基于数据增强和多解释方法融合的入侵检测方法

熊炫睿[1]，郭星佑[1]，宁兆龙[1]，张玉树[1]，周力[2]
（1. 重庆邮电大学通信与信息工程学院，重庆 400065；2. 国防科技大学电子科学学院，湖南 长沙 410073）

摘　要：为了解决入侵检测系统中现有解释方法给出不一致结果、模型决策缺乏可信度的问题，提出并设计了多解释方法融合技术。该技术通过一致性、聚焦性和稳定性指标，融合沙普利加性解释（SHAP）、局部可解释模型（LIME）和置换特征重要性（PFI）的优势，建立客观权重计算机制，生成更可靠的特征重要性解释结果。针对数据不平衡导致解释稳定性差及少数类检测性能低的问题，采用数据平衡技术提供稳定数据基础。实验结果表明，所提方法显著增强了模型解释的可靠性和一致性，并进一步提升了入侵检测性能。

关键词：入侵检测；可解释性；多解释方法融合；特征重要性；数据平衡
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000−436x.2025190

Intrusion detection method based on data augmentation and multi-explanation method fusion

XIONG Xuanrui[1], GUO Xingyou[1], NING Zhaolong[1], ZHANG Yushu[1], ZHOU Li[2]
1. School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: To address the issues of inconsistent explanations from existing interpretability methods and low decision credibility in intrusion detection systems, a multi-explanatory method fusion technique was proposed. The proposed method fused the advantages of shapley additive explanation (SHAP), local interpretable model-agnostic explanation (LIME) and permutation feature importance (PFI) by leveraging three evaluation metrics-consistency, focus, and stability-to establish an objective weighting mechanism, thereby generating more reliable feature importance interpretations. To tackle the poor explanation stability and low detection performance on minority classes caused by data imbalance, a data balancing technique was employed to provide a stable input foundation. Experimental results demonstrate that the proposed approach significantly enhances the reliability and consistency of model interpretations, while improving the overall performance of intrusion detection.

Keywords: intrusion detection, interpretability, multi-interpretation method fusion, feature importance, data balancing

0　引言

随着网络安全威胁的不断加剧，各类企业和组织对网络攻击防御与实时监测需求日益增强，这已成为网络安全领域需要考虑的重要问题之一[1]。在该背景下，基于深度学习的入侵检测（DLID，deep learning-based intrusion detection）技术因其能从大量网络流量中自动学习复杂特征模式，大幅提升了检测准确率和泛化能力，受到研究者的广泛关

收稿日期：2025−08−07；**修回日期：**2025−10−14
通信作者：宁兆龙，ningzl@cqupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62171449）；重庆市自然科学基金资助项目（No.CSTB2024NSCQ-JQX0013）
Foundation Items: The National Natural Science Foundation of China (No.62171449), The Natural Science Foundation of Chongqing (No.CSTB2024NSCQ-JQX0013)

注[2]。然而，DLID系统面临2个突出挑战：一是网络流量数据的严重不平衡分布问题，二是模型决策过程缺乏可解释性导致的可信度问题。

在实际应用环境中，正常流量通常占据绝大多数，而各类攻击流量相对稀少且种类多样，这种典型的类别分布不均衡，不仅导致模型训练过程偏向多数类，严重削弱了对少数类攻击的检测能力，还可能使模型的可解释性分析出现偏差，多数类特征在决策中占据主导地位，而反映攻击本质的少数类特征往往被忽视[3]。由于少数类样本稀少，解释方法在计算特征重要性时缺乏足够的样本支撑，导致解释结果的方差较大，不同解释方法在数据稀疏区域的表现差异更加明显。同时，现有解释方法如沙普利加性解释（SHAP, shapley additive explanation）、局部可解释模型（LIME, local interpretable model-agnostic explanation）和置换特征重要性（PFI, permutation feature importance）等往往给出不一致的解释结果，进一步削弱了系统的可信度[4]。在入侵检测场景中，这种解释不一致性不仅影响安全分析师对攻击模式的理解，还可能导致错误的防护策略制定，严重影响系统的实际部署效果。

数据平衡是解决不平衡分类问题的关键技术，通过调整各类别样本数量比例和优化样本分布特性可以有效提高分类器性能。在过采样方面，文献[5]提出了一种加权合成少数类过采样技术（SMOTE, synthetic minority over-sampling technique）。文献[6]将SMOTE过采样技术与深度循环神经网络相结合，用于处理物联网环境中的数据不平衡问题。近年来，生成模型也备受关注，文献[7]研究了扩散模型在类不平衡数据生成方面的应用，但该研究观察到其在处理长尾分布数据时存在生成多样性显著下降等问题。在欠采样策略方面，文献[8]基于重采样技术实现数据平衡。

除数据平衡外，模型可解释性也是入侵检测系统面临的重要挑战。现有解释方法中，SHAP[9]基于博弈论的Shapley值理论计算特征边际贡献；LIME[10]采用局部线性近似思想，在待解释样本附近构建简单可解释模型；PFI[11]通过随机打乱特征值评估特征对模型性能的影响。然而，这些方法在入侵检测应用中面临特殊挑战：网络流量数据的高维复杂性使解释结果难以直观理解；不同解释方法往往给出不一致的结果，削弱了解释的可信度；数据不平衡进一步恶化了解释方法在少数类样本上的稳定性。更重要的是，现有研究很少考虑数据平衡对解释方法性能的影响，缺乏数据平衡与解释方法的结合应用研究。

为了解决上述问题，本文提出了数据平衡方法和多解释方法融合技术，旨在提高网络入侵检测系统对少数类攻击的检测性能，同时增强模型决策的可解释性和可信度。

本文的主要贡献如下。

1) 设计多解释方法融合技术（MEMFT, multi-explanatory method fusion technology），通过一致性、聚焦性和稳定性3个评估指标融合SHAP、LIME和PFI等解释方法的优势。该融合基于三者在理论假设与输出特性上的互补性：SHAP提供全局一致性，LIME擅长局部可解释性，PFI反映特征扰动敏感性。一致性指标结合斯皮尔曼系数和Top-K重叠率评估解释稳定性；聚焦性指标通过熵值量化特征贡献度分布的集中程度；稳定性指标确保相似样本解释一致而不同样本体现差异。通过客观权重集成互补优势，生成更可靠的特征重要性排序。

2) 提出数据平衡方法，为解释方法提供更稳定的数据基础。采用基于注意力机制增强的卷积自编码器结合过采样技术生成高质量的少数类攻击样本，并通过混合欠采样策略实现数据类别平衡，增强解释方法在少数类样本上的稳定性。

3) 验证了数据平衡与多解释方法融合技术结合应用的有效性。通过在公开数据集上的实验，证明了本文方法在提升解释可靠性的同时显著改善了入侵检测性能，为网络安全领域的可解释人工智能应用提供了有价值的技术方案。

1 相关工作

入侵检测系统的性能优化涉及数据质量和模型可解释性2个核心问题。数据不平衡导致模型对少数类攻击检测能力不足，而深度学习模型的“黑箱”特性则带来了可解释性挑战。本节从数据平衡技术和可解释性方法2个方面回顾相关研究，分析现有方法的优势与局限，为本文方法提供理论基础。

1.1 入侵检测数据平衡方法

数据不平衡问题在入侵检测领域尤为突出，严

重影响模型对少数类攻击的检测能力。针对这一问题，研究者提出了多种数据平衡方法，主要分为过采样、欠采样和混合采样三大类。在过采样方面，经典的SMOTE算法[12]通过在少数类样本间进行线性插值生成合成样本，有效缓解了数据不平衡问题，但在高维复杂数据上容易产生噪声样本。自适应合成采样（ADASYN，adaptive synthetic sampling）算法[13]通过自适应密度分布，重点在少数类分布稀疏的区域生成更多样本。近年来，深度学习技术为数据平衡带来了新思路，条件生成对抗网络（CGAN，conditional generative adversarial network）[14]利用生成对抗网络通过条件标签指导生成过程，产生更加逼真的少数类样本。

在欠采样技术方面，主要通过减少多数类样本来实现数据平衡。Tomek Links[15]方法通过识别并删除类间边界的噪声样本，既减少了多数类样本数量又提高了数据质量。压缩最近邻（CNN，condensed nearest neighbour）算法采用原型选择策略，保留对维持决策边界至关重要的关键样本。混合采样策略综合了过采样和欠采样的优势，先通过过采样增加少数类样本，再通过欠采样清理重叠和噪声区域，取得了更好的平衡效果。

然而，现有数据平衡方法在网络入侵检测应用中仍存在不足：大多数方法未充分考虑网络流量数据的高维性和时序特性，缺乏对攻击行为完整性的保持；简单的几何插值难以捕捉复杂的攻击模式特征；在处理极度不平衡数据时效果有限。更重要的是，现有研究很少考虑数据平衡对后续解释方法性能的影响，缺乏数据平衡与解释分析的协同优化。

1.2　可解释性机器学习研究

随着深度学习模型在入侵检测中的广泛应用，其“黑箱”特性使模型决策过程难以被安全分析师理解和信任。为此，研究者提出了多种后验解释方法。Anagha等[16]提出基于SHAP的入侵检测可解释框架，通过生成局部解释并计算平均Shapley值，输出预测置信度，同时提供全局视角下的模型行为分析，但该方法在多分类场景下解释结果与专家知识的一致性仍有待提升。Guo等[17]针对复杂恶意行为的解释需求，提出LEMNA框架，通过构建代理模型模拟深度学习模型在决策边界附近的局部行为，有效捕捉特征依赖与非线性结构，但其解释局限于单一样本，缺乏全局可解释性。Pande等[18]利用局部归因技术量化各特征对预测的贡献度，为单个检测决策提供可追溯依据，但未涉及多方法结果的融合与一致性评估。Shakerin等[19]则探索了统计学习与逻辑学习的结合，提出SHAP-FOIL方法，利用SVM的支持向量和SHAP解释生成可读性强的逻辑规则，为深度学习模型提供了形式化、可验证的解释路径，但在处理高维网络流量数据时规则生成效率较低。然而，这些工作多聚焦于单一解释方法的应用，未充分考虑数据不平衡对解释稳定性的影响，且缺乏对不同解释结果的质量评估与融合机制，导致在关键特征排序上常出现不一致，影响了解释的可信度。

