Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
AI Energizes Process Manufacturing—Article

针对工业故障分类系统的单变量攻击及其防御

卓越[a], Yuri A.W. Shardt[b], 葛志强[a,*]

[a] State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
[b] Department of Automation Engineering, Technische Universität Ilmenau, Ilmenau D-98684, Germany

ARTICLE INFO

Article history:
Received 28 January 2021
Revised 19 June 2021
Accepted 13 July 2021
Available online 3 June 2022

关键词
对抗样本
黑盒攻击
工业数据安全
故障分类系统

摘要

近年来，工业过程故障分类系统主要是由数据驱动的，得益于大量的数据模式，基于深度神经网络的模型显著地提高了故障分类的准确性。但是，这些数据驱动模型容易受到对抗攻击，因此，在样本上的微小扰动会导致模型提供错误的故障预测。最近的研究已经证明了机器学习模型的脆弱性以及对抗样本的广泛存在。本文针对安全、关键的工业故障分类系统提出了一种具有极端约束的黑盒攻击方法：只扰动一个变量来制作对抗样本。此外，为了将对抗样本隐藏在可视化空间中，本文使用了雅可比矩阵来引导扰动变量的选择，使降维空间中的对抗样本对人眼不可见。利用单变量攻击（OVA）方法，文本探究了不同工业变量和故障类别的脆弱性，有助于理解故障分类系统的几何特征。基于攻击方法，文本还提出了相应的对抗训练防御方法，该方法能够有效地防御单变量攻击，并提高分类器的预测精度。在实验中，将本文所提出的方法在田纳西-伊士曼过程（TEP）和钢板（SP）故障数据集上进行了测试。本文探索了变量和故障类别的脆弱相关性，并验证了各种分类器和数据集的单变量攻击和防御方法的有效性。对于工业故障分类系统，单变量攻击方法的攻击成功率接近（在TEP上）甚至高于（在SP上）目前最有效的一阶白盒攻击方法（该方法需要对所有变量进行扰动）。

1. 引言

在故障分类领域，目前已经提出了许多数据驱动的机器学习方法，并且这些方法的准确率也较高[1–2]。其中一些来自传统的机器学习方法，如用于多故障分类的支持向量机（SVM）[3]、用于处理动态故障分类问题的线性动态系统（LDS）[4]、基于线性判别分析的迁移学习[5]，以及集成方法，如SVM-forest [6]。深度学习在故障分类方面也得到了广泛的研究，由于数据的收集量和计算能力的快速增长，模型性能也得了提高，其中包括具有时频表示的深度卷积神经网络（CNN）[7]、用于时序故障检测的双向深度递归神经网络（RNN）[8]，以及考虑动态信息的稀疏堆叠自动编码器（AE）[9]。

尽管机器学习的分类准确性很高，但最近的研究表明，这些分类器普遍存在一个令人深思的弱点：输入样本上难以察觉的扰动将导致分类器以高置信度输出错误的预测[10–12]。先前的许多工作都研究了图像分类深度模型的对抗攻击。2013年，Szegedy等[13]首次发现了深度网络的这一特性，Goodfellow等[14]的快速梯度符号方法（FGSM）是一项代表性的工作，该方法使用样本的梯度

* Corresponding author.
E-mail address: gezhiqiang@zju.edu.cn (Z. Ge).

英文原文：Engineering 2022, 19(12): 240–251
引用本文：Yue Zhuo, Yuri A.W. Shardt, Zhiqiang Ge. One-Variable Attack on the Industrial Fault Classification System and Its Defense. *Engineering*, https://doi.org/10.1016/j.eng.2021.07.033

上升方向来制作对抗样本，并用它们训练了鲁棒的深度神经网络（DNN）。基于FGSM，一些方法会多次迭代以计算对抗样本上的扰动并获得更好的性能，如投影梯度下降（PGD）[15]、自由对抗训练（FreeAT）[16]、单次传播对抗训练（YOPO）[17]。与上述修改每个像素的方法不同，Su等[18]声称可以使用测试图像成功地欺骗三个不同的网络，其中只扰动每个图像的一个像素。Papernot等[19]也通过限制扰动的像素数来创建对抗攻击。

对应地，一些工作研究了传统分类器的安全问题。Barreno等[20]研究了基于贝叶斯的垃圾邮件检测模型的安全性。Biggio等[21]攻击了SVM分类器。Hu和Tan [22]使用生成对抗网络来合成对抗样本，并以随机森林、线性回归和决策树的形式成功攻击了恶意软件检测器。

对于对抗样本的防御，最常见的方法是对抗训练，即在训练过程中添加对抗样本，使模型更加鲁棒和正则化[13–14,23]。另一种方法是使用收缩或去噪网络[24]或者屏蔽梯度传播[25]。此外，训练辅助模型来检测攻击也是一种有效的防御方法[26]。

由于故障分类系统对于工业安全至关重要，错误分类可能导致严重的不良后果[27–28]，因此研究工业系统的对抗攻击和防御问题至关重要。为此，本文提出一种仅扰动一个变量的工业故障分类系统的黑盒攻击方法，称为单变量攻击（OVA）。攻击方法的主要动机和显著特征如下。

（1）**攻击场景**：本文主要考虑工业故障诊断系统遭受恶意攻击的情况。一种可能的攻击场景是，恶意黑客通过物理方法访问现实世界的传感器或利用智能分类系统相关的电子技术漏洞来访问工业系统。例如，通过改变流程中的环节来扰乱工业制造过程，并对整个系统造成严重损坏。因此，本文研究了智能故障诊断系统中最隐蔽的攻击方法，黑客可以在单个传感器上设置微小的偏移量，如温度或视觉测量值。根据实验结果，这些无法检测到的单变量攻击严重威胁着故障诊断系统。

（2）**单变量**：与工业系统中可能和多个传感器相关的过程故障不同[29]，恶意攻击更有可能发生在单个传感器测量中。因此，假设一个变量受到扰动而不是多个传感器同时被攻击。此外，从理论角度来看，单变量攻击通过分别分析每个故障变量，可以更轻松地提供故障分类器输入空间中的几何信息，从而帮助人们更深入地了解故障分类系统。

（3）**黑盒**：在对抗学习中，黑盒攻击模式只能访问分类模型的输出，不需要模型的任何内部信息，如结构或参数。黑盒攻击是非常通用的，使得这些攻击可以被用于不同的情况。例如，出于安全原因，我们通常无法访问故障分类模型的内部。基于黑盒属性，单变量攻击还能在没有任何梯度信息的情况下攻击不可微分的分类器。

此外，本文采用了一种选择顺序，可以使对抗扰动在降维可视化空间中不可见，而降维可视化是观察抽象工业数据的一种标准方法。这是通过使用变量降维映射的雅可比矩阵来实现的，使得有较小导数的变量具有更高的扰动优先级。

对应地，本文还提出了一种对抗训练方法来防御对抗攻击，使用在部分梯度方向上的对抗样本来训练分类模型。本文中防御方法对于单变量攻击是鲁棒的，并且提高了故障分类的准确性，这对于一些现有的对抗训练方法来说是难以实现的。

本研究的主要贡献总结如下：

• 本文提出了一种更适用于实际工业故障分类场景的黑盒单变量攻击方法。单变量攻击仅扰动单个变量，可以在没有内部信息的情况下攻击不同类型的故障分类器，并且生成的对抗样本难以被检测到。

• 本文通过使用单变量攻击评估变量和故障类别的脆弱性和几何特征，可以更深入地了解工业变量和故障分类系统。

• 本文提出了一种防御单变量攻击的对抗训练方法，该方法训练了具有更高准确性的鲁棒分类系统。

据了解，这是第一次针对工业故障分类系统提出和分析对抗攻击和防御方法。本研究其余部分组织如下。第2节介绍了单变量攻击方法和相应的评估结果。第3节在变量和故障类别水平上探讨了工业数据的脆弱性和鲁棒性，分类边界直观地提供了对故障分类系统的深入了解。第4节介绍了针对单变量攻击的防御方法。第5节总结全文。

2. 单变量攻击

本节将详细介绍单变量攻击的方法论和评估结果。图1直观地展示了单变量攻击的概览，基于田纳西-伊士曼过程（TEP）数据集（有关数据集详细信息，请参见第2.2节），通过不同故障类型的数据点图，报道了被攻击的变量并给出相应的分类结果。该图显示，单变量攻击方法可以通过仅扰动样本中的一个变量来威胁故障分类器。以故障6样本的DNN分类器为例（图1左侧子图），以10%（或20%）的值偏移量攻击变量20（或17），误导故障分类器错误地将它们预测为故障17（或8），置信度接近100%。同时，可以看出，在低维可视化空间中，精心制作的对抗样本的偏移是微小且不明显的，这得益于提出的扰动变量搜索顺序。

