基于 CNN 的结膜苍白贫血检测的训练后量化性能分析

Sebastián A. Cruz Romero[1], Wilfredo E. Lugo Beauchamp[2]
Computer Science and Engineering,
University of Puerto Rico at Mayagüez,
Mayagüez, Puerto Rico
[1]sebastian.cruz6@upr.edu, [2]wilfredo.lugo1@upr.edu

摘要—贫血是全球广泛存在的健康问题，特别是在低资源环境中的幼儿中更为普遍。传统的贫血检测方法通常需要昂贵的设备和专业知识，这为早期和准确诊断设置了障碍。为了应对这些挑战，我们探索了通过结膜苍白使用深度学习模型来检测贫血的方法，重点关注 CP-AnemiC 数据集，该数据集包含 6 至 59 个月大的儿童的 710 张图像。数据集标注有血红蛋白水平、性别、年龄和其他人口统计数据，这使得开发用于准确诊断贫血的机器学习模型成为可能。我们使用 MobileNet 架构作为基础，因其在移动和嵌入式视觉应用中的高效性而闻名，并通过数据增强技术和交叉验证策略对我们的模型进行端到端微调。我们的模型实现达到了 0.9313 的准确率、0.9374 的精度以及 0.9773 的 F1 分数，展示了在数据集上的强劲性能。为了优化模型以适应边缘设备部署，我们进行了训练后量化，评估了不同位宽（FP32、FP16、INT8 和 INT4）对模型性能的影响。初步结果显示，尽管 FP16 量化保持了高准确率（0.9250）、精度（0.9370）和 F1 分数（0.9377），但更激进的量化（INT8 和 INT4）导致显著的性能下降。总体而言，我们的研究支持进一步探索量化方案和硬件优化，以评估移动医疗应用中模型大小、推理时间和诊断准确性之间的权衡。

***Index Terms*—计算机辅助诊断（CAD），贫血检测，卷积神经网络，训练后量化**

I. 介绍

贫血是全球范围内的一个普遍健康问题，主要影响低收入和中等收入国家（LMICs）的女性和儿童；症状如疲劳和免疫系统减弱可能会对儿童的发展产生影响。[1] [2] 贫血的标准诊断方法包括测量血液中的血红蛋白（Hb）水平，这需要专门的设备和人员——这些资源在农村和地区服务不足的地方通常有限。[1] [3] 根据世界卫生组织的数据，贫血影响全球 6 至 59 个月大的儿童中超过 40%，以及 37%的孕妇，其中最高比例出现在医疗保健获取受限的低收入和中等收入国家（LMICs）中。[2] 鉴于这些挑战，对于早期贫血检测的非侵入性、便携式诊断工具的兴趣日益增加，以实现预防干预。[4] [3]

一种非侵入性方法探讨了计算机辅助诊断（CAD）系统，该系统能够分析生理指标，如结膜苍白中的色素变化，作为贫血的诊断指标，因为它与血红蛋白水平有直接关系且易于使用。[5] 在评估苍白区域（指甲床、手掌、舌头）中，由于结膜的组织层较少且血管直接接触，它被认为是检测贫血最为敏感的部位。[5] [4] 然而，基于结膜苍白的贫血诊断工具在资源有限环境中的可行性仍然是一个研究兴趣增长领域，但尚未对最佳实施方法达成一致意见。[6]

随着人工智能（AI）和深度学习（DL）的进步，已经研究了各种非侵入性贫血诊断方法来解决临床方法的局限性。[4] 之前的工作引入了机器学习（ML）模型，通过基于苍白特征识别贫血状况或估计血红蛋白水平，尽管仍然存在依赖专有数据集和缺乏数据多样性等限制。[4] [7] 最近开发的 CP-AnemiC 数据集专注于通过结膜苍白来检测儿童贫血，通过提供一个大型、公开可用且平衡的数据集（包含来自加纳十个地区的多样样本），解决了其中的一些局限性。[7]

然而，在资源受限环境中引入计算开销较高的模型时，挑战依然存在。在这些情况下，高计算需求的模型往往不切实际，迫使研究转向更轻量、内存效率更高的模型，这些模型可能更适合边缘设备运行。[8] 主要缺