综上所述，现有入侵检测方法在处理数据不平衡和解释一致性问题时存在明显短板，数据平衡方法缺乏对网络攻击特征的深度理解，且未考虑与解释任务的协同优化；单一解释方法各有局限且在不平衡数据上表现不稳定，难以提供一致可信的特征重要性评估；而现有融合策略缺乏客观的质量评估机制，容易将低质量解释与高质量解释同等对待。因此，解决上述问题的关键在于构建数据平衡与解释融合的协同优化机制。为此，本文提出了一种基于数据增强和多解释方法融合的入侵检测方法。

2　数据平衡与多解释方法融合技术

2.1　SE-CAES数据平衡模型设计

本节提出了一种数据平衡方法，由2个核心模块构成。1) SE-CAES（squeeze-excitation convolutional autoencoder with SMOTE）生成模型采用基于注意力机制增强的卷积自编码器框架，通过自适应特征选择和跳跃连接设计提升特征提取能力。模型结合SMOTE过采样技术，在潜在特征空间进行插值采样，生成高质量的少数类攻击样本。2) 混合欠采样模块采用CNN与Tomek Links协同策略，通过原型选择保留关键决策边界样本，同时剔除类间重叠区域的冗余样本，实现多数类样本的有效欠采样。通过上述数据平衡处理，为后续的多解释方法融合提供类别均衡、特征空间可分性增强的稳定数据基础。

2.1.1　SE-CAES生成模型训练过程

SE-CAES生成模型的训练包括重构路径和扰动路径两部分。SE-CAES训练阶段模型如图1所示。在SE-CAES模型的训练阶段重构路径中，攻击样本以批次形式输入编码器，编码器通过多层卷

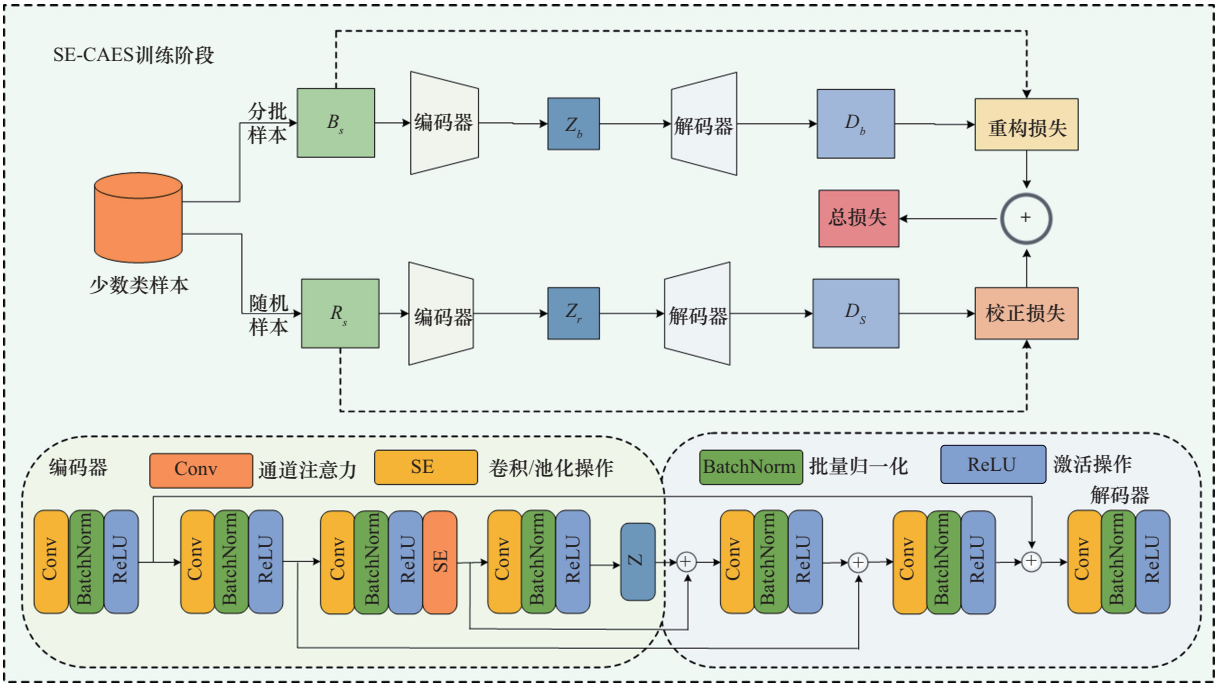


图 1　SE-CAES 训练阶段模型

积提取特征。在第 3 层卷积后，SE 注意力模块通过全局信息嵌入和自适应提取输入数据的关键信息，生成特征表示；然后，编码器通过全连接层输出潜在特征。与此同时，第 3 层卷积的输出通过跳跃连接直接传递至解码器，引导重构数据。解码器接收潜在特征后，先通过全连接层和 reshape 操作恢复形状，并结合跳跃连接的特征进行拼接。随后，通过多层转置卷积上采样，最终恢复与原始输入一致的重构数据。整个过程通过最小化重构损失优化，确保准确恢复输入数据。在本实验中，为评估模型重建输入数据的能力，模型采用了 Smooth L1 损失作为重构损失。Smooth L1 的核心优势在于：当预测误差 $|x|<1$ 时，其表现为二次函数形式 $(0.5x^2)$，此时损失曲线的梯度随误差减小而平滑下降，可引导模型对小误差样本进行精细化参数调整，避免传统 L2 损失因梯度爆炸导致的训练不稳定。当误差 $|x|>1$ 时，函数退化为线性形式，梯度大小恒定为 1。其数学表达式为

$$R_{\mathrm{L1}}=\mathrm{Smooth}_{\mathrm{L1}}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases} \tag{1}$$

其中，$x=D_b-B_s$ 是分批提取的少数类样本，D_b 是经过解码后重构的样本。

在 SE-CAES 模型训练阶段的扰动路径中，模型对随机采样的样本进行特征提取后，随机打乱潜在表示的顺序，增强模型对不同样本组合的适应能力，提升泛化性能。在这个过程中，依然采用 Smooth L1 损失函数作为重构损失。最终，将上下两部分的损失函数相结合，得到模型的总损失函数为

$$T_{\mathrm{L}}=R_{\mathrm{L1}}+R_{\mathrm{L2}} \tag{2}$$

SE-CAES 模型训练的算法流程如算法 1 所示。

算法 1　SE-CAES 模型训练的算法流程

输入　样本集 $M=\{M_1,M_2,\cdots,M_n\}$（含 n 个少数类别）

输出　训练好的 SE-CAES 生成模型

1)Begin

2) for 训练轮次 $e=1,2,\cdots,E$ do

3)　for 类别 $i=1,2,\cdots,n$ do

4)　　for 样本 $j=1,2,\cdots,l$ do

5)　　使用编码器对每个批次样本 B_s 进行特征编码，得到潜在表示 Z_b，同时通过跳跃连接将编码器中间层特征传递给解码器

6)　　解码器结合跳跃连接传递的特征信息，对 Z_b 进行重构，生成 D_b

7)　　计算重构损失函数 $R_{\mathrm{L1}}\leftarrow\mathrm{SmoothL1}(D_b,B_s)$

8)　　从每个批次样本中随机抽取与 B_s 同等数量的样本 R_s

9) 使用编码器对R_s进行特征编码，得到潜在表示Z_r，在编码过程通过跳跃连接的中间层特征传递给解码器

10) 使用置换操作打乱Z_r的样本顺序，得到Z_r^p

11) 解码器结合跳跃连接传递过来的特征信息，对Z_r^p进行重构，得到D_r

12) 计算重构损失函数$R_{\mathrm{L2}}\leftarrow$Smooth L1$(D_r,R_s)$

13) 计算模型的总损失函数$T_{\mathrm{L}}\leftarrow R_{\mathrm{L1}}+R_{\mathrm{L2}}$

14) 使用优化器根据T_{L}更新模型参数

15) end for

16) end for

17) end for

18) Return 训练好的SE-CAES生成模型

19) End

2.1.2 SE-CAES模型生成过程

模型训练完成后，SE-CAES模型在生成阶段通过深度过采样机制生成少数类样本。在训练过程中，模型利用随机置换操作引入数据的扰动；而在生成过程中，模型借助SMOTE过采样技术在潜在特征空间内进行精确插值。

在SE-CAES模型的生成阶段，训练好的生成模型被用来合成少数类攻击样本。合成过程如下：将少数类样本输入编码器，提取具有高度区分性的嵌入表示；基于这些嵌入，模型应用SMOTE算法生成合成数据；然后，使用解码器将生成的特征向量转换回原始数据空间，最终得到高质量的少数类攻击样本，为后续多解释方法融合提供稳定数据基础。SE-CAES模型的生成算法流程如算法2所示。

