arXiv:2602.01005v1 [cs.LG] 1 Feb 2026

利用机器学习和深度学习预测尼泊尔五岁以下儿童贫血

Deepak Bastola*[1]、Pitambar Acharya[2],Dipak Dulal[3],Rabina Dhakal[4],和 Yang Li[5]

[1, 5]佛罗里达大西洋大学，数学与统计系[2]阿拉巴马大学伯明翰分校，应用数学系[3]东新墨西哥大学，数学科学系[4]Apara Innovations

摘要

儿童贫血仍是尼泊尔面临的一项重大公共卫生挑战，并与生长发育受损、认知能力下降及发病率上升相关。依据世界卫生组织的血红蛋白阈值，我们界定了6至59个月儿童的贫血状况，并将所有严重程度的贫血合并为一类，构建了二元分类任务：贫血与非贫血。我们分析了尼泊尔人口与健康调查（NDHS 2022）的微观数据，其中包括1,855名儿童；起初考察了48项候选特征，涵盖人口统计、社会经济、孕产妇及儿童健康特征。为获得稳定且有充分依据的特征集，我们采用了四种特征选择方法（卡方检验、互信息、点二列相关和 Boruta），并优先纳入获得多方法共识支持的特征。儿童年龄、近期发热、家庭规模、孕产妇贫血和寄生虫驱虫这五项特征均被所有方法一致选中；闭经、族裔指标和省份也经常入选。随后，我们使用标准评估指标，对比了八种传统机器学习分类器（LR、KNN、DT、RF、XGBoost、SVM、NB、LDA）和两种深度学习模型（DNN 和 TabNet）；考虑到类别不平衡，重点关注 F1分数和召回率。在所有模型中，逻辑回归的召回率最高（0.701），F1分数也最高（0.649）；DNN 的准确率最高（0.709），而 SVM 的区分能力最强，曲线下面积最高（0.736）。总体而言，结果表明，机器学习和深度学习模型均可有效预测贫血；儿童年龄、感染替代指标、孕产妇贫血和驱虫史等可解释特征，对于尼泊尔的风险分层和公共卫生筛查至关重要。

关键词： 儿童贫血、尼泊尔人口与健康调查、机器学习、深度学习、尼泊尔

引言

贫血的特征是红细胞数量减少，或血细胞中的血红蛋白浓度降低。根据世界卫生组织分类，贫血严重程度依据血红蛋白水平划分：无贫血（$\geq$ 11.0 g/dL）、轻度贫血（10.0-10.9 g/dL）、中度贫血（7.0-9.9 g/dL）、重度贫血（¡7.0 g/dL） [39]。全球估计数据显示，15 至 49 岁女性中有 30%、所有孕妇中有 37%、6 至 59 个月儿童中有 40% 患有贫血 [39]。全球疾病负担研究估计，截至 2021 年，全球有 19 亿人受到贫血影响，其中 5 岁以下儿童、女性以及撒哈拉以南非洲和南亚等地区的人群受影响最为严重。5 岁以下儿童受影响最为普遍，全球患病率为 41.4%。当年南亚的贫血患病率为 43% [12]。儿童期贫血的后果包括认知和

*通讯作者：dbastola2022@fau.edu

运动发育受损、营养不良、生长迟缓以及发病率上升，这凸显了尽早干预的重要性 [39]。

在尼泊尔，儿童贫血仍是一个重大的公共卫生问题。根据《2022年尼泊尔人口与健康调查》（NDHS），约有43%的5岁以下儿童患有贫血。 [26]历史数据进一步凸显了令人担忧的趋势：6至59月龄贫血儿童的比例从2011年的46.4%上升至 2016 [25, 27]年的52.2%。尽管2022年这一年龄组的贫血率有所改善，但患病率仍然很高，构成了重大的公共卫生挑战。儿童贫血的负担在特定人群中尤为突出，因为贫困家庭、边缘化族群以及尼泊尔某些生态区域受到的影响尤为严重 [33]。同样，孕产妇贫血、孕期营养不良、产前保健不足和出生顺序等母亲因素，也与儿童贫血密切相关 [37, 7, 21]。相关研究还指出，年龄较小、营养不良、缺铁或感染蠕虫、居住在特莱地区以及属于达利特族群的儿童，贫血风险较高。地理差异也十分显著：特莱平原的贫血患病率明显高于丘陵和山区，这种差异通常归因于粮食不安全、气候变化以及卫生资源分布不均 [14]。高质量、具有全国代表性的数据集（如NDHS）的可用性，为大规模研究贫血提供了前所未有的机会。NDHS提供了涵盖人口、社会经济和健康指标的综合数据，可用于了解人群层面的贫血模式及其决定因素。此前的研究主要采用传统统计方法，考察尼泊尔贫血的患病率及相关因素 [21, 11]。然而，随着计算能力的提升，数据科学、机器学习（ML）和深度学习（DL）方法如今提供了有力的替代方案，能够对人口调查数据中的复杂非线性关系进行建模 [24]。这些方法能够开展高维分析，捕捉特征之间隐藏的相互作用。

尽管针对尼泊尔儿童贫血开展了大量流行病学研究，但迄今尚无研究使用 NDHS 数据，采用机器学习或深度学习方法进行预测建模。本研究通过系统评估十种机器学习和深度学习算法，预测尼泊尔 6 至 59 个月龄儿童的贫血状况，弥补了这一关键空白。我们使用分层交叉验证，严格比较传统分类器与神经网络架构的模型表现，并通过 SMOTE 处理类别不平衡，从而确定适用于资源有限环境下社区贫血筛查的最佳计算方法。我们的框架将分析范式从单纯解释关联转向可操作的风险预测，并表明健康调查中常规收集的社会人口学变量能够有效地按贫血风险对儿童进行分层。研究结果为开发实用的筛查工具奠定了基础，这些工具有望为有针对性的母婴健康干预提供依据，从而降低尼泊尔的贫血患病率及其对发育造成的后果。