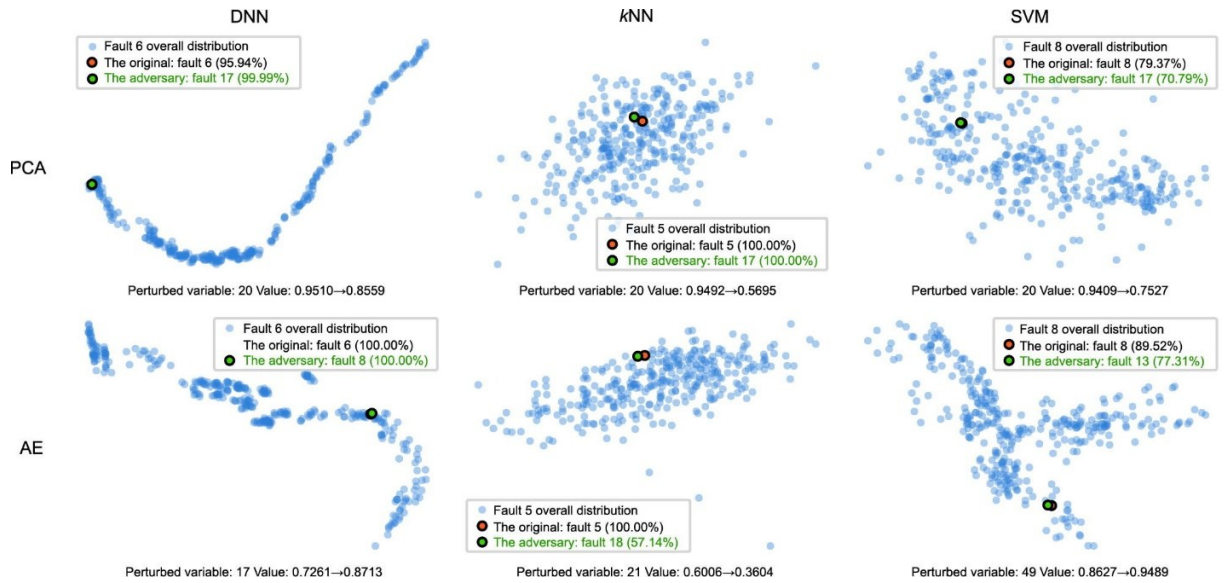

图1. 单变量攻击样本可视化。浅蓝色点表示某种故障类别的整体分布，橙色点表示原始样本，绿色点表示干扰原始样本的一个变量得到的对抗样本（一些点的距离非常接近以至于它们相互重叠）。每个图底部的注释给出了被攻击变量的原始值和目标值。图例显示了原始和目标的预测故障类型以及分类器的置信度。图中的行对应不同的二维（2D）降维技术：主成分分析（PCA，顶部）和自编码器（AE，底部），图中的列对应不同的分类器[DNN，k最近邻（kNN, $k=7$），SVM]。

2.1. 方法论

单变量攻击方法通过攻击原始样本来制作对抗样本，这些样本在特定范数测量中被限制在一定值以下。首先，训练工业数据的分类器，并选择在测试集上泛化最好的分类器（表示为f）进行攻击。正确预测的n维故障样本$\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$受到扰动。应该注意的是，只考虑在测试集中被$f$正确分类的样本。扰动$\boldsymbol{\eta}=(\eta_1, \eta_2, \ldots, \eta_n)$被定义为与输入样本大小相同。因此，单变量攻击为

$$f(\boldsymbol{x}+\boldsymbol{\eta}) \neq f(\boldsymbol{x})$$
$$\text{subject to } \|\boldsymbol{\eta}\|_0 = 1, \text{ and } \|\boldsymbol{\eta}\|_1 \leq \varepsilon \tag{1}$$

式中，ε是超参数；$\|\boldsymbol{\eta}\|_p (p=0 \text{ or } 1)$是$\boldsymbol{\eta}$的$l_p$范数。$l_0$范数计算向量的非零元素总数，用于约束要修改的变量数。l_1范数控制超参数以下的对抗样本和原始样本之间的距离ε，可以在区间[0, 1]中进行调整。为了更好地评估对抗样本的偏差，在以后的部分实验中，使用扰动比形式的失真代替绝对扰动大小$\boldsymbol{\varepsilon}$。对于扰动变量x_{pert}，失真被定义为原始变量值中的扰动比值，可以表述如下：

$$\text{Distortion} = \varepsilon / x_{\text{pert}} \tag{2}$$

为了使对抗样本在视觉空间中的位移尽可能小，在降维映射函数上计算雅可比矩阵的l_1列范数，以指导变量搜索期间的变量优先级。雅可比矩阵由多元函数的输入和输出之间的一阶偏导数组成。对于降维可视化映射函数$\boldsymbol{z}=F(\boldsymbol{x}), F:\mathbb{R}^n \to \mathbb{R}^m (m=2 \text{ or } 3)$，雅可比矩阵$\boldsymbol{J}_F$及其$l_1$列范数$\boldsymbol{v}$可以写成

$$\boldsymbol{J}_F = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_2}{\partial x_n} \\ \frac{\partial z_3}{\partial x_1} & \cdots & \frac{\partial z_3}{\partial x_n} \end{bmatrix} \tag{3}$$

$$\boldsymbol{v} = \|\boldsymbol{J}_F\|_{1-\text{col}} = \begin{bmatrix} \sum_{i=1}^{3} \frac{\partial z_i}{\partial x_1} & \cdots & \sum_{i=1}^{3} \frac{\partial z_i}{\partial x_n} \end{bmatrix} \tag{4}$$

按升序对$\boldsymbol{v}$中的元素进行排序，会产生一个可变序列，用于在攻击期间进行搜索。低维空间的较小梯度变量首先受到扰动以攻击分类器，保证生成的对抗样本在可视化空间中的偏移最小。

算法1显示了单变量攻击的过程。

Algorithm 1. One-variable attack for fault classification.

d is the predefined distortion; $\boldsymbol{v}^s$ is the index vector indicating the variable position in ascending order by sorting $\boldsymbol{v}$ (e.g. $\boldsymbol{v}^s(0)$ denotes the position index of the minimal variable, while $\boldsymbol{v}^s(n-1)$ denotes that of the maximal one);

Input: $\boldsymbol{x}$, f, d, $\boldsymbol{v}^s$

Output: adversarial samples $\boldsymbol{x}^*$

initial $\boldsymbol{x}^* \leftarrow \boldsymbol{x}$, $i=0$

repeat

$k = \boldsymbol{v}^s(i)$

$\varepsilon = d \times \boldsymbol{x}(k)$

$\boldsymbol{x}^*(k) = \boldsymbol{x}(k) \pm \varepsilon$

clip $\boldsymbol{x}^*(k)$ to [0, 1]

$i = i+1$

until $f(\boldsymbol{x}^*) \neq f(\boldsymbol{x})$ or $i=n$

2.2. 评估

为了全面验证单变量攻击对工业故障分类系统的有效性，选择了两个工业数据集：TEP [30]和钢板（SP）故障[31]。TEP是用于开发、研究和评估工业过程的公共基准数据集。TEP数据集由52个变量、28个故障类型和一个正常工作条件组成。TEP的流程图如图2所示。在实验中，选择了前21种故障类型的数据以及正常工作条件，每种故障类型大约有500个样本。SP数据集来自通信科学研究中心（Research Center of Sciences of Communication）Semeion [31]的研究，旨在正确分类不锈钢的表面缺陷类型。SP数据集中总共有1941个样本和7种故障类型。表1汇总了SP数据集中的属性和故障类型。SP数据集的27个属性根据其来源可以分为两种类型。使用视觉技巧，提取了钢缺陷的几何形状及其轮廓。此外，还描述了钢和输送机的内在特性。实际上，视觉传感器更容易受到攻击，小的扰动可能会对测量产生重大影响。因此，选择SP数据集进行故障分类系统中的单变量攻击评估。

在实验中，首先将所有样本归一化为从0到1的范围，并将数据集以3∶7的比例拆分为测试集和训练集。然后，利用训练集对分类器进行训练，选择测试集中正确预测的样本进行扰动攻击分类器。选择了三种类型的分类器：代表深度学习模型的DNN、SVM和用于传统机器学习模型的k最近邻（kNN）。表2报道了分类结果，包括分类准确性和置信度。分类准确性是衡量正确预测了多少测试样本的主要指标，置信度可以指示分类器f对其预测的信心，可以表述为

