融合多尺度 CNN 与 Transformer 的恶意软件行为检测方法*

刘　帅[1,2]，王小英[1,2]，戚盼盼[1,2]，崔方方[1,2]，谷瑞泽[1,2]
（1. 应急管理大学　计算机科学与工程学院，河北　廊坊　065201；
2. 廊坊市网络应急保障与网络安全重点实验室，河北　廊坊　065201）

摘　要：针对恶意软件行为轨迹隐蔽且长序列依赖难以建模的严重威胁，提出一种融合多尺度卷积神经网络与 Transformer 架构的恶意软件检测方法，此方法首先借助 Speakeasy 仿真日志去噪及复合事件标记化技术，将冗余日志转化为标准化语义序列，接着运用多层次卷积神经网络结构来提取局部攻击行为特征，在此基础上，将提取的局部攻击行为特征输入 Transformer 编码器，利用多头自注意力机制建模全局时序依赖关系。实验结果表明，该混合模型在 Speakeasy 数据集上的准确率和 F1-Score 分别达到 92.29% 和 92.48%。该方法显著降低了序列检测中的误报率，为复杂网络环境下的恶意软件检测提供了新的技术途径。

关键词：恶意软件检测；卷积神经网络；Transformer；多尺度特征提取；动态行为分析
中图分类号：TP309.5　　**文献标志码**：A　　**DOI**：10.19358/j.issn.2097-1788.2026.04.006

中文引用格式：刘帅，王小英，戚盼盼，等. 融合多尺度 CNN 与 Transformer 的恶意软件行为检测方法[J]. 网络安全与数据治理，2026，45(4)：45-50.
英文引用格式：Liu Shuai，Wang Xiaoying，Qi Panpan，et al. A malware behavior detection method based on the fusion of multi-scale CNN and Transformer[J]. Cyber Security and Data Governance，2026，45(4)：45-50.

A malware behavior detection method based on the fusion of multi-scale CNN and Transformer

Liu Shuai[1,2]，Wang Xiaoying[1,2]，Qi Panpan[1,2]，Cui Fangfang[1,2]，Gu Ruize[1,2]
（1. College of Computer Science and Engineering，University of Emergency Management，Langfang 065201，China；
2. Langfang Key Laboratory of Network Emergency Support and Cybersecurity，Langfang 065201，China）

Abstract：To address the severe threats posed by stealthy malware behavioral trajectories and the difficulty in modeling long-sequence dependencies，this paper proposes a detection method that fuses multi-scale Convolutional Neural Networks（CNN）with the Transformer architecture. First，the approach utilizes Speakeasy simulation logs denoising and composite event tokenization techniques to convert redundant logs into standardized semantic sequences. Next，it employs a multi-layer CNN structure to extract local attack behavior features. Subsequently，these extracted features are fed into a Transformer encoder to model global temporal dependencies via a multi-head self-attention mechanism. The experimental results show that the hybrid model has achieved an accuracy of 92.29% and an F1-Score of 92.48% on the Speakeasy dataset. This approach significantly reduces the false positive rate in sequence detection，providing a new technical pathway for malware detection in complex network environments.

Key words：malware detection；Convolutional Neural Network（CNN）；Transformer；multi-scale feature extraction；dynamic behavior analysis

0　引言

随着互联网技术快速发展，恶意软件数量不断增多，其攻击方式也变得更加隐蔽，这让基于动态行为分析[1]的技术成为检测未知威胁的关键手段。面对维度高且规模大的动态行为日志数据，实现恶意特征的高效提取与精准识别已经成为当前网络安全领域亟需突破的关键技术挑战。

*基金项目：中央高校基本科研业务费研究生科技创新基金（ZY20260317）；立德树人视域下 AI 赋能网络安全“赛-教-创-研-服”育人路径探索与实践（2026GJJG487）

目前学术界大多借助深度学习算法来达成此类任务的自动化处理，然而单一网络架构在处理高维长序列日志时仍存在缺陷，卷积神经网络（CNN）可以有效地借助滑动窗口机制来提取局部行为特征，但受限于卷积核感受野，很难有效地对长程语义关联进行建模，而基于自注意力机制的 Transformer 以及采用门控结构的 LSTM[2] 模型虽然在时序依赖关系建模方面有着出色表现，但是在面对高维长序列行为数据时，还是会面临计算并行性不足或者局部特征表征能力欠缺等技术难题。此外，仿真日志数据里普遍存在的随机噪声进一步降低了模型的鲁棒性。