数据集

数据预处理

数据预处理包括删除不需要的条目、剔除目标值缺失的观测记录，以及排除六个缺失值较多的变量（婴儿配方奶粉、果汁摄入、粪便处理、产后检查、首次母乳喂养时长、维生素A）。产前保健次数上限设为10次，并采用中位数插补；母乳喂养、孕期补铁和母亲驱虫则采用众数插补。根据人体测量指标创建了一个二元营养不良变量：按照世界卫生组织标准，只有所有 Z 分数均介于 -2 和 $+2$ 之间时，才将儿童归类为营养状况良好 [29]。预处理后，从原始的5,372条记录中保留了1,855条观测记录和13个分类特征。本研究将所有特征转换为分类变量。在模型实施过程中，本研究采用分层抽样，将数据按80：20划分为训练集和测试集，并对训练集应用合成少数类过采样技术（SMOTE），将其整合到流水线中以避免数据泄漏；此外，还进行了重复分层五折交叉验证，重复3次。

特征选择

响应变量为儿童贫血，根据世界卫生组织针对儿童制定的阈值，通过血红蛋白水平进行定义。本研究将任何阶段的贫血均归类为贫血，并旨在进行二元分类

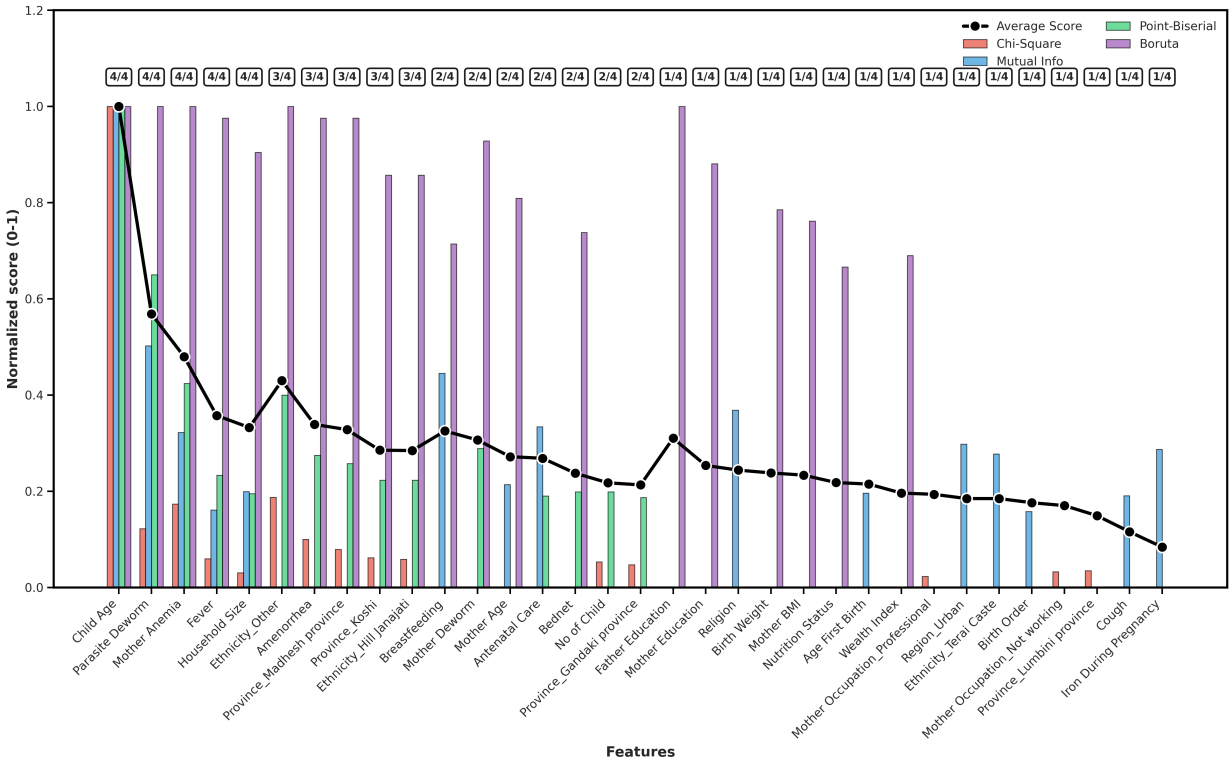


图1：各方法中归一化的特征重要性

，分为贫血和非贫血两类，分别涉及 1,085 名和 770 名儿童。首先，根据类似研究选取了 48 个候选特征，并采用四种不同的统计方法：卡方检验、互信息、点二列相关和 Boruta 算法，从候选集 [42] 中稳健地识别出 NDHS数据中最能反映贫血状况的特征。

图1显示，四种方法一致选出的五个特征为：儿童年龄、发热、家庭规模、母亲贫血和寄生虫驱虫，表明这些特征与儿童贫血密切相关。另外五个特征——闭经、族群（丘陵贾纳贾蒂人、其他）和省级指标（科西省、马德西省）——被三种方法选中。两种方法选中的变量，包括产前保健、母乳喂养和母亲驱虫，则根据近期尼泊尔人口与健康调查研究中已确立的重要性予以保留 [37, 21]。这种多方法共识策略可减少单一技术带来的偏差，并增强对最终特征集的信心。

