医疗病理预测：系统综述与拟议方法

Chaimae Taoussi*[a]，Imad Hafidi †[b]，以及 Abdelmoutalib Metrane ‡[c]

[a] 过程工程、计算机科学与数学实验室，苏丹穆莱·斯利曼大学，贝尼迈拉勒，摩洛哥

摘要

医疗保健领域是每个社区的重要支柱。为优化医疗流程、提升护理质量并促进患者管理，已有大量相关研究。本文分析并考察了多项研究，探讨如何利用大数据、人工智能、机器学习和深度学习等最新技术改善医疗保健，并据此提出一种总体方法，重点关注医疗数据的收集、预处理和聚类，以便在分析后更便捷地供患者和医疗专业人员使用，从而在较短时间内更准确地预测最常见的病理。

关键词：医疗保健、大数据、人工智能、自动语言处理、数据挖掘、预测模型。

1 引言

大数据与处理来自多种来源的数据所需的海量计算资源相关，这些数据包括结构化、半结构化或非结构化信息。海量数据处理系统需要高度的流程自动化，而随着计算机技术的最新发展，人工智能（AI）应运而生，并被应用于诊断决策，从而为此提供了助力 [1]。

在医疗保健领域，对包括医疗记录、患者记录和医学检查结果在内的多种大数据来源加以利用，使先进方法不仅可供医疗专业人员使用，也可供其患者使用 [2]。此外，研究表明，大规模电子健康记录（EHRs）数据能够彻底改变精准医学中的科学发现方式。因此，EHRs 包含极具价值的数据，应得到更充分的利用，以指导各个学科的生物医学研究 [3]。

近年来，临床数据一直在采用数据科学技术进行处理，使医疗专业人员能够创建多种分析模型。设计并部署能够表现得仿佛理解生物医学的计算机系统，是基于知识的生物医学数据科学的重要组成部分 [4]。过去几十年来，计算技术已应用于生物科学领域的疾病检测和诊断。由于基于人工智能技术的应用近期呈指数级增长，同时也因为

医生需要减少执业中的错误、意外和误诊，医疗行业已接纳并采用人工智能 [5]。

在当今快节奏的环境中，医疗专业人员在使用技术设备或提供技术支持服务时，需要具备数字化技能，因为医疗专业人员能力不足可能危及患者安全并增加出错的可能性 [6]。

疾病诊断是数据挖掘技术的应用领域之一，这些技术有助于从医学数据库中提取知识，并为预测各种疾病提供精准度较高的临床决策支持。由于能够发现隐藏在医疗数据中的模式和关系，这些技术在设计临床支持系统方面非常有效 [7]。

2 方法:

本系统综述的方法遵循 PRISMA 清单的原始格式。图1展示了该阶段的流程图。

2.1. 研究识别:

我们进行了详细的文献检索，重点关注 2018年 至 2022年 期间发表的文章，以便基于近期研究开展分析。我们制定了一份综述方案以尽量减少研究者偏倚，该方案包括研究问题的措辞、明确的检索策略以及纳入与排除标准。

2.2. 研究问题（RQs）:

为明确研究范围并使研究更具针对性，本文提出了研究问题（RQs），聚焦于五个主要方面：医学信息学中的数据收集、医学信息学中的数据预处理、医学信息学中的医疗数据映射、医学信息学中的分类与聚类，以及利用人工智能预测心理病理。然而，表1列出了本研究所考察的研究问题。

表 1 具体研究问题

RQ1	医学信息学中应用了哪些数据收集方法?
RQ2	医学信息学中应用了哪些数据预处理方法?
RQ3	医学信息学中应用了哪些数据映射技术?
RQ4	医学信息学中应用了哪些分类与聚类方法?
RQ5	如何利用人工智能预测医学病理?

2.3. 检索策略：

检索策略围绕检索问题展开，包括四个步骤：选择数字图书馆、选择其他检索来源、确定时间范围以及定义检索关键词。本研究选取了三个最受欢迎且覆盖面最广的在线数字图书馆：Scopus、PubMed 和 Google Scholar。选择这些数据库是因为其出版宗旨与我们的研究问题相关。检索范围限定为2018年至2022年，因为我们更关注近期文献，以了解最新的知识与研究贡献。

为查找所有相关文献，我们采用了全面的检索策略。本次检索所使用的检索词如下：

- “数据收集” AND “健康信息学”- “数据预处理” AND “健康信息学”- “医学数据映射” AND “健康信息学”- “分类与聚类” AND “健康信息学”- “人工智能” AND “医学病理预测”

2.4. 纳入、排除标准与数据提取：

经过前述流程，我们随后获得了一批文献。因此，我们制定了纳入与排除标准，以筛除无关论文：

- 若论文描述了健康信息学中的数据收集或预处理方法，或介绍了健康信息学中的数据映射或聚类技术，则纳入该研究。最后，若论文证明了人工智能在医学病理预测中的应用，则符合纳入条件。