NSF-EPSCoR Center for the Advancement of Wearable Technologies [NSF Award OIA-1849243]

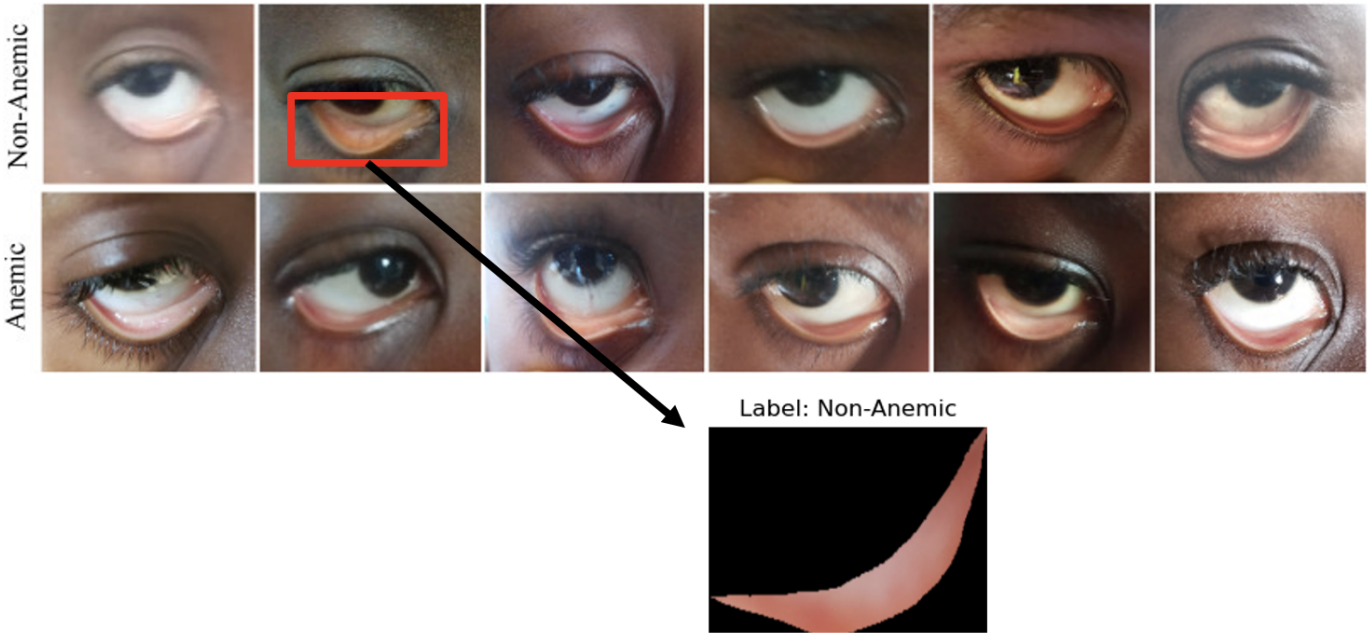

图 1. 结膜苍白的样本图像以及来自 CP-AnemiC 数据集的示例感兴趣区域。第一行代表非贫血患者的图像，第二行代表贫血患者的图像。

点是通常使用数十亿个参数进行计算，相比之下，传统算法则不然。[9] 因此，引入量化作为一种方法，在保持高性能准确率的同时减少神经网络架构的大小。深度学习涉及用浮点数表示的神经网络近似为位宽较低的神经网络。量化通过降低数值精度（通常从 32 位浮点数 (FP32) 降到 8 位整数 (INT8)），减少了内存占用并提高了计算效率。通过将原始连续值映射到一个离散、低分辨率范围内，量化在使用低精度算术运算时实现了存储效率和处理速度的显著提升。[10] [11] .

尽管这种转换引入了微小的近似误差，但几种量化策略有助于管理精度与准确率之间的权衡。训练后量化（PTQ）是一种广泛使用的方法，在模型训练完成后应用量化，无需额外的标记数据，因此在计算上较为经济，尽管会稍微牺牲一些准确性。相比之下，量化的感知训练（QAT）将量化整合到训练过程中，允许模型适应并保持较高的准确率，尽管这会增加训练过程中的计算需求。[10] [11] 量化可以遵循均匀或非均匀方案；均匀量化在区间内均匀分配值，提供简单性和速度，而非均匀量化则根据数据分布调整区间大小，虽然会提高计算复杂性。PTQ 可以采用静态或动态形式。训练后动态量化在推理过程中减少权重和激活的比特表示，降低计算负载和内存使用。这与训练后静态量化不同，后者使用校准数据集预先计算量化参数，包括部署前的权重和激活的比例因子。[10] [11] .

