一种基于多级特征选择的工业互联网入侵检测方法

贺舒畅，邓在辉*

武汉纺织大学电子与电气工程学院，湖北武汉 430200

摘要：针对当前工业互联网入侵检测中因冗余特征及特征表征能力不足而导致的检测准确率低等问题，本文提出一种基于多级特征选择的工业互联网入侵检测方法。首先，利用皮尔逊相关系数（PCC）计算特征之间的线性相关程度，通过设定相关系数阈值剔除高度相关的冗余特征，完成特征子集的初步降维；随后，引入带自适应交叉、变异概率与精英保留策略的自适应精英遗传算法（AEGA），并基于兼顾分类性能与特征数量的适应度函数，对初步筛选后的特征子集进行二次筛选，得到最优特征集合。实验基于公开的 TON-IoT 网络数据集进行验证，原 38 维特征经过两级特征筛选后缩减至 26 维，特征缩减率达到 31.58%，两级筛选协同作用，有效减小了特征冗余。并且在相同机器学习分类算法条件下，与互信息和粒子群优化等特征选择方法相比，本文方法结合随机森林和 K 近邻分类器均取得更优检测性能。为进一步提高分类准确率，充分挖掘优选特征的表征信息，基于优化后的特征集合构建 CNN-GRU 融合模型，检测准确率达到了 95.61%，相较于单一 CNN 和 GRU 网络模型分别提高了 2.19%和 0.47%，验证了融合模型在入侵检测中的有效性。实验结果表明，本文提出的多级特征选择方法能够有效减少数据冗余、降低模型计算开销，为资源受限工业终端提供轻量化、高精度的入侵检测解决方案。

关键词：入侵检测；多级特征选择；遗传算法；深度学习

DOI: 10.57237/j.cst.2026.04.001

An Intrusion Detection Method Based on Multi-level Feature Selection for Industrial Internet

He Shuchang, Deng Zaihui*

College of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China

Abstract: To address the limited detection accuracy of industrial Internet intrusion detection systems caused by feature redundancy and inadequate feature representation, this study proposes a multi-stage feature selection-based intrusion detection method for the industrial Internet. First, the Pearson correlation coefficient is employed to quantify the linear correlations among features, and highly correlated redundant features are eliminated according to a predefined correlation threshold, thereby achieving preliminary dimensionality reduction. Subsequently, an adaptive elite genetic algorithm, incorporating adaptive crossover and mutation probabilities together with an elite preservation strategy, is introduced to further optimize the feature subset. A fitness function that jointly considers classification performance and the number of selected features is designed to identify the optimal feature subset. Experiments are conducted on the publicly available TON-IoT network dataset. The proposed two-stage feature selection strategy reduces the original 38 features to 26,

*通信作者：邓在辉，zaihuideng@163.com

收稿日期: 2026-07-26；接受日期: 2026-08-17；在线出版日期: 2026-11-02

achieving a feature reduction rate of 31.58% while effectively alleviating feature redundancy. Under identical classification settings, the proposed method demonstrates superior detection performance with both Random Forest and K-Nearest Neighbors classifiers compared with mutual information- and particle swarm optimization-based feature selection methods. Furthermore, to exploit the discriminative information contained in the selected features and further improve detection accuracy, a CNN-GRU hybrid model is developed based on the optimized feature subset. The proposed CNN-GRU model achieves a detection accuracy of 95.61%, outperforming the standalone CNN and GRU models by 2.19% and 0.47%, respectively. These results demonstrate the effectiveness of the proposed framework in reducing feature redundancy and improving intrusion detection performance, while reducing the dimensionality of the model input and associated computational overhead. The proposed approach provides a potential solution for efficient and accurate intrusion detection in resource-constrained industrial environments.

