TwinExplainer：解释汽车数字孪生的预测

Subash Neupane计算机科学与工程 密西西比州立大学美国密西西比州斯塔克维尔
sn922@msstate.edu

Ivan A.Fernandez计算机科学与工程 密西西比州立大学美国密西西比州斯塔克维尔
iaf28@msstate.edu

Wilson Patterson计算机科学与工程 密西西比州立大学美国密西西比州斯塔克维尔
wep104@msstate.edu

Sudip Mittal计算机科学与工程 密西西比州立大学美国密西西比州斯塔克维尔
mittal@cse.msstate.edu

Milan Parmar计算机科学与工程 密西西比州立大学美国密西西比州斯塔克维尔
parmar@cse.msstate.edu

Shahram Rahimi计算机科学与工程系 密西西比州立大学美国密西西比州斯塔克维尔
rahimi@cse.msstate.edu

arXiv:2302.00152v1 [cs.LG] 1 Feb 2023

摘要—**车辆是复杂的信息物理系统（CPS），运行于各种环境中；一个或多个子系统（如发动机、变速器、制动器和燃油系统）发生失效的可能性，可能导致非计划停机，并产生高昂的维护或维修成本。为避免这些问题，持续监测各子系统的健康状况并识别异常的传感器通道行为至关重要。数据驱动的数字孪生（DT）系统能够完成此类任务。当前的数字孪生技术采用各种深度学习（DL）技术，但这些技术无法为其预测提供依据或解释。此类不透明系统的这一局限会影响决策，并引发对用户信任的担忧。本文提出一种解决方案：TwinExplainer系统通过其三层架构流程解释汽车数字孪生的预测。该系统可帮助汽车行业利益相关者从全局角度理解传感器通道，以及这些通道如何影响数字孪生的通用预测。TwinExplainer还可将数字孪生计算得出的正常和异常局部预测的解释可视化。**

索引词—**数字孪生、可解释性、车辆维护、深度学习**

I. 引言

现代车辆配备了各种传感器，可视为更广泛的信息物理系统（CPS）类别的一个子集，而整个类别会产生大量运行数据。借助当前的中央处理器和图形处理器，我们如今拥有了前所未有的计算能力。这些庞大的车辆数据集可以与运行于这些硬件上的人工智能（AI）和机器学习（ML）技术结合使用，以构建数据驱动的智能系统 [1]。数字孪生（DT）就是此类智能系统之一。顾名思义，数字孪生（DT）被定义为“对复杂产品进行多物理场、多尺度概率模拟的集成体，其作用是映照对应孪生体的生命周期” [2]。这项技术的主要目标是在数字世界中充当物理实体的全景映像。这具有许多内在优势，因为物理对象可以根据数字孪生提供的输入进行调整，以改变其实时行为。反过来，这种连接使模拟能够精确再现物理实体的现实世界状态。

为了实现完整的信息物理系统，数字孪生必须与物理模型动态耦合，这可以通过实时传感数据来实现。简单来说，数字孪生是物理对象的实时（虚拟）模型，它利用实时传感器数据来模拟行为、监控活动，并预测物理资产的未来状态。数字孪生技术运用机器学习算法，分析并识别海量传感器数据中的数据模式和传感器通道活动。人工智能和机器学习能够从数据中提取有用的见解，以优化性能、维护部件、减少排放并提高效率 [3]。

采用深度学习模型的一大缺点是，它们被视为具有不透明决策算法的黑箱。由于其嵌套式非线性结构，这些系统构成了最先进水平中的绝大多数，能够实现出色的预测准确率。然而，它们缺乏对其预测结果的解释 [4]。相比之下，不透明模型无法识别并解释数据中驱动其预测的确切信息，这会引发用户信任方面的问题 [5]。因此，若要在汽车领域中使用这些系统，就必须能够在一定程度上解释其决策过程。

为了解决这些问题，并秉持开发对汽车行业利益相关者而言具有可解释性的数字孪生的理念，我们提出了 *TwinExplainer*，它利用 SHapley 加性解释（SHAP） [6]这一可解释算法来近似并解释数字孪生的预测。目前，对于黑箱可解释性问题尚无定论。尽管如此，为不透明模型提供解释的各种可能方案已相继出现，它们探索并利用了机器学习过程的不同方面

为不透明模型 [7] 提供解释。一种基于特征的方法（例如 SHAP ）是这些可能方案之一，它以特征作为解释媒介。该方法旨在确定每个特征对输出预测的贡献程度。

本文扩展了我们此前的工作 [1] ，该工作旨在预测汽车系统中各个功能工作组（FWG）（亦称子系统）的异常传感器通道行为，并将发动机、变速器、燃油和制动子系统等四个主要 FWG 的组合视为一个数字孪生系统。在这一特定数字孪生系统中，我们采用时间卷积网络（ TCN）作为深度学习分类器，以监测和预测传感器通道的行为。本实验使用的数据集是车辆性能、可靠性与运行（VePRO）数据集，由美国陆军工程兵团和密西西比州立大学（MSU）创建，其中包含军用车辆车队的运行数据，例如时间序列传感器测量数据、检测到的故障以及维护报告。本文采用基于特征的方法来解释 TCN 模型所做的预测。

本文的主要贡献如下：

- 我们创建了一个名为*TwinExplainer*的数据驱动框架，用于从汽车数字孪生（DT）中生成解释。
- 通常，维修技师、操作人员等无法理解采用深度学习架构的预测型数字孪生模型所做出的预测。我们证明，可以为此类模型生成解释，进而增强用户信任并辅助决策。
- TwinExplainer 创建可视化内容，以解释数字孪生所做出的预测。我们聚焦于特征，通过展示特定预测（局部解释）以及整个系统（全局解释）中各特征的归因和重要性来解释预测。

本文其余部分安排如下：第 II 节介绍背景和相关工作。第 III 节介绍*TwinExplainer* 架构。第 IV 节讨论我们的实验和方法。第 V 节介绍如何生成解释。最后，第 VI 节总结全文。

II. 背景与相关工作

本节介绍数字孪生（DT）、预测性维护（PdM）、可解释人工智能（XAI）以及可解释数字孪生（X-DT）方面的一些相关工作。

A. 数字孪生（*DT*）

使用“物理孪生体”这一数字孪生（DT）的前身并非新概念。该概念最早于 20 世纪 70 年代在 NASA 阿波罗计划期间出现。Grieves 等人于 2002年 非正式地提出了数字孪生这一概念，随后在其白皮书 [8] 中对其进行了正式阐述。从本质上说，数字孪生是一个物理资产的虚拟原型，可模拟实时运行条件，从而表现得如同真实的物理资产。