算法2 SE-CAES模型生成新样本的算法流程

输入 样本集$M=\{M_1,M_2,\cdots,M_n\}$（含n个类别少数类别）、SE-CAES生成模型

输出 生成的少数类样本S_i

1) Begin

2) for i = 1,2,⋯,n do

3) 将少数类M_i输入训练好的编码器,得到编码后的特征Z_i，同时通过跳跃连接将编码器中间层特征传递给解码器

4) 对所有潜在特征表示使用SMOTE算法,生成新的合成潜在特征N_i

5) 将N_i输入训练好的解码器，并结合跳跃连接的特征信息，对N_i进行重构，生成S_i

6) end for

7) Return生成的少数类样本S_i

8) End

2.1.3 混合欠采样模型

CNN是一种基于原型选择的欠采样技术，旨在从多数类中筛选出对维持分类决策边界至关重要的关键样本。给定包含所有训练样本的数据集$D=(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，其中，$\boldsymbol{x_i}$是特征向量，$y_i$是标签。算法首先保留所有样本，然后迭代评估每个样本的重要性，逐步删除不影响分类性能的冗余样本，直到无法继续删除。对于每个样本x_i，通过计算与其他样本的距离来确定其最近邻集。

$$N(x_i)=\{x_j\in D\,|\,\mathrm{distance}(x_i,x_j)\leqslant\mathrm{distance}(x_i,x_k),\forall k\neq i\}\tag{3}$$

其中，$\mathrm{distance}(x_i,x_j)$是样本$x_i$和$x_j$之间的距离。

Tomek Links是一种用于处理不平衡数据集的技术，通过识别并删除类别间的边界样本，减少噪声并提高分类性能。如果2个互为最近邻的样本属于不同类别，并且一个来自多数类，一个来自少数类，则这2个样本构成一个“Tomek Link”。在Tomek Link中，通常将多数类样本视为冗余样本并予以移除，因为删除它们不会显著影响决策边界的完整性。Tomek Links的核心思想是寻找样本对(x_i,x_j)，其中x_i和x_j属于不同类别，并且它们彼此是最近邻。通过给定数据集$D=(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，且样本$x_i$和$x_j$之间的距离为$\mathrm{distance}(x_i,x_j)$。如果样本$x_i$和$x_j$分别属于不同类别，并且它们之间的距离是各自的最近邻距离，即

$$\mathrm{distance}(x_i,x_j)=\min(\mathrm{distance}(x_i,x_k),\mathrm{distance}(x_j,x_l))\tag{4}$$

若$k\neq i$且$l\neq j$，则称样本对(x_i,x_j)为一个Tomek Link。Tomek Links方法通过识别并删除边界样本对来清理数据集。

2.2 多解释方法融合技术

利用SE-CAES模型生成少数类样本和利用欠采样方法减少多数类样本数量以缓解数据不平衡问题后，模型的可解释性成为保障检测结果可信的关键环节。由于原始数据集中少数类样本稀疏，单一解释方法容易因输入扰动或模型结构敏感性产生不

一致的结果，影响分析人员对威胁来源的判断。现有融合方法多采用简单平均或经验赋权的方式，难以有效应对不同解释方法在稳定性、聚焦性等方面的差异，容易将低质量解释与高质量解释同等对待，导致融合结果失真。为克服这一局限，本文提出MEMFT，通过系统性评估与动态加权，提升融合解释的可靠性。

MEMFT的设计源于对SHAP、LIME与PFI内在机制差异的深入分析：SHAP基于Shapley值，理论严谨且满足一致性，但对特征独立性假设敏感，在高维冗余特征下易产生偏差；LIME通过局部线性近似提供直观解释，但结果依赖扰动策略，跨样本稳定性差；PFI评估特征对整体性能的影响，方法高效，但无法提供样本级解释，且在特征相关时可能低估重要性。这些方法在理论基础、解释视角和输出粒度上的本质差异，导致其在关键特征排序上常出现显著分歧，形成“解释冲突”。

为系统性应对这一问题，MEMFT从3个维度量化评估各方法在当前样本上的表现质量，并据此动态调整融合权重：一是一致性，衡量不同方法在关键特征排序上的吻合程度，反映解释共识性，采用斯皮尔曼相关系数与Top-K特征重叠率综合评估，量化SHAP与LIME等方法在排序上的分歧程度，并以此作为调整融合权重的依据；二是聚焦性，通过熵值量化特征重要性分布的集中程度，抑制冗余特征干扰，避免解释结果过度分散，针对SHAP在高维冗余下的偏差与LIME解释分散的弱点；三是稳定性，评估邻近样本间解释结果的波动程度，保障局部解释的鲁棒性，专门应对LIME与PFI在扰动或打乱下的不稳定性。这3项指标分别从“协同性”“判别性”和“鲁棒性”角度刻画了解释方法的可靠性。

基于上述评估结果，MEMFT通过非线性映射与归一化计算融合权重，使在一致性、聚焦性和稳定性方面表现出更优的解释方法，获得更高贡献。该“先评估、后融合”的策略，实现了从“差异识别”到“可信融合”的闭环，不仅增强了融合过程的系统性，也有效克服了传统方法因忽视方法性能差异而导致的解释偏差，提升了特征重要性评估的准确性与可信度。

2.2.1 评估指标设计

1) 一致性指标设计：一致性指标评估解释方法在多次应用时是否能保持一致的结果。若某一解释方法具有较高的一致性，说明模型在不同数据样本下给出的特征重要性解释比较稳定；相反，若多次计算中的特征贡献度差异过大，则说明该解释方法的可靠性较差。指标结合斯皮尔曼系数和重叠率，综合评估解释方法在不同维度上的一致性。斯皮尔曼系数用于评估特征重要性得分之间的排名一致性，其计算式为

$$\rho = 1 - \frac{6\sum_{i=1}^{n} r_i^2}{n\left(n^2 - 1\right)} \tag{5}$$

其中，r_i是样本i的排名差异，n是特征的数量。

由式(5)可知，如果一个解释方法在多次计算中给出的特征重要性排名几乎相同，它的ρ值越大，表明重要性排名相似度很高，解释方法的一致性越好；相反，如果排名变化显著，ρ值越小，说明解释结果一致性较差。Top-K重叠率则关注关键特征的一致性，确保模型对最重要特征的一致性判断，尤其是对前K个重要特征的稳定解释。其主要公式为

$$T=\frac{\text{Top-K中重叠的样本数量}}{\lambda} \tag{6}$$

其中，Top-K是每次计算排名中的前λ的特征，该比率表示每次计算中排名前λ的特征之间的重叠程度。

一致性指标的计算需要对同一解释方法进行至少3次特征重要性得分的计算，通过计算多次排名之间的平均值，可以客观评估解释方法的一致性。然后将求得的斯皮尔曼系数和Top-K重叠率相结合，得到解释方法的综合评估结果。其一致性计算式为

$$C_k^d = \frac{\rho_k^d + 1}{4} + \frac{T_k^d}{2} \tag{7}$$

其中，C_k^d为第d个样本对应第k种解释方法的一致性分数。

2) 聚焦性指标设计：聚焦性指标用于评估解释的清晰度和简洁性。随着特征数量的增加，解释方法可能依赖于众多特征，但通常只有少数关键特征对预测结果起决定性作用。理想的模型解释应突出这些关键特征，避免信息分散。指标通过量化特征贡献度分布的集中程度来衡量解释的聚焦性。其表达式为

$$B_k^d = 1 - \frac{H_k^d - \min\left(H_k^d\right)}{\max\left(H_k^d\right) - \min\left(H_k^d\right)} \tag{8}$$

其中，H_k^d表示样本d在解释方法k下的特征贡献度分布的熵值，计算式为

$$H_k^d = -\sum p_i^d \mathrm{lb}\left(p_i^d\right) \tag{9}$$

其中，p_i^d是归一化后的特征贡献度概率分布，定义为

$$p_i^d = \frac{f_i^d}{\sum_{j=1}^{n} f_j^d} \tag{10}$$

其中，f_i^d表示样本d第i个特征的特征贡献度，B_k^d为第d个样本对应第k种解释方法的聚焦性分数。熵值H_k^d用于衡量特征贡献度分布的混乱程度。当特征贡献度集中于少数几个特征时，熵值H_k^d趋近于0，此时聚焦性较高，模型解释更简洁且清晰。当特征贡献度均匀分散于多个特征时，熵值H_k^d趋近于$\log(n)$，此时聚焦性较低，解释的清晰度下降。最后通过归一化并反转熵值，得到的聚焦性指标B_k^d越大，则模型的显著性越好。

3) 稳定性指标设计：稳定性指标确保解释方法在相似样本上保持一致性，同时在不同样本上体现出区别。理想的解释方法应对预测结果相同且特征分布相似的样本给出相似的特征重要性评分，对预测结果不同的样本给出明显差异的评分。本文采用欧几里得距离量化样本间的相似性。对于预测结果一致的样本，选取特征空间中距离最近的15%样本作为对比对象，通过式(11)量化评估，即

$$U_{\text{same}}^d = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{N_{\text{same}}}\sum_{j=1}^{N_{\text{same}}}\left(E_i^d - E_{ij}^{\text{dnsame}}\right)\right] \tag{11}$$

其中，U_{same}^d表示第d个样本在预测结果相同条件下的可靠性指标；E_i^d是通过解释方法计算得到的第i个样本的特征重要性评分；E_{ij}^{dnsame}是第i个样本的第j个相邻样本的特征重要性评分；N_{same}是每个样本的相邻样本数量，即前15%的样本数。理想情况下，U_{same}^d的值应接近于0，意味着相邻样本间的特征重要性评分差异极小，解释方法在相似样本上表现出高度的稳定性。

同时考虑预测结果不同的样本，选取特征空间中距离最远的15%数据点进行对比。若这些远离样本的特征重要性评分与当前样本存在明显差异，则说明解释方法能够有效区分不同类型样本，具备较高的稳定性。计算式为

$$U_{\text{diff}}^d = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{N_{\text{diff}}}\sum_{j=1}^{N_{\text{diff}}}\left(E_i^d - E_{ij}^{\text{dndiff}}\right)\right] \tag{12}$$

