# 基于自注意力与多模态融合的电力系统攻防协同模型

吴润泽，张普阳，郭昊博，王嘉荣
（华北电力大学电气与电子工程学院，北京市 102206）

**摘　要：**【目的】针对新型电力系统数据驱动算法对抗攻击风险及攻防协同性不足的问题，搭建对抗攻击与防御协同优化理论框架，提升攻击靶向性、防御鲁棒性及复杂攻击特征辨识能力，建立攻防协同进化的闭环优化机制。【方法】在对抗攻击生成模块中，通过自注意力机制量化节点特征贡献度并结合Top-K策略筛选关键节点；利用编解码器与强化学习动态优化扰动策略，经过滤器保留关键节点扰动以提升攻击效率。在对抗防御模型中，采用栈式自编码器提取静态结构特征，卷积神经网络-长短期记忆网络融合时空特征，通过动态权重策略整合多模态特征后，经支持向量机分类器实现攻击样本与正常样本的辨识。【结果】相较于随机节点攻击、快速梯度符号法及投影梯度下降攻击方法，所提攻击方法在维持较高成功率的同时，其全攻击强度范围内的鲁棒性更贴合电力系统对抗攻击实际需求，且扰动可集中于关键节点，由此验证了攻击靶向性优势；防御层面，融合模型性能显著优于单一模型，凸显多模态特征融合对复杂攻击模式的强辨识能力。【结论】攻击侧融合自注意力与强化学习，实现了关键节点的靶向扰动；防御侧采用多模态特征融合，提升了复杂攻击的辨识能力；并通过闭环反馈机制，实现了攻防策略的动态协同进化。

**关键词：**对抗攻击；数据驱动算法；电力信息物理系统；攻击向量注入；攻击防御
**中图分类号：** TM73　　**文献标志码：** A　　**文章编号：**1000-7229(2026)04-0028-11
DOI：10.12204/j.issn.1000-7229.2026.04.003

# Power System Attack-Defense Collaborative Model Based on Self-Attention and Multi-Modal Fusion

WU Runze, ZHANG Puyang, GUO Haobo, WANG Jiarong
(School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China)

**ABSTRACT**: [**Objective**] Aiming at the problems of adversarial attack risks and insufficient offensive-defense coordination of data-driven algorithms in new power systems, a theoretical framework for co-optimization of adversarial attacks and defense is established. This framework aims to enhance attack targeting, defense robustness, and the capability to identify complex attack features, thereby establishing a closed-loop optimization mechanism for offensive-defense co-evolution. [**Methods**] In the adversarial attack generation module, a self-attention mechanism is utilized to quantify node feature contributions, and a Top-K strategy is combined to screen key nodes. An encoder-decoder architecture and reinforcement learning are employed to dynamically optimize perturbation strategies, and a filter retains perturbations on key nodes to improve attack efficiency. In the adversarial defense model, a stacked autoencoder extracts static structural features, while a convolutional neural network-long short-term memory network fuses spatiotemporal features. These multi-modal features are then integrated via a dynamic weighting strategy and fed into a support vector machine classifier to distinguish attack samples from normal samples. [**Results**] Compared with random node attacks, the fast gradient sign method, and projected gradient descent attacks, the proposed attack method maintains a high success rate while demonstrating robustness across the entire attack intensity range that better aligns with the practical requirements of power system adversarial attacks. Furthermore, perturbations can be concentrated on key nodes, verifying the advantage of attack targeting. On the defense side, the fusion model's performance significantly surpasses that of single models, highlighting the strong identification capability of multi-modal feature fusion for complex attack patterns. [**Conclusions**] On the attack side, the integration of self-attention and reinforcement learning achieves targeted perturbation on key nodes. On the defense side, the adoption of multi-modal feature fusion enhances the

**基金项目：**国家重点研发计划项目(2022YFB2402901)

identification capability for complex attacks. Furthermore, a dynamic co-evolution of offensive and defensive strategies is realized through a closed-loop feedback mechanism.

This work is supported by Key Research and Development Program of China (No. 2022YFB2402901).

**KEYWORDS**: adversarial attack; data-driven algorithm; power cyber-physical systems; attack vector injection; attack defense

## 0　引　言

随着新型电力系统建设推进，高渗透率新能源并网的高不确定性、大规模电力电子设备接入的强非线性、海量量测/控制节点的高维状态/动作空间等问题，导致电力系统分析与控制复杂度激增，传统模型驱动方法易出现维度灾难、不连续可微函数不可解等问题[1]。数据科学的发展推动数据驱动算法应用，基于大量数据构建的经验模型以挖掘特征、指导决策，其有效性已在电力系统多领域验证[2-5]。但数据驱动算法因设计特性，可能引入新型安全风险。由于其自身隐含的缺陷，可能暴露出缺乏鲁棒性、过拟合和泛化性能低等问题[6-7]。

对于数据驱动算法，其漏洞不仅可能产生于策略的设计过程中，在训练过程中也可能因为训练样本集的质量问题，导致策略出现漏洞[8-9]，且该漏洞更加隐蔽，攻击行为将更难被发现[10-11]。除此之外，已有数据驱动算法的模型和参数通常缺乏可解释性，并且训练样本数量通常无法达到理想值。由此导致训练出的模型准确率虽然在常规场景中能够满足要求，但通常无法对其分类边界处的样本进行有效判别[12]。攻击者可通过构造对抗样本，在边界附近的数据上做轻微扰动，即可诱导数据驱动算法输出错误结果。

在对抗攻击研究领域，文献[13]中给出了一个通过快速梯度符号法生成对抗样本的算法，利用在梯度上加入增量的行为造成模型对样本进行误分类。文献[14]提出了基础迭代法(basic iterative methods，BIM)攻击，将快速梯度符号法(fast gradient sign method，FGSM)攻击中的大扰动分解为多步的小扰动，并且每次小扰动都需要重新计算梯度来设置扰动方向。文献[15]提出的投影梯度下降(projected gradient descent，PGD)攻击对于多次迭代的梯度攻击设计了一个更加规范化的概念。文献[16]提出了一种基于生成对抗网络(generative adversarial network，GAN)的对抗攻击生成方法，通过识别电力系统中的薄弱节点，利用GAN的生成能力构造具有针对性的对抗样本。

在防御策略方面，文献[17]针对电力信息物理系统(cyber-physical system，CPS)中的虚假数据注入攻击(false data injection attack，FDIA)检测算法，提出了一种低成本对抗性FDIA方案，并通过对抗训练进行防御。文献[18]提出一种融合卷积神经网络的长短时记忆网络检测模型，根据实验结果可知，该模型相较其他单一入侵检测模型，可有效识别异常数据。文献[19]使用自动编码器作为自学习(self-taught learning，STL)模型，能够准确地对攻击类型进行分类。文献[20]提出的融合残差密集块自注意力机制与生成对抗网络的防御模型，通过PGD攻击生成对抗样本扩充训练集，有效增强了对图像对抗攻击的防御能力。

当前电力系统对抗安全研究热点集中于攻击侧基于网络拓扑的高效对抗样本生成与防御侧多模态特征融合的检测模型设计，现有研究存在攻击策略因未挖掘节点耦合关联而扰动分配盲目低效、防御模型单一特征提取难以协同表征数据静态结构与动态时序特征、攻防研究割裂缺乏动态协同机制等挑战。

本文旨在构建电力系统对抗攻击与防御协同优化框架，提升攻击靶向性与鲁棒性、增强防御对复杂攻击的特征辨识能力，构建攻防协同进化的闭环优化机制。与现有研究相比，本文创新性主要体现在三个方面：在攻击模型中，引入自注意力机制与强化学习，实现关键节点的精准筛选与靶向扰动生成，克服了传统方法的盲目性问题；在防御模型中，融合栈式自编码器(stacked autoencoder，SAE)提取的静态特征与卷积神经网络-长短期记忆网络(convolutional neural network-long short-term memory，CNN-LSTM)提取的时空动态特征，通过多模态互补提升了对复杂攻击的辨识能力；在框架层面，构建了“攻击—防御—反馈”闭环协同进化机制，突破了传统攻防割裂的研究模式，更贴合电力系统安全的动态博弈特性。最终通过新英格兰39节点系统仿真实验验证方法在攻击成功率、检测率与鲁棒性等方面的性能提升。