Keywords: Intrusion Detection; Multi-Level Feature Selection; Genetic Algorithm; Deep Learning

1 引言

近些年来，互联网技术快速发展，并广泛应用在各个行业和领域。在工业领域，互联网显著提升了工业生产效率与智能化水平，但与此同时，其开放互联互通的特性也使其面临日益严峻的网络安全威胁，各种网络入侵和网络攻击频发。一旦安全防护失效将引发生产线停滞、基础设施失控、企业核心数据泄漏等重大损失。因此，入侵检测技术[1]在保障工业互联网安全、抵御潜在攻击方面发挥着至关重要的作用。

在网络入侵检测模型的构建过程中，特征的有效性与安全相关性是决定模型性能的关键因素。当前工业网络安全数据集中普遍存在大量无关冗余特征，这些特征会对模型训练与性能优化带来一系列问题[2]。首先，冗余特征可能导致模型出现过拟合问题，致使模型仅能适配训练样本，泛化能力大幅下降。其次，高维冗余数据会显著提升模型训练复杂度，增加算力资源的消耗与训练耗时，给模型的实际开发与落地部署带来极大阻碍[3]。最后，冗余特征会干扰模型挖掘数据潜在有效规律的能力，无法精准捕捉网络攻击特征，最终大幅降低模型在真实场景中的检测预测精度。

因此，为了提高模型检测性能，有必要通过精细化的特征选择方法来剔除数据中的冗余特征与无效特征[4, 5]。特征筛选不仅可以缓解模型过拟合问题、简化模型结构，还能显著降低训练成本、提升模型收敛效率，进一步增强模型对未知网络流量的特征挖掘能力与泛化性能，从而实现更为精准、稳定的工业入侵检测效果。

Singh [6]提出了一种基于主成分分析（PCA）特征提取的多对数朴素贝叶斯算法，PCA 用于提取数据特征，然后将提取的特征使用基于多对数函数的朴素贝叶斯分类方法进行分类，以判别正常和异常数据。Ghasemi 等[7]提出一种基于支持向量机（SVM）和灰狼优化算法（GWO）的混合入侵检测方法，SVM 用于训练和区分正常和异常数据，GWO 用于来做特征选择，寻找 SVM 核函数以及调整最优参数，以提升分类效果。Benaddi 等[8]采用卡方特征选择方法，从 UNSW-NB15 数据集中筛选出 20 个最优相关特征以实现特征降维，并结合卷积神经网络（CNN）与双向长短期记忆网络（Bi-LSTM）构建混合入侵检测模型，有效提升了检测精度。Disha 等[9]采用了一种基于基尼指数的加权随机森林（RF）算法作为特征选择方法，该方法在 UNSW-NB15 和 TON_IoT 数据集上实验时分别将特征数量减少至 20 和 10，极大地降低了特征的冗余程度。Alshehri 等[10]采用互信息（MI）方法评估数据集中每个特征的重要性，选择熵值大于 0.1 的特征，以此来选择数据集中的有效特征以及剔除一些非必要的特征。

然而，他们的研究仅依赖单一的特征选择方法，而单一的特征选择方法可能无法产生最佳特征子集，这可能会影响模型的检测性能。通过结合其他特征可以解决单一特征选择方法的缺点，从而产生更好的结果。此外，现有特征选择技术中使用的大多数是单变量的，它们在评估每个特征时不考虑其他特征，并未考虑特征之间的相互作用，从而未能有效消除冗余特征。为了克服这一问题，可以使用多级特征选择方法。

Xu 等[11]将二进制灰狼优化算法（BGWO）与递归特征消除算法（RFE）结合做特征降维，筛选最优特征子集，采用参数调优的 XGBoost 构建入侵检测模型，多数据集实验验证该特征选择方法具备更优的检测性能。

Liu 等[12]提出一种融合信息增益与主成分分析的 IG-PCA 特征选择方法，基于半监督异常检测框架 SFSD，依托改进偏差网络与定制偏差损失函数优化异常评分，相比现有半监督方法准确率与检测率平均提升 1%~2%。Nandhini [13]结合 PCA 与改进哈里斯鹰优化算法（IHHO）实现特征筛选，通过 PCA 完成数据降维并依托 IHHO 算法优选关键特征，最后采用 SVM-KNN 两级分类模型对网络流量数据进行分类识别，有效区分正常流量与异常流量。Mishra 等[14]结合皮尔逊相关系数（PCC）和信息增益去除数据集中的冗余特征，在 BoT-IoT 数据集上选用基于 LightGBM 和随机森林的集成学习算法作为分类器进行实验，提升了检测的精度。王相月等[15]结合 PCC 和 RF 提出了一种多阶段特征选择方法，首先通过 PCC 剔除相关性高的特征，随后训练随机森林算法获得特征的重要程度值以此来过滤非重要特征。采用多级特征选择相对于单一特征选择更能体现出特征之间的关联性，提升模型训练的效率和检测的准确性。

