机器学习解释中的不确定性表达：一种面向预测性流程监控的可视化分析方法

Nijat Mehdiyev，德国人工智能研究中心（*DFKI*）和萨尔大学，萨尔布吕肯，*66123*，德国

Maxim Majlatow，德国人工智能研究中心（*DFKI*）和萨尔大学，萨尔布吕肯，*66123*，德国

Peter Fettke，德国人工智能研究中心（*DFKI*）和萨尔大学，萨尔布吕肯，*66123*，德国

arXiv:2304.05736v1 [cs.LG] 12 Apr 2023

摘要——随着数据驱动的智能系统不断发展，对可靠且透明的决策机制的需求日益重要。因此，整合不确定性量化与模型可解释性方法，对于促进可信的业务与运营流程分析至关重要。本研究探讨了如何在全局和局部事后解释方法中有效传达模型不确定性，例如部分依赖图（*PDP*）和个体条件期望图（*ICE*）。此外，本研究还考察了适当的可视化分析方法，以促进此类方法的整合。通过结合这两个研究方向，决策者不仅能够论证由解释驱动的可操作洞见的合理性，还能够验证其可靠性。最后，本研究通过专家访谈评估所提出的方法和设计的界面对于制造领域实际预测性流程监控问题的适用性。

The 机器学习（ML）模型在现实世界中涉及重大利害关系的决策场景中被广泛应用，因此必须确保其可靠性和可理解性，从而促进相关利益相关者对其建立信任 [1]。为此，近期提出了许多可解释人工智能（XAI）技术，包括事后解释方法，用于提供模型行为的全局和局部解释 [2]。这些方法旨在提升模型的透明度和可解释性，使利益相关者能够更好地理解模型的决策过程。

不确定性量化（UQ）是机器学习领域一个新兴的研究方向，侧重于估计并传播模型预测的不确定性。它可被视为一种补充性的透明度形式，能够提升决策任务中解决方案的可解释性；仅靠解决方案本身可能还不够 [3]。经过适当校准后并得到有效传播后，不确定性能够提升利益相关者判断何时应信任模型预测的能力，从而增强决策自动化或决策辅助 [4]。近期研究表明，结合适当的沟通策略提供模型置信度信息，也能为领域专家带来一定程度的透明度与信任 [5]。

选择并部署适当的方法来生成解释或量化不确定性，对机器学习模型检查至关重要。然而，如果未能向利益相关者清晰、简洁地呈现洞见，这一过程的效率可能会受到影响。因此，为了实现模型结果的最佳传播，必须构建与目标利益相关者的心智模型相适配、同时尽量降低其认知负荷的交互式界面。为此，交互式信息可视化已成为一个颇具前景的研究领域，它通过向用户提供相关的视觉表示 [6]，增强机器学习模型的可解释性、可信度和可靠性。仔细考察这一领域.....

相关文献表明，大量研究已经分别针对不确定性传达的可视分析[7]，或模型可解释性 [8]，开展，但二者彼此孤立。尽管这两个研究方向对于促进更加充分的信息支持决策都十分重要，但目前似乎仍缺乏将用于不确定性传达的信息可视化与模型可解释性相结合的整合方法，从而实现更全面的理解。

针对上述差距，我们提出了一种新方法，旨在将模型不确定性融入局部和全局事后解释中。我们将该方法部署在一个可视分析界面中，以便验证所生成解释的适当性和可用性。为确保解决方案满足目标用户的需求，我们在一项联盟研究项目中，针对设计流程和评估开展了全面的需求分析。我们的主要贡献总结如下：

- 我们探讨了多种呈现形式，以传达单个预测的模型不确定性，包括可视化、文本描述和表格描述。
- 我们提出并形式化了将模型不确定性融入全局事后解释方法的方案，具体包括部分依赖图（PDP）及其局部版本——个体条件期望图。
- 我们设计了可视化分析工具，以支持所提方法的实现，并在一个涉及预测性流程监控问题的真实应用案例中，与领域专家共同开展了全面评估。

背景与相关工作

在众多实际应用场景中，黑盒机器学习算法对于达到传统的内在可解释的机器学习方法无法实现的准确度至关重要。尽管如此，这类不透明方法往往无法说明其工作机制，导致分析人员难以验证其真实性 [2]。提供解释是促进人们接受智能系统所提供预测的有效方法。因此，可解释人工智能（XAI）已成为一个富有成果的研究领域，旨在通过让基于人工智能的系统所依赖的不透明算法变得易于理解，增强它们与人类用户之间的协作算法 [9]。可解释性的概念错综复杂且涉及多个方面，需要考虑决策环境中的各种因素。这些因素包括分析情境、用户属性、解释目标，以及一系列社会认知和流程特定因素。因此，在设计智能方法和界面时，必须将这些因素纳入考量，以确保具备充分的可解释性 [10]。

XAI 技术有多种分类法，其中一大类是事后解释技术。这些技术为人工智能模型预测提供解释，可分为局部解释和全局解释两类。局部技术（SHapley 加性解释（SHAP）、局部可解释模型无关解释（LIME）、ICE）用于解释单个预测，而全局方法（部分依赖图、SHAP 汇总图、基于置换的特征重要性）则用于解释人工智能模型的整体行为 [11]。由于易于理解，领域专家通常更倾向于使用事后解释技术进行论证和验证。然而，近期研究表明，大多数此类事后解释技术都存在不一致和不稳定的问题，且可能无法提供关于其可靠性的充分信息，这凸显了整合模型不确定性估计的必要性 [1]。尽管如此，迄今只有少数研究尝试对事后解释中的不确定性进行建模，我们的研究也开展了这一工作 [1]，[12]，[13]。此外，针对这一交叉领域开发融合可视化分析方法的用户界面，也一直未受到足够关注。

机器学习生命周期涵盖多个阶段，从数据收集到模型训练，每个阶段都会引入固有的不确定性。因此，人工智能模型生成的预测会受到各种不确定性的影响，例如数据收集错误、模型复杂性和算法局限性。在人工智能模型领域，通常区分两类不确定性：偶然不确定性和认知不确定性 [14]。偶然不确定性与底层数据的固有变异性或观测现象内在的随机性有关。相比之下，认知不确定性源于对所建模系统的认识不完整或不足，或模型本身的局限性。偶然不确定性描述的是无法降低的不确定性，通常与底层数据有关；认知不确定性则可以降低，可能是由于数据不完整，也可能是拟合模型的特性所致。用于量化并涵盖这两类不确定性的方法可分为

两大类：贝叶斯方法和频率学派方法 [3]。本研究概述了这两类方法中的技术 [14]。

在可解释性的背景下融入不确定性量化，是迈向全面、可信赖的人工智能（AI）的一步，尤其有助于建立用户对模型决策的信任与接受度。

研究方法

本节介绍如何构建不确定性感知的可解释人工智能解决方案及相应界面。本文采用设计科学研究方法，以系统、严谨的流程开展应用研究 [15]。该方法尤其适用于信息系统研究，包含六个步骤：识别问题并阐明动机、明确解决方案的目标、设计与开发、演示、评估和传播。

问题识别与动机：为确保情境适用性和方法严谨性，本文结合应用领域的意见以增强情境相关性，并借助现有科学知识基础确保方法严谨性。为确保研究的相关性，我们采用联合研究方法，邀请从业者共同识别尚待解决的问题并明确目标。此外，我们还开展了广泛的文献分析，以确保研究严谨性，并通过与从业者反复讨论对发现进行完善。

