基于 PBI-CLA 模型的糖尿病患者血糖浓度预测

肖泽秋[1]，李勇[1]，王霞[2]

（1.西北师范大学计算机科学与工程学院，甘肃 兰州 730000；2.甘肃省人民医院药剂科，甘肃 兰州 730000）

摘　要：糖尿病作为全球 4 种主要的非传染性疾病之一，死亡率连年上升。糖尿病患者若长期血糖偏高会引发一系列并发症，并产生严重的不良后果。预测和控制血糖浓度是糖尿病诊断和治疗中的一个关键问题，尽管近年来连续血糖监测（CGM）技术的发展已部分解决人工检测带来的不便，但是 CGM 设备不仅昂贵还易受外在因素干扰。基于深度学习方法提出一个患者血糖浓度水平预测模型 PBI-CLA。首先，在模型的卷积神经网络（CNN）层通过一维卷积将血糖浓度序列和胰岛素剂量序列的数据特征提取出来；其次，在模型的长短期记忆（LSTM）层学习时间序列步长的关联关系；最后，模型的注意力层对每个测定血糖浓度时间节点中注射胰岛素剂量赋予不同的权重，输出得到血糖浓度的预测值。实验结果表明，该模型的均方根误差（RMSE）、平均绝对误差（MAE）和平均绝对百分比误差（MAPE）3 个指标均有大幅下降，与其他血糖浓度预测模型相比，PBI-CLA 模型 1 h 血糖浓度预测的 RMSE 和 MAPE 分别下降了 12.82 和 10.24 百分点。

关键词：糖尿病；血糖浓度预测；深度学习；卷积神经网络；长短期记忆；注意力机制

源代码链接：https://physionet.org/content/mimiciv/2.2/

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0070527

Prediction of Blood Glucose Concentration in Diabetic Patients Based on PBI-CLA Model

XIAO Zeqiu[1], LI Yong[1], WANG Xia[2]

(1.School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730000, Gansu, China;
2.Department of Pharmacy, Gansu Provincial People's Hospital, Lanzhou 730000, Gansu, China)

【Abstract】Diabetes, one of four major global noncommunicable diseases, has seen an increase in mortality rates over the years. Patients with diabetes and chronically high blood glucose levels may experience various complications and serious adverse consequences. The accurate prediction and control of blood glucose levels are critical for the diagnosis and treatment of diabetes. Although Continuous Glucose Monitoring (CGM) technology has alleviated some challenges associated with manual detection, it remains costly and vulnerable to external interference. This study proposes the PBI-CLA model, which is based on deep learning, for predicting blood glucose concentration levels in patients. First, the Convolutional Neural Network (CNN) layer extracts data features from blood glucose concentration and insulin dose sequences through one-dimensional convolution. Subsequently, the Long Short-Term Memory (LSTM) layer learns the correlations between time series increments. Finally, the attention layer assigns different weights to the insulin dosage injected at each time node to measure the blood glucose concentration and outputs the predicted blood glucose concentration value. In extensive experiments, the model reduces the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) substantially. Compared with other glucose concentration prediction models, the RMSE and MAPE values achieved by PBI-CLA for one-hour glucose concentration prediction decrease by 12.82 and 10.24 percentage points, respectively.

【Key words】diabetes; blood glucose concentration prediction; deep learning; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); attention mechanism

0　引言

糖尿病是一种发病率较高的慢性非传染病，据国际糖尿病联盟（IDF）统计，2021 年，全世界有 5.37 亿人患糖尿病，预计到 2030 年，这一数字将达到 6.43 亿，到 2045 年将达到 7.83 亿[1]。2019 年，因糖尿病或其并发症而死亡的人数占总死亡人数的 11.3%。IDF 机构 2021 年统计显示，全球约有

基金项目：国家自然科学基金（62163033）；甘肃省科技计划项目（23JRZA397）。

作者简介：肖泽秋，男，硕士研究生，主研方向深度学习、人工智能；李勇（通信作者），副教授、博士；王霞，副主任药师、硕士。

收稿日期：2024-10-23　　**修回日期：**2024-12-05　　**E-mail：**facingworld@nwnu.edu.cn

计算机工程 www.ecice06.com

670 万成年人因糖尿病及其并发症死亡，相当于每 5 秒死亡 1 人[1]。预计到 2030 年，糖尿病将成为全球第七大死因。中国是糖尿病患病人数最多的国家，2021 年，中国 20～79 岁人群中糖尿病患者约有 1.4 亿例[2]，患病率约 10.6%，预计到 2045 年，患病人数将超过 1.74 亿例，每年糖尿病导致的死亡人数将高达 139 万例，占全球糖尿病死亡人数的 20%[1]。

糖尿病主要症状表现为由于患者体内葡萄糖代谢紊乱导致的长期血糖水平异常[3]。长期的高血糖症状会给患者的肾脏、心脑血管、神经系统等器官带来严重的危害，引发诸多并发症。糖尿病一般可分为两种主要类型：Ⅰ型糖尿病（T1DM）和Ⅱ型糖尿病（T2DM）[4]。Ⅰ型糖尿病是一种自身免疫性疾病，患者由于自身免疫系统紊乱导致胰岛 β 细胞的损害，引发胰岛素分泌水平的下降，使得患者不能产生足够的胰岛素来处理血液中的葡萄糖。Ⅱ型糖尿病患者尽管自身会产生胰岛素，但由于患者对胰岛素产生抵抗，导致胰岛素无法正常调节血糖水平[5]。

及时检测血糖浓度水平能帮助糖尿病患者有效维持正常血糖水平[4]，避免引发其他并发症。传统的血糖浓度预测通过连续血糖监测技术，这种技术能够预测患者短时间内的血糖浓度，但是无法满足重症监护室等医疗场景下的特定检测需求[6]，并且无法预测其他因素对血糖浓度的影响。因此，一些人工智能方法根据历史血糖浓度数据和患者身体信息进行血糖浓度预测。这种方法相比传统的血糖浓度预测技术进行了较大的改进，但仍存在一些不足。首先，血糖浓度数据具有明显的时序关联关系，现有部分研究未能捕捉到时序上的依赖关系，导致模型精确度不高；其次，大多数研究中患者的样例数量存在不足，少则几个患者样例，多则几十个样例，研究结果存在较大个体差异性；最后，已有研究所用的数据仅有患者的历史血糖监测数据，缺乏患者体征、人口等数据的支撑，导致模型的鲁棒性不强，预测精度不高。