本文主要包含三部分内容：首先对数据集进行预处理，包括数据清洗、数值化和归一化。工业互联网入侵数据集包含多种数据类型，包括非数值型和数值型。由于深度学习模型只能处理数值型数据输入，因此需要将非数值型数据转换为数值型数据。由于工业物联网入侵数据的每个特征具有不同的取值范围，这可能导致某些特征在模型训练过程中占据主导地位，而其他特征的影响则被削弱。所以需要进行归一化处理，消除入侵数据中的量纲影响，使得模型更加关注数据本身的内在关联结构和模式，而不被量纲所干扰。

然后是特征选择模块，利用皮尔逊相关系数和 AEGA 特征选择方法对预处理后的数据进行特征选择，得到一个最优特征子集。这个特征子集包含了最具代表性、最能区分正常数据和入侵数据的特征，使用这个特征子集来构建特征向量，得到最后输入模型的数据集。

最后是入侵检测模块，根据最优的特征子集得到入侵检测模型所需的数据集，并将其输入到训练好的 CNN-GRU 入侵检测模型中进行检测得到检测结果。将数据的真实标签与检测结果进行比较，通过计算精确率、准确率、召回率和 F1 值来评估模型的检测性能。

2 特征选择

2.1 特征初筛

皮尔逊相关系数（Pearson correlation coefficient，PCC）是衡量两个变量之间线性相关程度的重要指标，其取值范围为[-1,1]。若该系数绝对值越接近于 1，则表明两个变量间的线性关系就越强；而系数为 0 时，则表明这两个变量在线性关系上不存在相关性，即两个变量线性无关。特征 A 和特征 B 之间的皮尔逊相关系数可利用公式(1)计算:

$$Pcc = \frac{\sum_{i=1}^{n}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}} \tag{1}$$

其中，A_i和B_i分别表示特征A和特征B所对应的值，A和B分别表示特征A和特征B的平均值。

本实验基于皮尔逊相关系数的特征初筛来计算工业互联网入侵数据中每个特征之间的皮尔逊相关系数，通过删除每对强相关特征中的一个特征来选择最佳特征。

特征初筛具体流程为：首先设定皮尔逊相关系数阈值 u 的大小，其取值范围为[-1,1]。然后遍历工业互联网入侵数据集中所有可能的特征对组合，并计算所有特征对中两个特征之间的皮尔逊相关系数，并按照降序的顺序重新排列工业互联网入侵数据的特征。然后将计算所得工业互联网入侵数据集中的特征之间的皮尔逊相关系数与所设相关系数阈值 u 进行对比，对于皮尔逊相关系数大于阈值 u 的特征对，随机删除其中一个特征来降低数据维度，得到精简的特征子集。

2.2 特征优化

经过皮尔逊相关系数的特征初筛后，得到了一个精简且包含关键信息的特征子集集合。然后，为了进一步优化特征组合，将得到的精简特征子集利用自适应精英遗传算法（Adaptive Elite Genetic Algorithm，AEGA）进行特征优化，得到最优的特征子集。

2.2.1 编码和解码

利用 AEGA 对精简工业物联网入侵数据进行特征优化时，首先进行编码操作。本文采用了二进制编码方式，二进制编码方式简单直观且易于实现。AEGA 中的每个个体表示一个特征子集，每个基因位置代表一种特征。在二进制编码方式中，每个基因位置的取值只能为 0 或 1。0 和 1 具有特定的含义：0 表示未选取该基因位置所对应的特征，而 1 则表示该基因位置对应的特征被选中。例如，一个具有基因 1011000 的个体，表示在特征维度为 7 的入侵数据中，特征子集由第一个

