌。药物敏感型结核病的治疗成功率约为 88%，而多重耐药病例和利福平耐药病例的治疗成功率则降至 57–68%。因此，耐药菌株的出现和传播尤其阻碍全球根除工作；快速、准确地识别抗微生物药物耐药性（AMR），对于患者护理、治疗管理和公共卫生监测至关重要 [2]。

传统的抗生素药敏试验依赖培养法，通常需要四至六周才能得出结果，从而延误适当治疗，并导致患者预后不佳 [3]。GeneXpert 等分子诊断方法因快速、简便而得到广泛应用，但这些检测仅能识别靶向基因组区域中导致耐药性的特定突变，因此可能漏检新出现或罕见的变异，进而导致误诊和治疗无效 [4]。这些问题促使人们开发能够全面刻画药物耐药性并快速、准确地进行预测的计算工具。机器学习（ML）已成为一种强有力的计算方法，可根据基因组数据预测抗微生物药物耐药性 [5]–[7]。

用于预测结核分枝杆菌 抗生素耐药性的机器学习方法通常依赖耐药相关变异，例如单核苷酸多态性（TBProfiler [8]; Mykrobe [9]），或采用高维特征表示（MTB++ [10]）。基于耐药相关变异的方法需要预先了解耐药相关突变，因此限制了其识别新突变的能力 [11], [12]。高维表示虽能避免这一限制，但计算开销较大，因此不太适合部署在资源有限的环境中。在高负担、低资源环境中，这一限制尤为突出：测序能力日益普及，但持续使用高性能计算资源或通过可靠的网络连接访问远程服务器仍不可行。在此类环境中，设备端推理可使耐药性预测留在护理现场。此外，现有研究大多仅评估模型对有限的一线抗生素的预测性能，而且往往缺乏严谨的实验设计。

为解决这些局限性，我们提出了一种轻量级机器学习方法，利用紧凑的 4-mer 频率表示，从全基因组序列预测结核分枝杆菌的抗微生物药物耐药性，
专为资源受限的医院和诊所环境中的边缘部署而设计。我们的方法与以往研究有两点主要不同。首先，我们用 256 维的 4-mer 频率向量取代高维特征表示，将特征空间缩小了数个数量级，并免去了使用稀疏矩阵表示的需要

通讯作者：Faisal Ahmed

或进行大规模特征排序。其次，我们评估了针对 20 种抗生素的耐药性，其中包括若干在以往研究中较少受到关注的新型药物和再利用药物。具体而言，我们直接从全基因组序列中提取 4-mer 频率向量，按 k-mer 总数对其进行归一化，并将其作为针对特定抗生素的二分类随机森林（RF）分类器的输入。我们为每种抗生素训练一个分类器，并通过类别加权来补偿耐药分离株与敏感分离株之间的不平衡，这种不平衡影响了许多药物。我们的主要贡献如下：

1) 我们提出了一种用于结核分枝杆菌抗微生物药物耐药性预测的轻量级方法，将紧凑的 4-mer 频率表示与随机森林分类器相结合，旨在构建一个计算复杂度更低、更适合资源受限边缘计算环境的框架。

2) 我们使用严格的五折交叉验证和独立测试集，系统评估了所提出方法在多种抗生素上的性能，包括一线、二线以及新型和重新使用的药物。结果表明，对于特征表征充分的抗生素，4-mer 表示能够实现稳健的预测性能。

II. 方法

A. 特征编码

为构建预测模型，我们使用k-mer 特征编码 [13] with $k = 4$，将原始 DNA 序列转换为固定长度的数值特征向量。4-mer 定义为长度为四的连续核苷酸子串。对于每条 DNA 序列，使用长度为四、步长为一个核苷酸的滑动窗口，提取所有相互重叠的 4-mer；对于长度为 L 的序列，共得到（$L-3$）个 4-mer。随后，将提取出的每个 4-mer 映射到由四种 DNA碱基（A、C、G 和 T）构成的 $4^4 = 256$ 种可能核苷酸组合之一。最后，对所有可能的 4-mer 进行计数，生成一个 256 维特征向量，其中每个特征对应于特定 4-mer 的频率。

为消除序列长度的影响，我们将每个原始 4-mer 计数除以提取到的 4-mer 总数进行归一化，从而得到相对频率值。因此，每条 DNA 序列均表示为一个归一化的 256 维特征向量，其取值范围为 0 到 1。该表示能够有效捕捉基因组序列的局部组成特征，同时确保所有样本的特征维度固定。