针对以上问题，本文基于深度学习方法提出了一个血糖浓度水平预测模型 PBI-CLA（Physical-Body- Insulin and CNN-LSTM-Attention）。本文主要贡献有以下 3 个方面：

1）提出了血糖浓度预测模型 PBI-CLA。首先，通过卷积神经网络（CNN）与长短期记忆（LSTM）网络的有效结合，将多变量时间序列提取出来。然后，融合注意力机制对每个测定血糖浓度的时间节点赋予不同的权重。有效解决了血糖浓度预测中多变量数据难以建模预测的问题，最大程度地提升了模型的预测性能。

2）本文根据糖尿病患者数据特征采用随机森林和样条插值法进行数据的填补增强，对不规则样本数据进行生成拟合，有效解决了血糖浓度预测中数据的缺失值问题，同时对原始数据进行了归一化处理，提高了模型预测的准确率。

3）基于国际上公开的重症监护医疗数据集 MIMIC-IV 获得海量电子病历，采集了糖尿病患者的生命体征、24 h 间隔采样血糖浓度等多特征数据，有效解决了患者个体差异问题和样例不足问题，增强了预测模型的鲁棒性。

1 相关工作

1.1 传统的血糖浓度预测方法

传统的血糖浓度预测方法分为基于患者生理模型的预测方法和基于数据建模的预测方法。基于生理模型的预测方法需要大量参数且易受外界干扰，很难模拟人体血糖的调节代谢过程。连续血糖监测技术（CGM）是一种能够在短时间内多次测量患者血糖水平的技术[6]，CGM 设备可以以 5 min 为间隔测量血糖[7]，提供了比传统的手工血糖监测[8]更为准确可靠的血糖浓度数据[9-10]，CGM 的技术更新使其能够短时预测血糖浓度，但是也存在着无法对更长时间内血糖浓度进行预测的不足。

1.2 基于人工智能的预测方法

人工智能技术已在糖尿病诊断和预测领域得到广泛应用[11-13]，包括人工神经网络、支持向量回归、遗传规划等。文献[14]提出了一种基于历史血糖监测数据预测未来血糖浓度的人工神经网络模型，其预测结果比传统的方法更加准确，但是模型性能较差。文献[15]将支持向量回归方法应用于患者每日血糖预测。文献[16]提出了一种增强的遗传规划算法，以创建原始碳水化合物和胰岛素时间序列的平滑时间序列来进行血糖预测，但遗传算法的性能对参数高度敏感，导致优化效果的不稳定性。

1.3 融合时序建模的深度学习方法

融合时序建模的深度学习方法涉及多种模型的组合与优化，这类方法运用在血糖浓度数据中可以更准确地捕捉血糖浓度序列中的非线性关系以及复杂模式。文献[17]通过构建堆叠长短期记忆网络进行深度多任务学习以预测个性化血糖浓度，提前了解血糖趋势，避免高血糖和低血糖的发生。文献[18]通过改进粒子群优化与长短期记忆网络构建短期血糖预测模型，从而降低血糖时间序列的非平稳性。时序建模的深度学习方法能够适应不同患者的血糖变化特征，在处理时序预测问题上准确性更

计算机工程 www.ecice06.com

高。本文提出的模型 PBI-CLA 本质上也是融合时序建模的一种深度学习方法，通过构建血糖浓度序列和胰岛素序列进行血糖浓度预测。

2 模型与方法

2.1 问题定义

设给定糖尿病患者集合 $P=\{p_1,p_2,\cdots,p_n\}$，特征集合 $x_t=\{\boldsymbol{h}_t|\text{age, sex, height, glucose}_t,\cdots\}$，特征向量 $\boldsymbol{h}_t$ 表示糖尿病患者 p 在时间点 t 的生命体征、血糖浓度等监测数据，患者在 ICU 住院期间是一个连续的时间区间 T，$T=\{t_1,t_2,\cdots,t_n\}$，则 $\boldsymbol{x}_c=\bigcup_{t=1}^{n}\boldsymbol{h}_t$ 表示患者 c 在测量时间范围内的多变量时间序列特征向量 $\boldsymbol{x}$。糖尿病患者血糖预测问题可形式化描述为：已知糖尿病患者 p 在住院期间 T 的多变量时间序列特征向量 $\boldsymbol{x}$，目标是通过 PBI-CLA 模型求得患者 p 下一时间点的血糖浓度预测值 $y=P_{\text{PBI-CLA}}(p,x,T)$。

2.2 数据处理

由于设备故障、人工记录错误等多种原因造成 MIMIC 数据中有大量的缺失问题，并且缺失值呈现不规则性和不确定性。本文基于随机森林和样条插值的数据填补方法提高数据的完整性。

2.2.1 随机森林

随机森林是一种基于集成学习的决策树模型，该模型以其卓越的鲁棒性和广泛的适用性著称，特别是在处理包含缺失值的数据集时展现出显著优势。

针对患者体重数据存在部分缺失的情况，本文采取了一种新的策略：首先，将数据集细分为完整（无缺失值）与不完整（含缺失值）两部分。其次，利用无缺失值的体重数据子集训练随机森林模型，从而深入理解体重与其他相关因素之间的潜在关系。然后，利用已训练好的随机森林模型对含有缺失值的体重数据进行预测，生成准确的估计值。最后，将这些预测值无缝回填至原始数据矩阵的相应位置，实现了对缺失值的填补。

2.2.2 样条插值

样条插值是一种数学方法，用于在给定的数据点之间构造一个平滑函数曲线，本文利用这种方法填补患者血糖浓度序列中缺失的数据值。

首先设定血糖浓度序列为 $P=[P_0,P_1,\cdots,P_m]$，从糖尿病患者已有的血糖浓度序列中取最大值和最小值作为三次样条插值的插值区间 $[\min(P),\max(P)]$，给定 $n(1\leqslant n\leqslant 4)$ 个插值节点 $X_i(\min(P)=x_0<x_1<\cdots<x_{n-1}<x_n=\max(P))$。将数据点的坐标进行排序，每两个数据点间定义一个区间 $[x_i,x_{i+1}]$，其中 $i=0,1,\cdots,n-1$。