- 若研究问题无关、论文属于文献综述、无法获取全文，或论文未关注回答与医学信息学中健康数据的收集、预处理、映射和分类，以及使用人工智能模型预测医学病理相关的研究问题，则排除该研究。

2.5. 结果：

我们的检索共发现734篇发表于2018年至2022年间的论文。在分析研究问题的重要性时，根据标题或摘要排除了614篇论文（图1），剩下169篇符合条件。然而，根据排除标准，又排除了120篇论文，其中33篇包含无关的研究问题，28篇未关注在健康信息学中展示医疗数据映射，18篇未关注在健康信息学中展示数据收集，15篇为重复论文，10篇为提出新方法的方法学论文，7篇未关注在健康信息学中展示数据预处理，6篇未关注在健康信息学中展示聚类，3篇为简短通讯，另有3篇为更正声明、验证

研究。因此，最终选取了49篇文章用于本次系统综述，如图1所示。在下文各节中，我们将更深入地讨论数据分析的各项研究问题。

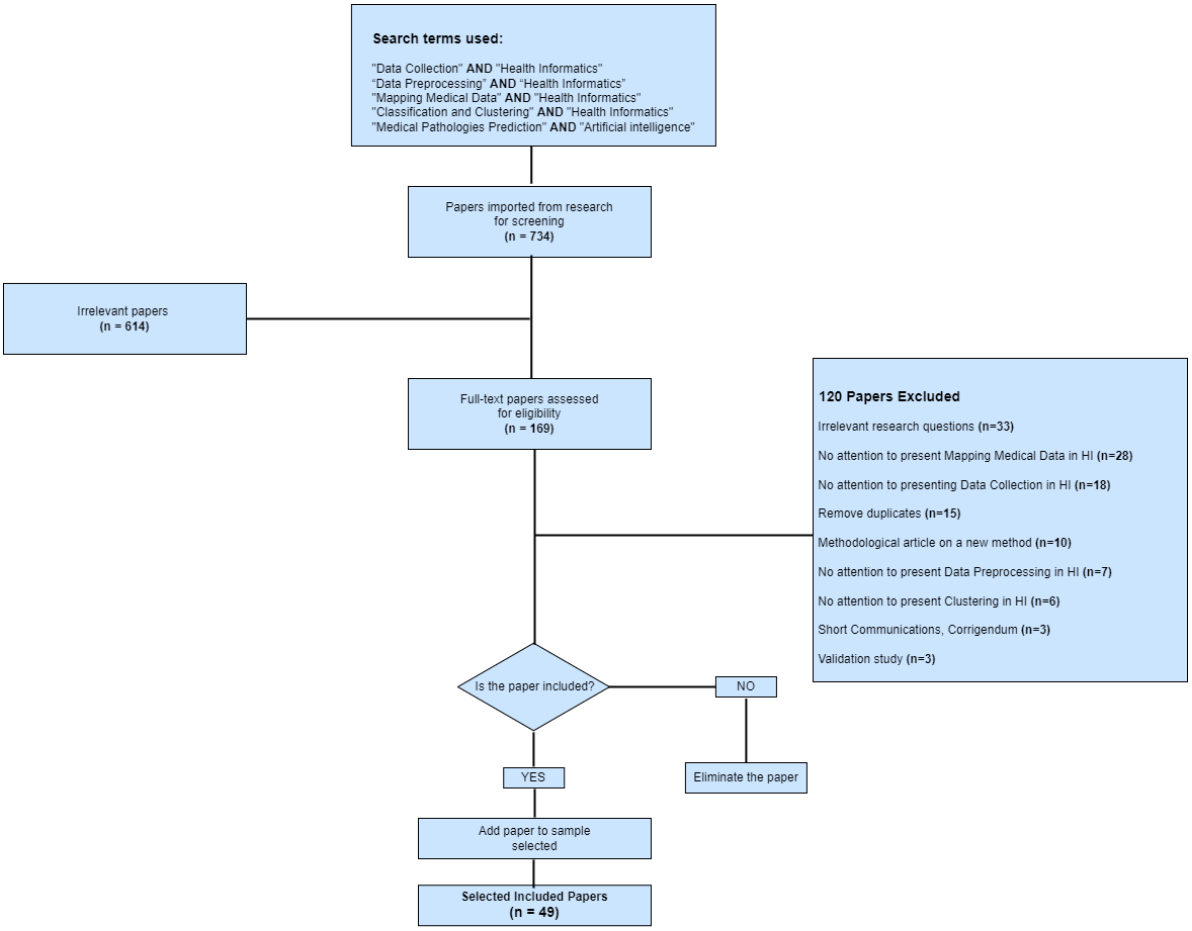


图1：PRISMA流程图

2.6. 文献综述：

2.6.1. RQ1：医学信息学中采用了哪些数据收集方法?

数据收集是至关重要的第一步，因为它能显著改善卫生部门的预防、诊断、治疗和医疗随访。Seol等人 [8] 开发了一种自然语言处理（NLP）解决方案，通过筛选分为两种文件类型的电子健康记录，检测用于流行病学目的的电子病历中的医疗事件：一种是文本文件的可扩展标记语言（XML）文件，另一种是包含病历元数据的XML文件。该方案通过应用全局数据模型来整合来自各种数据源的数据。

