面向深度学习侧信道分析的多属性决策模型评估方法研究

顾泽鹏 陈 琳* 蔡爵嵩 严迎建
(信息工程大学 郑州 450001)

摘 要：本文针对深度学习侧信道分析(DL-SCA)模型评估中存在的维度单一、公平性不足以及与工程场景脱节的问题，提出一种基于系统工程的多维度场景化评估框架。该框架首先构建了覆盖攻击效能、资源开销与环境适应性的层次化评估指标体系；其次，设计了一种CRITIC-AHP混合多属性决策机制，融合数据驱动的客观赋权与场景导向的主观权重，实现评估与不同应用需求的精准匹配；在此基础上，定义多维度攻击性能指标，融合多维度信息，生成直观、可比的综合评分，为模型优选提供统一量化依据。基于ASCAD数据集的实验表明，该框架在资源受限、高性能、高噪声及实时等典型场景下均能有效区分模型优势，如在资源受限场景中CNN综合评分最高(0.723)，在高噪声环境下CNN-LSTM表现最优(0.863)，显著提升了模型选型的科学性与可解释性。

关键词：侧信道分析；深度学习；模型评估；系统工程；多维度场景化评估
中图分类号：TP309 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-11
DOI: 10.11999/JEIT260198 **CSTR**: 32379.14.JEIT260198

1 引言

侧信道分析(Side-Channel Analysis, SCA)通过采集并分析密码设备运行中泄露的功耗、电磁辐射、时序等物理信息来推断其内部密钥，已成为硬件安全的核心威胁之一[1]。传统分析方法，如差分功耗分析(Differential Power Analysis, DPA)[2]和相关功耗分析(Correlation Power Analysis, CPA)[3]，在面对掩码、随机延迟等高级防护措施时攻击效率显著下降[4]。近年来，深度学习技术凭借其强大的自动特征提取能力，为破解防护实现提供了新途径，形成了深度学习侧信道分析(Deep Learning Side-Channel Analysis, DL-SCA)这一前沿分支[5–7]。

然而，深度学习模型具有“黑箱”特性，其性能受多重因素影响。随着DL-SCA研究的深入，如何系统、科学地评估模型性能，已成为制约当前研究从“攻击构建”迈向“工程部署”的主要障碍之一[8]。当前评估工作主要存在以下三方面不足。

评估维度单一，难以全面衡量模型实用价值。多数研究仍仅以猜测熵(Guessing Entropy, GE)[9]和成功率(Success Rate, SR)[10]作为核心评判指标。然而，模型的实用性是攻击效能、资源开销与环境适应性的综合体现，已有研究开始关注并探索面向不同部署上下文(Context)的可调谐评估策略[11]。Zaid等人[12]指出，忽视计算与存储成本的评估，可能导致所选模型无法在资源受限的嵌入式设备上实际部署。Benadjila等人[13]在ASCAD基准研究中也强调，仅凭攻击性能指标无法全面评估模型的工程适用性。此外，真实环境中的噪声干扰使得模型的鲁棒性(Robustness)至关重要[14]。

评估过程公平性缺失，对比基准不统一。深度学习模型的性能对超参数配置极为敏感[15]。Wu等人[16]的研究表明，未经充分超参数优化的模型间比较，其结论会有失公允。不同的超参数选择会导致同一架构的性能产生显著波动[17]。建立一个包含标准化超参数优化环节的公平评估基准，是进行有效比较的前提。

评估与多元化应用场景脱节，缺乏量化指导。现有研究普遍忽视了侧信道分析在不同应用场景下的差异化需求[18]。Kumar等人[19]的研究观察到，不同场景对模型特性的优先级要求截然不同。尽管在多目标安全评估中引入系统化权重分配的方法论已被证明有效[20]，但其在DL-SCA这一特定领域的场景化适配研究仍属空白。

针对上述不足，已有工作尝试同时考量模型的攻击成功率和参数量以权衡效能与开销[21]，或通过系统性的加噪测试来评估其鲁棒性[22]，还有一些研究开始关注评估流程的自动化与公平性[23]。然而，这些工作尚未形成一套系统化、可量化、可追溯的指标融合与场景化权重分配方法论[24]。具体而言，Perin等[21]仅考量攻击精度与参数量的两维权衡，Dubey等[22]聚焦噪声容限的单一维度，Lee与Park[23]主要解决评估流程的自动化与可复现性，Sánchez等[20]提出的MESA框架面向通用安全应用而未针对DL-SCA的特定泄露机制与部署约束进行定制，Wu等[16]虽揭示了超参数未优化导致的不公平比较问题，但未将其固化为评估流程的强制前置环节。鉴于此，本文旨在构建一个基于系统工程思

收稿日期：2026-02-14；改回日期：2026-06-24；网络出版：2026-07-04
*通信作者： 陈琳 chenlin916@163.com

想、面向DL-SCA模型的多维度场景化评估框架。本研究的核心贡献如下：

(1)构建一个覆盖攻击效能、资源开销与环境适应性三大准则的层次化量化评估指标体系，包含六项指标，为全面评估奠定基础。

(2)设计一种CRITIC-AHP混合多属性决策(Multiple Attribute Decision Making, MADM)机制，创新性地融合数据驱动的客观赋权与知识驱动的主观场景化调整，使评估结果与具体应用需求精准匹配。

(3)定义多维度攻击性能指标(Multi-dimensional Attack Performance Metric, MAPM)，通过对归一化指标值的加权融合，为不同场景下的模型优选提供直观、定量的决策依据。

(4)基于ASCAD基准数据集，以多层感知机(Multi-Layer Perceptron, MLP)、卷积神经网络(CNN)及CNN-长短期记忆网络(Long Short-Term Memory, LSTM)混合模型为评估对象，在确保各模型均经独立超参数优化的前提下，系统验证了框架的有效性，并输出了面向典型场景的工程选型指南。

本文其余部分的组织结构如下：第2节介绍系统工程思想及相关理论基础；第3节搭建基于系统工程的评估框架；第4节开展实验与结果分析；第5节总结本文的研究结论。

2 相关理论与方法基础

2.1 深度学习侧信道分析网络模型架构

深度学习模型能够自动从能量迹中提取特征，分析并恢复出密钥。本文选取三种最具代表性的架构：多层感知机(MLP)、卷积神经网络(CNN)以及CNN与长短期记忆网络混合模型(CNN-LSTM)。

(1)多层感知机

如图1所示，MLP由多个全连接层堆叠而成。输入层将整条能量迹展开为一维向量，经隐藏层逐层非线性变换后，由输出层完成密钥分类。该架构结构简单，但未利用能量迹的时序结构信息。