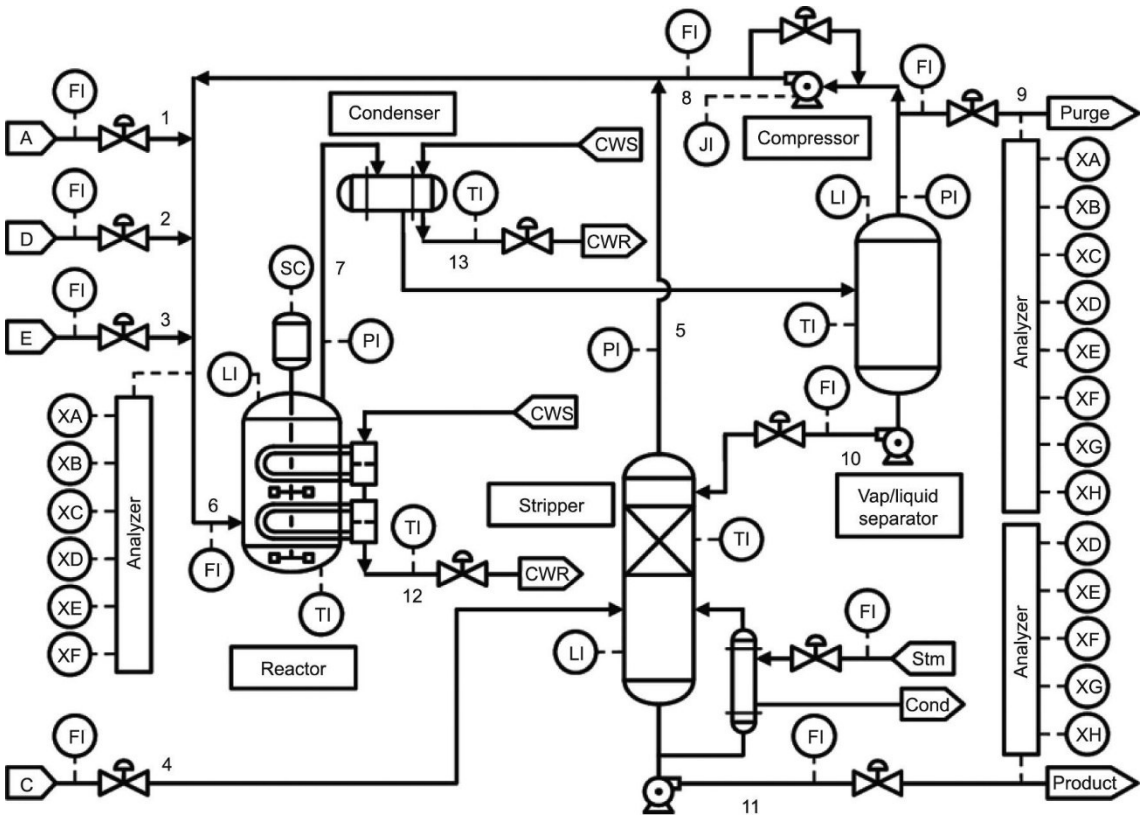

图2. TEP的流程图[30]。FI：流量指示器；Stm：蒸汽；Cond：冷凝器；LI：液位指示器；PI：压力指示器；TI：温度指示器；JI：压缩机功率指示器；SC：同步回旋加速器；XA、XB、XC、XD、XE和XF：分别为分析组分A、B、C、D、E和F；CWS：冷却水供应；CWR：冷却水回流。

表1 SP故障数据集摘要[31]

Attributes types	Attributes details	Fault types
Visual location (*X*, *Y*)	min, max, perimeter	1. pastry; 2. Z-scratch; 3. K-scratch; 4. stains; 5. dirtiness; 6. bumps; 7. others
Visual luminosity	min, max, sum	
Visual areas	pixels, sigmoid, log	
Visual index	edges, empty, square, outside, orientation	
Steel	type, thickness	
Conveyer	length	

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{I}\left\{ f\left(\boldsymbol{x}_{(i)}\right) = y_{(i)} \right\}}{N} \quad (5)$$

$$\text{Confidence} = \frac{\sum_{i=1}^{c} \Pr_f\left(y_{(i)} | \boldsymbol{x}_{(i)} \right)}{c} \quad (6)$$

式中，N是测试集中的样本编号；$\mathbb{I}$是指示函数；$\{\boldsymbol{x}_{(i)}, y_{(i)}\}$是一对样本及其真实标签。置信度是在正确预测的测试集上计算的，其样本数为$c = N \times \text{accuracy}$。

表2 测试集上的分类结果

Dataset	Accuracy (%)			Confidence (%)		
	DNN	*k*NN	SVM	DNN	*k*NN	SVM
TEP	75.8	57.4	56.2	81.7	62.0	54.0
SP	78.4	68.6	69.1	85.6	70.9	69.5

2.2.1. 攻击的有效性

在本节中，将研究单变量攻击在多大程度上可以误导故障分类。结果是通过两个工业故障数据和三个分类器的实验获得的。由于变量搜索的顺序不是本小节的重点（将在下一小节中讨论），因此所有单变量攻击实验都是使用恒定的随机搜索变量序列执行的。以下指标用于对抗攻击评估。

• **成功率**：在对抗攻击中，成功率是c个可攻击样本（定义见公式6）中成功攻击样本数量的比例，可以表示为

$$\text{Success rate} = \frac{\sum_{i=1}^{c} \mathbb{I}\left\{ f\left(\boldsymbol{x}_{(i)} + \boldsymbol{\eta}\right) \neq y_{(i)} \right\}}{c} \quad (7)$$

• **置信度**：对抗攻击的置信度与分类中的置信度略有不同。此处的置信度是根据成功扰动样本的错误预测概率计算的。置信度越高意味着分类器以更高的概率将对抗样本预测为不正确的故障类别。

$$\text{Confidence}_{\text{adv}} = \frac{\sum_{i=1}^{s} \Pr_f\left(f\left(\boldsymbol{x}_{(i)} + \boldsymbol{\eta}\right) \middle| \boldsymbol{x}_{(i)} + \boldsymbol{\eta} \right)}{s} \quad (8)$$

式中，$s = c \times \text{success rate}$。在下文中，所有置信度指标都按照公式（8）计算。

首先，验证了对于两个不同工业数据集的分类器，只需要修改样本的一个变量。在这种情况下，分类器可以以相对较高的置信度被成功攻击并且给出错误的故障类型预测。图3显示了结果。

根据成功率，DNN是三个分类器中最脆弱的，主要是因为深度模型的深度特征，导致小扰动被显著放大。对于TEP，即使失真相对较小，成功率也相当可观[10%（或20%）失真产生36.5%（或62.6%）的成功率]。另一方面，另外两个传统分类器比DNN的鲁棒性稍高，但仍然对于单变量攻击是脆弱的，尤其是当失真达到高水平时。可信度指标表明DNN通常更容易受到威胁。尽管失真和数据集不同，*k*NN和SVM输出的概率接近50%。相反，在大多数情况下，DNN的置信度压倒了其他两个模

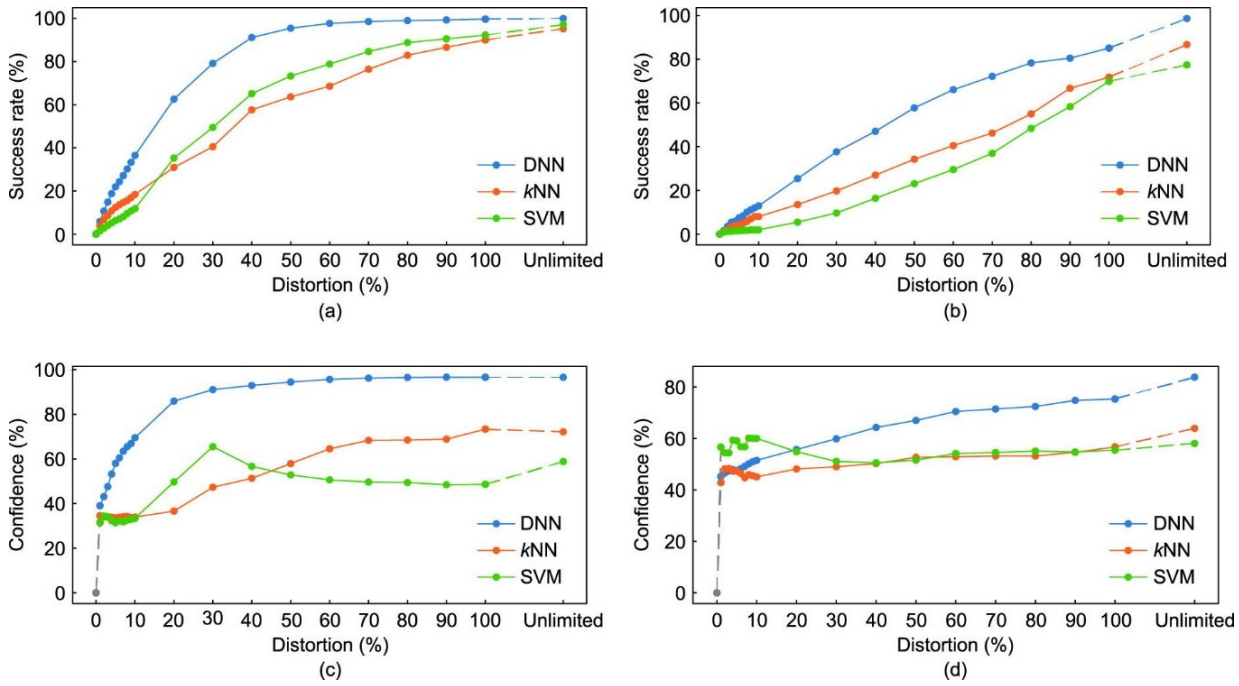

图3. 单变量攻击结果。图中显示了TEP [（a）、（c）]和SP [（b）、（d）]两个数据集的攻击成功率[（a）、（b）]和置信度[（c）、（d）]。失真值从0到无限，其中无限代表将变量扰动到最大边界[0, 1]†。

† For the variables less than 0.5, the maximal distortions are greater than 100%. For example, the maximal distortion on the variable of value 0.2 is 400% ((1–0.2)/0.2).