数字孪生技术已惠及医疗保健 [9], [10], 航空 [11]–[13], 制造 [1],[14], [15], 信息物理系统 [16], [17], 安全 [18], 等众多行业，并提供了多种应用。在汽车行业，数字孪生的许多功能都得到了应用，尤其是在车辆开发、运行和维护方面。制造商已成功将数字孪生集成到车辆开发过程中，从而降低开发成本。开发人员现在可以创建大量数字原型，在车辆实际制造出来之前，对其进行无数场景的模拟 [19]。数字孪生在车辆运行寿命期间还有许多其他车辆应用机会。安装在车辆上的传感器能够采集运行数据，并在车辆整个生命周期内更新其数字孪生。采集的数据会反馈给正在道路上行驶的车辆，从而实现实时诊断和评估。

B. 预测性维护（*PdM*）

维护是确保系统按预期模式运行的关键过程，无论是通过解决问题还是采取预防措施 [20]。文献中将维护策略大致分为三类，例如反应式维护（RM） [21]、预防性维护（PM） [21], [22],和预测性维护（PdM）。第一种策略也称为纠正性维护，是一种“运行至失效”的方法，旨在系统发生故障后对其进行维修。相比之下，PM 通常称为计划性维护，通过按固定时间间隔规划维护作业来降低故障发生的可能性。第三类 PdM 也称为基于状态的监测（CBM） [23]，旨在依据系统的健康状况、运行和维护数据，预测设备可能发生故障的时间。

PdM 利用系统和组件中嵌入式传感器经常采集的数据，通过预测算法进行预测分析。这些模型能够有效发现运行数据中的行为趋势和模式；这一任务与对系统正常行为进行建模密切相关，并可检测出偏差，即异常，这些异常可能表明当前或即将发生的故障 [20], [24]。异常是错误、故障和失效分类体系中的一种偏差或潜在错误，其中错误由故障引起，并可能导致失效。识别这些错误的过程称为“异常检测”，因此可用于 PdM 任务 [20]。在以下小节中，我们将讨论可解释方法。

C. 可解释人工智能（*XAI*）

近年来，人们对人工智能的可解释性和互操作性的关注显著增加，这些领域也被称为“可解释人工智能”（ XAI）。深度学习模型在经验上的成功催生了

不透明的决策系统，这些系统对于最终用户而言往往是复杂的黑箱模型。支持模型输出的解释对于向人工智能利益相关者提供有关预测的透明度至关重要 [7], [25]。解释是指智能系统为其决策提供依据的方式。可解释性的范围和模型依赖性是有关可解释性分类体系的文献中常见的类别 [7]。以下小节将对这些类别进行详细阐述。

1）局部可解释性：局部层面的可解释性是指系统向用户说明特定决策背后推理过程的能力。局部可解释模型无关解释（LIME） [26], 、Anchors [27] 和逐一剔除协变量（LOCO） [28] 都是局部解释技术的示例。LIME 使用局部代理模型来解释单个预测。另一个重要的解释器模块是 Lundberg 等人提出的 Shapley 加性解释（SHAP） [6]。它用于确定每个特征对每项预测的重要性。

2）全局可解释性：模型的全局可解释性使人们更容易理解所有潜在结果背后的逻辑。这些模型有助于理解各种输入数据的归因，并阐明模型整体的决策过程。全局解释器模块的示例包括 Submodular Pick-LIME（ SP-LIME）（LIME 模块的扩展）、概念激活向量（ CAV） [29], 、通过递归划分进行全局解释（GIRP） [30], 等。

3）模型无关：模型无关的方法不受任何特定类别的机器学习模型限制。因此，从模块化的角度来看，解释模块独立于其生成解释所针对的模型。可视化、知识提取、影响方法和基于示例的解释，都是模型无关可解释性技术的示例 [31]。

4）模型特定：模型特定的解释器方法专为某一特定模型类型设计。例如，模型特定的解释包括对线性模型中回归权重的解释。从这个意义上说，对回归、基于规则的模型和决策树等内在可解释模型的解释始终是模型特定的。

D. 可解释数字孪生（*X-DT*）

在由人工智能/机器学习/深度学习分类器赋能的数字孪生背景下，决策可能是自主推理过程的产物，因此可能需要解释。解释分类器的预测将有助于不同的利益相关者（如开发人员、操作人员、维修技师等）做出决策，因为这使他们能够理解各种计算所得决策的情境依据 [25]。

已有大量研究探讨了数字孪生的可解释决策。[32] 提出了一种风险诊断数字孪生系统，利用 XAI 改进肝脏疾病的决策。为此，

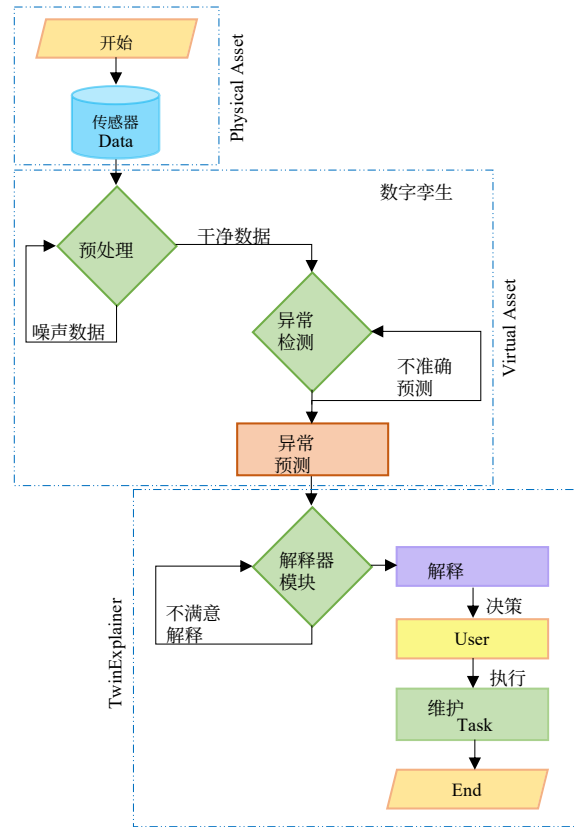

图 1：*TwinExplainer* 模型的流程图。车辆等物理资产收集的传感器数据构成模型的初始输入。对原始传感器数据进行预处理后，将其输入异常检测数字孪生模型，以预测未来的时间序列。解释器模块接收预测结果并生成解释。用户可根据对预测结果的解释，决定是否安排维护。