本研究所确定的主要挑战是开发一种能够克服黑箱机器学习技术不透明性的方案。这一挑战在高风险决策问题中尤为突出。为解决这一问题，此类系统的设计应纳入事后解释技术，使用户能够验证模型决策的有效性。此外，为确保解决方案不仅可解释，而且可靠且值得信赖，系统生成的解释还必须有效传达模型不确定性

解决方案的目标： 下一步是明确解决方案的目标。目标包含两个方面。首先，需要一种方法来传达模型不确定性，并将此信息纳入事后解释中。这将提高底层算法的可靠性和可信度。其次，应设计具备相关可视化技术的界面，以便向系统用户有效传达结果。

设计与开发：我们提出的人工智能

事实上，该方法由用于生成预测的深度前馈神经网络、用于估计模型不确定性的蒙特卡洛（MC）丢弃法，以及分别用于生成局部和全局解释的ICE图和部分依赖图方法构成。本研究的创新之处在于将不确定性信息融入这些基于可视化的事后解释方法。通过这种集成，决策者不仅能够论证由解释驱动的可行动见解的合理性，还能通过考察模型置信度来验证其可靠性。

演示：已针对预测性流程监控问题检验了所提出制品的适用性。预测性流程监控是过程挖掘的一个分支，它将先进的计算智能方法与流程建模方法相结合 [16]。其目标是从流程感知信息系统（PAIS）生成的事件日志中提取预测性、数据驱动且特定于流程的洞见，从而实现业务流程的持续改进 [17]。事件日志通过提供有关流程执行的必要细节，成为基于证据开展流程分析的重要基础。

更具体地说，我们探讨制造领域中的一个真实应用场景。所研究的问题涉及周期时间预测，重点是预测完成订单所需的各项制造活动的持续时间。数据来自联盟合作伙伴的制造执行系统（MES）。在实施我们提出的制品之前，我们利用融合了客户订单数据的流程专属数据，进行了严格的特征工程。

评估：评估阶段包括与两位领域专家开展半结构化访谈，他们针对系统设计和可用性提供了批判性且建设性的反馈，并提出了改进建议。评估结果为完善系统和提升其可用性提供了宝贵见解。

传播： 最后，传播阶段包括通过科学出版物和行业活动，向来自不同背景的广大研究人员和从业者分享发现和方法详情。

通过采用设计科学研究，我们可以确保以方法论严谨且系统化的流程开展与应用领域相关的应用研究。最终，采用这种方法开发的可解释人工智能解决方案将着重传达不确定性，并有望提升人工智能系统的可信度。

后验解释中的不确定性估计

本节概述了支撑我们提出的新颖不确定性感知可解释人工智能方法的数学基础。

使用蒙特卡洛丢弃进行不确定性量化

由 [18], 提出的 MC 丢弃技术是量化深度学习模型不确定性的最先进方法。通常，在神经网络训练阶段会采用丢弃技术，通过随机丢弃或忽略神经元来防止过拟合。在部署阶段启用丢弃正则化，可视为对概率深度高斯过程的一种贝叶斯近似。通过对网络执行 $T \in \mathbb{N}^+$ 次随机前向传播，可以计算模型预测方差，并将其用作不确定性的度量。

我们通过对深度前馈神经网络执行$T \geq 50$次随机前向传播来实现 MC 丢弃。该方法可以将模型预测映射到相应的方差。具体而言，可以先按升序排列训练数据的方差，再计算方差阈值，从而为训练数据创建不确定性档案。例如，可以计算第 25 和第 75 百分位数的阈值，从而得到三个不确定性档案。基于百分位数的估计为模型置信度档案的分类奠定了初步基础。最终，领域专家负责确定最终档案，可以在初始数据驱动估计的基础上进行改进，也可以自行定义范围或类别。

具备不确定性感知能力的部分依赖图。 [19], 提出的部分依赖图方法是一种全局解释方法，用于描绘拟合模型学习到的预测变量与目标变量之间的关系：给定一组预测变量，通过将这些预测变量的值替换为某个特定观测的对应值，创建一个新的数据集，并让拟合模型对该数据集进行评分。然后，对所得分数取平均，并针对其余观测到的特征值重复这一过程，即可将平均预测值绘制为观测特征值的函数，从而得到相应的部分依赖图。

设 $\mathbf{x} = (x_1, ..., x_p)$ 为一组预测变量，y为响应变量，设 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ 为相关 $(\mathbf{x}, y)$ 值的数据集，其中 N 表示观测值总数。该数据集被划分为三个数据集 $D = D_{train} \cup D_{val} \cup D_{test}$，其中 D_{train} 用于训练，D_{val} 用于超参数优化，D_{test} 用于评估，N_{train}、N_{val}、N_{test} 分别表示各子集的观测值数量。设 $\hat{F}(x)$ 为机器学习模型提供的函数，将 x 映射到 y，并在 D_{train} 上拟合。对于 $S \subset \{1, ..., p\}$，设 $x_S \subset x$ 为一组将用于生成部分依赖图（PDP）的预测变量，其中 $C \subset \{1, ..., p\} \setminus S$ 是 S 的补集，$x_C \subset x \setminus x_S$ 是所选预测变量 x_S 的补集。$\hat{F}$ 所选预测变量 x_S 上的 $\hat{F}(x) = \hat{F}(x_S, x_C)$ 的部分依赖定义为

$$\bar{F}_S(\mathbf{x}_S) = E_{\mathbf{x}_C}[\hat{F}(\mathbf{x})] = \int \hat{F}(\mathbf{x}_S, \mathbf{x}_C)\, p_C(\mathbf{x}_C)\, d\mathbf{x}_C \quad (1)$$

其中 $p_C(\mathbf{x}_C) = \int p(\mathbf{x}) d\mathbf{x}_S$ 表示$\mathbf{x}_C$的边际概率密度，而$p(\mathbf{x})$表示$\mathbf{x}$的联合密度。根据训练数据估计$p_C(\mathbf{x}_C)$，可将式（1）表示为

$$\bar{F}_S(\mathbf{x}_S) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \hat{F}(\mathbf{x}_S, \mathbf{x}_{C,i}). \quad (2)$$

纳入不确定性量化的部分依赖图的实际实现方式如下：

设$\mathbf{x}_S$ 为感兴趣的预测变量，其唯一取值为 $\{\mathbf{x}_{S,1}, ..., \mathbf{x}_{S,j}\}$，其中 j表示训练数据中观测到的唯一取值数量；设T 为随机前向传播的次数：

1) For $k \in \{1, 2, ..., j\}$:

   a) 创建 $\hat{D}_{train}$ 的副本，并将 D_{train} 中每个原始 $\mathbf{x}_S$ 值替换为特定值 $\mathbf{x}_{S,k}$。b) 根据 $\hat{D}_{train}$ 的预测值向量，计算平均模型预测$\bar{F}_S(\mathbf{x}_{S,k})$。c) 使用不确定性量化方法，通过$T$ </style}次随机前向传播计算方差向量v_k。d) 根据 v_k 计算不确定性配置文件成员的分布，并将代表性最强的群组标识保存到 $color_k$。