型。但是，对于SP数据集，当失真小于20%时，SVM的置信度高于DNN的置信度。这一有趣的观察结果表明，尽管只有极少数样本可以受到攻击，但成功攻击的样本对SVM具有很高的置信度。在两个工业数据集的比较中，对于所有三个分类器，TEP比SP更容易受到攻击。主要原因是SP上的分类器更加准确（如表2所示），这对应于更鲁棒的模型。因此，需要更大的扰动才能成功攻击SP的分类器。第3节中的内部数据分析证实了这一发现。

接下来，将本方法与对抗领域的一些有竞争力的攻击方法，如FGSM和PGD进行比较。两者都是白盒方法，它们根据梯度计算对抗样本。因为它们扰动了所有变量，为了公平竞争，对抗样本和原始样本之间的l_2距离范数被用作扰动的限制。攻击结果在相同l_2范数的扰动下进行比较。根据图4（a）、（b），对于工业故障分类，在相同的偏差距离限制下，仅扰动一个变量可以获得与扰动每个变量相似甚至更高的攻击成功率。与被认为是最强的一阶攻击方法PGD相比[15]，单变量攻击在SP数据集的成功率方面具有显著优势。然而，单变量攻击的置信度并不优于其他两种方法。

2.2.2. 变量搜索顺序的有效性

本小节展示了攻击期间变量的搜索顺序是如何影响可视化空间中对抗样本的分布。这些实验在没有失真限制的情况下攻击DNN分类器，并应用三维（3D）降维技术：主成分分析（PCA）和自编码器（AE）。图5比较了随机和基于雅可比的变量搜索顺序。从映射的最小梯度方向搜索变量可以将对抗样本人融合到原始分布中，而随机搜索可以使对抗样本在视觉上可区分。

此外，通过计算对抗样本与原始样本之间的l_2平均距离，比较两种变量搜索方法。从表3可以看出，对于基于雅可比的变量搜索，可视化空间中的偏差距离显著减小。

3. 工业数据脆弱性分析

之前在不同分类器和数据集上的结果表明，可以只用一个变量攻击工业故障分类系统，单变量攻击是一种通用的有竞争力的攻击方法。此外，还探讨了工业数据在变量和故障级别的脆弱性。研究目标是深入了解变量及其故障类型如何影响整个模型的脆弱性。

在DNN分类器和TEP数据集上的单变量攻击场景下分析了工业故障分类系统的漏洞（本节中的所有图示）。DNN和TEP是最主流的分类器和工业数据集，这种组合是具有代表性的。

3.1. 故障变量研究

与上一节不同，本节考虑了正确预测的故障样本中的所有变量，以探索哪些变量和故障类型容易受到攻击。首

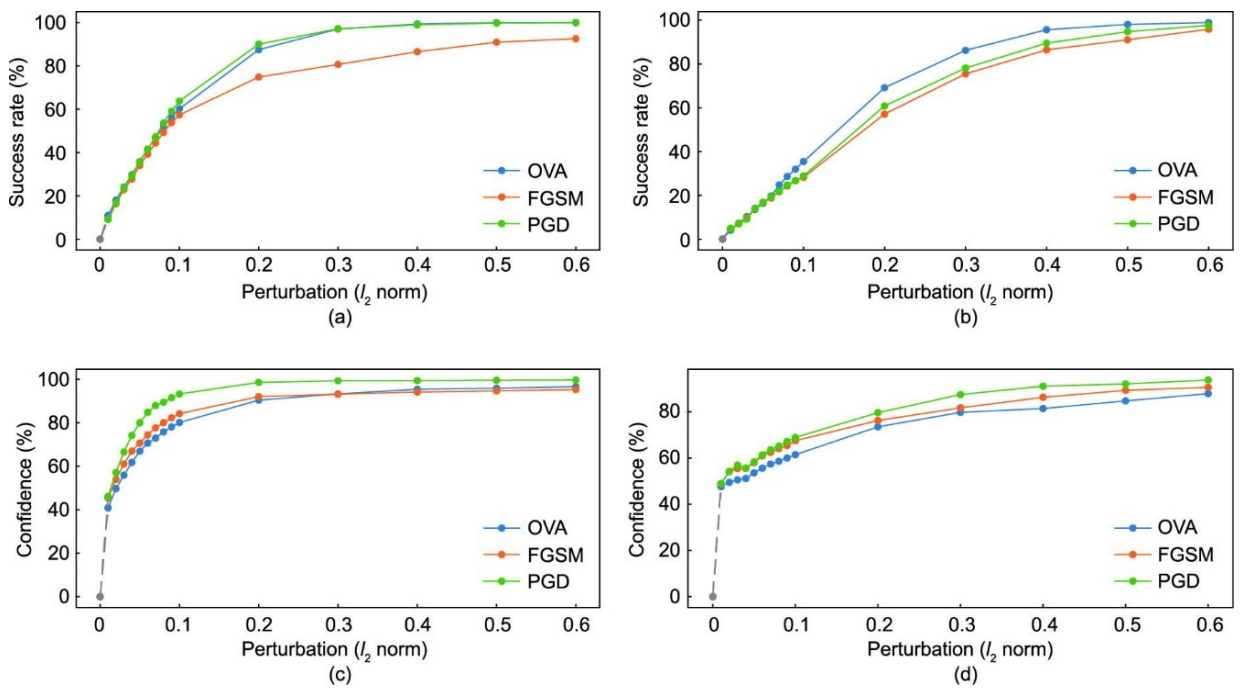

图4. DNN分类模型上的对抗攻击方法比较。这些数字显示了两个关于l_2扰动的范数的指标，即两个数据集TEP［（a）、（c）］和SP［（b）、（d）］的成功率［（a）、（b）］和置信度［（c）、（d）］。

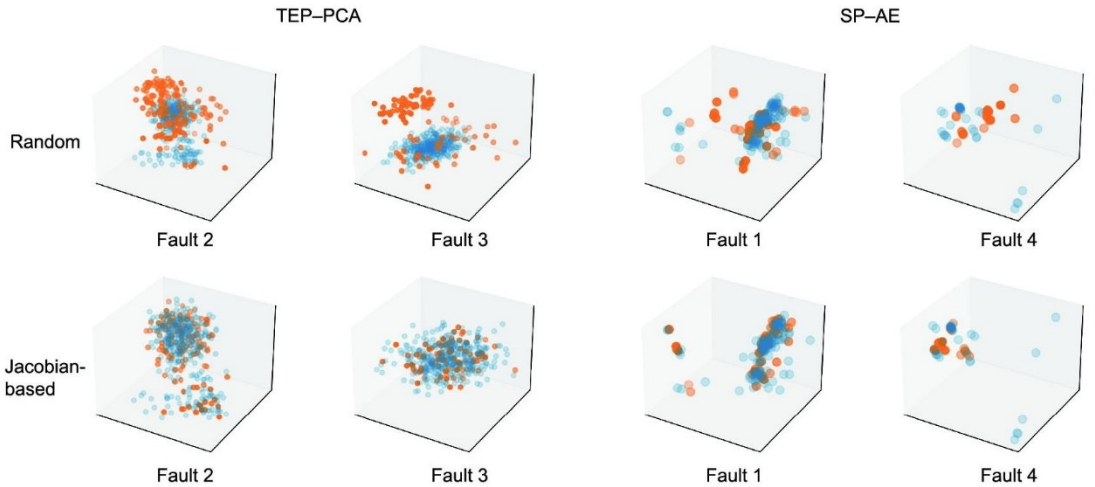

图5. 使用随机和基于雅可比变量搜索顺序的对抗样本可视化。蓝色点表示数据集的整体分布，橙色点表示生成的对抗。左半部分是来自PCA还原的TEP的故障数据，右半部分是具有AE还原的SP数据集。上行是使用随机搜索的对抗，下行是使用基于雅可比搜索的对抗。