作者使用最先进的局部可解释模型无关解释（LIME）方法，近似随机森林分类器所做出的决策。文献 [33]提供了相关解释，说明如何根据无人驾驶飞行器（UAV）的传感器测量数据，从组件化降阶模型库中动态选择数字孪生模型。研究人员考察了飞机机翼上多个不同严重程度的受损区域。另一项研究中，Zhang 等人 [25] 针对动态数据驱动数字孪生的可解释性问题，采用人在回路方法；作者提出了一种架构，用于确定解释在何时何地能够支持数据驱动或系统计算的决策。类似地，Wang 等人 [34]在论文中提出了一个可解释建模框架，以促进数字孪生生态系统的开发人员与其他利益相关者之间的协作和互动。作者以数字孪生工厂用例说明，如何使用三种可解释性评分（基于模型、基于场景和目标导向）来衡量易于理解的建模所具有的价值。

接下来，我们将介绍*TwinExplainer* 系统的数据流和架构。

III. TWINEXPLAINER 架构

本节介绍一种集成了 XAI 的深度学习数字孪生系统架构，称为*TwinExplainer*。大多数与车辆数字孪生相关的研究都侧重于提高预测准确率，却忽视了此类预测结果的可解释性。本文针对数字孪生系统预测结果的可解释性问题展开研究。*TwinExplainer* 系统旨在通过各种可视化技术和增强人工智能系统中的用户信任，帮助汽车维修技师理解数字孪生系统计算得出的预测决策。

图 1 总结了*TwinExplainer*系统中的数据流。首先对传感器数据进行预处理，以去除噪声、不完整数据和冗余观测。数据还会经过归一化和转换。此过程会反复进行，直至获得干净的数据子集。随后，将高质量的清理后数据输入数字孪生模块，该模块会预测传感器通道读数，并标记异常数据实例。如果数字孪生系统的异常预测有误，则会重复此过程。之后，预测结果会输入解释器模块，由该模块生成解释。如果提供的解释不尽如人意，则会重复此过程，以满足不同利益相关者的可解释性需求。例如，一组解释对于汽车操作员来说可能过于复杂。他们可能会要求提供更多易于理解的解释，或能帮助他们深入了解相关主题的资料。在这种情况下，开发人员可以修改解释器模块，以满足用户的需求 [7]。最后，用户可根据*TwinExplainer* 提供的解释，就故障诊断、故障预测和维修作出明智的决策，从而在一定程度上降低部件突然故障的风险。

图 2 展示了*TwinExplainer*架构的三个阶段：物理资产、虚拟资产和孪生解释器。这些阶段共同构成了我们的解释框架。本节其余部分将详细介绍各个阶段，并阐述我们采用的技术。

A. 物理资产

现实世界属性（如物体、设备、系统和组件）构成了物理资产。在车辆系统中，传感器和执行器是相互耦合的两种主要物理组件。传感器检测物理状态，并根据检测和测量到的特征向电子控制单元（ECU）传输电子信号。另一方面，执行器是主要负责移动和调节机构以完成任务的物理组件。例如，执行器可以根据传感器信号调整发动机怠速和凸轮轴气门的开闭，以优化发动机性能和燃油经济性。

现代汽车配备了多个传感器，可生成大量运行数据。这些数据以通道形式表示，例如车速、发动机转速、发动机扭矩百分比、制动开关、加速踏板位置、燃油率、发动机冷却液温度、变速器油温、所选挡位、喷油器控制压力等。通常，这些数据以时间序列格式采集。

在此阶段，会收集来自多个嵌入式传感器的多通道传感器数据，并将其存储在数据存储库中。利用这些数据集，可以构建智能系统，例如配备深度学习分类器的数字孪生（DT），以调查、分析和监控子系统或整个汽车系统。例如，可以为燃油子系统开发数字孪生，以预见并预测燃油系统相关通道的行为。这有助于利益相关者制定合理的维护计划、维修方案、故障诊断和故障预测策略，从而降低物理世界中发生意外运行中断的可能性 [35]。

B. 虚拟资产

在汽车行业，数字孪生通常用于完整或部分构建虚拟车辆模型。为了对数字孪生系统建模并评估车辆的实际性能，必须收集有关其运行、行为和功能的数据。*TwinExplainer* 采用车辆性能、可靠性与运行（VePRO）数据集，[1] 该数据集包含来自一支军用车辆车队的运行传感器数据，用于构建车辆数字孪生（有关数字孪生建模的更多信息，请参见第 IV-B 节）。有关该数据集的更多详情，请参见第 IV-A 节；表 I 汇总了 VePRO 数据集中包含的不同通道。

此阶段的第一步是数据处理，我们会移除不准确、不完整和重复的观测数据。随后对数据进行缩放和转换，以生成高质量的数据子集。接着将该数据集输入基于深度学习的异常检测模型。作为预测的一部分，该模型会对相连的传感器通道进行时间序列预测。然后比较预测观测数据与实际观测数据；如果实际观测数据超过预定义的阈值，系统就会将该观测数据标记为异常。本文采用的异常检测系统与我们在先前工作 [1] 中创建的系统相同。

C. TwinExplainer

了解汽车领域中传感器通道异常行为的“根本原因”，对利益相关者至关重要。最终，正是他们要及时决定安排维护、维修、诊断、故障检测和补救措施，这些措施可能有助于避免发动机或其他子系统即将发生的故障，并降低维护成本。因此，要让数字孪生对用户有用，关键在于将检测到或预测出的结果转化为人类易于理解的形式 [36]。然而，文献中通常会忽略这一阶段，而且大多数数字孪生

