使用生成式人工智能实时检测并解释产后抑郁症 [1]

Silvia García-Méndez[a,*], Francisco de Arriba-Pérez[a]

[a] *Information Technologies Group, atlanTTic, University of Vigo, Vigo, Spain*

Abstract

在许多母亲分娩后面临的挑战中，产后抑郁症（PPD）是严重的状况，显著影响她们的心理和身体健康。因此，快速检测 PPD 及其相关风险因素对于及时评估和通过专业预防措施进行干预至关重要。因此，本文旨在帮助从业者利用最新技术进步做出决策，以实现实时筛查和治疗建议。我们的工作主要贡献在于一种智能 PPD 筛查系统，该系统结合了自然语言处理、机器学习（ML）和大型语言模型（LLMs），实现一种经济、实时且非侵入性的自由语音分析。此外，它解决了黑箱问题，因为预测结果通过结合 LLM 与可解释的 ML 模型（即基于树的算法）使用特征重要性和自然语言向终端用户描述。获得的结果在所有评价指标上的 PPD 检测方面是 90％，优于文献中的竞争解决方案。最终，我们的解决方案有助于快速检测 PPD 及其相关风险因素，这对于及时和正确的评估和干预至关重要。

Keywords: eXplainable and generative Artificial Intelligence, healthcare case study, human-computer interaction, Large Language Models, Natural Language Processing, postpartum depression, stream-based Machine Learning

1. 引言

抑郁症是一个全球公共卫生问题，影响超过 1.5 亿人，并且在女性中更为常见 (Labaka et al., 2018; Moreira et al., 2019)。值得注意的是，女性的心理和身体健康受到怀孕的显著影响，尤其是在分娩时，当生殖激素水平迅速下降（例如，她们可能经历焦虑、食欲紊乱、失眠、易怒、注意力不集中、情绪障碍、压力）(Yang et al., 2022; Su et al., 2023)。在分娩后，母亲们面临的诸多挑战中，产后抑郁症（PPD）是一种通常需要医学干预的严重状况 (Falana & Carrington, 2019)。

主要地，PPD 是一种常见的非精神病性心理障碍，发生在产后第一年，这可能导致女性健康的严重并发症 (Abadiga, 2019)。目前数据显示，全球每年有 10％ 到 15％ 的母亲受到 PPD 的影响 (Fatima et al., 2019; Liu et al., 2023)。此外，目标人群中，只有 20％ 被及时诊断或治疗 (Mazumder & Baruah, 2021)。

[1] García-Méndez, S., & de Arriba-Pérez, F. (2025)。通过生成性人工智能实时检测和解释产后抑郁症。《应用人工智能》，39(1), 2515063。正式版本可在以下网址访问: https://www.tandfonline.com/doi/full/10.1080/08839514.2025.2515063。

*Corresponding author: sgarcia@gti.uvigo.es

Email addresses: sgarcia@gti.uvigo.es (Silvia García-Méndez), farriba@gti.uvigo.es (Francisco de Arriba-Pérez)

令人遗憾的是，她们的认知和情感状态直接影响到新生儿的母婴依附及其正常发展 (Slomian et al., 2019; Rogers et al., 2020; Andersson et al., 2021) 。

因此，快速检测 PPD 及其相关风险因素对于及时通过专业程序进行评估和干预至关重要，尤其是对于脆弱的妇女 (Wang et al., 2019) 。不幸的是，当前的解决方案影响了 PPD 的早期检测 (Nurbaeti et al., 2021) 。然而，可以通过节约成本且高效的系统实现后者，例如智能助手或聊天机器人 (Dergaa et al., 2024) 。然而，它们在心理健康诊断中的应用相对较新 (Cameron et al., 2019; Duvvuri et al., 2022) 。

值得注意的是，尽管当前技术有所进步，PPD 筛查仍然主要使用传统的基于调查的方法进行（例如，爱丁堡产后抑郁量表 - EPDS ，患者健康问卷-2/9 - PHQ -2/9），这些方法依赖于从产前期间收集的自我报告（即，主观数据）。不幸的是，后一种方法的结果可能无法反映由环境、遗传、激素和心理因素 (Moreira et al., 2019; Stewart & Vigod, 2019) 引起的 PPD 的高度复杂病理生成，或者可能因缺乏认识以及羞耻感等其他相关因素 (Gabrieli et al., 2020) 而产生偏差。

相比之下，机器学习（ML ）有潜力通过使用来自不同来源（例如，非正式笔记和电子健康记录 - EHR ）的最新数据进行精确推断，帮助医疗从业者在决策过程中检测有风险的患者。特别是，ML 算法可以有效地处理大量多维和非线性数据，这是由于它们比传统的线性分析方法具有更强的统计能力。实际上，ML 在个性化医疗保健中使用得越来越频繁，尤其是在心理健康领域。传统的 ML 技术（即监督式、半监督式、无监督式）可以应用于离线（批处理）或在线（流处理）。前者的批处理通过实验数据的训练和测试分区创建固定模型。相反，后者的流处理实时构建增量模型，从而形成一个最新的知识库。因此，我们特别关注于 PPD 的在线检测。

此外，ML 可以从自然语言处理（NLP ）技术中获益，以充分利用人类语言数据，即提取相关的高层推理特征，并将其以最合适的方式转化为计算机可解读的知识 (Sim et al., 2023) 。不幸的是，最新的技术进步，在大型语言模型（LLM ）方面，还未能被充分利用于检测和处理 PPD 。然而，LLM 利用深度学习技术的巨大潜力，在医疗实践中可以非常有用 (Jethani et al., 2023; Andargoli et al., 2024) 。

最后，应注意的是，目前用于 PPD 检测的大多数方法和技术都受到当前流行的黑箱问题的影响。解释智能系统如何得出特定预测可能会显著减少开发人员和最终用户的黑箱问题。后者在医疗保健领域尤其积极，能够生成纠正程序 (Agbavor & Liang, 2022) 。为了解决这一问题，研究界应在可能的情况下选择可解释的模型（例如，基于树的模型如随机森林 - RF ），或通过利用可解释的人工智能（XAI ）(Barredo Arrieta et al., 2020) 来创建专门的模块规避此问题。这对于医疗保健领域特别合适，因为理解诊断的决策过程可能是至关重要的，并且有利于个性化治疗 (Agbavor & Liang, 2022; Göndöcs & Dörfler, 2024) 。

因此，有机会利用最新技术为当前的 PPD 检测现状做出贡献。因此，我们的工作为从自然语言中结合 NLP 、ML 和 LLM 的智能 PPD 筛查系统做出了贡献，该系统实现了一种经济实惠、实时并且非侵入的过程。此外，它解决了黑箱问题，因为预测是通过特征重要性和自然语言解释的。

本文的其余部分组织如下。第 2 节概述了当前的相关竞争 PPD 解决方案，特别关注于应用先进技术。第 3 节介绍了提出的解决方案，而第 4 节描述了实验数据集、实现和设置，以及实证评估结果。最后，第 5 节总结并强调了成果和未来工作。

2. 相关工作

基于 AI 的医疗系统能够实现经济实惠、无创或友好且快速的筛查 (Yu et al., 2018; Aminizadeh et al., 2024) 。我们的研究主要集中在使用 LLM 在医疗领域中。请注意，LLM 展示了出色的自然语言生成（NLG ）和自然语言理解（NLU ）功能 (Agbavor & Liang, 2022) ，在医疗实践中可以得到高度重视 (Jethani et al., 2023) 。

在这一行中，GPT-3.5 、GPT-4 和 GPT-4 是近年来流行的 Chat GPT 助手中使用的模型，与先前提到的 NLG 和 NLU 能力相比，它们得到了改善，相较于 GPT-3 和其他早期模型。文献中已有关于这些改进版智能助手的具体应用，它们可以生成临床笔记 (Cascella et al., 2023) 并提供医学建议 (Ayers et al., 2023) 。然而，当前 LLM s 更注重语义学习，即一般的 NLP 能力。因此，当与 ML 结合使用以为其提供特定知识时，可以进一步增强这些助手，在我们的案例中，这是医学数据，同时克服其知识库中实时数据的缺乏 (Liao et al., 2023) 。

