文章编号：1006-3080(2026)02-0247-10 **DOI:** 10.14135/j.cnki.1006-3080.20250603001

基于深度学习的工控系统网络攻击多分类检测

王庚辰， 姜庆超， 颜学峰
（华东理工大学信息科学与工程学院, 上海 200237）

摘要：近年来，针对工业控制系统(Industrial Control System, ICS)的网络物理攻击事件频发，工控系统的异常检测成为安全防护的关键技术。传统的异常检测方法通常将问题简化为二元分类，难以满足实际需求。为了更精确地定位攻击源头并实现系统状态的快速恢复，需要对ICS异常状态进行更细致的划分。本文提出了一种基于深度学习的新型工控异常检测及攻击分类模型，结合卷积神经网络(CNN)、双向长短期记忆网络(BiLSTM)以及注意力(Attention)机制的优势，通过CNN提取数据包的空间特征，利用BiLSTM捕捉数据包间的时间依赖性，并引入注意力机制进一步聚焦关键的时间步信息，从而实现对工控系统网络攻击的高精度检测。实验结果表明，该模型在检测准确率等评价指标上优于现有的工业入侵检测系统，并且在处理不平衡数据集时表现出色，为工控系统的安全防护提供了新的解决方案。

关键词：工业控制系统；异常检测；网络攻击；攻击分类；深度学习

中图分类号：TP391 **文献标志码：**A

工业控制系统(Industrial Control Systems, ICS)是由硬件、软件、网络及操作人员协同构成的复杂系统，广泛应用于水处理、石油天然气、电力、交通等国民经济关键领域，承担着工业基础设施的实时监控、流程管理以及自动化控制等任务。随着工业4.0推动运营技术(OT)与信息技术(IT)网络的深度融合，传统封闭的工业控制系统逐步开放互联，但在提升效率的同时也暴露于数据篡改、恶意指令注入、通信协议劫持等新型网络攻击的威胁之下。此类攻击的破坏力远超常规数据泄露的风险[1-2]，针对ICS的网络攻击可能直接引发生产瘫痪、环境灾难以及公共安全危机。传统的基于规则库或静态阈值的检测机制难以适应工业场景中协议异构、设备多元的复杂环境，同时也无法有效地识别未知的攻击模式。为此，基于深度学习的异常检测技术成为工控系统网络安全防护体系的关键，通过持续学习网络流量与设备交互的正常基准，实时捕捉系统响应中偏离预期设定的细微异常，从而可以在攻击渗透时实现精准告警，为技术人员在不同网络攻击发生时及时处置风险争取黄金时间，将攻击发生后造成的损失降到最低。

为了应对以上问题，近年来以机器学习算法为代表的模型在工控系统网络攻击检测领域得到了广泛应用。Khan等[3]尝试通过机器学习方法设计一种入侵检测系统以检测攻击行为。在该系统中，通过主成分分析(PCA)、典型相关分析(CCA)和独立成分分析(ICA)等方法提取显著特征，随后将这些特征进行融合，并分别通过布隆过滤器(Bloom Filter)和K近邻算法(KNN)进行处理。Umer等[4]将用于工控系统入侵和异常检测的4种机器学习方法，即监督学习、半监督学习、无监督学习和强化学习进行比较，指出经典的机器学习方法依赖高质量标注数据、训练过程不稳定以及难以应对复杂入侵场景的缺点。

随着工业控制系统的复杂度不断提升，数据集的维度也在持续增长，传统的机器学习算法逐渐难以应对ICS中复杂的攻击场景。与传统机器学习相

收稿日期：2025-06-03
基金项目：国家自然科学基金(62433004)
作者简介：王庚辰(2000—)，男，河北石家庄人，硕士生，主要研究方向为基于深度学习的工控异常检测及攻击分类。E-mail: 15081191867@163.com
通信联系人：姜庆超, E-mail: qchjiang@ecust.edu.cn ；颜学峰, E-mail: xfyan@ecust.edu.cn
引用本文：王庚辰, 姜庆超, 颜学峰. 基于深度学习的工控系统网络攻击多分类检测 [J]. 华东理工大学学报(自然科学版), 2026, 52(2): 247-256.
Citation: WANG Gengchen, JIANG Qingchao, YAN Xuefeng. Multi-Class Detection of Cyber Attacks in Industrial Control Systems Based on Deep Learning[J]. Journal of East China University of Science and Technology, 2026, 52(2): 247-256.

比,深度学习作为机器学习的延伸,无需人工干预特征提取,能够自动从高维数据中学习到有效的特征。在ICS这种具有大量高维数据的场景中,基于深度学习的入侵检测系统展现出了更为优越的性能。Wang等[5]提出了一种堆叠深度学习方法来识别针对SCADA系统的恶意攻击,检测绕过传统IDS和防火墙的恶意入侵,所提出的方法优于独立的深度学习模型和一些先进的算法,包括最近邻、随机森林、朴素贝叶斯、Adaboost、支持向量机(SVR)和OneR(One Rule)。卷积神经网络(CNN)是一种常见的深度学习方法,也被广泛应用到工控系统网络入侵检测中。例如,张靖雯[6]提出了一种基于马氏距离的特征映射方法,将一维的工控流量特征数据映射成适合CNN处理的二维特征矩阵的形式,大大提高了攻击二分类和多分类问题的检测准确率。注意力(Attention)机制是一种常用于提高神经网络模型性能的深度学习技术。Liu等[7]利用Attention机制和长短期记忆网络(LSTM)对时间序列数据进行分析,生成安全态势预测结果并提高评估精度。

