

基于谱敏感性识别与稀疏图触发生成器的 负荷预测后门攻击方法

张腾飞, 王玮, 高煜超, 杨杨
(南京邮电大学自动化学院, 江苏 南京 210023)

摘要: 在新型电力系统中, 基于深度学习的负荷预测模型已被广泛应用于资源调度与市场响应, 后门攻击已成为更具隐蔽性的潜在威胁。然而, 在这类强耦合、多变量预测场景中, 后门触发不仅要实现对预测输出的定向操控, 还须兼顾关键样本、变量选择与跨变量依赖。针对上述问题, 提出一种基于Koopman谱敏感性识别与图结构扰动建模的后门注入方法KoopTrigger。该方法首先基于Koopman算子敏感性分析选取高脆弱输入窗口以生成高效投毒样本; 其次构建门控稀疏图卷积生成器, 在变量维度中注入低秩扰动, 实现跨窗口的结构化后门触发; 最后通过双阶段对抗训练机制稳定诱导模型学习攻击目标。实验在ETTh1、Ireland、Panama与Milan这4个多变量数据集上开展, 结果表明KoopTrigger在多个预测模型中均可实现高效、隐蔽的预测操控, 性能显著优于消融版本, 相较基线方法具有更优的攻击效果与泛化表现。

关键词: 电力负荷预测; 后门攻击; Koopman算子; 图神经网络; 对抗训练

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ25284

Backdoor attack method for load forecasting based on spectral sensitivity identification and sparse graph trigger generator

Zhang Tengfei, Wang Wei, Gao Yuchao, Yang Yang

School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract: In new-type power systems, deep learning-based load forecasting models have been widely applied to resource scheduling and market response, making backdoor attack a more insidious potential threat. However, in these strongly coupled, multivariate forecasting scenarios, backdoor triggering not only needs to achieve targeted manipulation of the forecast output but also needs to consider key samples, variable selection, and cross-variable dependencies. To address these issues, a backdoor injection method, KoopTrigger, based on Koopman spectral sensitivity identification and graph structure perturbation modeling was proposed. Firstly, this method selects highly vulnerable input windows based on Koopman operator sensitivity analysis to generate efficient poisoning samples; then, it constructs a gated sparse graph convolution generator to inject low-rank perturbations into the variable dimensions, achieving structured backdoor triggering across windows; finally, a two-stage adversarial training mecha-

收稿日期: 2025-12-24; 修回日期: 2026-05-22

通信作者: 杨杨, yyang@njupt.edu.cn

基金项目: 国家自然科学基金资助项目 (No.62473204); 江苏重点研发计划资助项目 (No.BE2023838); 南京邮电大学华礼人才支持计划; 江苏省研究生科研与实践创新计划项目 (No.KYCX24_1215)

Foundation Items: The National Natural Science Foundation of China (No.62473204), Key Research & Development Plan of Jiangsu Province (No.BE2023838), Huali Program for Excellent Talents in Nanjing University of Posts and Telecommunications, Postgraduate Research & Practice Innovation Program of Jiangsu Province (No.KYCX24_1215)

nism is used to stably induce the model to learn the attack target. Experiments were conducted on four multivariate datasets: ETTh1, Ireland, Panama, and Milan. The results show that KoopTrigger can achieve efficient and covert prediction manipulation in multiple prediction models, significantly outperforming the ablation version. It also shows better attack performance and generalization performance compared to the baseline method.

Key words: power load forecasting, backdoor attack, Koopman operator, graph neural network, adversarial training

0 引言

在“双碳”目标与能源结构转型的背景下，电力系统正加速演进为源网荷储深度耦合的新型电力系统^[1]。在这一架构中，负荷预测处于关键前置位置，其输出直接服务于发电计划、频率调控策略以及市场化响应决策，因此预测精度、运行稳定性与结果可信性已成为支撑系统安全运行的重要基础^[2-3]。

新型电力系统如图1所示，其呈现“感知-计算-决策”一体化协同特征，多源量测与边端数据经通信网络汇聚并在平台侧完成预测与优化，再转化为调度与控制指令下发至各类可控资源^[4]。《新型电力系统发展蓝皮书》明确指出，电网控制功能由调控中心向配电、负荷控制以及第三方平台前移，电网的攻击暴露面大幅增加，电力系统已成为网络攻击的重要目标^[5]。

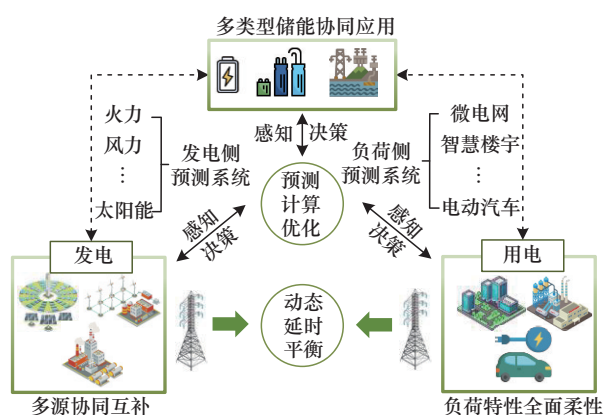


图1 新型电力系统

针对负荷攻击，已有研究主要关注运行阶段的数据层攻击，如负荷重分配攻击^[6]，其通过对局部负荷量测值的扰动误导状态估计。文献[7]构建了具备隐蔽性与资源限制的智能负荷攻击模型，并设计了基于预测与字典学习检测器的联合检测框架。文献[8]提出了一种不需要实时拓扑信息，仅操控局部负荷即可造成严重影响的数据

驱动负荷重分配攻击模型，突破了传统攻击对高计算成本与全局信息依赖的限制。文献[9]引入可解释增强型集成模型，在不平衡数据条件下提升了虚假数据注入检测的准确性与攻击定位能力。

尽管这些攻防方法在数据隐私层面^[10]已取得显著进展，但它们依赖于负荷观测数据，难以覆盖更隐蔽、更具持续性的模型级威胁。随着机器学习^[11]、深度学习模型^[12]在负荷预测中的广泛应用，攻击者的目标正逐步从数据干预演变为模型控制^[13]。尤其是在以多源信息融合^[14]为主流的背景下，这类模型虽具备强大的特征提取与时序建模能力^[15]，但也高度依赖数据驱动，结构复杂且可解释性弱^[16-17]，在训练、推理阶段极易被攻击。

相较于对抗样本^[18]或数据投毒^[19]等已被广泛关注的输入数据扰动类攻击，后门攻击^[20-21]因具备“训练期隐蔽植入、推理期定向激活”的特性，已成为研究前沿。该类攻击通过在模型训练阶段向部分样本中注入“触发器-目标输出”映射，实现了对预测输出的定向操控。一旦模型部署上线，只须触发信号便可精准干扰预测结果。

然而，相较于图像与文本分类等任务，负荷预测回归场景下的后门攻击设计更为复杂^[20-22]。一方面，负荷序列呈现出周期性、趋势性与节假日扰动等多尺度时序特征^[23]，攻击者须构造在这些统计规律下仍具有激活能力的扰动信号，以避免被模型的鲁棒自适应学习机制过滤。另一方面，负荷预测通常依靠多变量建模，涉及气温、设备运行状态、节假日等多个高度耦合变量^[24]，触发器不仅要作用于关键通道，还要保持变量间统计结构，以避免被检测为异常。

