第五届中国研究生人工智能创新大赛

[基于扩张 CNN 的可穿戴设备的高糖次数监测]

技术文档

[v23.8.06]

[2023.08.06]

[华为赛题三]

1 项目概况

1.1 背景和基础

糖尿病是一种以血糖升高为特征的全球性的慢性代谢疾病。据报道，2019 年全球约有 4.63 亿人患有糖尿病，160 万人直接死于糖尿病[1]。在过去的几十年里，糖尿病的病例数量和患病率都在稳步增加，每年有 150 万人的死亡直接归因于糖尿病[1]。同样，糖尿病会加剧心血管疾病、终末期肾病，导致视力障碍和截肢，从而给社会带来巨大的经济负担[2]。因此，实时监测血糖浓度，快速发现高糖状况，有效降低糖尿病患病率，对于个人、社会和国家都具有重大的意义。

而糖尿病血糖监测设备是帮助糖尿病患者管理血糖水平的重要工具之一。其中一个常见的监测方法是使用血糖仪，这种设备需要从指尖上采集一小份血液样本进行测试。采血时通常使用无菌针头，以确保卫生和准确性[3]。但糖尿病前期人群对于通过有创设备监测控糖水平的接受程度较低，而智能穿戴技术的持续进步为血糖监测领域带来了新的可能性[4]。从 1999 年的 GlucoWatch 到现今更强大的产品如 Dexcom、Medtronic 和 Abbott 等连续血糖监测(CGM)系统，智能穿戴技术已经取得了巨大的飞跃。因此，分析 CGM 相关监测数据，建立准确、稳定的血糖水平监测模型，对于更加高效、便捷的监测血糖水平至关重要。

我们经过调研发现，早期的血糖监测在机器学习中主要利用了随机森林、多元自适应样条回归、k 近邻、决策树、梯度增强回归树、支持向量回归(SVR)。然而，这些经典 ML 方法的主要缺点是很难发掘与时间顺序有关的特征，更多的是时间点建模，精度不高，并且它们的成功高度依赖于用于训练它们的数据的特征工程和统计分析的质量[5]。目前在这个领域比较流行的深度学习方法有 LSTM 和 transformer 等。LSTM 神经网络已成功应用于医疗预测等领域，基于 LSTM 网络的预测方法在血糖预测中也得到了很好的应用[5][6][7]。孙庆南等采用 LSTM 网络和双向 LSTM 网络预测血糖浓度序列，并对 LSTM 和双向 LSTM 的预测结果进行了对比分析。Aliberti 等人将 LSTM 网络与多患者数据驱动方法相结合，预测患者血糖浓度，取得了较好的预测效果[7]。尽管在不同的测试集上，这些最新的基于人工神经网络的研究都显示了更好的准确率值。但它们也有多个缺点:i)现有的 LSTM 以及 transformer 模型在训练时，需要依赖于长时间的有效记录数据（缺失值较多，缺失时间间隔较久），但这样的数据常常是难以获得的。很多时候我们仅能获得一小段时间内的有关特征数据，在 LSTM 以及 transformer 等模

[1] https://www.who.int/health-topics/diabetes#tab=tab_1 WHO/健康专题/糖尿病

型上的效果并不会特别显著，反而还会带来更大的训练和推理成本。;ii)它限制了系统的泛化能力并增加了过拟合的风险，因为如果训练数据过少或者噪声较多，LSTM 可能会学到训练数据中的细节和噪声，而忽略了通用的模式。且血糖浓度时间序列具有较强的时变性，属于典型的非线性非平稳序列，直接使用 LSTM 和 transformer 监测血糖浓度会在一定程度上影响监测精度[8]。同时我们的时序数据又存在缺失值较多,缺失时间间隔较久，噪声较大等问题，所以 LSTM 以及 transformer 并不适用于我们的数据。其他模型有一些侵入性设备提供的指标，但完全无侵入性指标的建模不足，需要建模来实现无侵入性指标的效果提升。