不幸的是，LLM 尚未被充分利用来检测和治疗 PPD 加上 (Liu et al., 2023) 。一个初步但具有代表性的例外是 Sezgin et al. (2023) 的工作。作者评估了 GPT-4 （采用 Chat GPT ）和 L a MDA （使用 Bard）在回答与 PPD 相关的常见问题时的表现。经过认证的医师评估了他们的临床准确性，而 Chat GPT 取得了最佳结果。

尽管缺乏利用上述 LLM 模型的研究，但在使用 ML 进行 PPD 检测方面已经存在扎实的研究，其中一些研究利用了文本输入数据。Moreira et al. (2019) 利用临床、人口和社会数据来预测妊娠期高血压妇女的 PPD 风险。同样，Andersson et al. (2021) 和 Zhang et al. (2021) 在 ML 模型中使用临床和人口特征进行 PPD 检测。第一项工作计算了特征重要性，从而在某种程度上提供了解释性。相反，Prabhashwaree & Wagarachchi (2022) 从一项调查中使用 ML 识别与家庭和社会相关的 PPD 风险因素。在这条线上，Raisa et al. (2022) 比较了不同的 ML 模型，以有效检测孟加拉国女性人口的具体情况。人口和社会数据通过特定的调查手动收集，以分析它们与 PPD 之间的相关性。相比之下，Yang et al. (2022) 开发了一种基于临床和人口的解决方案，用于经历剖宫产的女性的 PPD 筛查。他们证明了使用氯胺酮干预降低 PPD 发病率的有效性。最近，Nakamura et al. (2024) 利用人口信息和主观评级通过 ML 来评估 PPD 。

关于通过 ML 分析文本的 PPD 相关工作，Wang et al. (2019); Hochman et al. (2021); Zhang et al. (2021) 使用了 EHR 的数据，而 Fatima et al. (2019) 利用了从 Reddit 上的社交媒体帖子中提取的语言特征。后者解决方案的类似方法由 Shatte et al. (2020) 采用。在这方面，Trifan et al. (2020) 构建了一个来自 Reddit 的 PPD 相关帖子的语料库，并提出了一种基于 ML 的初步解决方案。

此外，Shin et al. (2020) 提出了一个混合系统。在这个研究中，作者除了借助两道适配自 PHQ-2 问卷关于女性当前情绪状态的问题外，还依赖于生活方式和母亲的人口统计特征来检测 PPD 。此外，Mazumder & Baruah (2021) 使用 ML 模型检测了 PPD ，但利用了装袋和提升技术，并以一项关于背景和心理状态的调查问卷的回答作为输入数据。

用于 PPD 筛查的应用存在，但没有 NLP 功能。这是 Nurbaeti et al. (2021) 的作品，他们开发了一款安卓应用，通过调查来检测 PPD 的早期症状。

尽管这项工作不属于 PPD 研究领域，但与我们自发言语分析的目标相关的是 Duvvuri et al. (2022) 的工作，他使用从 Discord 中收集的数据开发了一种基于 ML 的解决方案来检测抑郁。同样，Dergaa et al. (2024) 分析了在不同精神健康场景（例如，与教育培训相关的焦虑）中 Chat GPT 治疗建议的适当性，包括 PPD 。不幸的是，作者仅限于评估 Chat GPT 作为医疗专业人员协作工具的潜力，并且没有通过提示工程对其性能进行调整。此外，这些建议没有通过 XAI 进行解释。

经过对 PPD 现有技术的广泛修订，我们未能找到解决已经提到的黑箱问题或分析 PPD 的流媒体作品，尽管由于临床记录中大量的连续数据可用，它非常适合。在本工作的目标领域之外，研究界通过特征重要性 (de Arriba-Pérez et al., 2022) 、自然语言 (Ehsan et al., 2019) 和视觉解释 (Spinner et al., 2019) 来解决 AI 潜力的负责使用问题。

应用于 PPD 的 XAI 的少数例外是 Amit et al. (2021) 、Liu et al. (2023) 、Qi et al. (2025) 和 Zhang et al. (2025) 进行的研究，它们都利用了 SHapley Additive exPlanations (SHAP)。首先，Amit et al. (2021) 提出了一种基于 ML 的解决方案，该方案结合 EHR 进行了训练。除此之外，Liu et al. (2023) 创建了一种基于 ML 的 PPD 风险计算器。然而，他们仅考虑了接受剖腹产的女性，并且涉及的特征缩减为人口统计数据。如前所述，后者可能导致输入数据偏差，并在研究中有关的女性无意中引入了自我报告的障碍。最近，Qi et al. (2025) 评估了经典 ML 模型用于 PPD 预测的性能。作者使用了生物心理社会特征。最终，Zhang et al. (2025) 利用了 EMR 来训练不同的 ML 分类器。

令人惊讶的是，尽管 PPD 在我们的社会中十分严重，但在涉及最新 AI 技术（例如，LLM ，XAI ，在线处理）(Andersson et al., 2021; Raisa et al., 2022) 时，当前关于这一主题的研究数量却很少。因此，这项工作旨在通过最新的技术进步帮助从业者进行决策，以实现实时筛查和治疗建议。主要地，我们的工作贡献于一个智能 PPD 筛查系统，该系统结合了 NLP 、ML 和 LLM ，实现对即兴讲话的经济型、实时且非侵入性分析。此外，它解决了黑箱问题，因为通过将 LLM 与可解释的 ML 模型（即基于树的算法）结合使用特征重要性和自然语言，向最终用户描述了预测。最终，它有助于快速检测 PPD 及其相关的风险因素，这对于及时和适当的评估和干预至关重要。

2.1. 贡献

根据在表 1 中展示的与文献中最相关的工作的比较，我们的解决方案首次解决了同时考虑在线 PPD 检测和 LLM 以及提供可解释性这一研究空白。尽管使用了最新的技术，我们提出的流程仍然是这一工作的原创贡献。此外，我们的研究认识到透明度在心理健康筛查中的重要性，这是在临床环境中培养最终用户信任并促进医生做出知情决策的关键。

Table 1　比较最相关的 PPD 预测解决方案，考虑应用领域（QA：问答）、所采用的方法、数据处理（离线、在线）、所用输入数据（FS：自由语音，Demogr.：人口统计）、以及可解释性能力（Ex.，FI：特征重要性，NL：自然语言）。

Authorship	Domain	Approach	Processing	Data	Ex.
Amit et al. (2021)	PPD	ML	Offline	EHRs + EPDS	SHAP
Andersson et al. (2021)	PPD	ML	Offline	Clinical + demographic	FI
Duvvuri et al. (2022)	Depression	ML	Offline	FS	✗
Liu et al. (2023)	PPD	ML	Offline	Demographic	SHAP
Sezgin et al. (2023)	PPD-QA	ChatGPT	-	FS	✗
Qi et al. (2025)	PPD	ML	Offline	Biopsychosocial	SHAP
Zhang et al. (2025)	PPD	ML	Offline	EMR	SHAP
Proposed	PPD	ChatGPT+ML	Online	FS	FI,NL

3. 系统架构

图 1 展示了解决方案的方案。它包括一个聊天机器人应用程序（章节 3.1 ）用于与终端用户建立自由对话，以及一个通过提示工程提取特征的特征提取模块。后者的特征随后在流式数据处理模块中进行处理，该模块涉及特征工程（章节 3.2 ）和特征分析与选择（章节 3.3 ），并由流式分类模块（章节 3.4 ）进行分类。最后，利用最相关的特征为终端用户提供可解释的自然语言描述（章节 3.5 ）。

3.1. 聊天机器人应用 & 特征提取

该聊天机器人被开发为一个多平台应用程序（即，兼容 Android 和 iOS 的平板电脑和智能手机，以及适用于 macOS 和 Windows 的网页部署），能够进行自由和自然的对话。它可以通过提示工程技术生成有同情心和 PPD 导向的言语。更具体地，在与终端用户的对话中，聊天机器人提出的问题涉及以下主题：(i) 与婴儿的情感联系问题，(ii) 注意力集中和决策问题，(iii) 感到悲伤或容易流泪，(iv) 内疚，(v) 对婴儿或伴侣的易怒，(vi) 过度反应或食欲不振，(vii) 自杀行为，以及 (viii) 睡眠困难。

根据分析目标存在三种提示：(i) 主题发现，(ii) 根据主题进行移情文本生成，以及 (iii) 提出护理治疗方案。注意，当对话中已经涵盖某个主题时，该主题将从主题列表中移除以避免重复。此外，利用 LLM 的 `temperature` 参数来实现对话自然度。该参数支持对生成文本的随机性进行调整 [2] 。