针对上述问题，本文提出一种将多尺度 CNN 和 Transformer[3] 结合起来的恶意软件检测方法，该方法先采用基于核心行为的序列去噪以及复合事件标记化策略有效过滤冗余噪声，接着构建混合网络结构，利用多尺度 CNN[4] 捕捉局部攻击特征，并且借助 Transformer 机制挖掘全局长程依赖关系，最终实现对恶意代码的高效识别。在 Speakeasy 数据集上的实验验证了此方法在长序列建模问题方面的出色表现：检测准确率提升至 92. 29%，F1-Score 提升至 92. 48%，维持了较高的检测精度，又降低了漏报数量，为恶意软件检测[5] 提供了一种高效且可靠的技术解决方案。

1　数据获取与预处理

1. 1　数据集分布

本文选取公开的 Speakeasy 数据集[6] 作为实验依据，此数据集包含由 Windows 内核仿真器产生的 92 703个 PE 文件动态行为日志。为检测模型在面对恶意软件概念漂移现象时的鲁棒程度，本文按照时间顺序对数据进行了严谨划分，训练集选取 2022 年 1 月期间的数据，测试集则来源于同年 4 月。数据覆盖范畴包含良性程序以及后门、勒索软件、木马等七大常见恶意家族类型。各类别的具体样本数量分布如表 1 所示。

表 1　Speckeasy 数据集信息

序号	类型	含义	样本数	占比/%
1	Benign	良性软件	32 673	35. 24
2	Backdoor	后门程序	13 002	14. 03
3	Ransomware	勒索软件	11 766	12. 69
4	Trojan	特洛伊木马	9 818	10. 59
5	Coinminer	挖矿木马	8 575	9. 25
6	Dropper	投放器	8 495	9. 16
7	Keylogger	键盘记录器	5 419	5. 85
8	RAT	远程控制工具	2 955	3. 19

1. 2　特征提取与预处理

本文运用的原始数据集为非结构化的 JSON[7] 动态日志，为提取有效信息同时消除仿真过程中所产生的噪声干扰，本文设计了如下预处理流程。

1. 2. 1　核心行为特征提取

本文通过引入容错解析机制，从原始报告中提取出四类关键特征：API 调用（反映底层交互逻辑）、文件操作（涵盖读写与删除事件）、注册表行为（记录键值修改与持久化操作）以及网络行为（提取通信流量特征）[8]。这些特征共同构成了识别恶意意图的核心数据基础。

1. 2. 2　日志序列去噪

原始的 JSON 日志通常会包含大量与核心恶意行为逻辑无关的背景噪声。这些噪声主要来源于操作系统底层的常规调度、动态变化的运行参数（比如随机内存地址、动态时间戳等）以及冗余的特殊符号。若直接把这些未经清洗的长序列输入神经网络，极易导致特征空间维度爆炸，并且严重干扰模型对关键攻击序列的长程依赖建模。

为此，本文设计了针对性的序列去噪机制。首先，针对非结构化日志中存在的大量非规范化文本，系统进行了统一的符号过滤与格式归一化操作。例如：剔除破坏词汇语义连续性的无关特殊符号；将所有字符统一转换为小写格式，以消除大小写差异带来的语义割裂。通过上述去噪处理，有效滤除了动态干扰信息，为后续构建标准化的语义单元奠定了基础。

1. 2. 3　复合事件标记化

为了将纯文本格式的日志转化为神经网络可处理的数值输入，本文设计了如图 1 所示的复合标记化流程。

"api_name": "KERNEL32.GetSystemTimeAsFileTime"

(a) 去噪处理后

↓

api_kernel32 getsystemtimeasfiletime: 54

(b) 复合标记处理后

图 1　原始日志去噪与复合标记化处理前后对比

首先，对于原始 JSON 报告中出现的 API 调用（如 KERNEL32. GetSystemTimeAsFileTime），系统去除无关的特殊符号（如“.”）并且全部统一转为小写，然后将其重新组合成“行为类型 - 库名 - 函数名”结构的语义单元（如 api_kernel32 getsystemtimeasfiletime），以此来保留 API 调用的语义完整性[9]。随后，为了统

一特征维度，本文构建了容量为 $V=50\ 000$ 的动态词表，将低频 Token 映射为［UNK］标识。其次，将清洗后的序列映射为整数索引序列，设定最大长度 $L=2\ 048$；超长序列截取前段，不足序列末尾填充［PAD］。最终结合二分类标签生成模型输入矩阵。这种处理方式不仅大幅压缩了特征空间的维度，还解决了原始日志格式不统一的问题，为嵌入层处理提供了标准化输入。