然而，虽然某些研究表明量化模型可能保留与全精度模型相当的诊断准确性，但仍需进一步探索以确认这些结果在特定诊断环境中的一致性，例如基于结膜苍白的贫血检测。本文介绍了一个用于基于结膜苍白的贫血检测的 CNN 分类器，使用 MobileNet [12] 模型。该项目基于 CP-AnemiC 数据集，利用它来探讨采用 PTQ 比较 FP32、FP16、INT8 和 INT4 位表示下的推理性能和执行时间的可行性。

II. 方法论

A. 数据集

CP-AnemiC 数据集是一个大规模的公开数据集，旨在通过结膜苍白分析解决贫血检测中的挑战。它包括来自加纳十家医疗机构在 2022 年 1 月至 6 月期间收集的 710 张儿童（年龄为 6 至 59 个月）的结膜图像。其中，424 张图像（60%）被标记为贫血，286 张（40%）为非贫血，基于世界卫生组织将血红蛋白（Hb）水平低于 11 g/dL 作为贫血诊断阈值的标准。这种多样性旨在提高在此数据集上训练的模型的泛化能力。参与者的平均年龄为 31.58 个月，包括 306 名女性（43%）和 404

名男性（57%）。每张图像都标注了 Hb 水平、年龄、性别、采集地点以及实验室评估的备注。该数据集还提供了按年龄和性别划分的 Hb 浓度的人口统计分析，突出了贫血参与者较低的 Hb 水平。

表 I
数据集的患者水平特征汇总。

患者类别	**贫血的**	**非贫血的**	**总计**
Patients	424	286	710
Female	174	132	306
Male	250	154	404
Age (months)	31.04 ± 17.02	32.31 ± 16.46	31.58 ± 16.78
6 – 59 月龄儿童的贫血诊断			
Anemia Classification	Anemic	Non-anemic	
Hemoglobin Levels	<11 g/dL	≥ 11 g/dL	

B. 实验设置

我们利用 MobileNet [12] 架构作为贫血分类模型的骨干，如图 2 所示，对在 ImageNet [13] 数据集上预训练的权重进行微调。我们的实验设置复现了 Appiahene 等人 [7] 描述的基于 CNN 的贫血检测方法，包括随机水平翻转、随机旋转、随机平移和随机缩放等数据增强技术以提高模型泛化能力。

五折交叉验证策略被实施，其中四折用于训练，一折用于测试。模型权重在每个折叠开始时随机初始化，并使用批处理大小为 32 的 NVIDIA GeForce RTX 4090 GPU 进行训练。我们使用二元交叉熵作为损失函数，并通过 Adam 优化器对模型进行了优化，将学习率设置为 10^{-4}。全连接层被修改以输出一个二元分类目标，其中 1 代表贫血类，0 代表非贫血类，输出使用了 sigmoid 激活函数。训练过程从头到尾进行最多 150 个纪元，在验证步骤中如果连续 10 个纪元的 F1 分数没有提高，则应用提前停止。我们按照以下方式计算模型评估性能指标：

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}} \tag{1}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{tp}} \tag{2}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{3}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{AUC} = \int_0^1 \text{tp rate}\, d(\text{fp rate}) \tag{5}$$