[2] `Temperature` = 1 表示完全随机，而 0 代表确定性。

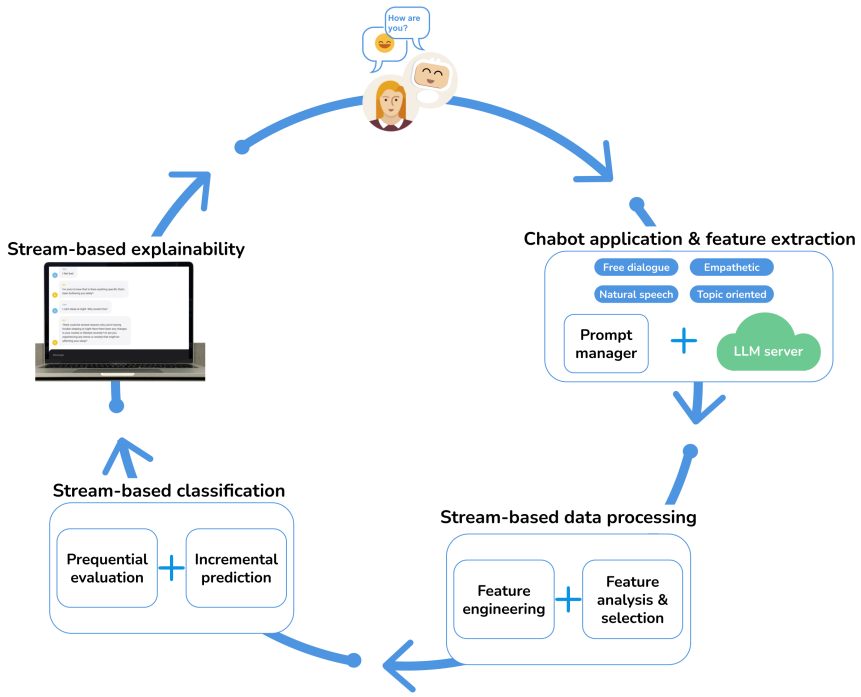

Figure 1　由（i）聊天机器人应用程序 & 特征提取，（ii）包含特征工程和特征分析 & 选择的流式数据处理，（iii）流式分类，和（iv）流式可解释性模块组成的系统图。

3.2. 特征工程

这个过程是通过提示工程来执行的。后者试图解释用户的回应并将其翻译为以下选项：`NA`、[3]、`yes`、`sometimes`、`often`、`no`、`unwilling to disclose`。然后，每个主题的解释都与一个分类值相关联。此外，聊天机器人应用程序会检索用户的年龄。

请注意，相较于转换为有序等级，二值化方法更适合于特征工程，因为所涉及的主题之间不存在代数关系。完整的过程详见算法 1 。

3.3. 特征分析 & 选择

ML 模型的性能直接受到实验数据质量的影响。后者在流处理场景中尤为重要，因为样本会不断到达并需要实时处理。此外，在在线分类问题中特征选择的具体情况下，相关特征可能随时间而变化。因此，需要适当的特征分析和选择程序。

在这项研究中，系统遵循方差阈值法删除不相关的特征。因此，方差值低于可配置阈值的特征会被舍弃，不进行进一步分析。阈值的计算考虑了实验数据中冷启动子集中特征的方差和一个可配置的百分位数（见算法 2 ）。

[3]不适用，超出范围。

Algorithm 1 用户响应编码。

Require: *dataset*

for *topic* in *dataset.columns* **do**
  unique_values = *set*(*dataset.get_column*(*topic*))
  for *option* in *unique_values* **do**
    column_name = *topic* + "_" + *option*
    dataset[*column_name*].*rows* = 0
  end for
  for *index* in *dataset.get_column*(*topic*) **do**
    value = *dataset*[*index*][*topic*]
    column_name = *topic* + "_" + *value*
    dataset[*index*][*column_name*] = 1
  end for
  dataset.remove_column(*topic*)
end for
return *dataset*

Algorithm 2 方差阈值计算。

Require: *dataset* , *percentile*

list_variances = []
for *column* in *dataset.columns* **do**
  variance = *dataset*[*column*].*var*()
  list_variances.add(*variance*)
end for
variance_threshold = *get_percentile*(*percentile*, *list_variances*)
return *variance_threshold*

3.4. 基于流的分类

我们基于流的分类问题的目标变量是 PPD 的缺失或存在。为此，使用以下 ML 模型：

- 高斯朴素贝叶斯（GNB ）(Tieppo et al., 2021) 。该模型基于传统的朴素贝叶斯算法，分析高斯概率分布，但配置为基于流的分类。它被用作基准参考。
- 逻辑回归（ LR ）(Montiel et al., 2021) 。该模型使用一个 S 型函数来评估样本属于特定类别的可能性。作为最简单的线性分类技术之一，尽管其准确性在复杂问题中可能有限，但它提供快速的执行时间。
- 近似大间隔算法（ALMA ）(Silva & Villela, 2021) 。它类似于支持向量机的批量版本。它通过近似超平面与范数之间的最大间隔（值为 $p \geq 2$ ）来计算一组线性可分数据的类别概率。
- Hoeffding 自适应树分类器（HATC ）(Mrabet et al., 2019) 。它也用于流式分类，并利用单一树结构中的分支性能评估机制。
- 自适应随机森林分类器（ARFC ）(Fatlawi & Kiss, 2020) 。该模型由于采用了集成策略和通过多数投票获得的预测，代表了后续 HATC 算法的改进。

准确率、曲线下面积（AUC ）、精确率、召回率和 F -测量用作评估用途。虽然精确率评估乐观预测的准确性（总正预测中真实正例的比例），但召回指标阐释了模型识别所有正实例的敏感性（检测出的真实正例占已确认正例总数的比例）(Rainio et al., 2024) 。因此，这两个指标之间存在权衡关系，在 F -测量中得到总结。一个更保守的模型就是精确率优于召回率的模型 (Ho et al., 2020) 。也就是说，为确保预测正确，即使未检测到所有正例。这种恢复率由召回率表示，可以扩大，但这样可能会增加假阳性的出现，从而降低精确率。在临床应用中，假阳性是错误识别疾病的情况。这种情况会导致不必要的治疗，可能影响患者的健康状况（例如，侵入性手术）及后续成本。相反，假阴性则是未检测到现有疾病的情况。后者的后果也需考虑（例如，缺乏早期和及时的治疗）。文献中的前人工作最大化了罕见疾病中的精确率，而在严重疾病中则倾向于召回率 (Lindsay et al., 2024) 。

3.5. 基于流的可解释性

可解释性数据包括有关提取出的最相关特征的信息，以便在预测透明度方面为终端用户提供有价值的见解。为了提取这些信息，我们应用了一种称为反事实解释的技术。它评估特征中的小扰动，这些扰动改变了预测的类别。在内部，这些排列在 ML 模型的决策路径中产生了变动（即，树路径中的大于和小于分叉），从而实现预测概率的改变。最终，系统考虑导致这种改变的特征为相关特征。算法 3 描述了反事实解释的过程。它迭代若干次，寻找使受影响特征数量最小化的配置。

Algorithm 3 反事实解释技术。

Require: $model$, x , $predicted\_label$, $n_{iterations}$

$x\_new = []$
$relevant\_features$
for $i = 1$ to $n_{iterations}$ **do**
  for each $feature$ in x **do**
    if predicted_label $= 0$ **then**
      $new\_value = random\_choice\_No()$
    else if predicted_label $= 1$ **then**
      $new\_value = random\_choice\_Yes()$
    end if
    $x\_new[feature] = new\_value$
  end for
  $pred\_new = model.predict\_one(x\_new)$
  $pred\_proba\_new = model.predict\_proba\_one(x\_new)$
  if $pred\_new != predicted\_label$ and $pred\_proba\_new > 0.5$ **then**
    $features\_modified = instance\_distance(x, x\_new)$
    if $len(features\_modified) < len(relevant\_features)$ **then**
      $relevant\_features = features\_modified$
      $pred\_proba\_final = pred\_proba\_new$
      $x\_final = x\_new$
    end if
  end if
end for
return $relevant\_features, pred\_proba\_final, x\_final$

4. 评估与讨论

所有实验均在具有以下规格的计算机上执行:

- 操作系统: Ubuntu 18.04.2 LTS 64 位
- 处理器: IntelCore i9-10900K 2.80 GHz
- 内存: 96 GB DDR4
- 磁盘: 480 GB NVME + 500 GB SSD

4.1. 实验数据集

使用了两个数据集 [4] : (i) 一个合成数据集 [5] , 由 200 个用户的与第 3.1 节描述的话题相关的语句组成, 并与用户的响应选项相关联, 排除了 NA 类别 [6] ; (ii) 一个公开可用的数据集 [7] , 其中包括 1491 用户对医

[4]本研究不涉及人类参与者。因此, 不需要伦理审批。
[5]可在 https://doi.org/10.5281/zenodo.14049633 处获取, 2025 年 5 月。
[6]每个主题和选项 5 个话语。
[7]可在 https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression 获取, 2025 年 5 月。

学调查的响应（排除 NA 类别）[8]，涵盖了所分析的话题。第一个数据集评估系统在解释用户响应时的 NLU 能力。相反，第二个数据集旨在用于基于流的 PPD 分类。

在第一个数据集中，分析的八种症状均匀分布，每种症状各有 25 个话语。更详细地说，图 2 详细说明了用户在第二个数据集中的响应分布。此外，关于第二个数据集中医疗记录的人口统计特征，研究中所涉及女性的年龄分布如图 3 所示。

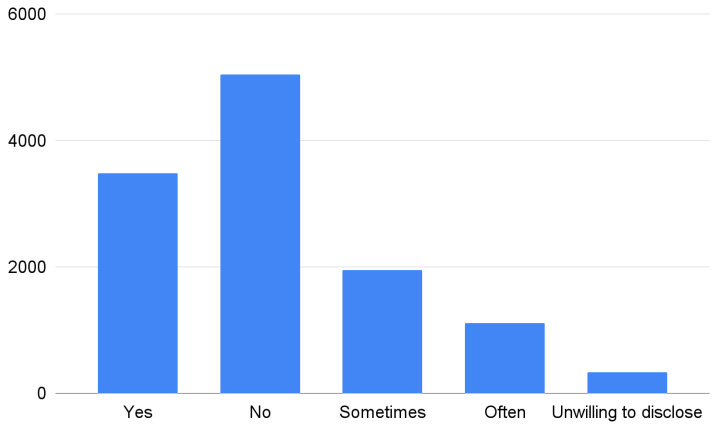

Figure 2　用户的响应分布。

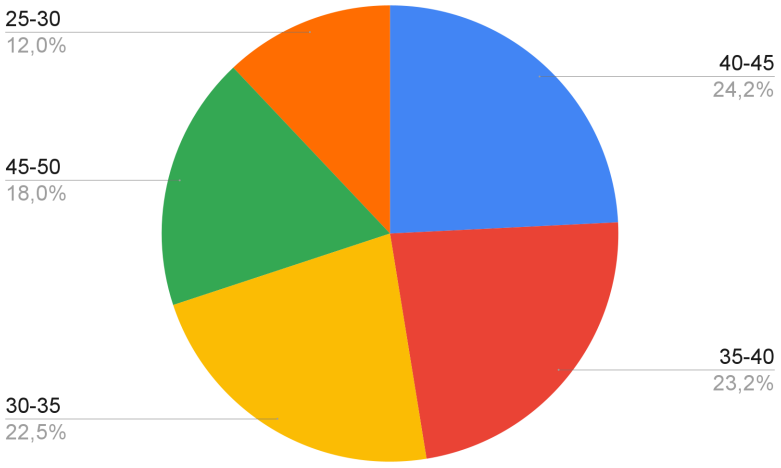

Figure 3　第二个数据集中的年龄分布。

最后，表格 **??** 显示了第二个数据集中的目标变量分布。

这个合成数据集是为了模拟通常在产后情况下观察到的语音模式和情绪表达生成的，特别关注与抑郁症状相关的词汇、韵律和情感特征。虽然它旨在接近真实世界的产后语音，合成数据可能无法完全捕获患者数

[8]注意，用户的响应遵循之前解释过的方案：`NA`，`yes`，`sometimes`，`often`，`no`，`unwilling to disclose`。

Table 2　第二个实验数据集中的类别分布。

Category	Number of samples
Absence of PPD	523
Presence of PPD	968
Total	1491

据的复杂性、细微差别和变异性。因此，我们结合了两个数据集，一个是合成生成的，另一个是公开可用的。关于数据集的代表性或普遍性局限性，尽管对控制实验有用，但合成数据可能无法完全捕捉真实产后语音的语言、情感或文化细微差别。相比之下，公共数据集可能在年龄、种族、语言种类和社会经济多样性方面存在限制，影响其对更广泛人群的适用性。

4.2. 聊天机器人应用 & 特征提取

多平台对话助手是使用 Flutter 编程框架设计的 [9] 。聊天机器人应用是在一个使用 Flask（版本 2.2.2）实现的服务器中用 Python （版本 3.8）[10] 进行编写的，该服务器包括一个使用 Gunicorn（版本 20.1.0）[11] 的流量均衡器，而 Flyer Chat 库 [12] 被用来为系统提供即时消息功能。Flask 服务器中的聊天机器人应用程序将提示和话语发送到使用的 LLM 模型，Chat GPT （版本 3.5）[13] ，通过 Open AI REST API [14] 。

请注意，保留最近 10 次交互的历史记录，以实现完全连贯一致的对话。一旦超过这个数量，最早的交互将被删除，类似于 FIFO 堆栈的类比。表 3 显示了设计的提示和温度配置。

此外，图 4 展示了一个真实的用户-聊天机器人对话，其中系统检测到用户有时感到悲伤、疲惫和入睡困难。

在利用第一个数据集解释用户响应时，系统性能的评估实现了 88.50 % 的准确率。此评估的混淆矩阵如图 5 所示。请注意，错误是有限的，系统仅在 4 个情况下未能理解用户的语句（即，将他们的响应与 NA 相关联）。表 4 详细列出了评估结果，总体性能约为 90 % 。

4.3. 特征工程

二值化过程使用了 Python `pandas` 库 [15] 中的 `get_dummies` 函数。总共生成了 53 个布尔特征，其中 48 个布尔特征对应 8 个主题、6 个用户响应解释选项和 5 个额外的年龄特征 [16] 。在图 6 中，我们可以观察到按分析主题划分的用户响应分布。

[9] 可在 `https://flutter.dev` 获得，2025 年 5 月。
[10] 可在 `https://www.python.org` 获取，2025 年 5 月。
[11] 可在 `https://gunicorn.org` 获取，2025 年 5 月。
[12] 将于 `https://flyer.chat` 在 2025 年 5 月提供。
[13] 可在 `https://platform.openai.com/docs/models` 查询，2025 年 5 月。
[14] 可在 `https://platform.openai.com` 获得，2025 年 5 月。
[15] 可在 `https://pandas.pydata.org` 获取，2025 年 5 月。
[16] 25 岁到 50 岁之间的 5 年间隔。

Table 3　提示工程

Prompt ID	Prompt content	Temperature
1	Analyze the user's responses in the dialogue and return per topic one of the following options: `NA`, `yes`, `sometimes`, `often`, `no`, `unwilling to disclose`. Topic 1: baby bonding issues, topic 2: concentration and decision-making problems, topic 3: feeling sad or tearful, topic 4: feeling guilty, topic 5: irritability towards the baby or the partner, topic 6: overreacting or loss of appetite, topic 7: suicide behavior, topic 8: trouble sleeping.	0
2	The following is a conversation with an AI assistant. The assistant is dynamic, never repeats the same thing twice, and is creative, intelligent, and kind. The assistant tries to establish a conversation, and it always asks a question when he finishes speaking about one of these topics: topic 1: baby bonding issues, topic 2: concentration and decision-making problems, topic 3: feeling sad or tearful, topic 4: feeling guilty, topic 5: irritability towards the baby or the partner, topic 6: overreacting or loss of appetite, topic 7: suicide behavior, topic 8: trouble sleeping. Return an utterance per topic.	1
3	Propose three care treatments for a user with PPD taking into consideration its responses into the following dialogue.	1