大数据在医疗保健领域的潜力在于能够识别模式，并将海量数据集转化为洞见，供决策者使用并用于精准治疗。根据 Hariri等人 [9], 的说法，从收集单项数据开始，

逐步综合来自不同来源的异构数据，将揭示全新的降低复杂性的新技术，从而提升疾病预测和预防能力。

2.6.2. RQ2：医学信息学中应用了哪些数据预处理方法?

准备数据并降低复杂度需要一个必要步骤，称为数据预处理。在许多科学和医学领域，我们都可以探讨多种干预措施。自然语言工具包（NLTK）是入门自然语言处理领域最有效的方式。NLTK 提供的一组应用程序包 [10] 深受自然语言处理、计算语言学和人工智能领域的研究人员及学生青睐。将第一种干预措施作为切入点，Jha 等人 [11] 展示了一种使用 NLTK 和 WordNet 将文本转换为表情符号以及反向转换的简便方法。他们认为，文本情感分析是自然语言处理领域的重要研究方向。在类似的背景下，Yao 等人 [12] 使用 NLTK 库开发了一种算法，可大致确定不同句子的情感分值，并将其与句子作者给出的分值进行比较。

由于全球范围内可获取的医疗信息数量不断增加，以及理解和挖掘医疗行业大数据的重要性日益凸显，自然语言处理的重要性也越来越高 [13]。ASSALE 等人 [14] 对将自然语言处理技术应用于自由文本电子患者记录的最新相关研究成果进行了综述。

近年来，人工智能（AI）在电子健康领域的应用受到广泛关注。利用智能界面提高患者参与度和治疗依从性，以及通过预测建模管理流程，都是人工智能分析非结构化数据的应用示例。Carchiolo 等人 [15] 提出了一个系统，该系统首先从数字化医疗处方中提取文本，然后应用自然语言处理和机器学习技术。

数据挖掘对于识别医疗保健领域的新兴趋势至关重要。从海量数据集中发现重要洞见是最具启发性的研究领域之一。电子健康记录数据由高维多变量信息构成，并使用数据挖掘方法创建统计模型。这类数据分析难度较高，也容易出错 [16]。通过将数据挖掘纳入数据框架，医疗保健机构可以更轻松地做出决策，同时获得重要的新医学洞见。医学领域预测性数据挖掘的目标是创建高效的预测模型，提供准确的预测，并帮助医生优化诊断和治疗规划流程 [17]。

文本工程通用架构（GATE）是一种用于语言工程的架构、框架和开发环境。近年来，它在医疗数据处理领域也日益受到欢迎，尤其适用于从英文文档中进行信息抽取。在一定程度上，它也支持其他语言，这主要得益于众多用户构建并在平台上共享的词典 [18]。GATE 是广为人知的文本分析开源工具之一，可用于多种格式和规模的

在 GATE 范式中，Amjad 等人[19] 提出了一款名为多语言情感分析统一方法（UAMSA）的全新插件，只需使用一个工具即可执行多种多语言情感分析任务。此外，UAMSA 还能帮助用户准确地对人们以多种语言发表的评论进行情感分类。

除了 GATE，PEZOULAS 等人 [20] 还开发了一种自动化医疗数据整理与清理系统，通过检测和监测不一致、缺失值、异常和相似性，并借助数据标准实现数据协调。因此，GOLDBERG 等人 [21] 指出，可以利用自然语言处理和机器学习来预测心理治疗过程中的一个重要组成部分。

2.6.3. RQ3：医学信息学中应用了哪些数据映射技术?

电子健康记录（EHRs）是一种特殊的健康信息技术，正越来越多地应用于各专科和临床场景。Topaz 等人 [22] 开发了一种名为 NimbleMiner 的新型快速临床文本挖掘系统，采用现代自然语言处理方法实现生物医学术语映射。文本数据量不断增长，因此需要创新的数据处理与分析方法。Al-Hroob 等人 [23]提出了一种新方法，可根据自然语言系统的要求，自动识别描述中的参与者和动作。相关讨论指出，自然语言处理对于提取人员和动作至关重要，也探讨了如何利用人工神经网络进行特定识别。此外，WANG 等人 [24] 介绍了一项近期发表的临床信息抽取（IE）应用研究，该研究能够自动从文本中提取并编码临床信息，以促进电子病历数据的二次利用。