其中 tp、tn、fp 和 fn 分别代表真正例、真负例、假正例和假负例。

C. 量化和推理

我们执行后训练量化（PTQ）以通过减少模型权重和激活的位宽来优化我们的微调后的 MobileNetV2 模型。模型权重被转换为开放神经网络交换（ONNX）[14] 格式，这是一种通用表示形式，确保了跨框架和优化工具的兼容性。此转换涉及导出 PyTorch 模型，并将功能操作符映射到 ONNX 等效项，同时保留计算图。ONNX 通过标准化模型表示来促进后端特定的优化。在我们的工作流程中，NVIDIA 的 ModelOpt [15] 提供量化配置，使用 TensorRT [16] 作为后端，根据计算效率和准确性要求动态选择每层操作符的最佳精度。TensorRT 逐层应用这些优化，利用张量核心进行 FP16 运算，并使用整数算术单元进行 INT8 和 INT4 运算。按需执行去量化以确保混合精度层之间的兼容性。对于 FP16 转换，每个 32 位浮点值 (x) 被截断为 16 位，通过保留较少的尾数位来表示：

$$\text{FP16 Value} = \text{Round}\left(\frac{x}{2^k}\right) \cdot 2^k$$

其中，

- k 确定精度范围。

对于 INT8 和 INT4 量化，激活值使用校准数据得出的缩放因子映射到整数范围。我们使用 ModelOpts INT8 默认配置来启用权重和激活值的 8 位精度。权重量化按通道进行，而激活量化采用整体张量的方法。MaxCalibrator 算法通过计算张量的最大绝对值来确定缩放因子，确保从 FP32 到 INT8 的稳健映射。此配置优先考虑最小化精度损失同时实现显著的计算效率。

$$q_x = \text{Round}\left(\frac{x}{s_x}\right), \quad s_x = \frac{\max(|x|)}{2^{b-1} - 1}$$

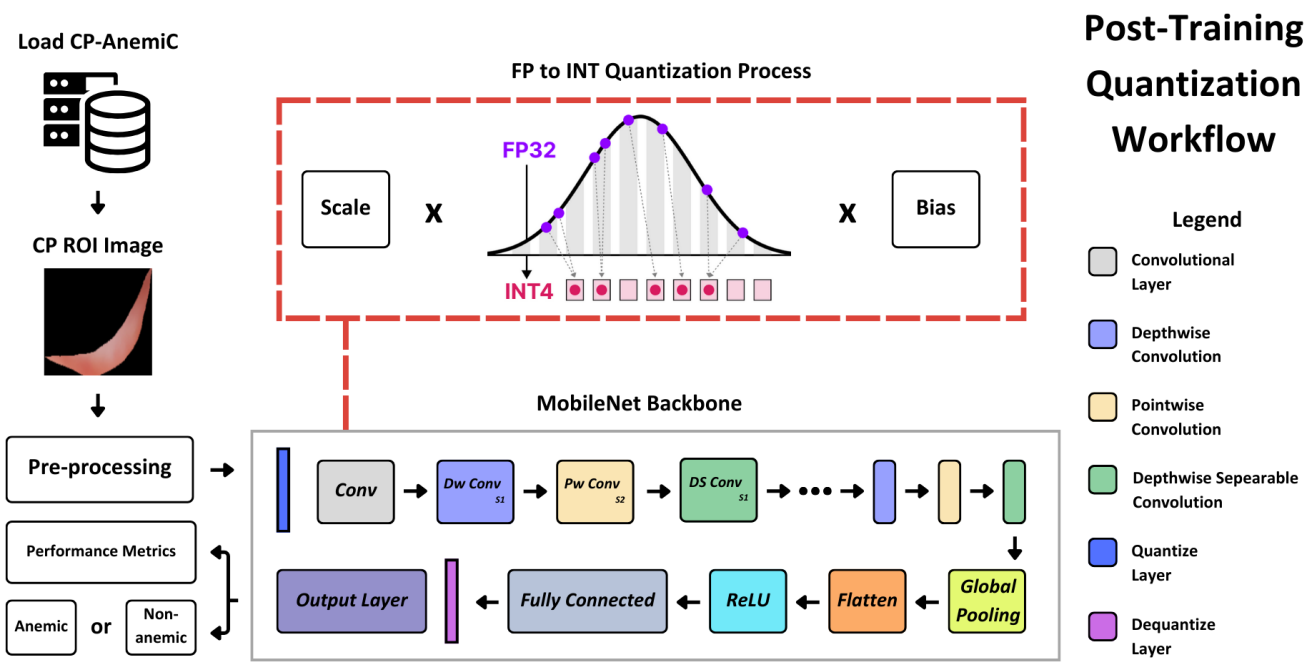

图 2. MobileNet 架构用作我们的贫血检测任务的主干。使用结膜苍白感兴趣区域作为输入，模型输出贫血和非贫血类别的数值表示。该模型对某些层进行了量化，以选择位宽表示。