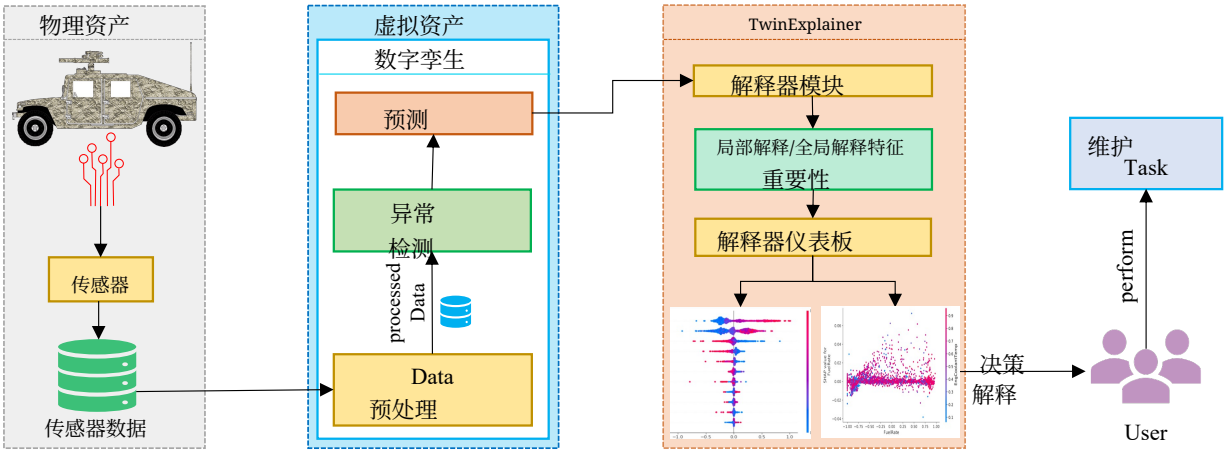

图 2：TwinExpainer 三阶段架构的图示。在物理资产阶段，系统收集并存储原始传感器数据。随后，在数字孪生阶段对数据进行清理、转换和缩放。此阶段会生成高质量数据，并将其传递给数字孪生模块，由该模块预测异常数据模式。在第三阶段，解释器模块以预测结果为输入，并利用可视化技术生成解释，帮助用户理解数字孪生为何做出特定预测。

系统更注重预测准确率，而非模型可解释性。

目前，机器学习/深度学习方法因能够实现前所未有的检测率，而被广泛用于构建检测（故障、异常、入侵等）系统。然而，尽管这些非透明模型的预测准确率令人瞩目，它们却无法说明其预测依据 [7]。这是因为机器学习/深度学习模型具有复杂的函数和非线性结构，因此难以确定影响其决策的具体数据信息 [4]。用户不了解不透明人工智能模型的内部工作机制，或无法将输出追溯至源数据，这会引发信任问题。 [5]。

在本文中，我们通过解释深度学习赋能的数字孪生系统所做出的预测，准确地解决了这一问题。数字孪生系统识别出异常后，维修技师或监督人员可能无法迅速确定导致该异常的底层传感器。为了解决这一问题，*TwinExplainer* 系统会检查特征归因，这些归因反映了模型中的每个特征对各个实例的预测所作贡献的程度。接下来，我们将讨论孪生解释器。

此阶段的输入是虚拟资产或物理系统克隆体所做出的预测。随后，这些预测会被提供给解释器模块，例如 LIME [26]和 SHAP [6]。解释器模块接着会模拟复杂的深度学习模块，并评估各项预测的特征归因。系统会评估局部特征和全局特征的重要性。解释器模块解释单项预测或决策的能力称为局部特征解释。它用于针对数字孪生所做出的特定决策生成独特的解释或理由。另一方面，全局特征解释 指的是解释器模块解释所有可能结果背后原因的能力

。从这个意义上说，此类解释有助于阐明模型整体的决策过程，从而理解不同输入数据的归因，并增强用户对系统的信任。

解释器模块旨在为最终用户生成系统的推理依据。最终用户依赖数字孪生所做出的预测和建议，或依赖其采取的行动，以便做出有把握的决策，并安排维护活动、维修、故障诊断与补救，以及风险缓解。例如，监控数据驱动车辆系统的操作员必须理解系统的决策模型，才能及时有效地安排故障排除、维修和维护活动，从而降低突发失效的风险。

接下来，我们将在第IV节介绍实验设置，并说明如何在第V节使用*TwinExplainer* 为汽车数字孪生计算出的预测生成解释。

IV. 实验设置

本节介绍用于生成解释的实验设置。在描述实验中使用的数据集之后，我们将介绍数字孪生建模方法。

A. 车辆性能、可靠性与运行（*VePRO*）

本研究采用了车辆性能、可靠性与运行数据集（Vehicle Performance, Reliability, and Operations dataset）。VePRO 项目由美国（U.S.）陆军工程兵团与密西西比州立大学（MSU）合作开展。有关该数据集的详细规格，请参阅我们先前的研究 [1]。该数据集收集了匿名化军用车辆车队的运行数据，且每

FWG	传感器通道	全名	Unit	描述
Time	UTC 1赫兹	1赫兹数据	秒	来自 1赫兹通道的时间序列数据。
发动机	发动机冷却液温度	发动机 冷却液温度	华氏度（°F）	反映发动机的工作负荷大小，并且与燃油率密切相关与燃油率。
	发动机负荷百分比	百分比 发动机 Load	百分比（%）	发动机行驶所需的负载量。
	发动机扭矩百分比	发动机 百分比 扭矩	百分比（%）	发动机的计算输出扭矩。
	增压压力	涡轮增压压力-sure	磅/平方英寸（磅/平方英寸）	表示任意时刻的气压信息。
	加速踏板位置	加速踏板 踏板位置	百分比（%）	涵盖加速、减速和稳态。
	进气歧管温度	进气 歧管温度	华氏度（°F）	表示进气歧管内空气的温度。
变速器	发动机车速	车速	英里每小时（m/s）	车辆速度——燃油消耗的主要影响因素。
	变速器油温	变速器 Oil 温度	华氏度（°F）	记录所有运行条件下的温度。
	变速器选定挡位	变速器所选挡位	无量纲	表示车辆当前所处的变速器挡位。
	TransTorqConv LockupEngaged	变速器扭矩锁止接合 转换器	基数为 10 的整数数值-ber	记录液力变矩器锁止是否已接合。
	变速器输出轴转速	变速器 Out-出轴转速	每分钟转数（转/分钟）	计算使用中的变速器传动比。
Fuel	燃油流量	燃油率	加仑（美国） Per 小时（加仑/小时）	燃油率本质上是车辆消耗燃油的速度。
	瞬时燃油经济性	瞬时 Fuel 经济性	加仑（美国） Per 小时（加仑/小时）	记录当前车辆速度下的当前燃油经济性。
	喷油控制压力	喷油器 控制压力	磅/平方英寸（磅/平方英寸）	可用作喷油器的健康状况指标。
制动器	制动开关	制动开关	基数为 10 的整数数值-ber	记录制动踏板的位置。