2 基于 CNN-Transformer 的检测模型

考虑到恶意代码行为序列存在局部组合模式与全局时序依赖并存的特性，本文设计了一种串行混合模型，主要依据多尺度局部特征提取、全局长程依赖建模以及特征融合与分类这三个核心要点来开展系统构建工作。检测模型具体架构如图 2 所示。

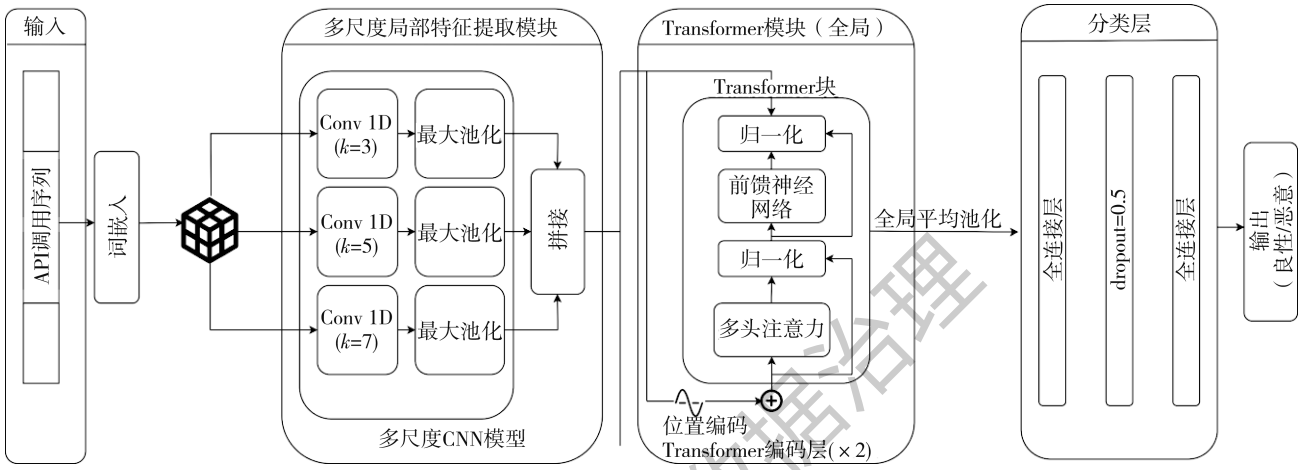

图 2 检测模型架构图

该架构的核心是使用并行的一维卷积层（$k\in\{3,5,7\}$）多尺度捕捉局部攻击模式，然后将融合的特征输入到 Transformer 编码器，利用其多头自注意力结构弥补传统 CNN 在长序列建模时的不足，最后经全连接层估计恶意概率。

2.1 多尺度局部特征提取

输入序列首先经嵌入层映射为 $d_{\text{model}}=128$ 维的稠密向量。为捕捉不同时间尺度下的恶意行为（如短促的文件操作或较长的网络握手），本文设计了一种并行的一维卷积结构。该结构采用三组不同尺寸的卷积核 $k\in\{3,5,7\}$，分别从输入序列里提取不同范围的局部特征。

在原理上，采用这种尺寸的组合是为了适配恶意软件的动态行为序列在不同时间尺度上的特征：小尺寸的卷积核（$k\in\{3\}$）能够提取以微观、紧凑的原子行为操作序列为基础的内容（例如连续文件的读写）；中等尺寸（$k\in\{5\}$）能够提取具有一定逻辑的中等长度的行为序列（例如标准的进程注入行为）；大尺寸（$k\in\{7\}$）能够覆盖更广的上下文，可以提取出宏观行为（例如完整的网络握手过程或完整复杂的注册表主动持久化）。因此，通过多尺度的并行提取，该模型能更全面地表征恶意行为在时域上的演化特性。

各卷积通道输出经 GELU 激活函数处理后，为了综合不同尺度的局部特征，模型将三组并行卷积分支的输出在特征维度上执行拼接，具体计算逻辑如下：由于输入层映射的嵌入维度 $d_{\text{model}}=128$，且三个并行分支的 1D 卷积层均保持输出通道数与输入维度一致（即 $d_{\text{out}}=128$），经特征融合后，特征向量在深度维度上完成线性堆叠。因此，最终输出的融合特征维度 $D_{\text{total}}=128\times3=384$。这种维度扩展策略旨在不损失各尺度信息的前提下，为后续 Transformer 编码器提供更丰富的多粒度语义表示。