基于上述问题，我们构建了一个血糖次数监测模型，其中，我们采用扩张 CNN 作为主要框架，在不同层级上同时关注细节和全局信息。其模型的简洁性将减少训练与推理的计算量，并同时也可以挖掘时间序列的影响。然后，我们使用 COB 对食物进行特征提取，最后实现了高糖次数监测。

作为一个深度学习模型，CNN（卷积神经网络）的每一层都进行卷积计算。对于时间序列数据，可以通过移位法来实现卷积的输出。与具有循环结构的神经网络相比，CNN 没有循环连接，因此在训练过程中更快。这对于训练长时间序列数据非常重要。

首先,CNN 相比其他现有算法具有更高的预测精度;其次,它易于实现，无需进行繁琐的参数调优。扩张型 CNN 的关键思想是在每层的输入值上跳过一定的步骤，使得网络可以在更大规模上运行，从而增加了接受域的层数数量级。扩张型 CNN 在处理具有更宽感受野的多维长信号方面表现出色[9]。

基于上述优点，本项目对传统的 CNN 进行改进，引入了扩展卷积神经网络（Dilation Convolution）来提高血糖预测的准确性和可靠性。通过扩张卷积，网络可以更好地捕捉时间序列数据中的长期依赖关系，从而提高预测性能[10]。

1.2 场景和价值

本项目适用于糖尿病前期患者的高糖预测任务，主要应用于健康管理、预防性医疗、生物医学研究、保险与健康政策等多个领域。通过使用智能穿戴设备进行糖尿病前期患者的高糖预测，旨在提高血糖监测的准确性和便捷性，帮助患者预警高血糖风险。市场上的健康管理应用和医疗设备虽然试图应用智能穿戴设备进行预测，但智能穿戴设备数据准确性、个体差异等问题尚待解决，算法和模型需优化以适应复杂情况。此项目通过进行数据特征处理，使用扩展卷积神经网络，针对糖尿病前期患者血糖波动特点和个体差异，进一步改进模型的预测性能。

1.3 所需支持

本项目所使用的服务器配置为：cpu 为 Intel(R) Xeon(R) CPU E5-2678 v3 @2.50GHz，内存 101.2GB，显卡为三张 Tesla V100，硬盘 2TB。

另外，本项目所需相关培训为对 pytorch、numpy 等 python 包的运用以及 Linux 系统服务器的使用。

2 项目规划

2.1 整体目标

本项目的整体目标在于基于现有可穿戴设备搜集到的存在缺失和短时无创特征与摄入食物记录，完善现有食物指标模型，并通过扩张 CNN 构建血糖浓度监测模型，从而提供高效准确的高糖次数预测。

2.2 技术创新点

在本次比赛中我们从模型和数据的角度出发，进行了以下创新：

(1)由于数据缺失较多，时序信息不完整等问题，我们采用 Dilation Convolution 来有效捕捉数据之间的关联性和交互模式，提高模型的准确性和泛化能力。Dilation Convolution 可以设置不同的扩张率，同时，扩张卷积的特点可以减少模型的参数数量，降低过拟合的风险，增强模型的泛化能力，使其在未见过的数据上表现更好[10]。通过去除缺失值较多的特征列，对标签前五分钟特征进行平均化，以及删除合并后标签行的空缺值，确保了数据的质量和完整性。采用这样的处理方法，我们只需利用前五分钟的十个 30 秒平均值点特征即可实时预测当前时间点的血糖值，无需使用长时间序列信息，从而提高了预测模型的准确性和泛化能力，为糖尿病患者提供了便捷、高效、及时的血糖监测方案。我们考虑引入了 Huber Loss，同时兼顾均方误差（Mean Squared Error，MSE）和平均绝对误差（Mean Absolute Error，MAE）的优点[11]。