表 I：概述了研究中纳入的车辆 FWG，以及相关传感器通道的说明[1]。

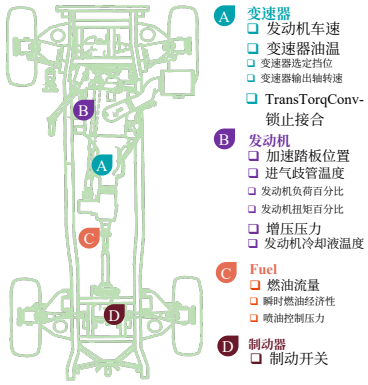

图 3：车辆中发动机、变速器、燃油和制动子系统及其传感器通道的示意图。根据参与同一数据集研究的机械工程师的建议，共选取了 15 个合适的传感器通道 [1], [37]。

辆车至少包含 106 个传感器通道。

我们利用与车辆系统主要部件（如发动机、变速器、燃油系统和制动器）相关的传感器通道来构建数字孪生。根据在同一数据集上开展工作的机械工程师的建议，我们选取了涉及四个子系统的 15 个传感器通道（相关子系统及其通道见图 3 示意图）。对于发动机子系统，我们使用发动机冷却液温度、发动机负载百分比、涡轮增压压力、加速踏板位置、进气歧管温度等传感器通道；对于变速器子系统，则考虑车速、变速器油温、变速箱当前挡位等。对于燃油子系统，我们选取燃油率、喷油器控制压力等通道。同样，对于制动子系统，我们选取制动开关传感器。这些通道的概况见表 I。

B. 数字孪生建模

现有数字孪生模型采用两阶段方法来识别异常，类似于单分类任务。第一阶段以时间序列预测的形式进行预测建模。数字孪生模型会读取来自各传感器通道的时间序列数据。由于数据具有流式特性，因此使用滑动窗口方法处理数据，创建由连续观测数据构成的输入序列。采用时序卷积网络（TCN），根据来自 t_1 至 t_{n-1} 的观测数据输入序列，预测 t_n 处的下一个观测数据（参见先前的工作 [1]）。第二阶段使用训练期间得到的TCN预测误差分布进行异常检测

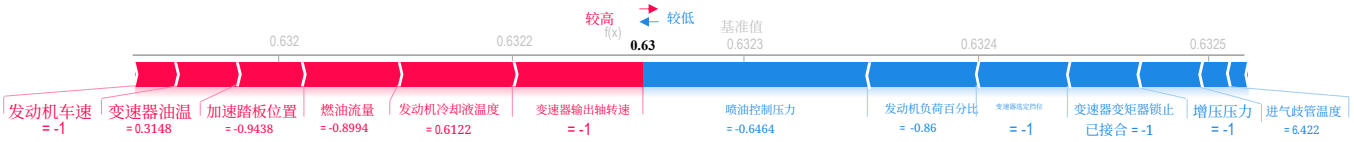

图4：使用力图作为可视化工具，对数字孪生的一次正常预测进行图形表示。图中的红色箭头表示对总体预测有正向贡献的特征，蓝色箭头则表示对预测有负向贡献的特征。传感器通道（特征）及其对应的值作为输入。箭头越大，表示 SHAP 值特征的影响越大。

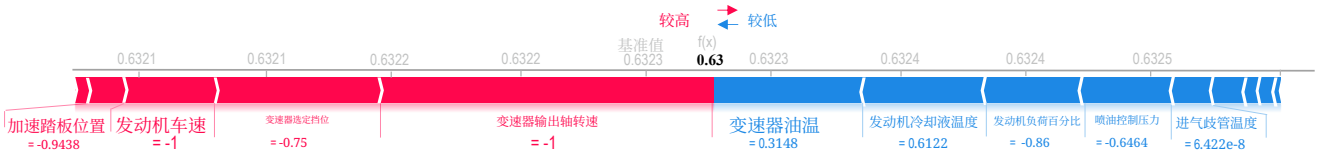

图5：使用力图示意图，将数字孪生的一次异常预测可视化。蓝色箭头表示对预测有负向贡献的特征，红色箭头则表示对预测有正向贡献的特征。图中所示的特征值（传感器通道）作为输入。影响较大的 SHAP 值特征以长度较长的箭头表示。

。当向数字孪生发起查询（即测试序列）时，会计算预测误差，并使用马氏距离（MD）将其与训练预测误差分布进行比较。较大的MD值表示与预测模型存在显著偏差。系统利用MD判断查询中是否包含异常观测。

V. 使用 TwinExplainer 生成解释

本节将介绍如何使用 *TwinExplainer* 生成解释。TwinExplainer 可帮助用户将数字孪生（DT）所作预测的全局解释和局部解释可视化。*TwinExplainer* 采用条形图、蜂群图、汇总图、依赖图和力图等可视化技术来展示数字孪生的预测。首先，我们将介绍全局解释，重点说明数字孪生的特征归因。这类解释有助于汽车行业利益相关者了解传感器通道的整体情况，以及这些通道如何影响数字孪生计算出的一般性预测。此外，我们还构建了局部解释，帮助汽车行业利益相关者理解特定数字孪生预测背后的推理过程。我们考察了两种局部预测场景。第一种场景是将正常预测的解释可视化。在流程的第二步中，我们为数字孪生产生的任何异常或非典型预测创建可视化表示。随后，我们比较这两种场景，并说明在每种场景下，哪些特征对预测的形成贡献最大。

A. 数字孪生的全局解释

全局解释旨在呈现模型整体的行为，并遵循其得出所有潜在结果的完整推理链条。它适用于群体层面的决策。因此，全局可解释性对于了解数字孪生针对输入数据分布的整体行为至关重要。

我们利用 SHAP [6] 图按重要性从高到低展示每个特征。为了说明全局特征对数字孪生预测的重要性，我们采用两种可视化工具：条形图（见图 6）和汇总图（见图 7）。此外，我们使用依赖图（见图 8）展示两个传感器通道之间的交互效应。