在提供生物医学术语资源方面，最有效的协作举措之一是统一医学语言系统（UMLS）。除了对已发表的 UMLS 文献进行详细的文献计量概况分析，并提出一种系统方法来映射科学的智力发展外，Kim 等人 [25] 还对未来可能的发展方向提出了一些建议。不过，在医疗数据映射框架内，GORRELL 等人 [26] 推出了全新的 Bio-YODIE 系统，该系统由两个主要组件构成：一条用于标注包含 UMLS 概念唯一标识符（CUI）及其他相关 UMLS 数据的文档处理流水线，以及一个负责准备相关资源的组件，以便在运行时处理 UMLS 和其他必要的信息资源。在同一背景下，我们也纳入了 ABBAS 等人 [27] 的研究。他们提出了一种实现 UMLS 术语服务（UTS）的算法，并对其进行了个性化调整，以从语录中使用的所有表达和术语中抽取概念，确定其语义类型和实体类型，从而对这些概念进行准确分类。由此我们得出结论：多种信息抽取方法与 UMLS 相结合，已被用于从多个医学领域的医疗数据源中标注并抽取具有临床意义的信息，例如心理学 [28] 和癌症 [29]。

2.6.4. RQ4：医学信息学中应用了哪些分类与聚类方法？

在医学领域，聚类用于识别具有相似特征的患者亚组或群体，根据疾病诊断史对异质性患者群体进行分类，并识别表型簇及其风险因素。Maurits 等人 [30] 开展了一项研究，旨在提出一个具有普适性的框架，用于利用纵向收集的电子健康记录，从而更便于开展研究，并为同质性更高的疾病人群提供治疗。在另一项研究中，Ricciardi 等 [31] 提出了一种机器学习方法，利用随机森林根据步态特征识别处于帕金森病不同阶段的患者，而不是比较患病与未患病的患者。

Wang 等人使用潜在狄利克雷分配（LDA）这一无监督概率生成模型（属于主题模型），研究了采用无监督机器学习技术从电子健康记录数据中识别患者亚组和潜在疾病簇。他们还提出了泊松狄利克雷模型（PDM），这是一种新颖的无监督机器学习技术。

为了提升文本分类系统的性能，Kadhimet 等提出了一套系统，用于文本预处理、基于阈值参数的特征提取，以及将文本分类到一个或多个类别中；他们证明，词频-逆文档频率结合余弦相似度得分的分类效果更好。在同一背景下，Kashina 等研究了文本准备的各个阶段如何影响分类结果，并发现逻辑回归是最准确且直接的方法。

近年来医疗数据大量生成，因此有必要开发大数据分类方法。Jerlin 等认识到，医疗数据分类可用于可视化隐藏模式，并据此判断医疗数据中是否存在疾病，随后提出了用于疾病分类的有效多核支持向量机（MKSVM）和果蝇优化算法（FFOA）。在同一框架下，Huang 等提出了一种基于社区的联邦学习（CBFL）算法，该算法整合来自不同社区的电子病历数据，并同时为每个社区训练一个模型，使学习过程比联邦机器学习（FL）算法更加高效。此外，Desai 等分别研究了基因组数据集，以及包含临床数据点的医学影像数据，并使用组合式深度学习架构执行分类任务。

临床医生将受益于能够在不侵犯患者隐私的情况下访问用于诊断的预测模型。使用全同态加密（FHE）可以计算加密医疗数据，同时还能保护数据机密性。Wood 等 [38] 开发了一种基于完全同态私钥加密的朴素贝叶斯隐私保护分类密码技术。该协议使数据所有者能够在无法直接访问训练得到的模型的情况下，对其数据进行隐私分类。

2.6.5. R Q5：如何利用人工智能预测医学病理 ?

精准医学是一种新兴的临床研究和患者治疗方法，旨在通过整合个体的多模态或多组学数据来理解和治疗疾病，从而作出适合患者的决策。为了对个人和群体的结果进行分类和预测，包括新型深度学习模型在内的多组学数据分析与处理技术至关重要 [39]。

如今，人们受到多种疾病的影响，因此疾病早期预测成为一项至关重要的任务。遗憾的是，医生仅凭症状无法准确预判疾病，这使得实现精准疾病预测成为最具挑战性的工作。为解决这一问题，数据挖掘在疾病预测中发挥着至关重要的作用。我们指出，Dahiwade等人的研究[40]发现，在疾病预测方面，卷积神经网络算法比K近邻算法更加精准高效。在另一项研究中，Shipe等人[41] 回顾了构建逻辑回归风险预测模型所涉及的步骤，包括从选择数据源和预测变量，到分析模型性能以及改进临床决策的全过程。最后，Nusinovici等人[42] 证明，在预测低发病率主要慢性疾病的风险时，逻辑回归与机器学习模型同样有效，即使临床预测因子较为简单。

医疗领域已开展了应用于多个健康领域的各种实验。RAMZAN等人[43] 发现，利用功能性磁共振成像和先进的深度学习方法对阿尔茨海默病等神经退行性脑疾病进行分类和预测，有利于临床决策，并有望帮助进行阿尔茨海默病预后。