其中，U_{diff}^d表示第d个样本在预测结果不同条件下的可靠性指标，E_{ij}^{dndiff}是第i个样本的第j个不相邻样本的特征重要性评分，N_{diff}是每个样本的不相邻样本数量。与U_{same}^d不同，在理想情况下，U_{diff}^d应该较大，表明不同类型样本的特征重要性评分存在显著差异，解释方法能够有效反映样本间的决策差异。

为了全面评估解释方法的稳定性，将U_{same}^d和U_{diff}^d结合进行分析。这2个指标的方向性相反，U_{same}^d越小越好，表示相邻样本的解释相似；而U_{diff}^d越大越好，表示不同类型样本的解释差异大。因此，需要对它们进行归一化处理，并调整方向以统一评估标准。最终的稳定性指标U_k^d计算式为

$$U_k^d = \frac{U_{\text{diff}}^d - \min\left(U_{\text{diff}}\right)}{\max\left(U_{\text{diff}}\right) - \min\left(U_{\text{diff}}\right)} + \left(1 - \frac{U_{\text{same}}^d - \min\left(U_{\text{same}}\right)}{\max\left(U_{\text{same}}\right) - \min\left(U_{\text{same}}\right)}\right) \tag{13}$$

将U_{same}^d和U_{diff}^d结合后，得到一个综合的稳定性指标。该指标值越大，表示解释方法在保持相似样本稳定性以及区分不同样本差异性方面的表现越好。

2.2.2　融合计算方法

在解释方法融合前，需要分析每个指标的变异性及与其他指标的相关性，计算稳定性、聚焦性和一致性的客观权重。这些权重将作为融合过程中的关键依据，确保每个指标在最终解释中的贡献与其信息价值相匹配。其权重计算式为

$$W_j = \frac{M_j}{\sum_{j=1}^{3} M_j} \tag{14}$$

每个指标的权重W_j是基于其信息量M_j与所有指标信息量之和的比值，确保权重的总和为1，并且指标的信息量越大，它在整个体系中的权重也越高。指标的信息量M_j是由指标的变异性和相关性冲突共同决定的。变异性越大、相关性冲突越大，说明该指标提供的信息越丰富。信息量的计算式为

$$M_j = V_j R_j \tag{15}$$

其中，V_j表示第j个指标的变异性，R_j表示第j指标的相关性冲突。

指标的变异性V_j反映了该指标在所有样本中的差异波动情况，通常使用标准差来衡量。V_j的计算式为

$$V_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2} \tag{16}$$

其中，x_{ij}为第j个指标的第i个样本值，$\overline{x_j}$为第j个指标的均值，n为样本数量。

指标的相关性R_j冲突反映了某一指标与其他指标的关系。如果2个指标的相关性较低，说明它们提供的信息较为独立，因此该指标的独立性较强，信息量较大。R_j的计算式为

$$R_j = \sum_{i=1}^{3}\left(1 - r_{ij}\right) \tag{17}$$

其中，r_{ij}为第j个指标与第i个指标的相关系数。

在确定了各指标的权重之后，接着对SHAP、LIME和PFI这3种解释方法进行融合计算。对于每种解释方法，分别计算其在一致性、聚焦性和稳定性上的样本均值，分别记为$\bar{C}_k$、$\bar{B}_k$和$\bar{U}_k$。基于这些均值，构造一个3×3矩阵$\boldsymbol{A}$，其行对应3种解释方法，列对应3种评估指标。矩阵元素a_{ij}表示第i种方法在第j个指标上的均值得分。为消除不同指标间的量纲差异，对矩阵$\boldsymbol{A}$的每一列进行min-max归一化，计算式为

$$\hat{a}_{ij} = \frac{a_{ij} - \min\left(a_{1j},a_{2j},a_{3j}\right)}{\max\left(a_{1j},a_{2j},a_{3j}\right) - \min\left(a_{1j},a_{2j},a_{3j}\right)} \tag{18}$$

归一化后，求对应每种解释方法的3个评价指标维度下的综合解释性表现。对于解释方法k在3个指标维度下的综合得分g_k计算式为

$$g_k = \sum_{j=1}^{3} W_j \hat{a}_{kj} \tag{19}$$

其中，g_k反映了方法k的整体解释能力的评分，W_j是第j个指标的权重，$\hat{a}_{kj}$是方法k在第j个指标上的归一化得分。

将综合得分g_k作为权重，应用于方法k给出的特征重要性得分向量$\boldsymbol{E}^k$，得到加权特征重要性S^k为

$$S^k = g_k \boldsymbol{E}^k \tag{20}$$

为确保不同方法间的特征重要性尺度一致，对S^k进行行归一化处理。对于方法k的第i个特征，其归一化后的重要性$\hat{s}_i^k$为

$$\hat{s}_i^k = \frac{s_i^k - \min\left(s_i^k\right)}{\max\left(s_i^k\right) - \min\left(s_i^k\right)} \tag{21}$$

式中，s_i^k是原始加权得分，$\min\left(s_i^k\right)$和$\max\left(s_i^k\right)$分别是s_i^k中的最小值和最大值。

将3种方法的归一化特征重要性$\hat{S}^k$相加，得到最终的综合特征重要性得分为

$$S_f = \sum_{k=1}^{3} \hat{S}^k \tag{22}$$

最后，对S_f进行排序，生成最终的特征重要性排名。

2.3 数据平衡与多解释方法结合应用

为实现数据平衡与解释融合的有效结合，本文构建了完整技术流程，基于数据平衡与多解释融合的入侵检测方法流程如图2所示。数据平衡模块首先对原始数据集进行处理，采用SE-CAES生成技术和混合欠采样策略，将不平衡的网络流量数据转换为类别均衡的数据集。基于平衡数据集，入侵检测模型模块完成模型训练、分类检测和结果输出的完整流程。多解释方法融合模块并行运行SHAP、LIME和PFI3种解释算法，通过融合权重计算生成综合解释结果。最终通过特征选择和性能评估，验证2种技术结合应用的效果。

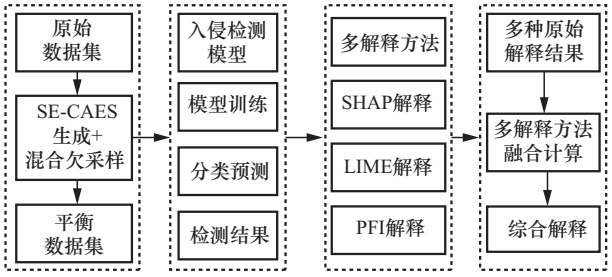


图2 基于数据平衡与多解释融合的入侵检测方法流程

该方法以数据平衡为解释融合提供更好的数据基础。在原始数据中，少数类样本稀疏且分布不均，导致SHAP、LIME等解释方法对相似样本的输出波动剧烈，评估指标的可靠性下降，进而影响融合权重的合理性。SE-CAES通过生成高质量的少数类样本并优化特征空间分布，显著提升了同类

样本的可分性与解释一致性。在此基础上，解释融合技术整合SHAP、LIME和PFI的优势，通过一致性、聚焦性和稳定性3个评估指标的客观权重计算，生成更可靠的特征重要性评估结果。通过数据平衡与多解释方法的结合应用，能够在保证分类性能的同时提升解释方法的效率和可靠性，实现检测性能与特征选择效果的双重提升。

3　实验与结果分析

3.1　实验设置

3.1.1　实验数据集及数据预处理

为验证本文方法在多样化网络环境下的普适性，实验在CICIDS-2017、UNSW-NB15、NSL-KDD和5G Threat这4个数据集上进行。CICIDS-2017由加拿大网络安全研究所创建，UNSW-NB15由澳大利亚新南威尔士大学发布，NSL-KDD是KDD Cup 99数据集的改进版本，三者均包含正常流量和多种攻击类型。2个大规模数据集（CICIDS-2017和UNSW-NB15）样本数均超250万，NSL-KDD包含约17万样本，且均存在显著的类别不平衡问题：CICIDS-2017中正常流量占比超过80%，部分攻击（如SQL Injection、Heartbleed）样本数极少；UNSW-NB15中正常流量占主导地位，部分攻击（如Generic、Analysis）样本数相对较少；NSL-KDD中正常流量占比约78%，DoS等攻击类别同样存在样本稀疏现象。5G Threat数据集是本文在5G半实物仿真环境中采集的数据集，包含正常与攻击流量，生成过程详见3.1.2节。数据预处理包括：对字符类型特征（“proto”“service”等）进行标签编码，将类别映射为整数值；对所有特征进行归一化处理，消除量纲差异对模型的影响。

本文采取的归一化方法计算式为

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{23}$$

其中，x'是经过归一化后的值，x是原始特征值，$x_{\min}$和$x_{\max}$分别是特征的最小值和最大值。

3.1.2　5G网络仿真环境构建与数据采集

为评估本文方法在新型5G网络环境中的适用性，本文基于Flex Stack5GC、free5GC与OAI-5G这3种开源核心网，并结合USRP-LW X310软件无线电平台构建端到端高保真仿真环境。可使用商用用户终端并通过USRP搭建的基站连接核心网平台与外部真实因特网服务器进行连接，实现文件下载、网页浏览、实时视屏观看等功能，并依托此平台生成了5G Threat数据集。经3GPP协议验证，正常信令流程在时序与状态转换上与真实网络表现出较高一致性。实验从用户侧、接入侧与核心侧构建5类典型5G安全威胁，分别是UE频繁上下线DDoS攻击、NGReset攻击、位置泄露攻击（LDA）、中间人攻击（MMA）和切片服务攻击（NSA）。5G网络仿真环境硬件配置如图3所示，5G仿真环境中USRP信号收发装置配置如图4所示，该平台支持多类型攻击注入与关键接口流量采集，构成完整的端到端测试环境。