现有面向回归预测的后门攻击主要围绕长预测窗下触发效应衰减与低投毒率下后门稀疏两类难点展开。文献[25]利用变量相关性与时间滞后构造分布式触发，并结合投毒样本筛选与解耦训练目标提升长跨度操控的隐蔽性。文献[26]则引入位置感知优化，使目标模式可在预测窗内实现

延迟且跨维度异步激活。同时,生成式触发结合梯度筛选的稀疏化已被证明可有效压缩触发规模并提高隐蔽部署性^[27]。然而,上述方法仍停留在时间域设计,缺少对预测数据动力学敏感性分析,尤其在电力负荷预测这类强耦合、多变量场景中,触发还需要兼顾关键样本、变量选择与跨变量依赖。

鉴于上述挑战,本文聚焦于面向电力负荷预测任务的结构化后门攻击设计,提出一种后门攻击方法 **KoopTrigger**: 以 Koopman 谱敏感性定位关键输入数据窗口,以图结构扰动建模在保持变量依赖结构的同时完成触发注入。本文主要贡献如下。

(1) 针对多变量负荷序列的多尺度动力学特征,提出基于 Koopman 谱敏感性的输入窗口选取机制,使用动态模态分解近似各窗口的 Koopman 线性演化算子,计算窗口谱差并排序,识别动力学高度敏感的负荷片段,使有限投毒率下的触发更易沿预测模型的时序依赖被放大,提升负荷预测场景的定向攻击效率。

(2) 针对负荷变量间的显著耦合结构,设计基于门控稀疏图卷积的扰动触发生成器,使用图神经网络建模负荷变量依赖,引入通道门控稀疏化,仅在高贡献变量上注入低秩扰动,既保留跨变量耦合关系,又将触发范围压缩,增加负荷后门攻击隐蔽性。

(3) 构建双阶段对抗训练框架,在干净负荷数据上更新预测模型以保证预测器精度,在触发数据上更新生成器实现目标输出,两阶段交替优化将性能保持与后门激活解耦,避免单阶段联合训练导致的负荷预测精度损失或触发失效。

(4) 在 ETTh1、Ireland、Panama、Milan 典型多变量电力负荷数据集上开展消融实验与基线对比,验证所提 KoopTrigger 方法在攻击有效性、隐蔽性与模型泛化性方面的综合优势。

1 多变量时间序列预测的后门攻击定义

1.1 负荷多变量时间序列预测

在负荷多变量时间序列预测^[28]中,训练数据集包含若干相关变量,可表示为:

$$\mathbf{Z} = \{\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(d)}\} \in \mathbf{R}^{L \times d} \quad (1)$$

其中, L 表示时间序列长度, d 表示响应负荷的变量数量, $\mathbf{Z}^{(j)}$ 为第 j 个变量的时间序列。在负荷预测任务中,通常采用滑动窗口方式构造训练样本。设输入窗口长度为 ω_{in} , 则样本在时间 τ 处的历史输入片段 $\mathbf{Z}_i^{\text{in}}$ 可表示为:

$$\mathbf{Z}_i^{\text{in}} = \mathbf{Z}[\tau - \omega_{\text{in}} : \tau] \quad (2)$$

对应的目标输出片段 $\mathbf{Z}_i^{\text{out}}$ 可表示为:

$$\mathbf{Z}_i^{\text{out}} = \mathbf{Z}[\tau : \tau + \omega_{\text{out}}] \quad (3)$$

其中, ω_{out} 为输出窗口长度。

预测模型记为

$$f_{\theta}: \mathbf{R}^{\omega_{\text{in}} \times d} \rightarrow \mathbf{R}^{\omega_{\text{out}} \times d} \quad (4)$$

其目标是最小化在训练集 $\mathcal{I} = \{1, 2, \dots, M\}$ 上的平均预测误差:

$$\min_{\theta} \mathcal{L}(\theta) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \ell(f_{\theta}(\mathbf{Z}_i^{\text{in}}), \mathbf{Z}_i^{\text{out}}) \quad (5)$$

其中, $\mathcal{I}$ 为训练样本索引集合, M 为训练样本数量, $\ell(\cdot)$ 表示误差损失函数, θ 为预测模型参数。

1.2 后门攻击

在负荷后门攻击训练阶段,攻击者向负荷训练样本中注入部分带有触发器的样本,构成投毒数据集。其目标是在不显著降低整体预测精度的前提下,使模型在特定触发条件下输出指定的目标预测结果。

设原始负荷训练样本索引集合为 $\mathcal{I} = \{1, 2, \dots, M\}$, 定义投毒比例 $\alpha \in [0, 1]$, 被投毒样本索引为 $\mathcal{I}_{\text{atk}} \subseteq \mathcal{I}$, 满足 $|\mathcal{I}_{\text{atk}}| = \alpha M$; 定义变量注入率 $\beta \in (0, 1]$, 被注入触发模式的变量索引集合为 $\mathcal{V}_{\text{atk}} \subseteq \{1, 2, \dots, d\}$, 满足 $|\mathcal{V}_{\text{atk}}| = \beta d$ 。

攻击者在被选中的负荷训练样本输入中注入触发器,并在训练样本输出中嵌入攻击目标模式 $\mathbf{Z}_{\text{tgr}}$ 构造投毒样本:

$$\mathbf{Z}_{\text{atk}} = \mathcal{P}\left(\mathbf{Z}; \mathcal{I}_{\text{atk}}, \mathcal{V}_{\text{atk}}, \left\{\mathbf{p}_i, \mathbf{Z}_{\text{tgr}}\right\}_{i \in \mathcal{I}_{\text{atk}}}\right) \quad (6)$$

其中, $\mathbf{p}_i$ 为第 i 个样本的触发器,注入操作 $\mathcal{P}(\cdot)$ 仅在索引属于 $\mathcal{I}_{\text{atk}}$ 的样本与 $\mathcal{V}_{\text{atk}}$ 对应的变量上生效。

基于上述定义,负荷后门攻击训练可形式化为一个联合优化问题,攻击者的优化目标是在保证正常样本预测精度稳定的同时,使注入样本在触发条件下输出尽可能接近目标模式,目标函数形式为:

$$\begin{cases} \min_{\mathcal{I}_{\text{atk}}, \mathcal{V}_{\text{atk}}, \{p_i, \mathbf{Z}_{\text{tgr}}\}} \mathbb{E}_{i \in \mathcal{I}_{\text{atk}}} [\mathcal{L}_{\text{atk}}(f_{\theta^*}(\tilde{\mathbf{Z}}_i^{\text{in}}), \tilde{\mathbf{Z}}_i^{\text{out}})] \\ \text{s.t. } \theta^* = \arg \min_{\theta} \mathbb{E}_{j \in \mathcal{I}} [\ell(f_{\theta}(\mathbf{Z}_j^{\text{in}}), \mathbf{Z}_j^{\text{out}})] \\ \|\mathbf{p}_i\|_{\infty} \leq \Delta_{\text{TR}}, \|\mathbf{Z}_{\text{tgr}}\|_{\infty} \leq \Delta_{\text{TGR}} \end{cases} \quad (7)$$

其中, θ^* 为在投毒训练集上训练得到的模型参数, $\mathcal{L}_{\text{atk}}(\cdot)$ 表示攻击损失, 用于衡量后门触发后模型输出 $f_{\theta^*}(\tilde{\mathbf{Z}}_i^{\text{in}})$ 与目标模式 $\tilde{\mathbf{Z}}_i^{\text{out}}$ 之间的差异, $\|\cdot\|_{\infty}$ 为无穷范数, Δ_{TR} 与 Δ_{TGR} 分别为触发扰动幅度和目标偏移范围的约束。