临床记录和统一医学语言系统资源是评估糖尿病重症监护患者死亡率的有效且关键的工具。在一项近期研究中，Ye等人 [44] 创建了多个预测模型，用于解读糖尿病患者的死亡率。该研究旨在利用统一医学语言系统资源（包括机器学习和自然语言处理方法）评估死亡风险。

因此，我们注意到ALAWAD等人 [45] 的研究，他们证实，利用统一医学语言系统词汇资源丰富卷积神经网络模型的词汇输入，在检测癌症病理方面，系统性地优于未集成预训练词的卷积神经网络（CNN）模型。此外，WANG 等人 [46] 使用深度卷积神经网络构建了一个针对肺癌病理图像的自动化肿瘤区域检测系统，使他们得以深入了解肿瘤形态与患者预后之间的关联。

随着数字技术的迅速发展，CHO等人 [47] 通过开发一种基于机器学习的情绪预测算法，为未来研究提供了一个有效模型；该算法甚至可能通过促进临床实际应用，改善情绪障碍患者的预后。此外，ALONSO等人 [48] 通过回顾有关心理健康领域数据挖掘技术和算法的现有文献，发现将数据挖掘技术应用于抑郁症等障碍，对于临床决策、诊断预测和患者生活质量都极为有益。

Th近期关于临床病理预测的大多数研究都集中在

预测糖尿病。LJUBIC 等人[49] 证实，循环神经网络模型最适合电子病历数据，可用于预测糖尿病并发症。该领域最新的研究来自 RASMY 等人[50] 他们评估了逻辑回归、循环神经网络以及统一医学语言系统在预测糖尿病患者心力衰竭风险方面的表现。

3 拟议方法

在分析和审阅了先前发表的、通过利用大数据、人工智能、机器学习和深度学习等最新技术来改善医疗保健并回答我们系统综述中各项研究问题的论文后，我们注意到，要在更短时间内准确预测医学病理，关键在于建立一个定义明确且可靠的流程，涵盖预测之前的所有步骤：首先收集医疗数据，然后对所收集的数据进行预处理，接着根据此前经过预处理的所收集数据对医疗数据进行映射，随后对患者画像进行分类和聚类，从而能够在较短时间内以高精度预测医学病理。为此，我们提出了一种能够在更短时间内精准预测临床病理的通用方法，该方法主要包括五个步骤：

3.1 数据收集：

我们拟议方法的第一步（图2）基于对不同格式的医疗数据（电子病历、医生观察记录.....）进行整合。我们设置了一个“收集”函数，该函数以文件目录路径（DME）作为输入，并返回每个文件的类型及其路径；随后创建一个“数据”类，其中包含每个文件的路径及其格式。