与 k 的较小取值相比，4-mer 能捕捉更丰富的序列模式和上下文信息。相反，k 的较大取值会显著扩大特征空间，导致更高的计算复杂度和更大的稀疏性。因此，选择 $k = 4$ 可在描述能力与计算效率之间实现有效平衡，非常适合用于训练轻量级预测模型。由全基因组序列生成的代表性 4-mer 频率特征向量见表 I。

表 I：从全基因组序列中提取的 4-mer 频率代表性 256 维特征向量。

基因组	1	2	3	⋯	255	256
1	0.042	0.038	0.041	⋯	0.042	0.038
2	0.039	0.040	0.038	⋯	0.040	0.039
⋮	⋮	⋮	⋮	⋱	⋮	⋮
598	0.039	0.041	0.040	⋯	0.040	0.039
599	0.042	0.040	0.039	⋯	0.038	0.041
600	0.038	0.041	0.042	⋯	0.041	0.040

B. 预测模型构建

为构建预测模型，我们使用了随机森林（RF）分类器，因为它在基因组学应用中已得到广泛验证，[14], [15]，比线性分类器更能抵抗共线特征，并且能够内置特征重要性估计。在训练过程中，它会构建多棵决策树，并通过多数投票汇总其输出。这种设计限制了过拟合，能够捕捉特征与耐药表型之间的非线性关系，并生成重要性得分，以表明哪些 k-mer 基序对预测贡献最大。

我们将每种抗生素的耐药性预测视为一个独立的二元分类问题，并针对每种抗生素训练一个分类器。输入为 256 维归一化 4-mer 频率向量，输出为相应的二元表型（耐药或敏感）。由于许多抗生素对应的两类样本分布不平衡，我们采用类别权重，提高将少数类（耐药）误分类的代价。

我们在训练集上通过初步网格搜索选择超参数，以平衡预测性能与计算成本。最终配置采用 100 个估计器、最大深度为 10，且每次分裂至少包含 5 个样本。

III. 实验设置

A. 数据集

本研究使用的数据集来源于细菌与病毒生物信息学资源中心（BV-BRC）[1]。我们提取了结核分枝杆菌的AMR表型记录，得到329,980条耐药记录和1,069,517条敏感记录。这些记录涵盖755条基因组序列和20种抗生素。每条记录包含21个属性，例如基因组信息、抗菌药物及相关耐药表型。本研究仅重新训练了四个属性：基因组ID、基因组序列、抗生素和耐药表型。

随后，我们按抗生素对记录进行分组，得到20个不同的子集，每个子集对应一种特定抗生素。每个子集都包含755条基因组序列中与相应抗生素相关的耐药表型标签。

表II中全面汇总了这些数据集。

[1]https://www.bv-brc.org/

表II：各抗生素耐药（R）和敏感（S）分离株的分布

抗生素	R	S	R:S 比率
阿米卡星	101	654	1:6.48
卷曲霉素	137	618	1:4.51
环丙沙星	128	592	1:4.63
氯法齐明	81	639	1:7.89
环丝氨酸	82	638	1:7.78
德拉马尼	448	272	1.65:1
乙胺丁醇	342	413	1:1.21
乙硫异烟胺	143	577	1:4.03
异烟肼	408	347	1.18:1
卡那霉素	165	590	1:3.58
左氧氟沙星	285	435	1:1.53
莫西沙星	117	603	1:5.15
烟酰胺	150	570	1:3.80
氧氟沙星	163	592	1:3.63
对氨基水杨酸	62	583	1:9.40
丙硫异烟胺	445	275	1.62:1
吡嗪酰胺	331	389	1:1.18
利福布汀	278	442	1:1.59
利福平	471	284	1.66:1
链霉素	400	355	1.13:1