2 KoopTrigger方法

KoopTrigger方法流程如图2所示, 为实现针对负荷多变量时间序列预测模型的精准操控与高隐蔽性后门注入, 本文提出一种基于Koopman算子识别与结构门控触发器注入的后门攻击方法, 命名为KoopTrigger。该方法通过以下3个阶段实现: (1) 基于Koopman算子识别负荷攻击输入窗口; (2) 通过图神经网络与门控机制生成可控稀疏触发器; (3) 采用对抗训练策略优化预测模型与触发生成器。

2.1 基于Koopman算子的投毒样本选择

在后门攻击中, 对于目标输出敏感的输入窗

口对预测模型参数更新的影响显著, 其能够显著放大扰动触发对预测结果的操控效果, 提升后门攻击训练效率。本文引入Koopman算子理论^[29], 以近似线性演化算子及其谱特征表征负荷输入窗口敏感度, 直接从数据层面捕捉非平稳、高敏感片段。

具体而言, Koopman算子可视为局部一阶演化算子, 当在输入窗口 $\mathbf{Z}_i$ 注入微扰 $\delta\mathbf{Z}_i$ 时, 其在特征空间诱发 $\delta\phi(\mathbf{X}_i)$, 并按以下方式在窗口内传播:

$$\delta\phi(\mathbf{X}_i') \approx K_i \delta\phi(\mathbf{X}_i) \quad (8)$$

其中, $\mathbf{X}_i$ 、 $\mathbf{X}_i'$ 分别表示由第 i 个输入窗口构造的前、后相邻快照矩阵, K_i 表示第 i 个输入窗口对应的Koopman算子。

因此, 扰动能否被放大取决于 K_i 的谱结构, 若主导特征值 λ_i 更大, 则同等幅值扰动在该窗口内更易积累为更大的状态偏移并映射为更显著的预测输出偏移。为捕捉这类高增益片段, 本文定义相邻窗口特征值谱差 $\Delta\lambda_i$, 对其排序选取窗口可稳定定位高脆弱负荷片段, 从而在低投毒率下获得更高的注入效率。考虑到Koopman算子本质上为无限维线性算子, 不可直接求解, 本文采用动态模态分解^[30]对其进行有限维近似, 得到

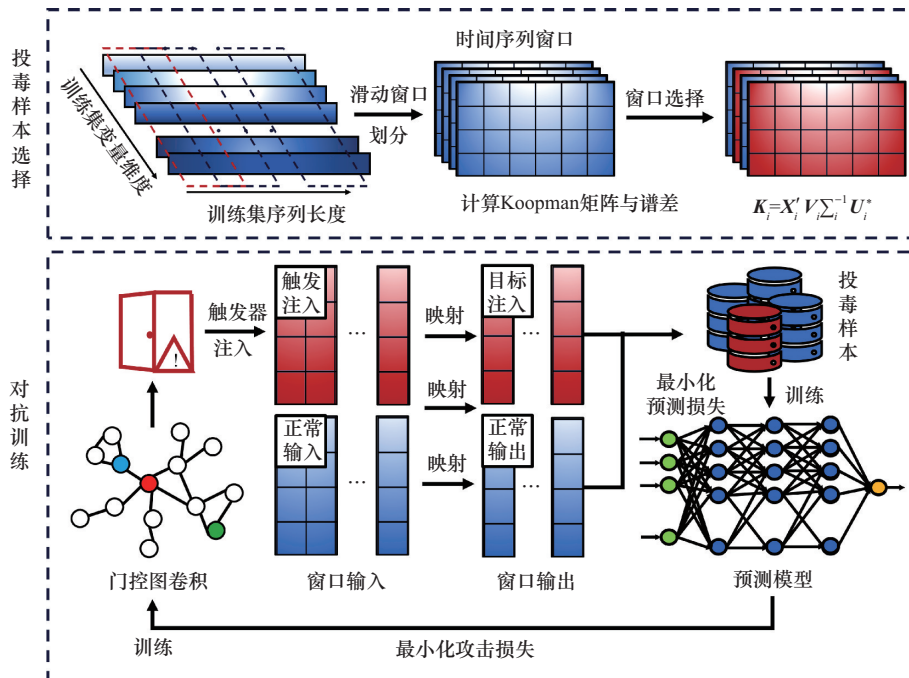


图2 KoopTrigger方法流程

每个窗口对应的 Koopman 算子。

具体地, 对于所有样本索引 $i \in \mathcal{I}$, 其输入窗口 $\mathbf{Z}_i^{\text{in}} \in \mathbf{R}^{\omega_{\text{in}} \times d}$, 记其在时间轴上的起始位置为 t_i 。对每个窗口构造两个快照矩阵:

$$\mathbf{X}_i = [\mathbf{x}_{t_i}, \mathbf{x}_{t_i+1}, \dots, \mathbf{x}_{t_i+\omega_{\text{in}}-2}] \in \mathbf{R}^{d \times (\omega_{\text{in}}-1)} \quad (9)$$

$$\mathbf{X}'_i = [\mathbf{x}_{t_i+1}, \mathbf{x}_{t_i+2}, \dots, \mathbf{x}_{t_i+\omega_{\text{in}}-1}] \in \mathbf{R}^{d \times (\omega_{\text{in}}-1)} \quad (10)$$

对 $\mathbf{X}_i$ 进行奇异值分解:

$$\mathbf{X}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^* \quad (11)$$

其中, $\mathbf{U}_i$ 为左奇异向量矩阵, $\mathbf{V}_i$ 为右奇异向量矩阵, $\mathbf{\Sigma}_i$ 为奇异值对角矩阵, $*$ 表示共轭转置, 则样本输入窗口的一阶 Koopman 算子可通过以下形式计算得到:

$$\mathbf{K}_i = \mathbf{X}'_i \mathbf{V}_i \mathbf{\Sigma}_i^{-1} \mathbf{U}_i^* \quad (12)$$

该矩阵即为窗口 i 上的线性动力学近似表达, 对每个 $\mathbf{K}_i$ 计算其特征值:

$$\lambda_i = \text{eig}(\mathbf{K}_i) = (\lambda_i^{(1)}, \lambda_i^{(2)}, \dots, \lambda_i^{(d)})^T \quad (13)$$

定义相邻窗口的谱差为:

$$\Delta \lambda_i = \|\lambda_i - \lambda_{i-1}\|_2, \quad i = 2, 3, \dots, M \quad (14)$$

将所有窗口谱差标准化后, 按从大到小排序, 选取前 aM 个窗口所属样本作为候选注入样本, 得到投毒样本索引集合:

$$\mathcal{I}_{\text{atk}} = \text{TopK}(\{\mathbf{Z}_i\}_{i=1}^M, [aM]) \quad (15)$$

其中, $\text{TopK}(\cdot)$ 表示在全部训练样本窗口上按照谱差从大到小排序并选取前 $[aM]$ 个窗口, a 为投毒比例, M 为训练样本总数。该集合将作为注入算子 $\mathcal{P}(\cdot)$ 的注入样本索引集, 从而可生成更加聚焦于模型脆弱性的投毒样本。

需要指出的是, 该方法所依据的 Koopman 算子相比基于残差、梯度等指标构建的敏感性评分^[31], 依据负荷数据本身的动力学属性, 其触发模式在训练中更容易被预测器吸收为稳定的响应映射, 具有更好的跨预测器泛化能力。