1) 数字孪生特征重要性：为了说明数字孪生如何得出其预测结果，我们构建了 SHAP 特征重要性图，该图基于特征归因的大小；因此，具有较大绝对值 Shapley值的特征被视为重要特征。为了计算全局特征重要性，我们计算数据集中每个特征的绝对 Shapley值平均值。然后，根据每个变量对预测模型的贡献，使用这些值生成条形图，并按与预测模型相关性递减的顺序排列。条形图（见图 6）描绘了每个特征的相对重要性，有助于汽车行业利益相关者了解最具影响力和影响力最小的传感器通道。在此实例中，影响力最大的三个传感器通道是 *VehSpeedEng*、*TransOilTemp*、*EngCoolant-Temp*，而 *BoostPres*、*InstFuelEco* 和 *Brakeswitch*的影响力最小。

数字孪生特征重要性图虽然提供了有用的信息，但除了特征的重要性之外，并未揭示太多内容。下一节将深入介绍更高级、信息更丰富的图表，例如汇总图。

2) 数字孪生汇总图：汇总图（蜂群图（见图7））结合了特征显著性和特征效应，展示了每个特征对数字孪生

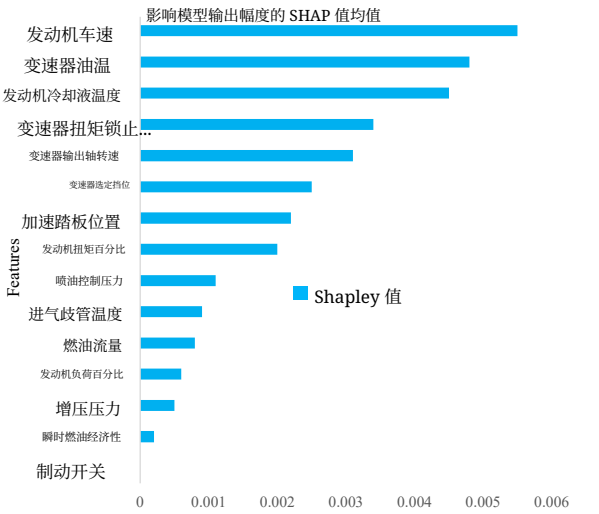

图 6：条形图的 x 轴显示数字孪生各特征的 SHAP值（或贡献度），y 轴显示按重要性从高到低排列的特征。最重要的三个全局特征是 发动机车速、变速器油温和发动机冷却液温度，而 增压压力、瞬时燃油经济性和制动开关对数字孪生所作预测的影响最小。通道说明见表 I。

预测的贡献，同时考虑所有特征值，并按照SHAP值绝对值均值排序，最显著的特征位于顶部。蜂群图中的每个点代表数字孪生预测和特征对应的单个SHAP值。水平轴表示数字孪生用于进行预测的每个特征的SHAP值。右侧的颜色图例表示特征的大小，红色代表较大的特征值，蓝色代表较小的特征值。红色和蓝色点的分布表明了特征的方向性及其对数字孪生预测的整体影响。

图 7 展示了数字孪生预测的汇总图。在这种情况下，我们可以观察到，发动机车速 传感器通道中的大多数数据点为红色，表示其推动预测结果趋于正向；而发动机冷却液温度中的大多数数据点为蓝色，表示其对预测结果产生负向影响。其余传感器通道也适用相同的分析。在此语境中，正向和负向仅表示数字孪生预测受影响的方向，与性能无关。

3）数字孪生依赖图：依赖图是一种实用的图表（见图 8），展示单个特征如何影响整个数据集上的数字孪生预测。每个点代表数据集中的一行。x 轴表示特征值，y 轴表示该特征的 SHAP值。图 8 展示了传感器通道燃油流量和发动机冷却液温度的依赖图。从图中可以推断，随着燃油流量增加，总体预测得分会降低。同样，随着发动机冷却液温度升高，总体预测也会降低。

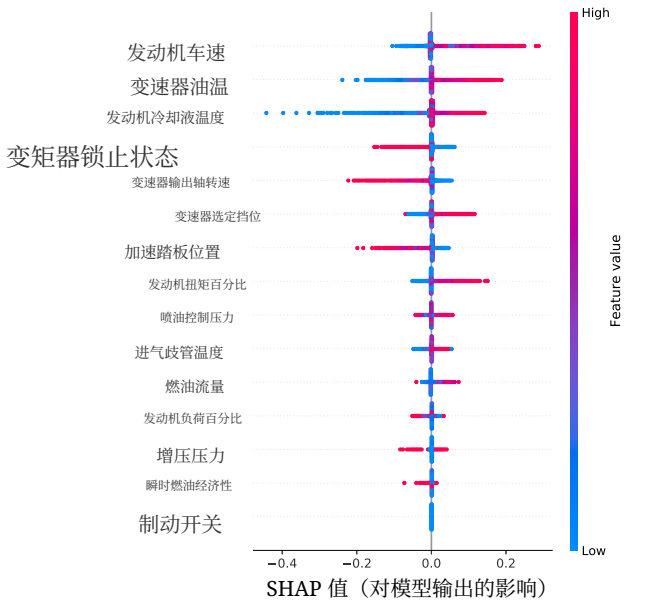

图 7：蜂群图的图形表示，它以信息密集的方式概述数据集中排名靠前的特征如何影响模型输出。在每个特征行中，一个点代表给定解释的一个实例。该特征的 SHAP值决定点的 x 坐标，而点沿每个特征行堆叠以显示密度。特征按 SHAP 值绝对值均值从高到低排序。在本例中，对总体预测有正向贡献的前三个传感器通道是*VehSpeedEng*、*TransOilTemp*和*EngCoolant-Temp*；而对预测产生负向影响且影响最小的传感器通道是*InstFuelEco* 和 *Brakeswitch*。

B. 数字孪生的局部解释

局部解释旨在帮助利益相关者说明数字孪生所做的特定预测。该技术通过回答“如果输入发生显著变化，模型的预测会如何改变？”等问题，帮助理解单个预测。此外，这些解释基于一个在特定观测附近能够紧密近似预测模型的模型。汽车行业的利益相关者可以运用这一策略，了解传感器通道的特征，以及这些特征如何促成某些预测。这将使他们能够重点关注产生异常读数的通道传感器，并提出维护计划来修复有问题的通道。

通过力图，*TwinExplainer* 解释器对数字孪生在正常预测和异常预测情况下的局部预测进行解释。我们首先介绍数字孪生所作的正常预测，然后介绍异常预测。这样做的目的是深入了解哪些特征归因导致了正常预测或异常预测。