(2)卷积神经网络

如图2所示，通过卷积层在能量迹上滑动卷积核，自动提取局部功耗特征；池化层进行降维，保留主要特征；全连接层完成分类。其局部连接与权重共享特性，使其对能量迹中的对齐误差具有一定鲁棒性。

(3)CNN-LSTM混合模型

如图3所示，首先由卷积层提取能量迹的局部特征，输出特征序列；由LSTM层对序列进行时序依赖建模，捕捉长距离上下文关系；最后经全连接层输出分类结果，兼顾局部特征提取与全局时序建模能力。

2.2 基于系统工程的侧信道模型评估方法

深度学习模型评估面临多维度、多目标、多场景的复杂权衡，传统单一维度的评估方法难以应对。系统工程为此提供了系统化的顶层方法论。

(1)系统思维与层次化分解：将模糊的“模型优选”问题，分解为目标层、准则层、指标层、方案层构成的清晰层次结构，使复杂问题可管理、可分析。

(2)生命周期模型与流程化保证：借鉴“V模

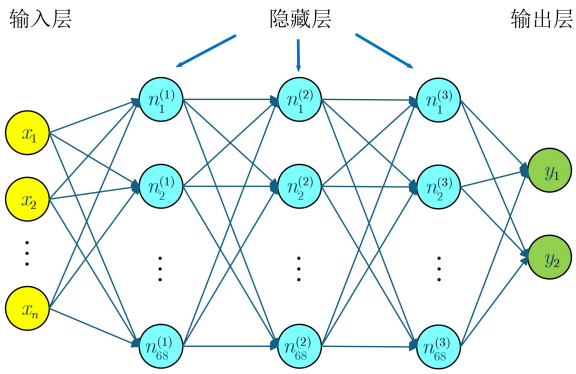


图 1 MLP结构示意图

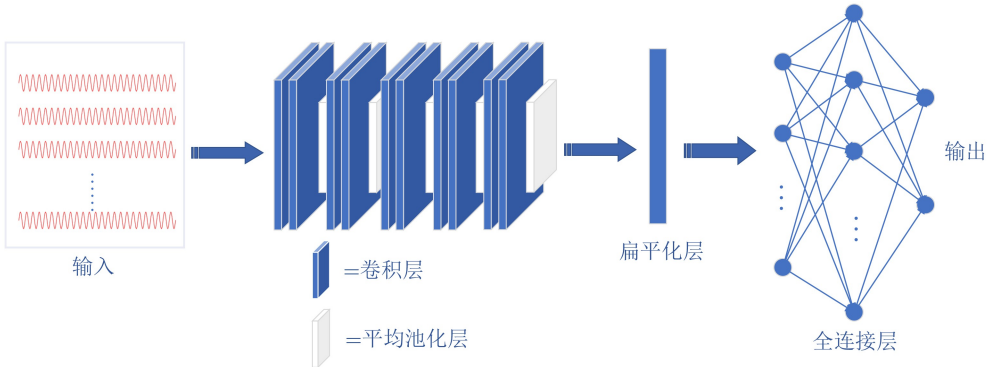


图 2 CNN结构示意图

型”，将评估活动规划为需求定义、系统分解、组件实现、集成验证的标准化流程，确保过程严谨、结果可追溯。

(3)多属性决策分析：研究采用融合客观(CRITIC法)与主观(AHP法)赋权的混合MADM方法，生成兼具客观公平性与场景适配性的综合权重。

(4)需求追溯与场景化适配：强调评估的最终价值在于满足具体应用需求，必须建立从“场景需求”到“评估准则”再到“评估结果”的可量化映射链路。

3　多维度场景化评估框架设计

3.1　框架总体设计

本研究旨在将系统工程思想转化为一个可执行、可量化的DL-SCA模型评估框架。框架以“V模型”为流程骨架，以层次化指标体系为评估维度，并以CRITIC-AHP混合决策引擎为核心，最终输出多维度攻击性能指标(MAPM)，为场景化模型优选提供定量决策依据。本研究框架的层次化结构如图4所示。

框架的设计遵循基于系统工程的四方面原则，据此指导框架的总体设计。

(2)V模型流程。评估遵循V模型的左右分支。左支(定义与分解)涵盖数据准备、模型独立优化与多维数据采集；右支(集成与验证)涵盖数据融合、权重决策、综合评分与结果输出，形成从需求到验证的闭环。