Table 4　用户响应的解释结果。

Class	**Accuracy**	**Precision**	**Recall**	**F-measure**
Overall	88.50	92.13	88.50	89.45
Yes	-	**100.00**	75.00	85.71
Sometimes	-	76.92	**100.00**	86.96
Often	-	83.72	90.00	86.75
No	-	**100.00**	82.50	90.41
Unwilling to disclose	-	**100.00**	95.00	**97.44**

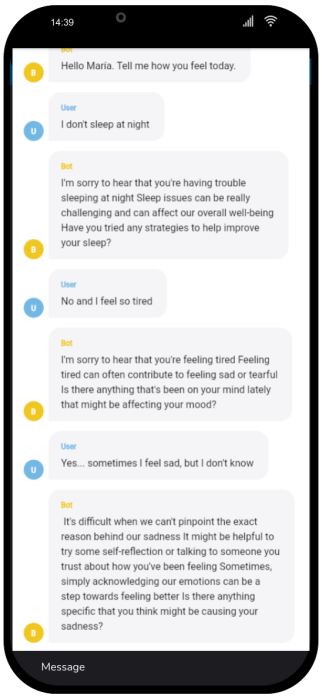

Figure 4　真实对话示例。

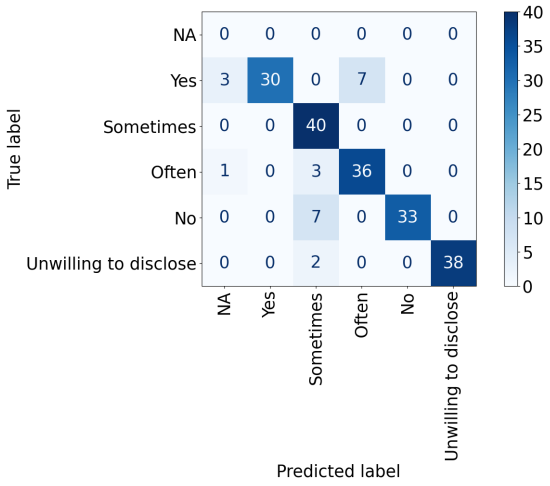

Figure 5　用户响应解释评估的混淆矩阵。

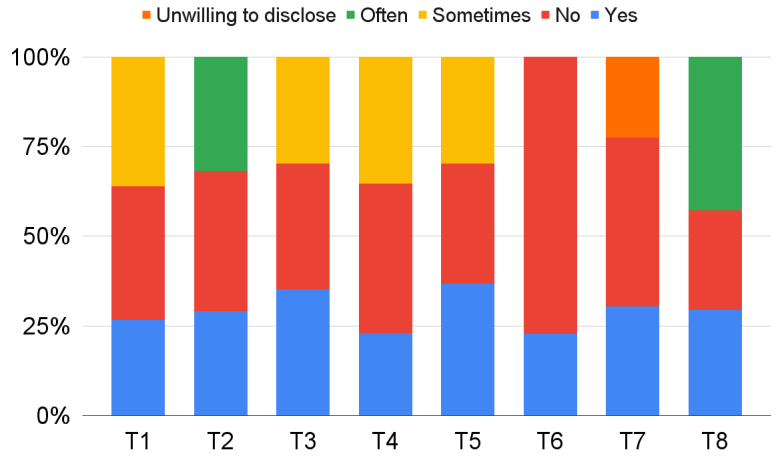

Figure 6　主题和响应分布。

4.4. 特征分析 & 选择

注意，方差阈值被设定为实验数据集 (Cao et al., 2022) 中 10 % 所含特征方差的第 5 百分位数值 (T. et al., 2021) 。后者代表特征选择方法的冷启动设置。因此，阈值被设为 0.079 以剔除方差波动低的特征。

`River` 库中的 `VarianceThreshold` [17] 方法用于计算特征方差。

在图 7 中，我们展示了在线系统最后处理样本中各种特征的方差值。可以看出，所有特征都符合资格标准，并被认为是相关的。

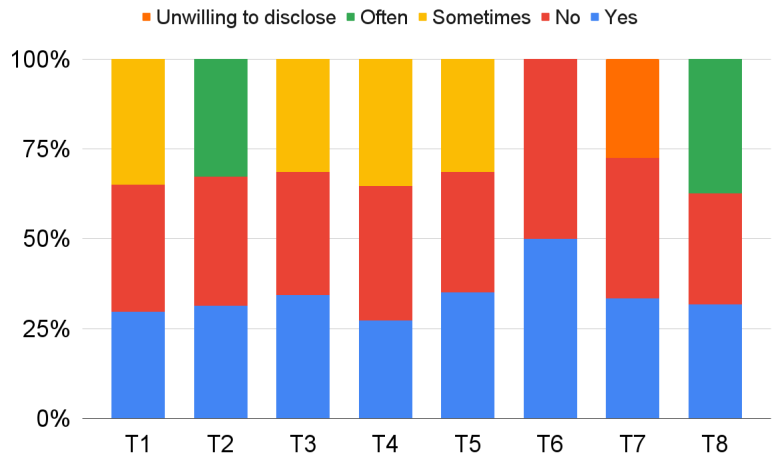

Figure 7　每个主题的方差和用户响应。

[17] 可在 `https://riverml.xyz/0.11.1/api/feature-selection/VarianceThreshold` 获得，2025 年 5 月。

4.5. 基于流的分类

基于河流库实现了基于流的 ML 模型：GNB [18] 、LR [19] 、ALMA [20] 、HATC [21] 和 ARFC [22] 。清单 1 、2 、3 和 4 详细列出了用于优化超参数的范围，其中加粗的值用于 LR 、ALMA 、HATC 和 ARFC 算法，分别为 [23] 。

Listing 1 LR hyperparameter configuration.

```
l2 = [0.0, 0.1, 1.0]
interceptlr = [ 0.001, 0.01, 0.1]
```

Listing 2 ALMA hyperparameter configuration.

```
alpha = [0.5, 0.7, 0.9]
B = [0.6, 1.0, 1.4]
C = [1.0, 1.4, 1.8]
```

Listing 3 HATC hyperparameter configuration.

```
depth = [None, 50, 200]
tiethreshold = [0.5, 0.05, 0.005]
maxsize = [50, 100, 200]
```

Listing 4 ARFC hyperparameter configuration.

```
models = [10, 50, 100]
features = [sqrt, 5, 50]
lambda = [10, 50, 100]
```

表 5 比较了三个模型的结果。可以观察到，GNB 在 70 % 到 80 % 范围内的评估指标中表现出一致的值。ALMA 分类器在 80 % 到 90 % 范围内显示了较好的常规结果。更不稳定的结果由 LR 获得（例如，PPD 缺失的召回率和 F -衡量），后者也适用于 HATC 。最终，ARFC 是获得最佳结果的模型，超过了 90 % 阈值。

[18] 可在 https://riverml.xyz/dev/api/naive-bayes/GaussianNB 获得，2025 年 5 月。
[19] 可在 https://riverml.xyz/0.11.1/api/linear-model/LogisticRegression 获取，2025 年 5 月。
[20] 可在 https://riverml.xyz/0.11.1/api/linear-model/ALMAClassifier 获取，2025 年 5 月。
[21] 可在 https://riverml.xyz/0.13.0/api/tree/HoeffdingAdaptiveTreeClassifier 获取，2025 年 5 月。
[22] 可在 https://riverml.xyz/0.11.1/api/ensemble/AdaptiveRandomForestClassifier ，2025 年 5 月获得。
[23] 请注意，后面的超参数优化过程不适用于基线 GNB 算法。

Table 5　在最后一个样本之后进行连续预测结果（# 0 和 # 1 分别表示 PPD 的不存在和存在）。

Model	Acc.	AUC	Precision			Recall			F-measure			Time (s)
			Macro	# 0	# 1	Macro	# 0	# 1	Macro	# 0	# 1	
GNB	80.05	77.26	78.32	73.35	83.28	77.26	67.88	86.65	77.72	70.51	84.93	0.26
LR	78.52	70.94	81.88	87.18	76.58	70.94	45.51	96.38	72.57	59.80	85.35	0.15
ALMA	86.85	85.70	85.51	80.91	**90.11**	85.70	**81.84**	89.56	85.60	81.37	89.83	0.10
HATC	75.35	70.00	73.65	70.10	77.20	70.00	52.01	87.99	70.98	59.71	82.24	0.49
ARFC	**91.40**	**88.82**	**92.32**	**94.58**	90.06	**88.82**	80.11	**97.52**	**90.19**	**86.75**	**93.64**	34.28