方法

支持向量机

支持向量机旨在寻找一个最优分离超平面，其定义为 $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$，以最大化类别间的间隔。该优化问题旨在最小化 $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$，并满足条件 $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0$。核函数可通过将数据隐式映射到更高维空间，实现非线性分类 [8]。

决策树

决策树是一种监督学习模型，通过决策规则递归地将特征空间划分为互不相交的区域。给定数据 $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$，特征空间被划分为多个区域 $\{R_k\}_{k=1}^{K}$，并通过 $\hat{y}(\mathbf{x}) = \sum_{k=1}^{K} c_k \mathbb{I}(\mathbf{x} \in R_k)$ 进行预测，其中 c_k 表示分配给区域 R_k[5] 的类别标签。



图 2：工作流图

随机森林

随机森林是一种集成学习方法，它使用自助样本，并在每次分裂时进行随机特征选择，从而构建多棵决策树。最终预测通过多数投票得出：

$\hat{}y_{\mathrm{RF}}(\mathbf{x})$ = 模式$\{T_1(\mathbf{x})\ldots,\ T_N(\mathbf{x})\}$。该策略可降低方差并改善泛化能力，尤其适用于高维和噪声场景 [4]。

K近邻

K近邻（KNN）是一种非参数监督学习算法，根据特征空间中 k 最近邻样本的多数标签对样本进行分类 [9]。给定一个带标签数据集$\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$，使用欧几里得度量计算距离 $d(\mathbf{x},\ \mathbf{x}_i) = \sqrt{\sum_{k=1}^{N}(x_k - x_{ik})^2}$。预测类别通过 $\hat{}y = \arg\max_c \sum_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} \mathbb{I}(y_i = c)$获得。

逻辑回归

逻辑回归（LR）将二元结果的概率建模为 $p(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$，其中$\sigma(\cdot)$是Sigmoid函数。通过最小化负对数似然$L(\mathbf{w}\ b) = -\sum_{i=1}^{N}[y^{(i)} \log\ p^{(i)} + (1 - y^{(i)}) \log(1 - p^{(i)})]$, 来估计模型参数，并可选择使用正则化以防止过拟合。

深度神经网络

深度神经网络（深度神经网络）通过纳入多个隐藏层来扩展人工神经网络架构，以学习层次化表示。对于输入$\mathbf{x}$，前向传播定义为$\mathbf{h}^{(0)} = \mathbf{x}$ 和 $\mathbf{h}^{(\ell)} = \sigma^{(\ell)}(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)})$，从而能够提取高层抽象特征 [30]。

TabNet

TabNet是一种专为表格数据设计的深度学习架构，采用序列注意力来执行逐实例特征选择。在每个决策步骤中，稀疏掩码 $M_i \in \mathbb{R}^{B\times D}$ 被应用于输入~特征 f，从而产生 $f_i = M_i \odot f$。该掩码通过 $M_i = \mathrm{sparsemax}(P_{i-1} \odot h_i(a_{i-1}))$生成，借助稀疏特征使用促进可解释性。正则化项对过度特征使用加以惩罚，以促进稀疏性 [2]。

线性判别分析

线性判别分析（线性判别分析）是一种有监督降维技术，通过求解广义特征值问题 $S_B w = \lambda S_W w$来最大化类别可分性，其中 S_B 和 S_W 表示类间散布矩阵与类内散布矩阵。所得投影增强了类别判别，对结构化的社会经济及人口统计数据集尤为有效。

朴素贝叶斯

朴素贝叶斯（NB）是一种基于贝叶斯定理的概率分类器，并假设各特征之间条件独立。后验概率表示为 $P(C_k|\mathbf{x}) = {}^{P}{}_{(C_k)} \prod_{i=1}^{n} P(x_i|C_k)\, P(\mathbf{x})$。尽管 NB 简单，但它在文本分类和分类数据分类等高维场景中仍具有较强的竞争力 [3]。

极端梯度提升

XGBoost 是一种集成提升方法，通过梯度下降优化并构建加性决策树。目标函数 $\mathcal{L}(\theta) = \sum_{i=1}^{n} L(y^{(i)}, f(x^{(i)};\ \theta)) + \Omega(T) + R(w)$ 综合考虑损失最小化和正则化，以控制模型复杂度，从而实现出色的预测性能，并增强对过拟合的鲁棒性。

评估指标

表1概述了用于评估所提出模型性能的评估指标。这些指标有助于系统地比较机器学习和深度学习模型，从而能够.....

对其在数据集中区分贫血病例与非贫血病例这一分类任务中的优势和局限性进行详细分析。

S.N.	指标	公式	描述
1	混淆矩阵		A 通过统计模型表现进行汇总的矩阵— 正确和错误分类的贫血与非 贫血病例分别为 TP , TN, FP, 和 FN。
2	准确率	$\frac{TP+TN}{TP+TN+FP+FN}$	它衡量贫血预测中 模型准确识别为贫血或非贫血的比例 贫血。
3	精确率	$\frac{TP}{TP+FP}$	它衡量被预测为贫血病例中 真正为贫血者所占的比例。
4	召回率	$\frac{TP}{TP+FN}$	它衡量模型正确识别贫血的能力 mic个体。高召回率意味着低假阴性。
5	F1分数	$\frac{2（精确率）（召回率）}{精确率+召回率}$	精确率与召回率的调和平均数，提供一个 平衡的性能度量。
6	平均精确率	$\sum_{k=1}^{N} P(k)\ [R(k)\ - R(k-1)]$	它衡量精确率-召回率曲线下的面积, 用于表明模型在不同 召回率水平。
7	AUC	$\Pr(y_i > y_j)$	模型赋予更高分数的概率 在所有阈值下，赋予贫血病例的分数高于非贫血病例 。
8	Cohen's Kappa	$\frac{p_0 - p_e}{1 - p_e}$	衡量模型与真实标签的一致性超出随机机会的程度, 并对类别不平衡进行调整 ance.