尽管上述方法在工控系统网络攻击入侵检测中作出了一定贡献,但仍存在明显的不足和缺陷。首先,这些方法在捕获复杂数据的能力上存在局限性,难以从多协议混合通信流量或设备状态时序变化的数据中充分提取非线性特征。其次,部分方法依赖于人工特征选择,易受到先验知识偏差的影响,导致模型在新型攻击场景下的泛化能力不足。此外,传统算法的计算效率在高维工控数据场景下存在瓶颈,比如在处理每秒数万条SCADA系统日志时响应延迟过高,这将严重影响系统的实时检测性能。为解决这些问题,亟需引入更高效的特征提取技术,如基于深度学习的更优解决方案。

本文首先通过CNN自动捕获数据中的空间关联特征,并引入双向长短期记忆网络(BiLSTM),从CNN提取的特征中学习前后双向时间依赖关系(如攻击指令的阶段性传递行为);随后利用注意力机制动态分配不同时间步特征的权重,将融合时空特征的数据输入分类网络,完成工控系统网络攻击的多分类检测任务;设计一种检测工控系统网络攻击的多分类模型,基于CNN-BiLSTM-Attention混合架构实现多层次特征融合。

1 基于CNN-BiLSTM-Attention的工控系统网络攻击多分类检测模型

1.1 CNN

在经典的CNN结构中,通常包含卷积层、池化层和全连接层3大核心组件。

卷积层通过局部感知野的权值共享机制,逐层提取输入数据的空间关联特征,其中每个神经元的激活响应仅与前一层的局部邻域信息相关联[8]。通常情况下,卷积层可以表示为:

$$x_j^l = f\left(\sum_{i\in M_j} x_j^{l-1}k_{ij}^l + b_j^l\right) \tag{1}$$

其中,l表示所处的层数,k代表该层的卷积核,M_j为输入数据,b_j代表偏置,f表示所使用的激活函数,常用的激活函数有Sigmoid、ReLU等。

池化层则采用非线性压缩策略,在保留关键特征的同时降低特征图的分辨率,从而增强模型对输入数据几何形变的鲁棒性。池化层可以表示如下:

$$x_j^l = f\left(\beta_j^l \mathrm{down}\left(x_j^{l-1}\right) + b_j^l\right) \tag{2}$$

其中,$\mathrm{down}(\cdot)$表示所使用的池化函数,如平均池化(Average Pooling)和最大池化(Max Pooling)。

全连接层通过全连接结构对高层抽象特征进行全局整合,最终输出目标类别的概率分布。这种分层递进的结构设计使得CNN在图像识别等任务中展现出优异的特征表达能力和计算效率。CNN的典型结构如图1所示。

CNN卷积层采用权值共享机制,使卷积核在整个输入数据范围内滑动时保持相同的参数,从而有效降低计算复杂度[9]。此外,CNN的网络结构具备良好的扩展能力,可适应不同格式的数据输入。本文利用CNN的特征自动学习、参数共享以及扩展性优势,将其应用于工控系统网络攻击分类任务,通过分析网络流量、系统日志等数据,挖掘关键特征,以识别异常行为和潜在安全威胁。然而,由于CNN主要关注局部空间特征,在检测涉及时间序列相关性的异常行为时仍存在一定局限性,所以本文引入了BiLSTM。

1.2 BiLSTM

长短期记忆网络(LSTM)是一种高效的深度学习模型,专门设计用于处理具有长期依赖关系的序列数据。与传统循环神经网络(RNN)不同,LSTM通过设计记忆单元(Memory Cell),利用输入门(Input Gate)、遗忘门(Forget Gate)和输出门(Output Gate)控制信息的流动,从而有效缓解RNN存在的梯度消失和梯度爆炸问题[10]。LSTM通过门控机制对信息进行筛选,使得网络可以长期存储重要的信息,同时丢弃不必要的信息。LSTM的整体结构如图2所示。

输入门决定当前时刻的新信息x_t是否应该被写入记忆单元。输入门i_t可以被描述为:

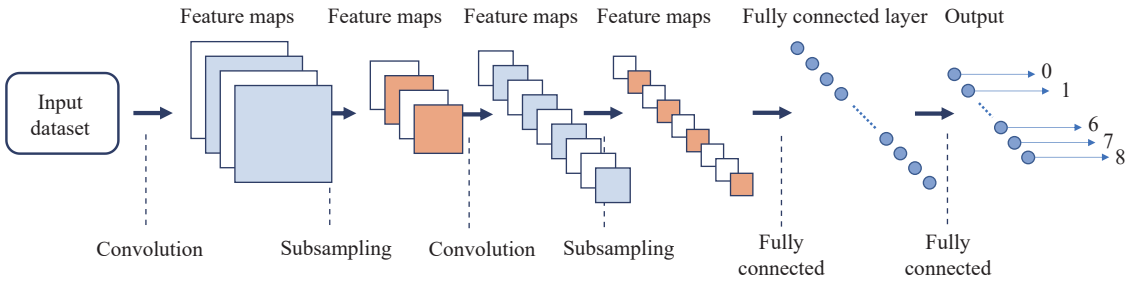

图 1　CNN 结构图

Fig. 1　Structure diagram of CNN

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

其中，x_t 是当前时间步的输入数据；h_{t-1} 是上一时间步的隐藏状态，即记忆单元的输出；$\boldsymbol{W}_i$ 和 b_i 分别是输入门的权重矩阵和偏置项，通过 sigmoid 激活函数 $\sigma(\cdot)$ 将计算结果映射到 0 ~ 1 之间，从而决定信息被保留的比例。