(3)MADM决策引擎。框架的核心是一个CRITIC-AHP混合的MADM引擎，系统化地处理多维度指标的冲突与权衡。

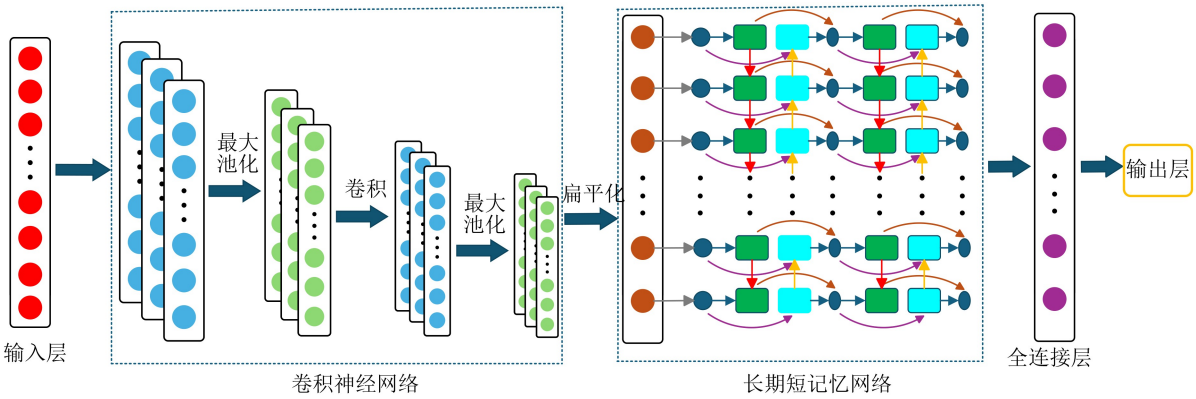


图 3 CNN-LSTM混合结构示意图

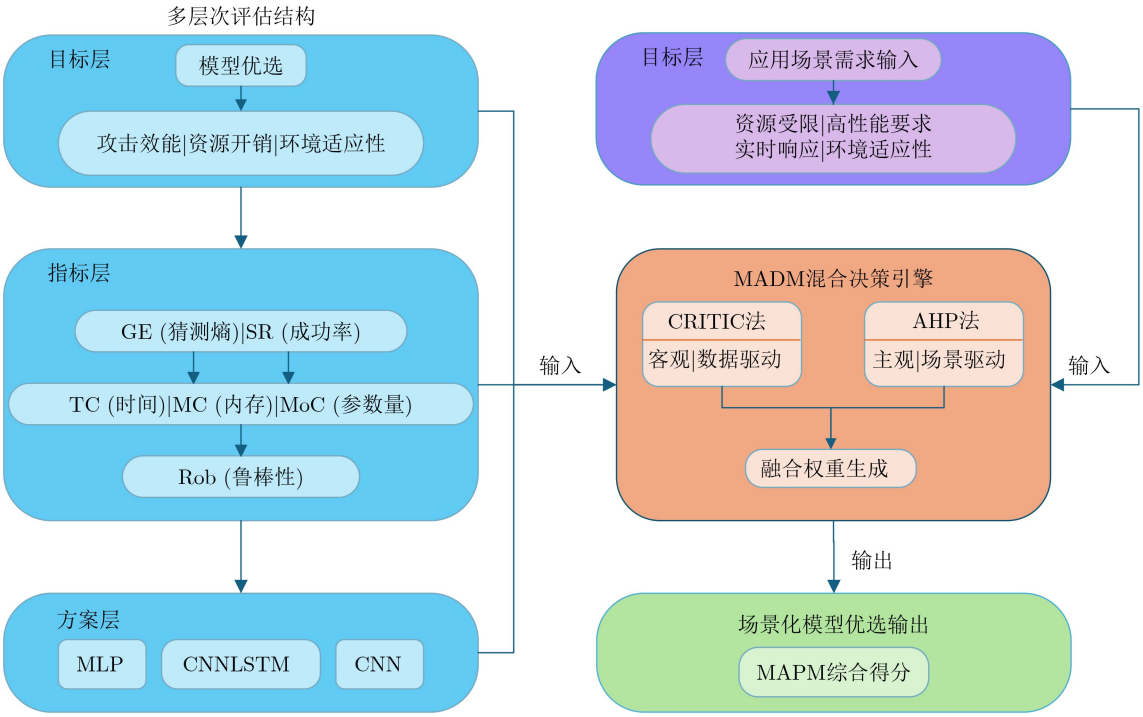


图 4 框架总体设计图

(4)需求追溯。通过“场景-需求-权重”的链路，确保评估结果精准反映预设场景的核心诉求。

上述设计原则与系统工程的核心理念一一对应：层次化分解原则对应评估指标体系的四层结构(目标层—准则层—指标层—方案层)；V模型原则对应标准化评估流程的左右分支(左支定义与分解，右支集成与验证)；MADM决策引擎对应CRITIC-AHP混合赋权机制；需求追溯原则对应“场景—需求—权重”的可量化映射链路。四项原则共同确保了框架的系统性与可追溯性。

3.2 基于V模型的标准化评估流程

框架执行流程标准化为七个步骤，如图5所示，确保评估的系统性与可重复性。

阶段一：定义与分解(V模型左支)

此阶段目标是将总体评估需求分解为可独立执行和验证的子任务。

步骤1：需求分析与输入定义。明确场景，输入预处理后的标准数据集。

步骤2：模型子系统独立优化。框架采用网格搜索方法，对每个候选模型执行独立的超参数优化。

步骤3：多维度性能数据采集。对模型执行三类测试，生成覆盖所有模型和维度的标准化数据矩阵。

①功能性能测试：在测试集上进行攻击，采集攻击效能维度数据(猜测熵、成功率)。

②非功能性能测试：在训练过程中，采集资源开销维度数据(训练时间、峰值内存、参数量)。

③环境适应性测试：在加噪测试集上评估，计算噪声鲁棒性。

阶段二：集成与验证(V模型右支)

此阶段目标是将分解阶段采集的数据进行系统化集成、融合，并验证，形成决策闭环。

步骤4：系统级数据集成。汇集所有模型在六项指标上的数据，构建标准化评估矩阵。

步骤5：多属性决策融合。应用CRITIC-AHP混合权重分配机制，生成场景化权重。

步骤6：综合验证与评分。利用所得权重计算每个模型的MAPM值。

步骤7：决策输出与反馈。输出场景化模型排名与分析结果，完成需求验证闭环。

3.3 层次化评估指标体系构建

依据“层次化分解”原则，将“模型优选”目标分解为表1所示的四层结构。该体系全面覆盖了模型的功能性(攻击效能)、非功能性(资源开销)及环境属性(适应性)，确保了评估的全面性。

3.4 基于CRITIC-AHP的多属性决策融合

为保证评估的客观公平性与场景适配性，设计CRITIC-AHP混合多属性决策(MADM)机制。核心思想是利用CRITIC法从性能数据中提取客观权重作为基准，采用AHP法融入具体应用场景的主观偏好进行校准，通过乘法合成得到匹配场景需求的综合权重，生成指导模型优选的多维度性能指标。

3.4.1 多属性决策问题形式化

将模型评估问题形式化为一个典型的多属性决策问题。定义有m个待评估的深度学习模型构成方案集A；$n=6$个评估指标构成属性集C。原始评估数据矩阵为X。目标是求解一个与场景S对应的权重向量W_s，进而通过线性加权和的方式，将多维