其中，

- x 表示张量（权重或激活值），
- b 是比特的数量（例如，b=8 对于 INT8 或 b=4 对于 INT4）。

对于 INT4 量化，我们采用了高级权重量化（AWQ）配置，该配置利用块状量化对权重进行处理，块大小为 128 个元素。激活值被排除在量化之外以减少精度损失。AWQ 与 “awq_lite” 方法一起使用，通过迭代调整缩放参数（α），逐步减小量化误差。该方法实现了极端压缩，目标是针对内存和计算资源受限的环境。

$$\alpha^{(t+1)} = \alpha^{(t)} - \eta \cdot \nabla_{\alpha} \mathcal{L}(w, q_w),$$

其中，

- η 是步长，
- $\mathcal{L}(w, q_w)$ 是测量量化误差的损失函数。

III. 结果

A. 贫血检测

模型通过 5 折交叉验证初始化训练超过 150 个 epoch，由于数据集大小有限。监控验证 F1 分数以保存在不同位宽下运行推理的最佳模型权重。如表 III 所示，在 26 个 epoch 时观察到最高的验证 F1 分数为 0.9773，验证准确率为 96.88%，精度为 97.13%。然而，为了避免过拟合，在 26 个 epoch 后触发了提前停止，表明模型收敛，因为进一步的 F1 分数改进微乎其微。

B. 不同量化级别下的性能比较

FP32 达到了最佳性能，损失为 0.2141，准确率为 93.13%，精确率为 93.74%，召回率为 95.00%，F1 分数为 0.9428，AUC 分数为 0.9657，如表 VI 所示。FP16 紧随其后，性能仅略有下降，损失为 0.2149，准确率为 92.50%，精确率、召回率和 F1 分数相似，AUC 维持在较强的水平 0.9654。相比之下，INT8 量化导致准确性大幅下降（71.25%）和 AUC（90.05%），损失显著增加至 0.7441，表明尽管计算需求较低，模型性能仍