表3 降维空间中对抗样本和原始样本之间的距离

Searching method	TEP–PCA	TEP–AE	SP–PCA	SP–AE
Random	0.1095	0.1725	0.1356	0.3521
Jacobian-based	0.0068	0.0702	0.0203	0.1342

先，设计了一组扰动极限值（$\varepsilon \in \{0.01, 0.02, \ldots, 0.19, 0.2, 0.3, \ldots, 1.0\}$ [方程（1）中的超参数）]，旨在测试对某个变量的最小扰动。由于在较小的扰动极限值下攻击成功率迅速上升，因此设计了ε小于0.2时的较小测试值间隔（0.01）。

每个变量相对于故障的平均最小扰动如图6所示，其中较小的扰动值表示这些变量的微小变化会使分类器错误地预测该故障样本，反之亦然。由于无法成功攻击故障样本的某些变量，因此图7也报道了故障变量的攻击成功率，表示扰动此变量会导致样本的部分漏洞。两个热力图是负相关的，说明每种故障类型的变量的脆弱性。由于最大扰动的成功率往往代表难以攻击的变量，因此在图8中绘制了在相对较小的扰动（0.1）限制下的攻击成功率，其中强调了易受攻击的变量。此外，图9计算了沿图6的行和列的平均最小扰动。

从图6~8和图9（a）的三个热图的行可以看出，故障7是TEP分类中最鲁棒的，而故障15、故障16和故障21更容易制作对抗样本。这与分类器对这些故障的置信度相关。DNN对故障7的置信度较高，而对易受攻击的故障类型的置信度却较低。这符合TEP测试集上DNN的混淆矩阵（见附录A中的图S1），其中对于易受攻击的故障也难以进行分类。

基于三个热力图的列和图9（b），变量17和变量48比其他变量更容易受到攻击，而变量3和变量26对扰动的抵抗力相对较强。这对应于每个变量的分类器损失梯度。如附录A中的图S2所示，从DNN的损失函数到变量17和变量48的梯度比其他梯度重要得多，意味着这些变量的扰动将显著影响分类器。

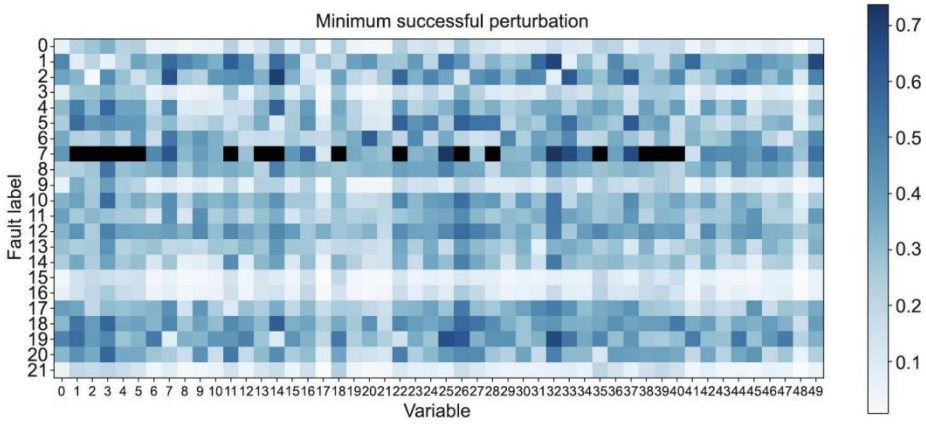

图6. 故障变量的平均最小成功扰动。颜色较深的点表示变量更难攻击。黑色点表示在最大扰动下甚至一次都无法攻击的变量。

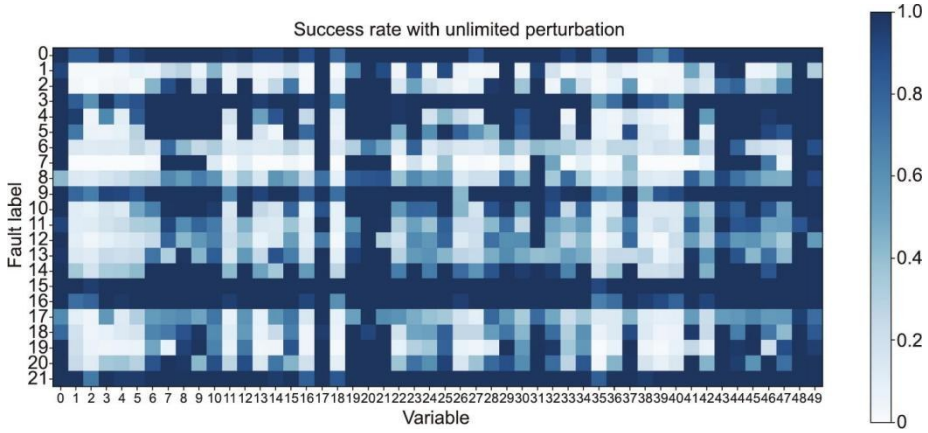

图7. 故障变量最大扰动的成功率。颜色较浅的点表示变量更难攻击。

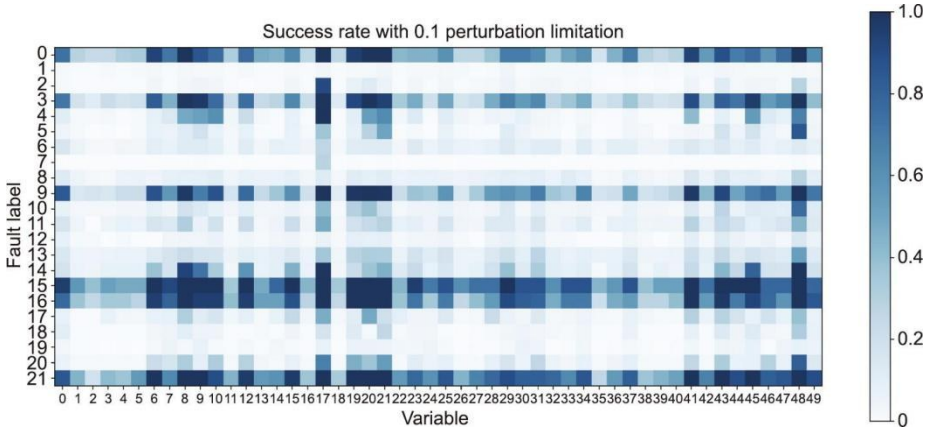

图8. 故障变量扰动为0.1的成功率。颜色较浅的点表示变量更难攻击。

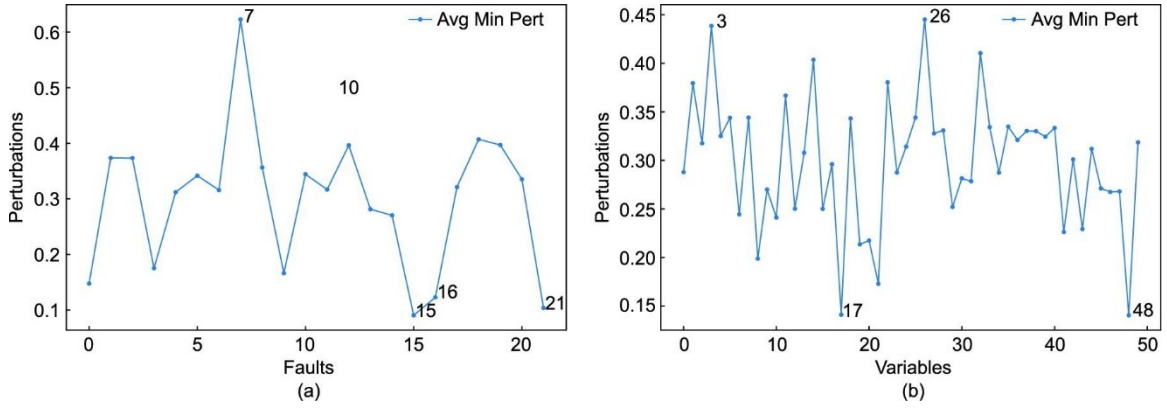

图9. 故障（a）和变量（b）的平均最小扰动，这是图6中行和列的平均值。

3.2. 故障对研究

本小节分析故障级别的漏洞。将故障对A-B定义为：故障A是真正的故障类型，故障B是受攻击分类器预测的错误类型。这意味着，给定扰动样本（故障A）和攻击，分类器将故障A错误地诊断为故障B。和上一节一样，实验是用TEP和DNN分类器进行的。