## 1　数据与目标设定

### 1.1　攻击场景与数据体系搭建

基于新英格兰39节点系统构造攻击算例，包含39条母线、19个负荷和10台发电机，负荷、发电机、线路、变压器参数根据文献[21]提供的数据进行

设置。

在10个发电机上分别安装1个测量装置，分别对每一台发电机的电压、功角、有功和无功进行测量。电网运行初期负载波动范围为80%~120%，且任何一条线路上都可能出现三相短路故障，且远端故障的切除时间比近端多0.01 s。采用蒙特卡罗方法对各种操作模式及故障情景进行仿真，产生1500个样本，将其随机分为训练与测试两个部分。

极限故障切除时间（critical clearing time，CCT）是指当电网出现故障时，允许故障存在的最大时限，以确保系统的瞬态稳定。新英格兰39节点系统的CCT标准数据集参照文献[22]，包含故障后的发电机量测数据以及利用简化的综合扩展等面积法则计算得到的CCT值。

$$\boldsymbol{X}=\begin{bmatrix} x_{1,1} & \cdots & x_{1,P} \\ \vdots & x_{i,j} & \vdots \\ x_{B,1} & \cdots & x_{B,P} \end{bmatrix} \tag{1}$$

$$X_{\text{norm}}=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

$$T_{\text{norm}}=\frac{T_{\text{CCT}}-T_{\min}}{T_{\max}-T_{\min}} \tag{3}$$

式中：$\boldsymbol{X}$为电力系统的原始量测数据矩阵；$x_{i,j}$为第$i$个节点的第$j$个特征；$B$为量测节点总数；$P$为特征总数；$X_{\text{norm}}$和$T_{\text{norm}}$分别为归一化后的量测数据和临界切除时间；$T_{\text{CCT}}$为原始临界切除时间值；$T_{\min}$和$T_{\max}$分别为CCT的最小值和最大值；$x_{\max}$和$x_{\min}$分别为量测数据的最大值和最小值。

### 1.2 目标建模

攻击目标借助训练极限学习机（extreme learning machine，ELM）开展CCT预测，将故障后各发电机量测数据作为输入，输出为预测结果。构建含283个隐藏层节点的ELM，以训练集训练模型，再用测试集测试CCT预测准确性，验证ELM实现CCT准确预测。CCT正常范围为200~800 ms。针对该模型设定对抗攻击目标为使CCT预测值偏离真实值至少40 ms，以验证攻击方法对数据驱动算法的威胁性。

$$\boldsymbol{H}=\text{sigmoid}\left(\boldsymbol{X}_{\text{norm}}\cdot\boldsymbol{W}+\boldsymbol{b}\right) \tag{4}$$

$$\boldsymbol{\beta}=\left(\boldsymbol{H}^{\text{T}}\cdot\boldsymbol{H}\right)^{-1}\cdot\boldsymbol{H}^{\text{T}}\cdot\boldsymbol{T}_{\text{norm}} \tag{5}$$

$$\boldsymbol{Y}=\boldsymbol{H}\cdot\boldsymbol{\beta} \tag{6}$$

$$\boldsymbol{Y}_{\text{original}}=\boldsymbol{Y}\cdot\left(Y_{\max}-Y_{\min}\right)+Y_{\min} \tag{7}$$

式中：$\boldsymbol{H}$为ELM模型的隐藏层输出特征；$\boldsymbol{W}$为输入层到隐藏层的权重矩阵；$\boldsymbol{b}$为隐藏层的偏置向量；$\boldsymbol{\beta}$为输出层权重矩阵；$\boldsymbol{H}^{\text{T}}$为转置矩阵，$\left(\boldsymbol{H}^{\text{T}}\cdot\boldsymbol{H}\right)^{-1}$表示求逆矩阵；$\boldsymbol{Y}$为模型的预测输出矩阵；$\boldsymbol{Y}_{\text{original}}$为反归一化后的最终预测输出矩阵；$Y_{\max}$和$Y_{\min}$分别为归一化过程中的最大值和最小值。

$$J\left(\boldsymbol{X},\boldsymbol{Y}_{\text{true}}\right)=\left(\boldsymbol{Y}_{\text{pred}}-\boldsymbol{Y}_{\text{true}}\right)^{2} \tag{8}$$

$$\nabla_{\boldsymbol{X}}J=\left(\frac{\partial J}{\partial\boldsymbol{Y}_{\text{pred}}}\right)\cdot\left(\frac{\partial\boldsymbol{Y}_{\text{pred}}}{\partial\boldsymbol{H}}\right)\cdot\left(\frac{\partial\boldsymbol{H}}{\partial\boldsymbol{X}}\right) \tag{9}$$

$$\frac{\partial J}{\partial\boldsymbol{Y}_{\text{pred}}}=2\left(\boldsymbol{Y}_{\text{pred}}-\boldsymbol{Y}_{\text{true}}\right) \tag{10}$$

$$\frac{\partial\boldsymbol{Y}_{\text{pred}}}{\partial\boldsymbol{H}}=\boldsymbol{\beta}^{\text{T}} \tag{11}$$

$$\frac{\partial\boldsymbol{H}}{\partial\boldsymbol{X}}=g'\left(\boldsymbol{X}\cdot\boldsymbol{W}+\boldsymbol{b}\right)\odot\boldsymbol{W} \tag{12}$$

式中：$J$为损失函数；$\boldsymbol{Y}_{\text{pred}}$和$\boldsymbol{Y}_{\text{true}}$分别为CCT的预测向量和真实向量；$\nabla_{\boldsymbol{X}}J$为损失函数关于$\boldsymbol{X}$的梯度；$g'$为sigmoid函数的导数。

## 2 电力CPS对抗攻防模型

### 2.1 框架总体设计与技术目标

攻防协同模型基于马尔可夫决策过程与特征融合理论构建，如图1所示，其包含三大核心模块：攻击生成模块基于强化学习生成对抗样本，通过自注意力机制筛选关键节点并结合近端策略优化（proximal policy optimization，PPO）算法优化扰动策略；防御检测模块通过多模态特征融合技术，利用SAE和CNN-LSTM分别提取静动态特征，经支持向量机（support vector machine，SVM）分类实现攻击检测；数据反馈模块构建“效果评估—防御响应—策略更新”闭环机制，推动攻防策略动态协同进化以适应复杂对抗环境。

技术目标包括：攻击端在满足电力系统物理约束下最大化CCT预测偏差并最小化扰动成本；防御端提升复杂攻击检测准确率，降低漏检误报率；通过闭环反馈机制实现攻防策略的动态协同进化，使攻击策略适应防御机制迭代，推动防御模型持续学习新型攻击特征。

### 2.2 基于强化学习的对抗攻击生成模型

#### 2.2.1 马尔可夫决策过程建模

本文设计了一种基于PPO的强化学习模型[23]，通过智能选择关键节点和生成最优扰动向量来最大化攻击效果。

状态空间$\boldsymbol{S}$包含原始量测数据$\boldsymbol{X}_{\text{norm}}$、历史选择的攻击节点$\boldsymbol{n}_{\text{prev}}$和生成的扰动矩阵$\boldsymbol{r}'_{\text{prev}}$：

$$\boldsymbol{s}_t=\left\{\boldsymbol{X}_{\text{norm}},\boldsymbol{n}_{\text{prev}},\boldsymbol{r}'_{\text{prev}}\right\}\in\boldsymbol{S} \tag{13}$$

式中：$\boldsymbol{s}_t$代表$t$时刻的状态空间。

动作空间$\boldsymbol{A}$包含两个子动作：

$$\boldsymbol{a}_t=\left(\boldsymbol{a}_{\text{select}},\boldsymbol{a}_{\text{perturb}}\right)\in\boldsymbol{A} \tag{14}$$

式中：$\boldsymbol{a}_t$代表$t$时刻的动作空间；$\boldsymbol{a}_{\text{select}}=\left\{n_1,n_2,\cdots,n_k\right\}$，$\left|\boldsymbol{a}_{\text{select}}\right|=k$表示从$B$个节点中选择$k$个