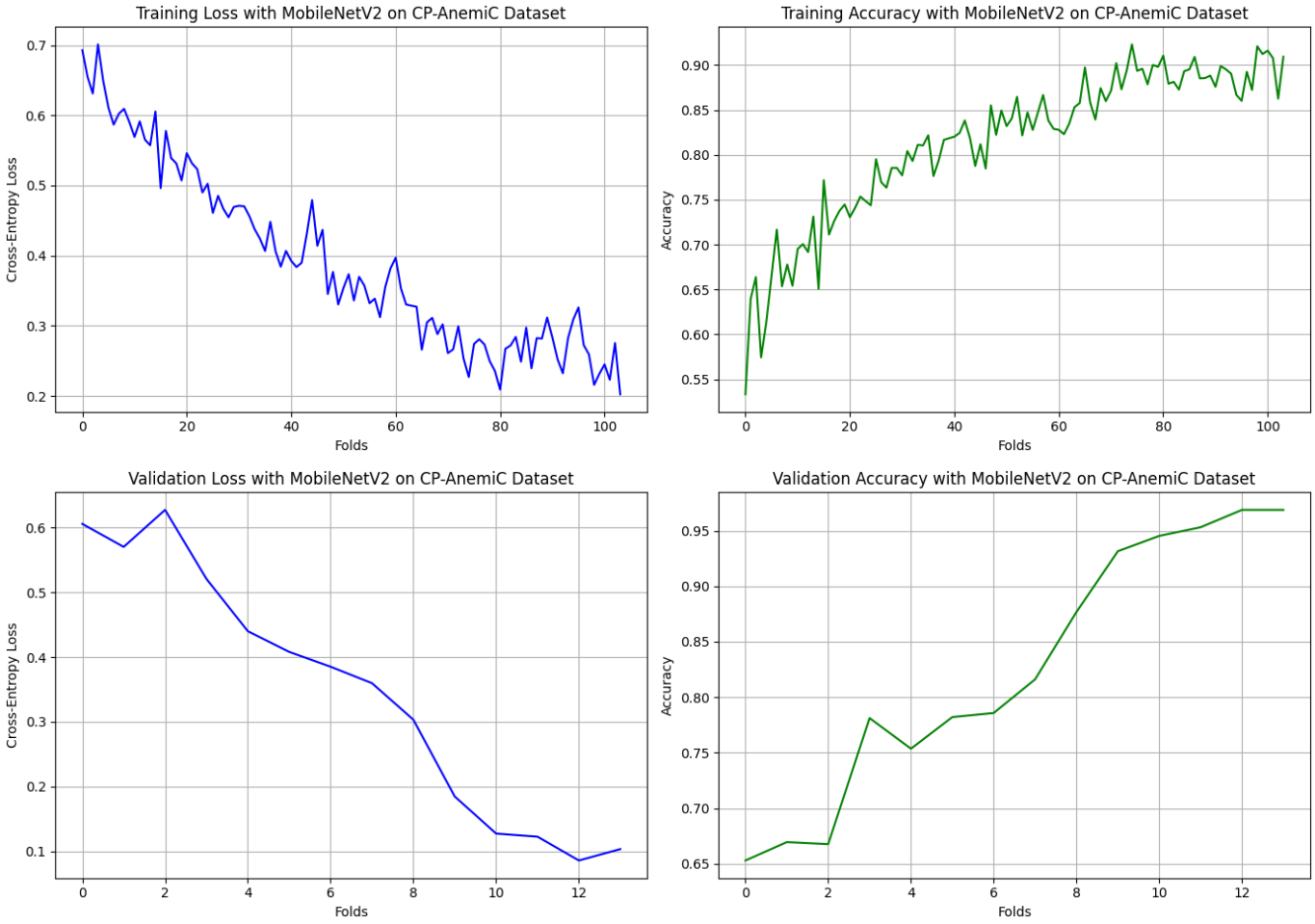

图 3. 经过 26 个周期训练后，我们的微调 MobileNet 在贫血检测中的训练和验证性能。

受到影响。INT4 量化导致了严重的性能下降，损失为 2.3136，准确率仅为 43.13%，精确率（20.00%）和召回率（1.00%）也较差。

C. 量化层用于整数算术运算

表 IV 展示了量化级别在内存和执行时间上的权衡。FP16 达到了显著的内存减少（4.61 MB）和最快的执行时间（37.4 ms）。INT8 尽管精度降低，但模型大小增加（9.24 MB）且延迟较高（91.9 ms），这可能是因为量化参数增加了现有架构上的开销。INT4 虽然实现了极端的压缩，但导致了最大的模型尺寸（17.75 MB）和与 FP32 相当的延迟（49.5 ms），表明在处理超低精度时存在效率问题。表 V 总结了受 INT8 和 INT4 量化影响的关键层，并突出了它们的权重量化特性。

D. 量化层用于整数算术运算

表 VI展示了在不同量化级别上的内存和执行时间的权衡。FP16 实现了显著的内存减少（4.61 MB）和最快的执行时间（37.4 ms）。INT8 尽管精度降低，但显示模型大小增加（9.24 MB）和延迟（91.9 ms），这可能是由于量化参数带来的开销叠加在现有架构之上所致。INT4 虽然实现了极端压缩，但导致了最大的模型大小（17.75 MB）和与 FP32 相当的延迟（49.5 ms），表明处理超低精度时存在效率低下问题。表 V 汇总了受 INT8 和 INT4 量化影响的关键层，突出了它们的权重量化特性。