特征、第三个特征以及第四个特征构成。

因此，一个可行的候选特征子集集合可以用公式(2)表示。

$$X=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D-1} & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D-1} & x_{2,D} \\ \vdots & \vdots & x_{m,d} & \vdots & \vdots \\ x_{M-1,1} & x_{M-1,2} & \cdots & x_{M-1,D-1} & x_{M-1,D} \\ x_{M,1} & x_{M,2} & \cdots & x_{M,D-1} & x_{M,D} \end{bmatrix} \tag{2}$$

其中X表示候选特征子集集合，M表示候选特征子集集合的大小，D表示精简工业互联网入侵数据的特征数，$x_{m,d}$表示第m个候选特征子集的d个特征。则第i个特征子集X_i可用公式(3)表示：

$$X_i=\{x_{i,1},x_{i,2},\ldots,x_{i,d},\ldots,x_{i,D}\} \tag{3}$$

其中，$i\in\{1,2,3,\ldots,M\}$, $x_{i,d}\in\{0,1\}$，$x_{i,d}=1$表示特征子集包括第d个特征，$x_{i,d}=0$表示特征子集不包括第d个特征。在进行解码时，仅需要关注那些基因位置取值为 1 的特征。

2.2.2 特征子集评估指标

为了评估所选特征子集的好坏，可以从特征子集分类准确率和特征子集所包含的特征数两方面进行评估。本文采用一个适应度函数作为特征子集的评估指标，该适应度函数的目标是最大化分类准确率和最小化特征数，适应度函数的计算方式如公式(4)所示：

$$fitness=\alpha\times Accuracy+(1-\alpha)\times\frac{s}{S} \tag{4}$$

其中，$Accuracy$表示所选特征子集的分类准确率，本文选择随机森林作为分类器。α表示分类准确率的权重系数，是一个常数。s表示所选特征子集中包含的特征数量，S则表示工业互联网入侵数据中的特征总数。

2.2.3 自适应策略

交叉操作主要促进不同特征子集之间的特征交流。本文的方法是两点交叉方式，这个过程不仅继承了父代特征子集的优良基因，同时也引入了新的特征组合，增加了特征子集集合的多样性。通过不断地进行两点交叉操作，能够逐步找到最优特征子集。

变异操作主要在候选特征子集中引入新的特征信息，从而生成更多具有多样性的候选特征子集。本文的方法是单点变异方式，通过这种方式，能够产生具有新特征组合的特征子集，这些新的特征子集可能具有更好的评估指标，从而有助于跳出局部最优解，向全局最优解逼近。

然而交叉变异率过大可能会影响特征选择效率，交叉变异概率过小可能会提前结束特征选择。为了避免这两种情况的发生，AEGA 采用自适应策略动态调整交叉概率和变异概率，计算公式如(5)和(6)所示：

$$P_c=\begin{cases} P_{c\min}+\dfrac{(f_{\max}-f_{im})\times(P_{c\max}-P_{c\min})}{f_{\max}-f_{avg}} & ,f_{im}>f_{avg} \\ P_{c\max} & ,f_{im}\le f_{avg} \end{cases} \tag{5}$$

$$P_m=\begin{cases} P_{m\min}+\dfrac{(f_{\max}-f)\times(P_{m\max}-P_{m\min})}{f_{\max}-f_{avg}} & ,f>f_{avg} \\ P_{m\max} & ,f\le f_{avg} \end{cases} \tag{6}$$

其中P_c表示交叉概率，f_{im}表示当前待进行交叉操作的特征子集中较高的评价指标值。P_m表示变异概率，f表示当前待进行变异操作的特征子集的评价指标值。$f_{\max}$表示特征子集集合中的最大评价指标值，f_{avg}表示特征子集集合的平均评价指标值。$P_{c\min}$和$P_{c\max}$分别表示最小交叉概率和最大交叉概率，$P_{m\min}$和$P_{m\max}$分别表示最小变异概率和最大变异概率。

2.2.4 精英个体保留策略

为确保特征筛选过程中，既不会丢失已经找到的最优特征子集，又能加速搜索过程并提高算法的收敛性和稳定性，在利用 AEGA 对工业互联网入侵数据进行特征筛选时，使用精英个体保留策略。