2.2 基于门控稀疏图卷积的扰动触发生成

为实现自适应高隐蔽性的后门扰动触发生成, 本文设计了一种基于门控稀疏图卷积的扰动触发生成方法。该方法通过图卷积网络 (graph convolutional network, GCN)^[32] 捕捉负荷变量之间的依赖关系, 并通过门控稀疏网络^[33] 动态选择关键变量通道, 从而有效提升后门激活的隐蔽性。

在具体实现上, 扰动触发生成依赖于投毒样本输入窗口前一时刻的状态信息。设某个投毒样本的输入窗口为 $\mathbf{Z}_i^{\text{in}} \in \mathbf{R}^{\omega_{\text{in}} \times d}$, 其前一输入窗口记为 $\mathbf{Z}_{i-1}^{\text{in}}$ 。扰动生成网络 G_ϕ 以 $\mathbf{Z}_{i-1}^{\text{in}}$ 为输入, 首先通过 GCN 提取变量之间的结构依赖, 生成初始扰动触发矩阵:

$$\mathbf{A}_i = \sigma(\tilde{\mathbf{A}} \mathbf{Z}_{i-1}^{\text{in}} \mathbf{W}_g) \in \mathbf{R}^{\omega_{\text{in}} \times d} \quad (16)$$

其中, $\tilde{\mathbf{A}} \in \mathbf{R}^{d \times d}$ 为对称归一化邻接矩阵, $\mathbf{W}_g$ 为图卷积权重, $\sigma(\cdot)$ 为非线性激活函数, $\mathbf{A}_i[t, j]$ 表示对第 j 个变量在窗口内第 t 个时刻的扰动触发幅度。

为了增强扰动触发隐蔽性, 本文引入变量通道上的门控稀疏机制, 首先对扰动触发矩阵沿时间维度平均池化, 提取通道级特征向量:

$$\bar{\mathbf{A}}_i = \frac{1}{\omega_{\text{in}}} \sum_{t=1}^{\omega_{\text{in}}} \mathbf{A}_i[t] \in \mathbf{R}^d \quad (17)$$

该特征向量经过一组门控权重参数 $\mathbf{W}_m \in \mathbf{R}^{d \times d}$, 偏置 $\mathbf{b}_m \in \mathbf{R}^d$ 和 Sigmoid 激活, 生成门控概率向量:

$$\mathbf{g}_i = \sigma(\mathbf{W}_m \bar{\mathbf{A}}_i + \mathbf{b}_m) \in [0, 1]^d \quad (18)$$

将 $\mathbf{g}_i$ 中概率最大的前 k 个通道保留, 其余置零, 得到稀疏掩码 $\mathbf{m}_i \in \{0, 1\}^d$:

$$\mathbf{m}_i = \text{TopK}(\mathbf{g}_i, k) \quad (19)$$

最后通过广播将掩码扩展至整个时间窗口, 形成与扰动触发矩阵同形的稀疏门控结构:

$$\tilde{\mathbf{A}}_i = \mathbf{A}_i \odot (\mathbf{1}_{\omega_{\text{in}} \times 1} \cdot \mathbf{m}_i^T) \quad (20)$$

其中, $\mathbf{1}_{\omega_{\text{in}} \times 1}$ 为全 1 列向量, 用于在时间维度上广播门控, $\odot$ 表示逐元素乘。

该稀疏扰动触发仅作用于被选中的关键变量通道, 有效提升了触发器激活的针对性与可解释性。本文将上述扰动矩阵叠加至投毒样本输入窗口, 得到注入扰动触发后的输入片段:

$$\tilde{\mathbf{Z}}_i^{\text{in}} = \mathbf{Z}_i^{\text{in}} + \tilde{\mathbf{A}}_i \quad (21)$$

为了诱导预测模型在该输入下输出指定的后门响应, 本文为每个投毒样本预定义一个目标输出模式 $\mathbf{Z}_{\text{tgr}}^{\text{out}} \in \mathbf{R}^{\omega_{\text{out}} \times d}$, 并叠加至原始输出端, 形成最终标签:

$$\tilde{\mathbf{Z}}_i^{\text{out}} = \mathbf{Z}_i^{\text{out}} + \mathbf{Z}_{\text{tgr}}^{\text{out}} \quad (22)$$

通过上述过程, 可构造一组包含触发器扰动与目标响应的投毒样本对:

$$\mathcal{D}_{\text{atk}} = \left\{ \left(\tilde{\mathbf{Z}}_i^{\text{in}}, \tilde{\mathbf{Z}}_i^{\text{out}} \right) \middle| i \in \mathcal{I}_{\text{atk}} \right\} \quad (23)$$

这些样本将与正常训练样本共同用于模型训练, 通过联合优化预测模型 f_θ 与扰动触发生成网络 G_ϕ , 实现在保持正常样本预测准确性的前提下, 对投毒样本后门目标的精准诱导。

2.3 双阶段对抗训练方法

在完成投毒样本选择与门控稀疏扰动触发生成网络设计后, 本文进一步提出一种双阶段对抗训练策略, 以优化扰动触发生成器参数, 使生成的扰动触发在保持攻击隐蔽性的同时, 实现对模型输出的有效操控。该训练过程采用预测模型 f_θ 与扰动触发生成器 G_ϕ 的交替优化框架, 分为两个阶段: 正常样本预测能力强化与投毒样本后门目标对齐, 如算法1所示。

算法1 KoopTrigger方法

输入 时间序列数据集 $\mathbf{Z}$ 、注入率 α 、最大训练轮数 E 。

- 1) 初始化预测模型 f_θ 与触发器生成器 G_ϕ ;
- 2) 将数据集划分为 M 个滑动窗口 $\{\mathbf{Z}_i^{\text{in}}, \mathbf{Z}_i^{\text{out}}\}$;
- 3) **for each** 窗口 $i = 1, 2, \dots, M$ **do**
- 4) 通过式(9)、式(10)构造快照对 $(\mathbf{X}_i, \mathbf{X}_i')$;
- 5) 通过式(11)、式(12)计算 Koopman 近似矩阵;
- 6) 通过式(13)、式(14)计算谱差 $\Delta\lambda_i$;
- 7) **end for**
- 8) 选取谱差最大的前 αM 个样本作为投毒样本集合 $\mathcal{I}_{\text{atk}}$;
- 9) **for** epoch = 1 $\rightarrow E$ **do**
- 10) **for** $i \in \mathcal{I}_{\text{atk}}$ **do**
- 11) 通过式(16)构造扰动触发;
- 12) 通过式(17)~式(21)生成稀疏扰动触发;
- 13) 通过式(22)、式(23)构建投毒样本对;
- 14) **end for**
- 15) 固定 G_ϕ 更新 θ 最小化预测损失式(24);
- 16) **if** epoch ≥ 10
- 17) 固定 f_θ 更新 ϕ 最小化攻击损失式(25);
- 18) **end if**
- 19) epoch $\leftarrow$ epoch + 1;
- 20) **end for**
- 21) **返回**: 预测模型参数 θ^* 与触发生成器

参数 ϕ^* 。

具体地, 在第一阶段, 固定扰动触发生成器 G_ϕ 的参数, 通过最小化正常样本上的预测误差来更新预测模型参数 θ , 以保持模型在非投毒数据上的准确性。设正常负荷训练样本集合为 $\mathcal{D}_{\text{clean}} = \left\{ \left(\mathbf{Z}_i^{\text{in}}, \mathbf{Z}_i^{\text{out}} \right) \middle| i \in \mathcal{I}_{\text{clean}} \right\}$, 预测损失函数定义为:

$$\mathcal{L}_{\text{pred}}(\theta) = \frac{1}{|\mathcal{I}_{\text{clean}}|} \sum_{j \in \mathcal{I}_{\text{clean}}} \left\| f_\theta(\mathbf{Z}_j^{\text{in}}) - \mathbf{Z}_j^{\text{out}} \right\|_1 \quad (24)$$