定义一个三次多项式 $S(x)$，可表示为：

$$S(x)=a+b(x-x_i)+c(x-x_i)^2+d(x-x_i)^3 \quad (1)$$

式中：a、b、c、d 是每个区间的多项式系数，并且要保证数据点函数值和一阶导数要连续。通过使用边界条件和区间连续性条件建立线性方程组求解区间的多项式系数，根据建立好的三次多项式填补血糖浓度序列缺失值。

2.3 模型总体架构

本文 PBI-CLA 模型框架结构如图 1 所示。首先将患者的血糖浓度和注射胰岛素剂量数据作为 CNN 模块的输入，有效降低了特征提取和数据重建的复杂度，提高了数据特征的质量。然后，将不同时间的特征向量形成一个新的时间序列，输入到 LSTM 模块以学习血糖浓度和胰岛素时间序列的关联关系。最后，引入注意力机制来分析每个特征对血糖浓度趋势影响的重要性。通过注意力层，捕获历史血糖浓度数据和胰岛素剂量数据之间的时间相关性，进一步提高模型的预测精度。

2.4 CNN 模块

本文提出的 PBI-CLA 模型中 CNN 模块由两个一维卷积层、两个池化层（最大池化）和一个全连接层组成，通过全连接层提取输入患者血糖浓度序列特征和患者注射胰岛素的序列特征。经过卷积层和池化层的处理后，将原始数据映射到隐藏层的特征空间，捕获时间窗口内的血糖浓度波动，然后构建全连接层进行结构转换输出，提取特征向量。

CNN 模块的输入包括糖尿病患者的身高、体重、年龄、性别、24 h 的血糖浓度和注射胰岛素序列。经过一维卷积层后输出 $\boldsymbol{C}$，如式(2)所示：

$$\boldsymbol{C}=f(\boldsymbol{W}\otimes\boldsymbol{x}+\boldsymbol{b}) \quad (2)$$

式中：$\boldsymbol{C}$ 是经过一维卷积之后的输出表示；$\boldsymbol{W}$ 是权重矩阵；$\boldsymbol{b}$ 是卷积层偏置参数；$\boldsymbol{x}$ 是本文中的血糖浓度序列与胰岛素剂量序列的联合向量表示，$\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$，$n=6$。

通过一维卷积后，已获得患者的血糖浓度和胰岛素序列特征，再输入到池化层经过对数据中的重要信息进行提取，得到池化后的输出 $\boldsymbol{P}$，如式(3)所示：

$$\boldsymbol{P}=\max(\boldsymbol{C})+\boldsymbol{b}' \quad (3)$$

式中：max 表示最大池化；$\boldsymbol{b}'$是池化层的偏置参数。

经过两轮卷积最大池化操作，输出 CNN 模块的特征向量 $\boldsymbol{H}_c$，如式(4)所示：

$$\boldsymbol{H}_c=f(\boldsymbol{W}'\otimes\boldsymbol{P}+\boldsymbol{b}'') \quad (4)$$

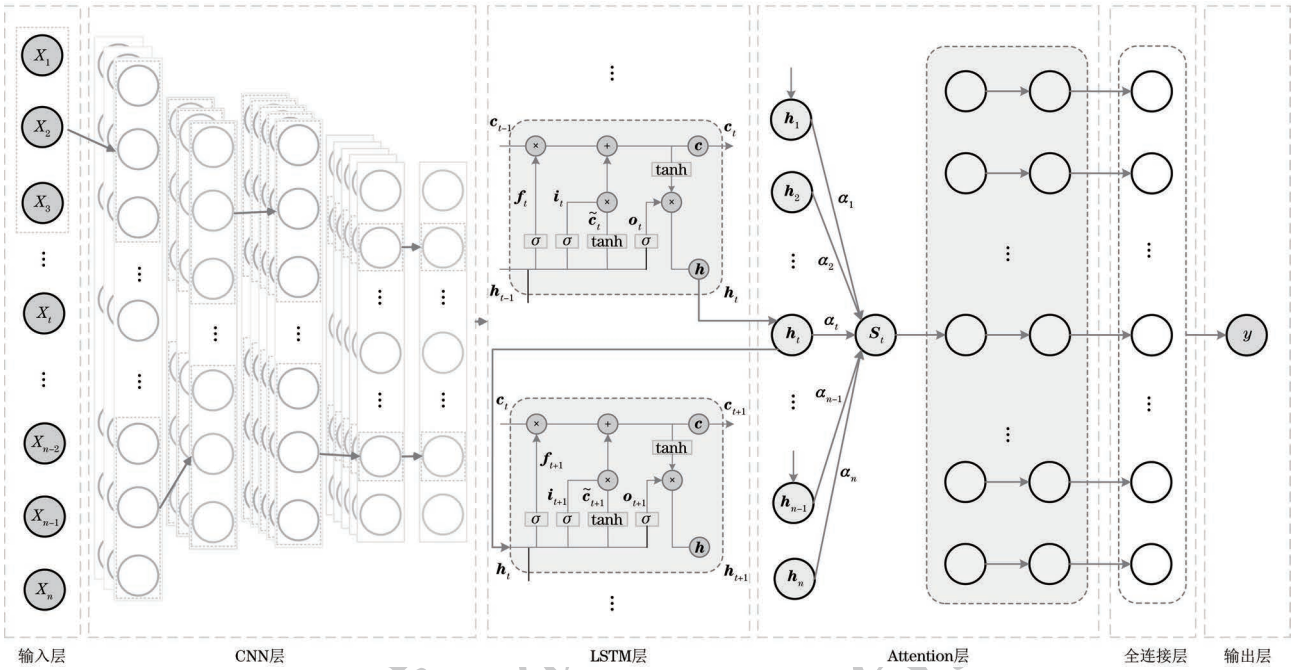


图 1 PBI-CLA 模型框架结构

Fig. 1 Structure of the PBI-CLA modeling framework

最后，CNN 模块的输出为 $\boldsymbol{H}_c=[\boldsymbol{h}_{c_1},\boldsymbol{h}_{c_2},\cdots,\boldsymbol{h}_{c_i}]^{\mathrm{T}}$，其中 i 是输出长度。

2.5 LSTM 模块

PBI-CLA 模型的 LSTM 模块用于学习 CNN 模块提取的特征以捕获其内部变化，通过处理提取到的特征序列，捕获其中的时间依赖性。LSTM 模块能够学习血糖浓度随时间的长期变化趋势，以及外部因素（注射胰岛素剂量）对血糖的影响。