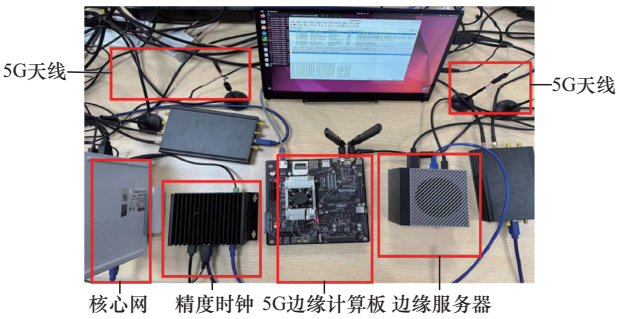


图3　5G网络仿真环境硬件配置

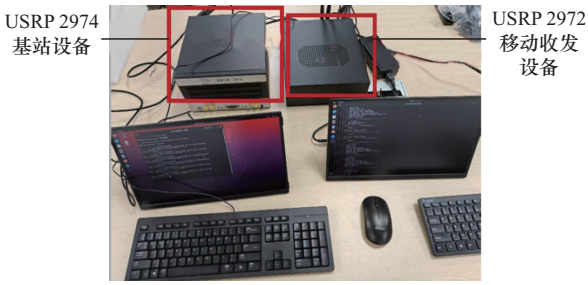


图4　5G仿真环境中USRP信号收发装置配置

5G Threat数据集通过在N2/N3等关键接口使用TShark捕获原始流量，随后基于CICFlowMeter工具将流量划分为双向流。该数据集共包含40 571条标注样本。其中正常流量32 675条，DDoS攻击3 755条，LDA攻击1 323条，NGReset攻击306条，NSA攻击1 033条，MMA攻击1 479条，整体呈现显著的类别不平衡特性，符合真实网络中攻击稀发的实际情况。该数据集不仅涵盖了5G特有攻击，也包含在多代移动网络中长期存在的安全威胁，为评估入侵检测方法在复杂、异构网络环境下的解释能力与检测性能提供了高保真验证基础。

3.1.3　评价指标

在评估入侵检测系统的性能时，考虑到数据的复杂性和不平衡性，本文采用多维度评价指标体系。微观层面使用精确率P（Precision）、召回率R（Recall）和F1（F1-score）分数评估各类别表现。TP、FP、FN分别表示真正例、假正例和假负例。微观层面评估指标的计算式为

$$\begin{cases}\text{Precision}=\dfrac{\text{TP}}{\text{TP+FP}}\\ \text{Recall}=\dfrac{\text{TP}}{\text{TP+FN}}\\ \text{F1-score}=\dfrac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}\end{cases}\tag{24}$$

宏观层面使用宏观精确率（Macro_P）、宏观召回率（Macro_R）和宏观F1值（Macro_F1）评估整体性能，通过对所有类别性能平均化处理，弥补微观评估在处理不平衡数据时的局限性。宏观层面评估指标的计算式为

$$\begin{cases}\text{Macro\_P}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}P_i\\ \text{Macro\_R}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}R_i\\ \text{Macro\_F1}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\text{F1}_i\end{cases}\tag{25}$$

其中，n为类别总数，P_i、R_i和F1_i分别为第i类样本的精确率、召回率和F1-score。

3.2　数据平衡方法性能验证

3.2.1　数据平衡后的实验对比分析

为有效验证所提出的数据平衡方法，本文采用逻辑回归、随机森林、决策树和XGBoost这4种具有代表性的机器学习模型进行实验。选用Macro_P、Macro_R和Macro_F1作为核心评价指标，这些宏观指标通过对各类别指标取算术平均得到，能充分反映数据不平衡下各类别的性能表现。

实验分2个阶段：第1阶段在原始数据集训练测试，第2阶段在平衡数据集进行测试，数据平衡前后各分类器在测试集上分类对比如图5所示。数据平衡后，4种分类器的Macro_R分别提升18.09%、20.91%、22.40%和17.47%，其中决策树提升最显著，表明对少数类检测能力增强。Macro_F1明显提高，LR、RF、DT、XGBoost分别提升15.32%、19.12%、19.79%和17.33%，表明模型在攻击精确率与召回率上更趋均衡。

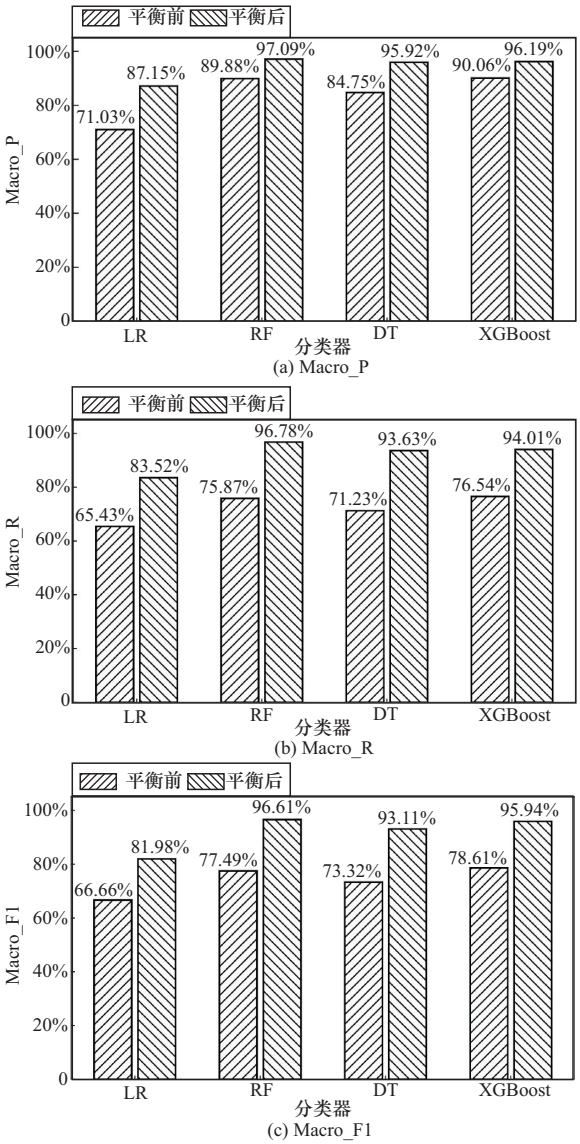


图5　数据平衡前后各分类器在测试集上分类对比

3.2.2　消融实验分析

为验证各模块有效性，本文设计消融实验对比4种方法：方法1为基准线（无数据平衡），方法2采用SE-CAES生成模型，方法3采用混合欠采样技术，方法4结合2种技术。消融结果对比如表1所示，方法2相比基准线在Precision、Recall和F1-score上分别提升5.81%、16.05%和15.73%，其中Recall提升最显著，证明SE-CAES有效提升少数类检测性能；方法3的Precision提升最为明显；方法4实现最佳综合性能，验证了2种技术的组合优势。结果表明各模块均有效贡献数据平衡性能，完整方案实现最优效果。

表1　消融结果对比

方法	模块		宏平均		
	SE-CAES	混合欠采样	Precision	Recall	F1-score
方法1	—	—	89.88%	75.87%	77.49%
方法2	√	—	95.69%	91.92%	93.22%
方法3	—	√	96.53%	83.56%	84.98%
方法4	—	√	96.32%	96.78%	96.61%

3.2.3　不同数据平衡方法性能对比

SE- CAES 与其他数据平衡方法的综合指标对比如图 6 所示，本文 SE-CAES 方法与 ADASYN、CGAN 和 VAE-WGAN[20]等主流数据平衡方法对比中，在 3 个指标上均取得最佳性能。SE-CAES 优势在于：SE 注意力机制自适应识别关键特征，卷积自编码器与 SMOTE 结合增强样本多样性，混合欠采样策略保留决策边界关键信息。

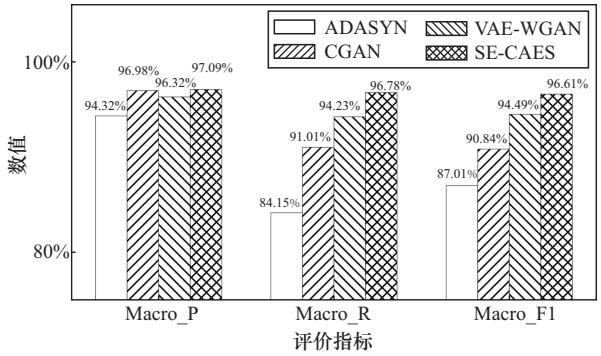


图6　SE- CAES 与其他数据平衡方法的综合指标对比

3.3　多解释方法融合技术性能验证

3.3.1　MEMFT 模型提取的前 20 个最重要特征

本文多解释方法融合技术在不同数据集上提取了前 20 个最重要的特征，并针对每个数据集中的各类攻击类别，统计了这些特征的重要性排序，特征重要性从左到右依次递减。MEMFT 在未平衡的 CICIDS-2017、UNSW-NB15 和 5G Threat 这 3 个数据集上提取的最重要特征分别如表 2、表 3 和表 4 所示。

根据表2结果，在CICIDS-2017数据集中，多个攻击类型的重要特征高度重叠，但由于攻击方式的不同，其特征重要性排序存在差异。特征65、67、64和15在多种攻击场景下均表现突出，可视为该数据集的关键指标。MEMFT模型在分类时并非依赖少量特征，而是结合多种特征信息进行综合判断。部分攻击类别展现出独有的特征偏好，反映了特定攻击模式的差异性。