2.2 全局长程依赖建模

为捕捉序列起始点与终点之间存在的复杂逻辑联系，将卷积层产生的输出结果输入到 Transformer 编码器模块中。为了适配 Transformer 的输入标准，首先通过线性映射将 CNN 输出的高维特征投影为固定的序列特征向量。

由于自注意力机制本身属于排列等变性结构，不具备处理序列位置信息的能力，若直接输入特征，会丢失恶意 API 调用之间至关重要的先后因果逻辑。因此，在进入注意力层之前，引入了正余弦位置编码来显式地注入时序信息，其计算公式如下：

$$\text{PE}_{(\text{pos},2i)}=\sin(\text{pos}/10\ 000^{2i/d_{\text{model}}}) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos(pos/10\ 000^{2i/d_{model}}) \quad (2)$$

其中，pos 表示当前 Token 在行为序列中的绝对位置，i 表示特征维度的索引。

叠加位置信息后的序列特征随后进入由 2 层堆叠构成的 Transformer Encoder。该模块的核心为多头自注意力机制。对于输入序列，首先通过不同的权重矩阵将其线性映射为查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$。标准的缩放点积注意力计算过程如式（3）所示：

$$\text{Attention}(\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (3)$$

其中，d_k 为键向量的维度。设定注意力头数 $h=8$，使得模型能够同时从 8 个不同的特征子空间中[10]，并行挖掘 API 调用序列在不同上下文环境下的隐含关联规律（例如独立捕捉文件释放行为与后续网络回连行为之间的依赖关系）。此架构有效突破了传统 CNN 感受野受限的瓶颈，实现了对全局长序列依赖的深度建模。

2.3 特征聚合与分类

在 Transformer 编码器完成全局语义建模之后，模型需要将序列级特征进一步转换成一种可用于分类的固定长度表示。由于行为序列的长度可能不同，直接基于全部时间步的特征进行分类不仅计算成本较高且容易引入冗余信息。因此采用全局平均池化来对各 Transformer 输出的特征聚合[11]。

具体而言，全局平均池化是指沿时间维度对所有 Token 的隐藏表示求均值，将原始的二维特征矩阵压缩为固定长度的一维向量，该向量可以认为是整个行为序列在语义空间的全局表示，它反映了恶意软件在整个运行过程中呈现出的多种行为模式[12]。相比于简单地选择序列末尾特征或最大池化方法，平均池化能更稳定地保留整体的行为分布特征，同时减少单一异常行为对最终判决结果的干扰，从而提升模型的鲁棒性。

得到序列级特征向量后，模型将其输入到由两层全连接网络构成的分类模块中。第一层全连接层负责对高维特征进行非线性映射，并通过 GELU 激活函数增强模型对复杂特征边界的表达能力[13]；随后引入 Dropout 机制随机丢弃部分神经元，以减少模型在训练过程中的过拟合风险并提升其泛化能力。第二层全连接层则进一步压缩特征维度，并通过 Sigmoid 函数输出最终的恶意概率值。

3 实验结果与分析

3.1 评估指标

对于分类问题，本文使用 Accuracy（准确率）、Precision（精确率）、Recall（召回率）和 F1-Score（F1 得分）来评价模型分类性能。计算公式如式（4）~式（7）所示：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

式中：TP（真正例）指将恶意软件正确识别为恶意样本的数量；TN（真负例）指将良性软件正确识别为良性样本的数量；FP（假正例）指将良性软件错误识别为恶意样本的数量（即误报）；FN（假负例）指将恶意软件错误识别为良性样本的数量（即漏报）。

3.2 消融实验分析

为探究 CNN 局部特征提取与 Transformer 全局建模的具体贡献，设计了如表 2 所示的消融实验。从表中可以清晰地看出单一的 CNN 或 Transformer 模型均无法对长序列日志进行有效处理（准确率仅在 85% ~88% 左右）。本文采用 CNN-Transformer 混合模型的效果在准确率和 F1-score 上均有较大幅度提高，尤其是准确率提高到了 92.29%。这一性能提升表明：多尺度 CNN 在局部特征提取上的优势与 Transformer 在全局长程依赖建模上的优势形成了互补效应，二者的级联有效突破了单一架构的表征极限。

表 2 不同网络架构的消融实验结果对比

模型架构	准确率/%	F1-Score/%
CNN	85.59	85.26
Transformer	88.36	88.99
CNN + Transformer	92.29	92.48