本文使用了最小成功扰动和成功率作为指标，其结果基于原始目标故障对。搜索所有可攻击故障样本的所有变量，并记录可以欺骗分类器的最小扰动ε，结果如图10（a）所示。对于被攻击样本的成功率，如果故障A可以在不同的变量上攻击故障B，则故障A的每个样本只计一次攻击成功次数。图10（b）表示扰动极限，用于显示易受攻击的故障对。变量攻击成功率统计了故障样本中每个变量被成功扰动的频率，如图第10（c）所示。与第3.1节一样，本文还计算了平均最小扰动，以便对原始故障对和目标故障对进行严格分析，如图10（d）所示。

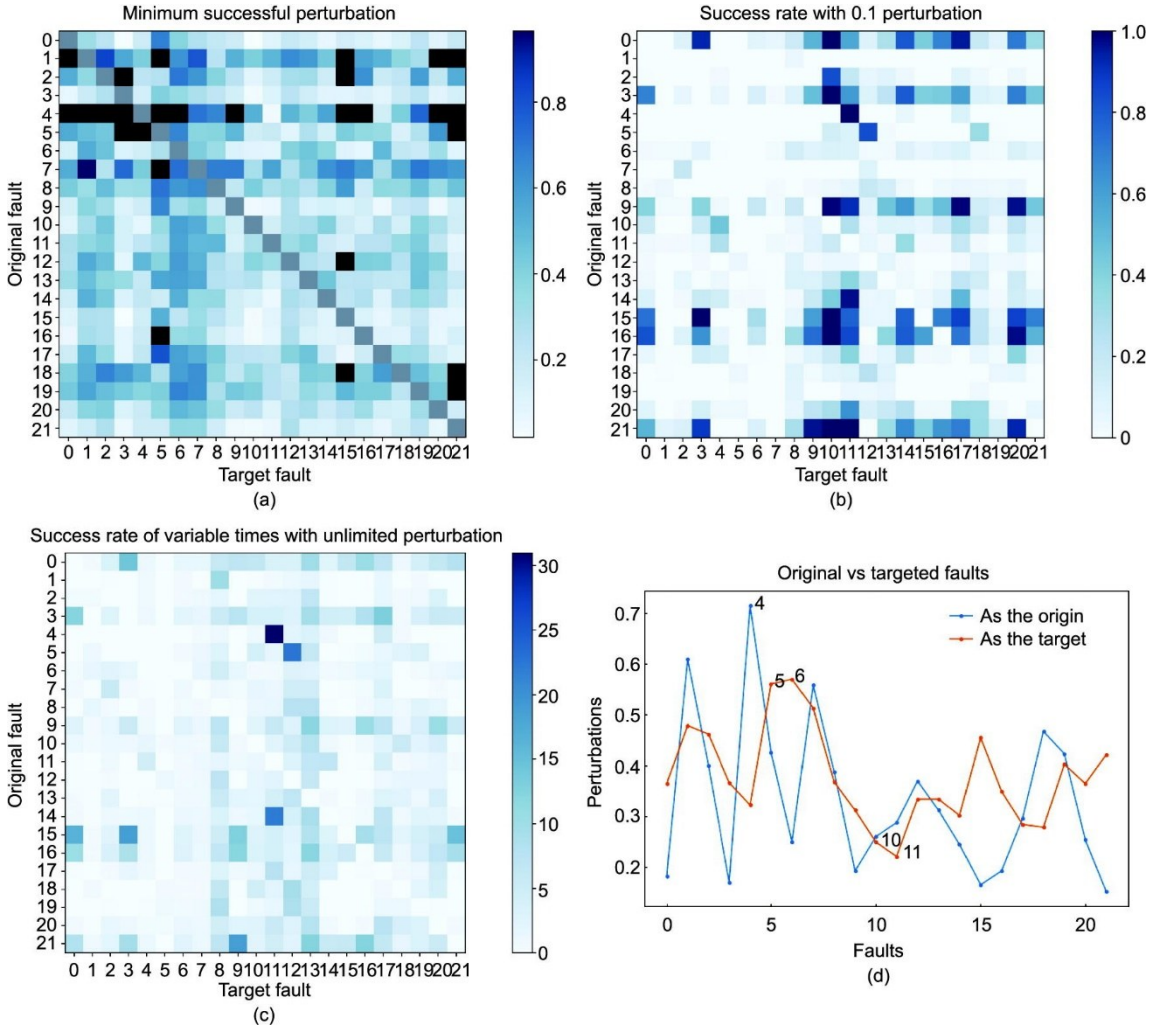

图10.（a）故障对的平均最小成功扰动。该扰动是原始目标故障对之间最小值的平均值ε。颜色较深的点表示故障对更难攻击。黑色点表示在最大扰动下甚至一次也无法攻击的故障对。（b）扰动为0.1的故障对攻击成功率。攻击成功率是原始目标故障对之间成功攻击的样本占原始故障样本总数的百分比。颜色较浅的点表示故障对更难攻击。（c）具有无限扰动的变量成功攻击率。攻击成功一个扰动变量计数一次。颜色较浅的点表示故障对更难攻击。（d）原始故障对和目标故障对的平均最小扰动，来自该图（a）中的行和列。

对于原始故障，在第3.1节针对变量分析了鲁棒性和脆弱性。从故障对的角度来看，结果非常相似，唯一的区别是故障4最难被错误地分类为其他大多数故障。

对于目标故障，故障10和故障11更容易成为目标，而故障5和故障6的目标较少。结合分类器混淆矩阵（见附录A中的图S1），可以找到一个有趣的联系：鲁棒故障和脆弱故障分别是测试集中分类精度最高和最低的故障。一种可能的几何解释是，对于这些脆弱故障，分类决策区域较小，并且它们更接近边界，而鲁棒故障远离边界，这使得它们更难受到攻击。

此外，本文还发现了一些有趣的模式。

（1）**不对称性**：直观地说，如果故障A很容易被攻击为故障B，那么故障B也应该容易被扰动为故障A，这意味着故障对A-B和B-A同样容易受到攻击，热力图应该围绕对角线对称。对称模式符合直觉预期，因为易受攻击的故障对在输入空间中通常距离更近。但实际上，三张热力图中只有少数故障对是对称的，存在大量的不对称模式，如原始-目标对21-20、21-17、5-12等。这种不对称性耐人寻味，表明某些故障（如故障21）类似于许多其他故障，如故障20和故障17，但反之则不然。

（2）**集中性**：两张热力图表明，与原始故障一样，故障4总体上是鲁棒的，它与大多数其他故障之间的转移是困难的。但是，根据上一节的结论，故障4的攻击成功率并不是很低，因此故障4的目标集中在一个故障上，即故障11。这在计算变量攻击成功率（即故障中可以攻击其他故障的变量百分比）时更明显，如图10（c）所示。故障对4-11的攻击成功的平均变量数为30.98，这意味着故障4的每个样本有近31个变量（总共50个）可以被扰动以使DNN分类器将故障4预测为故障11。可以注意到，对于原来的故障4，最大的目标故障11占总数的

68.55%。

对这两种模式最直接的解释可以在以下小节中从分类的几何角度给出。

3.3. 故障分类的几何解释

本小节使用几何方法解释单变量攻击的工作原理以及为什么故障和变量之间的相关性如前两小节所示。基于几何方法，还提供了对工业故障分类系统的更深入见解。

本小节的主要方法是绘制分类边界，以演示故障分类的几何特征。由于视觉维度的数量有限，分类边界的变化仅针对特定类型故障的两个变量绘制。由于其他变量的值会影响分类边界的形状，因此其余变量的值由该故障的平均值表示。

首先，对于变量0和变量46，绘制故障20和故障21的DNN决策边界，如图11所示。由于其他变量的值是通过均值近似的，因此故障点的位置存在一些偏移。事实上，分类器对故障样本进行了正确分类。比较这两个结果可以得出以下几点：

（1）为什么单个变量中的扰动可以成功攻击故障分类系统？如图11所示，分类器的输出沿一个变量变化，特别是对于故障21（如图11左边子图所示），水平和垂直扰动的样本都会使分类器错误地预测至少三种不同的故障类别。

（2）为什么不同的故障对同一个变量有不同的脆弱性？与故障21相比，故障20更容易被分类器分类，其输出置信度也更高，因此，分类区域更宽，样本离分类边界更远。另外，分类器对故障21的置信度较低，分类区域较窄，样本非常接近分类边界。这意味着只有轻微的扰动才能使分类器输出故障21样本的错误预测，而故障20的样本对扰动的鲁棒性更强。这与第3.1节中的热力图和混淆矩阵一致（见附录A中的图S1）。