2) 绘制对 $\{x_{S,k}, \bar{F}_S(x_{S,k})\}$ 的成对数据，以获得部分依赖图。3) 根据相应的颜色$_k$-值，为 $\{x_{S,k}, \bar{F}_S(x_{S,k})\}$ 的成对数据着色。 4) 对于每一对 $\{x_{S,k}, \bar{F}_S(x_{S,k})\}$，将方差$v_k$ 的分布绘制为密度图或直方图，并在用户选择某个点时使图表可访问。 5) 对于每一对 $\{x_{S,k}, \bar{F}_S(x_{S,k})\}$，使用来自 v_k 的不确定性档案成员的分布

圆环图或直方图，并在用户选择某个点时让图表可供其访问。

步骤 c) 和 d) 对于纳入所提出的不确定性量化方法是必要的，这是对 [19], 的方法的扩展；而步骤 3)、4) 和 5) 则是用于人在回路分析的所提出的可视化结果（见图 2 和图 3）。该部分依赖图描绘了一组所选特征对拟合模型预测的平均影响。然而，正如 [20], 所指出的，对特定特征值的预测取平均，仍可能掩盖感兴趣特征与目标变量之间的局部关系。为揭示此类关系，我们提出将不确定性量化纳入个体条件期望图。

不确定性感知的个体条件期望图

ICE图（由 [20], 提出）是一种全局解释方法，侧重于单个特征，并展示当该特征发生边际变化时，拟合模型的预测会如何变化。具体而言，[20] 提出了以下方法：给定一组所选预测变量（其k 个取值各不相同）以及一条来自底层数据集的观测，将该观测复制k次，构造一个新的人工数据集，并依次将所选预测变量的值替换为其观测到的不同取值。随后，让拟合模型对这个新数据集进行评分，得到一个预测分数向量；该向量可与相应的预测变量不同取值相对应，从而绘制出它们之间的关系。如果仅选择一个预测变量，例如可以通过折线图将这种关系可视化。对底层数据集中的其余观测重复这一过程，就可以在图中为每条观测添加一条线。对这些线取平均，即可得到所选预测变量对应的部分依赖图。因此，对单条观测使用 ICE图属于一种局部解释方法。

将不确定性量化纳入个体条件期望图，作为一种局部解释方法，聚焦于底层数据集中的单个观测值以及一个感兴趣的预测变量，其实现方式如下：令 x_S 为感兴趣的预测变量，其唯一值为 $\{x_{S,1}, ..., x_{S,j}\}$，其中 j 表示训练数据中的唯一观测值数量；令 $(x^{(i)}_S, \mathbf{x}^{(i)}_C)$, $i \in \{1, ..., N\}$ 为从底层数据集中选取的单个观测值：

1) For $k \in \{1, 2, ..., j\}$:

a) 创建所选实例的副本，并将每个原始值 $x^{(i)}_S$ 替换为特定值 $x_{S,k}$。b) 计算模型预测 $\hat{f}(x_{S,k}, \mathbf{x}^{(i)}_C)$。c) 使用不确定性量化方法，通过$T$ 次随机前向传播，计算 $(x_{S,k}, \mathbf{x}^{(i)}_C)$ 的方差 $v^{(i)}_k$。d) 计算$v^{(i)}_k$ 在不确定性配置文件中的归属，并将组的标识符保存在 $color^{(i)}_k$ 中。

2) 绘制点对 $\{x_{S,k}, \hat{f}(x_{S,k}, \mathbf{x}^{(i)}_C)\}$，以获得 ICE图。3) 根据相应的 $color^{(i)}_k$ 值，为点对 $\{x_{S,k}, \hat{f}(x_{S,k}, \mathbf{x}^{(i)}_C)\}$ 着色。4) 分别绘制特征值和预测的分布，以堆叠直方图的形式呈现，并根据其所属的相应不确定性档案进行分组和着色，显示在相应的坐标轴上。

界面概述

本节主要介绍该解决方案的核心界面组件。这些组件旨在有效传达每项模型决策所产生的不确定性及相应的解释。除用于总体概览和实例选择的组件（**图 1**）外，该解决方案还由三个界面组件构成，分别是“不确定性估计与可视化”、“ICE图中的不确定性传达”和“部分依赖图中的不确定性传达”。

本项目的核心机器学习组件使用 R 中的 *“keras”* 库构建。用于数据准备、界面创建和可视化的其他重要库包括 *“data.table”*、*“dplyr”*、*“ggplot2”*、*“plotly”*、*“vip”*、*“shinydashboard”* 等。

不确定性估计与可视化

我们提出的解决方案中的“不确定性估计与可视化”组件旨在通过多种呈现形式向系统用户传达模型不确定性。具体而言，该组件采用了两种不同的可视化技术——密度图和箱线图，并辅以文本和表格描述，传达所选预测的具体模型不确定性信息。

首先，第一种可视化方法通过密度图（**图 2**，A）呈现采用所选不确定性量化方法（例如蒙特卡洛丢弃）生成的可能模型预测的分布。这使用户能够直观地查看分布细节，并了解模型预测主要落在哪些范围内。或者，我们也可以使用

箱线图来呈现相同的信息（图 2，C），其中铰链之间显示四分位距，竖线表示中位数，须线延伸至距离四分位距范围不超过 1.5 倍的最低和最高数据点。这两种可视化方法还辅以其他信息。例如，图中加入了预测区间，显示预测有 95 % 的概率落入的范围。每个图下方的标签以箭头指向红色竖线，作为视觉辅助来标示置信区间。图表依据定性的不确定性描述着色，分别以绿色、黄色和红色表示“低”“中”和“高”置信度档案。

除了这些可视化方法之外，我们还提供文本描述（图 2，B），其中包含有关不确定性量化方法的信息、对预测区间的文字说明，以及显示特定预测的模型置信度的不确定性档案。最后，我们向用户呈现一个表格（图 2，D），其中展示了模型预测、标准差和其他相关信息。

个体条件期望图中的不确定性传达

我们所提出的不确定性感知局部解释方法的这一组件提供了一个界面，用于可视化新合成实例的预测分数，并传达其不确定性信息（**图 3**，数值变量对应 B；**图 4**，分类变量对应B）。这是通过在图中使用颜色编码的置信区间实现的。此外，每个合成实例的不确定性还通过前一组件中描述的各种呈现形式进行传达。

此外，我们在ICE图的适当坐标轴上加入直方图作为补充图表，以呈现预测分数和特征值的分布，使用户能够辨别所检查预测实例的原始值在更广泛数据集中的位置。通过这项分析，用户可以确定该预测实例的值是否处于常见值范围内，或是否属于非典型异常值。

部分依赖图中的不确定性传达

为提高部分依赖图的透明度和可解释性，我们为部分依赖图界面引入了一个新组件，用于提供两类不确定性信息（**图 5**，B）。第一类不确定性信息通过补充密度图（图 5，C）呈现，该图显示预测的分布及其周围的 95 % 置信水平带。虽然可以直接在主图中可视化这些信息，但我们发现这样会让用户更难阅读和解读图表。我们方法的主要贡献，是在部分依赖图分析中，对感兴趣特征每个受检取值的模型不确定性进行可视化。为此，我们使用圆环图（图 5，D）显示不确定性档案的分布，让用户能够概览模型在该受检取值上的全局可靠性。