此外，为验证卷积核尺寸组合的有效性，测试了不同卷积核尺寸组合对模型性能的影响，结果如表 3 所示。

表 3 卷积核尺寸组合性能对比

卷积核组合	准确率/%	F1-Score/%
$k \in \{2, 4, 6\}$	90.15	90.42
$k \in \{3, 5, 7\}$（本文）	92.29	92.48
$k \in \{5, 7, 9\}$	91.03	91.26

上述实验的结果显示，卷积核尺寸 $k \in \{3, 5, 7\}$ 时模型检测结果准确率、F1-Score 最高。这表明该尺

寸可以更好、更直观地表达 Speakeasy 日志中绝大部分恶意行为发生的时间尺度相关特征。卷积核过小，容易丢失信息；而卷积核过大则可能提取一些无意义的信息。

综上，CNN 模型虽然计算高效，但缺乏长程依赖捕捉能力；Transformer 模型虽对此有所改善，但仍低于融合模型。此外，通过对不同尺寸卷积核组合的实验发现，选取适中的卷积核尺寸（$k \in \{3, 5, 7\}$）能够更平衡地兼顾恶意行为的时间局部相关性。本文提出的串行融合架构实现了“1 +1 >2”的效果，充分证明了同时兼顾局部细节与全局语境是提升检测性能的关键。

3.3 对比实验分析

为验证方法的优越性，将本文模型与 KNN、决策树（Decision Tree）、随机森林（Random Forest）及朴素贝叶斯（Naive Bayes）四种主流基准算法进行对比，如图 3 所示。相比其他算法本文模型在四项指标上均实现显著提升，这表明该模型不仅准确率远超基准算法，且对恶意代码的识别精准度（精确率）与覆盖完整性（召回率）更均衡，综合性能（F1 分数）优势突出。

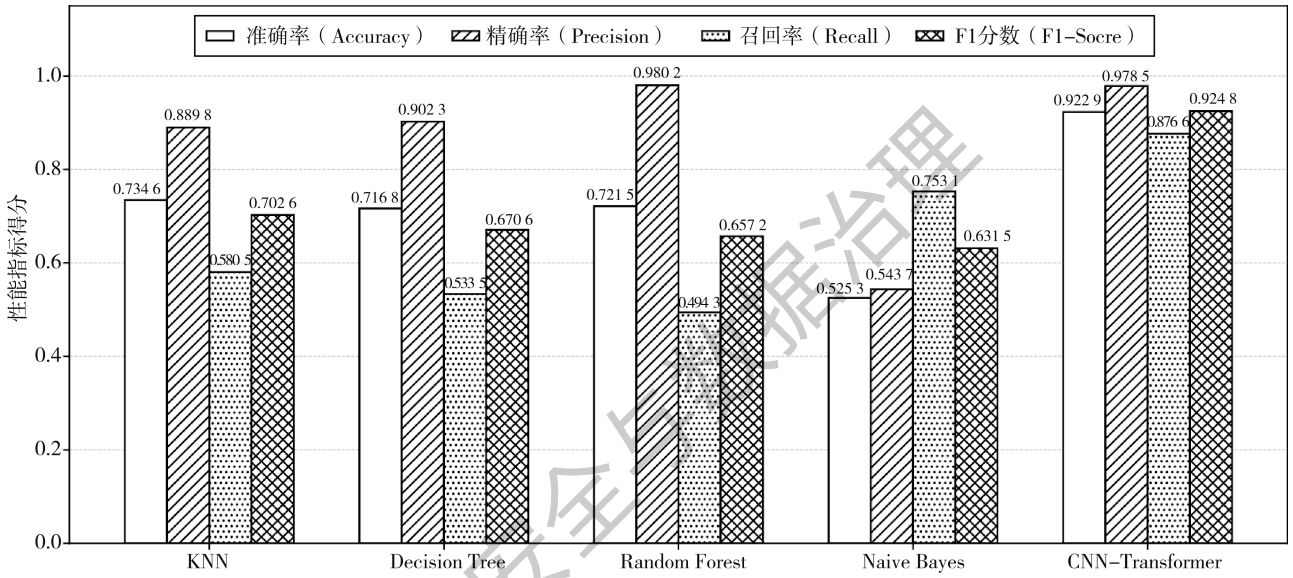

图 3　检测模型性能对比图

传统算法性能不佳的深层原因分析如下：

（1）特征工程的局限性。传统算法如随机森林、决策树等只是对特征频率、特征统计分布的捕获，对本文所描述的高维长程的“行为”序列是难以捕获的，其召回率较低是因为这类算法不能发掘隐藏在 API 调用“行为”序列中的时序相关信息，故而对复杂灵活的恶意变体漏报严重。