(2)在血糖监测任务中，我们在关注的三轴加速度计特征中，为了综合考虑

加速度在平面方向上的正负值对运动量影响的相似性，我们采用特殊处理方式，将三个方向的加速度值综合考虑并降低特征维度，同时有效地将运动对血糖的影响融入模型。而对于食物的处理，我们提取对血糖影响最大的碳水量这一列，并采用到食物刚摄入体内先递增后递减的策略。这种处理方式使得预测模型更全面、准确地捕捉运动与血糖之间的关系，为实时血糖监测提供更可靠、高效的方案，对糖尿病患者的健康管理具有重要意义。

3 实施方案

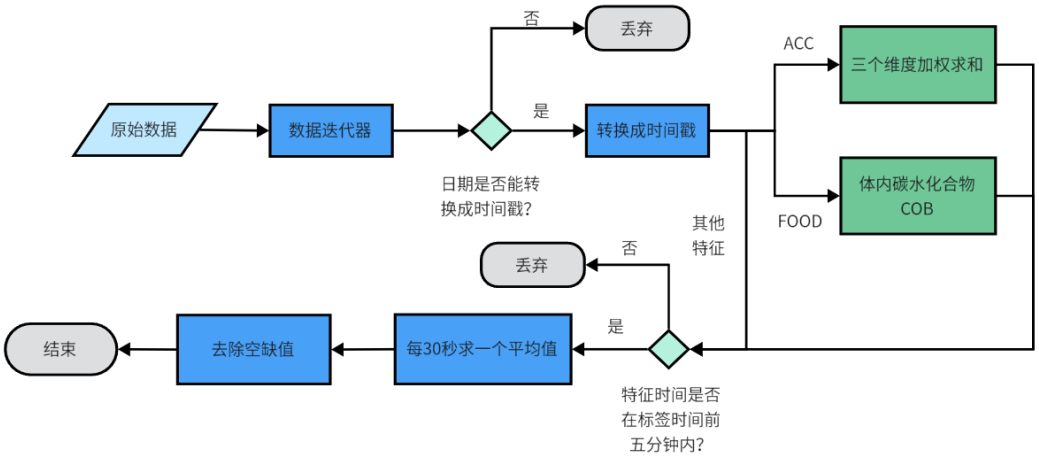
3.1 技术可行性分析

本项目所需数据主要来源于比赛方公开数据集，其与可穿戴设备的实际搜集到的情况十分接近，从数据层面具有推广的可行性；本项目组在模型训练阶段，自身有对应的服务器可以使用，以保证有足够的算力、硬件支持；另外本项目采用的基准模型扩张 CNN 计算花费较低，从技术上容易落地应用。综上所述，本项目组的技术可行性极高。