表1：性能指标概览

结果

描述性统计

表2显示，6–59个月儿童的总体贫血患病率为38.5%。贫血患病率随儿童年龄增长而下降。6–12个月儿童的贫血患病率最高（75.3%）。与6–12个月的婴儿相比，25–36个月儿童患贫血的几率低72%（AOR = 0.28；95% CI：0.17- 0.46； $p < 0.001$）。同样，37–48个月（AOR = 0.20；95% CI：0.12-0.32； $p < 0.001$）和49–59个月儿童（AOR = 0.14；95% CI：0.08-0.22； $p < 0.001$）患贫血的几率也显著较低。与未服用驱虫药片的儿童相比，服用过驱虫药片的儿童患贫血的几率较低（AOR =0.65；95% CI：0.48–0.88； $p = 0.006$）。与没有发热的儿童相比，在调查前两周内发过热的儿童患贫血的几率较高（AOR = 1.41；95% CI：1.12-1.77； $p = 0.003$）。

一些母亲因素也与儿童贫血有关。与非贫血母亲的儿童相比，贫血母亲的儿童发生贫血的几率高出 1.62 倍（AOR = 1.62；95% CI：1.30-2.02； $p < 0.001$）。与母乳喂养儿童相比，未接受母乳喂养的儿童发生贫血的几率更高（AOR = 1.27；95% CI：0.83-1.93； $p = 0.28$）；然而，这种关联在多变量回归分析中并不显著。母亲在妊娠期间接受过驱虫治疗的儿童，发生贫血的几率低于母亲未接受驱虫治疗的儿童（AOR = 1.06；95% CI：0.71-1.58；$p = 0.795$）。

不同省份、家庭规模和族群等人口统计学因素下，儿童贫血患病率存在差异。马德什省儿童贫血患病率最高（50.1%）。与马德什省儿童相比，科西省儿童患贫血的几率显著较低

特征	类别	贫血	非贫血	总患病率（%）		粗比值比	调整比值比	调整95%置信区间	p
儿童年龄	6–12	125	41	166	75.30	1.00			
	13–24	213	134	347	61.38	0.52	0.642	0.404–1.021	0.061
	25–36	160	226	386	41.45	0.23	0.279	0.169–0.458	<0.001
	37–48	154	332	486	31.69	0.15	0.196	0.121–0.317	<0.001
	49–59	118	352	470	25.11	0.11	0.137	0.084–0.223	<0.001
发热	No	540	835	1375	39.27	1.00			
	Yes	230	250	480	47.92	1.42	1.408	1.119–1.772	0.003
家庭规模	中等	415	568	983	42.22	1.00			
	大	100	104	204	49.02	1.32	1.306	0.933–1.829	0.120
	小	255	413	668	38.17	0.85	1.040	0.832–1.300	0.731
母亲贫血	非贫血	448	776	1224	36.60	1.00			
	贫血	322	309	631	51.03	1.81	1.621	1.299–2.022	<0.001
寄生虫驱虫	No	231	139	370	62.43	1.00			
	Yes	539	946	1485	36.30	0.34	0.650	0.479–0.882	0.006
闭经	No	668	1001	1669	40.02	1.00			
	Yes	102	84	186	54.84	1.82	1.119	0.783–1.599	0.536
族群	丘陵婆罗门/切特里人	211	366	577	36.57	1.00			
	丘陵达利特人	96	152	248	38.71	1.10	1.210	0.861–1.699	0.272
	丘陵贾纳贾蒂人	126	242	368	34.24	0.90	0.982	0.703–1.372	0.915
	其他	200	167	367	54.50	2.08	1.974	1.391–2.803	<0.001
	特莱种姓	137	158	295	46.44	1.50	1.449	0.961–2.185	0.077
省	马德西省	176	175	351	50.14	1.00			
	巴格马蒂省	83	130	213	38.97	0.64	0.942	0.589–1.507	0.803
	甘达基省	48	105	153	31.37	0.46	0.707	0.423–1.181	0.186
	卡纳利省	114	180	294	38.78	0.63	0.902	0.569–1.430	0.661
	科西省	96	194	290	33.10	0.49	0.584	0.392–0.869	0.008
	蓝毗尼	132	143	275	48.00	0.92	0.978	0.668–1.432	0.910
	远西省	121	158	279	43.37	0.76	0.962	0.623–1.485	0.861
产前保健	充分产前保健	716	1040	1756	40.77	1.00			
	产前保健不足	54	45	99	54.55	1.74	0.727	0.458–1.154	0.177
母乳喂养	Yes	707	1022	1729	40.89	1.00			
	No	63	63	126	50.00	1.45	1.265	0.828–1.933	0.277
母亲驱虫	Yes	688	1025	1713	40.16	1.00			
	No	82	60	142	57.75	2.04	1.06	0.71–1.58	0.795

表2：与儿童贫血相关的因素

（AOR $= 0.58$；95% CI：0.39–0.87；$p = 0.008$）。同样，“其他”民族群体儿童患贫血的几率几乎是丘陵婆罗门/切特里人儿童的两倍（AOR $= 1.97$；95% CI：1.39–2.80；$p < 0.001$）。不同家庭规模的儿童贫血患病率略有差异；但在多变量回归模型中，这种关联并无统计学意义。