评估

使用场景

本使用场景展示了所设计界面的有效性，该界面包含可视化分析组件。该界面帮助流程专家（PE）估算完成给定客户订单所需的周期时间。根据订单规格，所需生产活动的顺序已预先确定。然而，流程专家仍须确定每项活动的持续时间，而我们的数据驱动方法可以为此提供帮助。由于按时完成高优先级订单至关重要，流程专家有责任确保系统提供的数据驱动指导有效。

总体概览：系统交互始于引导流程专家进入专为预测性流程监控用例设计的“总体概览”页面。流程专家可在此页面使用下拉菜单选择感兴趣的相关案例，该菜单对应客户订单的生产活动序列（图1，A）。页面还重点展示客户订单的规格，例如产品重量和尺寸，这些信息也会作为机器学习算法的输入。图 1，B 展示了每个生产步骤的时长预测相关信息，以及不确定性档案信息。动画流程图和甘特图（图1，C）直观呈现了活动序列，后者还显示了每个流程步骤的预测时长。甘特图中的每个步骤均根据三种不确定性类别之一进行颜色编码：“高”用红色表示，“中”用橙色表示，“低”用绿色表示。

这一特定场景中的生产活动包括 3D 切割（“3D_Microstep_Circ”）、在压凹机上作业（“Dishing_Press_1”）、滚边（“Bead_S”）、在锻造过程中调整形状（“Forge”），

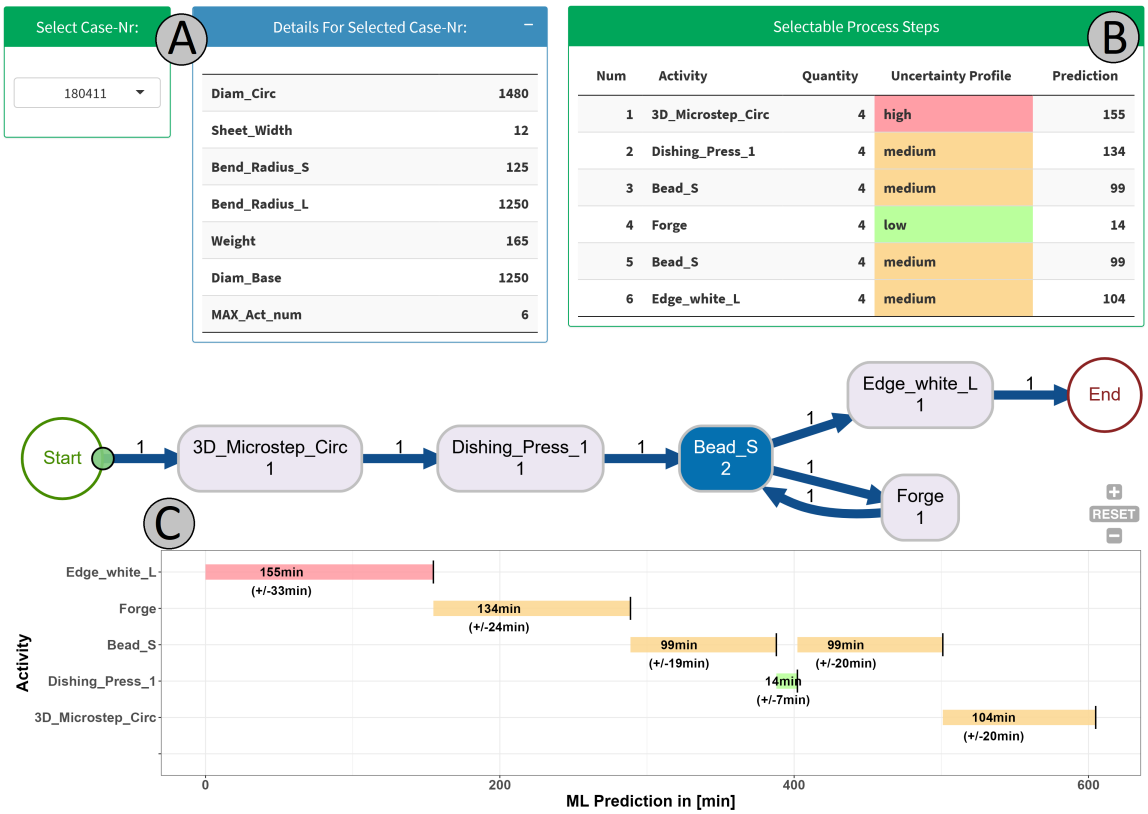


图 1。总体概览。A：左上方的下拉菜单可供用户选择一个生产案例；相应的详细信息显示在右侧。B：表格视图可帮助用户进一步了解生产案例的各个步骤，以及有关活动时长预测和不确定性的信息。可选择流程活动进行进一步分析。C：动态流程图和甘特图展示了计划的顺序安排和周期时间预测。

另一次珠饰处理（“Bead_S”）和边缘精修（“Edge_white_L”）。我们的解决方案将活动 2、3、5 和 6 归类为具有“中等”不确定性档案，而活动 3 具有“低”不确定性档案。第一项活动 3D 切割是个例外，属于“高”不确定性类别。甘特图显示，预计该活动需要 155 分钟，标准差为 33 分钟。流程专家决定调查这项生产活动，因为在这一早期阶段发生的任何中断都可能对后续流程产生连锁影响。

活动层级的不确定性分析：流程专家可以点击感兴趣的活动来分析其不确定性，本例中为 3D-cutting。此操作会显示蒙特卡洛丢弃预测的分布，如图 2 所示。模型预测值 154.9 分钟落在高亮显示的置信区间 70.7 分钟至 199.6 分钟内（图 2，A），该区间涵盖 95% 的数值，表明模型预测值不是异常值。

然而，箱线图的上四分位数（图 2，C）和置信区间的上限表明，延迟近 45 分钟并非不可能。还提供了文本描述（图 2，B），以确保正确解读。最后，图 2，D 提供了表格摘要，便于快速查阅所选活动的持续时间预测和不确定性信息。

不确定性感知的ICE图：为了解数值型和类别型特征对模型预测的影响，并制定避免不良结果的计划，流程专家会参考不确定性感知的ICE图（图3，A）。通过分析与产品规格相关的变量，例如弯曲半径（"Bend_Radius_S"），流程专家发现，在原始值（125 cm）附近，3D切割活动的持续时间预测具有“高”不确定性（图3，B）。然而，提高该特征的值会缩短预测的活动持续时间，并提高模型对其预测的置信度，从而使不确定性降至“中”等级。这张不确定性感知的ICE图