B. 评估指标

我们使用以下标准分类指标评估了该方法，其中 TP、TN、FP 和 FN 分别表示真正例、真负例、假正例和假负例。耐药分离株始终被视为正类。

灵敏度（*Sn*），或称召回率，用于衡量被正确识别的耐药分离株所占的比例：

$$Sn = \frac{TP}{TP + FN}$$

特异度（*Sp*）用于衡量被正确识别的敏感分离株所占的比例：

$$Sp = \frac{TN}{TN + FP}$$

精确率（*Pr*）用于衡量预测为耐药且实际耐药的分离株所占的比例：

$$Pr = \frac{TP}{TP + FP}$$

F1 分数（*F1*）是精确率与召回率的调和平均数，将两种错误类型合并为一个数值：

$$F1 = 2 \times \frac{Pr \times Sn}{Pr + Sn}$$

马修斯相关系数 *(MCC)*汇总混淆矩阵中的四个条目，并且在类别不平衡情况下仍具有信息量：

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

*ROC*曲线下面积 *(*曲线下面积*)*衡量模型在所有分类阈值下区分耐药分离株和敏感分离株的能力。

IV. 结果与讨论

A. 交叉验证性能与稳健性分析

首先，我们使用 5 折分层交叉验证评估模型的稳健性，并报告了六项评估指标在各折中的均值和标准差（表 III）。模型性能因抗生素而异，表明不同药物的预测难度存在差异。利福平、环丝氨酸和丙硫异烟胺的性能最高，其 F1 分数均超过 83%。相比之下，对氨基水杨酸、烟酰胺和其他几种药物的预测效果较差。这种差异表明，在该数据集中，模型性能因抗生素和耐药表型而异。

汇总结果显示，模型性能存在一致的不对称性：在所有抗生素中，特异度均高于灵敏度。这表明，即使采用类别加权，模型通常也更擅长识别敏感分离株，而不是耐药分离株。对于耐药分离株较少的抗生素，这种不平衡尤为明显：极高的特异度与相对较差的耐药性检出率同时出现。在临床上，这一区别十分重要，因为耐药假阴性预测可能导致不恰当的治疗，因此灵敏度是评估这些模型时尤其重要的指标。

交叉验证结果的变异性进一步揭示了在哪些情况下应谨慎解读性能估计。利福平和异烟肼的预测结果在各验证折之间相对稳定，而卷曲霉素、环丙沙星，尤其是阿米卡星则表现出更大的折间变异。特别是阿米卡星，尽管其精确率均值处于中等水平，但精确率却呈现出显著的变异性。这种变异性与某些抗生素可获得的耐药分离株数量有限相一致。当少数类样本量较小时，将耐药分离株分配到五个验证折中，可能导致个别预测结果对折级指标产生不成比例的影响，进而造成较大的标准差。因此，对于这些抗生素，性能估计的确定性较低，并且可能对训练折和验证折的构成更加敏感。

总之，交叉验证结果表明，该模型能够稳健地预测一线抗生素的耐药性；然而，其在不同药物中的表现并不一致。主要局限仍是灵敏度与特异度之间持续存在的差距，尤其是在耐药分离株数量有限的抗生素中。这些发现凸显了按单种抗生素评估耐药性预测，并报告类别敏感指标及其变异性的重要性，而不能仅依赖总体表现。此外，增加耐药分离株的代表性，特别是针对较少见的耐药表型，可能是提升耐药检测能力和模型泛化能力的关键。