遗忘门决定前一时间步的记忆单元状态 c_{t-1} 对当前时刻的影响程度。遗忘门 f_t 可以被描述为：

$$f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

其中，$\boldsymbol{W}_f$ 和 b_f 分别是遗忘门的权重矩阵和偏置项。当 f_t 取值接近 1 时，表示保留更多的过去信息；而当 f_t 取值接近 0 时，表示遗忘大部分的过去信息。

输出门决定当前时刻的隐藏状态 h_t（即输出值）有多少信息应该被输出。输出门 o_t 可以被描述为：

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

其中，$\boldsymbol{W}_o$ 和 b_o 分别是输出门的权重矩阵和偏置项。

LSTM 的核心是记忆单元，它的更新公式如下：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, x_t] + b_c) \tag{6}$$

其中，c_t 是当前时间步的记忆单元状态，$\boldsymbol{W}_c$ 和 b_c 分别是用于计算新的候选记忆的权重矩阵和偏置项，非线性激活函数 tanh 将候选记忆单元的值缩放到 $(-1, 1)$ 之间，使得 LSTM 具有非线性建模的能力。记忆单元的更新就是先通过遗忘门决定旧记忆 c_{t-1} 需要保留多少信息，再通过输入门决定新的信息是否应该加入，通过两者的加权求和得到当前时间步的记忆状态 c_t [11]。

隐藏状态 h_t 的更新由 c_t 以及 o_t 共同决定，表示如下：

$$h_t = o_t \odot \tanh(c_t) \tag{7}$$

其中，$\odot$ 代表了点积计算。这一机制确保了 LSTM 能够灵活选择要输出的内容，同时避免信息过载。

在传统的 LSTM 中，信息是按照时间步的顺序从前往后传播的，即网络在计算当前时间步的隐藏状态 h_t 时，仅依赖于过去时间步的信息 $x_1, x_2, \cdots, x_t$。

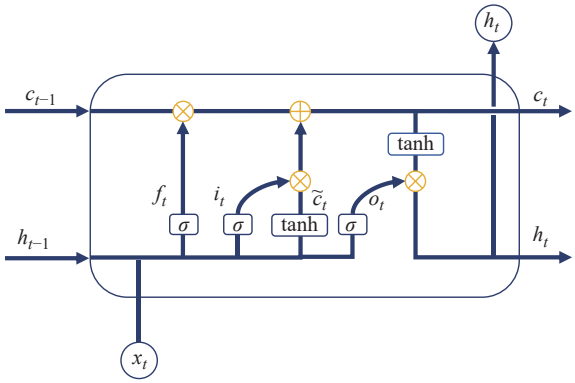

图 2　LSTM 结构图

Fig. 2　Structure diagram of LSTM

然而，很多实际任务（例如时间序列预测任务中）需要用到整个序列甚至包括未来的信息以获得更精准的预测，所以 BiLSTM 通过在 LSTM 结构的基础上添加一个逆向传播的 LSTM 层，通过双向信息流动达到获取双向信息精准预测的目的[12]。本文 BiLSTM 结合前向和后向 LSTM，能够捕获序列中过去和未来的信息，它不仅可增强模型的上下文理解能力，还在复杂的时间序列预测任务中表现出色。BiLSTM 的整体结构如图 3 所示。

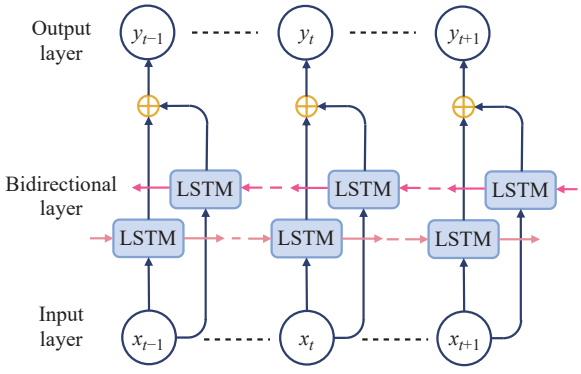

图 3　BiLSTM 结构图

Fig. 3　Structure diagram of BiLSTM

1.3　系统模型

本文提出的基于 CNN-BiLSTM-Attention 的工控系统网络攻击多分类检测模型的整体结构如图 4 所示。由图可知，模型由输入模块、CNN 模块、BiLSTM

模块、Attention 模块、全连接输出模块 5 部分组成，分别用序号①到⑤表示，图中粗箭头表示模块间的传递顺序，细箭头表示模型中数据的传递方向。

(1)输入模块

输入模块是模型的起始，主要对输入模型的数据进行预处理的操作。本文中的数据预处理部分主要对原始训练数据进行特征与标签分离、标签编码、数据清洗、数据标准化、数据批量划分、数据格式转换等操作，保证输入数据的格式和数据质量符合下一模块中卷积神经网络的输入要求。