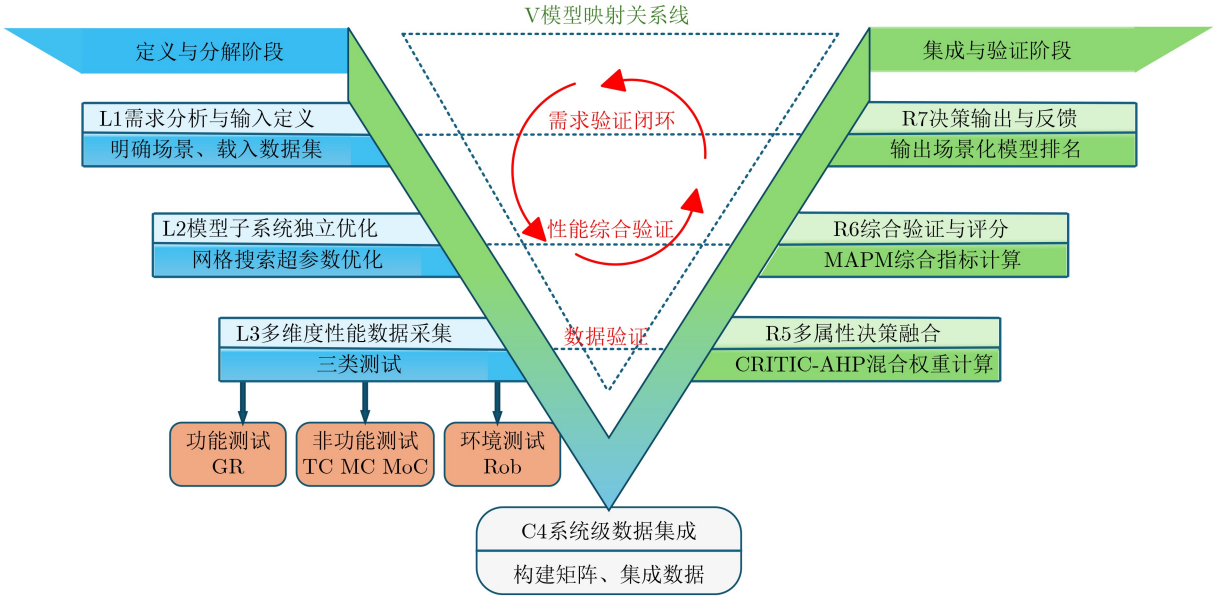


图 5 基于V模型的评估流程闭环图

表 1　基于层次化分解的评估指标体系

层次	名称	要素	说明
目标层	模型优选	选择最优DL-SCA模型	评估的最终目标
准则层	攻击效能	恢复密钥的核心能力	功能性准则
	资源开销	计算与存储成本	非功能性准则
	环境适应性	噪声下的稳定性	环境性准则
指标层	猜测熵 (GE)	负向指标	衡量攻击效率
	成功率 (SR)	正向指标	衡量即时有效性
	时间复杂度 (TC)	负向指标	训练时间成本
	空间复杂度 (MC)	负向指标	峰值内存成本
	模型复杂度 (MoC)	负向指标	参数量(存储成本)
	噪声鲁棒性 (Rob)	正向指标	抗干扰能力
方案层	候选模型	MLP, CNN, CNN-LSTM	待评估的实体

指标数据聚合为单一的综合评分，为决策提供直观依据。

3.4.2 CRITIC客观权重计算

CRITIC法的权重由评估数据矩阵X的统计特性决定。核心原理是：指标的权重应与其所提供的信息量成正比，而信息量取决于该指标取值的对比强度及其与其他指标的冲突性。计算步骤如下：

(1)数据标准化：采用极差法将原始矩阵X归一化，得到：$Z=[Z_{ij}]_{m\times n}$。

对于正向指标(SR、Rob)：$z_{ij}=(x_{ij}-min_j)/(max_j-min_j)$。

对于负向指标(GE、TC、MC、MoC)：$z_{ij}=(max_j-x_{ij})/(max_j-min_j)$。

(2)计算指标信息量：第j个指标的信息量：$I_j=\sigma_j.\sum_{k=1}^{n}1-r_{jk}$，其中$\sigma_j$为标准差(对比强度)，$r_{jk}$为指标间的皮尔逊相关系数。

(3)生成客观权重：对信息量归一化，得到客观权重向量：$W_c=(w_{c1},w_{c2},\cdots,w_{cn})$，$w_{cj}=I_j/\sum_{k=1}^{n}I_k$。

3.4.3 AHP场景化偏好权重计算

采用AHP法将依据场景的判断转化为结构化权重。针对每个预设场景S：

(1)构建判断矩阵：依据场景目标，使用1-9标度法进行两两比较，形成判断矩阵$B_s=(b_{jk})_{n\times n}$。(2)计算场景因子：计算B_s最大特征值对应的特征向量，得到场景因子权重${W_s}^a=({w_{s1}}^a,\cdots,{w_{sn}}^a)$。

3.4.4 权重融合与MAPM评分

为生成基于客观事实与主观导向的最终权重，融合客观基准权重W_c与场景因子${W_s}^a$：

$$W_s=(W_c\circ{W_s}^a)/\sum_{k=1}^{n}(W_{ck}\circ{W_{sk}}^a)\qquad(1)$$

其中$\circ$表示逐元素相乘。W_s即为场景S下的最终评估权重。

多维度攻击性能指标作为V模型验证阶段的输出，对模型A_i在场景S下的综合性能进行量化评分：

$$\mathrm{MAPM}_s(A_i)=\sum_{j=1}^{n}w_{sj}\cdot z_{ij}\qquad(2)$$

MAPM通过系统化的加权融合，将复杂的多维度权衡转化为一个直观可比的标量，从而为特定工程场景下的模型优选提供明确、定量的依据。

4　实验验证与结果分析

为验证评估框架的有效性与科学性，本节设计了系统的对比实验。实验围绕业界标准基准数据集展开，选取三种代表性深度学习模型架构，在确保公平比较的前提下，全面测量其在攻击效能、资源开销及环境适应性三大维度下的表现，并通过多维度攻击性能指标(MAPM)进行场景化综合评估与决策分析。

4.1　实验设置

4.1.1　数据集与预处理

本研究采用ASCAD数据集中的固定密钥版本(ASCADf)。该数据集包含50,000条用于训练和10,000条用于测试的能量轨迹，每条轨迹包含700个时间采样点。攻击目标为恢复AES-128算法第3轮第2个字节的密钥。所有能量迹在输入模型前均进行幅度归一化处理，以加速训练收敛。

4.1.2　实验环境配置

实验的软硬件配置如表2所示。

4.1.3　模型实现与超参数优化

为实现公平比较，本研究选取MLP、CNN及CNN-LSTM三种架构作为评估对象，并为每种架

构执行独立的超参数优化。优化采用网格搜索方法，以确保在预定义的搜索空间内找到各架构的最优配置。优化目标为最小化验证集上的猜测熵。超参数优化得到的最佳配置如表3所示。

4.2 多维度评估结果分析

在获得各模型最优配置后，重新训练并测量所有评估指标。

4.2.1 攻击效能分析

攻击效能直接衡量模型恢复密钥的能力。表4展示了在测试集上，各模型的猜测熵(GE)和成功率(SR)随能量迹数量增加的变化情况。猜测熵收敛曲线如图6(a)所示，成功率变化曲线如图6(b)所示。