表 III：5 折交叉验证下的性能，以均值 ± 标准差（%）表示

Drug	Pr	Sn	Sp	F1	MCC	AUC
异烟肼	78.45 ± 2.72	75.16 ± 2.95	79.46 ± 3.78	76.70 ± 1.71	54.78 ± 3.43	85.34 ± 1.75
利福平	88.22 ± 2.99	81.94 ± 3.74	84.45 ± 4.98	84.84 ± 1.06	65.97 ± 1.90	91.36 ± 2.19
环丝氨酸	80.88 ± 6.73	88.36 ± 7.12	97.95 ± 0.92	84.20 ± 5.24	82.87 ± 5.73	98.55 ± 0.60
丙硫异烟胺	84.33 ± 3.94	83.15 ± 6.03	75.89 ± 7.44	83.55 ± 3.12	59.36 ± 7.11	89.60 ± 3.84
链霉素	75.38 ± 2.90	69.77 ± 4.11	78.30 ± 3.99	72.34 ± 1.96	48.43 ± 3.56	84.04 ± 2.56
吡嗪酰胺	76.20 ± 2.52	67.37 ± 5.01	85.67 ± 0.91	71.48 ± 3.95	54.29 ± 5.48	85.60 ± 3.32
乙胺丁醇	72.65 ± 3.91	69.91 ± 5.72	81.64 ± 3.31	71.14 ± 4.12	51.95 ± 6.11	86.07 ± 4.34
氯法齐明	66.96 ± 9.00	61.67 ± 6.67	96.41 ± 1.38	63.89 ± 6.60	60.19 ± 7.33	93.12 ± 3.34
左氧氟沙星	73.35 ± 3.28	60.90 ± 5.12	87.49 ± 2.02	66.45 ± 3.86	50.76 ± 5.23	82.92 ± 2.97
卡那霉素	76.06 ± 8.13	62.94 ± 5.05	95.53 ± 1.75	68.68 ± 5.21	63.10 ± 6.46	87.13 ± 3.08
利福布汀	76.09 ± 4.16	66.62 ± 6.79	88.89 ± 3.14	70.73 ± 3.70	57.63 ± 4.17	88.09 ± 0.94
环丙沙星	49.47 ± 10.40	44.41 ± 12.57	92.50 ± 2.90	45.91 ± 9.94	38.52 ± 10.77	78.47 ± 3.83
德拉马尼	82.09 ± 4.60	77.81 ± 3.10	70.16 ± 8.25	79.87 ± 3.62	47.16 ± 10.19	81.59 ± 4.26
乙硫异烟胺	88.06 ± 4.52	65.33 ± 5.52	98.15 ± 0.77	74.93 ± 4.93	71.79 ± 5.42	81.53 ± 7.51
氧氟沙星	52.46 ± 4.82	47.00 ± 5.10	91.20 ± 2.32	49.13 ± 1.31	39.96 ± 1.45	78.82 ± 4.31
卷曲霉素	61.12 ± 11.08	50.00 ± 12.17	94.12 ± 2.70	54.23 ± 10.74	47.90 ± 11.78	83.97 ± 5.08
莫西沙星	48.67 ± 9.28	39.56 ± 6.53	94.23 ± 2.19	43.11 ± 6.25	37.02 ± 7.07	74.49 ± 7.81
烟酰胺	47.65 ± 5.82	37.11 ± 8.81	90.32 ± 2.04	41.31 ± 6.54	30.21 ± 6.71	73.77 ± 2.82
阿米卡星	67.41 ± 17.67	45.00 ± 8.50	97.04 ± 1.70	52.25 ± 6.10	50.03 ± 7.12	81.18 ± 6.50
对氨基水杨酸	53.57 ± 11.07	29.72 ± 6.31	97.52 ± 1.05	37.81 ± 7.11	35.75 ± 7.84	76.08 ± 6.32
平均值	69.95 ± 6.48	61.19 ± 6.35	88.85 ± 2.88	64.63 ± 4.85	52.38 ± 6.24	84.09 ± 3.87

B. 独立测试集上的性能分析

我们还在一个包含 155 个基因组的独立测试集上评估了该模型，这些基因组未用于训练。

表 IV：独立测试集上的性能（数值以 % 表示）

Drug	Pr	Sn	Sp	F1	MCC	受试者工作特征曲
异烟肼	89.19	90.00	73.33	89.59	63.76	89.36
利福平	90.35	84.43	66.67	87.29	47.42	87.85
环丝氨酸	89.29	80.65	97.03	84.75	80.54	95.23
丙硫异烟胺	83.91	82.02	67.44	82.95	48.91	84.11
链霉素	87.50	77.06	73.91	81.95	47.96	84.30
吡嗪酰胺	90.28	70.65	82.50	79.27	49.06	84.58
乙胺丁醇	80.00	75.00	69.49	77.42	43.78	82.57
氯法齐明	64.29	85.71	90.99	73.47	68.63	93.52
左氧氟沙星	84.62	60.27	86.44	70.40	47.53	79.35
卡那霉素	81.40	61.40	91.84	70.00	57.34	85.05
利福布汀	76.56	63.64	72.73	69.50	35.87	81.63
环丙沙星	83.87	55.32	94.12	66.67	55.84	84.24
德拉马尼	67.61	64.86	60.34	66.21	25.10	69.79
乙硫异烟胺	91.30	50.00	97.78	64.62	58.67	79.09
氧氟沙星	78.57	52.38	90.22	62.86	47.08	81.19
卷曲霉素	69.70	48.94	90.74	57.50	44.55	78.69
莫西沙星	76.19	32.65	93.98	45.71	35.17	77.53
烟酰胺	46.43	35.14	84.21	40.00	21.25	71.44
阿米卡星	58.82	24.39	93.86	34.48	25.76	71.54
对氨基水杨酸	42.86	16.67	95.96	24.00	19.21	79.04
平均值	76.64	60.56	83.68	66.43	46.17	82.01