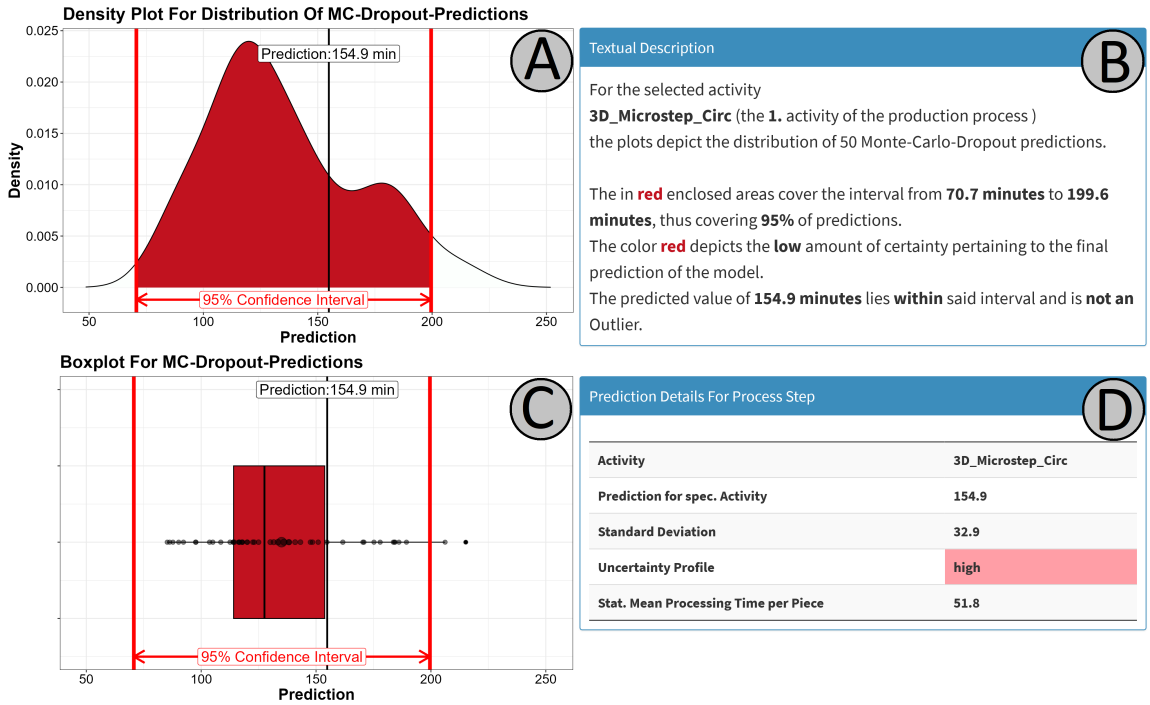


图 2. 用于实例级不确定性分析的仪表板界面。A：密度图展示了蒙特卡洛丢弃预测的分布。B：将 A 中的相同数据以箱线图形式呈现。C：文本描述总结了 A 和 B 中的重要发现。D：表格展示了有关所检查生产活动的其他信息。红色编码表示与活动时长预测相关的“高”不确定性等级。

使流程专家能够了解感兴趣特征与模型结果之间的关系，并理解模型置信度。然而，在规划流程中无法调整该特征的值，因为它是固定且预先确定的产品规格。

流程专家发现了一个可操纵且不会影响订单要求的分类变量“工人”（图 4，A），并筛除了不可用人员。在此情景中，流程专家注意到，系统分配给该订单的工人（通过标识符“751”匿名化）的预测属于“高”不确定性类别（图 4，B），而其他工人的预测不确定性为“中”。在预测得分最低的区域中，流程专家选出另外三名可用工人（“114”、“736”和“797”），其预测时长均属于“中”不确定性类别，以作进一步考察。这三名工人“797”、“736”和“114”的预计时长分别为 ~103、~118 和 ~123 分钟。

流程专家通过将模型输出与其领域专业知识进行交叉比对，来验证模型输出的准确性。由于更换选定的工人，预测得到了改善，这归因于备选工人在执行生产活动方面经验更丰富。

由于模型估算的预测时长和不确定性符合流程专家的预期，他们对模型可靠性的信心有所增强。

不确定性感知的PDP： 鉴于该订单的优先级很高，流程专家切换到“全局解释”选项卡，以确保所选替换工人中至少有一人在总体上表现良好。此选项卡向流程专家展示了基于置换的特征重要性图（图 5，A），有助于理解某些变量的整体影响。此外，还向流程专家展示了不确定性感知的PDP（图 5，B），以供进一步分析。

流程专家选择分类变量“工人”，并反复检查三个可用的工人。通过使用蒙特卡洛丢弃预测的分布（图5，C），流程专家得出结论：工人“797”的置信区间上界比当前分配的工人低约30分钟，其平均预测分数（82 分钟）低 22 分钟。此外，圆环图（图 5，D）表明，工人“797”是

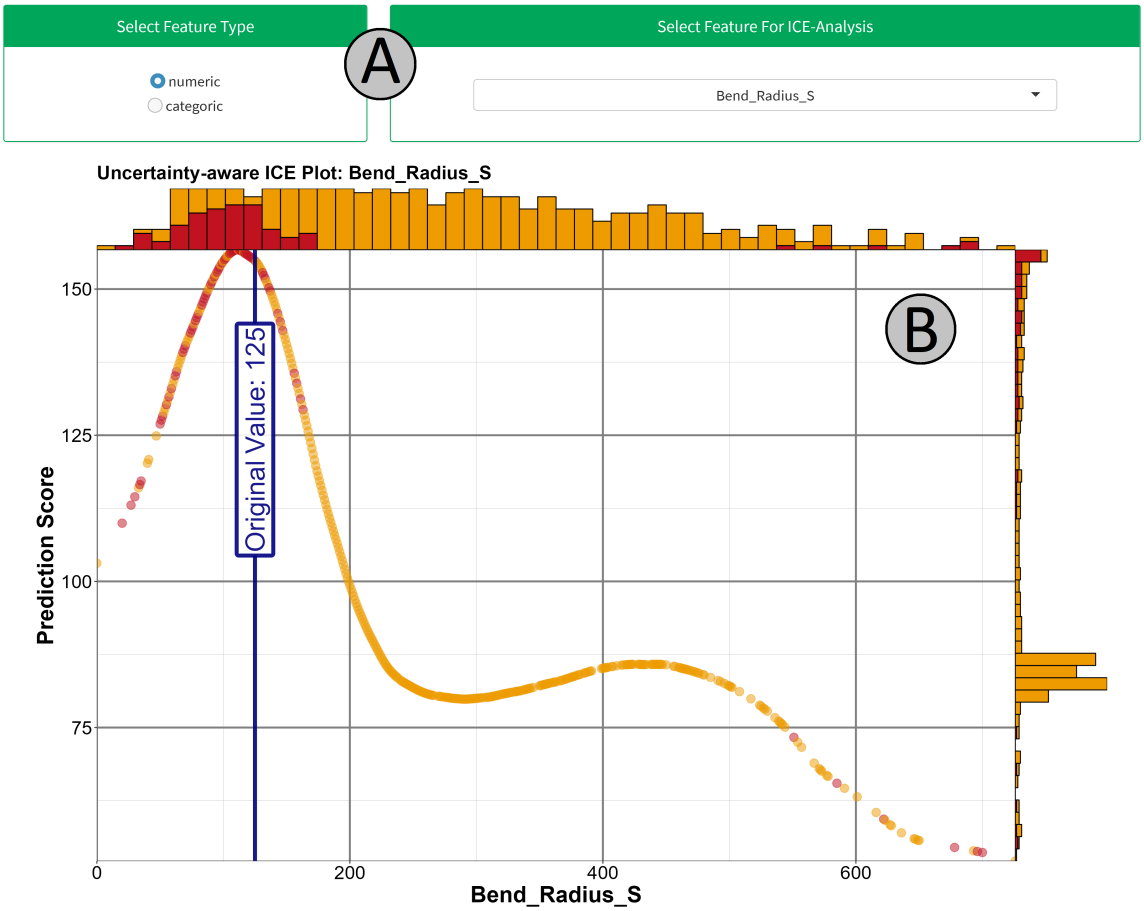


图3。 面向不确定性的ICE图仪表板界面。A：用户在左上方选择“数值”作为变量类型，随后在其右侧的下拉菜单中更新为所选类型的变量。B：此处显示了针对数值变量的面向不确定性的个体条件期望图。