当 AEGA 进行到第k代时，将特征子集集合$X(k)$中评价指标最优的特征子集$X_i(k)$保留到数组中，然后将$X_i(k)$与第$k+1$代特征子集集合$X(k+1)$进行比较，将$X(k+1)$中评价指标最差的特征子集替换为$X_i(k)$。由于特征子集评价指标与分类器的分类准确率关系较大，所以精英个体保留策略保留的精英个体是评价最大的候选特征子集。

3 数据集准备与预处理

3.1 数据集介绍

TON-IoT 数据集是目前工业物联网入侵检测领域应用最广泛的数据集之一，由澳大利亚新南威尔士大学堪培拉物联网实验室与澳大利亚国防军学院网络靶场共同收集提出。该数据集是从整个工业物联网系统收集的各种数据源，包括来自 IoT 设备的遥测数据、Linux 与 Windows 操作系统记录以及工业物联网系统

的网络流量。本文使用的是 TON-IoT 数据集中的 Train_Test_datasets 文件。TON-IoT 数据集中包含 9 种攻击数据类型和 1 种正常数据类型，共有 461043 条数据记录，其中正常数据共有 300000 条，9 种攻击数据共有 161043 条。将 TON-IoT 数据集数据按 7：3 的比例划分为训练集和测试集，具体分布如表 1 所示。

表 1 TON-IoT 数据集样本分布

类型	训练集	测试集
backdoor	14057	5943
ddos	13990	6010
dos	13979	6021
injection	13956	6044
mitm	727	316
xss	14046	5964
password	14106	5894
ransomware	14028	5972
scanning	13982	6018
normal	210016	89984

3.2 数据预处理

由于 TON-IoT 数据集中存在一些不规范数据，需要对不规范数据进行清理，例如在 TON-IoT 数据集中，特征 dns_aa 存在非布尔型的值 T 和 F，需要将其更改为 True 和 False。通过对数据集的理解和观察删除不必要的特征，TON-IoT 数据集中的特征数变为 38。TON-IoT 数据集和数据集中均存在数值型数据和非数值型数据，而非数值型数据不能够作为模型的输入，因此需要将其转换为数值型数据。本实验预处理数值化方法采用的是标签编码方式。TON-IoT 数据集的数据标签数值化的结果如表 2 所示。

表 2 TON-IoT 数据集标签数值化结果

编码	数据集类别
0	backdoor
1	ddos
2	dos
3	injection
4	mitm
5	normal
6	password
7	ransomware
8	scanning
9	xss

数值化之后的数据集中不同特征之间的量纲不同，量纲差异过大会影响之后的实验结果，为了消除量纲影响，将数据集进行归一化处理，使所有特征处于同一数量级。归一化计算公式(7)所示：

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{7}$$

其中，x'表示归一化后的值，x表示原始特征值，x_{max}表示该特征值中的最小值，x_{min}表示该特征值中的最大

4 入侵检测模型

4.1 CNN 和 GRU 模型

在经过皮尔逊相关系数与精英遗传算法的特征选择后，得到了一个表征能力最优的特征子集，基于此特征子集得到的数据集作为模型的输入，本文构建了 CNN-GRU 工业互联网入侵检测模型，提取工业互联网入侵数据集的时空特征。

CNN 用于提取输入数据的局部空间特征，卷积层通过滑动窗口和卷积核的操作，对输入数据进行特征提取；池化层则通过下采样减少输出数据特征的维度。

GRU 用于捕捉时间序列中的长期依赖关系，其结构如图 1 所示。作为长短期记忆神经网络（Long-short term memory，LSTM）的一个简化版，与 LSTM 结构相似，拥有更新门和重置门，但 GRU 一个门控就能实现 LSTM 需要多个门控才能实现的遗忘和记忆操作，在保持 LSTM 性能优势的同时，降低其复杂性和计算成本。