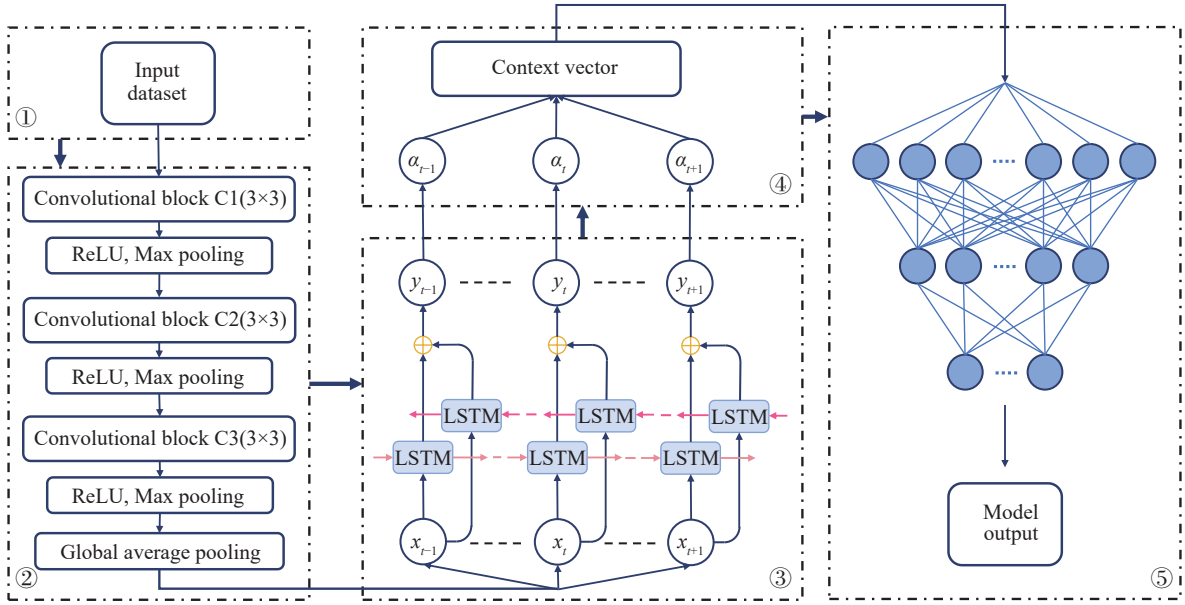

图 4 CNN-BiLSTM-Attention 模型结构图

Fig. 4 Structure diagram of CNN-BiLSTM-Attention model

(2)CNN 模块

CNN 模块主要对经过输入模块预处理后的输入数据进行局部特征提取的操作。数据依次通过卷积层、ReLU 激活函数和池化层，考虑到模型过拟合的风险以及模型的实效性，本文搭建了一个相对简单的卷积模块，并在每个卷积层之后添加了一个 ReLU 激活函数，该函数具有计算简单、收敛速度快的优点，能够有效缓解梯度消失的问题。

(3)BiLSTM 模块

卷积模块的输出经过形状变换调整为时间步 × 特征维度的形式，以适配 BILSTM 模块的输入格式，同时要确保每个时间步的特征维度的一致。经过该模块处理后，模型获得了融合局部空间模式与全局时序关系的高阶特征表示，能够更有效刻画工控网络流量在攻击发生过程中的连续变化特征与阶段性演化规律，为后续 Attention 机制模块提供更加充分的判别依据。

(4)Attention 机制模块

Attention 机制模块中集成的加权注意力机制可以对 BiLSTM 模块提取的时序特征进行动态加权，以突出关键的时间步信息，使得模型能够动态关注最重要的时间步信息，提升分类效果的同时还能够高效提取关键特征，计算复杂度低且计算速度快，适合于工控系统故障检测中时序数据分类的任务[13]。

Attention 模块首先将 BiLSTM 模块输出的隐藏状态序列送入一个全连接层，通过线性变换计算每个时间步的注意力分数，然后利用 Softmax 函数将注意力分数在时间维度上进行归一化处理，得到反映各时间步重要性的注意力权重分布。随后利用该权重对原始隐藏状态进行加权求和，生成一个融合了全局时序信息的上下文向量，该向量更能突出关键时间步的信息[14]。最后，将该上下文向量传递至全连接输出模块，以在下一模块实现对输入样本的精确分类。

注意力权重计算部分表示如下：

$$e_t = \boldsymbol{W}_a \cdot h_t + b_a \tag{8}$$

其中，e_t 表示注意力分数，h_t 表示时间步 t 的隐藏状态，$\boldsymbol{W}_a$ 表示全连接层的权重矩阵，b_a 是全连接层的偏置项，通过全连接层将每个时间步的隐藏状态 h_t 映射为注意力分数 e_t。

注意力权重归一化部分表示如下：

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t'=1}^{T} \exp(e_{t'})} \tag{9}$$

其中, α_t 表示时间步 t 的归一化注意力权重, T 表示时间步总数。使用 Softmax 函数将注意力分数 e_t 转换为概率分布 α_t, 并确保所有权重之和为 1, 满足概率分布的性质。

上下文向量计算部分表示如下:

$$\boldsymbol{c} = \sum_{t=1}^{T} \alpha_t \cdot h_t \tag{10}$$

其中, $\boldsymbol{c}$ 表示上下文向量。对 h_t 进行加权求和得到 $\boldsymbol{c}$, 即将时间序列每个时间步的隐藏状态这一局部特征聚合为上下文向量这一全局特征[15]。

(5)全连接输出模块

全连接输出模块主要是将注意力机制模块生成的上下文向量 $\boldsymbol{c}$ 映射为最终的分类结果。上下文向量进入本模块后, 被输入到全连接层进行进一步的特征提取与分类, 并使用 ReLU 激活函数增强特征表达能力。最后通过全连接层将特征映射到最终的类别空间, 以获得预测结果。模型在训练过程中采用交叉熵损失函数, 并通过反向传播和优化算法不断优化网络参数, 以提高分类性能。

2　实验结果及分析

本实验在一台配置有 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz 的处理器、64 GB 的机带 RAM 以及 windows 10 操作系统的计算机上进行。实验软件方面采用 python 3.13.0 版本进行编程, 使用 Anaconda3 构建 python 环境, 将 Pycharm 作为集成开发环境, 并基于 Pytorch 平台搭建模型框架。