模型中 LSTM 模块由 2 个 LSTM 层组成。第 t 步的输出如式(5)所示：

$$\begin{cases}\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{b}_{t-1},\boldsymbol{b}_{ct}]+\boldsymbol{b}_i)\\ \boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{b}_{t-1},\boldsymbol{b}_{ct}]+\boldsymbol{b}_f)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{b}_{t-1},\boldsymbol{b}_{ct}]+\boldsymbol{b}_o)\\ \boldsymbol{c}_t=\tanh(\boldsymbol{W}_c\cdot[\boldsymbol{b}_{t-1},\boldsymbol{b}_{ct}]+\boldsymbol{b}_c)\\ \boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{c}_t\\ \boldsymbol{b}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t)\end{cases}\tag{5}$$

式中：$\boldsymbol{h}_t$ 是 LSTM 层在 t 时刻的输出隐藏状态；σ 是 Sigmoid 函数；tanh 是激活函数；$\boldsymbol{W}$ 是权重系数矩阵，$\boldsymbol{b}$ 是偏置项；$\odot$ 是哈达玛积；$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$、$\boldsymbol{c}_t$ 和 $\boldsymbol{c}_t$ 分别是输入门、遗忘门、输出门、单元的候选状态和单元的当前状态。LSTM 模块经输入门得到前一时刻 $\boldsymbol{h}_{t-1}$ 的隐藏状态和当前时刻 $\boldsymbol{h}_t$ 的输入状态。当前时刻的状态和输入分别经激活函数 tanh 和 Sigmoid 映射，然后求和通过输出门输出。在这个过程中，LSTM 模块能够有效地提取序列数据中的时间特征，并将这些特征以隐藏状态 $\boldsymbol{h}_t$ 的形式表示出来，最后输出表示为一个时间节点数量×步长维度的矩阵传递到 Attention 模块。

2.6 Attention 模块

在 Attention 模块中，注意力机制对每个时刻的隐藏状态进行加权以突出重要的历史时刻，并且忽略不相关数据的影响。在 LSTM 的输出基础上添加注意力机制，用来识别特定时间步长对预测当前血糖浓度更为重要的信息。这有助于模型专注于那些与预测目标最相关的输入数据部分。

通过构建可学习参数矩阵与 LSTM 模块输入的隐藏状态 $\boldsymbol{h}_t$ 计算，得到长度为 T 的注意力相似度 $\boldsymbol{e}_t$。要计算数据节点间相似度，首先对相似度 $\boldsymbol{e}_t$ 使用 Sigmoid 函数进行归一化处理，得到节点间注意力权值系数 $\boldsymbol{\alpha}_t$，如式(6)所示：

$$\boldsymbol{\alpha}_t=\mathrm{Softmax}(\boldsymbol{e}_t)=\frac{\exp(\boldsymbol{e}_t)}{\sum_{k=1}^{t}\boldsymbol{e}_k}\tag{6}$$

式中：$\boldsymbol{e}_t$ 为 t 时刻确定的注意力权重。

在计算 t 时刻的注意力权重时，本文采用加性模型将 $\boldsymbol{h}_t$ 时刻状态进行加性融合，即对应元素相加，并通过激活函数 tanh 来增强表达能力，如式(7)所示：

$$\boldsymbol{e}_t=\boldsymbol{u}^{\mathrm{T}}\tanh(\boldsymbol{W}_e\boldsymbol{h}_t+\boldsymbol{b}_e)\tag{7}$$

式中：$\boldsymbol{u}^{\mathrm{T}}$ 和 $\boldsymbol{W}_e$ 分别为权重系数矩阵；$\boldsymbol{b}_e$ 为偏置系数矩阵。

如式(8)所示，最终对所有节点加权求和得到注意力层 t 时刻的输出 $\boldsymbol{s}_t$：

$$\boldsymbol{s}_t=\sum_{t=1}^{n}\boldsymbol{\alpha}_t\boldsymbol{h}_t\tag{8}$$

计算机工程 www.ecice.com

加权求和得到的上下文向量通过 dropout 层进一步减少模型对特定上下文信息的依赖，提高模型的泛化能力，然后由全连接层进行输出，最终得到关键节点特征。

3 实验与结果分析

3.1 数据来源

本文的数据来自国际上公开的大型公共电子病历数据集 MIMIC-IV，该数据集收集了美国马萨诸塞州波士顿贝斯以色列女执事医疗中心 2008 年至 2019 年超过 18 万名患者，43 万人次住院记录的临床数据，包括以患者为中心的临床记录，如人口统计学信息、生命体征测量、实验室检查以及药物管理等内容[19]。

3.2 数据筛选

本文选取了 MIMIC-IV 数据库中 37 157 例各型糖尿病患者，为保证患者测量血糖浓度次数和注射胰岛素次数，进一步筛选出了具有以下 3 个条件的糖尿病患者：

1)第一次进入 ICU。

2)年龄大于 18 岁。

3)血糖浓度缺失值小于 50%。

糖尿病患者病例的数据选择流程如图 2 所示，最终选取糖尿病患者病例 2 020 例，选取的患者病例信息包括患者编号、身高、体重、年龄、性别、24 h 内每隔 2 h 的患者血糖平均浓度以及 24 h 内每隔 2 h 的患者注射胰岛素剂量水平，MIMIC-IV 纳入糖尿病患者统计信息如表 1 所示。