攻击效能结果表明，CNN在猜测熵收敛速度与成功率提升方面最好。这源于一维卷积神经网络固有的局部连接与平移不变性特性，使其能高效提取能量迹中的局部时序特征。MLP因将能量迹扁平化处理而初期表现滞后，但随迹数量增加亦能收敛，体现全局拟合能力。CNN-LSTM混合模型收敛所需轨迹数最多，说明其复杂结构在ASCAD固定密钥泄露模式下未展现出预期优势，反而引入了更高的训练成本。

4.2.2 资源开销分析

表5统计了各模型的复杂度与训练成本。图7展示了资源开销的对比图。

资源开销测量结果表明，CNN拥有最大的参

表 2 实验环境配置

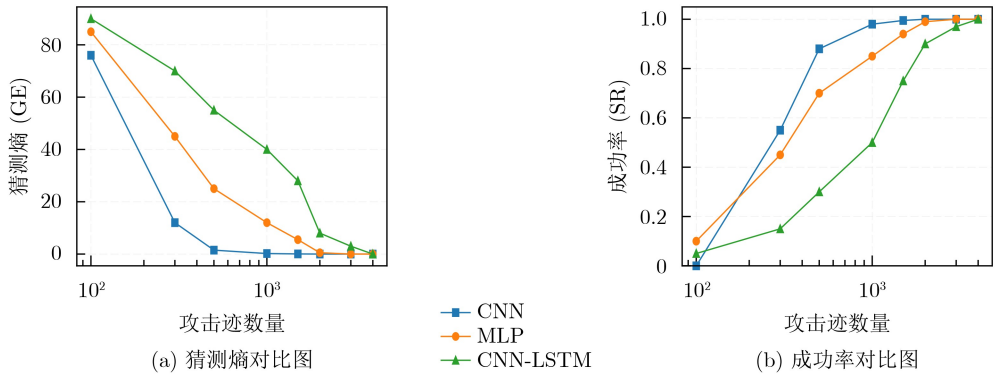
配置项	参数
CPU	Intel Core i7-11700K
GPU	NVIDIA GeForce RTX 4090 (24 GB)
内存	128GB
操作系统	Windows11
深度学习框架	TensorFlow 2.10.0,Keras 2.10.0
CUDA/cuDNN	11.2/8.1.0

表 3 超参寻优网络配置表

Hyperparameters	MLP	CNN	CNNLSTM
FC Layers	5	2	-
Neurons	200	[512,256]	-
Filters	-	[64,128,256,512]	4
Kernel size	-	3	50
Conv layers	-	4	1
Lstm units	-	-	128
Batch size	100	300	200
Activation	ReLU	ReLU	ReLU
Learning rate	1e-5	1e-4	1e-5
Epoch	300	200	500
Optimizer	RMSprop	RMSprop	RMSprop

表 4 攻击效能指标对比(测试集：N=10,000条轨迹)

迹数量	CNN_GE	CNN_SR	MLP_GE	MLP_SR	CNN-LSTM_GE	CNN-LSTM_SR
100	76.0	0.0	85.0	0.0	90.0	0.0
300	12.0	0.55	45.0	0.10	70.0	0.05
500	1.5	0.88	25.0	0.45	55.0	0.15
1000	0.2	0.98	12.0	0.75	40.0	0.30
1500	0.05	0.995	5.5	0.99	28.0	0.50
2000	0.0	1.000	0.3	1.00	8.0	0.90
3000	0.0	1.000	0.0	1.00	3.0	0.97
4000	0.0	1.000	0.0	1.00	0.0	1.00



(a) 猜测熵对比图 (b) 成功率对比图

图 6 攻击效能曲线对比图

表 5　资源开销指标对比

模型	参数量(MoC/万)	训练时间(TC/s)	峰值内存(MC/MB)
MLP	35	568	1950
CNN	425	2850	520
CNN-LSTM	78	1450	3050

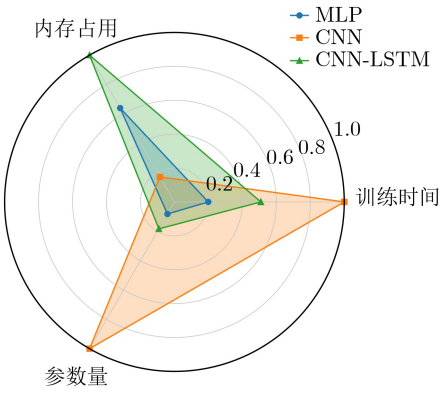


图 7 资源开销对比图

数量与最长的训练时间，这与其复杂的卷积层结构直接相关。然而，其峰值内存占用却最低，这得益于卷积操作的参数共享特性，大幅减少了前向传播中的中间激活存储开销。MLP模型总参数量与训练时间最小，但峰值内存占用最高，这是全连接层矩阵运算的固有特点。CNN-LSTM在各项开销上居中。此结果凸显了单一指标的局限性，实际部署约束需区别考量。

4.2.3　噪声鲁棒性分析

为评估模型在真实环境中的稳定性，向测试集添加高斯噪声模拟非理想条件，并计算鲁棒性指标。噪声水平定义为所添加高斯噪声的标准差σ_noise与原始能量迹信号标准差σ_signal的比值，即噪声水平为$\sigma_noise/\sigma_signal$。0.0表示未添加噪声的原始信号，0.7表示噪声强度为信号标准差的70%。表6统计了随着噪声增加，各模型的成功率和猜测熵改变情况。图8展示了鲁棒性的对比图。

噪声鲁棒性测试结果表明，所有模型性能均随噪声增强而衰减，但衰减幅度差异显著。CNN-LSTM混合模型表现出最强的鲁棒性，其成功率随噪声水平上升下降最为平缓，猜测熵增长最慢。这得益于其结构中LSTM层所具备的时序建模与记忆能力，能有效过滤噪声干扰并保持对关键时序依赖的捕获。CNN对噪声最为敏感，其优异的局部特征提取能力同时放大了噪声影响。此结果证明，环境适应性是与攻击效能、资源开销相互独立且关键的评价维度，而CNN-LSTM在复杂噪声环境下体现出独特的稳定优势。

4.3　场景化综合评估(MAPM)

本节内容结合CRITIC-AHP决策和MAPM指标，为不同场景提供量化选型建议。

(1)客观基础权重计算：基于表4、5、6的全部数据，应用CRITIC法计算得到客观基础权重为：W_critic=[0.17,0.19,0.15,0.21,0.14,0.14](对应GE, SR, TC, MC, MoC, Rob)