模型性能

表 3 展示了八种机器学习模型和两种深度学习模型的表现。鉴于类别不平衡以及识别贫血儿童在临床上的重要性，F1分数和召回率被优先选为主要评估指标。

逻辑回归的整体表现更佳，F1分数最高（64.9%），召回率也最高（70.1%），因此在正确识别贫血病例方面最为有效。支持向量机和深度神经网络的表现也颇具竞争力，F1分数分别为63.6%和63.3%。支持向量机的曲线下面积最高（73.5%），表明其区分贫血儿童与非贫血儿童的能力更强。深度神经网络虽然取得了最高的准确率（70.9%）和精确率（66.4%），但其召回率（60.4%）低于逻辑回归，说明它漏掉了更多真实的贫血病例。随机森林和XGBoost的准确率相同，均为

69.0%，但召回率较低，因此在这一类别不平衡分类任务中的有效性有限。K近邻在所有指标上的表现均最差（F1分数：57.0%），而决策树的曲线下面积最低（66.4%），并呈现出过拟合迹象。大多数模型的训练集与测试集F1分数基本一致（约为60%–63%），表明其泛化能力良好。

图3展示了表现最佳的三个模型——逻辑回归、支持向量机和深度神经网络——的 ROC-AUC 曲线。所有三个模型在区分贫血儿童与非贫血儿童方面均展现出相当的判别能力，其曲线明显高于随机分类器的对角线。所有模型的训练集与测试集 AUC 得分高度一致，有力地证明了模型具有稳健的泛化能力且未发生过拟合，同时验证了我们的交叉验证策略和 SMOTE 在处理类别不平衡问题方面的有效性。

	机器学习								深度学习	
指标	LR	KNN	DT	RF	XGB	SVM	NB	LDA	DNN	TabNet
准确率	0.685	0.625	0.644	0.69	0.69	0.679	0.668	0.666	**0.709**	0.663
精确率	0.603	0.544	0.56	0.636	0.633	0.601	0.594	0.586	**0.664**	0.582
召回率	**0.701**	0.597	0.662	0.591	0.604	0.675	0.636	0.662	0.604	0.669
F1分数	**0.649**	0.57	0.607	0.613	0.618	0.636	0.614	0.622	0.633	0.622
平均精确率	**0.679**	0.577	0.598	0.663	0.667	**0.684**	0.676	**0.684**	0.671	0.628
AUC	**0.733**	0.677	0.664	0.722	0.724	**0.736**	**0.733**	**0.733**	**0.731**	0.709
Cohen's Kappa	0.365	0.239	0.286	0.355	0.357	0.351	0.324	0.324	**0.393**	0.321
训练集F1分数	0.614	0.634	0.634	0.607	0.606	0.611	0.588	0.609	0.607	0.607

表3：模型性能

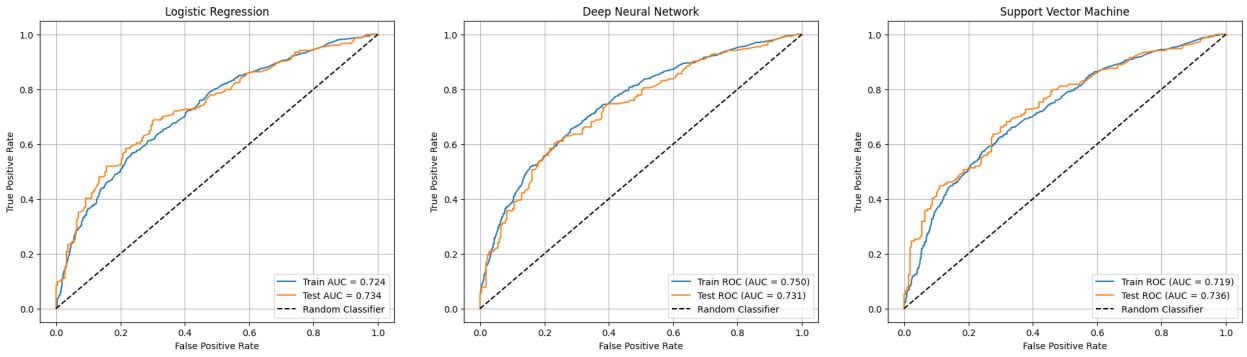


图3：训练集和测试集的ROC曲线

讨论

本研究结果表明，机器学习算法能够预测尼泊尔5岁以下儿童是否患有贫血，进一步支持了数据驱动建模方法可补充传统流行病学方法的证据 [31, 18]。我们的研究显示出明显的年龄相关趋势：年龄较小的儿童比年龄较大的儿童更容易贫血。随着年龄增长，贫血发生几率持续下降，这可能反映了膳食多样性的改善、免疫系统的成熟，以及对微量营养素缺乏的易感性降低 [40]。同样，本研究表明，儿童贫血与寄生虫驱虫、孕产妇驱虫、所在省、近期发热及其他族群等因素有关，这与既往研究结果一致 [37, 7, 21]。鉴于我们的数据集存在类别不平衡，我们将F1分数和召回率作为主要评估指标。F1分数综合平衡了精确率与召回率，而召回率则专门衡量模型正确识别贫血病例的能力

即在所有实际患有贫血的儿童中正确识别病例的能力。这对于公共卫生筛查至关重要，因为漏诊贫血病例会带来严重的临床影响。本研究中相对较低的精确率凸显了对假阳性病例进行确认性血红蛋白检测的必要性，这一点也得到采用其他筛查方法的研究所强调 [20, 32]。