（2）长程依赖建模能力的缺失。恶意行为通常跨度较大。KNN 是一种距离度量方法，不反映行为的前后逻辑；而朴素贝叶斯法（Naive Bayes）受限于特征独立性假设，忽略了 API 之间的联系，导致其准确率处于所有算法中最低水平。

（3）局部与全局特征的失衡。传统模型通常只能捕捉全局统计特征（如 API 出现次数），丢失了局部行为片段（如特定堆栈操作）的语义。

相比之下，本文模型各项指标均表现最优，F1-Score 显著高于其他基准模型，证明了“多尺度局部特征 + 全局时序建模”架构在处理复杂恶意行为序列时的鲁棒性。

4　结论

针对恶意软件动态检测面临的长序列建模与噪声干扰难题，本文提出了一种融合多尺度 CNN 与 Transformer 的混合检测模型。该方法通过复合事件标记化与串行融合架构，有效兼顾了局部行为特征提取与全局时序依赖挖掘。实验表明，模型在 Speakeasy 数据集上的准确率与 F1-Score 分别达到 92.29% 和 92.48%，相较于传统机器学习方法与单一深度学习模型展现出显著性能优势，证明了局部与全局特征级联融合在应对恶意软件时的有效性。未来的工作将致力于解决 Transformer 架构计算开销大的问题，通过模型剪枝与量化技术探索其在资源受限环境下的部署潜力，并进一步增强模型应对对抗样本攻击的鲁棒性。

参考文献

[1] 陈冲，朱晓旭，万林葳，等. 融合深度学习技术的多模态安全管理应用[J]. 吉林大学学报（信息科学版），2025，43（6）：1430－1440.

[2] 杨伟青，周游，张婉澂，等. 基于 LLM 与 Transformer 机制的网络安全防御技术与应用研究[J]. 铁路计算机应用，2025，34（10）：37－41.

[3] 李沼洁，郭家毅，杨英东，等. 融合 Transformer 和 CNN 的伪装目标检测[J]. 福建理工大学学报，2025，23（6）：557－563.

[4] 李滨. 基于 Transformer 和卷积神经网络的图像去模糊算法研究［D］. 天津：天津工业大学，2023.

[5] 孙润康，彭国军，李晶雯，等. 基于行为的 Android 恶意软件判定方法及其有效性[J]. 计算机应用，2016，36（4）：973－978.

[6] TRIZNA D，DEMETRIO L，BIGGIO B，et al. Nebula：self-attention for dynamic malware analysis[J]. IEEE Transactions on Information Forensics and Security，2024，19：6155－6167.

[7] HERRERA-SILVA J A，HERNÁNDEZ-Álvarez M. Dynamic feature dataset for ransomware detection using machine learning algorithms[J]. Sensors，2023，23（3）：1053.

[8] SACHDEV H，CHEN L，REBMAN C. A new framework for securing，extracting and analyzing big forensic data[J]. Journal of Digital Forensics，Security and Law，2018，13（2）：6.

[9] LI X，JIANG H，KAMEI Y，et al. Bridging semantic gaps between natural languages and APIs with word embedding[J]. IEEE Transactions on Software Engineering，2018，46（10）：1081－1097.

[10] FAN X，GONG M，XIE Y，et al. Structured self-attention architecture for graph-level representation learning[J]. Pattern Recognition，2020，100：107084.

[11] 顾兆军，王亚飞，刘春波，等. 基于时序卷积网络的轻量级日志异常检测[J]. 计算机工程与设计，2025，46（8）：2272－2279.

[12] 李志新，王爽. 基于行为分析的恶意软件动态检测技术研究[J]. 科技与创新，2026（4）：155－157.

[13] TEREFE G，ASEFA H S，ASSABIE Y. Obfuscated malware detection using a hybrid of CNN and GRU models[J]. Journal of Computer Virology and Hacking Techniques，2026，22（1）.

（收稿日期：2026－02－24）

作者简介：

刘帅（2001－），男，硕士研究生，主要研究方向：灾害信息处理技术。

王小英（1979－），女，硕士，教授，主要研究方向：无线通信、WSN 技术、网络安全。

戚盼盼（2001－），女，硕士研究生，主要研究方向：灾害信息处理技术。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部