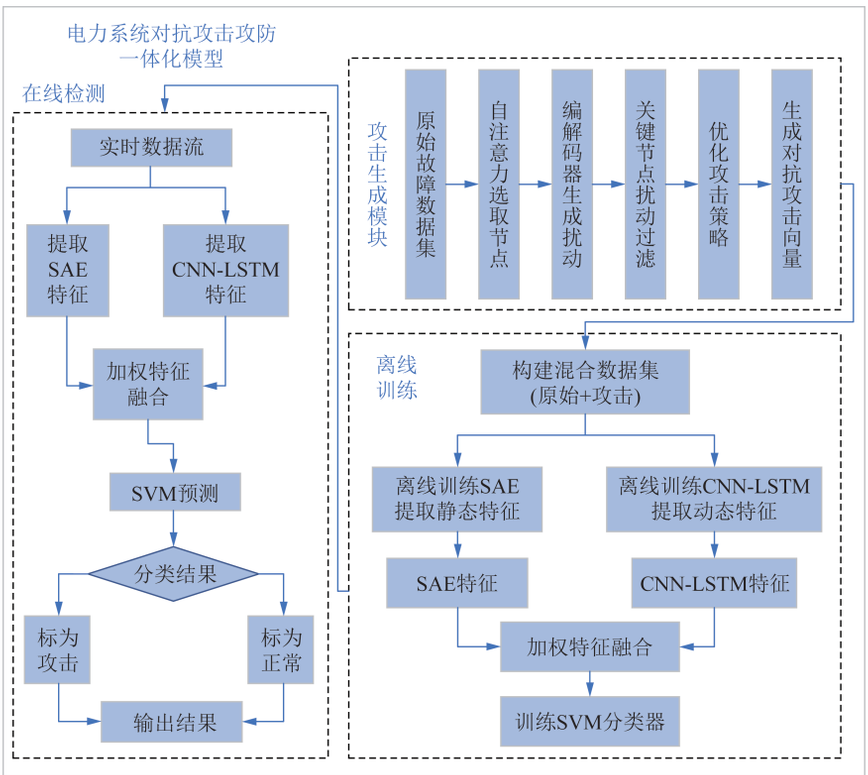


**图 1　对抗攻击攻防一体化模型**

**Fig. 1　Integrated adversarial attack-defense model**

关键节点；$\boldsymbol{a}_{\text{perturb}} \in \mathbb{R}^{B\times P}$，$\left\|\boldsymbol{a}_{\text{perturb}}\right\|_{\infty} \leqslant \varepsilon$ 表示生成扰动矩阵 $\boldsymbol{r}' \in \mathbb{R}^{B\times P}$。

奖励函数通过多目标加权量化攻击效果、隐蔽性和资源利用率：

$$r_t = R\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) = \sum_{i=1}^{4} \omega_i \cdot R_i \tag{15}$$

式中：$\omega_i$ 表示第 $i$ 个奖励分量的权重。

攻击有效性奖励 $R_1$ 通过攻击前后 CCT 预测值的绝对偏差衡量，攻击对系统决策的影响程度。扰动隐蔽性奖励 $R_2$ 基于对抗样本与原始数据平均绝对差异设计，鼓励低扰动攻击以规避检测。特征平衡性奖励 $R_3$ 通过特征扰动的标准差与均值之比构建，抑制单一特征过扰动，提升攻击的多维一致性。$R_4$ 为防御检测到攻击时的负奖励。

$$R_1 = \frac{\left(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\right)^2}{K} \tag{16}$$

$$R_2 = 1 - \left(\frac{1}{B} \times P\right) \sum_{i=1}^{B} \sum_{j=1}^{P} \left| x_{i,j}^{\text{adv}} - x_{i,j}^{\text{orig}} \right| \tag{17}$$

$$R_3 = 1 - \left[\frac{\text{std}\left(\Delta \boldsymbol{x}_t\right)}{\text{mean}\left(\Delta \boldsymbol{x}_t\right)} + \varepsilon\right] \tag{18}$$

$$R_4 = -\left(\hat{y} = 1\right) \tag{19}$$

式中：$\boldsymbol{Y}$ 和 $\overline{\boldsymbol{Y}}$ 分别表示攻击前后 CCT 预测值 $K$ 归一化常数，确保奖励值在合理范围；$x_{i,j}^{\text{adv}}$ 和 $x_{i,j}^{\text{orig}}$ 为攻击前后的量测值；$\text{std}\left(\Delta \boldsymbol{x}_t\right)$ 和 $\text{mean}\left(\Delta \boldsymbol{x}_t\right)$ 分别为各节点在第 $i$ 个特征上的平均扰动量的标准差和平均值；$\varepsilon$ 为极小量；$\hat{y} = 1$ 表示被检测到。

### 2.2.2　自注意力关键节点筛选

通过自注意力机制量化节点特征贡献度，生成节点重要性评分，结合 Top-K 策略筛选关键节点以提升攻击效率。

1）特征关联评估。利用可学习矩阵 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，计算节点 $i$ 与 $j$ 在特征 $p$ 上的注意力权重：

$$a_{i,j,p} = \frac{\exp\left[\text{sim}\left(\boldsymbol{Q}_i, \boldsymbol{K}_j\right)/\sqrt{d}\right]}{\sum_{k=1}^{B} \exp\left[\text{sim}\left(\boldsymbol{Q}_i, \text{K}_k\right)/\sqrt{d}\right]} \tag{20}$$

式中：$d$ 是一个模型超参数，称为注意力维度；$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$ 是查询矩阵；$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$ 是键矩阵；$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$ 是值矩阵；sim(·) 为相似度函数。$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$ 是可学习参数矩阵[24]，分别将输入节点特征向量转换为查询、键、值向量，维度均为 $\left(p, d\right)$。

2）重要性评分生成。对节点 $i$ 的多特征注意力权重取绝对值累加，得到重要性评分：

$$s_i = \sum_{p=1}^{P} \sum_{j=1}^{B} \left| a_{i,j,p} \right| \tag{21}$$

式中：$s_i$ 反映节点与全局网络的关联及对任务的贡献，值越大表明节点 $i$ 对攻击目标的影响越显著。

3）靶向节点筛选。采用 Top-K 策略选取评分最高的 $k$ 个节点，生成二值掩码：

$$\boldsymbol{n} \in \{0,1\}^B,\ n_i = \begin{cases} 1, & i \in \operatorname{argmax}_{i=1}^{B}\{s_i\} \\ 0, & \text{其他} \end{cases} \tag{22}$$

式中：$\boldsymbol{n} \in \{0,1\}^B$为二值数组，表示选中的攻击节点。

确保扰动集中于高贡献节点，提升攻击效率与隐蔽性。

### 2.2.3 扰动生成

编码器负责提取输入数据中的系统特征及节点关联关系，以量测数据$\boldsymbol{X}$为输入，输出对应特征向量。编码器基于深度卷积神经网络架构，用3×3卷积核（64通道），经ReLU激活、2×2最大池化，加高斯噪声四层处理，将原始电力系统量测数据映射至低维特征空间。

解码器以编码器输出的特征向量为输入，其在攻击向量生成模型中承担扰动生成器的角色，核心任务是将抽象的特征向量转换为具体、可作用于原始数据的扰动。采用反卷积网络结构，通过3×3转置卷积核（64通道）、tanh激活层、注入层和维度适配层4层，将低维特征重构为与原始数据同维度的扰动数据$\boldsymbol{r}'$。生成扰动向量后，对特征维度及系统全局指标施加硬编码的物理约束，确保输出的攻击向量符合电力系统运行规则。量测幅值约束表示为：

$$\Delta \boldsymbol{X}_{i,p} = \operatorname{clip}\left(\Delta \tilde{\boldsymbol{X}}_{i,p}, \Delta X_p^{\min}, \Delta X_p^{\max}\right) \tag{23}$$

式中：clip(·)表示对扰动向量进行逐元素裁剪，确保扰动后量测值不越界；$\Delta \tilde{\boldsymbol{X}}_{i,p}$表示第$i$个量测、第$p$类特征最终的扰动值；$\Delta X_p^{\min}$和$\Delta X_p^{\max}$分别表示第$p$类特征扰动最小值和最大值。

局部功率平衡约束为基于电气距离的高斯滤波平滑有功/无功扰动，近似满足基尔霍夫定律。

$$\Delta \boldsymbol{P}_i^{\text{filtered}} = \left[\sum_{j \in N(i)} \Delta \tilde{\boldsymbol{P}}_j \cdot \exp\frac{-d_{ij}^2}{2\sigma^2}\right] \cdot \left[\sum_{j \in N(i)} \exp\frac{-d_{ij}^2}{2\sigma^2}\right]^{-1} \tag{24}$$