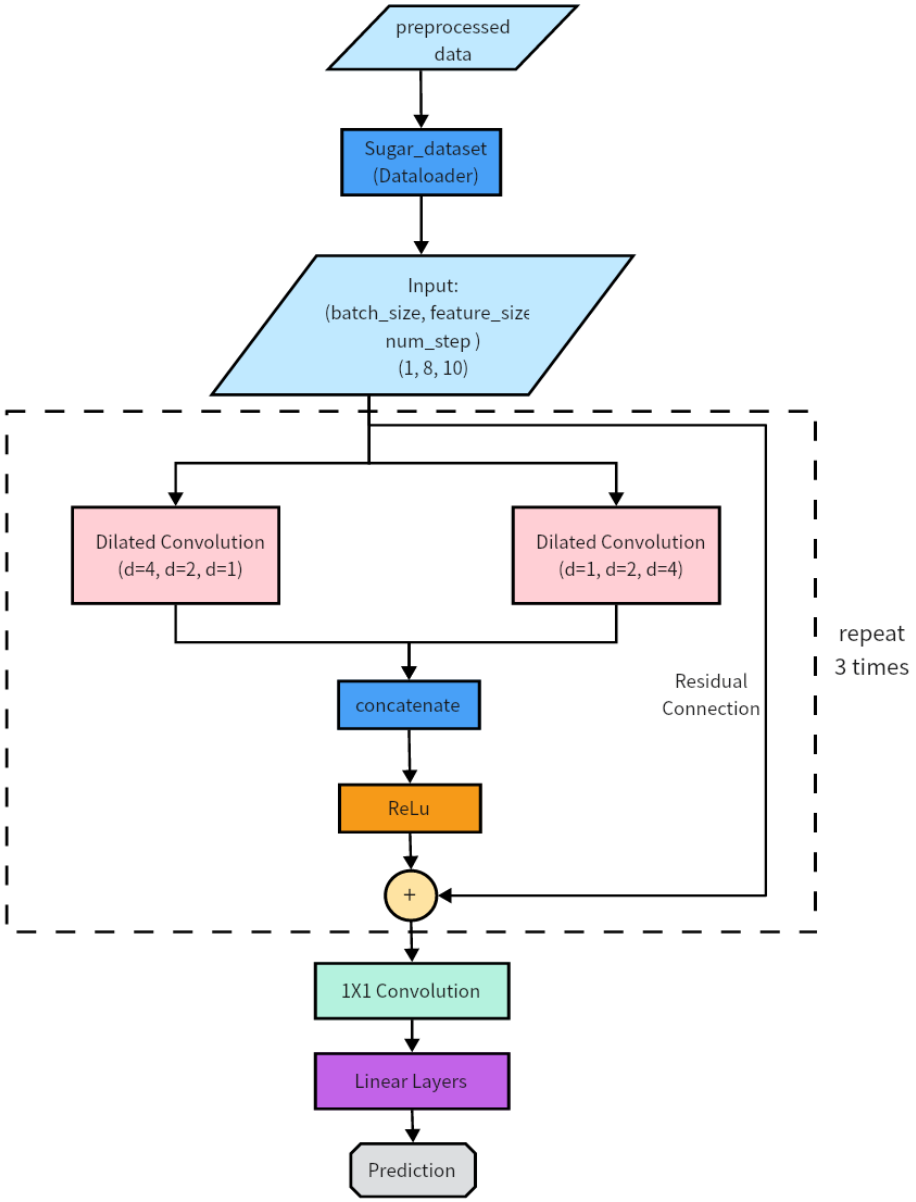
3.2 技术细节

我们使用扩张 CNN 对模型进行改进，增强了时序性捕捉和泛化能力，选取 HuberLoss 作为损失函数，加入先验知识处理加速度计和食物特征，降低特征维度，实现实时血糖监测。扩张 CNN 块对不同时间步的数据进行卷积，最后融合多尺度特征，增强特征表达能力。我们的模型训练 300 轮后，在预测高糖次数上获得良好结果。

数据预处理流程图：



模型整体结构流程图：



在数据预处理方面：我们对 8 个特征分分别进行预处理，在这里，我们以 30s 为一个时间块，以当前五分钟的前十个时间块的特征来预测当前血糖值，而每个时间块的特征我们定义如下：

bvp_average：这里取的是每个时间块的 bvp 指标的平均值
eda_average：这里取的是每个时间块的 eda 指标的平均值
hr_average：这里取的是每个时间块的 hr 指标的平均值
ibi_average：这里取的是每个时间块的 ibi 指标的平均值
temp_average：这里取的是每个时间块的 temp 指标的平均值

然后我们对 acc 和 foodt 特征进行了特殊的预处理。

为了综合考虑加速度在平面方向上的正负值对运动量影响的相似性，我们采用特殊处理方式，将三个方向的加速度值综合考虑并降低特征维度，同时有效地将运动对血糖的影响融入模型。所以我们对 acc 进行特有的预处理方式，具体地，我们首先考虑到 acc_z 向上对血糖的影响更大，向下对血糖影响相对较小，所以我们分别对 acc_z>0 乘以了一个 2 倍的权重，对 acc_z<0 乘以了一个 1/2 倍的权重，并考虑到平移对血糖的影响相同，所以我们对 acc_x 和 acc_y 取绝对值，为了避免 acc 三个特征 acc_x，acc_y 和 acc_z 对血糖影响分配了太大的权重，导致其他特征对血糖影响分配不平衡，我们定义了如下的 acc_average 作为每个时间块的输入：

$$acc\_average = |acc\_x| + |acc\_y| + \delta$$
$$\delta := \begin{cases} -\frac{1}{2}acc\_z, & if \quad acc\_z < 0 \\ 2acc\_z, & if \quad acc\_z \geq 0 \end{cases}$$