IV. 结论与未来工作

本研究展示了轻量级架构如 MobileNet 在通过结膜苍白分析检测贫血方面的潜力。我们的模型在 CP-AnemiC 数据集上达到了最先进的性能，在未进行量化

表 II
贫血分类训练性能

Fold	损失	准确率	精度	回忆	F1 分数	AUC 分数
74	0.2269	0.9229	0.9252	0.9531	0.9383	0.9667
98	0.2157	0.9208	0.9305	0.9402	0.9331	0.9680
100	0.2448	0.9160	0.9370	0.9267	0.9257	0.9636
99	0.2314	0.9122	0.8869	0.9372	0.9090	0.9759
80	0.2090	0.9104	0.9170	0.9293	0.9219	0.9705
103	0.2023	0.9092	0.9312	0.9229	0.9249	0.9755
86	0.2394	0.9090	0.9160	0.9271	0.9208	0.9710
101	0.2230	0.9076	0.9598	0.8934	0.9225	0.9767
71	0.2663	0.9021	0.9071	0.9307	0.9176	0.9519
78	0.2495	0.9000	0.9244	0.9155	0.9166	0.9573
91	0.2521	0.8988	0.9085	0.9252	0.9132	0.9614
79	0.2356	0.8979	0.9046	0.9286	0.9127	0.9633
65	0.2660	0.8972	0.9329	0.8979	0.9115	0.9591
76	0.2808	0.8958	0.9107	0.9162	0.9072	0.9464
85	0.2973	0.8951	0.9047	0.9191	0.9061	0.9527
92	0.2322	0.8951	0.9057	0.9213	0.9097	0.9677
73	0.2531	0.8938	0.9204	0.9063	0.9104	0.9571
75	0.2741	0.8935	0.9265	0.9098	0.9147	0.9585
84	0.2487	0.8931	0.8993	0.9221	0.9071	0.9624
96	0.2725	0.8924	0.9124	0.9144	0.9064	0.9588
93	0.2823	0.8903	0.9084	0.9045	0.9053	0.9653
89	0.3117	0.8882	0.9086	0.9010	0.9021	0.9559
88	0.2817	0.8854	0.8947	0.9079	0.8992	0.9504
87	0.2823	0.8851	0.8797	0.9345	0.9001	0.9526
82	0.2719	0.8813	0.9103	0.8968	0.9001	0.9502

表 III
贫血分类验证性能

Fold	损失	准确性	精度	回忆	F1 分数	AUC 得分
12	0.0857	0.9688	0.9713	0.9722	0.9705	0.9978
13	0.1033	0.9688	0.9773	0.9773	0.9773	0.9923
11	0.1225	0.9531	0.9565	0.9659	0.9602	0.9920
10	0.1273	0.9453	0.9268	0.9747	0.9481	0.9910
9	0.1846	0.9315	0.8970	1.0000	0.9453	1.0000
8	0.3033	0.8768	0.9147	0.8728	0.8915	0.9497
7	0.3598	0.8162	0.8481	0.8302	0.8382	0.9189
6	0.3852	0.7858	0.7981	0.8589	0.8239	0.9160
5	0.4081	0.7822	0.7570	0.9025	0.8181	0.8923
3	0.5210	0.7813	0.7964	0.8160	0.7993	0.8048
4	0.4401	0.7537	0.7365	0.8927	0.8017	0.8865
1	0.5704	0.6696	0.6731	0.8441	0.7435	0.7548
2	0.6277	0.6677	0.6700	0.8615	0.7458	0.6965
0	0.6060	0.6530	0.6696	0.8242	0.7292	0.6981

的情况下推理时，F1 得分为 0.9428，准确率为 93.13%。训练后量化进一步优化了模型，使其适合跨不同量化级别的边缘设备部署，并揭示了计算效率和预测准确性之间的微妙权衡。FP16 量化保持了强大的性能，准确率

表 IV
不同量化水平的性能比较

位宽	损失	准确性	精度	回忆	F1 分数	AUC 得分
FP32	0.2141	0.9313	0.9374	0.9500	0.9428	0.9657
FP16	0.2149	0.9250	0.9370	0.9400	0.9377	0.9654
INT8	0.7441	0.7125	0.7697	0.7519	0.7607	0.9005
INT4	2.3136	0.4313	0.2000	0.0100	0.0196	0.6387

表 V
关键层的量化总结

层名称	位宽	量化方法	amax 范围
features.0.0	INT8	Per-axis	[0.0039, 1.4840]
features.1.conv.0.0	INT8	Per-axis	[0.0036, 2.6928]
features.10.conv.1.0	INT8	Per-axis	[0.0103, 0.6126]
features.0.0	INT4	Block-wise	[0.0005, 1.4840]
features.1.conv.0.0	INT4	Block-wise	[0.0014, 2.6928]
features.10.conv.1.0	INT4	Block-wise	[0.0035, 0.5219]