其中, $\|\cdot\|_1$ 为预测与观测值之间的平均误差。

在第二阶段, 固定预测模型参数 θ , 更新触发扰动生成器 G_ϕ 的参数, 使其生成的扰动在注入后能驱动预测模型输出预设目标模式。设投毒样本集合为 $\mathcal{D}_{\text{atk}} = \left\{ \left(\tilde{\mathbf{Z}}_i^{\text{in}}, \tilde{\mathbf{Z}}_i^{\text{out}} \right) \middle| i \in \mathcal{I}_{\text{atk}} \right\}$, 攻击损失函数定义为:

$$\mathcal{L}_{\text{atk}}(\phi) = \frac{1}{|\mathcal{I}_{\text{atk}}|} \sum_{i \in \mathcal{I}_{\text{atk}}} \left\| f_\theta(\tilde{\mathbf{Z}}_i^{\text{in}}) - \tilde{\mathbf{Z}}_i^{\text{out}} \right\|_2 \quad (25)$$

其中, $\|\cdot\|_2$ 为预测与观测值之间的平方误差。

上述两个阶段交替进行, 每次固定一组参数, 优化另一组。完成对抗训练后, 对应的预测模型参数 θ^* 与扰动触发生成器参数 ϕ^* , 能实现对后门目标的精准诱导。

3 实验设计与效果评估

3.1 数据集与实验设置

本文在 ETTh1^[34]、Ireland^[35]、Panama^[36]、Milan^[37]这4个典型负荷数据集上验证所提攻击方法。上述4个数据集均来自真实环境的公开采集。ETTh1数据集来自变压器运行监测场景, 采用15 min采样, 记录油温等设备状态量及与负荷相关的多变量序列。Ireland数据集由爱尔兰输电系统运营方 EirGrid 公开发布, 同样为15 min粒度的负荷时间序列, 反映国家级电网的实际负荷波动。Panama数据集是公开的短期负荷预测数据, 以小时为单位提供负荷及温度、湿度、风速等典型气象协变量。Milan数据集源自意大利电信大数据挑战赛的开放数据体系, 面向城市级多源输入, 可用于检验在城市环境多源信息融合条件下跨变量依赖模式的保持能力, 数据集的详细

描述见表 1。4 个数据集均按 6 : 2 : 2 的比例划分训练集、验证集与测试集。

表 1 数据集描述

数据集	采样量/个	特征维度	时间跨度
ETTh1	17 420	7	2016-07-01—2018-06-26
Ireland	11 612	9	2020-01-01—2020-04-30
Panama	48 048	13	2015-01-03—2020-06-27
Milan	4 464	184	2013-12-01—2013-12-31

实验选取了 4 类代表性负荷预测模型, 包括序列建模能力强的长短期记忆网络 (long short term memory, LSTM)、基于注意力机制的 FEDformer 与 Autoformer 以及具备结构建模能力的图神经网络 (graph neural network, GNN), 在统一参数配置下进行训练。所有模型采用 12 步输入与 12 步预测输出, 编码层数设为 2, 丢弃率 (dropout) 为 0.1, 学习率固定为 1×10^{-5} , 轮次 (epoch) 为 100, 批次数为 64。LSTM 的嵌入维度为 256, 前馈网络维度为 256; GNN 的嵌入维度为 384, 前馈网络维度为 1 024; FEDformer 与 Autoformer 的嵌入维度为 512, 前馈网络维度为 2 048。本文采用平均绝对误差 (mean absolute error, MAE) 作为统一评估指标, 设 MAE_C 反映负荷和预测结果与观测值之间的整体平均偏差, MAE_A 用于评估在触发器激活后, 负荷预测结果与攻击目标模式之间的偏移程度。

3.2 模块消融评估

本文在 ETTh1 与 Ireland 数据集上开展了消融实验, 实验统一采用 LSTM 作为预测模型, 投毒比例 α 设为 0.03, 对比以下 3 种配置。

(1) w/o Koopman & Gate: 不进行投毒样本选择, 同时移除门控稀疏部分。

(2) w/o Gate: 进行基于 Koopman 矩阵的投毒样本选择, 但移除门控稀疏部分。

(3) w/o Koopman: 使用基于门控稀疏图卷

积的扰动触发, 但扰动触发注入窗口随机。

模块消融评估实验结果见表 2。其中, MAE_C 表示 LSTM 在正常训练样本下预测值与观测值之间的 MAE, 反映其在未受攻击时的预测性能; MAE_A 表示 LSTM 输出与攻击目标之间的 MAE, 主要用于衡量后门攻击是否成功将模型预测引导至预设目标。KoopTrigger 在两组数据集中均取得最低 MAE_A , 且 MAE_C 接近正常训练模型, 说明该方法在保证预测精度的同时具备良好的攻击隐蔽性与有效性。从对比组来看, 移除任一模块均会显著削弱攻击性能。以 ETTh1 为例, MAE_A 分别上升至 1.38 (w/o Gate)、0.77 (w/o Koopman) 和 1.52 (w/o Koopman & Gate)。Ireland 数据集亦呈现相似趋势。

这一结果验证了基于 Koopman 算子的投毒样本选择能够利用 Koopman 算子挖掘敏感的输入窗口, 提升扰动触发注入的效率; 而门控稀疏扰动生成器通过引入门控稀疏图卷积结构, 在变量维度上压缩扰动范围, 从而提高触发器的隐蔽性。

3.3 攻击泛化评估

为了进一步验证 KoopTrigger 攻击方法在不同预测模型架构下的泛化能力, 本文在消融实验的基础上扩展了模型对比实验。除基础实验中采用的 LSTM 模型外, 进一步引入 GNN、Autoformer 和 FEDformer 作为预测模型进行对比分析, 以评估 KoopTrigger 攻击的泛化性。

攻击泛化评估结果见表 3, KoopTrigger 在不同模型下整体表现稳定。在 ETTh1 数据集中, 当 GNN、FEDformer 分别作为预测模型时, KoopTrigger 的 MAE_A 分别为 0.63 与 0.21, 均优于对应消融版本。在 Ireland 数据集中, KoopTrigger 的 MAE_A 在 GNN、Autoformer 与 FEDformer 这 3 种预测模型下分别为 179.81、112.02 与 88.03, 整体表现最佳。唯一例外是 ETTh1 中的 Autoformer 模

表 2 模块消融评估实验结果

数据集	干净模式	KoopTrigger		w/o Gate		w/o Koopman		w/o Koopman & Gate	
	MAE_C	MAE_C	MAE_A	MAE_C	MAE_A	MAE_C	MAE_A	MAE_C	MAE_A
ETTh1	1.04	1.44	0.50	1.42	1.38	1.62	0.77	1.63	1.52
Ireland	152.23	241.51	106.25	296.04	139.69	290.58	114.69	347.99	181.79