如表 **??** 所示，在实验数据中，相较于对照组，阳性病例的数量显著增加。为了分析所提出的解决方案是否受到可能导致潜在误报病例的数据不平衡的影响，表 6 详细展示了在均衡设置下获得的结果。为此，我们采用对少数类进行下采样技术的均衡重采样方法（即 523 个有和没有 PPD 的病例）。如预期的结果，在均衡情境下结果更好。在 GNB 和 HATC 模型的情况下，只有检测到 PPD 的精度受到影响，而对于 LR ，则是召回率受到影响。相比之下，ALMA 模型获得了更稳定的数值。类似地，ARFC 在评估指标上有轻微的变化，证明其相较其他模型更具鲁棒性。更重要的是，评估结果在不平衡和均衡设置中是一致的。因此，由于所采用的通过多数投票获得预测的集成策略，我们的方案避免了潜在的误报问题。

Table 6　在流式环境下，最后一个样本后的平衡设置中的预测结果（# 0 和 # 1 分别表示 PPD 不存在和存在）。

Model	Acc.	AUC	Precision			Recall			F-measure			Time (s)
			Macro	# 0	# 1	Macro	# 0	# 1	Macro	# 0	# 1	
GNB	78.64	78.64	79.09	82.71	75.47	78.64	72.41	84.87	78.56	77.22	79.89	0.18
LR	78.66	78.65	79.23	83.30	75.17	78.65	71.65	85.66	78.55	77.03	80.07	0.11
ALMA	84.59	84.59	84.59	84.51	84.67	84.59	84.67	84.51	84.59	84.59	84.59	0.07
HATC	75.10	75.10	77.29	85.03	69.55	75.10	60.92	89.27	74.59	70.98	78.19	0.34
ARFC	**90.13**	**90.13**	**90.18**	**91.49**	**88.87**	**90.13**	**88.51**	**91.76**	**90.13**	**89.97**	**90.29**	21.79

最后，表格 7 将这些结果与文献中最相关的竞争工作进行了比较。需要注意的是，Sezgin et al. (2023) 的研究是非 ML 研究，并且没有提供分类指标。考虑到之前提到的理论差异（例如，没有研究进行在线处理或用自然语言提供可解释性），我们的提议是第一个结合 LLM 和 ML ，并使用自由语音输入数据进行 PPD 检测的方案。

准确性最高的差异在于 Andersson et al. (2021) 的工作，+ 17.90 % 分。在 AUC 指标 (+ 21.68 % 分相比于 Duvvuri et al. (2022))、召回率 (+ 26.02 % 分相比于 Duvvuri et al. (2022)) 和精确度 (+ 28.77 % 分相比于 Andersson et al. (2021)) 上，性能提升更大。而在 F -测量上相较于 Liu et al. (2023) 取得了更大的提升 (+ 40.19 % 分)。相反，最小的差异在于与 Amit et al. (2021) 和 Zhang et al. (2025) 关于 AUC 指标的研究，虽然对于最后一项工作，准确性差异超过了 10 % 分，有利于我们的方案。关于 Qi et al. (2025) 的研究，精确度、召回率和 F -测量的提高分别为 + 14.52 % 分、+ 9.07 % 分和 + 18.69 % 分。

Table 7　与相关研究结果的比较。

Authorship	Accuracy	AUC	Precision	Recall	F-measure
Amit et al. (2021)	-	84.40	-	78.00	-
Andersson et al. (2021)	73.50	80.95	63.55	73.50	-
Duvvuri et al. (2022)	88.21	67.14	-	62.80	-
Liu et al. (2023)	74.20	-	-	73.10	50.00
Qi et al. (2025)	-	85.80	77.80	79.75	71.50
Zhang et al. (2025)	81.30	84.90	78.65	79.00	-
Proposed	**91.40**	**88.82**	**92.32**	**88.82**	**90.19**

4.6. 基于流的可解释性

为了生成反事实解释，我们使用那些通过 `Predict_Proba_One` 函数预测结果超过 80 % 的样本。这些样本通过在每个特征上引入变化来进行修改，直到预测结果改变为超过 50 % 。此外，这些排列经过精心设计，以避免与模型逻辑相矛盾的变化。

列表 5 显示了使用反事实解释算法评估的两个样本的结果。第一个例子显示了将 PPD 的正预测转为不存在所需的变化。相反，第二个例子说明了相反的情况。最重要的特征是那些使分类器预测超过 50 % 阈值的特征（用粗体表示），而其余特征对输出的影响可以忽略不计。

Listing 5　Counterfactual explanations examples.

```
Presence of PPD (84.26


Age : 30-35 -> 35-40
Baby bonding issues : Sometimes -> No
Concentration and decision-making problems: No
Feeling sad or tearful: Sometimes
Feeling guilty: No
Irritable towards the baby or the partner : Yes -> No
Overeating or loss of appetite : Yes -> Sometimes
Suicide behavior: Unwilling to disclose
Trouble sleeping : Sometimes -> No


Absence of PPD (83.75
```

Age: 30-35

Baby bonding issues : No -> Yes

Concentration and decision-making problems : No -> Yes

Feeling sad **or** tearful: Yes

Feeling guilty : Yes -> Sometimes

Irritable towards the baby **or** the partner: Sometimes

Overeating **or** loss of appetite: Yes

Suicide behavior : No -> Yes

Trouble sleeping: Often

图 8 展示了一个与终端用户的真实对话，包括具体情况的可解释性数据。请注意，系统提供了关于决策过程中最相关特征的信息。解释性结果作为对话的一部分被插入，最后使用表 3 中的第三个提示生成建议，以在检测到 PPD 时降低压力水平。

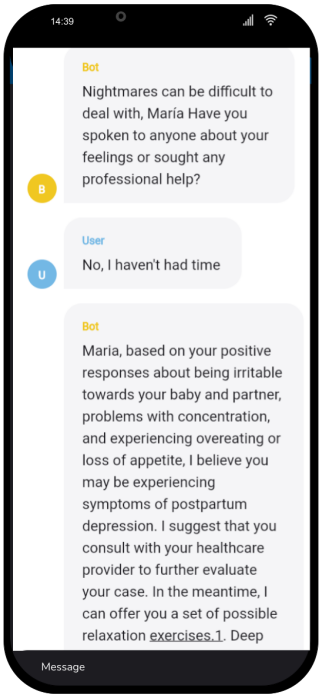

Figure 8　可解释性信息插入到实际对话中。

5. 结论

女性的身心健康受到怀孕和分娩的显著影响。然而，目前的解决方案在早期检测 PPD 方面存在妥协。令人惊讶的是，尽管 PPD 在我们的社会中十分严重，但当前关于这一主题的研究数量很少，尤其是在最新的 AI 技术方面（例如，LLM 、XAI 、在线处理）。

因此，我们的研究通过结合 NLP 、ML 和 LLM s 开发了一种智能的 PPD 筛选系统，该系统能够从自发的言语中以经济实惠、实时、有创伤方式实施程序。此外，它解决了黑箱问题，因为预测通过特征重要性和自然语言来解释。结果表明，在 PPD 检测的所有评价指标上都取得了优异的 90 % ，超越了文献中的竞争解决方案。最终，我们的解决方案有望通过精确的估计帮助医疗从业人员在决策过程中检测高风险患者。

我们的工作在医疗保健方面的应用指向孕产妇健康护理环境，特别是在产前诊所和社区基础的孕产妇项目中。该系统可以集成到移动健康平台中，使得医疗保健专业人员在患者访问期间能够实时使用。系统的输出将作为决策支持工具，帮助临床医生筛查早期不良信号，优先处理患者分诊，甚至在风险水平超过预定阈值时指导干预。这种集成通过关注高风险案例，尤其是在服务不足或资源有限的环境中，提高了早期检测能力并帮助优化临床资源。从患者的角度来看，所提议的系统提供了几个关键优势。首先，通过持续监测实现较早且更精确的潜在健康风险检测可以导致及时的临床干预。其次，将系统集成到健康平台中可能增加对医学建议的参与和遵从。第三，在资源有限或乡村地区无法获取专科孕产妇护理时，该系统可以作为远程监控工具，减少不必要的诊所访问，同时确保高风险案例得到及时关注。