表2　MEMFT 在未平衡的 CICIDS-2017 数据集上提取的最重要特征

类别	提取出的前 20 个最重要特征
DoS Hulk	65,64,9,36,53,41,11,15,38,17,10,67,20,61,12,14,5,35,40,22
Port Scan	10,61,51,65,39,36,13,15,3,17,20,67,11,53,14,45,22,0,38,19
DDoS	67,65,9,5,8,61,53,11,38,64,3,7,66,4,63,40,36,41,51,52
DoS GoldenEye	9,67,15,65,53,17,14,36,64,40,38,12,11,27,41,25,16,20,21,26
FTP-Patator	65,15,67,14,8,38,17,36,10,20,51,41,40,52,21,3,5,39,53,7
SSH-Patator	65,67,36,64,13,15,10,17,14,20,47,5,2,16,50,35,51,40,28,26
DoS Slowloris	65,67,15,36,20,14,17,11,10,38,53,9,41,16,40,22,64,21,28,33
DoS Slow httptest	15,36,67,71,65,20,14,17,68,16,53,9,51,10,34,11,39,38,35,21
Bot	67,65,10,15,51,11,17,53,36,14,13,61,39,20,28,63,35,9,38
Brute Force	63,17,36,15,67,20,14,64,16,21,10,23,35,19,22,9,18,0,53,13
XSS	65,17,15,36,20,67,14,23,64,0,16,21,19,22,18,3,35,10,39,33
Infiltration	67,10,53,11,65,15,19,0,9,72,36,2,61,22,17,24,3,37,68,4
SQL Injection	65,27,64,15,17,67,25,0,24,11,53,12,9,20,23,36,14,21,41,10
Heartbleed	64,53,7,67,65,61,9,15,11,5,52,41,1,13,35,38,40,26,19,34
综合	65,64,15,67,39,5,61,52,19,41,40,8,20,36,9,7,51,38,14,13

根据表 3 结果，在 UNSW-NB15 数据集中，特征 6、26、40、35 和 9 在几乎所有攻击类型中都被选中，表明这些特征具有较强的判别能力。不同攻击类别呈现特征独特性，如 Exploits 类别中特征 10 频繁出现，Worms 类别中则为特征 25，反映了不同攻击的行为差异。此外，特征 3、18、20、21、22、23、28、36、37 几乎未被选中，表明这些特征在该场景下对攻击区分的贡献较小。

表3 MEMFT在未平衡的UNSW-NB15数据集上提取的最重要特征

类别	提取出的前20个最重要特征
Fuzzers	6,40,9,2,26,35,30,34,7,27,13,33,4,1,11,24,12,16,17,5
Analysis	6,26,40,35,34,1,30,11,2,9,33,7,27,39,15,29,19,32,13,0
Backdoor	6,26,1,2,34,40,35,9,39,11,30,7,13,15,27,19,33,8,32,4
DoS	6,26,40,1,9,30,35,34,19,33,11,2,27,7,16,13,31,15,39,24
Exploits	40,6,9,27,7,26,30,14,31,35,2,10,1,11,33,12,5,13,0,8
Generic	6,2,40,33,26,34,35,30,32,11,39,9,0,8,1,4,31,13,19,16
Reconnaissance	6,26,40,35,27,9,30,17,7,33,11,2,14,34,12,13,1,5,4,0
Shellcode	35,6,26,40,2,33,1,34,27,17,30,7,9,11,14,4,0,13,15,8
Worms	6,26,9,40,2,35,30,7,13,38,27,24,5,11,4,12,25,10,14,31
综合	6,26,40,35,9,11,30,2,34,33,12,1,7,27,8,31,15,0,39,32

根据表4结果，在5G Threat数据集中，特征65、67、36和15在多数攻击中排名靠前，表明其在5G环境下的普遍有效性。不同攻击的特征排序存在差异，如LDA攻击中特征10靠前，MMA攻击则对特征20更为敏感。MEMFT能结合多特征进行综合判断，并有效捕捉攻击行为的特异性与关键模式。

表4 MEMFT在未平衡的5G Threat数据集上提取的最重要特征

类别	提取出的前20个最重要特征
DDoS	65,67,15,36,20,14,38,64,53,17,11,10,9,40,41,5,61,39,63,22
NGReset	67,65,36,20,38,15,17,14,64,53,11,9,10,40,41,63,5,61,39,7
LDA	10,65,17,67,15,20,36,14,64,11,13,9,5,61,0,19,35,40,21,49
MMA	20,38,65,67,36,15,17,14,64,53,11,10,9,40,41,63,5,61,39,22
NSA	65,15,67,36,17,14,20,38,64,53,11,10,9,5,61,39,40,41,63,42
综合	65,67,36,15,20,17,14,64,53,11,10,9,38,61,5,40,41,63,13,22

3.3.2 MEMFT模型筛选的少量特征与原始特征维度进行多分类实验对比

为验证本文MEMFT方法解释性的优越性，在CICIDS-2017、UNSW-NB15和5G Threat这3个数据集上进行了多分类实验。在CICIDS-2017数据集上，对比了使用MEMFT提取的30维重要特征（性能达到峰值）与原始76维特征的分类性能；在UNSW-NB15数据集上，对比了使用11维特征（峰值）与原始43维特征的性能；在5G Threat数据集上，对比了使用40维特征（峰值）与原始76维特征的性能。基于未平衡CICIDS-2017、UNSW-NB15和5G Threat这3个数据集多分类实验结果分别如表5、表6和表7所示。

表5 基于未平衡CICIDS-2017数据集多分类实验结果

类别	Precision		Recall		F1-score	
	76维	30维	76维	30维	76维	30维
BENIGN	99.86%	99.96%	99.84%	99.93%	99.85%	99.89%
DoS Hulk	98.99%	99.82%	99.58%	99.91%	99.28%	99.86%
PortScan	99.39%	99.39%	99.96%	99.97%	99.67%	99.68%
DDoS	99.94%	99.97%	99.91%	99.92%	99.93%	99.95%
DoS Golden-Eye	99.83%	100%	98.57%	98.77%	99.20%	99.38%
FTP-Patator	99.91%	99.96%	99.87%	99.87%	99.89%	99.92%
SSH-Patator	100%	100%	93.95%	98.98%	96.88%	99.49%
DoS slow-loris	99.71%	99.71%	98.85%	98.91%	99.27%	99.31%
DoS Slowhttpt-est	95.19%	95.26%	98.36%	98.67%	96.75%	96.93%
Bot	98.73%	96.61%	39.83%	62.71%	56.76%	76.05%
Brute Force	69.11%	96.15%	87.61%	16.59%	77.26%	28.30%
XSS	87.50%	100%	3.57%	2.55%	6.86%	4.97%
Infiltration	100%	100%	18.18%	63.64%	30.76%	77.78%
SQL In-jection	0	0	0	100%	0	0
Heart-bleed	100%	100%	100%	100%	100%	100%
宏平均	89.88%	**92.45%**	75.87%	**76.03%**	77.49%	**78.77%**

根据表5的分析，采用MEMFT方法精选出30维特征进行训练后，分类器的整体性能相较于使用原始76维特征显著提升。其中，Bot攻击的Recall从39.83%提升至62.71%；Infiltration攻击的Recall从18.18%大幅提升至63.64%。在30维特征下，11个类别的F1-score达到最高，模型复杂度显著降低。总体性能方面，Macro_F1从77.49%提升至78.77%，Macro_P从89.88%升至92.45%，均优于全特征模型，表明MEMFT在特征精简的同时有效保留了关键判别信息，提升了分类性能与可解释性。

根据表6的分析，在UNSW-NB15数据集上，采用MEMFT精选出11维特征后，分类器性能显著提升，尤其在少数类攻击检测中表现突出。Worms

攻击的Recall从13.46%提升至34.61%，F1-score从21.53%提高至43.90%。与使用43维特征相比，11维特征的Macro_P、Macro_R和Macro_F1分别达到71.76%、57.25%和60.23%，均有所提高。实验表明，MEMFT能够提供高解释性的特征子集，有效保留关键信息，使分类器在更少特征下实现更优性能，兼顾效率与准确性。

表6　基于未平衡UNSW-NB15数据集多分类实验结果

类别	Precision		Recall		F1-score	
	43维	11维	43维	11维	43维	11维
Normal	95.25%	94.97%	95.19%	95.16%	95.22%	95.06%
Fuzzers	74.75%	74.67%	73.75%	72.53%	74.25%	73.59%
Analysis	67.08%	69.07%	13.20%	13.07%	22.06%	21.99%
Backdoor	53.95%	56.80%	10.73%	10.15%	17.90%	17.23%
DoS	35.87%	39.90%	25.31%	24.56%	29.68%	30.40%
Exploits	64.09%	63.89%	82.82%	84.49%	72.26%	72.76%
Generic	99.73%	99.63%	97.89%	97.90%	98.80%	98.76%
Reconnais-sance	92.08%	91.02%	75.64%	76.35%	83.05%	83.04%
Shellcode	66.35%	67.63%	63.08%	63.67%	64.67%	65.59%
Worms	53.84%	60.00%	13.46%	34.61%	21.53%	43.90%
宏平均	70.30%	**71.76%**	55.11%	**57.25%**	57.94%	**60.23%**

根据表7的分析，在5G Threat数据集上，采用MEMFT精选出40维特征后，分类器性能显著提升。NGReset攻击的Recall从28.00%提升至40.30%，F1-score从42.30%提高至56.10%。总体性能方面，Macro_F1从78.45%提升至79.73%，且稀有攻击检测能力明显增强。结果表明，MEMFT在复杂仿真网络环境中仍能有效识别关键特征，有效提升分类性能。