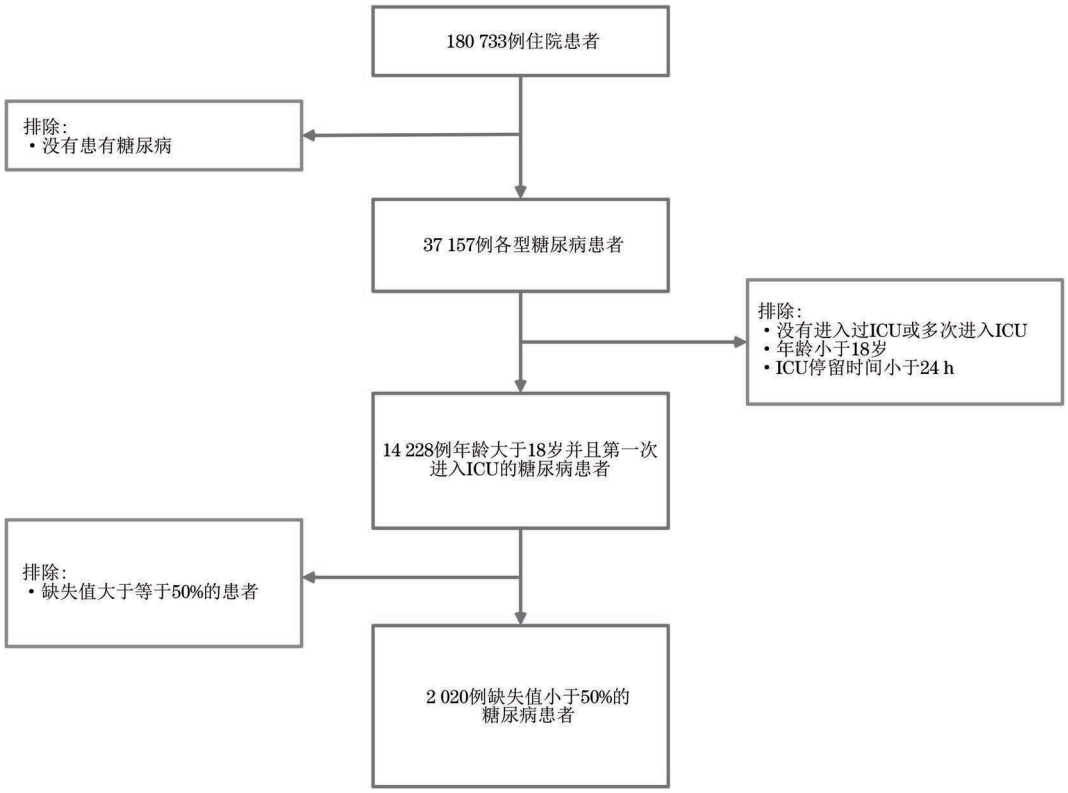


图 2 糖尿病患者病例筛选过程

Fig. 2 The data selection process of diabetic patients

表 1 MIMIC-IV 纳入糖尿病患者统计信息

Table 1 MIMIC-IV include statistical information on patients with diabetes

字段名	单位	类型	数量/例	统计信息(均值)
患者编号	—	int	2 020	—
身高	cm	float	2 020	169.1(122.0～203.0)
体重	kg	float	2 020	86.0(38.7～226.0)
年龄	—	int	2 020	63.3(18～96)
性别	—	char	2 020	—
血糖平均浓度	mg/dl	float	24 240	208.3(25.6～978.5)
注射胰岛素剂量	ml	float	24 240	—

3.3 数据填补方法对比

本文分别使用均值填充、随机森林、样条插值、随机森林与样条插值结合插补方法处理数据在 6 种模型上进行 1 h 血糖浓度预测，并对 4 组方法处理的结果进行比较。

实验通过采用随机森林结合样条插值法插补缺失值 3 691 个，不同数据补全方式进行血糖浓度预测实验对比，实验结果如图 3 所示。

本文提出的 PBI-CLA 模型由于结合了随机森林和样条插值法，其预测性能提升较为显著。这一

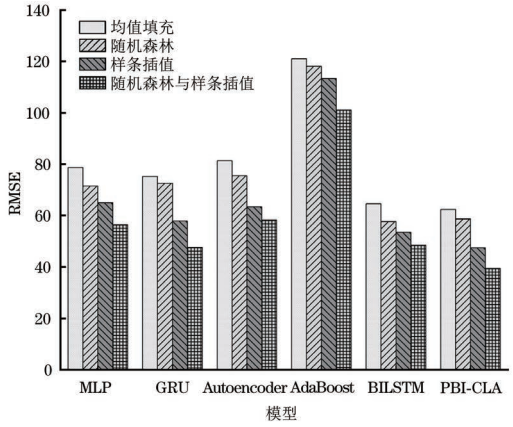


图 3 模型数据补全方式的 RMSE 对比

Fig. 3 RMSE comparison of model data completion mode

显著改进主要是结合了随机森林预测精度高的特点和样条插值法在处理非规则历史血糖数据时的卓越表现，能够更有效地拟合出数据的内在趋势和变化，从而为模型提供更加准确、平滑的输入，进而提升整体预测的准确性。

3.4 参数设置与基准模型

超参数优化是机器学习和深度学习方法中一个至关重要的环节，它直接影响到模型的性能和效果。常用的超参数优化方法有网格搜索、随机搜索、贝叶斯优化等。网格搜索方法具有简单直观且不会遗漏任意一组超参数组合的优势，更易找到全局最优解，且该方法在超参数较少的情况下优化更为有效。本文采用 10 倍交叉验证的网格搜索方法来优化 PBI-CLA 模型的超参数，从而使模型预测误差降至最小。根据文献[20-21]，以及实验中采用 RMSE 指标评估超参数的预测效果，得到 PBI-CLA 模型的最优超参数如表 2 所示。

表 2 PBI-CLA 模型超参数设置

Table 2 PBI-CLA model hyperparameter settings

模型结构	超参数	最优值	参数设置范围
CNN 层	kernel_size	3	[1,2,3]
	hidden_size	64	[32,64,128]
	num_layers	2	[2,4,6,8]
	pooling_layers	2	[1,2,3]
LSTM 层	num_layers	2	[1,2,3,4,5]
	hidden_size	128	[32,64,128,256]
Attention 层	drop_out	0.2	[0.1,0.2,0.3,0.4]
PBI-CLA	num_epochs	100	[50,100,150]
	learning_rate	0.001	[0.001,0.005,0.01,0.1]
	batch_size	32	[16,32,64]
	optimizer	Adam	[SGD, Adam]

为了验证 PBI-CLA 模型的预测效果，本文选取了 5 种代表性模型进行对比实验：

1) AdaBoost[22]。AdaBoost 是一种自适应的，能够将多个弱分类器组合成一个强分类器的集成学习算法。

2) Autoencoder[23]。Autoencoder 是一种通过无监督学习的方式，将输入数据编码为一个更小的表示，然后再从这个表示中重构出原始数据的方法。