表 VI
不同量化级别下的内存消耗和执行时间

位宽	模型大小	执行时间
FP32	9.13 MB	48.6 ms ± 235 μs
FP16	4.61 MB	37.4 ms ± 1.01 ms
INT8	9.24 MB	91.9 ms ± 1.92 ms
INT4	17.75 MB	49.5 ms ± 1.34 ms

仅略有下降（92.50%），F1 得分略微降低（0.9377）。更激进的量化技术导致性能退化明显。模型在不同量化比特宽度上的初步结果表明，激进的量化会严重损害模型的预测能力。

未来的工作包括在逐层计算中系统地实现完整的整数算术。此外，我们将研究这些优化措施对推理延迟的影响，特别是在边缘设备如 NVIDIA Jetson Xavier NX 和 TX2 NX 上的影响，因为这些设备具有小巧的外形尺寸和低功耗。使用 TensorRT 作为操作符转换后端，我们旨在利用其按层选择最优精度的能力，通过混合精度算术来提高执行速度同时保持诊断完整性。

致谢

本工作得到了波多黎各梅亚圭兹大学研究与发展中心以及国家科学基金会 EPSCoR 可穿戴技术进步中心的支持，国家科学基金会资助号 OIA-1849243。

参考文献

[1] C. M. Chaparro and P. S. Suchdev, "Anemia epidemiology, pathophysiology, and etiology in low- and middle-income countries," *Annals of the New York Academy of Sciences*, vol. 1450, no. 1, pp. 15–31, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6697587/

[2] World Health Organization, "Anaemia," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/anaemia

[3] M. N. Garcia-Casal, O. Dary, M. E. Jefferds, and S. R. Pasricha, "Diagnosing anemia: Challenges selecting methods, addressing underlying causes, and implementing actions at the public health level," *Annals of the New York Academy of Sciences*, vol. 1524, no. 1, pp. 37–50, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10880862/

[4] R. An, Y. Huang, Y. Man, R. W. Valentine, E. Kucukal, U. Goreke, Z. Sekyonda, C. Piccone, A. Owusu-Ansah, S. Ahuja, J. A. Little, and U. A. Gurkan, "Emerging point-of-care technologies for anemia detection," *Lab on a Chip*, vol. 21, no. 10, pp. 1843–1865, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8875318/

[5] T. N. Sheth, N. K. Choudhry, M. Bowes, and A. S. Detsky, "The relation of conjunctival pallor to the presence of anemia," *Journal of General Internal Medicine*, vol. 12, no. 2, pp. 102–106, 1997.

[6] M. W. Merid, D. Chilot, and A. Z. Alem, "An unacceptably high burden of anaemia and its predictors among young women (15 – 24 years) in low and middle income countries; setback to sdg progress," *BMC Public Health*, vol. 23, no. 1292, 2023. [Online]. Available: https://doi.org/10.1186/s12889-023-16187-5

[7] G. Dapena and et al., "Cp-anemic: A conjunctival pallor dataset and benchmark for anemia detection in children," *Pattern Recognition Letters*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590093523000395

[8] R. Sharma, A. Alharbi, A. Alsarhan *et al.*, "Internet of intelligent things: A convergence of embedded systems, edge computing and machine learning," *Journal of King Saud University - Computer and Information Sciences*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660524000945

[9] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *CoRR*, vol. abs/1605.07678, 2016. [Online]. Available: http://arxiv.org/abs/1605.07678

[10] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," *CoRR*, vol. abs/2106.08295, 2021. [Online]. Available: https://arxiv.org/abs/2106.08295

[11] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," 2021. [Online]. Available: https://arxiv.org/abs/2103.13630

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[14] O. R. developers, "Onnx runtime," https://onnxruntime.ai/, 2021, version: x.y.z.

[15] NVIDIA, *TensorRT Model Optimizer: PyTorch Quantization Guide*, 2024, accessed: 2024-11-27. [Online]. Available: https://nvidia.github.io/TensorRT-Model-Optimizer/index.html

[16] NVIDIA Corporation, *TensorRT Developer Guide*, 2024, accessed: 2024-11-23. [Online]. Available: https://docs.nvidia.com/deeplearning/tensorrt/pdf/TensorRT-Developer-Guide.pdf