表 IV 报告了独立测试集上全部 20 种抗生素的完整性能指标。在这 20 种抗生素中，模型的平均准确率为 77.50%，F1 分数为 0.664，ROC 曲线下面积为 0.820。不同抗生素的表现各异，其中异烟肼和利福平的表现最佳（F1 分数分别为 89.59% 和 87.29%，

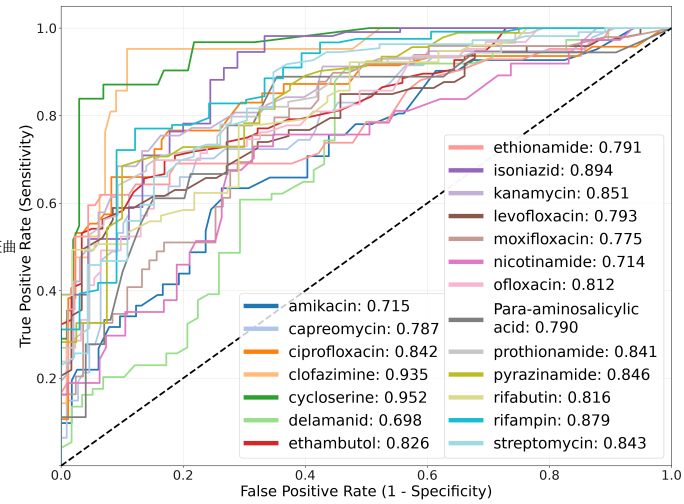

图 1：在独立测试集上评估的 20 种抗生素的 ROC 曲线。

而对氨基水杨酸和阿米卡星的表现则明显较差。

环丝氨酸仍属于表现最佳的抗生素之一，并展现出较强的整体区分能力，这与其交叉验证性能一致。这些模式表明，预测性能反映了耐药表型的可获得性及其分布。

总体而言，在独立测试集上的表现低于交叉验证估计值，准确率从 84.30% 降至 77.50%。这一差异在预期之中，因为交叉验证评估的是训练

数据分布内的表现，而独立测试集则对此前未见过的基因组提供了更严格的评估。

为直观补充表 IV 中呈现的定量结果，图 1 展示了全部 20 种抗生素的 ROC 曲线。对于环丝氨酸、氯法齐明和异烟肼，我们的方法表现出优异的判别性能，曲线下面积均大于 0.89。相比之下，德拉马尼、烟酰胺和阿米卡星的 ROC 曲线更接近对角参考线，表明其判别性能较弱。这种性能下降可能归因于训练数据中严重的类别不平衡。

C. 评估边缘设备上的资源需求

我们在具有代表性的商用边缘硬件上测量推理延迟、内存消耗、能耗和模型存储需求，以评估所提出模型进行边缘部署的可行性。该硬件为搭载 Ubuntu 24.04.5 LTS 的 Intel Core i5-1035G1 CPU（4 核、8 线程），配备 12 GB RAM。

该模型需要 16.20 MB 的存储空间，在独立测试集上的平均推理延迟为 71.57 ms。它使用 55.46 MB 的 RAM，并在对测试集所有样本进行推理时消耗 4327.23 J 能量，表明该模型适合部署在资源受限的边缘设备上，以快速进行即时检验抗微生物药物耐药性检测。

V. 结论

我们评估了一种轻量级机器学习方法，利用全基因组序列预测结核分枝杆菌的抗生素耐药性。基于紧凑的256维4-mer表示训练的随机森林分类器，在特征明确的一线抗生素上表现出稳健的预测性能，表明大幅缩减特征空间仍能保留预测能力。然而，对于新型和经重新定位用途的药物，预测性能有所下降，因为这些药物的耐药表型数据稀缺，且类别不平衡问题严重。这证实，对这些药物而言，主要限制因素是数据可用性，而非特征表示。