3) MLP[24]。多层感知机是一种基本的人工神经网络模型，是由多个神经元组成的多层结构。

4) GRU[25]。GRU 是一种带有门控机制的循环神经网络单元，用于处理变长序列数据。

5) BiLSTM[26]。BiLSTM 是一种循环神经网络变体，能够同时考虑输入序列的正向和反向信息，捕捉长距离的依赖关系。

3.5 评价指标

本文的目标是为了预测糖尿病患者未来时间的血糖浓度水平，本质上是一项多变量回归任务。考虑到血糖浓度预测的性能要求较高，并且数据样本存在异常值较多的情况，本文选择均方根误差(RMSE，E_{RMSE})、平均绝对误差(MAE，E_{MAE})作为评估模型性能的主要评价指标，并以平均绝对百分比误差(MAPE，E_{MAPE})作为辅助指标。RMSE 的优点在于对误差进行了平方处理，使较大误差对整体评估的影响更大，这有助于模型更加关注那些预测误差较大的样本。MAE 优点在于计算简单且对所有误差给予同等权重，对异常值的敏感性相对较低。两者相结合，更能反映模型的总体性能。此外，Clarke 误差网格分析(EGA)法[17]在医学领域时常被用于评估血糖测量预测模型准确性，因此，本文也选取了 Clarke 误差网格作为判断实验结果准确性的度量指标。

RMSE 的计算公式如式(9)所示：

$$E_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{9}$$

MAE 的计算公式如式(10)所示：

$$E_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{10}$$

MAPE 的计算公式如式(11)所示：

$$E_{\mathrm{MAPE}}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{11}$$

式中：$\hat{y}_i$ 表示模型预测值；y_i 表示实际真实值；n 表示比较次数；RMSE 表示预测值与真实值偏差的平方与比较次数 n 比值的平方根；MAE 表示预测值与真实值的绝对误差；MAPE 表示预测值与真实值偏差与真实值比值的绝对值。RMSE、MAE、MAPE 的值越小，表示模型实际误差越小，预测结

果更准确。

Clarke 误差网格在纵轴上对应模型预测值，在横轴上对应参考真实值，由多条线段划分成 5 类区域，区域 A 被认为是临床上准确的，区域 B 表示误差结果不大，在临床可接受的范围内，区域 C 表示误差结果较大，区域 D 和 E 表示预测值与真实值存在极大误差，没有任何实际意义。

3.6 对比实验分析

本文分别进行了 PBI-CLA 模型、MLP 模型、Autoencoder 模型、GRU 模型、AdaBoost 模型和 BiLSTM 模型提前 1 h 的血糖浓度预测，Clarke 误差网格预测结果如图 4 所示。

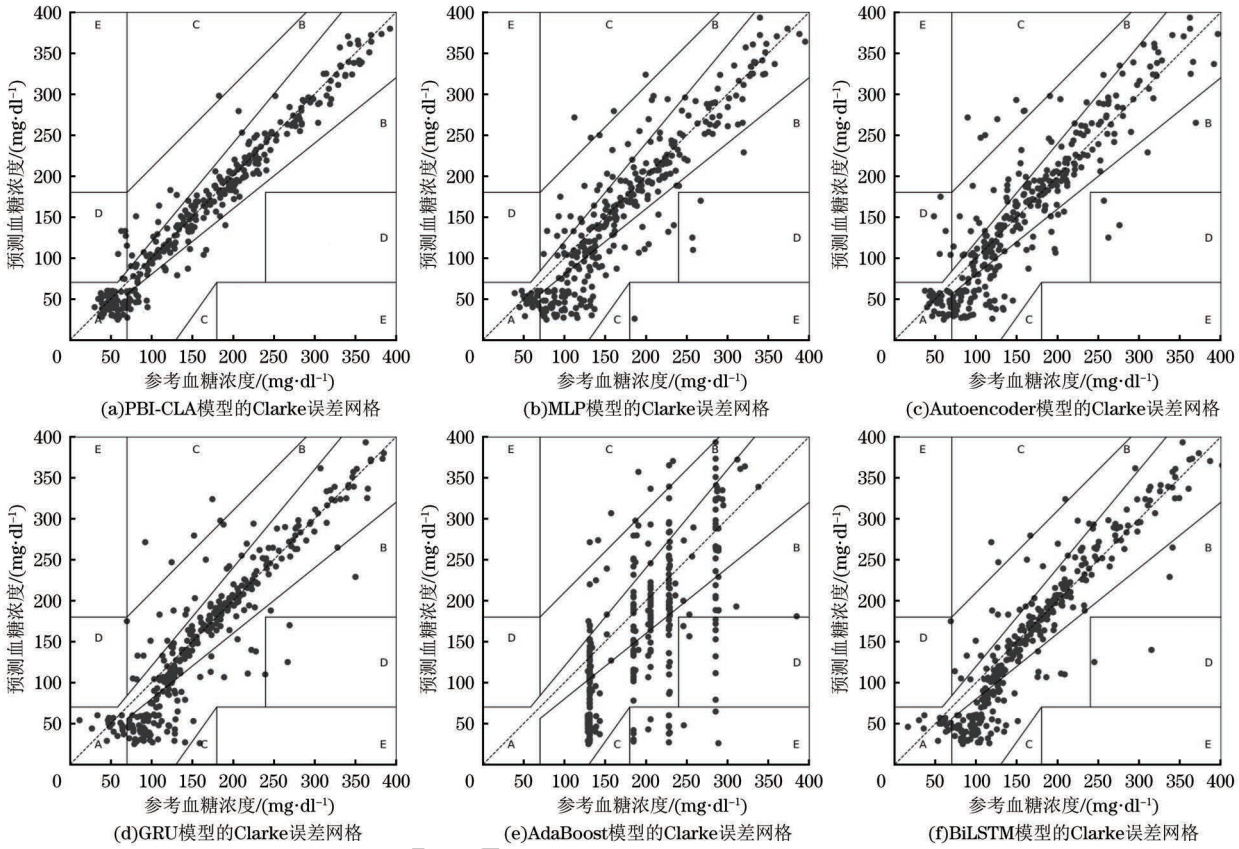


图 4 Clarke 误差网格预测结果

Fig. 4 The prediction results of the Clarke error grid

图 4 的结果显示，MLP 模型、Autoencoder 模型和 BiLSTM 模型部分预测值要明显小于实际真实值，样本点偏离到 B 区域较多。与其他 4 种方法相比，PBI-CLA 模型提前 1 h 预测血糖浓度的准确度更好，其误差范围明显小于其他 5 种模型。