结果表明，机器学习和深度学习方法的预测性能相当。逻辑回归表现最佳，F1分数（64.9%）和召回率（70.1%）均为最高，其次是支持向量机（F1分数：63.6%，召回率：67.5%）和深度神经网络（F1分数：63.3%，召回率：60.4%）。逻辑回归表现优异，得益于其凸优化特性，可确保收敛至全局最小值；此外，其参数空间有限，可降低过拟合风险 [15, 13]。在传统流行病学研究中，逻辑回归也一直是首选方法，用于确定低收入和中等收入国家的人口与健康调查数据中各变量与贫血预测之间的关联[7]。训练集 F1分数（61.4%）与测试集 F1分数（64.9%）相近，表明模型具有稳定的泛化能力，且未出现过拟合。

深度神经网络表现出较强的竞争力，准确率（70.9%）和精确率（66.4%）均为最高，因为它能够对复杂的非线性决策边界和特征交互进行建模 [22, 41]。本研究的特征空间相对较小（13 个特征），减轻了维度灾难并有助于提升泛化能力，这一点也得到了训练集（60.7%）和测试集（63.3%）F1分数稳定一致的印证。我们采用 SMOTE，通过合成扩增少数类来解决类别不平衡问题，从而帮助模型学习更均衡的决策边界，并提高对贫血病例的灵敏度 [6, 16]。随机森林、XGBoost、支持向量机和线性判别分析在多个指标上均表现稳定且出色（准确率：66.6-69.0%，曲线下面积：72.2-73.5%），表明多种稳健的学习算法均能基于给定的特征表示实现有效泛化。

K近邻的表现最差（F1：57.0%，召回率：59.7%），这很可能是因为SMOTE会扭曲基于距离的学习器所依赖的局部邻域结构 [6], ，并且以分类特征为主会造成距离度量的模糊性 [17, 1]。TabNet的表现也较差（F1：62.2%），因为其针对复杂特征交互设计的架构在小样本量（n = 1,855）以及以类别数据为主的情况下优势有限 [2, 34]。

我们的模型所观察到的中等性能水平（各项关键指标约为60-70%）符合预期，因为所使用的是间接的社会人口学和家庭层面特征，而非直接的临床测量。使用临床和实验室数据的研究报告了显著更高的贫血分类准确率（¿96%），因为这些数据源提供了血红蛋白水平和铁状态的直接生理指标 [23, 10]。相比之下，我们的方法依赖于儿童年龄、近期发热、家庭规模、族群和驱虫等代理变量，这些变量捕捉的是上游社会决定因素，而非即时生物标志物。特征类型上的这一根本差异为社会人口学模型设定了预期的性能上限。然而，此类基于无创预测的方法，可能为尼泊尔等因资源限制而筛查稀缺的中低收入国家提供一种更经济的贫血筛查途径 [35]。无创方法处于中间位置：基于指甲图像的方法准确率约为70% [28], ，使用血液学特征的弹性网络逻辑回归在区分缺铁性贫血与 β-地中海贫血性状方面达到97%的准确率 [23], ，基于面部图像的筛查则显示出84%的准确率和93%的灵敏度 [43]。

基于 DHS 数据的机器学习研究表明，在不同国家，模型表现会因样本量、特征和类别可分性而异。一项孟加拉国 DHS 2011 研究（$n = 2,$ 013 个样本，包含 24 个变量）报告称，随机森林在预测儿童贫血方面的准确率为 68.5%、召回率为 70.7%、精确率为 66.4% [19], ，与我们的结果相当。埃塞俄比亚 DHS 分析（$n = 8,$ 482 个样本，包含 13 个特征）发现，逻辑回归是表现最佳的方法，准确率为 66%、召回率为 82% [36], ，与我们的研究结果高度一致。不同 DHS 研究得出的表现范围相近，进一步表明社会人口学特征能够稳定地产生准确率为 60%–75% 的模型，这反映了间接指标固有的预测局限性。

尽管样本量较小（n =1，855），特征也较少（13 个特征），我们基于 NDHS 的分析仍展现出有竞争力的预测性能。以往使用 NDHS 的研究侧重于贫血的空间模式和决定因素 [33, 21], ，而我们的研究则独特地将机器学习应用于预测。研究结果表明，可利用易于获取的特征开发基于社区的风险评分，包括儿童年龄、近期发热、孕产妇贫血、家庭规模、省、驱虫状况、闭经、产前保健充分性以及母乳喂养做法，这与使用 DHS 数据集的研究一致 [7, 21, 18]。尽管

这些模型无法取代实验室确诊，但在无法进行常规血红蛋白检测的环境中，可为初步筛查和资源分配提供有价值的工具 [35]。被识别为高风险的儿童可转介接受血红蛋白检测，随后实施有针对性的干预措施，包括铁剂补充和发热预防。我们的结果强调了采取综合母婴健康措施的重要性，因为母亲因素（贫血状况、妊娠期间驱虫以及充分的产前保健）会显著影响儿童贫血风险 [7, 21]。

局限性

排除缺失值后，我们的分析保留了1,855条观测数据，这可能限制结果的普遍适用性。模型表现中等（F1分数：64.9%），反映出社会人口学预测因子的固有限制；这些模型可用作筛查工具，但仍需进行确认性血红蛋白检测，不能替代诊断。横断面设计无法进行因果推断，而由SMOTE生成的合成样本可能会引入人为模式。