(2)场景化权重生成：针对四个预设场景，采用AHP法设定场景偏好。AHP方法要求对判断矩阵进行一致性检验，以验证两两比较过程中不存在逻辑矛盾。一致性比率CR<0.1即认为判断矩阵通过检验，权重分配合理可信。以高噪声场景为例，依据1-9标度法构建六项指标的判断矩阵，如表7所示。在高噪声环境下，噪声鲁棒性(Rob)是衡量模型可用性的首要指标，其重要性显著高于其他指标；猜测熵(GE)与成功率(SR)仍保留一定参考价值；训练时间(TC)、内存占用(MC)与模型复杂度(MoC)在该场景中非核心考量。

一致性检验结果：λ_max=6.153，CI=0.031，RI(n=6)=1.24，CR=0.025<0.1，通过一致性检验，证明该判断矩阵逻辑合理。

将上述AHP场景因子与CRITIC客观权重按式(1)进行乘法融合，同理可得其余三个场景的判断矩阵及融合权重(均通过一致性检验，CR<0.1)。四个场景的最终权重分配如表8所示。

(3)MAPM计算与场景化排名：归一化后的指

表 6　鲁棒性(Rob)指标对比

噪声水平	MLP_SR	MLP_GE	CNN_SR	CNN_GE	CNN-LSTM_SR	CNN-LSTM_GE
0.0	0.992	0.08	0.784	10.25	0.865	6.45
0.1	0.980	0.15	0.770	12.50	0.860	7.20
0.2	0.960	0.30	0.750	15.80	0.855	8.50
0.3	0.930	0.65	0.720	20.10	0.850	10.20
0.4	0.890	1.20	0.680	25.50	0.845	12.80
0.5	0.840	2.10	0.630	32.00	0.840	16.00
0.6	0.780	3.50	0.570	40.20	0.835	20.50
0.7	0.710	5.80	0.500	50.10	0.830	25.80

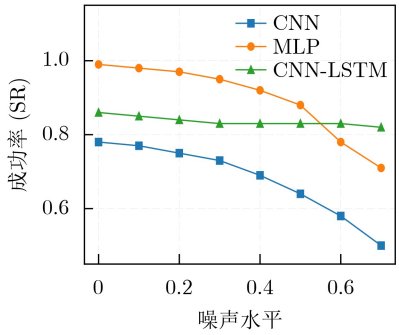


图 8 鲁棒性对比图

标值与表8的权重代入公式计算，结果如表9所示。综合性能对比如图9所示。

表10建立了典型场景与其核心约束、关注指标的映射关系，将场景诉求转化为指标优先级。

需说明的是，当前DL-SCA领域尚缺乏统一的场景定量分类体系。本文对场景的界定重心不在于具体的阈值数值，而在于指标的优先级结构：资源受限场景始终将MC、MoC置于最高优先级，高噪声场景始终将Rob置于最高优先级，高性能场景始终将GE、SR置于最高优先级，实时场景始终将TC置于最高优先级。这一优先级结构跨数据集稳定，在后续的AHP主观赋权和CRITIC权重融合中，将直接决定各场景下的权重分配方向。

基于上述结果，CRITIC-AHP融合权重的MAPM评分结果有力验证了本框架的场景化决策能力。权重的动态分配反映不同场景的核心诉求：在“资源受限”场景中，内存占用(MC)权重主导，使峰值内存最低的CNN排名第一，尽管其总参数量和训练时间最大；在“高性能”场景中，攻击效能(GE, SR)权重占优，CNN同样胜出；在“高噪声”场景中，鲁棒性(Rob)权重至高，使在此维度表现最佳的CNN-LSTM混合模型脱颖而出；在“实时”场景中，训练时间(TC)权重最大，训练最快的MLP位列榜首。这些排名变化直观证明，不存在普适的“最优模型”，而本框架提供的MAPM指标能系统化地量化多维度权衡，为特定工程约束下的模型选型提供清晰的决策依据。

5 结论

本文针对深度学习侧信道分析模型评估中存在的维度单一、公平性不足以及与工程场景脱节的问题，构建了一个基于系统工程的、多维度场景化评估框架。该框架通过“层次化指标体系—CRITIC-

表 7 高噪声场景AHP判断矩阵及一致性检验

指标	GE	SR	TC	MC	MoC	Rob	权重
GE	1	1	3	3	3	1/4	0.15
SR	1	1	3	3	3	1/4	0.15
TC	1/3	1/3	1	1	1	1/5	0.06
MC	1/3	1/3	1	1	1	1/5	0.06
MoC	1/3	1/3	1	1	1	1/5	0.06
Rob	4	4	5	5	5	1	0.52

表 8 各场景最终权重分配

评估指标	资源受限场景	高性能场景	高噪声场景	实时场景
猜测熵(GE)	0.04	0.37	0.06	0.09
成功率(SR)	0.06	0.41	0.04	0.11
训练时间(TC)	0.11	0.05	0.12	0.38
内存占用(MC)	0.52	0.06	0.13	0.29
模型复杂度(MoC)	0.12	0.04	0.08	0.08
鲁棒性(Rob)	0.15	0.07	0.57	0.05

表 9 多维度场景化综合评估(MAPM)结果与排名

模型	资源受限场景	排名	高性能场景	排名	高噪声场景	排名	实时场景	排名
CNN	0.723	1	0.894	1	0.382	3	0.501	2
MLP	0.608	2	0.832	2	0.785	2	0.758	1
CNN-LSTM	0.289	3	0.214	3	0.863	1	0.324	3

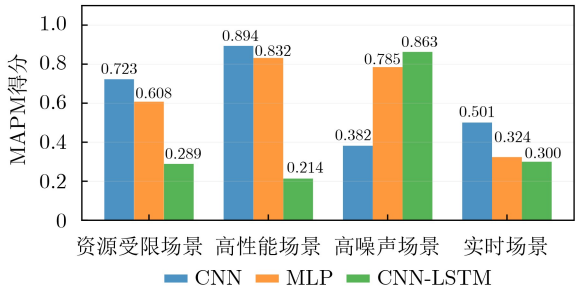


图 9 综合性能对比图

AHP混合决策—V模型标准化流程”的三层融合，实现了模型综合性能的结构化、可量化评估，并建立了基于MAPM的量化选型模型。从设计原理上看，本框架的指标体系源于DL-SCA的通用任务属性，CRITIC客观赋权基于信息量准则、具备根据数据分布自适应调整的能力，AHP场景化赋权基于领域先验知识，三层架构均与具体数据集解耦，用户可根据不同数据集特征设置不同权重，作为多属性决策评估模型的输入，对不同神经网络模型进行评估。

本研究存在一定局限：当前DL-SCA领域尚缺乏统一的场景定量分类标准[25]，本文场景的定量准则基于领域经验设定，未来需推动领域内场景分类共识的形成；部分指标精细化不足。未来工作将聚焦于构建标准化评估基准、推动权重生成的自适应优化。