表7　基于未平衡5G Threat数据集多分类实验结果

类别	Precision		Recall		F1-score	
	76维	40维	76维	40维	76维	40维
Normal	99.80%	99.92%	99.82%	99.90%	99.81%	99.91%
DDoS	98.50%	99.10%	97.20%	96.80%	97.84%	97.93%
NGReset	82.00%	94.50%	28.00%	40.30%	42.30%	56.10%
LDA	96.00%	97.50%	88.50%	88.00%	92.10%	92.48%
MMA	94.80%	96.30%	91.20%	90.80%	92.90%	93.42%
NSA	90.20%	95.60%	85.40%	85.60%	87.70%	90.28%
宏平均	91.88%	**94.65%**	71.70%	**73.50%**	78.45%	**79.73%**

3.3.3　MEMFT模型与其他模型解释性能对比

在该实验中，对于每种方法，选出其评估结果中最重要的前N个特征，形成特征子集。基于不同解释方法，生成多个特征子集。使用随机森林分类器分别在这些特征子集上进行训练和测试，比较不同子集的分类性能。评估指标采用Macro_F1，各模型在未平衡CICIDS-2017、UNSW-NB15、NSL-KDD、5G Threat数据集上Macro_F1值对比结果分别如图7~图10所示。根据图7，在CICIDS-2017数据集上，不同解释方法的前N个特征分类性能趋势相似但存在差异。使用前5个特征时，SHAP和PFI的Macro_F1均超68%，解释能力较强；LIME为61%，较弱，表明前者在少量特征下更有效。特征数从25增至30时，SHAP和LIME的Macro_F1下降，可能因新增特征与已有特征高度相关，引发多重共线性。相比之下，MEMFT仅需30个特征即达峰值78.77%，稳定性与效率更高；SHAP和LIME需40个特征才达峰值。

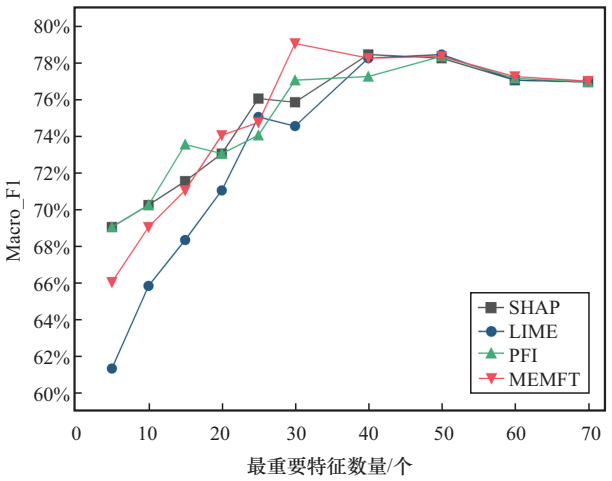


图7　各模型在未平衡CICIDS-2017数据集上Macro_F1值对比结果

根据图8，在UNSW-NB15数据集上，不同解释方法随特征数量变化的性能趋势如下：使用前1个特征时，SHAP和PFI的Macro_F1显著高于其他方法，且均超过30%，表明初始特征区分能力强。当特征数增至5个时，MEMFT超过SHAP、LIME和PFI，开始主导。随着特征增加，MEMFT的重要性优势显现，性能显著提升。仅用前11个特征即达Macro_F1峰值60.23%，明显优于其他方法。结果表明，MEMFT解释更全面准确。

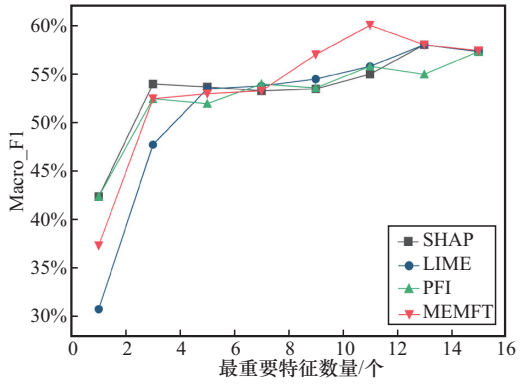


图 8　各模型在未平衡 UNSW-NB15 数据集上 Macro_F1 值对比结果

根据图 9，在 NSL-KDD 数据集上，不同解释方法的性能趋势与前述数据集相似。MEMFT 在 17 个特征达到 Macro_F1 峰值 77.81%，而 SHAP 和 LIME 需更多个特征才接近该性能，表明 MEMFT 在更少特征下即可收敛。

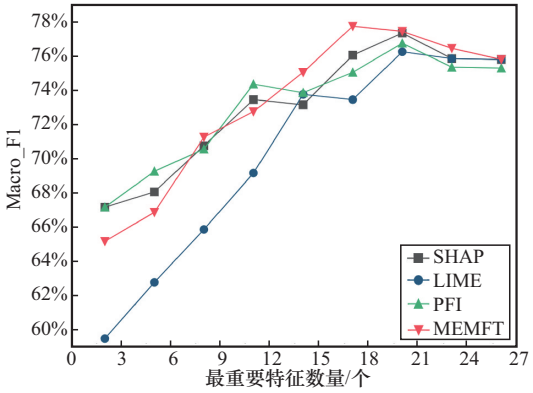


图 9　各模型在未平衡 NSL-KDD 数据集上 Macro_F1 值对比结果

根据图 10，在 5G Threat 数据集上，MEMFT 在 40 个特征时达到 Macro_F1 峰值 79.42%，高于 SHAP 和 LIME 在相同特征数下的表现。结果表明，MEMFT 在不同网络环境下均能高效识别关键特征，提供更具判别性的解释。

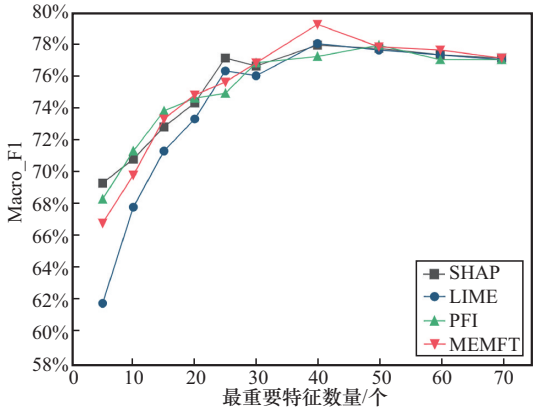


图 10　各模型在未平衡 5G Threat 数据集上 Macro_F1 值对比结果

3.4　资源消耗分析

为评估本文方法的资源消耗，实验从处理效率、内存占用和 CPU 使用率 3 个方面对比了 SHAP、LIME、PFI 及 MEMFT。不同解释方法的计算资源消耗与效率比较如图 11 所示，各方法资源利用存在差异。SHAP 处理效率最优，MEMFT 因需并行执行多算法并融合评估，处理时间较长。在内存占用和 CPU 使用率上，MEMFT 需求较高，主要因需维护多

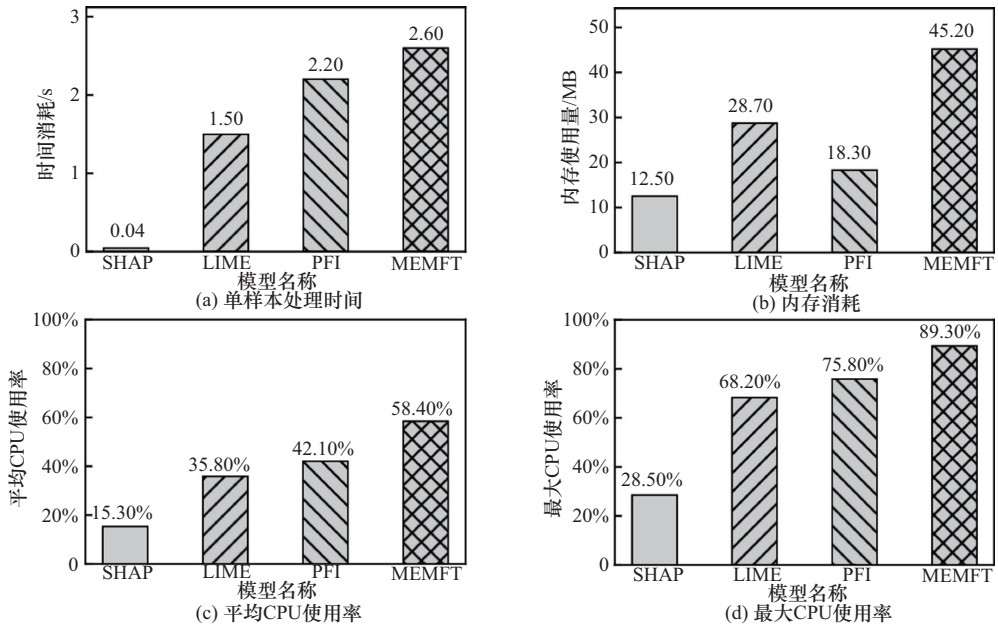


图 11　不同解释方法的计算资源消耗与效率比较

个模型中间结果、指标缓存及动态权重融合的矩阵运算。综合表明，MEMFT在增加一定开销下，显著提升解释质量，通过融合多方法优势，仅用更少特征即达最优性能，提升了可解释性与决策效率，验证了其在解释性能与资源消耗间的合理平衡，适用于高要求的网络安全分析，具备较高的实用价值。

3.5　多解释方法融合技术与数据平衡结合应用效果分析

本文多解释方法融合技术与数据平衡方法通过结合应用实现了性能提升。各模型在平衡后 CICIDS-2017 数据集上 Macro_F1 值结果如图 12 所示，MEMFT 仅需 20 个特征即可使分类器 Macro_F1 达到峰值，而 SHAP 和 PFI 需 30 个特征，LIME 则需更多。与原始数据集相比，数据平衡后各方法达到峰值性能所需的特征数显著减少，表明其有效降低了冗余特征干扰，为解释提供了更稳定的基础。