（3）为什么故障对是不对称的？主要原因是分类的输入空间是高维的，故障样本只占整个空间的一小部分，而一些故障的分类面积很大，占据了很大的空间。这意味着，尽管不同故障之间的距离非常远，但在输入空间的某些区域中，这两种故障的分类边界可能是相邻的。在这两个图中，对于故障21的样本，它们接近这两个变量投影中的故障20决策区域。两种故障的分类边界在空间的某些部分相邻，但与故障20样本附近的空间部分不相邻。将z轴投影到故障11上，显示故障20的决策区域在该轴上扩展并挤压故障21、故障3、故障18和故障14的区域。故障20的样本聚集在z轴上的某个位置，因此在对变量0和变量46的投影中，故障20与故障21不相邻。实际上，z轴表示了其他变量。

接下来，考虑图12，试图解释为什么某些故障对非常脆弱。从故障对4-11的样本来看，在4个子图的8个变量上，故障4的决策区域与故障11紧密包围。因此，可以假设在高维输入空间中，故障4决策区域的很大一部分也被故障11包围。因此，故障4的对抗样本很有可能被错误地归类为故障11。

最后，从梯度角度研究了鲁棒变量和脆弱变量。针对变量2和变量48绘制故障5的样本分布，并在每个点计算分类损失作为这两个变量函数的梯度，如图13所示。DNN的损失值由输出和真实标签之间的交叉熵计算，交叉熵用于衡量两个概率分布之间的差异。就梯度值而言，变量48比变量2高一个数量级。在从样本开始的方向上，沿变量48方向的梯度变化明显大于变量2，沿变量2几乎没有梯度。这意味着对变量48应用扰动可以使分类器的输出发生更剧烈的变化，并且更容易获得分类的错误输出，这也验证了第3.1节中的结果。此外，结合梯度图和

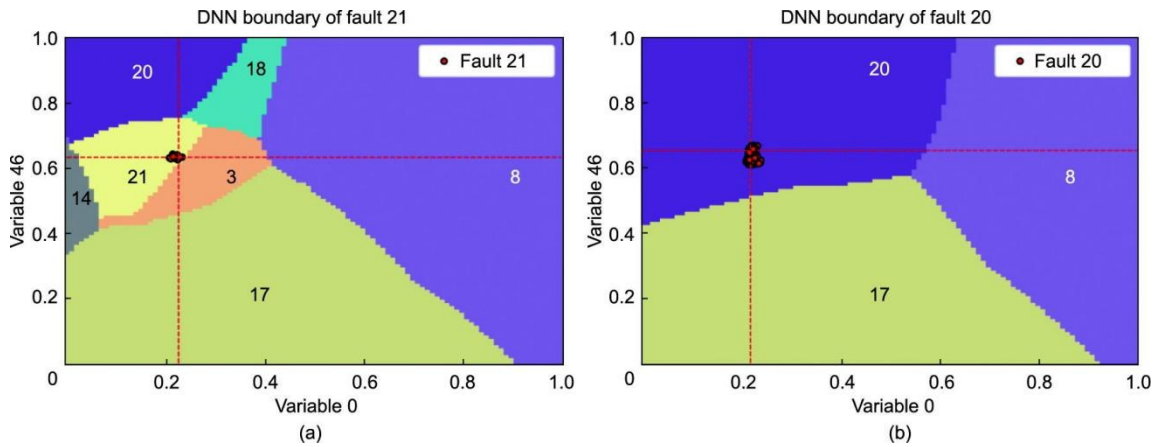

图11. 故障21和故障20的DNN分类边界与变量0和变量46的函数关系。红色点是某种故障类型的样本。不同的颜色表示不同故障的分类区域，其中数字表示其故障类型。

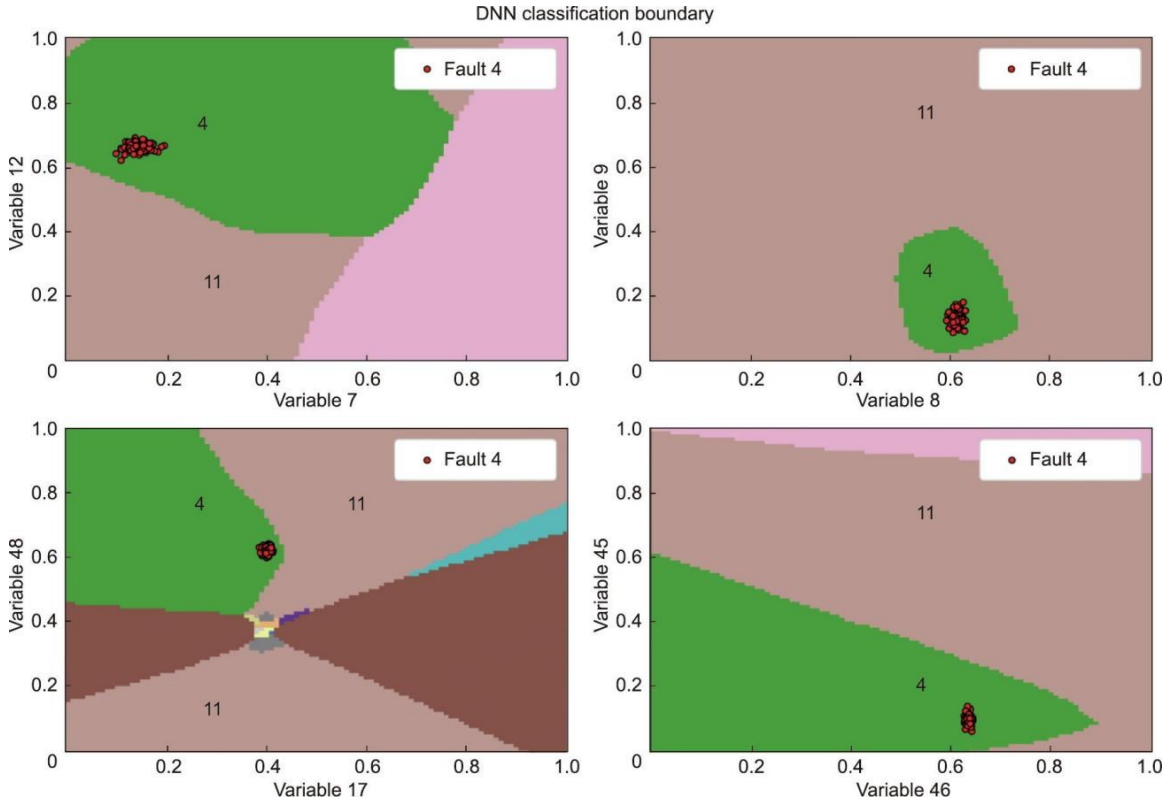

图12. 故障4的DNN分类边界，作为8个变量的函数。

分类边界可以发现，故障越接近分类边界，梯度就越高。这意味着决策边界附近的样本更容易受到攻击，反之亦然，这与第3.2节中的几何解释一致。

4. 单变量防御

为了防御单变量攻击生成的对抗样本并提高DNN的鲁棒性，本文提出了一种对抗训练方法，在DNN训练过程中添加对抗样本，以训练更鲁棒的分类器。与大多数现有的对抗训练方法（大多数现有的对抗训练方法都会扰动每个变量以获得对抗训练样本）不同，本文中的方法只会将扰动添加到梯度较高的变量中。对抗训练方法仅适用于迭代训练模型，如DNN，因此本文中的防御方法不适用于SVM和*k*NN。

为了实现这一点，在训练过程中计算每个故障变量的平均梯度的梯度表，这是一个类似于附录A中图S2的表。仅选择表中排名前K的故障变量进行扰动以生成对抗样本，其中K是可调超参数，用于控制对抗训练期间的扰动变量。符号函数用于确定所选变量的方向，沿上升梯度方向施加扰动。

通过在模型训练期间添加与原始样本具有相同标签的对抗样本，可以获得更鲁棒的分类器。图14显示了三个分类器的单变量攻击成功率，即原始DNN，具有FGSM训练的DNN和采用所提出方法（单变量攻击训练）的