2.1　数据集

为证明本文提出的模型在工控系统异常检测及攻击分类上的有效性, 本实验使用密西西比大学莫里斯等开发的天然气管道测试平台中的 SCADA 系统记录的真实数据构成的数据集[16]。该数据集共包含 27 个工控流量特征, 这些特征数据根据数据包属性可大致分为网络流量特征和有效载荷内容特征两类。

网络流量特征描述了 SCADA 系统的通信模式, 这部分特征包括数据包中用于请求与响应的设备地址、存储器位置以及字节长度等信息, 同时还记录了通信请求发送到响应接收之间的时间间隔。SCADA 系统网络的拓扑结构和服务较为固定, 但一些针对 SCADA 系统的网络攻击可能会改变系统的网络通信模式, 所以网络流量特征常用于描述系统正常的流量模式, 以便于检测恶意行为的发生。

有效载荷内容特征则描述了 SCADA 系统的当前状态, 这部分特征包括响应/命令功能码, 气体压力初始值、当前压力测量值以及系统控制模式等, 有效载荷内容特征常用于检测导致设备(如工业控制系统核心设备 PLC)异常行为的攻击。

除了上述两类特征外, 数据集还包括一个标签特征, 用于区分正常样本和攻击样本。数据集样本中包括系统正常运行情况的数据和 7 类不同网络攻击发生时记录的工控流量数据, 样本数共 97019 条, 其中正常样本共 61156 条, 攻击样本共 35863 条, 数据集中的攻击分类如表 1 所示。

2.2　数据预处理

2.2.1　*数据清洗*　由于数据集中的记录存在特征值不完整或缺失的情况, 数据预处理的第一步是使用适当的值来填补这些空缺。本文采取的方法是保留先前值的方法, 这是由于记录中未给出数值特征受到时间依赖性的影响, 所以空缺或不完整的值应当与先前值相关且相等, 所以这部分数据使用了保留先前值的方法来补全, 这不仅符合实际情况还能保持原始数据的分布[17]。

2.2.2　*特征与标签提取*　数据集共包含 27 个特征, 其中前 26 个是工控流量特征, 最后一个是标签特征, 利用数据切片操作, 将前 26 个特征作为特征矩阵, 标签特征作为标签向量, 由于标签向量以字符串的形式存在, 所以将它们编码成数字 0 到 7, 以便于后续模型的处理。

2.2.3　*数据归一化*　本实验使用 z-score 归一化方法对特征进行标准化, 该方法将原始数据中的每个特征值减去该特征的均值, 然后除以标准差, 通过这种变换将每个特征转换为均值为 0、标准差为 1 的标准正态分布。归一化后的数据可以让模型更准确地捕捉数据的空间和时间特征, 有效提高模型训练的效率和稳定性。

2.2.4　*基于滑动窗口的多通道时间序列构造*　在使用深度递归网络时, 单个数据包的特征可能不足以揭示攻击行为的规律, 但如果将多个连续的数据包构造成一个时间序列后, 模型中的 BiLSTM 模块就能够学习到特征随时间的变化和演变趋势, 从而更有效地识别异常状态和攻击行为。

本文将每 8 个连续的数据包组成一个时间序列, 并利用滑动窗口的方式将原始特征数据转换为多通道时间序列数据。滑动窗口的方法可以从连续的数据包中提取出相互重叠的固定长度的子序列, 这样可以确保样本中局部的时序信息不被遗漏, 并且每个窗口内的记录之间存在紧密的时间关联, 同时还可以从有限的数据中生成更多的样本, 提高模型训练的有效数据量。通过以上处理, 每个时间序

表1 数据集中的攻击分类

Table 1 Attack classification in the dataset

Abbreviation	Label	Label type	Number
Normal	b0	Normal behavior	61 156
NMRI	b1	Naive malicious response injection attack	2 763
CMRI	b2	Complex malicious response injection attack	15 466
MSCI	b3	Malicious state command injection attack	782
MPCI	b4	Malicious parameter command injection attack	7 637
MFCI	b5	Malicious function command injection attack	573
DOS	b6	Denial-of-service attack	1 837
Recon	b7	Reconnaissance attack	6 805

列包含8个时间步以及3个连续窗口组成的多通道,再将3个通道的数据堆叠起来,构成形状为[3, 8, 26]的单个样本数据。

2.2.5 *数据转换* 为了适应基于Pytorch搭建的系统模型的输入形式,将生成的多通道时间序列数据转换为模型可以处理的Pytorch张量。

整个预处理流程确保了SCADA数据经过适当的特征提取、归一化以及时序构造后,能够以合适的形状和格式输入到后续的系统模型中,从而有效地捕捉数据中的空间和时序信息。

2.3 评价指标

本文中选取的模型性能评价指标有准确率(Accuracy)、精确率(Precision)、召回率(Recall)和F_1值。各评价指标[18-19]的计算方法如下所示:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{11}$$

$$\text{Precison} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$F1 = \frac{2\cdot\text{Recall}\cdot\text{Precision}}{\text{Recall}+\text{Precision}} \tag{14}$$

其中,TP表示样本本身为正常样本,模型也将其划分为正常样本的样本数;TN表示样本本身为攻击样本,模型也将其划分为攻击样本的样本数;FP表示样本本身为攻击样本,但模型将其分为正常样本的样本数;FN表示样本本身为正常样本,但模型将其划分为攻击样本的样本数[18]。

2.4 结果与分析

本文将预处理后的数据按80%和20%的比例划分为训练集和测试集,训练批次大小设置为32,学习率设置为0.001,训练周期设置为100。根据上文选取的模型评价指标,本模型对于工控系统网络攻击分类在多分类场景下的实验评估结果如表2所示。