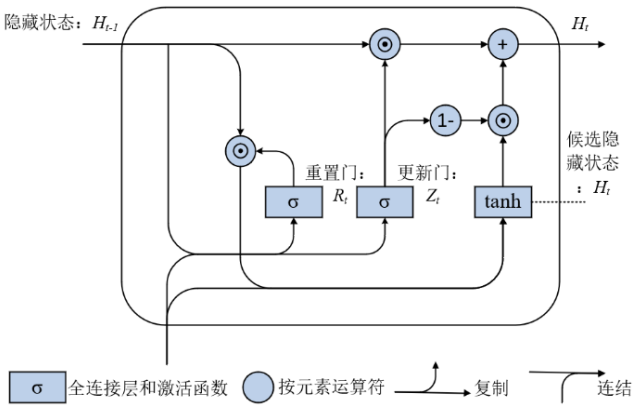


图 1 GRU 模型结构

4.2 CNN-GRU 模型

CNN-GRU 模型的具体结构如图 2 所示。经过特征选择后的所得到的数据作为 CNN-GRU 模型的输入，然后输入到 CNN 结构，通过两层卷积层-归一化层-最大池化层提取数据的空间特征。

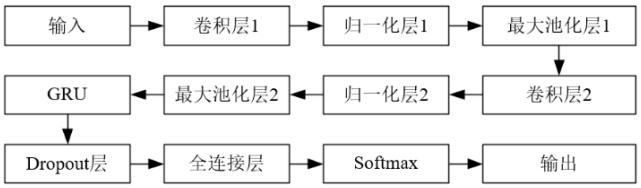


图 2 CNN-GRU 模型结构

CNN 得到的输出特征序列输入到 GRU 结构，并利用其独特的门控机制捕捉时间依赖性，从而进一步提取数据中的时间特征。

为了防止过拟合加入 Dropout 层，并通过全连接层的非线性变换增强特征的表达能力。最终，经过全连接层处理后的特征输入到 softmax 分类器，然后输出入侵检测结果。

5 实验与分析

5.1 实验参数

皮尔逊相关系数阈值 u 为 0.95，AEGA 参数设置为：最大迭代次数 T 为 50，特征子集集合大小 M 为 20，最小交叉概率$P_{c\min}$为 0.7，最大交叉概率 $P_{c\max}$ 为 0.9，最小变异概率 $P_{m\min}$为 0.01，最大变异概率$P_{m\min}$为为 0.1，特征子集评估指标中的权重α为 0.99。

CNN-GRU 模型的参数为：CNN 卷积核数量为 64，卷积核大小为 5，步长为 1，dropout 率为 0.001，GRU 层神经元数量 64，优化器为 Adam，激活函数选择 Relu，学习率为 0.001。

5.2 评估标准

为验证模型性能所使用的指标为入侵检测领域常用的评价标准，包括准确率、精确率、 召回率、F1 值。将预测为正例，实际也为正例的用 TP 表示、预测为负例，但是实际为正例的用 FP 表示、预测为正例，但是实际为负例的用 TN 表示、预测为负例，实际也为负例的用 FN 表示。准确率、精确率、 召回率、F1 值如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$F1 = \frac{2Recall \cdot Precision}{Recall+Precision} \tag{11}$$