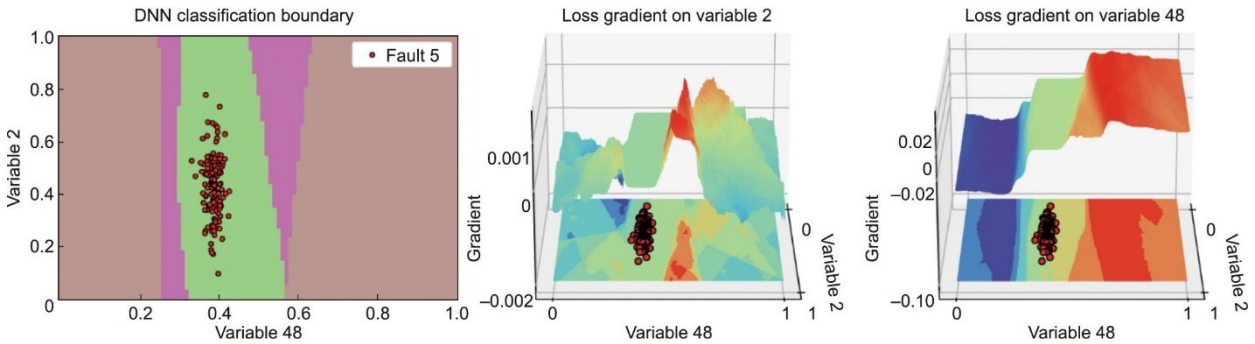

图13. 故障5的DNN分类边界，作为变量2和变量48的函数，以及相应的梯度。梯度从分类器故障5的交叉熵损失分别到变量2和变量48，并绘制梯度等高图的三维和二维投影。

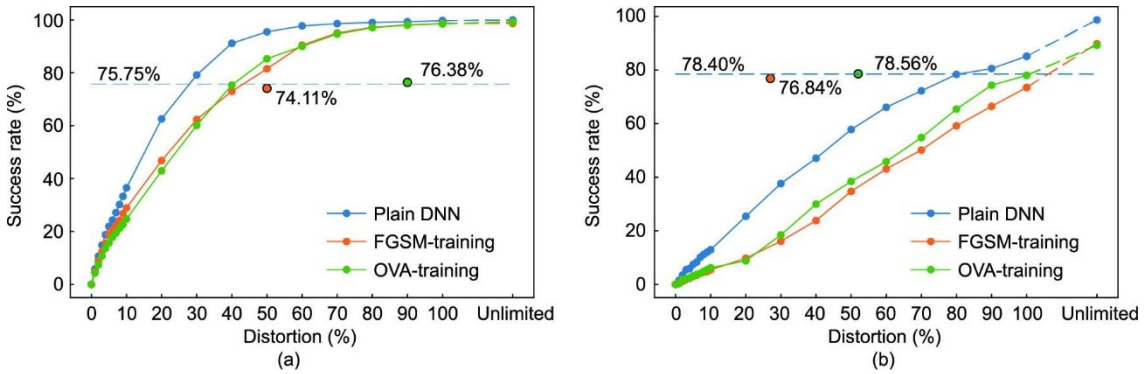

图14. 不同训练方法的成功率和测试准确性。

DNN。在实验中，将K设置为50%，意味着一半梯度的故障变量在所提出的方法中受到扰动。

两个数据集上的结果表明，所提方法能够有效降低单变量攻击的成功率。然而，与FGSM相比，当失真增加时，单变量攻击训练并不那么鲁棒，但在测试集上获得了很高的准确度，比普通DNN模型高0.63%（SP中为0.16%），比具有FGSM对抗训练的DNN高2.27%（或1.76%）。这主要是因为跨不同故障的全局变量选择。向每个变量添加扰动会降低攻击成功率，但作为权衡，高置信变量中的一些扰动会降低测试精度。本文中的方法只在难以分类的变量上添加扰动，以帮助模型学习这些变量的更明确的边界，从而提高测试集的准确性。同时，由于攻击方法仅针对一个变量，因此在对抗训练期间减少扰动变量不会降低鲁棒性。

5. 结论

本文研究了工业故障分类系统的安全性，提出了单变量攻击在仅扰动单个变量的条件下攻击故障分类模型。结果表明，仅扰动一个变量就足以攻击工业故障分类器。即使扰动限制在很小的值，攻击成功率也很高。在TEP中，单个变量的10%（或20%）失真干扰了36.5%（或62.6%）的样本，从而成功攻击DNN分类系统。

利用单变量攻击方法，本文还探索了由DNN代表的工业故障分类模型的几何特征。绘制了分类边界和梯度图，以深入了解工业故障分类系统的脆弱性和鲁棒性。最后，为了最小化对抗攻击的影响，提出了一种利用部分变量的对抗训练方法。展示了在对抗鲁棒性的小幅下降之间的权衡，以提供更高的预测准确性（比没有对抗训练的DNN高0.63%，比所有变量对抗训练的DNN高2.27%）。

致谢

本工作由国家自然科学基金（92167106、62103362、61833014）和浙江省自然科学基金（LR18F030001）资助。

Compliance with ethics guidelines

Yue Zhuo, Yuri A.W. Shardt, and Zhiqiang Ge declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2021.07.033.

References

[1] Ge Z. Semi-supervised data modeling and analytics in the process industry: current research status and challenges. IFAC J Syst Control 2021;16:100150.
[2] Ge Z, Song Z, Ding SX, Huang B. Data mining and analytics in the process industry: the role of machine learning. IEEE Access 2017;5:20590–616.
[3] Dash PK, Samantaray SR, Panda G. Fault classification and section identification of an advanced series-compensated transmission line using support vector machine. IEEE Trans Power Deliv 2007;22(1):67–73.
[4] Chen X, Ge Z. Switching LDS-based approach for process fault detection and classification. Chemom Intell Lab Syst 2015;146(C):169–78.
[5] Wang Y, Wu D, Yuan X. LDA-based deep transfer learning for fault diagnosis in industrial chemical processes. Comput Chem Eng 2020;140:106964.
[6] Chen G, Ge Z. SVM-tree and SVM-forest algorithms for imbalanced fault classification in industrial processes. IFAC J Syst Control 2019;8:100052.
[7] Zhao D, Wang T, Chu F. Deep convolutional neural network based planet bearing fault classification. Comput Ind 2019;107:59–66.
[8] Chadha GS, Panambilly A, Schwung A, Ding SX. Bidirectional deep recurrent neural networks for process fault classification. ISA Trans 2020;106:330–42.
[9] Jiang L, Ge Z, Song Z. Semi-supervised fault classification based on dynamic sparse stacked auto-encoders model. Chemom Intell Lab Syst 2017;168:72–83.
[10] Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. Engineering 2020;6(3):346–60.
[11] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 2018;6:14410–30.

[12] Xu H, Ma Y, Liu H, Deb D, Liu H, Tang J, et al. Adversarial attacks and defenses in images, graphs and text: a review. Int J Autom Comput 2020;17(2): 151–78.
[13] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations; 2014 Apr 14–16; Banff, Canada; 2014.
[14] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA; 2015.
[15] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, Canada; 2018.
[16] Shafahi A, Najibi M, Ghiasi MA, Xu Z, Dickerson J, Studer C, et al. Adversarial training for free! In: Proceedings of Advances in Neural Information Processing Systems 32; 2019 Dec 8–14; Vancouver, Canada; 2019.
[17] Zhang D, Zhang T, Lu Y, Zhu Z, Dong B. You only propagate once: accelerating adversarial training via maximal principle. In: Proceedings of Advances in Neural Information Processing Systems 32; 2019 Dec 8–14; Vancouver, Canada; 2019.
[18] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput 2019;23(5):828–41.
[19] Papernot N, McDaniel PD, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 1st IEEE European Symposium on Security and Privacy; 2016 Mar 21–24; Saarbrücken, Germany; 2016.
[20] Barreno M, Nelson B, Joseph AD, Tygar JD. The security of machine learning. Mach Learn 2010;81(2):121–48.
[21] Biggio B, Corona I, Maiorca D, Nelson B, Srndic N, Laskov P, et al. Evasion attacks against machine learning at test time. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Heidelberg: Springer, Berlin; 2013. p. 387–402.
[22] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. 2017. arXiv:1702.05983.
[23] Sankaranarayanan S, Jain A, Chellappa R, Lim SN. Regularizing deep networks using efficient layerwise adversarial training. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[24] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015); 2015 May 7–9; San Diego, CA, USA; 2015.
[25] Papernot N, Mcdaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security; 2017 Apr 2–6; Abu Dhabi, United Arab Emirates. New York: Association for Computing Machinery; 2017. p. 506–519.
[26] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Jun 18–23; Salt Lake City, UT, USA; 2018.
[27] Shang C, You F. Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering 2019;5(6):1010–6.
[28] Chen Y. Integrated and intelligent manufacturing: perspectives and enablers. Engineering 2017;3(5):588–95.
[29] Yi TH, Huang HB, Li HN. Development of sensor validation methodologies for structural health monitoring: a comprehensive review. Measurement 2017;109: 200–14.
[30] Downs JJ, Vogel EF. A plant-wide industrial process control problem. Comput Chem Eng 1993;17(3):245–55.
[31] Research center of sciences of communication [Internet]. Rome: Semeion Communication Science Research Centre; 2022 April 19 [cited 2022 Apr 30]. Available from: https://www.semeion.it.