结论

本研究使用NDHS 2022数据，对十种用于贫血预测的机器学习和深度学习算法进行了系统基准评估，作出了方法学上独特的贡献。我们整合了涵盖儿童人口统计特征、孕产妇健康和家庭特征的13项特征，并采用分层划分、SMOTE处理类别不平衡，以及通过GridSearchCV和重复分层k折交叉验证进行严格的超参数优化。逻辑回归表现最佳，F1分数为64.9%，召回率为70.1%。各模型的表现均稳定在60%至70%之间，这反映了社会人口学代理指标相较于直接临床测量的预测能力，且与类似的人口与健康调查研究相比仍具竞争力。该框架可利用常规健康调查中易于获取的信息进行早期风险分层，从而在无法普及血红蛋白检测的情况下开展有针对性的筛查。社区卫生工作者可利用风险评分识别需要接受确认性检测的儿童，从而优化铁剂补充和营养干预的资源分配。将该框架整合到母婴健康服务中，有助于改善产前保健并支持家庭层面干预的数据驱动决策 [42, 38]。

未来工作

未来研究应整合多个尼泊尔人口与健康调查（NDHS）轮次，以扩大样本量并开展时间验证。纳入膳食多样性、微量营养素依从性、蠕虫负担和地理信息系统等补充数据，或可提高准确率。开发适用于移动健康平台的可解释风险评分，将有助于在社区推广；同时，将研究扩展至多分类严重程度预测，可支持差异化诊疗路径和有针对性的干预措施。

利益冲突

作者声明不存在利益冲突。

数据和代码可用性

本研究分析的数据集可通过人口与健康调查（DHS）资料库公开获取（https://dhsprogram.com/data/）。有意获取本分析所用代码的研究人员可向通讯作者提出申请。

参考文献

[1] H.A.A. Alfeilat、A.B.A. Hassanat、O. Lasassmeh、A.S. Tarawneh、M.B. Alhasanat、H.S. Eyal Salman和V.B.S. Prasath。距离度量选择对K近邻分类器性能的影响：综述。*Big Data*，7(4):221–248，2019。

.[2] Sercan O. Arik和Tomas Pfister。TabNet：注意力可解释表格学习。AAAI人工智能大会论文集，35(8):6679–6687，2021。[3] Christopher M Bishop和Nasser M Nasrabadi。模式识别与机器学习，第4卷。Springer，2006。[4] Leo Breiman。随机森林。机器学习，45(1):5–32，2001。[5] Leo Breiman、Jerome Friedman、Richard A Olshen和Charles J Stone。分类与回归树。Chapman and Hall/CRC，2017。[6] Nitesh V. Chawla、Kevin W. Bowyer、Lawrence O. Hall和W. Philip Kegelmeyer。SMOTE：合成少数类过采样技术。人工智能研究杂志，16:321–357，2002。[7] Mohammad Rocky Khan Chowdhury、Md. Mobarak Hossain Khan、Hafiz T.A. Khan、Md. Shafiur Rahman、Md. Rashedul Islam、Md. Moinul Islam和Baki Billah。尼泊尔儿童贫血的患病率与危险因素：多层次分析。PLOS ONE，15(10):e0239409，2020。[8] Corinna Cortes和Vladimir Vapnik。支持向量网络。机器学习，20(3):273–297，1995。[9] Thomas Cover和Peter Hart。最近邻模式分类。IEEE信息论汇刊，13(1):21–27，1967。[10] Prakriti Dhakal、Santosh Khanal和Rabindra Bista。使用机器学习算法进行贫血预测。国际计算机科学与信息技术杂志，15:15–30，02 2023。[11] Ram Prasad Dhakal。尼泊尔人口与健康调查（NDHS）2006、2011和2016年流行病学研究综述。TUTA期刊，第153–158页，2022。[12] William M Gardner、Christian Razo、Theresa A McHugh、Hailey Hagins、Victor M Vilchis-Tella、Conor Hennessy、Heather Jean Taylor、Nandita Perumal、Kia Fuller、Kelly M Cercy等。1990–2021年按严重程度和病因划分的贫血疾病负担患病率、伤残生存年数及趋势：2021年全球疾病负担研究结果。《柳叶刀·血液学》，10(9):e713–e734，2023。[13] Ian Goodfellow、Yoshua Bengio和Aaron Courville。深度学习。MIT出版社，马萨诸塞州剑桥，2016。[14] Kassandra L Harding、Vıctor M Aguayo、Grace Namirembe和Patrick Webb。尼泊尔和巴基斯坦妇女和儿童贫血决定因素：近期全国调查数据分析。母婴营养，14:e12478，2018。[15] Trevor Hastie、Robert Tibshirani和Jerome Friedman。统计学习的要素：数据挖掘、推断与预测。Springer，纽约，第2版，2009。[16] Haibo He和Edwardo A. Garcia。从不平衡数据中学习。IEEE知识与数据工程汇刊，21(9):1263–1284，2009。[17] Liang-Yi Hu、Ming-Wen Huang、Shih-Wei Ke和Chih-Fong Tsai。距离函数对医学数据集k近邻分类的影响。SpringerPlus，5，2016。[18] Abdulaziz Kebede Kassaw、Ali Yimer、Wondwosen Abey、Tibebu Legesse Molla和Alemu Birara Zemariam。应用机器学习方法确定埃塞俄比亚五岁以下儿童贫血预测因素。《科学报告》，13(1):22919，2023。[19] Jahidur Rahman Khan、Srizan Chowdhury、Humayera Islam和Enayetur Raheem。用于预测孟加拉国儿童贫血的机器学习算法。数据科学杂志，17(1):195–218，2022。[20] Abhishek Kumar、Rakesh Singh和Priya Sharma。使用机器学习方法进行血红蛋白检测。见计算与通信进展国际会议论文集，第115–126页。Springer，尚，2025。[21] Yasutaka Kuniyoshi。孕产妇贫血作为尼泊尔儿童贫血的预测因素：人口与健康调查分析。Cureus，17(8)，2025。[22] Yann LeCun、Yoshua Bengio和Geoffrey Hinton。深度学习。《自然》，521(7553):436–444，2015。[23] Salma Mahmood、Asaad Abdulameer Khalaf和Saad S Hamadi。巴士拉评分：一种利用红细胞指数区分缺铁性贫血与β地中海贫血携带状态的新型基于机器学习的评分。Frontiers in Big Data，8:1634133，2025。[24] Abraham Keffale Mengistu。理解埃塞俄比亚药浸蚊帐使用的决定因素：一项ma-