在实际应用中，必须明确获得所有用户的知情同意后才能收集或处理数据。这包括沟通系统的目的、所收集数据的性质、其使用方式、存储和保护措施，以及提供用户随时撤回的选项。关于隐私，设备上的数据处理和任何传输数据的强加密协议确保没有个人可识别信息被泄露。为了确保责任归属，系统决策，尤其是可能影响临床护理的决策，应始终由合格的健康专业人员进行审核。在这方面，可解释性模块允许用户和临床医生理解输出结果背后的理由。最后，通过对来自不同用户群体的模型性能进行评估并不断进行新数据更新，有助于减轻偏见。

尽管系统强调可解释性，我们也认识到确保积极和敏感的用户体验至关重要，特别是对于经历情绪困扰的产后女性。在这一方面，我们的提案旨在作为辅助决策工具而不是诊断工具来运行。系统不会存储可识别的个人数据，从而降低隐私风险。在未来的工作中，我们计划通过从终端用户收集的实验数据分析解决方案的同理心能力。此外，我们计划考虑其他语言，比如西班牙语，以及新功能如何影响 PPD 预测。此外，将考虑替代分类视角（例如，强化学习、无监督分类）。将分析与系统当前版本相关的计算和存储负载，以确保其有效性能。还应考虑衡量系统的可接受性和安全性。最终，临床专家将被邀请验证特征或聊天机器人提示的解释，以解决数据集代表性或普适性方面的限制。

6.

披露声明

作者声明与本文内容相关的无竞争利益。

7.

资助

本工作部分得到了加利西亚自治区政府资助项目 ED481B-2022-093 和 ED481D 2024/014 的支持。

8.

开放研究者与贡献者 ID

Silvia García Méndez `https://orcid.org/0000-0003-0533-1303` , Francisco de Arriba Pérez `https://orcid.org/0000-0002-1140-679X` 。

9.

作者贡献声明

所有作者都已阅读并批准了稿件的最终版本。Silvia García-Méndez：概念化、方法学、软件、验证、正式分析、调查、资源、数据管理、撰写 - 原始草稿、撰写 - 审阅以及编辑、可视化、资金获取。Francisco de Arriba-Pérez：概念化、方法学、软件、验证、正式分析、调查、资源、数据管理、撰写 - 原始草稿、撰写 - 审阅以及编辑、可视化、资金获取。

10.

数据和代码可用性声明

实验数据可以在 `https://doi.org/10.5281/zenodo.14049633` 和 `https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression` 公开获取。代码可以在 `https://github.com/nlpgti/postpartum_stress` 公开获取。

References

Abadiga, M. (2019). Magnitude and associated factors of postpartum depression among women in Nekemte town, East Wollega zone, west Ethiopia, 2019: A community-based study. *PLOS ONE*, *14*, 1–15. doi:`10.1371/journal.pone.0224792`.

Agbavor, F., & Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, *1*, 1–14. doi:10.1371/journal.pdig.0000168.

Aminizadeh, S., Heidari, A., Dehghan, M., Toumaj, S., Rezaei, M., Navimipour, N. J., Stroppa, F., & Unal, M. (2024). Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service. *Artificial Intelligence in Medicine*, (pp. 102779–102804). doi:10.1016/j.artmed.2024.102779.

Amit, G., Girshovitz, I., Marcus, K., Zhang, Y., Pathak, J., Bar, V., & Akiva, P. (2021). Estimation of postpartum depression risk from electronic health records using machine learning. *BMC Pregnancy and Childbirth*, *21*, 630. doi:10.1186/s12884-021-04087-8.

Andargoli, A. E., Ulapane, N., Nguyen, T. A., Shuakat, N., Zelcer, J., & Wickramasinghe, N. (2024). Intelligent decision support systems for dementia care: A scoping review. *Artificial Intelligence in Medicine*, (pp. 102815–102826). doi:10.1016/j.artmed.2024.102815.

Andersson, S., Bathula, D. R., Iliadis, S. I., Walter, M., & Skalkidou, A. (2021). Predicting women with depressive symptoms postpartum with machine learning methods. *Scientific Reports*, *11*, 7877–7881. doi:10.1038/s41598-021-86368-y.

de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F. J., & González-González, J. (2022). Explainable machine learning multi-label classification of Spanish legal judgements. *Journal of King Saud University - Computer and Information Sciences*, *34*, 10180–10192. doi:10.1016/j.jksuci.2022.10.015.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, *183*, 589–596. doi:10.1001/jamainternmed.2023.1838.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi:10.1016/j.inffus.2019.12.012.

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O' Neill, S., Armour, C., & McTear, M. (2019). Assessing the Usability of a Chatbot for Mental Health Care. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 121–132). Springer volume 11551 LNCS. doi:10.1007/978-3-030-17705-8_11.

Cao, J., Sheng, J., Cong, X., Liu, T., & Wang, B. (2022). Cross-Domain Recommendation to Cold-Start Users via Variational Information Bottleneck. In *Proceedings of the IEEE 38th International Conference on Data Engineering* (pp. 2209–2223). IEEE volume 2022-May. doi:10.1109/ICDE53745.2022.00211.

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *Journal of Medical Systems*, *47*, 1–5. doi:10.1007/s10916-023-01925-4.

Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., Ben Aissa, M., Souissi, N., Guelmami, N., Swed, S. et al. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, *14*, 1277756–1277767. doi:10.3389/fpsyt.2023.1277756.

Duvvuri, V., Guan, Q., Daddala, S., Harris, M., & Kaushik, S. (2022). Predicting Depression Symptoms from Discord Chat Messaging Using AI Medical Chatbots. In *Proceedings of the International Conference on Machine Learning and Soft Computing* (pp. 111–119). Association for Computing Machinery. doi:10.1145/3523150.3523168.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the International Conference on Intelligent User*

Interfaces (pp. 263–274). Association for Computing Machinery volume Part F147615. doi:10.1145/3301275.3302316.

Falana, S. D., & Carrington, J. M. (2019). Postpartum Depression: Are You Listening? *Nursing Clinics of North America*, *54*, 561–567. doi:10.1016/j.cnur.2019.07.006.

Fatima, I., Abbasi, B. U. D., Khan, S., Al-Saeed, M., Ahmad, H. F., & Mumtaz, R. (2019). Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, *36*, 1–13. doi:10.1111/exsy.12409.

Fatlawi, H. K., & Kiss, A. (2020). *On Robustness of Adaptive Random Forest Classifier on Biomedical Data Stream* volume 12033 LNAI. Springer. doi:10.1007/978-3-030-41964-6_29.

Gabrieli, G., Bornstein, M. H., Manian, N., & Esposito, G. (2020). Assessing Mothers' Postpartum Depression From Their Infants' Cry Vocalizations. *Behavioral Sciences*, *10*, 55–64. doi:10.3390/bs10020055.

Göndöcs, D., & Dörfler, V. (2024). AI in medical diagnosis: AI prediction & human judgment. *Artificial Intelligence in Medicine*, *149*, 102769–102780. doi:10.1016/j.artmed.2024.102769.

Ho, S. Y., Wong, L., & Goh, W. W. B. (2020). Avoid Oversimplifications in Machine Learning: Going beyond the Class-Prediction Accuracy. *Patterns*, *1*, 100025. doi:10.1016/j.patter.2020.100025.

Hochman, E., Feldman, B., Weizman, A., Krivoy, A., Gur, S., Barzilay, E., Gabay, H., Levy, J., Levinkron, O., & Lawrence, G. (2021). Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. *Depression and Anxiety*, *38*, 400–411. doi:10.1002/da.23123.

Jethani, N., Jones, S., Genes, N., Major, V. J., Jaffe, I. S., Cardillo, A. B., Heilenbach, N., Ali, N. F., Bonanni, L. J., Clayburn, A. J. et al. (2023). Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores. *medRxiv*, (pp. 1–27). doi:10.1101/2023.07.10.23292373.

Labaka, A., Goñi-Balentziaga, O., Lebeña, A., & Pérez-Tejada, J. (2018). Biological Sex Differences in Depression: A Systematic Review. *Biological Research For Nursing*, *20*, 383–392. doi:10.1177/1099800418776082.

Liao, Z., Wang, J., Shi, Z., Lu, L., & Tabata, H. (2023). Revolutionary Potential of ChatGPT in Constructing Intelligent Clinical Decision Support Systems. *Annals of Biomedical Engineering*, (pp. 1–5). doi:10.1007/s10439-023-03288-w.