表3 攻击泛化评估结果

数据集	预测模型	干净模式	KoopTrigger		w/o Gate		w/o Koopman		w/o Koopman & Gate	
		MAE _C	MAE _C	MAE _A	MAE _C	MAE _A	MAE _C	MAE _A	MAE _C	MAE _A
ETTh1	GNN	1.08	1.09	0.63	1.07	0.69	1.09	0.75	1.12	0.69
	Autoformer	1.15	1.29	0.86	1.08	0.90	1.16	0.53	1.26	0.89
	FEDformer	1.06	1.03	0.21	0.96	0.68	1.03	0.26	1.06	0.27
Ireland	GNN	139.58	196.01	179.81	181.01	212.23	184.25	229.05	198.03	182.91
	Autoformer	109.99	145.23	112.02	134.73	141.90	148.37	238.62	123.68	161.26
	FEDformer	103.79	106.88	88.03	120.80	88.36	123.37	90.13	144.11	90.62

型，KoopTrigger（w/o Koopman）的 MAE_A 为 0.53，略优于 KoopTrigger 的 0.86。但整体结果表明，KoopTrigger 能在多种时序预测模型上保持良好的攻击性能，具备较强的泛化能力。

为进一步探讨双阶段对抗训练方法是否稳定有效，本文绘制了各预测模型训练阶段的损失收敛曲线，如图3所示。图3中 L1 与 L2 曲线分别表示训练中两个阶段对应的损失函数变化趋势。从 L1 曲线可知，在前10个训练轮次，模型处于正常训练阶段，损失稳定下降。第11个训练轮次开始注入触发器，导致损失出现突升，随后模型逐渐适应扰动信号，损失再次下降，最终趋于稳定。这表明模型在经历攻击扰动后，能够重新优化权重、学习攻击目标分布，实现对攻击目标的逐步对齐。此外，本文绘制了预测结果散点图以直观展示预测模型的拟合效果，预测散点如图4所示，FEDformer 模型在攻击目标拟合上表现最

优，其他模型同样展现出良好趋势，整体与表3中的定量指标一致。

本文进一步将 KoopTrigger 与后门攻击基线时间解耦后门攻击（temporally decoupled backdoor attack, TDBA）、随机触发基线方法（Random）使用 LSTM 在4种典型电力负荷数据集上进行对比。后门攻击基线对比见表4，在 ETTh1、Ireland、Panama 数据集上，KoopTrigger 的攻击目标拟合效果更强，而在 Milan 数据集上，TDBA 依靠异步激活隐蔽性更高。由于 TDBA 侧重时间域的位置感知触发优化，而 KoopTrigger 结合谱敏感性与跨变量结构约束进行触发构造，其优势主要体现在强非平稳或长期依赖更显著的多变量电力负荷数据集上。

此外，图5展示了 KoopTrigger 方法在 Ireland 数据集上注入率为 0.03、预测模型为 LSTM 时的攻击。由图5可知，KoopTrigger 在触发器生

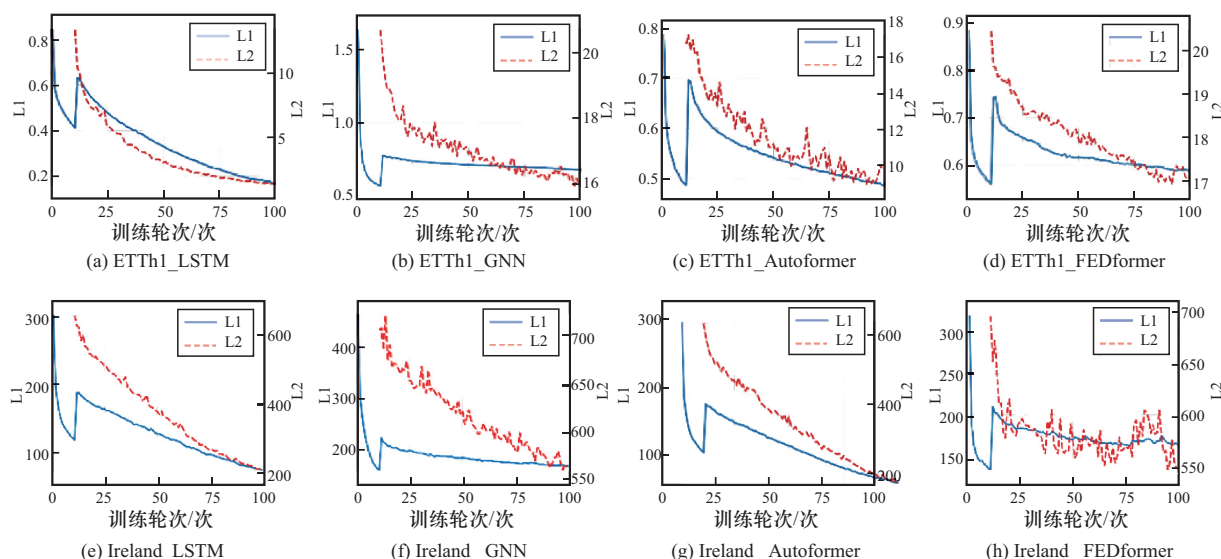


图3 损失收敛曲线

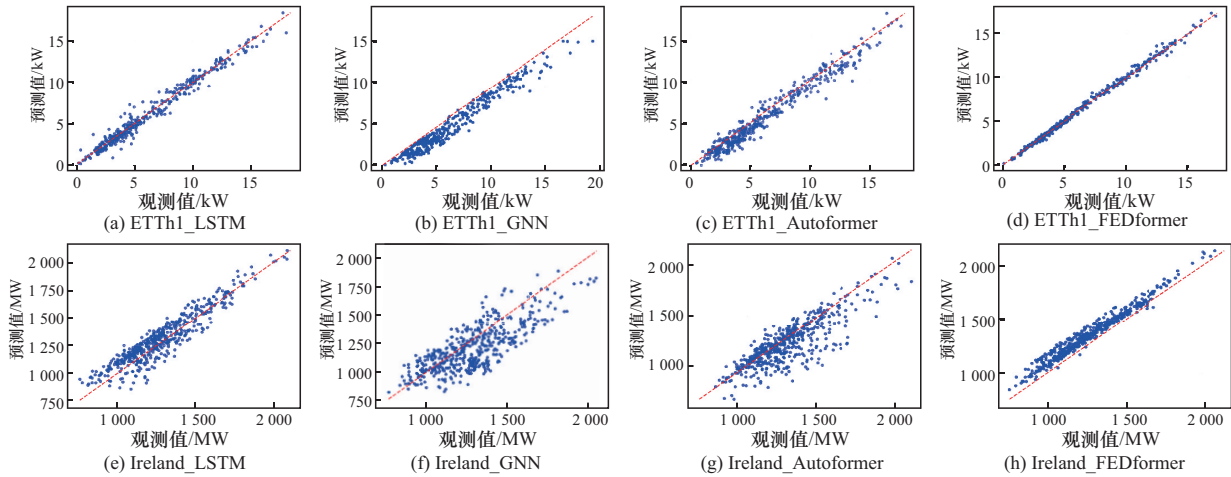


图4 预测散点

表4 后门攻击基线对比

数据集	KoopTrigger		Random		TDBA	
	MAE _C	MAE _A	MAE _C	MAE _A	MAE _C	MAE _A
ETTh1	1.44	0.50	1.51	1.42	1.42	1.38
Ireland	241.51	106.25	255.81	218.75	296.04	139.69
Panama	67.21	35.54	60.24	50.50	55.16	42.83
Milan	8.87	8.31	8.00	9.27	7.26	7.05