与当前工人相比，这与更多的“低”不确定性（33.1%）和更少的“高”不确定性（16.6%）相关（当前工人的“低”不确定性为 21.2%，“高”不确定性为 22.7%）。因此，流程专家有充分理由修改生产计划，将给定流程活动中的当前工人“751”替换为工人“797”，从而缩短预测交付周期并降低模型不确定性。

专家访谈

在评估所提出的面向不确定性的解释方法时，我们对两位流程专家进行了访谈，其中包括一位工厂经理和一位生产经理。两位专家都对底层数据拥有丰富的知识，对相关流程有深入的理解，并且在所研究特征与目标变量。

在访谈过程中，流程专家选择了一个示例案例ID，并查看了相应生产流程的可视化结果，以此作为进一步分析的起点。他们很快便熟悉了此选项卡，并深入分析了各个生产步骤。在示例案例中，他们选择了一个流程步骤，并详细查看了不确定性分析。随后，他们以类似的方式展示并讨论了不确定性感知的ICE图。接下来，他们检查了包含不确定性感知的PDP的全局解释。他们针对示例案例中的其他流程步骤重复了这些步骤。然后，他们对其他随机选取的案例执行了相同步骤，并被要求反馈：如果该仪表板用于生产计划，他们会如何与之交互。最后，他们

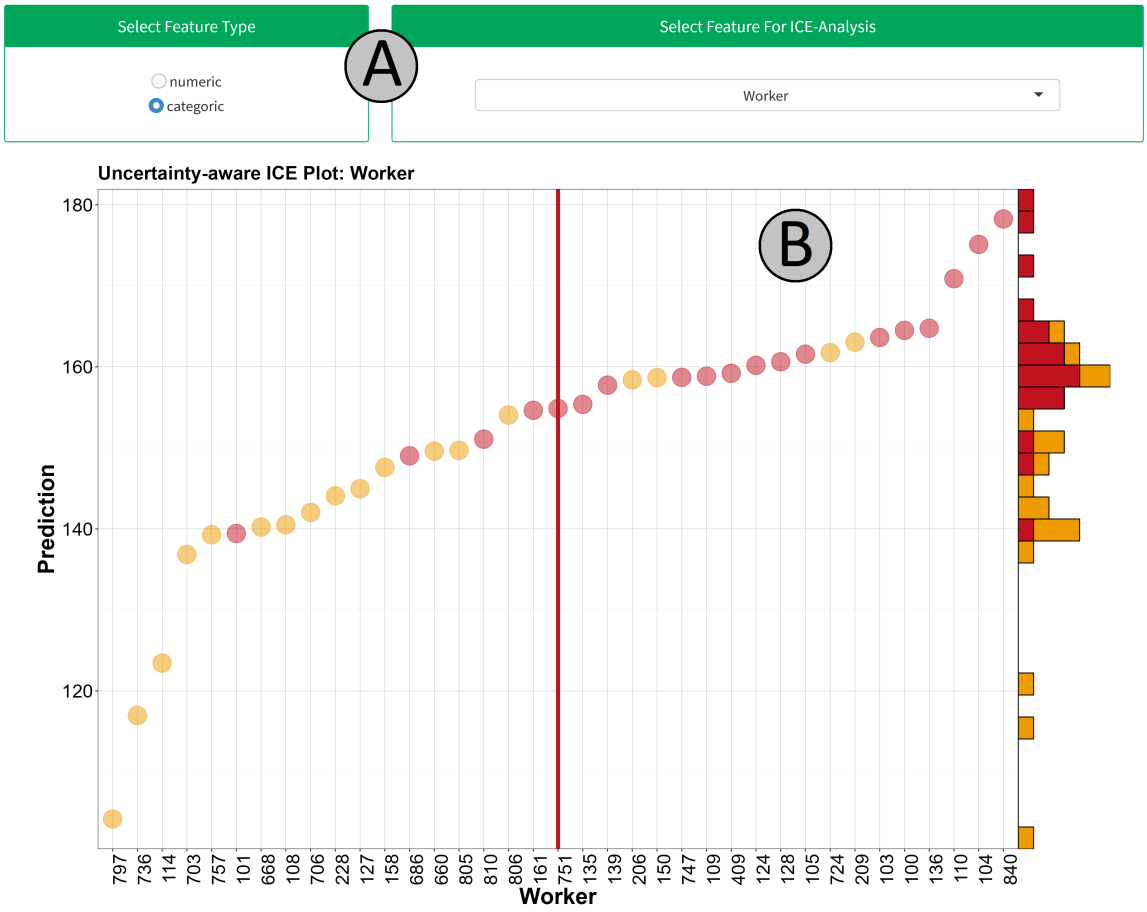


图 4. 不确定性感知的ICE图仪表板界面。A：用户在左上方选择“分类”作为变量类型，旁边的下拉菜单随即更新，显示所选类型的变量。B：此处显示所提出的适用于分类变量的不确定性感知ICE图；由于绘图时每个值只出现一次，因此省略了分类值的分布。红色竖线表示所分析实例的原始变量值。

详细讨论了仪表板，并记录了专家反馈。访谈持续了一个小时，上文描述的每个步骤约耗时 15 分钟。受访专家针对所提出的不确定性感知的解释方法中采用的可视化和解释技术的可用性与设计提供了宝贵反馈，并提出了改进建议。

设计与可用性： 在设计方面，专家对所提出的不确定性感知的ICE图和不确定性感知的PDP表示认可，并在初次介绍后不久便能够推导并验证这些可视化结果所呈现的关系。此外，用于快速区分不确定性档案的集成配色方案很快便被专家掌握，并在第三个真实使用场景中得到应用。

面试步骤。

对于呈现蒙特卡洛丢弃预测的分布，密度图被评为最易理解且最有效的可视化方法，其次是作为补充可视化方法的箱线图。专家们指出，额外的文本解释是防止错误解读、提升整体用户接受度的有力工具。经过简短的初步说明后，专家们很容易理解预测分数与蒙特卡洛丢弃预测分布之间的关系。

改进建议： 尽管专家总体上对所提出的解释方法感到满意，但他们希望能够比较相似案例。这样用户就可以筛选底层数据，以构建不确定性感知的个体条件期望图和受限的不确定性感知的部分依赖图