Lindsay, M. E., de Oliveira, S., Sciacca, K., Lindvall, C., & Ananth, P. J. (2024). Harnessing Natural language processing to assess quality of end-of-life care for children with cancer. *JCO Clinical Cancer Informatics*, *8*, e2400134. doi:10.1200/CCI.24.00134.

Liu, H., Dai, A., Zhou, Z., Xu, X., Gao, K., Li, Q., Xu, S., Feng, Y., Chen, C., Ge, C., Lu, Y., Zou, J., & Wang, S. (2023). An optimization for postpartum depression risk assessment and preventive intervention strategy based machine learning approaches. *Journal of Affective Disorders*, *328*, 163–174. doi:10.1016/j.jad.2023.02.028.

Mazumder, P., & Baruah, S. (2021). A Community Based Study for Early Detection of Postpartum Depression using Improved Data Mining Techniques. In *Proceedings of the IEEE International Conference on Computation System and Information Technology for Sustainable Solutions* (pp. 1–7). IEEE. doi:10.1109/CSITSS54238.2021.9682941.

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H. M., Read, J., Abdessalem, T., & Bifet, A. (2021). River: Machine learning for streaming data in python. *Journal of Machine Learning Research*, *22*, 1–8.

Moreira, M. W., Rodrigues, J. J., Kumar, N., Saleem, K., & Illin, I. V. (2019). Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Information Fusion*, *47*, 23–31. doi:10.1016/j.inffus.2018.07.001.

Mrabet, Z. E., Selvaraj, D. F., & Ranganathan, P. (2019). Adaptive Hoeffding Tree with Transfer Learning for Streaming

Synchrophasor Data Sets. In *Proceedings of the IEEE International Conference on Big Data* (pp. 5697–5704). IEEE. doi:10.1109/BigData47090.2019.9005720.

Nakamura, Y., Ueno, T., Takahashi, N., Ichikawa, D., Yamauchi, A., & Ozaki, N. (2024). Early identification of postpartum depression using machine learning. *Psychiatry and Clinical Neurosciences*, *78*, 372. doi:10.1111/pcn.13659.

Nurbaeti, I., Syafii, M., & Lestari, K. B. (2021). Developing an android-based application for early detection of postpartum depression symptoms in Indonesia. *Belitung Nursing Journal*, *7*, 118–124. doi:10.33546/bnj.1308.

Prabhashwaree, T. H. K. R., & Wagarachchi, N. M. (2022). Towards Machine Learning Approaches for Predicting Risk Level of Postpartum Depression. In *Proceedings of the International Conference on Artificial Intelligence* (pp. 1–6). IEEE. doi:10.1109/SLAAI-ICAI56923.2022.10002477.

Qi, W., Wang, Y., Wang, Y., Huang, S., Li, C., Jin, H., Zuo, J., Cui, X., Wei, Z., Guo, Q. et al. (2025). Prediction of postpartum depression in women: development and validation of multiple machine learning models. *Journal of Translational Medicine*, *23*, 291. doi:10.1186/s12967-025-06289-6.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*, 6086. doi:10.1038/s41598-024-56706-x.

Raisa, J. F., Kaiser, M. S., & Mahmud, M. (2022). A Machine Learning Approach for Early Detection of Postpartum Depression in Bangladesh. In *Proceedings of the International Conference on Brain Informatics* (pp. 241–252). Springer volume 13406 LNAI. doi:10.1007/978-3-031-15037-1_20.

Rogers, A., Obst, S., Teague, S. J., Rossen, L., Spry, E. A., Macdonald, J. A., Sunderland, M., Olsson, C. A., Youssef, G., & Hutchinson, D. (2020). Association Between Maternal Perinatal Depression and Anxiety and Child and Adolescent Development. *JAMA Pediatrics*, *174*, 1082–1092. doi:10.1001/jamapediatrics.2020.2910.

Sezgin, E., Chekeni, F., Lee, J., & Keim, S. (2023). Clinical Accuracy of Large Language Models and Google Search Responses to Postpartum Depression Questions: Cross-Sectional Study. *Journal of Medical Internet Research*, *25*, 1–4. doi:10.2196/49240.

Shatte, A. B., Hutchinson, D. M., Fuller-Tyszkiewicz, M., & Teague, S. J. (2020). Social Media Markers to Identify Fathers at Risk of Postpartum Depression: A Machine Learning Approach. *Cyberpsychology, Behavior, and Social Networking*, *23*, 611–618. doi:10.1089/cyber.2019.0746.

Shin, D., Lee, K. J., Adeluwa, T., & Hur, J. (2020). Machine Learning-Based Predictive Modeling of Postpartum Depression. *Journal of Clinical Medicine*, *9*, 1–14. doi:10.3390/jcm9092899.

Silva, W. A., & Villela, S. M. (2021). Improving the one-against-all binary approach for multiclass classification using balancing techniques. *Applied Intelligence*, *51*, 396–415. doi:10.1007/s10489-020-01805-1.

Sim, J.-a., Huang, X., Horan, M. R., Stewart, C. M., Robison, L. L., Hudson, M. M., Baker, J. N., & Huang, I.-C. (2023). Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. *Artificial intelligence in medicine*, (pp. 102701–102711). doi:10.1016/j.artmed.2023.102701.

Slomian, J., Honvo, G., Emonts, P., Reginster, J.-Y., & Bruyère, O. (2019). Consequences of maternal postpartum depression: A systematic review of maternal and infant outcomes. *Women's Health*, *15*, 1–55. doi:10.1177/1745506519844044.

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2019). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, *26*, 1064–1074. doi:10.1109/TVCG.2019.2934629.

Stewart, D. E., & Vigod, S. N. (2019). Postpartum Depression: Pathophysiology, Treatment, and Emerging Therapeutics.

Annual Review of Medicine, *70*, 183–196. doi:10.1146/annurev-med-041217-011106.

Su, X., Zhang, Y., Chen, M., Wang, H., & Liu, G. (2023). Influencing factors and risk prediction modeling of maternal postpartum depression: a cross-sectional study in Chinese puerperal women of sitting the month. *Frontiers in Psychiatry*, *14*, 1–9. doi:10.3389/fpsyt.2023.1252789.

T., J. M. C., Starmans, M. P. A., Arif, M., Niessen, W. J., Klein, S., Bangma, C. H., Schoots, I. G., & Veenland, J. F. (2021). A Multi-Center, Multi-Vendor Study to Evaluate the Generalizability of a Radiomics Model for Classifying Prostate cancer: High Grade vs. Low Grade. *Diagnostics*, *11*, 369. doi:10.3390/diagnostics11020369.

Tieppo, E., Barddal, J. P., & Nievola, J. C. (2021). *Classifying Potentially Unbounded Hierarchical Data Streams with Incremental Gaussian Naive Bayes* volume 13073 LNAI. Springer. doi:10.1007/978-3-030-91702-9_28.

Trifan, A., Semeraro, D., Drake, J., Bukowski, R., & Oliveira, J. L. (2020). Social Media Mining for Postpartum Depression Prediction. *Studies in Health Technology and Informatics*, *270*, 1391–1392. doi:10.3233/SHTI200457.

Wang, S., Pathak, J., & Zhang, Y. (2019). Using electronic health records and machine learning to predict postpartum depression. *Studies in Health Technology and Informatics*, *264*, 888–892. doi:10.3233/SHTI190351.

Yang, S.-T., Yang, S.-Q., Duan, K.-M., Tang, Y.-Z., Ping, A.-Q., Bai, Z.-H., Gao, K., Shen, Y., Chen, M.-H., Yu, R.-L., & Wang, S.-Y. (2022). The development and application of a prediction model for postpartum depression: optimizing risk assessment and prevention in the clinic. *Journal of Affective Disorders*, *296*, 434–442. doi:10.1016/j.jad.2021.09.099.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*, 719–731. doi:10.1038/s41551-018-0305-z.

Zhang, R., Liu, Y., Zhang, Z., Luo, R., & Lv, B. (2025). Interpretable Machine Learning Model for Predicting Postpartum Depression: Retrospective Study. *JMIR Medical Informatics*, *13*, e58649. doi:10.2196/58649.

Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021). Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *Journal of Affective Disorders*, *279*, 1–8. doi:10.1016/j.jad.2020.09.113.