效后成功诱导模型输出向攻击目标模式偏移, 预测结果与目标模式高度接近, 说明扰动具有显著的引导能力。

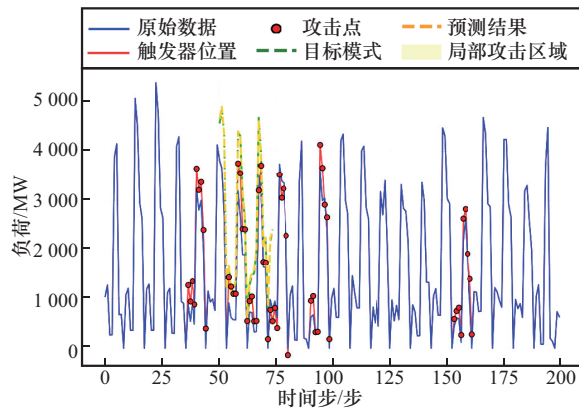


图5 KoopTrigger攻击示意

3.4 注入率敏感性评估

为探究不同注入率 α 对KoopTrigger攻击效果的影响, 本文在模块消融评估实验的基础上设置了3种注入率配置(α 分别为0.01、0.03、0.05), 并在ETTh1与Ireland数据集上分别评估了4种预

测模型的表现。

实验所得注入率敏感性热图如图6所示。总体来看, 随着注入率的增加, 模型的MAE_C略有上升, 但整体变化幅度较小, 说明KoopTrigger在不同注入强度下均能保持较高的隐蔽性和预测稳定性。

ETTh1数据集上模型的MAE_A整体随注入率增加而降低, 表明更高的注入比例能够增强攻击强度, 使模型输出更趋近于目标模式。以LSTM为例, 当注入率从0.01提升至0.05时, MAE_A由1.52降至0.47, 攻击误差显著减小。GNN、Autoformer与FEDformer模型均呈现类似趋势, 其中FEDformer在 $\alpha = 0.05$ 时表现最优, MAE_A仅为0.20。

在Ireland数据集上, LSTM的MAE_A随注入率增加先降后升: 当 α 从0.01增加至0.03时, MAE_A由226.29降至106.25, 随后在 $\alpha = 0.05$ 时略回升至158.58。Autoformer亦出现类似波动, 而GNN与FEDformer模型整体表现平稳, 且在 $\alpha = 0.05$ 时分别取得最优MAE_A值162.33与

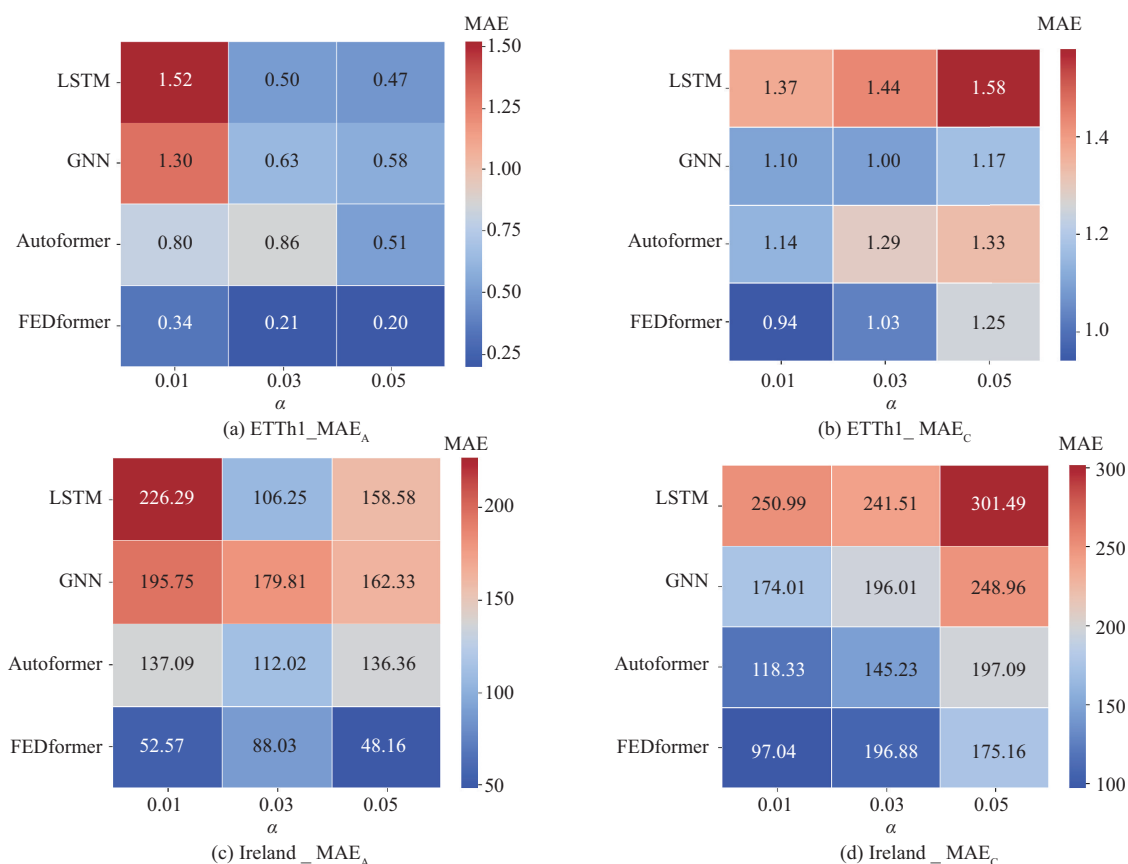


图6 注入率敏感性热图

48.16。

综上所述，KoopTrigger方法在不同注入率下均展现出良好的适应性与稳定性，说明较低的注入率有助于保持模型的原始预测性能，增强攻击的隐蔽性；而适度提高注入率则能显著提升攻击效果，使模型输出更接近预设目标。

4 结束语

本文聚焦于新型电力系统中负荷预测模型面临的后门攻击问题，提出了一种后门攻击框架KoopTrigger，利用Koopman算子谱敏感性分析选择关键输入窗口，结合稀疏图扰动生成机制与双阶段对抗训练，实现了对预测输出的定向控制。在ETTh1、Ireland、Panama、Milan这4个典型多变量电力负荷数据集上开展消融实验与基线对比，验证了所提KoopTrigger方法在攻击有效性、隐蔽性与模型泛化性方面的综合优势。未来工作将进一步拓展KoopTrigger在其他关键时序任务中的迁移能力，引入模型层级的防御机

制，并在实际环境中进行测试，推动可信负荷预测系统的构建。

参考文献：

- [1] 韩富佳, 王晓辉, 乔骥, 等. 基于人工智能技术的新型电力系统负荷预测研究综述[J]. 中国电机工程学报, 2023, 43(22): 8569-8591.
Han F J, Wang X H, Qiao J, et al. Review on artificial intelligence based load forecasting research for the new-type power system[J]. Proceedings of the CSEE, 2023, 43(22): 8569-8591.
- [2] 左宇, 秦大林, 杜尔顺, 等. 交互式负荷预测: 研究框架与展望[J]. 中国电机工程学报, 2026, 46(6): 2381-2394.
Zuo Y, Qin D L, Du E S, et al. Interactive load forecasting: research framework and outlook[J]. Proceedings of the CSEE 2026, 46(6): 2381-2394.
- [3] Hasan M, Mifta Z, Papiya S J, et al. A state-of-the-art comparative review of load forecasting methods: characteristics, perspectives, and applications[J]. Energy Conversion and Management: X, 2025, 26: 100922.
- [4] Liu M X, Teng F, Zhang Z Y, et al. Enhancing cyber-resiliency of DER-based smart grid: a survey[J]. IEEE Transactions on Smart Grid, 2024, 15(5): 4998-5030.
- [5] 《新型电力系统发展蓝皮书》编写组. 新型电力系统发展蓝皮书