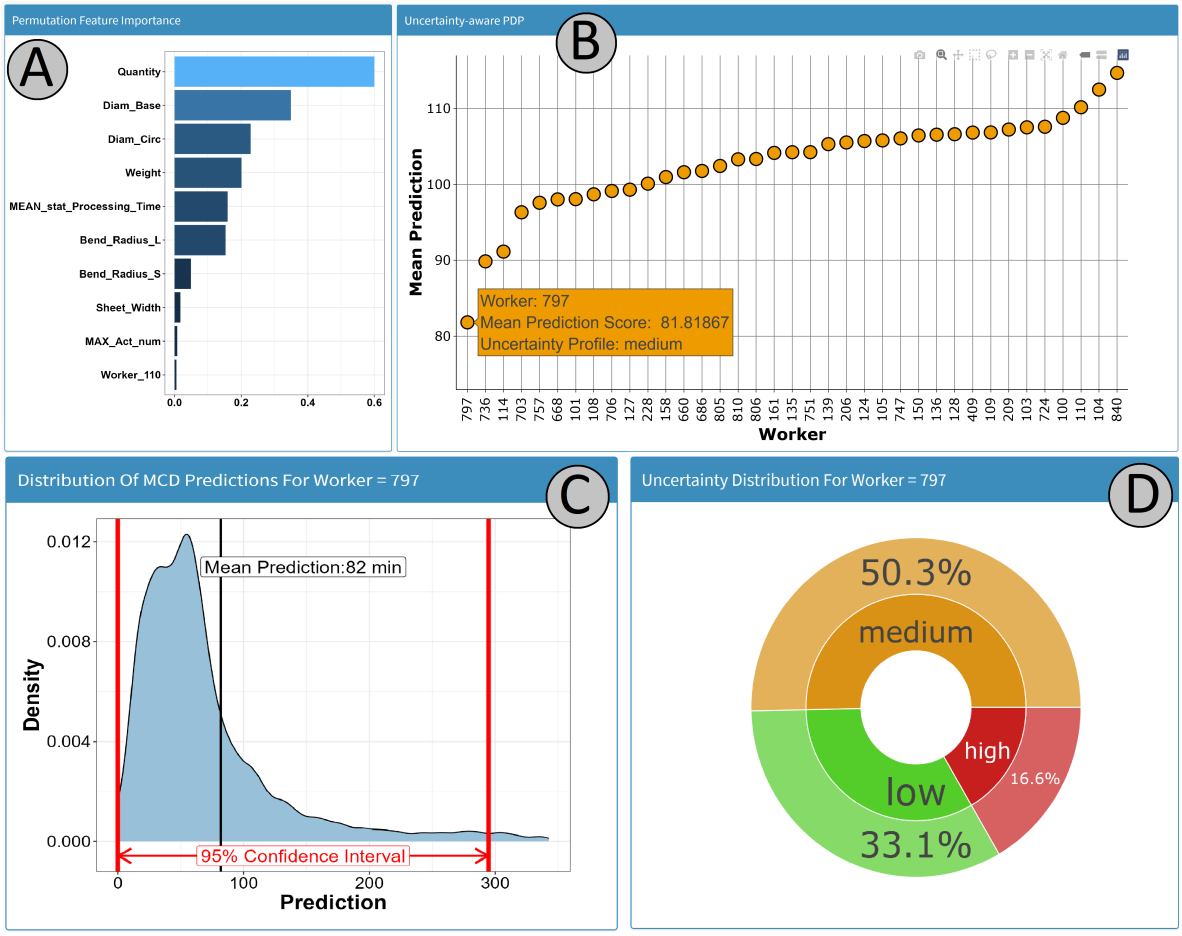


图 5。不确定性感知的PDP仪表板界面。A：将基于置换的特征重要性显示为辅助工具，用于提供方向指引和全局可解释性。B：此处显示了针对分类变量提出的不确定性感知的PDP。将鼠标悬停在某个点上，可显示其平均预测分数以及对应的主导不确定性组。C：点击 B 中的某个点可更新此密度图，该图显示预测值的分布。D：点击 B 中的某个点可更新此圆环图，该图显示对应预测值所属不确定性组的分布。

基于筛选后的数据。将此功能纳入所提出的仪表板，可提升用户体验，并提高所提出的解释方法对生产计划的实用性。

该方法在真实制造场景中的有效性，展示了流程专家使用该方法的实用价值。此外，我们还通过专家访谈，评估了所介绍的方法集成到原型仪表板后的效果和可用性。下文将讨论研究发现、挑战和未来工作。

讨论与结论

在本研究中，我们提出了一种在可解释人工智能（XAI）背景下整合并传达模型不确定性的方法，重点关注将可视化作为传递信息的媒介。具体而言，我们介绍了不确定性感知的ICE图这一局部解释方法，以及不确定性感知的部分依赖图这一全局解释方法，并通过各种视觉属性和功能加以增强，以便开展更深入的不确定性分析。我们展示了效-

灵活性与可迁移性：我们的研究采用了深度神经网络，并利用蒙特卡洛丢弃准确量化模型预测中的不确定性。这些步骤中使用的数据驱动方法并不局限于蒙特卡洛丢弃，还可以替换为其他方法，例如使用极端梯度提升（XGBoost）生成预测，或使用自助法进行不确定性量化。

我们的方法具有灵活性，因此适用于各种涉及表格数据的分类任务或回归问题，而不仅限于预测性流程监控场景。

结合局部解释与全局解释：研究发现，结合基于实例的解释与全局解释能够提升用户对人工智能模型的整体信任与接受度。通过结合全局解释和局部解释，用户既能从高层次了解模型，也能深入探究特定实例的细节。与仅提供全局解释或局部解释相比，这种方法更有助于增进用户对人工智能模型的理解与信任。我们的评估进一步凸显了使用这两种解释来探索模型不确定性的重要性。

可扩展性：不确定性量化和可解释人工智能技术（例如蒙特卡洛丢弃、ICE 和部分依赖图）的集成带来了重大的计算挑战，尤其是在可扩展性方面。例如，为具有高维输入的模型生成 ICE 和部分依赖图可能需要进行海量评估；当蒙特卡洛丢弃为每个输入生成多个预测时，这一过程的成本可能高得难以承受。此外，将这些技术结合起来的计算成本，可能会使其难以扩展到更大的数据集或更复杂的模型。为缓解这些挑战，人们已经提出了几种方法。一种潜在的解决方案是跨多个 GPU 并行化计算，这可以显著提升速度。对于蒙特卡洛丢弃而言，这种方法很实用，因为每个输入所需的大量前向传递可以高效地分配到多个处理器上。此外，可以使用分箱值而不是所有唯一值来应用可解释人工智能技术，这能够显著减少所需的评估次数。

未来工作： 基于领域专家的建议以及在研究和构思阶段汲取的经验，我们确定了若干未来研究方向。其中一个方向涉及评估不确定性量化和可解释人工智能方法。鉴于目前尚无评估这些技术组合的标准流程或机制，这构成了一项特殊挑战。解决这一挑战的一条潜在未来研究路径，是扩展仪表板，整合预测区间覆盖概率（PICP）和平均预测区间宽度（MPIW）等不确定性可靠性度量。根据这些度量的结果，可以采用不同的校准技术来提高模型的可靠性。此外，仪表板界面。可以扩展为针对不同不确定性档案，纳入平均绝对误差（MAE）或均方根误差（RMSE）等模型准确性指标。这将使用户更全面地了解模型在不同不确定性水平下的性能。未来研究的另一个有趣方向，是利用所生成的具备不确定性感知能力的解释，设计决策增强和自动化场景。通过利用这些解释，用户可以在高风险场景中做出更明智的决策。最后，还可以探索结合不确定性量化和可解释人工智能的方法，以考察高风险应用场景中决策制定的算法公平性。鉴于基于人工智能的决策系统可能存在偏见和歧视，这一研究方向具有重大的社会和伦理影响。总体而言，这些未来研究方向将深化我们对不确定性量化和可解释人工智能方法的性能与可靠性的理解，并为开发更加稳健、公平的决策系统提供依据。