式中：$\Delta \boldsymbol{P}_i^{\text{filtered}}$为节点$j$滤波后的有功功率扰动；$N(i)$表示节点$i$的电气邻居集合；$\Delta \tilde{\boldsymbol{P}}_j$为节点$j$的原始有功功率扰动；$d_{ij}$为节点$i$与$j$之间的电气距离；$\sigma$为平滑系数。

全局能量约束为限制扰动向量的L2范数，防止运行点过度偏离。

$$\|\Delta \boldsymbol{X}\|_2 \leqslant \varepsilon_{\max} \tag{25}$$

式中：$\varepsilon_{\max}$为最大允许的全局扰动能量阈值。

依据自注意力机制的选择结果，过滤解码器输出的原始扰动$\boldsymbol{r}'$，仅保留被选中节点的扰动数据，最终得到满足攻击资源约束的扰动$\boldsymbol{r}$，将其叠加至原始量测数据，获得攻击向量。过滤器的过滤过程可表示为：

$$\boldsymbol{r} = \boldsymbol{r}' \odot \operatorname{diag}(\boldsymbol{n}) \tag{26}$$

$$\overline{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{r} \tag{27}$$

### 2.2.4 基于PPO的对抗攻击策略自适应优化

策略网络以系统量测数据、历史攻击节点掩码及扰动信息为输入，通过自注意力机制筛选关键节点，经编解码器提取抽象特征并生成符合物理约束的扰动参数，再通过掩码过滤聚焦高贡献节点，形成“状态感知—节点筛选—扰动生成”的端到端决策流程以平衡攻击效率与效果。同时，价值网络以相同状态向量为输入，通过时序差分学习输出状态价值估计以量化长期奖励期望，为策略网络提供优势函数基准信号，通过构建目标函数优化参数，形成“价值评估—反馈优化”闭环以提升攻击策略的稳定性与环境适应性，如表1所示。其中，$\rho_t(\theta)$表示在新策略参数$\theta$下选择动作$\boldsymbol{a}_t$的概率与在旧策略参数下$\theta_{\text{old}}$选择同一动作的概率的比值。$L^{\text{CLIP}}(\theta)$裁剪损失函数，目标是在鼓励策略改进的同时，通过裁剪策略比率来限制更新幅度，避免训练不稳定。添加熵正则项$L^{\text{entropy}}$可以鼓励智能体探索环境，防止策略过早收敛到局部最优。$L^{\text{value}}$用于学习状态值函数，最小化这个损失可以使价值网络更准确地估计状态值。

**表1 基于PPO的对抗攻击策略自适应优化**

**Table 1 Adaptive optimization of adversarial attack strategy based on PPO**

输入：初始策略网络$\pi_\theta$，价值网络$V_\phi$，经验缓冲区容量$N_{\text{buf}}$，更新次数$K$，截断阈值$\varepsilon$。
输出：优化后的策略网络$\pi_\theta$。
while 未达到训练终止条件：
1. 数据采集阶段：
for $t$ in 1 to $T$：
根据$\pi_\theta$采样动作$\boldsymbol{a}_t$，执行获取奖励$r_t$与下一状态$\boldsymbol{s}_{t+1}$，将$\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$存入经验缓冲区。
if 缓冲区数据量≥$N_{\text{buf}}$：break
2. 策略与价值网络更新阶段：
for _ in 1 to $K$：
从缓冲区随机采样批量数据$\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$
计算优势函数$A_t$与目标价值$V_{\text{target}}$
计算策略比率$\rho_t(\theta) = \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)/\pi_{\theta_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$
计算策略损失$L^{\text{policy}} = -L^{\text{CLIP}}(\theta) + L^{\text{entropy}}$
计算价值损失$L^{\text{value}} = \text{MSE}(V_\phi(\boldsymbol{s}_t), V_{\text{target}})$
更新参数$\theta = \text{Adam}(L^{\text{policy}})$、$\phi = \text{Adam}(L^{\text{value}})$
$\pi_{\theta_{\text{old}}} = \pi_\theta$
end while

截断阈值$\varepsilon = 0.2$以限制策略更新幅度，经验缓冲区容量$N_{\text{buf}}$=10 000存储轨迹数据，每轮更新$K = 10$次提升参数优化稳定性，学习率设为0.001，折扣因子$\gamma = 0.99$权衡即时与未来奖励，采用Adam优化器（$\beta_1 = 0.9, \beta_2 = 0.999$）自适应调整梯度，平衡探索与利用以适配电力系统攻防场景。

### 2.3　多模态特征融合的防御模型

#### 2.3.1　混合数据集构建

构建包含原始运行状态、对抗攻击样本的多模态数据集 $\boldsymbol{D}=\{\boldsymbol{D}_{\text{norm}},\boldsymbol{D}_{\text{att}}\}$。其中，$\boldsymbol{D}_{\text{norm}}$ 为量测节点的原始运行数据；$\boldsymbol{D}_{\text{att}}$ 为由攻击生成模块输出的对抗样本，形成“真实数据—对抗扰动”的对偶训练空间。样本结构定义为 $\boldsymbol{D}\in\mathbb{R}^{N\times T\times F}$，$N$ 为样本总数，$T$ 为时间序列长度，$F=B\times P$ 为特征维度。

对混合集进行归一化处理，同时准备对应的标签，1 表示攻击样本，−1 表示正常样本。运用分层抽样法划分数据集为训练集与测试集，训练集占比 70%，测试集占比 30%，且两类数据中对抗样本与正常样本的比例保持一致。

#### 2.3.2　多模态特征处理

在电力系统攻击检测领域，单一特征难以全面表征复杂动态特性。本文采用 SAE 与 CNN-LSTM 模型构建协同防御架构。SAE 通过无监督降维弥补了 CNN-LSTM 对系统静态结构特征的建模缺失，CNN-LSTM 则通过时空动态特征提取完善了 SAE 对数据时序演化规律的表征盲区。

SAE 通过三层非对称非线性压缩架构及无监督学习机制，实现对数据内在静态结构特征的逐层提取与降维提纯，通过重构误差最小化剥离时序动态信息，显式建模系统稳态运行的本质结构。单层预训练时，AE 模型借助逐层贪婪算法对训练样本开展无监督学习，获取网络层初始参数；微调训练阶段，SAE 模型将样本数据作为输入，对模型整体进行训练[25]。SAE 网络结构设计如图 2 所示。

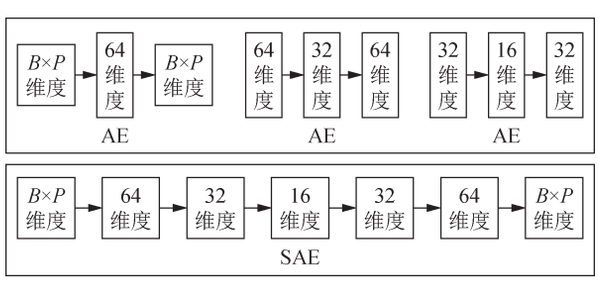


**图 2　SAE 模型结构**

**Fig. 2　SAE model structure**

CNN 用于提取数据的局部特征[26]，LSTM 用于处理时间序列特征[27]。利用 CNN 自动提取数据的空间特征，并将输出的特征向量构造成时间序列形式作为 LSTM 网络的输入[28]。CNN 模型由输入层、卷积层、批归一化层和池化层组成，选用 ReLU 激活函数。LSTM 模块包含独立的 LSTM 层、Dropout 层和全连接层，LSTM 层的门控机制采用 sigmoid 和 tanh 激活函数，全连接层采用 ReLU 激活函数。

CNN 输入层接收维度为 $(N,T,F)$ 的时序电力数据，经第一层 32 个 3×1 卷积核（步长 1）卷积运算提取短期局部特征，通过 ReLU 激活函数增强非线性表达，利用批归一化层稳定训练过程，再经 2×1 最大池化层降维，第二、三层卷积层分别采用 64 个和 128 个 3×1 卷积核提取中高层抽象特征，最后通过全局平均池化层聚合成 128 维向量；LSTM 层接收 CNN 输出的三维张量 $(N,T',C)$，以 128 个隐藏单元进行时序建模捕获长程依赖，通过丢弃率 0.4 的 Dropout 层防过拟合，经 128 维全连接层非线性变换后，通过时间维度全局平均池化聚合成 128 维向量。