表2 系统模型的评价指标

Table 2 Evaluation indicators for system model

Accuracy/%	Precision/%	Recall/%	F_1/%
99.05	99.05	99.05	99.04

表2的实验结果表明,本文提出的方法在工控系统多分类任务中展现出优异的分类能力,整体分类准确率可以达到99.05%,精确率99.05%,召回率99.05%,F_1值99.04%,这表明绝大多数工控流量样本均被正确分类,其中指标数值最低的F_1值也达到了99.04%,均高于工业场景安全阈值的要求。综上所述,本文提出的模型性能能够满足实际应用场景下工控系统对于网络攻击造成的异常检测要求。

表3给出了每个样本的分类精确率、召回率和F_1值,从实验结果来看,各个样本的分类指标表现存在一定差异。针对侦察攻击(Recon),其精确率、召回率和F_1值均达到了100.00%,表现出完美的分类效果。而对于DOS攻击,其精确率和F_1值分别为99.72%和97.77%,但召回率仅为95.90%,表明有少部分DOS样本未被全部捕获。对于MFCI攻击,其精确率达到了100.00%,召回率为98.26%,F_1值为99.12%,表现较为优秀。相比之下,NMRI攻击的分类指标较低,MSCI攻击的各项指标也略显不足。因此,本文进行对比实验与消融实验,以探讨上述部分指标偏低的原因。

2.5 对比实验

在对比实验中,将本文提出的CBA模型与近年来用于工控系统异常检测的传统机器学习方法以及

表 3　样本的精确率、召回率和 F_1 值

Table 3　Precision, Recall, and F_1 of the samples

Sample class	Precision/%	Recall/%	F_1/%
Normal	99.12	99.39	99.25
NMRI	99.42	92.41	95.78
CMRI	98.97	99.51	99.24
MSCI	97.40	96.15	96.77
MPCI	97.59	98.03	97.81
MFCI	100.00	98.26	99.12
DOS	99.72	95.90	97.77
Recon	100.00	100.00	100.00

深度学习模型在准确率、精确率、召回率和 F_1 值上进行对比, 结果如表 4 所示。

对比的机器学习方法包括 K 最邻近分类算法(KNN)[20]、K 均值聚类算法(K-means)[21]、SVM 集合方法[22]3 种, 其中 KNN 和 SVM 是经典的机器学习算法, K-means-CAE 算法是 Chang 等[21]在 2019 年提出的一种基于半监督技术的工业控制系统异常检测方法, 是将 K-means 和卷积自编码器方法相结合的方法。对比的深度学习模型包括邓志刚等[23] 在 2021 年提出的 WGAN-GP 数据增强模型, 该模型利用梯度惩罚扩充稀有攻击样本并结合多层感知机进行模型训练; Sokolov 等[24]在 2019 年提出的 LSTM/GRU 模型, 该模型在 ICS 的入侵检测问题中展示了所考虑的递归神经网络体系结构的能力; Mohammad 等[25] 在 2022 年提出的 CNN-LSTM模型, 结合了一维 CNN 与 LSTM 的优势, 在处理不平衡数据集问题上表现优异; Yanika 等[26] 在 2023 年提出的 HEDLF(混合集成深度学习框架)模型, 它结合了深度学习与集成学习, 以提高入侵检测的准确性和鲁棒性, 可用于检测复杂和隐蔽的网络攻击; Lu 等[27] 在 2025 年提出的 BPSO-AHDL-IDS 模型, 是一种创新的自动化混合深度学习入侵检测方法, 将 CNN 与循环神经网络结合作为混合深度学习模型, 用于提取物联网数据集的特征, 以实现入侵的准确检测; Xue 等[28] 在 2025 年提出了一种基于可解释深度学习的工业互联网入侵响应方法 xIIRS, 该方法通过改进解释方法确定防御规则范围, 结合安全约束生成细粒度防御规则, 从而实现入侵响应。

从实验结果可以看出, 本文提出的模型在所有评价指标上均优于其他模型, 表明在使用本模型进行多分类任务时, 能够对不同类型的网络攻击行为进行精准识别。

表 4　对比实验结果

Table 4　Comparative experimental results

Comparative model	Accuracy/%	Precision/%	Recall/%	F_1/%
KNN[20]	91.00	92.00	81.00	85.00
K-means-CAE[21]	95.53	95.43	83.52	89.08
SVM[22]	92.56	92.47	92.56	92.50
WGAN-GP[23]	95.30	96.52	90.55	93.43
LSTM/GRU[24]	91.70	91.77	91.70	91.73
CNN-LSTM[25]	97.75	97.54	95.47	96.49
HEDLF[26]	94.60	94.23	94.60	94.37
BPSO-AHDL-IDS[27]	98.07	98.43	96.50	97.44
xIIRS[28]	98.05	96.70	98.00	97.34
Ours	99.05	99.05	99.05	99.04

传统机器学习方法在本实验中表现较为逊色。KNN 和 SVM 作为经典分类方法, 在小规模数据集上可能表现良好, 但在处理复杂的工控流量数据时, 由于缺乏对深层次特征的建模能力, 其检测性能明显落后。K-means-CAE 方法结合了无监督聚类和自动编码器(CAE), 能够在一定程度上学习数据分布, 但由于未能有效建模时序特征, 其召回率仅有 83.52%, 影响了整体检测能力。