对于 6 种模型，本文继续对延后时间（提前多小时）进行了血糖浓度预测。不同预测方法的 RMSE、MAE 和 MAPE 的性能指标如表 3 所示。

与其他 5 个预测模型相比，PBI-CLA 模型的 RMSE、MAE 和 MAPE 最小。提前 2 h 和 3 h 的血糖浓度预测结果表明，PBI-CLA 模型无论是在短期（1 h 内）还是在延后时间均具有良好的预测精度。

同时，通过对延后时间的血糖浓度预测进行实验，发现随着时间的延长，模型预测的准确率在不断下降，这是因为长距离的时间序列导致序列中相邻结点的依赖性不断降低，时间关联性逐渐下降。

3.7 消融实验分析

为了验证模型中每个模块在整体模型中的预测效果，本文采取分别移除每个模块的方式进行实验。

实验结果显示，当移除 PBI-CLA 模型中任意模块后，模型在血糖浓度预测上准确率会有不同程度的下降，其中移除 CNN 模块和 LSTM 模块后 RMSE 值升高最为明显。但在移除 Attention 层后，模型准确率虽有所下降，但幅度较为有限，这主要归因于数据集中胰岛素剂量序列包含大量零值，限制了注意力机制在特定情境下发挥其应有的聚焦与增强作用。

去掉 LSTM 模块的模型预测效果最差，说明 LSTM 模块在预测血糖浓度过程中是最为重要的。其次是 CNN 模块，这归因于 CNN 模块能够有效捕捉患者的静态特征和动态序列特征，预测结果如表 4 所示。

表 3 预测模型结果比较

Table 3 Comparison of predictive model results

模型	评价指标	预测时间为 1 h	预测时间为 2 h	预测时间为 3 h
MLP	RMSE	56.506 3	64.315 6	75.262 5
	MAE	37.343 8	49.563 2	56.628 4
	MAPE/%	34.935 1	44.162 4	49.516 4
GRU	RMSE	47.675 1	56.741 2	61.231 5
	MAE	29.182 4	33.549 2	39.135 9
	MAPE/%	30.255 4	34.648 5	38.956 4
Autoencoder	RMSE	58.277 4	71.562 1	74.246 2
	MAE	34.752 0	58.652 3	65.546 8
	MAPE/%	28.852 0	36.560 4	48.265 4
AdaBoost	RMSE	101.018 6	105.326 4	112.653 2
	MAE	73.381 9	78.365 1	84.623 5
	MAPE/%	74.089 5	76.415 2	81.634 1
BiLSTM	RMSE	48.542 2	53.628 9	62.625 4
	MAE	31.954 0	38.954 5	49.265 1
	MAPE/%	33.227 1	37.628 5	47.612 3
PBI-CLA	RMSE	**35.724 1**	**40.516 7**	**47.315 6**
	MAE	**19.845 9**	**26.871 9**	**35.959 1**
	MAPE/%	**18.612 5**	**23.945 3**	**30.173 9**

表 4 消融实验模型预测结果比较

Table 4 Comparison of ablation experiment model prediction results

模型	评价指标	预测时间为 1 h	预测时间为 2 h	预测时间为 3 h
PBI-CLA	RMSE	**35.724 1**	**40.516 7**	**47.315 6**
	MAE	**19.845 9**	**26.871 9**	**35.959 1**
	MAPE/%	**18.612 5**	**23.945 3**	**30.173 9**
PBI-CLA w/o LSTM	RMSE	52.481 1	65.954 8	78.264 8
	MAE	35.747 5	43.578 4	53.790 6
	MAPE/%	35.336 4	41.628 5	49.945 6
PBI-CLA w/o CNN	RMSE	47.675 1	59.648 4	74.246 2
	MAE	29.182 5	42.890 8	51.648 1
	MAPE/%	30.255 4	36.560 4	48.265 4
PBI-CLA w/o Attention	RMSE	41.551 9	49.089 1	59.767 2
	MAE	22.452 4	29.763 5	38.198 2
	MAPE/%	20.972 7	28.083 5	37.890 8

由表 5 可以看出，PBI-CLA 模型预测准确度明显高于其他模型，实验中 Clarke 误差网格中的样本点分布在 A 区范围内的数量较多且较为集中。在多个小时的预测中，误差率也比其他 3 个模型要更小，因此，PBI-CLA 模型中的 CNN 模块、LSTM 模块和 Attention 模块都不可或缺。

通过分析消融实验的结果表明：具有所有模块组合的 PBI-CLA 模型在预测性能上是最优的；模型中 LSTM 模块的结构能帮助模型更好地捕捉血糖浓度序列时间上的联系；模型的 CNN 模块和 Attention 模块对特征的选取具有重要作用。

表 5 消融实验模型 Clarke 误差网格样本点占比比较

Table 5 Comparison of the percentage of sample points in the Clarke error grid of the ablation experiment model %

模型	区域	预测时间为 1 h	预测时间为 2 h	预测时间为 3 h
PBI-CLA	A	94.10	89.45	83.94
	B	5.01	3.15	10.43
	C	0.23	3.76	3.45
	D	0.66	3.64	2.18
	E	0.00	0.00	0.00
PBI-CLA w/o LSTM	A	79.24	74.08	67.25
	B	16.94	17.74	14.63
	C	2.18	3.85	3.68
	D	1.32	3.97	8.96
	E	0.32	0.36	5.48
PBI-CLA w/o CNN	A	82.15	77.85	71.17
	B	14.60	9.67	16.75
	C	2.40	4.68	5.43
	D	0.85	3.76	4.60
	E	0.00	4.04	2.05
PBI-CLA w/o Attention	A	85.78	83.16	75.20
	B	13.12	11.47	14.34
	C	0.34	1.60	3.60
	D	0.42	2.57	4.86
	E	0.34	1.20	2.00

3.8 模型复杂度分析

本节通过模型参数总量、浮点数运算量和运行时间 3 个指标进行定量分析，所有模型均在同一数据集进行实验。由表 6 可以看出，在保证模型误差的基础上，PBI-CLA 模型在浮点数运算量和运行时间上比其他深度学习模型具有优势，在模型参数总量上比 MLP 要稍差一些，这是由于卷积层的卷积操作和注意力层提高了整体的参数量。实验结果表明，PBI-CLA 模型复杂度较低，能够完成血糖浓度预测这样的时序回归任务。