5.3 特征选择结果

针对本文特征选择方法在缩减工业互联网入侵数据集冗余特征方面的效果，对比了特征筛选前后特征属性个数的变化，对比结果如图 3 所示。

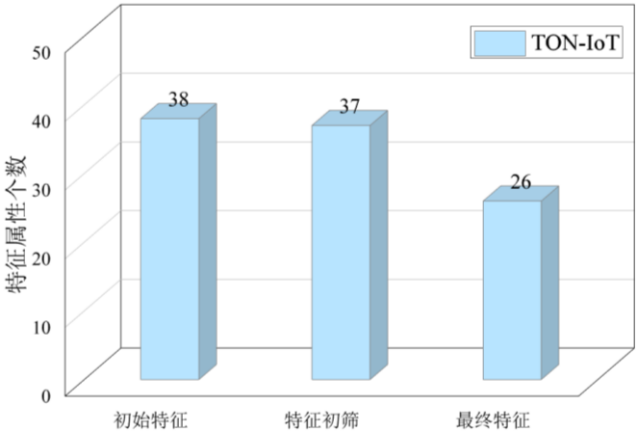


图 3 特征子集数目

从图中可以看出，对于 TON-IoT 数据集，最初特征属性个数经过皮尔逊相关系数第一次筛选后，从原始的 38 个减少到了 37 个。然后，对得到的精简特征子集继续进行 AEGA 特征筛选，最后得到了最优特征子集。为了更加直观地展示特征缩减的效果，计算了特征缩减率，即减少的特征数量占原始特征总数的百分比。在 TON-IoT 数据集上，对于皮尔逊相关系数特征初筛，特征缩减率为 2.63%，对于 AEGA，特征缩减率达到了 29.73%，对于整个特征选择过程，特征缩减率达到了 31.58%，有助于降低入侵检测模型的复杂度。

5.4 特征选择效果

本文采用对比实验来验证所用特征筛选方法的效果，选择两个特征选择方法作为对比对象，分别是互信息（MI）、粒子群算法（PSO）。同时，为了控制实验中的其他变量因素，统一采用了两种广泛应用的模型-随机森林（RF）、K 邻近算法（KNN）来进行检测，不同的特征选择方法检测结果如下表 3 所示：

表 3 TON-IoT 数据集检测结果

方法	准确率	精确率	召回率	F1 值
MI-RF	0.9214	0.9346	0.9143	0.9244
PSO-RF	0.9263	0.9428	0.9186	0.9306
本文方法	0.9354	0.9465	0.9354	0.9407
MI-KNN	0.9316	0.9196	0.8680	0.8929
PSO-KNN	0.9321	0.9266	0.8956	0.9106
本文方法	0.9422	0.9468	0.9183	0.9322

在 TON-IoT 数据集上，从表 3 中可以看出，本文所用特征选择方法在 RF、KNN 两个分类模型的准确率、精确率、召回率和 F1 值上均表现更优。在 RF 分类器上，本文方法准确率为 0.9354、精确率为 0.9465、召回率为 0.9354、F1 值为 0.9407。比 PSO 的准确率高了 0.91%，比 MI 的准确率高了 1.40%。召回率分别比 PSO、MI 高了 1.68%、2.11%。在 KNN 分类器上，本文方法的准确率为 0.9422、精确率为 0.9468、召回率为 0.9183、F1 值为 0.9322。召回率分别比 MI、PSO 高了 5.03%、2.27%。这进一步验证了本文特征选择方法在工业互联网数据集上的泛化能力和稳定性。

5.5 入侵检测模型效果

为验证本文所提特征选择方法在工业互联网入侵数据检测上的优越性，将本文提出的 CNN-GRU 入侵检测方法与其他入侵检测方法，如 CNN、GRU，分别在 TON-IoT 数据集上进行对比，结果如表 4 所示：

表 4 TON-IoT 不同模型检测结果

检测方法	准确率	精确率	召回率	F1 值
CNN	0.9342	0.9509	0.9346	0.9427
GRU	0.9514	0.9547	0.9514	0.9531
CNN-GRU	0.9561	0.9572	0.9561	0.9567

在数据集 TON-IoT 上，从表 4 结果可以看出，CNN-GRU 混合模型在四项评估指标上均优于单独使用的 CNN 与 GRU 模型。CNN-GRU 模型取得准确率 95.61%、精确率 95.72%、召回率 95.61%、F1 值 95.67%。其中准确率相较 CNN 提升 2.19%，相较 GRU 提升 0.47%。这表明 CNN-GRU 融合网络结构在互联网流量检测场景下性能表现更优，模型具备更强的特征提取与分类识别能力。

6 结论

针对工业互联网环境下入侵检测存在特征冗余、有效特征提取不足，导致精度低的问题，提出一种特征筛选策略与时序空间融合模型检测框架。先利用皮尔逊相关系数剔除线性冗余特征，降低特征维度与数据冗余程度；再引入自适应精英遗传算法完成特征寻优，在进行特征压缩的同时，最大化子集分类准确率，降低模型运算开销。在模型层面，结合 CNN 局部特征提取优势与 GRU 时序依赖捕捉能力搭建 CNN-GRU 融合网络，弥补单一深度学习模型仅能挖掘单一维度信息的缺陷，充分挖掘工业流量数据中的空间分布与时序变化规律，相较于传统机器学习模型与单一网络模型具备更强的泛化能力，可为资源受限的工业终端入侵检测任务提供轻量化、高精度的可行思路。