深度学习模型中评价指标较为优秀的 CNN-LSTM 模型虽然能够有效提取时空特征, 但相比本文模型, 在信息提取能力和全局特征学习能力上仍存在不足。HEDLF 模型采用集成学习的方法, 虽然具有一定的鲁棒性, 但在复杂的工控流量数据场景下, 其分类性能仍然逊色于本文模型。而 LSTM/GRU 模型仅依赖于序列建模, 未能充分结合空间特征的提取, 因此在多层次特征学习方面存在一定不足, WGAN-GP 模型采用生成对抗网络(GAN)进行数据增强, 在缓解数据不平衡问题方面具有一定优势, 但以上两种类型仍不及本文模型的综合性能。BPSO-AHDL-IDS 方法虽然能自动搜索最优超参数和神经网络架构, 并利用混合模型提高特征提取能力, 但存在计算复杂度较高、可解释性较弱的缺点; xIIRS 方法虽然有较强的可解释性且防御规则细粒度与适应性较强, 但解释过程依赖额外计算且防御规则生成较为复杂, 所以仍逊于本文模型的综合性能。

综上, 本文提出的 CBA 模型在各项指标上均达到最优, 验证了该方法在工控系统网络攻击入侵检测任务中的优越性。相比于传统机器学习方法和其他深度学习方法, 本模型在准确率、精确率、召回率、F_1 值等关键指标上均有明显提升, 适用于安全性

要求较高的网络攻击检测任务。

2.6 消融实验

首先对系统模型进行训练以获得基准性能,然后确定需要消融的目标组件,对目标组件进行移除,并在保持其他部分不变的前提下,利用相同的数据集和数据预处理方法对模型重新进行训练,最后对比消融前后模型在准确率、召回率、精确率和 F_1 值评价指标上的差异[29]。

在本实验中,我们针对 CNN、BiLSTM 和 Attention 这 3 个关键模块开展了消融实验,搭建了 CNN、BiLSTM、CNN-BiLSTM、CNN-Attention 4 个消融实验模型,实验结果如表 5 所示。

表 5 消融实验评价指标结果

Table 5 Ablation study evaluation metrics results

Model type	Accuracy/%	Precision/%	Recall/%	F_1/%
CNN	95.49	95.64	95.49	94.33
BiLSTM	95.68	95.82	95.68	94.53
CNN-BiLSTM	98.30	97.85	98.24	98.04
CNN-Attention	97.15	97.30	96.85	97.07
CNN-BiLSTM-Attention	99.05	99.05	99.05	99.04

从表 5 的实验结果可以看出,各模块对模型性能的贡献明显且互补。在单独使用 CNN 模块时,模型的准确率为 95.49%,精确率、召回率和 F_1 值也都在 95.00% 左右,这表明仅依靠 CNN 的局部特征提取存在一定的局限。而单独使用 BiLSTM 模块时,模型准确率略有提高,精确率和召回率也相应提高,这说明时序信息的捕捉对模型具有正向作用。将前两个模块结合后,CNN-BILSTM 模型的整体性能有了显著提升,准确率提高到 98.30%,精确率、召回率和 F_1 值均接近 98%,这表明二者结合可以更全面地提取数据中的局部与时序特征,从而实现更精准的分类。将 CNN 和 Attention 模块结合的模型虽然整体性能较单一模块有所提升,但效果仍不如 CNN-BiLSTM 的组合,这表明单独依靠局部特征加上注意力机制而忽视了时序依赖不利于模型性能的提升。

本文提出的 CNN-BiLSTM-Attention 模型达到了消融实验中的最高准确率(99.05%),精确率、召回率和 F_1 值均在 99.00%左右,说明在融合了局部特征提取、时序建模和关键特征加权之后,模型能够充分利用各模块的优势,显著提升整体的分类性能。消融实验结果验证了本文系统模型中的各模块在工控系统异常流量检测任务中的有效性和协同作用,为模型结构的设计和优化提供了有力支持。

3 结 论

为改善工业控制系统网络攻击检测在数据不平衡、数据量大及数据存在高维特征情景下的检测性能,并提高入侵检测系统对于稀有攻击类别的识别能力,本文提出了一种基于 CNN-BiLSTM-Attention 混合架构的工控系统网络攻击多分类检测模型。模型采用 CNN 提取数据的局部时空特征,BiLSTM 建模时间序列依赖,Attention 机制增强了关键特征的权重,三者的融合实现了在复杂工控流量数据下对不同网络攻击类型的精确识别,准确率达到了 99.05%。对比实验和消融实验结果表明,所提出的模型在准确率、精确率、召回率及 F_1 值等指标上均表现优越,与其他模型相比显著提升了稀有攻击检测的能力,并且各个模块的存在均有其合理性。在未来的研究工作中,本文将针对其他深度学习方法进行研究,探讨不同方法在工控系统异常检测领域应用的可能性。

参考文献:

[1] LANGNER R. Stuxnet: Dissecting a cyberwarfare weapon[J]. IEEE Security & Privacy, 2011, 9(3): 49-51.

[2] LEE R M, ASSANTE M J, CONWAY T. Analysis of the cyber attack on the Ukrainian power grid[EB/OL]. (2016-03-18). https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf.

[3] KHAN I A, PI D, KHAN Z U, *et al*. HML-IDS: A hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems[J]. IEEE Access, 2019, 7: 89507-89521.

[4] UMER M A, JUNEJO K N, JILANI M T, *et al*. Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations[J]. International Journal of Critical Infrastructure Protection, 2022, 38: 100516.

[5] WANG W, HARROU F, BOUYEDDOU B, *et al*. A stacked deep learning approach to cyber-attacks detection in industrial systems: Application to power system and gas pipeline systems [J]. Cluster Computing, 2022: 1-18.

[6] 张靖雯. 基于深度学习的工控异常检测及攻击分类方法研究 [D]. 北京:北京工业大学, 2019.