为降低海量医疗数据的复杂性，我们实现了“数据分析器”功能，该功能以“数据”类别为输入，并沿着两个主要维度对所有文件进行分类：

- 文本（RTF、PDF）：包含所有需要进一步分析和处理的非结构化文件。- 关系数据库（.csv、.db）：包含所有只需较少分析和处理的半结构化文件。

我们将设置一组包装器，用于在不同数据源中执行子查询，并将子查询结果转换为 JSON 文档，随后存储在“数据集”中。

在此数据收集步骤结束时，我们将设置一个“解析器”，它以数据集中的数据为输入，并将其作为输出返回：

- 元数据：包含每个文件最重要的信息，以便于查找和归档。

- **新的“电子病历”数据集：**包含处理和分析所有患者医疗数据所需的全部字段，以便进行数据挖掘，并通过仅分析所需数据来降低复杂度、节省时间。

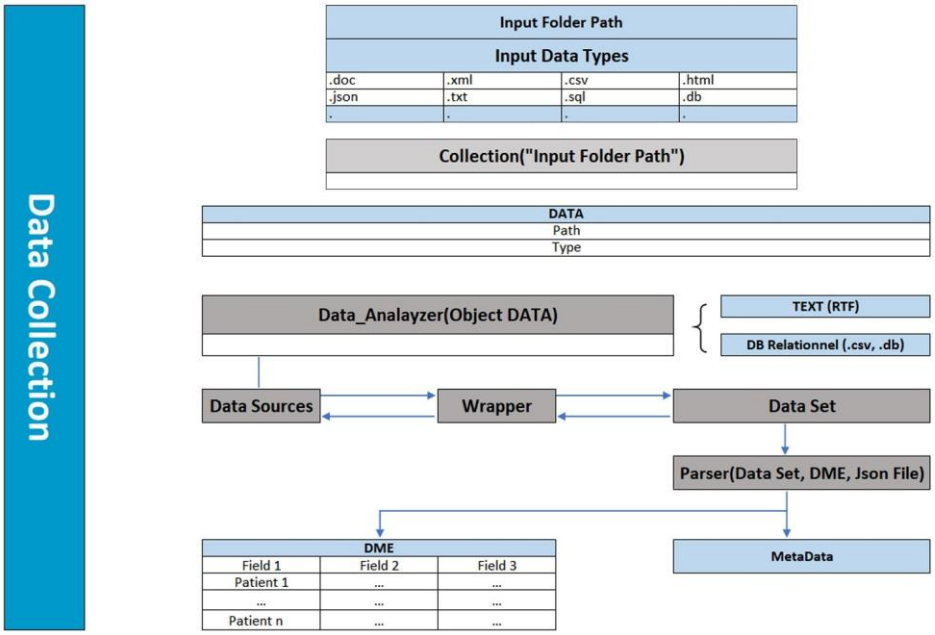


图2：数据收集

3.2 数据预处理：

我们拟议方法的第二步（图3）是对“DME”数据集应用预处理流程。该流程使用以 Python 编写的自然语言工具包（NLTK）[51] 软件库，对医疗数据进行数据清洗、集成、归约和转换。该库依据 GNU通用公共许可证（GPL）分发，包含程序模块、数据集和教程，用于支持计算语言学和自然语言处理的学习与教学；它还支持多项功能，例如透明的语法、有效的字符串管理功能以及简洁性，并使用树库分词器和词性标注器。所使用的 NLTK 库能够自动处理数据，从而更高效地生成结构化且经过预处理的“SDME”数据集。

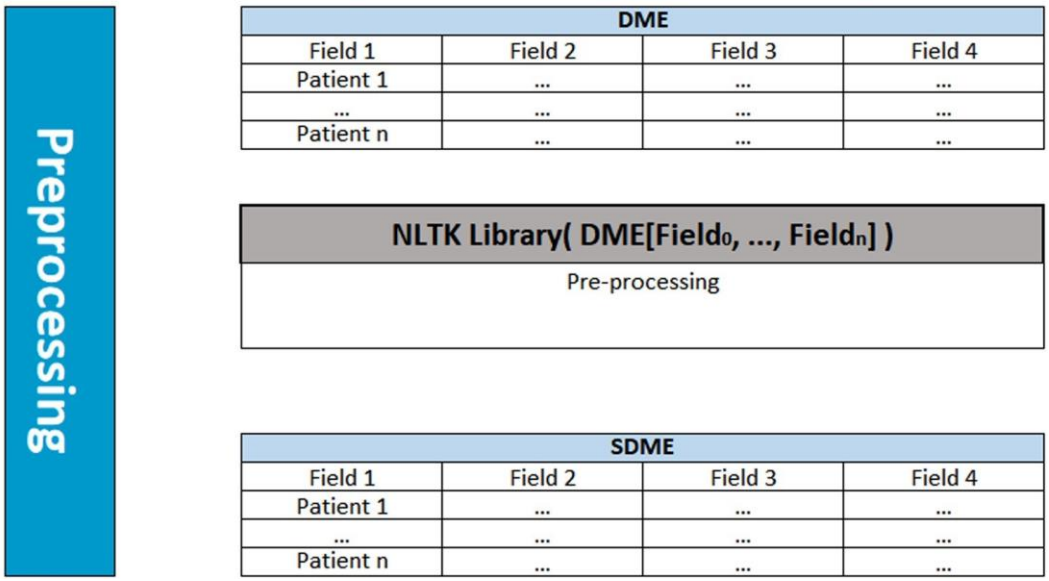


图3：数据预处理

3.3 医学数据映射：

我们拟议方法的第三步（图4）是使用“UMLS” API 对结构化数据集“SDME”中包含的生物医学术语进行映射。该 API 将多种健康与医学词汇表及标准整合为一组文件和软件，以促进与计算机系统的互操作性。此外，我们还使用生物医学实体“GATE (Bio-YODIE)” API。它是一种命名实体识别与消歧系统，能够识别文本中的不同类型生物医学实体，并尝试将其关联到 UMLS 中最合适的概念标签，以发现文本中提及的元词库概念。由此可实现更精细的理解，并输出一个包含以下字段的新数据集“映射后的 SDME”：

- 必选术语：表示一个词语或症状列表，其中每一项都映射到统一医学语言系统元词库。- 概念提取：概念是医学术语的含义，每种含义都包含不同的名称。- 语义类型提取：统一医学语言系统的另一种知识来源是语义网络，其中包含 135 种语义类型和 54 种语义关系，用于对统一医学语言系统元词库中表示的所有概念进行分类和归类。- 实体类型提取：实体类型显示概念的父级关系；与语义类型相比，它以更标准、更直观的方式呈现概念的含义。