关于食物这块，我们选取食物指标中对血糖影响最大的碳水化合物量，关于碳水量特征的处理，我们参考了文献[12]，其变化方式为：以人吃下食物的事件节点为起点，经过 15 分钟后，此特征增加对应食物的碳水量，然后以每分钟 0.5 克的速率开始下降。按照如下公式：

$$\mathrm{COB}_t = \max(0, C - R_{\mathrm{COB}} * (t - t_C - \Delta_{\mathrm{COB}}))$$

COB_t 是 t 时刻体内所剩碳水化合物含量，C 表示 t_c 时刻的摄入碳水化合物含量，R_{COB} 表示食物每分钟的衰减速率，单位为 g/min，Δ_{COB} 表示食物从 t_c 时刻摄入食物起到开始消化的间隔时间，这里等于 15。

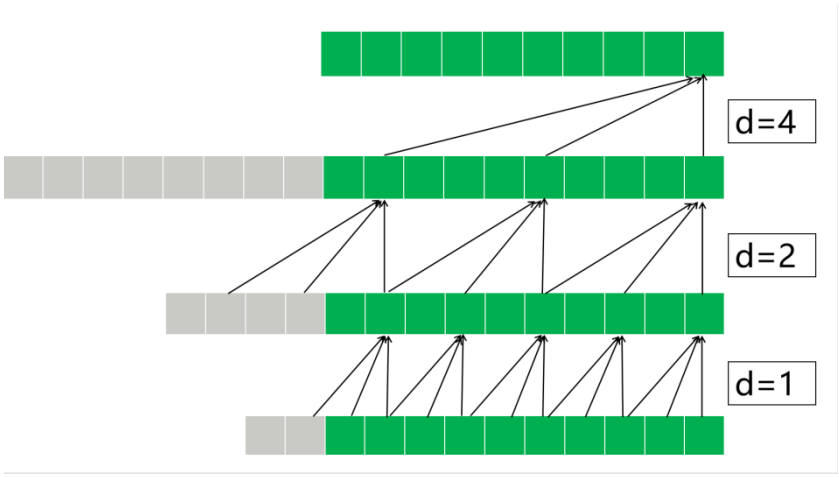
最终，我们对数据进行预处理，以适应模型的输入形状(batch_size, feature_size, num_steps)。在这里，我们将批量大小(batch_size)设置为 1，特征数(feature_size)设置为 8，时间步长(num_steps)设置为 10。这意味着我们将以每次处理一个样本的方式输入模型，并且每个样本将包含 10 个时间步长的特征。

为了更好地表示时间序列数据，我们将每个金标的前 5 分钟划分为 10 个时间块，每块 30 秒。这样，我们可以更好地捕捉到时间上的动态变化。通过这种划分方式，我们可以将每个时间块的特征作为一个时间步长输入到模型中，共有 10 个时间步长。