1) 正常预测：力图展示了每个特征（传感器通道）对数字孪生所计算出的特定预测的贡献。

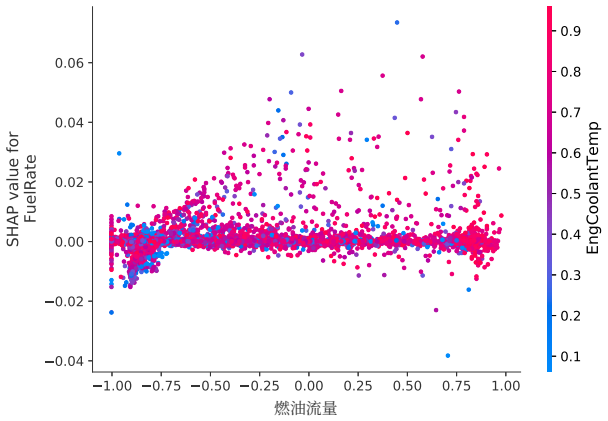

图8：展示传感器通道燃油流量 主要与发动机冷却液温度交互的部分依赖图。图中的每个点代表一个数据实例。随着燃油流量 和发动机冷却液温度增加，总体预测值降低。

图4和图5分别展示了正常预测和异常预测的力图可视化结果。在这些图中，红色箭头表示使预测值增大的特征效应（SHAP值），而蓝色箭头表示使预测值减小的特征效应。每个箭头的大小表示相应特征的影响程度。数字孪生在整个训练集上的平均预测值由“基准值”（左上角的灰色数字）表示。图表底部显示了实际特征值及其影响。

图4展示了数字孪生以0.63的概率进行正常预测时的力图。在此情况下，输入由所有传感器通道（特征）组成，其数值列在各自名称旁边。对做出该预测起正向作用的最重要特征是变速器输出轴转速、发动机冷却液温度、燃油流量、加速踏板位置、变速器油温和发动机车速；而按影响程度排序，对预测产生负向影响的特征是喷油控制压力、发动机负荷百分比、变速器选定挡位、变矩器锁止状态、增压压力和进气歧管温度。

2）异常预测：为了解释数字孪生的异常预测，我们采用了相同的可视化技术。图5以力图的形式展示了一个异常预测的图示。与正常预测类似，此处解释所用的输入由相同的特征构成，但特征值有所不同，如上方图表中各特征名称旁所示。在这种情况下，各特征的贡献与正常预测不同。变速器输出轴转速、变速器选定挡位、发动机车速、加速踏板位置等特征对0.63的总体预测产生了正向影响，而变速器油温、发动机冷却液温度、发动机负荷百分比、喷油控制压力、进气歧管温度、发动机扭矩百分比等特征则对预测产生了负向影响。对两种预测产生最显著正向影响的特征都是变速器输出轴转速；而对正常预测产生最显著负向影响的特征是喷油控制压力，对异常预测产生最显著负向影响的特征则是变速器油温。

六、结论与未来工作

随着先进传感技术和智能系统的发展，数字化在未来将至关重要。因此，数字孪生可能是最具前景的赋能技术之一，也是决定网络（虚拟）世界与物理世界如何相互作用的关键因素。通常，数字孪生系统利用深度学习模型进行分析、洞察和商业智能；然而，数字孪生系统所采用的深度学习方法，其决策本身并不具备可解释性，这成为其应用的障碍，因为它会引发对决策透明度的质疑，并导致用户信任方面的担忧。

为解决这些问题，本文提出了*TwinEx-plainer*，用于解释数字孪生对车辆的预测。合理的预测解释有助于汽车行业利益相关者理解决策结果，并协助其完成预测性维护、故障检测与诊断、优化和维修等后续任务，同时增强其对人工智能系统的信任。我们利用该流程自动生成数字孪生的解释。该流程包含三个不同阶段：首先收集、存储和处理传感器通道输入；然后利用能够预测传感器通道读数的深度学习分类器构建数字孪生；最后为预测结果生成解释。为论证预测结果，*TwinExplainer* 采用了一种基于特征的可解释方法。基于特征的解释以特征作为解释手段，用于确定每个特征对输出预测的贡献程度。该方法以可视化形式提供两类解释（全局解释和局部解释），包括条形图、汇总图、依赖图和力图，用于展示特征归因、特征影响，以及特征如何促使模型做出特定决策。

未来，我们希望评估这些解释的性能。为此，我们计划开发以用户为导向的评估指标，以探究解释的影响。

致谢

本研究得到密西西比州立大学计算机科学与工程系的PATENT 实验室（预测分析与技术集成实验室）、美国国防部的一项资助，以及美国国家科学基金会（# 1565484）的支持。文中观点和结论仅代表作者个人意见。

参考文献

[1] Subash Neupane、Ivan A Fernandez、Wilson Patterson、Sudip Mittal 和 Shahram Rahimi。利用功能工作组和传感器通道的车辆时间异常检测系统。第八届 *IEEE* 协作与互联网计算国际会议（*IEEE CIC 2022*），2022年。[2] Edward Glaessgen 和 David Stargel。面向未来 NASA 和美国空军车辆的数字孪生范式。载于第*53*届 *AIAA/ASME/ASCE/AHS/ASC*结构、结构动力学与材料会议、第*20*届 *AIAA/ASME/AHS*自适应结构会议第*141*818，2012届 *AIAA*，页码未提供。