本研究在标准公开的流量数据集上完成验证，模型参数优化以及针对新型未知攻击的泛化能力仍有提升空间，后续研究将面向真实工业生产环境，采集多设备、多协议混合现场流量，基于实际业务场景下的流量，进行实时性流量分析。

参考文献

[1] Nuaimi M, Fourati L C, Hamed B B. Intelligent approaches toward intrusion detection systems for Industrial Internet of Things: A systematic comprehensive review [J]. Journal of Network and Computer Applications, 2023, 215: 103637. https://doi.org/10.1016/j.jnca.2023.103637

[2] Venkatesan S. Design an intrusion detection system based on feature selection using ML algorithms [J]. Mathematical Statistician and Engineering Applications, 2023, 72(1): 702-710. https://philstat.org/index.php/MSEA/article/view/2000

[3] Thakkar A, Lohiya R. A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions [J]. Artificial Intelligence Review, 2022, 55(1): 453-563. https://doi.org/10.1007/s10462-021-10037-9

[4] Turukmane A V, Devendiran R. M-multisvm: An efficient feature selection assisted network intrusion detection system using machine learning [J]. Computers & Security, 2024, 137: 103587. https://doi.org/10.1016/j.cose.2023.103587

[5] Barhoush M, Abed-alguni B H, Al-qudah N E A. Improved discrete salp swarm algorithm using exploration and exploitation techniques for feature selection in intrusion detection systems [J]. The Journal of Supercomputing, 2023, 79(18): 21265-21309. https://doi.org/10.1007/s11227-023-05444-4

[6] Singh S. Poly logarithmic naive Bayes intrusion detection system using linear stable PCA feature extraction [J]. Wireless Personal Communications, 2022, 125(4): 3117-3132. https://doi.org/10.1007/s11277-022-09701-2

[7] Ghasemi H, Babaie S. A new intrusion detection system based on SVM–GWO algorithms for Internet of Things [J]. Wireless Networks, 2024, 30: 2173-2185. https://doi.org/10.1007/s11276-023-03637-6

[8] Benaddi H, Jouhari M, Elharrouss O. A lightweight hybrid approach for intrusion detection systems using a chi-square feature selection approach in IoT [J]. Internet of Things, 2025, 32 101624-101624. https://doi.org/10.1016/j.iot.2025.101624

[9] Disha R A, Waheed S. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique [J]. Cybersecurity, 2022, 5(1): 1 https://doi.org/10.1186/s42400-021-00103-8

[10] Alshehri M S, Saidani O, Alrayes F S, et al. A Self-Attention-Based Deep Convolutional Neural Networks for IIoT Networks Intrusion Detection [J]. IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3380816

[11] Xu B, Sun L, Mao X, et al. IoT intrusion detection system based on machine learning [J]. Electronics, 2023, 12(20): 4289. https://doi.org/10.3390/electronics12204289

[12] Liu Y, Deng W, Liu Z，et al. Semi-supervised attack detection in industrial control systems with deviation networks and feature selection [J]. The Journal of Supercomputing, 2024, 80(10): 14600-14621. https://doi.org/10.1007/s11227-024-06018-8

[13] Nandhini U, SVN K S. An improved Harris Hawks optimizer based feature selection technique with effective two-staged classifier for network intrusion detection system [J]. Peer-to-Peer Networking and Applications, 2024, 17 (5): 2944-2978. https://doi.org/10.1007/s12083-024-01727-6

[14] Mishra A K, Paliwal S. Mitigating cyber threats through integration of feature selection and stacking ensemble learning: the LGBM and random forest intrusion detection perspective [J]. Cluster Computing, 2023, 26(4): 2339-2350. https://doi.org/10.1007/s10586-022-03735-8

[15] 王相月，赵利辉. 基于多阶段特征选择和 CNN-GRU 的网络入侵检测模型 [J]. 中北大学学报 (自然科学版), 2024, 45(01): 6673. https://doi.org/10.3969/j.issn.16733193.2024.01.009