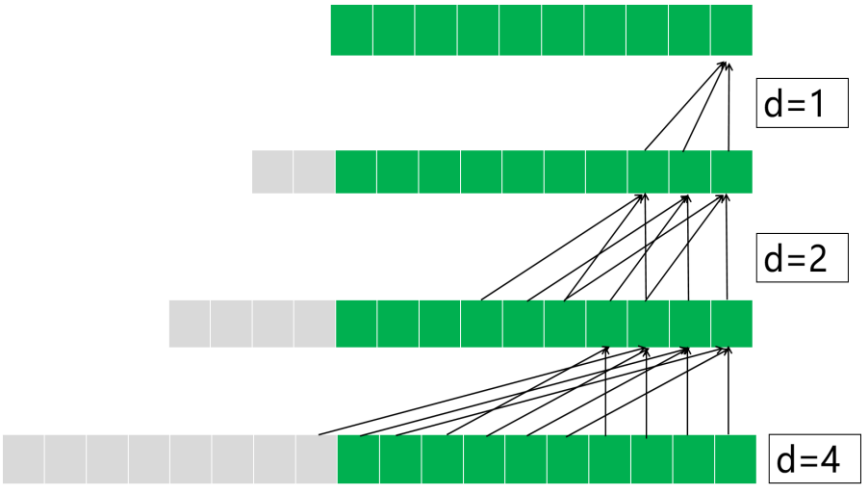
在模型的选择方面，我们采用了一种基于传统卷积神经网络(CNN)修改后的结构，即 Dilation Convolution。为了捕捉到不同时间尺度上的特征，我们选

择了 kernel_size 为 3 的卷积核。这种修改后的 CNN 结构具有更大的感受野，可以在不增加网络参数的情况下提高模型的表示能力。

第一类型结构为：



第二类型结构为：



其中绿色格子代表不同时间步的值，灰色格子代表 padding 步数。

此结构的意义是使得最终输出的每个时间步上的值受到初始输入中此时间步及之前时间步所有值的影响[13][14]。

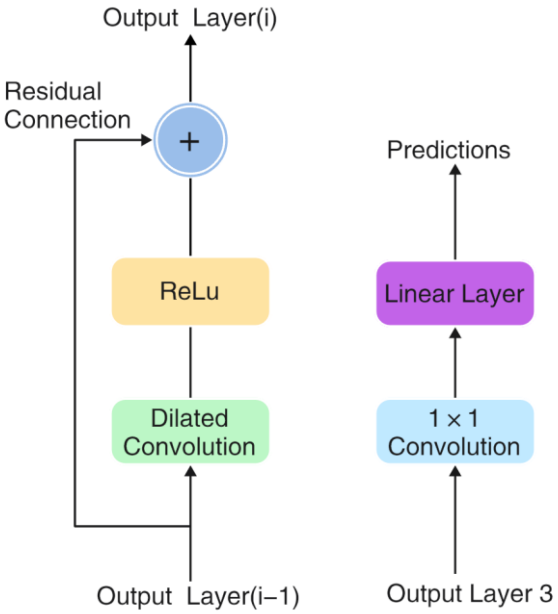
第一类型的中间块，输入数据依次经过 dilation=1,dilation=2，dilation=4 的卷积层，由于 dilation=1 的卷积层在感受野上不进行扩张，可以更好地捕捉局部细节信息。而后续的 dilation=2 和 dilation=4 的卷积层逐渐扩张感受野，这样有助于我们捕捉更大范围的上下文信息。通过这种组合，可以在不同层级上同时关注细节和全局信息[15]。

第二类型的中间块，则为输入数据依次经过 dilation=4,dilation=2,dilation=1 的卷积层，然后也依次通过 ReLu 函数以及残差连接。dilation=4，2，1 的顺序会导致较大的 dilation 值先进行特征提取，这可能有助于捕捉更宽广

的上下文信息。而较小的 dilation 值则在后面的层级进行特征提取，更关注细节信息。

我们输入数据复制为相同的两组，一组依次经过三个第一类型的中间块，一组依次经过三个第二类型的中间块，然后将两组的输出进行拼接，再通过 1x1 的卷积核进行卷积，最后进入若干带有激活函数的全连接层得到最后的回归值结果。回归值高于阈值 7.8mmol/L 的视为高血糖，否则视为正常值。这样将不同 dilation 值的结果进行拼接，并接全连接层进行回归任务，可以实现多尺度特征的融合，增强特征表达能力，增加非线性能力，并减少过拟合问题，从而提高回归任务的性能和泛化能力[16][17]。

模型整体结构如下：



实验设置:我们选取了 HuberLoss 作为损失函数，其中的超参数 delta 设置为当前训练集的金标的标准差，Adam 作为优化算法，beta1，beta2 分别设置为 0.9,0.999,eps=1e-08,epoch 设置为 500,kernel_size 设置为 3,学习率 lr=0.0001[13]。HuberLoss 定义如下：