[3] 亚马逊。什么是数字孪生技术？什么是数字孪生技术？ https://aws.amazon.com/what-is/digital-twin/，2022年。访问日期：2022-12-22。[4] Wojciech Samek、Thomas Wiegand 和 Klaus-Robert M¨uller。可解释人工智能：理解、可视化和解读深度学习模型。arXiv 预印本 arXiv:1708.08296，2017年。[5] Alaa Marshan。人工智能：可解释性、伦理问题与偏差。《机器人与自动化年鉴》，5(1):034–037，2021年。[6] Scott M Lundberg 和 Su-In Lee。解读模型预测的统一方法。《神经信息处理系统进展》，第30卷，2017年。[7] Subash Neupane、Jesse Ables、William Anderson、Sudip Mittal、Shahram Rahimi、Ioana Banicescu 和 Maria Seale。可解释入侵检测系统（x-ids）：当前方法、挑战与机遇综述。IEEE Access，10:112392–112415，2022年。[8] Michael Grieves。数字孪生：通过虚拟工厂复制实现卓越制造。白皮书，1(2014):1–7，2014年。[9] Ying Liu、Lin Zhang、Yuan Yang、Longfei Zhou、Lei Ren、Fei Wang、Rong Liu、Zhibo Pang 和 M Jamal Deen。利用数字孪生为老年人医疗服务构建新型云端框架。IEEE Access，7:49088–49101，2019年。[10] Haya Elayan、Moayad Aloqaily 和 Mohsen Guizani。面向智能情境感知物联网医疗系统的数字孪生。IEEE《物联网期刊》，8(23):16749–16757，2021年。[11] Mike Shafto、Mike Conroy、Rich Doyle、Ed Glaessgen、Chris Kemp、Jacqueline LeMoigne 和 Lui Wang。建模、模拟、信息技术与处理路线图。美国国家航空航天局，32(2012):1–38，2012年。[12] Eric J Tuegel、Anthony R Ingraffea、Thomas G Eason 和 S Michael Spottswood。利用数字孪生重新设计飞机结构寿命预测。《国际航空航天工程期刊》，2011，2011年。[13] Shimin Liu、Jinsong Bao、Yuqian Lu、Jie Li、Shanyu Lu 和 Xuemin Sun。基于仿生学的航空航天部件加工数字孪生建模方法。《制造系统期刊》，58 195，2021 180: –。[14] Agniva Banerjee、Raka Dalal、Sudip Mittal 和 Karuna Pande Joshi。利用知识图谱为工业生产线生成数字孪生模型。第届 ACM Web Science 国际会议 2017，2017年。[15] Fei Tao、Jiangfeng Cheng、Qinglin Qi、Meng Zhang、He Zhang 和 Fangyuan Sui。大数据驱动的数字孪生产品设计、制造与服务。《国际先进制造技术期刊》，94(9):3563–3576，2018年。[16] Kazi Masudul Alam 和 Abdulmotaleb El Saddik。C2PS：面向云端信息物理系统的数字孪生架构参考模型。IEEE Access，5:2050–2062，2017年。[17] Fei Tao、Qinglin Qi、Lihui Wang 和 AYC Nee。面向智能制造与工业4.0的数字孪生和信息物理系统：关联与比较。《工程》，5(4):653–661，2019年。[18] Jesse Ables、Thomas Kirby、William Anderson、Sudip Mittal、Shahram Rahimi、Ioana Banicescu 和 Maria Seale。使用自组织映射创建可解释入侵检测系统。IEEE计算智能研讨会系列（IEEE SSCI 2022），2022年。[19] Florian Biesinger 和 Michael Weyrich。生产与汽车行业中数字孪生的各个方面。载于《2019年第23届机电一体化技术国际会议（ICMT）》，第1–6页。IEEE，2019年。[20] Andreas Theissler、Judith P´erez-Vel´azquez、Marcel Kettelgerdes 和 Gordon Elger。机器学习助力预测性维护：汽车行业中的应用案例与挑战。《可靠性工程与系统安全》，215:107864，2021年。[21] R Keith Mobley。《预测性维护导论》。Elsevier，2002年。[22] Jiafu Wan、Shenglong Tang、Di Li、Shiyong Wang、Chengliang Liu、Haider Abbas 和 Athanasios V Vasilakos。面向主动预防性维护的制造业大数据解决方案。《IEEE工业信息学汇刊》，13(4):2039–2047，2017年。[23] John H Williams、Alan Davies 和 Paul R Drake。《基于状态的维护与机器诊断》。Springer Science & Business Media，1994年。[24] Yongyi Ran、Xin Zhou、Pengfeng Lin、Yonggang Wen 和 Ruilong Deng。预测性维护综述：系统、目的与方法。arXiv 预印本 arXiv:1912.07383，2019年。

[25] Nan Zhang、Rami Bahsoon、Nikos Tziritas 和 Georgios Theodoropou- los。可解释的人在回路中的动态数据驱动数字孪生。*arXiv* 预印本 *arXiv:2207.09106*，2022 年。[26] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。“我为什么应该相信你？”解释任意分类器的预测。载于第 *22* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 1135–1144 页，2016 年。[27] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。Anchors：高精度模型无关解释。载于*AAAI* 人工智能会议论文集，第 32 卷，2018 年。[28] Jing Lei、Max G'Sell、Alessandro Rinaldo、Ryan J Tibshirani 和 Larry Wasserman。用于回归的无分布假设预测推断。美国统计学会会刊，113 1111，2018 523 1094( ): – 。[29] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 等。超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于国际机器学习会议，第 2668–2677 页。PMLR，2018 年。[30] Chengliang Yang、Anand Rangarajan 和 Sanjay Ranka。通过递归划分进行全局模型解释。载于*2018 IEEE* 第 *20* 届高性能计算与通信国际会议；*IEEE* 第 *16* 届智慧城市国际会议；*IEEE* 第 *4* 届数据科学与系统国际会议（*HPCC/SmartCity/DSS*），第 1563–1570 页。IEEE，2018 年。[31] Amina Adadi 和 Mohammed Berrada。窥探黑箱内部：可解释人工智能（XAI）综述。*IEEE Access*，6：52138–52160，2018 年。[32] Dattaraj Jagdish Rao 和 Shraddha Mane。采用可解释人工智能的临床决策支持系统数字孪生方法。*arXiv* 预印本 *arXiv:1910.13520*，2019 年。[33] Michael G Kapteyn、David J Knezevic 和 Karen Willcox。通过基于组件的降阶模型和可解释机器学习实现预测型数字孪生。载于*AIAA Scitech 2020* 0418，2020论坛，页码未详。[34] Lu Wang、Tianhu Deng、Zeyu Zheng 和 Zuo-Jun Max Shen。数字孪生中的可解释建模。载于*2021* 年冬季仿真会议（*WSC*），第 1–12 页。IEEE，2021 年。[35] Dimitrios Piromalis 和 Antreas Kantaros。汽车行业中的数字孪生：迈向物理—数字融合之路。*Applied System Innovation*，5(4):65，2022 年。[36] Fei Song、Yanlei Diao、Jesse Read、Arnaud Stiegler 和 Albert Bifet。ExAD：一种用于大数据轨迹可解释异常检测的系统。载于*2018 IEEE* 国际数据挖掘会议研讨会（*ICDMW*），第 1435–1440 页。IEEE，2018 年。[37] Wilson Patterson、Ivan Fernandez、Subash Neupane、Milan Parmar、Sudip Mittal 和 Shahram Rahimi。针对数字孪生的白盒对抗攻击。*2022* 年 *ACM* 年度计算机安全应用会议（*ACSAC 2022*），2022 年。