参考文献

[1] D. Slack、A. Hilgard、S. Singh 和 H. Lakkaraju，“可靠的事后解释：对可解释性中的不确定性进行建模”，神经信息处理系统进展，第34卷，第9391–9404页，2021年。[2] A. B. Arrieta、N. Dı´az-Rodrı´guez、J. Del Ser、A. Bennetot、S. Tabik、A. Barbado、S. Garcı´a、S. Gil-López、D. Molina、R. Benjamins、等，“可解释人工智能（xai）：概念、分类体系、机遇及迈向负责任人工智能的挑战”，信息融合，第58卷，第82–115页，2020年。[3] U. Bhatt、J. Antorán、Y. Zhang、Q. V. Liao、P. Sattigeri、R. Fogliato、G. Melançon、R. Krishnan、J. Stanley、O. Tickoo、等，“不确定性作为透明度的一种形式：衡量、传达和利用不确定性”，载于*2021*年 *AAAI/ACM* 人工智能、伦理与社会会议论文集，2021年，第401–413页。[4] R. Tomsett、A. Preece、D. Braines、F. Cerutti、S. Chakraborty、M. Srivastava、G. Pearson 和 L. Kaplan，“通过可解释且具备不确定性感知能力的人工智能快速校准信任”，*Patterns*，第1卷，第4期，文章编号100049，2020年。[5] Y. Zhang、Q. V. Liao 和 R. K. Bellamy，“置信度与解释对人工智能辅助决策中准确性及信任校准的影响”，载于*2020*年公平性、问责制与透明度会议论文集，2020年，第295–305页。

[6] J. Choo 和 S. Liu，“面向可解释深度学习的可视分析”，*IEEE* 计算机图形学与应用，第38卷，第4期，第84–92页，2018年。[7] J. Hullman、X. Qiao、M. Correll、A. Kale 和 M. Kay，“探寻误差：不确定性可视化评估综述”，*IEEE* 可视化与计算机图形学汇刊，第25卷，第1期，第903–913页，2018年。[8] G. Alicioglu 和 B. Sun，“面向可解释人工智能方法的可视分析综述”，计算机与图形学，第102卷，第502–520页，2022年。[9] R. Guidotti、A. Monreale、S. Ruggieri、F. Turini、F. Giannotti 和 D. Pedreschi，“黑箱模型解释方法综述”，*ACM* 计算调查（*CSUR*），第51卷，第5期，第1–42页，2018年。[10] N. Mehdiyev 和 P. Fettke，“过程挖掘中的可解释人工智能：概述及一种面向预测性流程监控的新型局部解释方法的应用”，可解释人工智能：粒计算视角，第1–28页，2021年。[11] M. R. Islam、M. U. Ahmed、S. Barua 和 S. Begum，“从不同应用领域和任务角度对可解释人工智能进行的系统综述”，应用科学，第12卷，第3期，文章编号1353，2022年。[12] J. Moosbauer、J. Herbinger、G. Casalicchio、M. Lindauer 和 B. Bischl，“通过部分依赖图解释超参数优化”，神经信息处理系统进展，第34卷，第2280–2291页，2021年。[13] G. Cafri 和 B. A. Bailey，“理解黑箱预测中的变量效应：使用部分依赖量化树集成模型中的效应”，数据科学期刊，第14卷，第1期，第67–95页，2016年。[14] J. Gawlikowski、C. R. N. Tassi、M. Ali、J. Lee、M. Humt、J. Feng、A. M. Kruspe、R. Triebel、P. Jung、R. Roscher、M. Shahzad、W. Yang、R. Bamler 和 X. Zhu，“深度神经网络中的不确定性综述”，*ArXiv*，卷号 abs/2107.03342，2021年。[15] K. Peffers、T. Tuunanen、M. A. Rothenberger 和 S. Chatterjee，“面向信息系统研究的设计科学研究方法论”，管理信息系统期刊，第24卷，第3期，第45–77页，2007年。[16] P. Fettke，“概念建模与人工智能：从预测性业务流程管理视角出发的概述与研究挑战”，载于伴随论文集《*Modellierung 2020* 短文、研讨会及工具与演示论文集》，与 *Modellierung 2020* 同期举办，奥地利维也纳，*2020* 年 *2* 月 *19—21* 日，2020 年，第 157–164 页。[17] W. van der Aalst，“过程挖掘：概述与机遇”，《*ACM* 管理信息系统汇刊（*TMIS*）》，第 3 卷，第 2 期，第 1–17 页，2012 年。[18] Y. Gal 和 Z. Ghahramani，“将 Dropout 作为贝叶斯近似：表示深度学习中的模型不确定性”，载于第 *33* 届国际机器学习会议论文集——第 *48* 卷，ICML’16 系列，纽约州纽约市：JMLR.org，2016 年，第 1050–1059 页。[19] J. H. Friedman，“贪心函数近似：梯度提升机”，《统计年刊》，第 29 卷，第 5 期，第 1189–1232 页，2001 年。[20] A. Goldstein、A. Kapelner、J. Bleich 和 E. Pitkin，“窥探黑箱内部：通过个体条件期望图可视化统计学习”，《计算与图形统计学杂志》，第 24 卷，第 1 期，第 44–65 页，2015 年。

Nijat Mehdiyev 是德国人工智能研究中心（DFKI）的高级研究员。他拥有锡根大学商业信息系统学士学位，以及慕尼黑工业大学和奥格斯堡大学金融与信息管理（FIM）专业的荣誉硕士学位。Nijat 在萨尔兰大学完成了博士论文，研究主题为“面向预测性流程分析的可解释人工智能——概念框架、方法与应用”。他的研究兴趣包括以人为本的可解释人工智能、信息系统研究、预测性流程分析和数据驱动决策。他积极参与工业4.0领域的国家级研究项目，并合著了多篇学术论文。联系方式：nijat.mehdiyev@dfki.de。

MaximMajlatow 自2022年起在德国人工智能研究中心（DFKI）担任研究员。他于2021年在萨尔大学获得商务信息学学士学位，专攻机器学习。他的主要研究兴趣包括真实工业应用场景中的可解释人工智能、可视化和用户界面开发。联系方式：maxim.majlatow@dfki.de。

Peter Fettke 是萨尔大学商业信息学教授，也是德国萨尔布吕肯德国人工智能研究中心（DFKI）的首席研究员、研究员及研究组组长。在面向应用的基础研究和面向基础的应用研究中，他与约30人的研究团队致力于探讨商业信息学与人工智能（AI）学科交界处的问题，尤其是计算机集成系统建模、系统挖掘、流程预测和机器人流程自动化。他的研究成果是商业信息学领域国际权威期刊中被引用次数最多的论文之一，他也是DFKI被引用次数排名前五的科学家之一。他是期刊《企业建模与信息系统架构》（EMISAJ）的联合主编。
PeterFettke创立了依托DFKI设立的“税务技术能力中心”和“审计转型能力中心”。联系方式：
peter.fettke@dfki.de。