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

当 $\delta \to 0$ 时，Huber 损失会趋向于 MAE；当 $\delta \to \infty$，Huber 损失会趋向于 MSE。

模型评估方面，我们随机选取 8 个人的数据进行训练，用剩下的人的数据进行测试，从分类准确率，以及预测高糖次数的相对误差来评价模型。

可行性分析：

数据可行性：通过数据统计分析，发现第一个用户的 hr 特征的时间与金标

无法对齐，因此我们将其弃用。最终数据集由 15 个糖尿病前期用户组成，包含智能穿戴设备记录的多个生理数据，具备实现目标的基本数据条件。

技术可行性：扩展卷积神经网络等深度学习方法在处理时间序列数据和传感器数据方面已经取得了很好的成果，因此在此任务中也有很大的可行性。

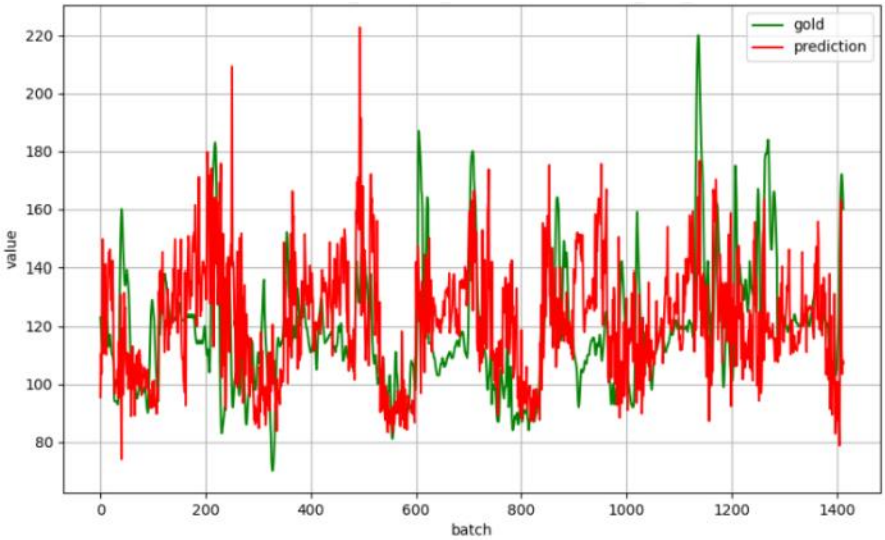
算法可行性：预测糖尿病前期患者一天内高糖发生次数是一个有挑战性的问题，但通过合理的数据预处理和特征提取，结合深度学习模型的优势，可以有效地解决这一问题。

预期技术指标：

高糖次数：对测试集中高糖发生次数的预测是核心指标之一。我们希望模型能够以较高的准确度监测高糖发生的次数，为糖尿病患者提供更加个性化和精准的血糖控制建议，提高生活质量和健康水平，我们选取了在五个测试用户上测试的表现情况。如图所示：

	003	008	014	015	016
真实高糖次数	78	129	174	13	107
预测高糖次数	92	83	161	20	157
准确率	0.8816	0.9101	0.8195	0.9325	0.8555
MSE	513.3376	443.6053	511.9233	487.9582	684.3259

实时性：由于糖尿病前期患者需要实时监测血糖情况，模型需要具备实时性，能够及时预测高糖发生次数，以便患者及时调整血糖管理策略，下图为 014 用户的血糖预测值与金标的对比图。其中横坐标表示每个 batch 节点，纵坐标 value 表示每个 batch 上的平均血糖值，单位为 mg/dL。



可以看到，我们的模型在金标属于高糖情形时大多可以监测出属于高糖的血糖值，曲线拟合较好，并且预测高糖次数和实际高糖次数较为吻合，不足之处在于部分高糖处的预测值和真实值有一定的差距。