表 6 模型复杂度分析结果

Table 6 Model complexity analysis results

模型	参数总量/10^6	浮点数运算量/10^9	运行时间/s
MLP	1.35	1.28	65.3
GRU	1.56	1.47	70.6
Autoencoder	1.41	1.56	67.5
AdaBoost	0.86	0.87	56.9
BiLSTM	1.95	1.93	97.5
PBI-CLA	1.35	1.27	64.4

计算机工程 www.ecice06.com

4 结束语

针对传统模型缺少时序上的关联关系的问题，本文提出多变量血糖浓度预测模型 PBI-CLA。该模型在 LSTM 模型基础上引入 CNN，通过一维卷积提取患者的血糖浓度序列和胰岛素序列特征，输出到 LSTM 后捕捉序列特征的潜在时序关系，并通过注意力机制赋予血糖浓度值不同的权重达到准确预测下一时间节点血糖浓度的效果。实验结果表明，PBI-CLA 模型在误差和稳定性方面均优于现有模型，并且在模型性能和复杂度方面与其他深度学习模型相比具有较大优势。但是本文模型目前只输入了患者基本信息、血糖和胰岛素数据，易受其他因素干扰，未来将引入患者各项生理指标进行预测，提高模型的泛化能力。

参考文献

[1] SUN H, SAEEDI P, KARURANGA S, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045[J]. Diabetes Research and Clinical Practice, 2022, 183: 109119.

[2] 彭艳琼, 谢楠, 敬敏, 等. 基于信息化血糖监测系统建立血糖基准报告研究[J]. 中国全科医学, 2021, 24(33): 4255-4260.
PENG Y Q, XIE N, JING M, et al. Development of a blood glucose benchmark report based on the information glucose monitoring system[J]. Chinese General Practice, 2021, 24(33): 4255-4260. (in Chinese)

[3] WADGHIRI M Z, IDRI A, EL IDRISSI T, et al. Ensemble blood glucose prediction in diabetes mellitus: a review[J]. Computers in Biology and Medicine, 2022, 147: 105674.

[4] KONONENKO I V, SMIRNOVA O M, MAYOROV A Y, et al. Classification of diabetes, world health organization 2019. what's new[J]. Diabetes Mellitus, 2020, 23(4): 329-339.

[5] RODEN M, SHULMAN G I. The integrative biology of type 2 diabetes[J]. Nature, 2019, 576(7785): 51-60.

[6] KRINSLEY J S, CHASE J G, GUNST J, et al. Continuous glucose monitoring in the ICU: clinical considerations and consensus[J]. Critical Care, 2017, 21: 1-8.

[7] KHADILKAR K S, BANDAR T, SHIVANE V, et al. Current concepts in blood glucose monitoring[J]. Indian Journal of Endocrinology and Metabolism, 2013, 17(3): 643-649.

[8] TEMURTAS H, YUMUSAK N, TEMURTAS F. A comparative study on diabetes disease diagnosis using neural networks[J]. Expert Systems with Applications, 2009, 36(4): 8610-8615.

[9] RODBARD D. Continuous glucose monitoring: a review of successes, challenges, and opportunities[J]. Diabetes Technology & Therapeutics, 2016, 18(2): 203-213.

[10] CHEN C, ZHAO X L, LI Z H, et al. Current and emerging technology for continuous glucose monitoring[J]. Sensors, 2017, 17(1): 182.

[11] 张彤, 孟亮. 基于注意力神经网络的糖尿病视网膜病变识别[J]. 计算机工程与科学, 2022, 44(3): 479-485.
ZHANG T, MENG L. Recognition of diabetic retinopathy based on attention neural network[J]. Computer Engineering & Science, 2022, 44(3): 479-485. (in Chinese)

[12] WALLIS C. How artificial intelligence will change medicine[J]. Nature, 2019, 576(7787): 48-48.

[13] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[14] PEREZ-GANDIA C, FACCHINETTI A, SPARACINO G, et al. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring[J]. Diabetes Technology & Therapeutics, 2010, 12(1): 81-88.

[15] GEORGA E I, PROROPAPPAS V C, ARDIGO D, et al. Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression[J]. IEEE Journal of Biomedical and Health Informatics, 2012, 17(1): 71-81.

[16] HIDALGO J I, COLMENAR J M, KRONBERGER G, et al. Data based prediction of blood glucose concentrations using evolutionary methods[J]. Journal of Medical Systems, 2017, 41: 1-20.

[17] SHUVO M M H, ISLAM S K. Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(3): 1612-1623.

[18] WANG W B, TONG M, YU M. Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization[J]. IEEE Access, 2020, 8: 217908-217916

[19] JOHNSON A E W, BULGAERLLI L, SHEN L, et al. MIMIC-IV, a freely accessible electronic health record dataset[J]. Scientific Data, 2023, 10(1): 221-232.

[20] MAHARDIKA T N Q, FUADAH Y N, JEONG D U, et al. PPG signals-based blood-pressure estimation using grid search in hyperparameter optimization of CNN-LSTM[J]. Diagnostics (Basel, Switzerland), 2023, 13(15): 2566.

[21] ADNAN M, ALAROOD A A S, UDDIN M I, et al. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models[J]. Peerj Computer Science, 2022, 8: e803.

[22] BARROW D K, CRONE S F. A comparison of AdaBoost algorithms for time series forecast combination[J]. International Journal of Forecasting, 2016, 32(4): 1103-1119.

[23] KAMYSHANSKA H, MEMISEVIC R. The potential energy of an autoencoder[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(6): 1261-1273.

[24] LI B Y, LENG L, SHEN S, et al. Efficient deep spiking multilayer perceptrons with multiplication-free inference[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(4): 7542-7554.

[25] KHODABANDELOU G, JUNG P G, AMIRAT Y, et al. Attention-based gated recurrent unit for gesture recognition[J]. IEEE Transactions on Automation Science and Engineering, 2020, 18(2): 495-507.

[26] HOU L, ZHANG J, WU O, et al. Method and dataset entity mining in scientific literature: a CNN+BiLSTM model with self-attention[J]. Knowledge-Based Systems, 2022, 235: 107621.

文字编辑 索书志
栏目编辑 宋 圆

计算机工程 www.ecice06.com