[7] LIU Y, SUN Y, LIU C, *et al*. Industrial internet security situation assessment method based on self-attention mechanism[C]//2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC). Wuhan, China: IEEE, 2024: 148-151.

[8] ALZUBAIDI L, ZHANG J, HUMAIDI A J, *et al*. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions[J]. Journal of Big Data, 2021, 8: 1-74.

[9] GUPTA J, PATHAK S, KUMAR G. Deep learning (CNN) and transfer learning: A review[J]. Journal of Physics: Conference Series, 2022, 2273: 012029.

[10] AL-SELWI S M, HASSAN M F, ABDULKADIR S J, *et al*. RNN-LSTM: From applications to modeling techniques and beyond—Systematic review [J]. Journal of King Saud University-Computer and Information Sciences, 2024: 102068.

[11] WEN X, LI W. Time series prediction based on LSTM-attention-LSTM model[J]. IEEE Access, 2023, 11: 48322-48331.

[12] SEABE P L, MOUTSINGA C R B, PINDZA E. Forecasting cryptocurrency prices using LSTM, GRU, and bidirectional LSTM: A deep learning approach[J]. Fractal and Fractional, 2023, 7(2): 203.

[13] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[14] DAI Z, LIU H, LE Q V, *et al*. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in neural information processing systems, 2021, 34: 3965-3977.

[15] CHAUDHARI S, MITHAL V, POLATKAN G, *et al*. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[16] MORRIS T, GAO W. Industrial control system traffic data sets for intrusion detection research[C]//8th International Conference on Critical Infrastructure Protection (ICCIP 2014). Berlin, Germany: Springer, 2014: 65-78.

[17] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述 [J]. 计算机工程与应用, 2021, 57(14): 27-38.

[18] 李鹏威, 郑红, 单蓉胜. 基于多通道图神经网络和 CNN-BiLSTM 的漏洞检测方法 [J]. 华东理工大学学报 (自然科学版), 2025, 51(6): 835-842.

[19] POWERS D M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation [EB/OL]. (2020-10-11). https://doi.org/10.48550.arXiv.2010.16061.

[20] KUMAR A, CHOI B J. Benchmarking machine learning based detection of cyber attacks for critical infrastructure[C]//2022 International Conference on Information Networking (ICOIN). [s.l.]: IEEE, 2022: 24-29.

[21] CHANG C P, HSU W C, LIAO I E. Anomaly detection for industrial control systems using K-means and convolutional autoencoder[C]//2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). Split, Croatia: IEEE, 2019: 1-6.

[22] PEREZ R L, ADAMSKY F, SOUA R, *et al*. Machine learning for reliable network attack detection in SCADA systems[C]//2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). New York, USA: IEEE, 2018: 633-638.

[23] 邓志刚, 孙子文. 工业信息物理系统攻击检测增强模型 [J]. 信息与控制, 2021, 50(4): 410-418.

[24] SOKOLOV A N, ALABUGIN S K, PYATNITSKY I A. Traffic modeling by recurrent neural networks for intrusion detection in industrial control systems[C]//2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). [s.l.]: IEEE, 2019: 1-5.

[25] MOHAMMAD R M, SEYED M F. Development of intrusion detection in industrial control systems based on deep learning[J]. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2022, 46(3): 641-651.

[26] YANIKA K, PAKARAT M, PHET A. *et al*. An intrusion detection and identification system for internet of things networks using a hybrid ensemble deep learning framework[J]. IEEE Transactions on Sustainable Computing, 2023, 8(4): 596-613.

[27] LU K D, YANG Y W, ZENG G Q, *et al*. BPSO-AHDL-IDS: Binary particle swarm optimization-based automated hybrid deep learning model for intrusion detection of internet of things[C]//IEEE Transactions on Automation Science and Engineering. [s.l.]: IEEE, 2025: 15859-15877.

[28] XUE Q, ZHANG Z, FAN K, *et al*. xIIRS: Industrial internet intrusion response based on explainable deep learning[J]. Electronics, 2025, 14(5): 987.

[29] FOSTIROPOULOS I, ITTI L. ABLATOR: Robust horizontal-scaling of machine learning ablation experiments[J]. Proceedings of Machine Learning Research, 2023, 224: 1-15.

Multi-Class Detection of Cyber Attacks in Industrial Control Systems Based on Deep Learning

WANG Gengchen, JIANG Qingchao, YAN Xuefeng
(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: As a core component of national critical infrastructure, the security of Industrial Control Systems (ICS) is of paramount importance. With the widespread application of information technology, the efficiency of ICS operations has significantly improved, but new security risks have also emerged. In recent years, the frequent occurrence of cyber-physical attacks targeting ICS has made anomaly detection a key technology in safeguarding such systems. Traditional anomaly detection methods often reduce the problem to binary classification, which is insufficient for practical needs. To more precisely locate attack sources and facilitate rapid system recovery, a finer-grained classification of ICS anomalies is required. This paper proposes a novel deep learning-based model for ICS anomaly detection and attack classification. The model leverages the strengths of Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM) networks, and the Attention mechanism. CNN is used to extract spatial features of data packets, BiLSTM captures temporal dependencies between packets, and the Attention mechanism focuses on critical time-step information to achieve high-precision detection of ICS network attacks. Experimental results demonstrate that the proposed model outperforms existing industrial intrusion detection systems in terms of detection accuracy and performs well on imbalanced datasets, offering a new solution for ICS security protection.

Key words: industrial control system; anomaly detection; cyber attack; attack classification; deep learning

（责任编辑：李娟）