表 10　场景-指标约束映射表

场景	典型部署环境	核心约束	对应指标	数据来源
资源受限	物联网边缘节点、智能卡、低功耗MCU	存储与算力严格受限	峰值内存(MC)、参数量(MoC)	来自训练过程测量
高性能	GPU服务器、云端计算平台	资源充裕，唯攻击效能论	猜测熵(GE)、成功率(SR)	来自测试集攻击实验
高噪声	工业现场、电磁泄露远距离/非侵入攻击	环境信噪比低，干扰强烈	噪声鲁棒性(Rob)	来自加噪测试
实时攻击	在线攻击系统、支付终端、车载/物联网IDS	训练或推理时延严格受限	训练时间(TC)	来自训练过程测量

面对深度学习侧信道分析在实际部署中复杂的约束条件与多样化需求，本文提出的面向深度学习侧信道分析的多属性决策模型评估方法具有重要的现实意义，通过CRITIC-AHP融合实现主客观权重的平衡，能够有效识别不同模型在攻击效能、资源开销与环境适应性间的差异化优势，为侧信道分析领域提供了可扩展、可复现的评估工具体系。

参 考 文 献

[1] KOCHER P C. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems[C]. Advances in Cryptology - CRYPTO'96, 16th Annual International Cryptology Conference, Santa Barbara, USA, 1996: 104–113. doi: 10.1007/3-540-68697-5_9.

[2] KOCHER P, JAFFE J, and JUN B. Differential power analysis[C]. Advances in Cryptology - CRYPTO'99, 9th Annual International Cryptology Conference, Santa Barbara, USA, 1999: 388–397. doi: 10.1007/3-540-48405-1_25.

[3] BRIER E, CLAVIER C, and OLIVIER F. Correlation power analysis with a leakage model[C]. Cryptographic Hardware and Embedded Systems - CHES 2004, 6th International Workshop, Cambridge, USA, 2004: 16–29. doi: 10.1007/978-3-540-28632-5_2.

[4] ZHANG Fan, DONG Xiaofei, YANG Bolin, *et al*. A systematic evaluation of wavelet-based attack framework on random delay countermeasures[J]. *IEEE Transactions on Information Forensics and Security*, 2020, 15: 1407–1422. doi: 10.1109/TIFS.2019.2941774.

[5] MAGHREBI H, PORTIGLIATTI T, and PROUFF E. Breaking cryptographic implementations using deep learning techniques[C]. Security, Privacy, and Applied Cryptography Engineering, 6th International Conference, SPACE 2016, Hyderabad, India, 2016: 3–26. doi: 10.1007/978-3-319-49445-6_1.

[6] HETTWER B, GEHRER S, and GUNEYSU T. Deep neural network based cryptanalysis of lightweight block ciphers[J]. *IACR Transactions on Symmetric Cryptology*, 2020, 2020(3): 49–78. doi: 10.46586/tosc.v2020.i3.49-78.

[7] BENADJILA R, PROUFF E, STRULLU R, *et al*. Deep learning for side-channel analysis and introduction to ASCAD database[J]. *Journal of Cryptographic Engineering*, 2020, 10(2): 163–188. doi: 10.1007/s13389-019-00220-8.

[8] MARTINAZZI S, ZANKL A, SCHILLING M, *et al*. A systematic review of deep learning for side-channel analysis: Challenges and opportunities[J]. *IEEE Transactions on Information Forensics and Security*, 2024, 19: 1058–1073. doi: 10.1109/TIFS.2023.3321123.

[9] GOHR A. Improving attacks on round-reduced speck32/64 using deep learning[C]. Advances in Cryptology - CRYPTO 2019, 39th Annual International Cryptology Conference, Santa Barbara, USA, 2019: 150–179. doi: 10.1007/978-3-030-26951-7_6.

[10] WOUTERS L, ARRIBAS V, GIERLICHS B, *et al*. Revisiting a methodology for efficient CNN architectures in profiling attacks[J]. *IACR Transactions on Cryptographic*

Hardware and Embedded Systems, 2020, 2020(3): 147–168. doi: 10.13154/tches.v2020.i3.147-168.

[11] PARK D, LEE K, and KIM H. CASTLE: A context-aware strategy for tunable evaluation of deep learning SCA models[J]. *IEEE Transactions on Dependable and Secure Computing*, 2024, 21(3): 2145–2159. doi: 10.1109/TDSC.2023.3258246.

[12] ZAID G, BOSSUET L, HARBRECHT H, *et al*. Towards efficient and scalable side-channel attacks modeling using convolutional neural networks[C]. 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 2020: 133–138. **(查阅网上资料, 未找到本条文献信息, 请确认)**.

[13] BENADJILA R, PROUFF E, STRULLU R, *et al*. Study of deep learning techniques for side-channel analysis and introduction to ASCAD database[J]. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2018, 2018(3): 1–35. doi: 10.46586/tches.v2018.i3.1-35.

[14] WANG Z, LIU Y, SONG N, *et al*. Robustness assessment of deep learning-based side-channel analysis against adversarial trace perturbations[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024, 43(5): 789–802. doi: 10.1109/TCAD.2023.3334567.

[15] RIJSDIJK J, WU Lichao, PERIN G, *et al*. Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis[J]. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021, 2021(3): 677–707. doi: 10.46586/tches.v2021.i3.677-707.

[16] WU L, PERIN G, and PICEK S. The (un)fairness of deep learning-based side-channel analysis: A large-scale benchmarking study[J]. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2024, 2024(2): 1–30. doi: 10.46586/tches.v2024.i2.1-30.

[17] ZHANG J, WANG H, LIU Z, *et al*. Evaluating robustness of deep learning-based side-channel attacks against adversarial traces[C]. 2023 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 2023: 1–8. **(查阅网上资料, 未找到本条文献信息, 请确认)**.

[18] CHEN L, WANG Y, LIU J, *et al*. A lightweight CNN architecture for side-channel analysis on embedded devices[J]. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2024, 71(2): 456–460. doi: 10.1109/TCSII.2023.3329876.

[19] KUMAR A, ZHOU Y, and BHATTACHARYA S. On the trade-offs between model complexity and attack efficiency in deep learning-based SCA[C]. Proceedings of the 2023 ACM Workshop on Attacks and Solutions in Hardware Security (ASHES'23), New York, USA, 2023: 45–52. **(查阅网上资料, 未找到本条文献信息, 请确认)**.