采用 Adam[29] 优化器训练，SAE 预训练阶段以学习率 0.01、$\beta_1=0.9$、$\beta_2=0.999$ 迭代 200 次完成单层参数优化，微调阶段学习率降至 0.001 并结合 L2 正则化；CNN-LSTM 则用初始学习率 0.001、$\beta_1=0.9$，$\beta_2=0.999$，批次大小 32，结合 L2 正则化和早停机制，自适应特性提升训练稳定性。

#### 2.3.3　特征融合与分类决策

针对电力系统数据的静态结构与动态时空双重特征，通过 SAE 与 CNN-LSTM 提取多模态特征并融合，结合 SVM 构建分类器。

SAE 第三层编码器输出的特征经时间维度均值池化剥离时序信息，提取静态结构特征：

$$\bar{\boldsymbol{h}}_{\text{SAE}}=\frac{\sum_{t=1}^{T}\boldsymbol{h}_{i,t}^{(3)}}{T}\tag{28}$$

式中：$\boldsymbol{h}_{i,t}^{(3)}$ 为第 $i$ 个样本在第 $t$ 时间步的 16 维特征向量。CNN-LSTM 输出的时序特征通过时间维度全局平均池化聚合动态信息：

$$\bar{f}_{\text{CNN-LSTM}}=\frac{\sum_{t=1}^{T}\boldsymbol{h}_{i,t}^{\text{lstm}}}{T}\tag{29}$$

式中：$\boldsymbol{h}_{i,t}^{\text{lstm}}$ 为 LSTM 层输出的维时序特征。采用动态权重分配策略实现 SAE 与 CNN-LSTM 特征的自适应融合。基于两者在验证集上的准确率 $A_{\text{SAE}}$ 和 $A_{\text{CNN-LSTM}}$ 计算 $\alpha$ 和 $\beta$，SAE 和 CNN-LSTM 特征的融合权重系数。

$$\alpha=\frac{A_{\text{SAE}}}{A_{\text{SAE}}+A_{\text{CNN-LSTM}}}\tag{30}$$

$$\beta=\frac{A_{\text{CNN-LSTM}}}{A_{\text{SAE}}+A_{\text{CNN-LSTM}}}\tag{31}$$

对 SAE 的特征和 CNN-LSTM 特征进行 Z-score 标准化处理后按权重拼接生成融合特征向量：

$$\boldsymbol{F}_{\text{fused}}=\left[\alpha\cdot\text{Norm}\left(\bar{\boldsymbol{h}}_{\text{SAE}}\right),\beta\cdot\text{Norm}\left(\bar{f}_{\text{CNN-LSTM}}\right)\right]\tag{32}$$

将加权融合的特征作为输入，选择径向基函数作为核函数[30]：

$$K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\exp\left(-\gamma\left\|\boldsymbol{x}_j-\boldsymbol{x}_i\right\|^2\right)\tag{33}$$

通过5折交叉验证网格搜索优化惩罚参数$C$与核系数$\gamma$，以hinge损失为目标，利用序列最小优化（sequential minimal optimization，SMO）算法求解对偶问题，实现攻击样本分类。

$$L\left[y,f(x)\right]=\max\left[0,1-yf(x)\right] \tag{34}$$

式中：$y$为样本的真实标签；$f(x)$为模型的预测得分。

### 2.4 闭环反馈机制与协同进化

针对电力系统对抗攻击的动态博弈特性，构建“攻击生成—防御检测—策略迭代”的闭环框架如图3所示，通过攻防双方的策略互馈与增量学习，实现对抗能力的持续进化。

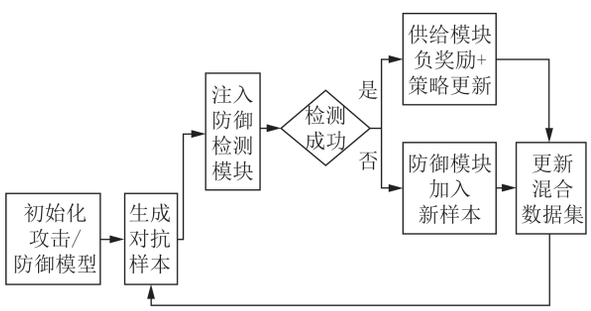


**图3 攻防对抗训练迭代流程图**

**Fig. 3 Flowchart of attack-defense adversarial training iteration**

防御检测模块对每一批输入样本进行实时检测，并将所有漏检样本（即攻击成功样本）及其对应的对抗样本向量存入一个固定容量的经验缓冲区$D$（容量设置为$N_{\text{buf}}$）。该缓冲区采用先进先出策略进行更新，以确保其中存储的始终是最新、最相关的攻击模式，同时防止缓冲区无限膨胀占用过多计算资源。

防御模型的增量学习采用定期触发机制，以平衡模型更新效率与计算开销。具体而言，每完成$K$轮完整的攻防对抗训练后，触发一次增量学习；每次从缓冲中随机采样$M$个样本（不足则全部使用）与当前训练集混合，构成增量训练批次；防御模型基于该混合数据集进行一个训练轮的微调，学习率设置为初始训练率0.1倍，以避免破坏已有知识表征。

攻击方的策略优化与防御方同步进行：在防御模型每次增量学习的同期，攻击方利用近$K$轮训练中积累的经验数据更新其PPO策略网络与价值网络；同时，防御模块提供的检测结果作为奖励函数中的惩罚，直接引导攻击策略向更隐蔽的方向演化，从而实现攻防双方的动态协同进化。

## 3 仿真分析

### 3.1 实验目标

攻击向量生成验证，用于验证自注意力结合强化学习生成的攻击向量对CCT预测的有效性。对比随机节点攻击、FGSM和PGD等传统攻击方法的性能。

防御方法验证，用于评估SAE-CNN-LSTM结合SVM模型对CCT攻击的检测能力。

### 3.2 ELM基准模型验证

测试集样本的CCT真实值与ELM模型预测值在分布趋势上高度契合，如图4所示。ELM模型对训练数据的特征表征能力显著，能够有效学习CCT相关特征的内在映射关系，精准捕捉系统暂态稳定性指标的动态变化规律，为后续对抗攻击实验提供了可靠的基准模型支撑。

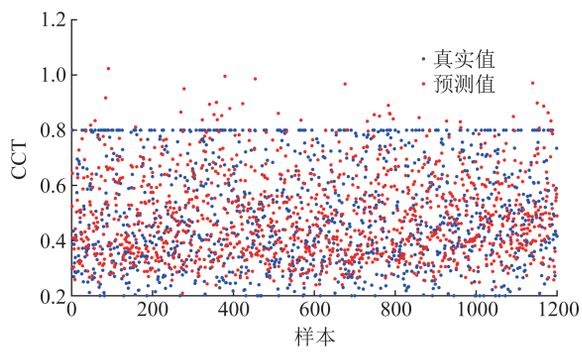


**图4 CCT真实值与预测值对比散点**

**Fig. 4 Scatter plot comparing actual values and predicted values of CCT**

### 3.3 攻击效果分析

为体现攻击向量对CCT预测算法的影响，分别统计正常条件下的CCT预测值及注入攻击向量后的预测偏差值。

攻击前后的CCT概率密度分布呈现显著差异，如图5所示。多数样本受攻击后，预测值分布向长时区间偏移；仅少数样本分布与正常预测接近，判定为攻击失效。且正常预测峰值对应的样本，受攻击后分布扰动更明显，反映攻击对系统关键工况的干扰效应更强，直观呈现攻击对CCT预测概率特性的影响，为量化攻击危害提供依据。

为验证本文方法在电力系统数据对抗攻击中的优越性，与随机节点攻击、FGSM和PGD方法进行对比实验，对比攻击成功率。

FGSM攻击和PGD攻击如下：

$$\boldsymbol{X}_{\text{adv}}=\boldsymbol{X}+\varepsilon'\cdot\text{sign}\left[\nabla_{X_t}J\left(\boldsymbol{X}_t,\boldsymbol{Y}_{\text{true}}\right)\right] \tag{35}$$

$$\boldsymbol{X}_{t+1}=\text{Clip}_{X,\varepsilon}\left\{\boldsymbol{X}_t+\alpha'\cdot\text{sign}\left[\nabla_{X_t}J\left(\boldsymbol{X}_t,\boldsymbol{Y}_{\text{true}}\right)\right]\right\} \tag{36}$$

式中：$\boldsymbol{X}_{\text{adv}}$为最终生成的对抗样本；$\varepsilon'$攻击强度参数；$J\left(\boldsymbol{X}_t,\boldsymbol{Y}_{\text{true}}\right)$为损失函数梯度；$\alpha'$是迭代步长。