- [M]. 北京:中国电力出版社,2023.
- The compilation team for the blue book on the development of new-type power systems. Blue book on the development of new-type power systems[M]. Beijing:China Electric Power Press,2023.
- [6] Verma P, Chakraborty C. Load redistribution attacks against smart grids – models, impacts, and defense: a review[J]. IEEE Transactions on Industrial Informatics, 2024, 20(8): 10192-10208.
- [7] Deng W F, Xiang Z L, Huang K K, et al. Detecting intelligent load redistribution attack based on power load pattern learning in cyber-physical power systems[J]. IEEE Transactions on Industrial Electronics, 2024, 71(6): 6285-6293.
- [8] Verma P, Dasgupta P, Chakraborty C. DeepLR: deep learning assisted load redistribution attack on cyber-physical power system[J]. IEEE Transactions on Smart Grid, 2026, 17(1): 551-561.
- [9] Dehbozorgi M R, Rastegar M, Arani M F M. False data injection attack detection and localization framework in power distribution systems using a novel ensemble of CNNs and explainable artificial intelligence[J]. IEEE Transactions on Industry Applications, 2025, 61(3): 4801-4811.
- [10] 刘圣龙, 黄秀丽, 江伊雯, 等. 面向多方数据融合分析的隐私计算技术综述[J]. 网络与信息安全学报, 2024, 10(6): 24-36.
- Liu S L, Huang X L, Jiang Y W, et al. Review of privacy computing techniques for multi-party data fusion analysis[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 24-36.
- [11] Ibrahim M S, Dong W, Yang Q. Machine learning driven smart electric power systems: current trends and new perspectives[J]. Applied Energy, 2020, 272: 115237.
- [12] Eren Y, Küçükdemir İ. A comprehensive review on deep learning approaches for short-term load forecasting[J]. Renewable and Sustainable Energy Reviews, 2024, 189: 114031.
- [13] 迟欢欢, 熊平, 刘恒竹, 等. 激励驱动的联邦学习研究综述: 隐私与安全[J]. 网络与信息安全学报, 2025, 11(3): 19-44.
- Chi H H, Xiong P, Liu H Z, et al. Survey on incentive-driven federated learning: privacy and security[J]. Chinese Journal of Network and Information Security, 2025, 11(3): 19-44.
- [14] Li Q, Xu P, He D Q, et al. Multi-source information fusion graph convolution network for traffic flow prediction[J]. Expert Systems with Applications, 2024, 252: 124288.
- [15] Kim N, Park H, Lee J, et al. Short-term electrical load forecasting with multidimensional feature extraction[J]. IEEE Transactions on Smart Grid, 2022, 13(4): 2999-3013.
- [16] Rahman M A, Islam M R, Hossain M A, et al. Resiliency of forecasting methods in different application areas of smart grids: a review and future prospects[J]. Engineering Applications of Artificial Intelligence, 2024, 135: 108785.
- [17] 陈晋音, 瞿康赞, 郑海斌. 基于模仿学习的深度强化学习训练数据推断攻击[J]. 网络与信息安全学报, 2025, 11(5): 50-67.
- Chen J Y, Qu K Y, Zheng H B. Membership inference attack based on imitation learning for deep reinforcement learning[J]. Chinese Journal of Network and Information Security, 2025, 11(5): 50-67.
- [18] Jiang W Z, Liu H, Liu F. Practical adversarial attacks on spatiotemporal traffic forecasting models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates, Inc., 2022: 19035-19047.
- [19] Takiddin A, Ismail M, Atat R, et al. Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(3): 1287-1301.
- [20] Fu D Q, Lin X, Liu Z N, et al. BackTime: backdoor attacks on multivariate time series forecasting[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024: 131344-131268.
- [21] 汪旭童, 尹捷, 刘潮歌, 等. 神经网络后门攻击与防御综述[J]. 计算机学报, 2024, 47(8): 1713-1743.
- Wang X T, Yin J, Liu C G, et al. A survey of backdoor attacks and defenses on neural networks[J]. Chinese Journal of Computers, 2024, 47(8): 1713-1743.
- [22] Luo J, Hong T, Fang S C. Benchmarking robustness of load forecasting models under data integrity attacks[J]. International Journal of Forecasting, 2018, 34(1): 89-104.
- [23] Zhang Y, Liu L F, Yuan F M, et al. Multifactor and multiscale method for power load forecasting[J]. Knowledge-Based Systems, 2023, 268: 110476.
- [24] Zhang Q F, Yan H, Liu Y M. Power generation forecasting for solar plants based on dynamic bayesian networks by fusing multi-source information[J]. Renewable and Sustainable Energy Reviews, 2024, 202: 114691.
- [25] Xiang K L, Yang H M, Hao M, et al. BadTime: an effective backdoor attack on multivariate long-term time series forecasting[PP]. arXiv preprint, 2025, arXiv: 2508.04189.
- [26] Liu Z X, Liu X L, Xu S H, et al. Beyond immediate activation: temporally decoupled backdoor attacks on time series forecasting[PP]. arXiv preprint, 2026, arXiv:2601.04247.
- [27] Chen J Y, Xiong H Y, Zheng H B, et al. Dyn-backdoor: backdoor attack on dynamic link prediction[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(1): 525-542.
- [28] Cai W L, Liang Y X, Liu X G, et al. MSGNet: learning multi-scale inter-series correlations for multivariate time series forecasting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(10): 11141-11149.
- [29] Otto S E, Rowley C W. Koopman operators for estimation and control of dynamical systems[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2021, 4: 59-87.
- [30] Baddoo P J, Herrmann B, McKeon B J, et al. Physics-informed dynamic mode decomposition[J]. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2023, 479 (2271): 20220576.
- [31] Jiao R H, Han Z T, Liu X, et al. A gradient-based wind power forecasting attack method considering point and direction selection[J]. IEEE Transactions on Smart Grid, 2024, 15(3): 3178-3192.

- [32] Hu Y Q, Cheng X Y, Wang S H, et al. Times series forecasting for urban building energy consumption based on graph convolutional network[J]. Applied Energy, 2022, 307: 118231.
- [33] Hong J C, Li K R, Liang F W, et al. A novel state of health prediction method for battery system in real-world vehicles based on gated recurrent unit neural networks[J]. Energy, 2024, 289: 129918.
- [34] Zhou H Y. Electricity transformer dataset (ETDataset)[DS]. GitHub [2025-10-23].
- [35] Irish social science data archive. ISSDA: Irish social science data archive[DB]. University College Dublin[2025-11-23].
- [36] Aguilar Madrid E. Short-term electricity load forecasting (Panama) [DS]. Kaggle[2025-11-01].
- [37] Abdullah M. Milan dataset[DS]. IEEE DataPort.[2025-11-01].

[作者简介]



张腾飞 (1980–), 男, 博士, 南京邮电大学自动化学院教授、博士生导师, 主要研究方向为微电网运行与控制技术、智能信息处理。



王玮 (2002–), 男, 南京邮电大学硕士生, 主要研究方向为电力负荷预测。



高煜超 (2000–), 男, 南京邮电大学博士生, 主要研究方向为新型电力系统智能控制技术。



杨杨 (1986–), 男, 博士, 南京邮电大学自动化学院教授、博士生导师, 主要研究方向为非线性控制理论及其在电力系统中的应用。