[20] SÁNCHEZ P, ROJAS E, and ROY D B. MESA: A multi-objective evaluation framework for security applications using systematic weighting[J]. *ACM Transactions on Privacy and Security*, 2023, 26(4): 1–30. doi: 10.1145/3592612. **(查阅网上资料,未找到本条文献信息,请确认)**.

[21] PERIN G, WU L, and PICEK S. Exploring the trade-offs: Model accuracy vs. complexity in deep learning SCA[C]. Constructive Side-Channel Analysis and Secure Design – COSADE 2022, Milan, Italy, 2022: 189–209. **(查阅网上资料, 未找到本条文献信息, 请确认)**.

[22] DUBEY A and MUKHOPADHYAY D. Noise tolerance of deep learning based side channel attacks: An experimental study[J]. *Journal of Cryptographic Engineering*, 2023, 13(4): 431–449. doi: 10.1007/s13389-022-00311-x.

[23] LEE K and PARK D. AutoSCA: An automated framework for fair and reproducible side-channel analysis with deep learning[J]. *IEEE Access*, 2024, 12: 45678–45692. doi: 10.1109/ACCESS.2024.3369876.

[24] DUBOIS S, NAJM Z, and DANGER J L. One metric to rule them all? A critical discussion on the evaluation of deep learning-based side-channel attacks[C]. Constructive Side-Channel Analysis and Secure Design – COSADE 2023, Munich, Germany, 2023: 89–110. **(查阅网上资料, 未找到本条文献信息, 请确认)**.

[25] LIU Weifeng, LI Wenchang, CAO Xiaodong, *et al*. Full-element analysis of side-channel leakage dataset on symmetric cryptographic advanced encryption standard[J]. *Symmetry*, 2025, 17(5): 769. doi: 10.3390/sym17050769.

顾泽鹏：男，硕士生，研究方向为侧信道分析与防护.
陈　琳：女，副教授，研究方向为芯片安全.
蔡爵嵩：男，博士生，研究方向为侧信道分析与防护.
严迎建：男，教授，研究方向为芯片安全.

责任编辑：马秀强

A Multi-Dimensional Scenario-Based Evaluation Method for Deep Learning Side-Channel Analysis Using a Multi-Attribute Decision Model

GU Zepeng　CHEN Lin　CAI Juesong　YAN Yingjian

(*College of Cryptography Engineering*, *Information Engineering University*, *Zhengzhou* 450001, *China*)

Abstract:

Objective The application of deep learning has significantly advanced side-channel analysis (DL-SCA), enabling attacks against protected implementations. However, transitioning DL-SCA models from research to practical deployment is hindered by the lack of systematic, fair, and scenario-aware evaluation methodologies. Current evaluations predominantly rely on one-dimensional metrics like Guessing Entropy (GE) and Success Rate (SR), neglecting critical practical dimensions such as resource consumption and environmental robustness. Furthermore, comparisons are often unfair due to inconsistent hyperparameter optimization, and they fail to provide quantifiable guidance for model selection tailored to diverse real-world constraints (e.g., resource-limited devices, high-noise environments, or real-time requirements). This paper aims to address these gaps by proposing a comprehensive, systems engineering-based evaluation framework that enables holistic, quantifiable, and scenario-adaptive assessment of DL-SCA models.

Methods A multi-dimensional, scenario-based evaluation framework is constructed based on systems engineering principles. First, a hierarchical evaluation index system is established, encompassing three criteria (attack efficacy, resource overhead, and environmental adaptability) and six specific metrics (GE, SR, training time TC, peak memory consumption MC, model complexity MoC, and noise robustness Rob). Second, a standardized evaluation process following the "V-model" is designed to ensure fairness. This process mandates independent hyperparameter optimization for each candidate model (MLP, CNN, CNN-LSTM) using grid search before comprehensive multi-dimensional data collection. Third, the core of the framework is a hybrid CRITIC-AHP (Criteria Importance Through Intercriteria Correlation - Analytic Hierarchy Process) Multi-Attribute Decision Making (MADM) engine. The CRITIC method derives objective weights from the statistical characteristics (contrast intensity and conflict) of the measured data matrix. The AHP method incorporates subjective, scenario-specific preferences (e.g., prioritizing low memory or high robustness) through pairwise comparison matrices. These weights are fused to generate final scenario-adapted weights. Finally, a Multi-dimensional Attack Performance Metric (MAPM) is defined as the linear weighted sum of normalized metric values using the fused weights, providing a single, comparable score for each model under a given scenario.

Results and Discussions The framework is rigorously validated using the standard ASCAD fixed-key dataset. After independent optimization, the three model architectures are evaluated across all six metrics. The CRITIC method yields an objective base weight vector: W_critic = [0.17, 0.19, 0.15, 0.21, 0.14, 0.14]. For four predefined scenarios (Resource-Constrained, High-Performance, High-Noise, Real-Time), specific AHP judgments are made and fused with the objective weights to produce the final adapted weights (Table 8). For instance, in the Resource-Constrained scenario, memory consumption (MC) receives the highest weight (0.52), while in the High-Noise scenario, robustness (Rob) is dominant (0.57). The calculated MAPM scores (Table 9, Fig.9, Fig.10) clearly quantify the differentiated advantages of each model and demonstrate the framework's scenario-aware decision capability: CNN achieves the highest score in High-Performance scenarios (0.894), MLP excels in Real-Time scenarios (0.758) due to its lowest training time, and CNN-LSTM performs best in High-Noise scenarios (0.863) owing to its superior robustness, despite its higher resource cost. These results effectively prove that there is no universally "best" model and that the proposed MAPM provides a clear, quantitative basis for model selection under specific engineering constraints.

Conclusions This paper proposes a novel systems engineering-based, multi-dimensional evaluation framework to address the key limitations in current DL-SCA model assessment. By integrating a hierarchical index system, a fair V-model process, and a hybrid CRITIC-AHP MADM engine, the framework successfully quantifies and balances the trade-offs between attack efficacy, resource cost, and environmental adaptability. The experimental results on the ASCAD benchmark demonstrate its practical utility in generating clear, quantifiable, and scenario-aware model selection guidelines. The proposed MAPM offers a direct decision basis for engineers facing diverse deployment contexts, bridging the gap between academic attack construction and practical model deployment in DL-SCA. Future work may involve extending the evaluation to more model architectures and datasets, enhancing the automation of the framework, and validating it in real-world deployment scenarios.

Key words: Side-channel analysis (SCA); Deep learning; Model evaluation; Systems engineering; Multi-dimensional scenario-based evaluation