基于PMA埃塞俄比亚2023年调查数据的机器学习分类方法。PLOS ONE，20(7):e0327800，2025。[25] 尼泊尔卫生与人口部（MOHP） [尼泊尔] 和 New ERA、尼泊尔及 ICF International。尼泊尔2011年人口与健康调查：最终报告。技术报告，DHS项目，ICF International，尼泊尔加德满都，2012年。访问日期：2026年2月3日。[26] 卫生与人口部 [尼泊尔] 和 New ERA 及 ICF。尼泊尔2022年人口与健康调查：摘要报告。摘要报告（sr275），DHS项目，ICF，尼泊尔加德满都，2023年。[27] 尼泊尔卫生部（MOH）和 New ERA [尼泊尔] 及 ICF。尼泊尔2016年人口与健康调查：最终报告。技术报告，DHS项目，ICF，尼泊尔加德满都，2017年。访问日期：2026年2月3日。[28] Jorge Raul Navarro-Cabrera、Miguel Angel Valles-Coral、Mar´ıa Elena Farro-Roque、Nelly Re´ategui- Lozano 和 Lolita Ar´evalo-Fasanando。利用指甲图像的机器视觉模型：对大学生缺铁性贫血进行无创检测。《Frontiers in Big Data》，第8卷，2025，2025年。[29] 世界卫生组织。《世卫组织儿童生长标准技术报告》，日内瓦：世卫组织出版社，2006年。世卫组织出版社，日内瓦，2006年。[30] David E Rumelhart、Geoffrey E Hinton 和 Ronald J Williams。通过反向传播误差学习表征。《自然》，323:533–536，1986年。[31] Nahid Salma 和 Majid Khan Majahar Ali。利用机器学习算法评估影响因素并预测儿童贫血。《马来西亚基础与应用科学杂志》，21(1):1529–1541，2025年。[32] Neha Sharma 和 Anil Gupta。学龄儿童贫血检测的临床体征：可靠性及确认性血红蛋白检测的必要性。《国际当前药学评论与研究杂志》，17(1):100–105，2025年。[33] Shristi Sharma、Bipin Kumar Acharya 和 Qian Wu。尼泊尔五岁以下儿童贫血的空间差异及决定因素：人口与健康调查（2006–2016）。《国际环境研究与公共卫生杂志》，19(14):8664，2022年。[34] Ravid Shwartz-Ziv 和 Amitai Armon。表格数据：深度学习并非你所需的一切。《信息融合》，2022年。[35] Shyamkumar Sriram 和 Soni Sharma。提升印度贫血诊断的水平与可及性：有效贫血管理的政策建议。Frontiers in Health Services，5:1529094，2025年。[36] Solomon Hailemariam Tesfaye、Binyam Tariku Seboka 和 Daniel Sisay。应用机器学习方法预测儿童贫血：对埃塞俄比亚2016年人口与健康调查的分析。PLOS ONE，19:1–14，2024年4月。[37] Hari Prasad Upadhyay、Nirajan Budhathoki、Ram Prasad Khatiwada 和 Bijay Lal Pradhan。尼泊尔6至23个月龄儿童贫血的决定因素：一种替代性贝叶斯建模方法。《BMC公共卫生》，25:3329，2025年。发表于2025年10月3日。[38] Henriette Usanzineza、Etienne Nsereko、Jean Pierre Niyitegeka、Aline Uwase、Jean de Dieu H Tuyishime、Francois Xavier Sunday、Christian Mazimpaka 和 Jeanine Ahishakiye。卢旺达儿童贫血的患病率及危险因素：基于卢旺达2019—2020年人口与健康调查。Public Health Challenges，3(1):e159，2024年。[39] 世界卫生组织。贫血诊断及严重程度评估的血红蛋白浓度。技术报告 WHO/NMH/NHD/MNM/11.1，世界卫生组织，日内瓦，2011年。网址：https://www.who.int/publications/i/item/WHO-NMH-NHD-MNM-11.1。[40] 世界卫生组织。《婴幼儿及儿童每日补铁：指南》。世界卫生组织，瑞士日内瓦，2016年。访问日期：2025年12月15日。[41] Han-Jia Ye、Si-Yang Liu、Hao-Run Cai、Qi-Le Zhou 和 De chuan Zhan。深入探讨表格数据集上的深度学习方法。载于 ACL 2024会议论文集，2024年。[42] Alemu Birara Zemariam、Ali Yimer、Gebremeskel Kibret Abebe、Wubet Tazeb Wondie、Biruk Beletew Abate、Addis Wondmagegn Alamaw、Gizachew Yilak、Tesfaye Masreshaw Melaku 和 Habtamu Setegn Ngusie。运用监督式机器学习算法对贫血进行分类和预测

埃塞俄比亚青年女性。《科学报告》，14(1):9080，2024年4月。[43] Aixian Zhang、JingjiaoLou、ZijiePan、Jiaqi Luo、Xiaomeng Zhang、Han Zhang、Jianpeng Li、Lili Wang、Xiang Cui、Bing Ji 和 Li Chen。在急诊科利用面部图像和深度学习技术预测贫血。《公共卫生前沿》，第10卷，2022年，2022年。