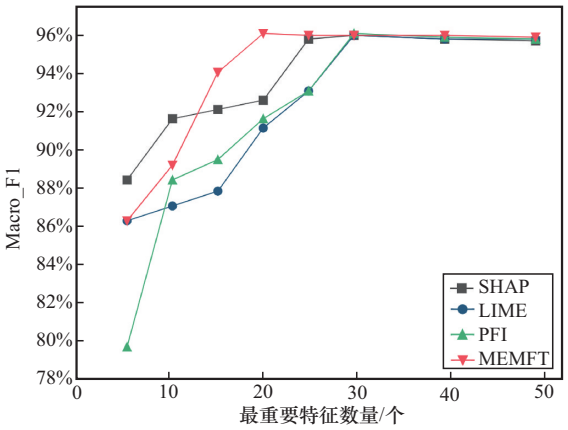


图 12　各模型在平衡后CICIDS-2017数据集上Macro_F1值结果

各模型在平衡后 UNSW-NB15 数据集上 Macro_F1 值结果如图 13 所示，MEMFT 方法同样表现优异，仅需 5 个特征即可达到 Macro_F1 峰值，而 SHAP 和 PFI 需 8 个特征才取得相似性能。与未平衡数据集相比，分类器性能达峰值时，数据平衡后特征需求减少近一半。

各模型在平衡后 NSL-KDD 和 5G Threat 数据集上 Macro_F1 值结果分别如图 14 和图 15 所示，MEMFT 均展现出显著的性能优势。在 NSL-KDD 数据集上，MEMFT 仅需 11 个特征即可使分类器 Macro_F1 达到峰值，而 SHAP 和 PFI 需要 17 个特征，LIME 则需超过 20 个特征。在 5G Threat 数据集上，MEMFT 在 25 个特征时即达到 Macro_F1 峰值，较未平衡状态减少了 15 个特征，而 SHAP 和 PFI 仍需 40 个特征才能收敛。在 2 个数据集中，与未平衡数据相比，各方法在平衡后所需特征数显著减少，表明数据平衡不仅有效压缩了冗余特征空间，还进一步提升了特征解释的效率与稳定性。

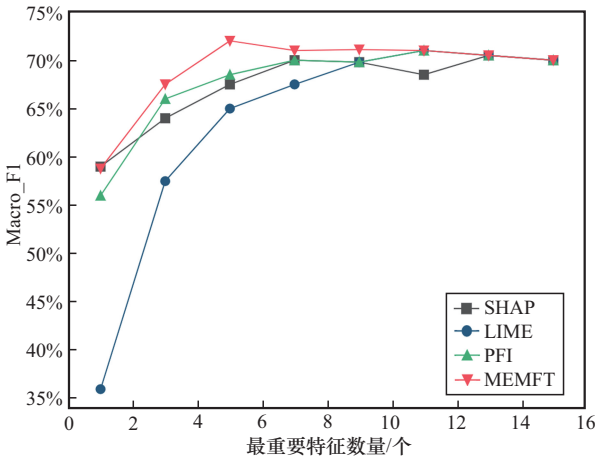


图 13　各模型在平衡后UNSW-NB15数据集上Macro_F1值结果

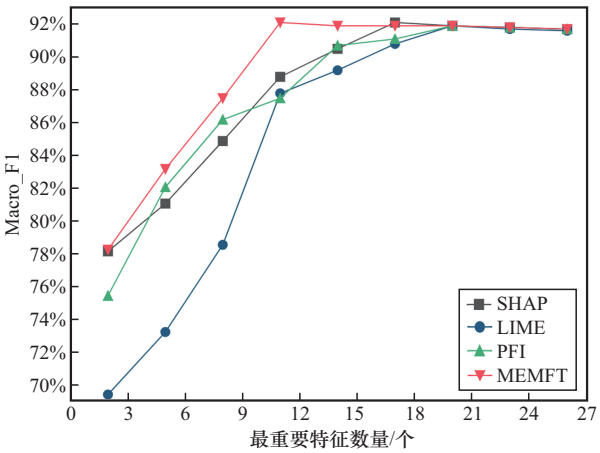


图 14　各模型在平衡后NSL-KDD数据集上Macro_F1值结果

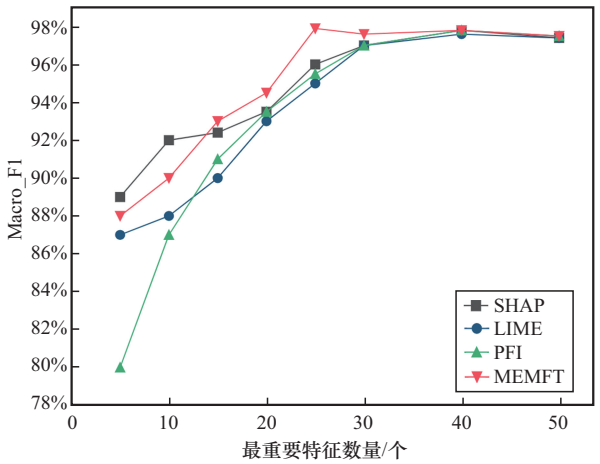


图 15　各模型在平衡后5G Threat数据集上Macro_F1值结果

实验结果表明，数据平衡与解释融合方法的结合使用产生良好效果。一方面，数据平衡为解释方法提供了更均衡的数据基础，有效改善解释方法的特征选择效率，使MEMFT模型以更少的特征达到理想性能；另一方面，解释融合技术通过多方法协同，生成了更稳定的特征重要性评估结果。2种技术的联合应用实现了检测性能与解释效果的同步改善，验证了本文方法的有效性。

4 结束语

本文针对入侵检测系统中现有解释方法给出不一致结果、模型决策缺乏可信度的问题，提出了多解释方法融合技术和数据平衡方法。该方法通过多解释方法融合技术的核心驱动，结合数据平衡方法的辅助支撑，实现了解释可信度与检测性能的协同提升。实验结果表明，MEMFT方法显著提升了解释结果的可靠性和一致性。在4个数据集上，MEMFT在数据平衡后所需特征数显著减少，实现了解释可靠性与检测性能的双重优化，验证了其在多样化网络环境中的高效性。

参考文献：

[1] KHRAISAT A, ALAZAB A, SINGH S, et al. Survey on federated learning for intrusion detection system: concept, architectures, aggregation strategies, challenges, and future directions[J]. ACM Computing Surveys, 2024, 57(1): 1-38.

[2] NIE L S, SUN W T, WANG S P, et al. Intrusion detection in green Internet of Things: a deep deterministic policy gradient-based algorithm[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(2): 778-788.

[3] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature Machine Intelligence, 2019, 1(5): 206-215.

[4] TJOA E, GUAN C T. A survey on explainable artificial intelligence (XAI): toward medical XAI[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(11): 4793-4813.

[5] ZHU T F, LIU X W, ZHU E. Oversampling with reliably expanding minority class regions for imbalanced data learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(6): 6167-6181.

[6] KHAN S H, HAYAT M, BENNAMOUN M, et al. Cost-sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(8): 3573-3587.

[7] QIN Y M, ZHENG H J, YAO J C, et al. Class-balancing diffusion models[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 18434-18443.

[8] SUN J, LI H, FUJITA H, et al. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting[J]. Information Fusion, 2020, 54: 128-144.

[9] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 4768-4777.

[10] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1135-1144.

[11] RONG Y, LEEMANN T, NGUYEN T T, et al. Towards human-centered explainable AI: a survey of user studies for model explanations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(4): 2104-2122.

[12] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics, 2009, 39(2): 539-550.

[13] LIN M L, TANG K, YAO X. Dynamic sampling approach to training neural networks for multiclass imbalance classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(4): 647-660.

[14] DABLAIN D, KRAWCZYK B, CHAWLA N V. DeepSMOTE: fusing deep learning and SMOTE for imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(9): 6390-6404.

[15] ABDI L, HASHEMI S. To combat multi-class imbalanced problems by means of over-sampling techniques[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(1): 238-251.

[16] ANAGHA S A, THOMAS C, BALAKRISHNAN N. Optimized intrusion predictions through feature selection methods[J]. Computers & Security, 2025, 157: 104541.

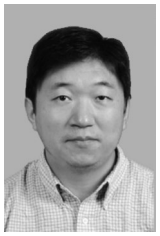
[17] GUO W B, XU J, WANG G, et al. Explaining deep learning based security applications[M]. Berlin: Springer, 2023.

[18] PANDE S, KHAMPARIA A. Explainable deep neural network based analysis on intrusion detection systems[J]. Computer Science, 2023, 24(1): 102983.

[19] SHAKERIN F, GUPTA G. White-box induction from SVM models: explainable AI with logic programming[J]. Theory and Practice of Logic Programming, 2020, 20(5): 656-670.

[20] ZHOU Y, LIANG X M, ZHANG W, et al. VAE-based Deep SVDD for anomaly detection[J]. Neurocomputing, 2021, 453: 131-140.

[作者简介]



熊炫睿（1976– ），男，四川德阳人，博士，重庆邮电大学副教授，主要研究方向为人工智能应用、网络入侵检测、车联网网络安全、5G/6G网络安全。

郭星佑（2000– ），男，四川巴中人，重庆邮电大学硕士生，主要研究方向为5G通信安全。

宁兆龙（1986– ），男，辽宁沈阳人，博士，重庆邮电大学教授，主要研究方向为移动边缘计算、应急网络、机器学习、资源管理。

张玉树（1998– ），男，四川成都人，重庆邮电大学硕士生，主要研究方向为移动物联网与终端技术。

周力（1988– ），男，湖北汉川人，博士，国防科技大学副研究员，主要研究方向为无线网络、软件定义网络、异构网络。