未来，我们将通过纳入CRyPTIC等额外的公共数据库来扩充训练集，并探索针对不平衡表型的重采样和代价敏感学习策略，以应对影响新型药物的数据稀缺问题。

参考文献

[1] 世界卫生组织，《2024年全球结核病报告》，世界卫生组织报告，2024年。

[2] L. B. Reller、M. Weinstein、J. H. Jorgensen 和 M. J. Ferraro，《抗菌药物敏感性试验：一般原则与当代实践综述》，临床传染病，第 49 卷，第 11 期，第 1749–1755 页，2009 年。[3] M. A. Salam、M. Y. Al-Amin、J. S. Pawar、N. Akhter 和 I. B. Lucy，《抗菌药物敏感性试验的传统方法与未来趋势》，沙特生物科学杂志，第 30 卷，103582，2023 年 3 月，第 . 页。[4] A. L. Haas、A. Ma、J. Pham、P. Verma、U. Malhotra、E. C. Church、M. Narita、V. Escuyer 和 S. M. Shakir，《MTB/RIF检测的局限性：4个临床病例的 Xpert 综述》，开放论坛传染病，第 12 卷，第 4 期，第 ofaf132 页，2025 年 4 月。[5] Y. Ren、T. Chakraborty、S. Doijad、L. Falgenhauer、J. Falgenhauer、A. Goesmann、A.-C. Hauschild、O. Schwengers 和 D. Heider，《基于全基因组测序和机器学习的抗微生物药物耐药性预测》，生物信息学，第 38 卷，第 2 期，第 325–334 页，2022 年 1 月。 [在线]。获取地址：https://doi.org/10.1093/ bioinformatics/btab681[6] M. I. Gr¨oschel、M. Owens、L. Freschi、R. Vargas Jr、M. G. Marin、J. Phelan、Z. I qbal、A. Dixit 和 M. R. Farhat，《Gentb：由机器学习驱动、易于使用的结核病耐药性基因组预测器》，基因组医学，第 13 卷，第 1 期，第 138 页，2021 年。[7] G. Paredes-Gutierrez、R. Perea-Jacobo、H.-G. Acosta-Mesa、E. Mezura- Montes、J. L. Morales Reyes、R. Zenteno-Cuevas、M.-´A. Guerrero- Chevannier、R. Mu˜niz-Salazar 和 D.-L. Flores，《结核分枝杆菌药物耐药性预测：基于机器学习的基因组突变分析方法》，诊断学，第 15 卷，第 3 期，第 279 页，2025 年。[8] J. E. Phelan、D. M. O'Sullivan、D. Machado、J. Ramos、Y. E. Oppong、S. Campino、J. O'Grady、R. McNerney、M. L. Hibberd、M. Viveiros 等，《整合信息学工具与便携式测序技术，快速检测抗结核药物耐药性》，基因组医学，第 11 卷，第 1 期，第 41 页，2019 年。[9] M. Hunt、P. Bradley、S. G. Lapierre、S. Heys、M. Thomsit、M. B. Hall、K. M. Malone、P. Wintringer、T. M. Walker、D. M. Cirillo 等，《利用基因组序列数据和 Mykrobe 预测结核分枝杆菌的抗生素耐药性》，*Wellcome*开放研究，第 4 卷，第 191 页，2019 年。[10] M. Serajian、S. Marini、J. N. Alanko、N. R. Noyes、M. Prosperi 和 C. Boucher，《结核分枝杆菌抗生素耐药性的可扩展从头分类》，生物信息学，第 40 卷，增刊 1，第 i39–i47 页，2024 年。[11] L˜opez-Fern´andez、Galo 等，《首个 WHO突变目录在西班牙瓦伦西亚地区结核分枝杆菌抗生素耐药性诊断中的作用：一项回顾性基因组分析》，柳叶刀微生物，2024 年。[12] A. Pal 和 D. Mohanty，《基于机器学习的方法：从结核分枝杆菌全基因组序列中识别新的耐药相关突变》，生物信息学进展，第 5 卷，第 1 期，第 vbaf050 页，2025 年。[13] P. Barua、F. Ahmed、M. A. Azim 和 M. K. Islam，《利用 k-mer 特征和多头卷积神经网络预测增强子DNA》，载于*2025*年第*28*届国际计算机与信息技术会议（*ICCIT*），2025 年，第 1–6 页。[14] D. Y. Lim、J. Khanal、H. Tayara 和 K. T. Chong，《 iEnhancer-RF：利用随机森林和增强特征表示识别增强子及其强度》，化学计量学与智能实验室系统，第 212 卷，第 104284 页，2021 年。[15] O. A. Montesinos L´opez、A. Montesinos L´opez 和 J. Crossa，《用于基因组预测的随机森林》，载于用于基因组预测的多元统计机器学习方法。Springer，2022 年，第 633–681 页。