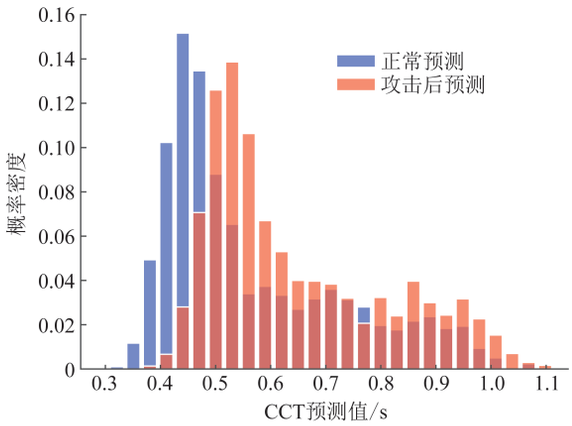


**图5　攻击前后CCT预测值概率密度对比**

**Fig. 5　Probability density comparison of predicted CCT values before and after attack**

图6呈现了四种对抗攻击方法成功率随攻击强度的变化曲线。Ours方法在攻击强度达一定值后增速加快，全范围成功率较优，体现策略有效与鲁棒；Random攻击成功率平缓上升、效率低，节点选择策略未结合特征敏感性；PGD和FGSM演化特征相似，低强度时成功率低，高强度后因梯度攻击强干扰性快速上升，但高扰动或违反电力系统物理规则。Ours方法在不同强度下的表现，贴合电力系统对抗攻击场景对鲁棒性和实用性的需求，为电力系统数据安全防御策略设计提供参考。

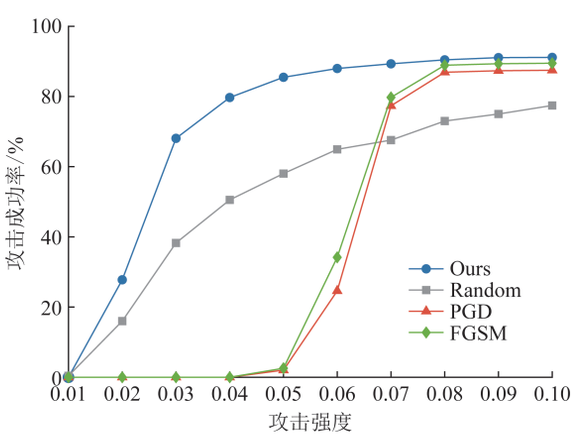


**图6　不同攻击强度下攻击成功率对比**

**Fig. 6　Comparison of attack success rates under different attack intensities**

图7为对抗攻击训练中，选择的关键节点在四个特征维度上的平均扰动幅度随训练回合的变化。每个子图对应一个特征，不同颜色曲线代表不同关键节点。直观呈现扰动强度变化，为分析攻击有效性、节点重要性及特征敏感性提供支撑。

采用归一化互相关（normalized cross-correlation，NC）和峰值信噪比（peak signal-to-noise ratio，PSNR）量化隐蔽性，其中NC用于表征两者的相似程度（数值越接近1相似性越高），PSNR用于衡量攻击向量中的扰动大小（数值越大扰动越小，通常30 dB以下扰动明显、30~40 dB扰动较小且不易察觉），具体计算

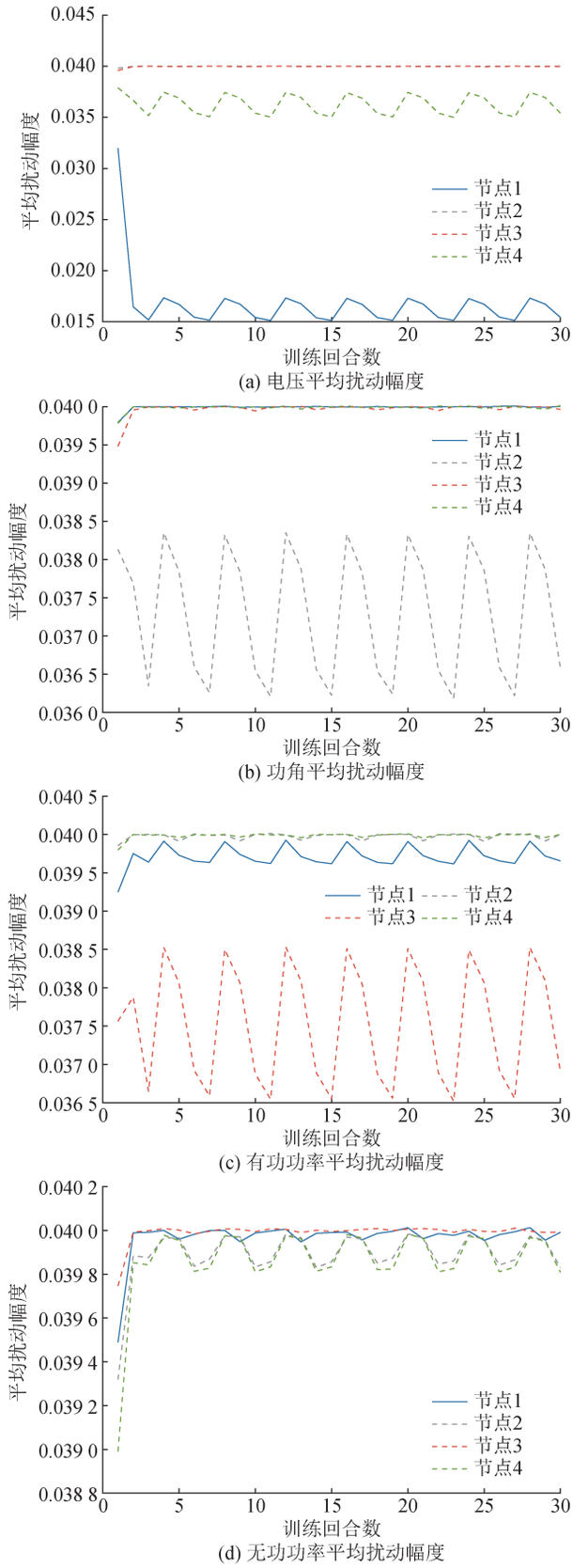


**图7　关键节点特征扰动变化曲线**

**Fig. 7　Feature perturbation variation curves for key nodes**

方法如下：

$$N_{Ci'}=\frac{\sum_{z=1}^{Z}\sum_{u=1}^{U}\bar{\boldsymbol{x}}_{i'}(z,u)\boldsymbol{x}_{i'}(z,u)/ZU}{\sqrt{\sum_{z=1}^{Z}\sum_{u=1}^{U}\bar{\boldsymbol{x}}_{i'}(z,u)^2/ZU}\sqrt{\sum_{z=1}^{Z}\sum_{u=1}^{U}\boldsymbol{x}_{i'}(z,u)^2/ZU}} \tag{37}$$

$$P_{\mathrm{SNR}i'}=20\log_{10}\left\{\frac{X_{\max}}{\sqrt{\sum_{z=1}^{Z}\sum_{u=1}^{U}\left[\bar{\boldsymbol{x}}_{i'}(z,u)-\boldsymbol{x}_{i'}(z,u)\right]^2/ZU}}\right\} \tag{38}$$

式中：$Z$和$U$分别代表输入数据矩阵的长与宽；$X_{\max}$是输入数据的最大可能值；$i'$为样本编号；$\boldsymbol{x}_{i'}$表示样本中的正常数据；$\tilde{\boldsymbol{x}}_{i'}$则是针对该样本生成的攻击向量。

$\varepsilon=0.04$时，为验证攻击隐蔽性，结合电力系统量测规范：同步相量测量装置（phasor measurement unit，PMU）电压量测误差≤0.5%、功角≤0.5°、功率≤1%，数据采集与监视控制系统（supervisory control and data acquisition，SCADA）核心电气量误差≤2%。本文攻击的关键节点中，电压扰动对应实际偏差0.2%~0.45%，功角、有功及无功扰动≤0.4%，均低于量测误差容限；本文攻击的NC达0.848 8，相似性高，PSNR为32.14 dB，扰动小且难察觉；FGSM（NC为0.581 5、PSNR为27.96 dB）与PGD（NC为0.637 6、PSNR为28.10 dB）则相似性欠佳、扰动明显。可见本文方法隐蔽性更优，能平衡相似性与扰动控制，为隐蔽性场景攻击策略提供参考。