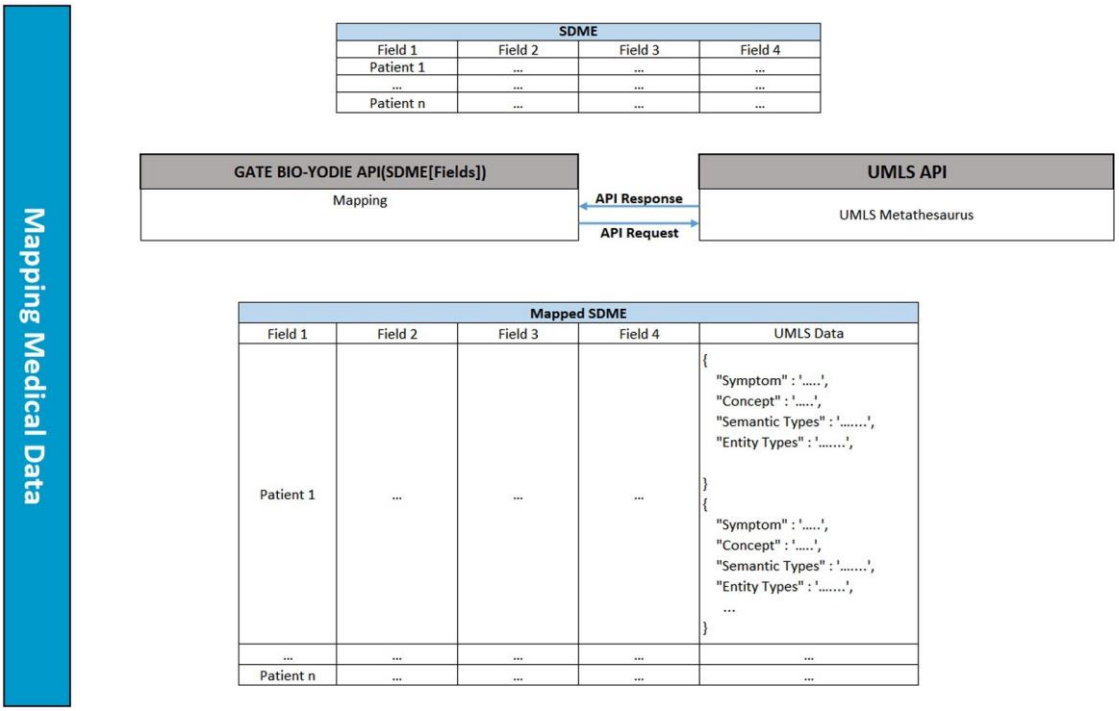


图 4：医学数据映射

3.4 分类与聚类：

拟议方法的第四步（图 5）涉及患者画像的分类与聚类。因此，我们设置了一个“聚类”函数，以“映射后的 SDME”数据集中的数据为输入，并使用数据挖掘中常见的 K-means 聚类算法，根据医疗专业人员预先制定的规则，定义不同的患者画像/患者簇。该算法通常用于将大量数据划分为 k 个不同的簇，并根据对象是否符合特定类别的要求对其进行分类[52]。在我们的方法中，K-means 算法根据“UMLS 数据”字段中的数据，将每位患者分配到一个或多个簇中，输出一个新的“Clustered Mapped SDME”数据集，其中包含“Cluster”字段，该字段由每位患者所属的不同簇组成。

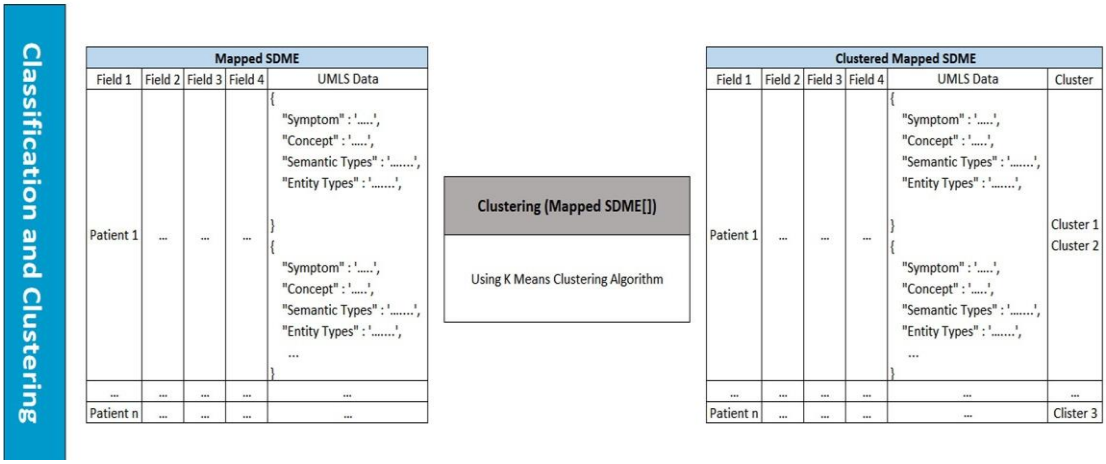


图 5：分类与聚类

3.5 病理预测：

拟议方法的最后一步（图6）涉及使用著名的循环神经网络（RNN）算法开发人工智能模型，以利用隐藏单元分析数据流，重点对序列数据进行建模和处理；该模型包含循环结构和记忆机制，因此能够记住以往的计算结果，这与其他前馈神经网络不同[53], 拟议的人工智能模型以从“Clustered Mapped SDME”数据集获取的数据作为输入，并生成最终的“病理预测”数据集，其中包含以下三个新字段：