图8为模型训练进程的监控曲线，横轴为训练回合数，左侧纵轴表示平均奖励，右侧纵轴表示攻击成功率。其中，蓝色实线代表的平均奖励，第5回合后稳定于3.1~3.2，波动极小，体现攻击策略在训练中快速收敛且稳定，反映强化学习策略梯度算法对攻击参数优化的高稳定性；红色实线代表的攻击成功率，前5回合升至85%，随后在85%~87%区间震荡，表明攻击策略可行性强，早期快速提升验证自注意力特征筛选与编解码器扰动生成协同作用，后期波动体现策略在探索-利用平衡下对不同样本适应性稳定，攻击泛化能力逐步增强。

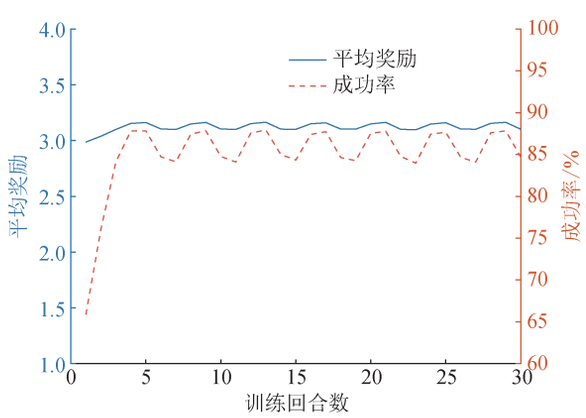


**图8 平均奖励与攻击成功率动态变化曲线**

**Fig. 8 Dynamic variation curves of average reward and attack success rate**

### 3.4 防御效果分析

在闭环训练中，防御模型增量学习的触发频率$K$设为5轮，反馈数据批量大小$M_{\mathrm{SIZE}}$设为512，经验缓冲区容量$N_{\mathrm{buf}}$为2000。这些参数经初步试验确定，能有效平衡学习效率与计算成本。

用训练好的SVM分类器对测试集融合特征向量预测，计算准确率$A$、精确率$P_{\mathrm{r}}$、召回率$R_{\mathrm{e}}$等指标，评估模型性能。

$$A=\frac{T_+ + T_-}{T_+ + T_- + F_+ + F_-} \tag{39}$$

$$P_{\mathrm{r}}=\frac{T_+}{T_+ + F_+} \tag{40}$$

$$R_{\mathrm{e}}=\frac{T_+}{T_+ + F_-} \tag{41}$$

式中：$T_+$为模型正确预测为正类的样本数；$T_-$为模型正确预测为负类的样本数；$F_+$为模型错误预测为正类的样本数；$F_-$为模型错误预测为负类的样本数。

表2为三种模型分别在准确率、精确率、召回率和F1分数的表现。SAE+SVM各项指标较低（准确率为66.83%、精确率为62.08%、召回率为77.54%、F1分数为0.69），特征表征与分类能力局限。CNN-LSTM+SVM性能显著提升（准确率为87.17%、精确率为83.12%、召回率为91.58%、F1分数为0.87），其处理时序与空间特征的判别力更强。融合模型表现最优（准确率为92.33%、精确率为90.51%、召回率为93.68%、F1分数0.92），通过整合SAE与CNN-LSTM的特征优势，增强了复杂模式识别能力，验证了特征融合策略提升模型综合性能的有效性与优越性。

**表2 不同模型的分类性能指标对比**

**Table 2 Comparison of classification performance metrics for different models**

| 模型名称 | 准确率/% | 精确率/% | 召回率/% | F1分数 |
|---|---|---|---|---|
| SAE+SVM | 67.83 | 70.83 | 77.54 | 0.69 |
| CNN-LSTM+SVM | 87.17 | 83.12 | 91.58 | 0.87 |
| SAE+CNN-LSTM+SVM | 92.33 | 90.51 | 93.68 | 0.92 |

受试者工作特性（receiver operating characteristic，ROC）曲线用于评估二分类模型性能，通过绘制不同分类阈值下真阳性率$T_{\mathrm{PR}}$与假阳性率$F_{\mathrm{PR}}$的关系来体现模型表现。

$$F_{\mathrm{PR}}=\frac{F_+}{F_+ + T_-} \tag{42}$$

$$S_{\mathrm{AUC}}=\sum_{q=1}^{Q}\left(F_{\mathrm{PR}_q}-F_{\mathrm{PR}_{q-1}}\right)\times\frac{T_{\mathrm{PR}_q}+T_{\mathrm{PR}_{q-1}}}{2}\tag{43}$$

式中：$S_{\mathrm{AUC}}$为ROC曲线下面积，用于量化模型的分类性能。

图9为ROC曲线对比，横轴为假阳性率，纵轴为真阳性率，展示三种模型的分类性能。ROC曲线反映模型在不同分类阈值下对正负样本的区分能力，$S_{\mathrm{AUC}}$越大性能越优。融合模型的ROC曲线最贴近左上角，$S_{\mathrm{AUC}}$值显著高于其他模型，凸显其在识别正例与区分负例上的卓越性能。

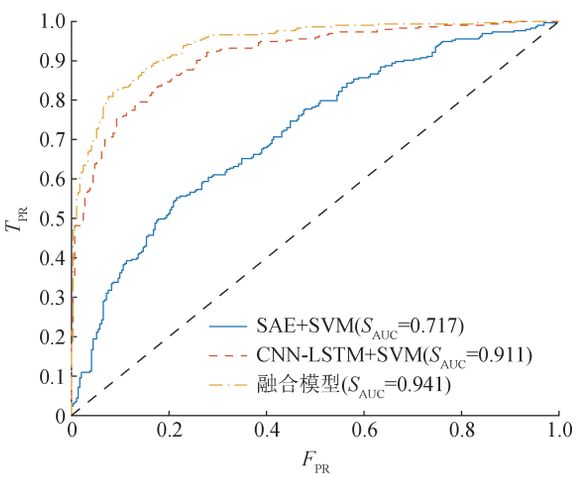


**图9　ROC曲线对比图**

**Fig. 9　ROC curve comparison plot**

## 4　结　论

本文聚焦新型电力系统对抗攻击与防御的协同优化，构建闭环防护体系，实现攻击靶向性、防御鲁棒性及特征辨识能力的突破。得出以下结论：

1）提出了一种基于自注意力与强化学习的协同攻击生成模型。通过自注意力机制量化节点重要性，并利用强化学习动态生成靶向扰动，克服了传统攻击策略的盲目性问题，提高了攻击的效率和隐蔽性。

2）构建了一种多模态特征融合的协同防御模型。通过融合SAE提取的静态结构特征和CNN-LSTM网络提取的时空动态特征，并结合动态权重策略，突破了单一模态的局限性，显著增强了对复杂对抗攻击的辨识能力与系统鲁棒性。

3）设计了一种攻防一体的闭环协同进化机制。构建了“攻击—防御—反馈”的动态优化框架，实现了攻防策略的持续迭代与协同演化。

未来可深化高维系统攻击泛化、实时防御轻量化及多类型攻击协同进化研究，进一步强化新型电力系统安全韧性。

**利益冲突声明（Conflict of Interests）：**

所有作者声明不存在利益冲突。

**作者贡献声明（Authors' Contributions）：**

吴润泽进行了研究设计，提出核心方法，构建攻防协同优化理论框架；张普阳负责论文撰写与修订，建立模型并进行数据分析，参与论文框架设计；郭昊博确定研究对象范围，收集相关数据；王嘉荣进行文献调研与整理，分析实验数据，并参与设计研究思路。所有作者均阅读并同意了论文终稿内容。

## 5　参考文献

[1] 汤奕，崔晗，李峰，等. 人工智能在电力系统暂态问题中的应用综述[J]. 中国电机工程学报，2019，39(1)：2-13，315.
TANG Yi，CUI Han，LI Feng，et al. Review on artificial intelligence in power system transient stability analysis [J]. Proceedings of the CSEE，2019，39(1)：2-13，315.

[2] 李峰，王琦，胡健雄，等. 数据与知识联合驱动方法研究进展及其在电力系统中应用展望[J]. 中国电机工程学报，2021，41(13)：4377-4389.
LI Feng，WANG Qi，HU Jianxiong，et al. Combined data-driven and knowledge-driven methodology research advances and its applied prospect in power systems [J]. Proceedings of the CSEE，2021，41(13)：4377-4389.

[3] 张怡，张恒旭，李常刚，等. 深度学习在电力系统频率分析与控制中的应用综述[J]. 中国电机工程学报，2021，41(10)：3392-3406.
ZHANG Yi，ZHANG Hengxu，LI Changgang，et al. Review on deep learning applications in power system frequency analysis and control[J]. Proceedings of the CSEE，2021，41(10)：3392-3406.

[4] 和敬涵，罗国敏，程梦晓，等. 新一代人工智能在电力系统故障分析及定位中的研究综述[J]. 中国电机工程学报，2020，40(17)：5506-5516.
HE Jinghan，LUO Guomin，CHENG Mengxiao，et al. A research review on application of artificial intelligence in power system fault analysis and location[J]. Proceedings of the CSEE，2020，40(17)：5506-5516.

[5] 冯双，崔昊，陈佳宁，等. 人工智能在电力系统宽频振荡中的应用与挑战[J]. 中国电机工程学报，2021，41(23)：7889-7904.
FENG Shuang，CUI Hao，CHEN Jianing，et al. Applications and challenges of artificial intelligence in power system wide-band oscillations [J]. Proceedings of the CSEE，2021，41(23)：7889-7904.

[6] 朱卫平，汤奕，魏兴慎，等. 针对电力CPS数据驱动算法对抗攻击的防御方法[J]. 中国电力，2024，57(9)：32-43.
ZHU Weiping，TANG Yi，WEI Xingshen，et al. Defense methods for adversarial attacks against power CPS data-driven algorithms[J]. Electric Power，2024，57(9)：32-43.

[7] 王新宇，张明月. 基于改进PSO优化RBF神经网络的新型电力系统虚假数据攻击检测研究[J]. 山东电力技术，2025，52(5)：50-56.
WANG Xinyu，ZHANG Mingyue. Research on detection of false data attack detection based on improved PSO and clustering optimization RBF in smart grid[J]. Shandong Electric Power，2025，52(5)：50-56.

[8] REN C, DU X N, XU Y, et al. Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples [J]. IEEE Transactions on Smart Grid, 2022, 13(2): 1622-1632.
[9] WANG J Y, XU G Q, LEI W Q, et al. CPFL: an effective secure cognitive personalized federated learning mechanism for industry 4.0 [J]. IEEE Transactions on Industrial Informatics, 2022, 18(10): 7186-7195.
[10] TIAN J W, WANG B H, LI J, et al. Adversarial attacks and defense for CNN based power quality recognition in smart grid[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(2): 807-819.
[11] ZHANG J L, GE C P, HU F, et al. RobustFL: robust federated learning against poisoning attacks in industrial IoT systems [J]. IEEE Transactions on Industrial Informatics, 2022, 18(9): 6388-6397.
[12] 陈洪, 陈惠文, 冯良坤, 等. 面向恶意攻击与级联失效的电力调度数据网络鲁棒性增强方法[J]. 广东电力, 2025, 38(2): 28-37.
CHEN Hong, CHEN Huiwen, FENG Liangkun, et al. Robustness enhancement method for power dispatching data network against malicious attacks and cascading failures [J]. Guangdong Electric Power, 2025, 38(2): 28-37.
[13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. [2020-06-22]. http://www.arxiv.org/pdf/1312.6199.pdf.
[14] 常颢, 徐俊俊, 王晓兵, 等. 基于对抗性自动编码器的城市配电网虚假数据注入攻击检测[J]. 山东电力技术, 2024, 51(3): 18-26.
CHANG Hao, XU Junjun, WANG Xiaobing, et al. Detection of false data injection attack in urban distribution network based on adversarial autoencoder [J]. Shandong Electric Power, 2024, 51(3): 18-26.
[15] 陶磊, 罗萍萍, 林济铿. 基于深度学习的直流微电网虚假数据注入攻击二阶段检测方法[J]. 中国电力, 2024, 57(9): 11-19.
TAO Lei, LUO Pingping, LIN Jikeng. Two-stage detection method for DC microgrid false data injection attack based on deep learning [J]. Electric Power, 2024, 57(9): 11-19.
[16] 刘增稷, 王琦, 薛彤, 等. 电力系统中数据驱动算法安全威胁分析及应对方法研究[J]. 中国电机工程学报, 2023, 43(12): 4538-4553.
LIU Zengji, WANG Qi, XUE Tong, et al. Research on security risks and defense methods of data-driven algorithms in power systems [J]. Proceedings of the CSEE, 2023, 43(12): 4538-4553.
[17] 黄冬梅, 丁仲辉, 胡安铎, 等. 低成本对抗性隐蔽虚假数据注入攻击及其检测方法[J]. 电网技术, 2023, 47(4): 1531-1540.
HUANG Dongmei, DING Zhonghui, HU Anduo, et al. Low-cost adversarial stealthy false data injection attack and detection method [J]. Power System Technology, 2023, 47(4): 1531-1540.
[18] YAO R Z, WANG N, LIU Z H, et al. Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion CNN-LSTM-based approach[J]. Sensors, 2021, 21(2): 626.
[19] LEE B, AMARESH S, GREEN C, et al. Comparative study of deep learning models for network intrusion detection [J]. SMU Data Science Review, 2018, 1(1).
[20] 赵玉明, 顾慎凯. 融合残差密集块自注意力机制和生成对抗网络的对抗攻击防御模型[J]. 计算机应用, 2022, 42(3): 921-929.
ZHAO Yuming, GU Shenkai. Adversarial attack defense model with residual dense block self-attention mechanism and generative adversarial network [J]. Journal of Computer Applications, 2022, 42(3): 921-929.
[21] ATHAY T, PODMORE R, VIRMANI S. A practical method for the direct analysis of transient stability [J]. IEEE Transactions on Power Apparatus and Systems, 1979, PAS-98(2): 573-584.
[22] LI F, WANG Q, TANG Y, et al. An integrated method for critical clearing time prediction based on a model-driven and ensemble cost-sensitive data-driven scheme [J]. International Journal of Electrical Power & Energy Systems, 2021, 125: 106513.
[23] HANNA J P, NIEKUM S, STONE P. Importance sampling in reinforcement learning with an estimated behavior policy [J]. Machine Learning, 2021, 110(6): 1267-1317.
[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
[25] 张海洋, 张斌. 结合堆栈自编码器和FSVM的入侵检测方法[J]. 桂林电子科技大学学报, 2024, 44(6): 621-627.
ZHANG Haiyang, ZHANG Bin. Intrusion detection method combining stacked autoencoder and FSVM [J]. Journal of Guilin University of Electronic Technology, 2024, 44(6): 621-627.
[26] 何俊鹏, 罗蕾, 肖塑, 等. 基于特征值分布和人工智能的网络入侵检测系统的研究与实现[J]. 计算机应用研究, 2021, 38(9): 2746-2751.
HE Junpeng, LUO Lei, XIAO Kun, et al. Framework for building network intrusion detection system based on feature distribution and AI[J]. Application Research of Computers, 2021, 38(9): 2746-2751.
[27] YU J J Q, HILL D J, LAM A Y S, et al. Intelligent time-adaptive transient stability assessment system [J]. IEEE Transactions on Power Systems, 2018, 33(1): 1049-1058.
[28] ZHAO Y X, XU Y B, YE J D, et al. Urban water supply forecasting based on CNN-LSTM-AM spatiotemporal deep learning model[J]. IEEE Access, 2023, 11: 144204-144212.
[29] KINGMA D P, BA J. Adam: a method for stochastic optimization [C]. Proceedings of the 2017 International Conference on Learning Representations, 2017: 71-80.
[30] AL-MEJIBLI I S, ALWAN J K, ABD D H. The effect of gamma value on support vector machine performance with different kernels [J]. International Journal of Electrical and Computer Engineering (IJECE), 2020, 10(5): 5497.

**收稿日期**: 2025-07-09 **修回日期**: 2025-09-28


吴润泽

**作者简介**:
吴润泽(1975),女,博士,副教授,主要研究方向为电力系统通信与信息技术;
张普阳(2000),男,硕士研究生,通信作者,主要研究方向为电力信息物理系统,E-mail:zhang_puyang_ok@163.com;
郭昊博(1994),男,博士,讲师,主要研究方向为电力系统通信与信息技术;
王嘉荣(2001),男,硕士研究生,主要研究方向为新型通信技术在物联网中的应用。

(编辑 孙静琳)