- **“病理”字段：**可用于预测最常见且最严重的病理。- **“最佳预测”字段：**可更准确地预测病理。- **“最佳精度”字段：**可在重要时间范围内预测病理。

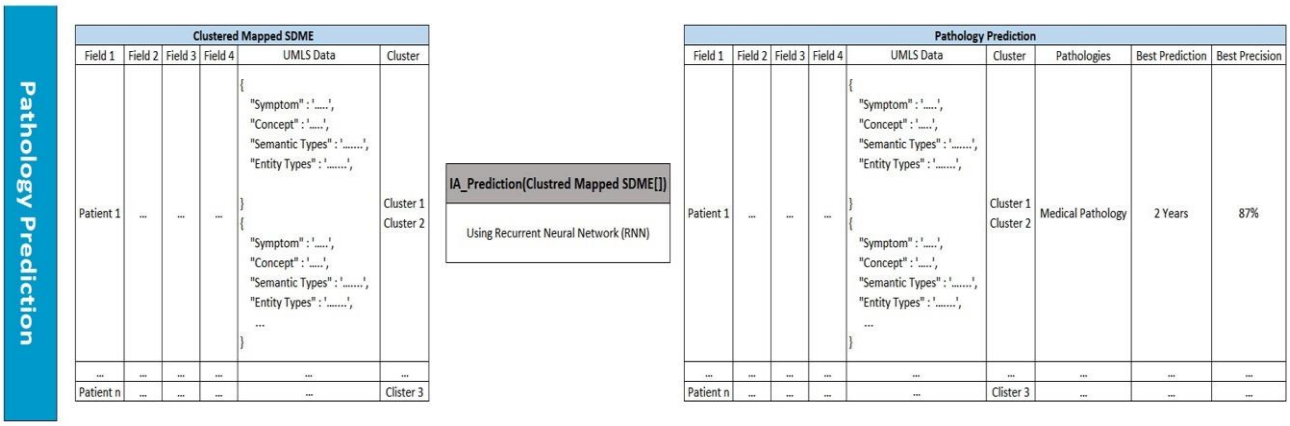


图6：病理预测

4 讨论与局限性

本文的主要目标是考察最具代表性的研究案例，展示运用大数据、人工智能、机器学习和深度学习等最新技术，通过预测医学病理来改善医疗保健的价值。根据我们的调查，许多研究人员目前都在积极利用数字革命，并将源自计算机科学的方法应用于医疗行业，一方面提升医疗护理水平，另一方面通过准确、快速地预测最常见的医学疾病，提出创新的医疗治疗方案，从而减轻患者和医务人员的负担。为此，本文提出了一种可用于预测各种医学病理的通用方法，内容不止于所呈现的系统综述。我们所提出方法的主要优势在于，它给出了一个完整且详尽的流程，涵盖从数据收集到在显著延迟内预测医学病理的所有阶段。

根据上述讨论，可以得出结论：值得注意的是，由于医学信息学领域的文献发展迅速，我们的系统综述可能未能纳入若干已发表的研究。本研究分析了来自三个数据库的研究，包括 Scopus、PubMed 和 Google Scholar。未来的研究可以使用 Medline 和 Web of Science 等不同数据库中的论文，对研究结果进行比较。

5 结论

电子健康正在呈指数级发展，数字技术是创新的核心，既推动临床研究，也促进住院患者护理与支持。大数据在医疗领域的应用带来了众多新可能，并改变了我们收集数据的方式，尤其是因为健康数据规模过于庞大且结构过于复杂，无法通过传统的管理和治疗预约方法加以利用。

我们的文章提出了新的证据，表明要在更短时间内准确预测医学病理，关键在于建立一个定义明确且可靠的流程，涵盖预测之前的所有步骤，并采用大数据、人工智能、机器学习和深度学习等最新推荐技术。49篇入选文章为我们提供了充分论据，用以阐述回答五个研究问题的观点。这些结果将有助于研究人员加深对健康信息学中数据收集或预处理方法、健康信息学中的数据映射或聚类技术，以及如何利用人工智能预测健康信息学中的医学病理的理解。

基于本文开展的系统综述，我们成功提出了一种适用于医学各领域的通用方法。该方法首先从不同的医院和诊所收集非结构化和半结构化数据；随后，使用 Python 自然语言工具包（NLTK）库对收集到的医疗数据应用数据挖掘技术进行预处理；之后，参照统一医学语言系统（UMLS）对医疗数据进行映射，以便于数据分析并降低复杂性和所耗费的时间；接着，我们应用聚类对不同病理进行分类，从而便于建立患者画像；最后，开发人工智能，以更高的精度在显著缩短的时间内预测最常见的病理。

声明

- 资金支持：本稿件的撰写未获得任何资金支持。- 利益冲突：作者声明不存在任何利益冲突。- 伦理审批：本文不包含任何作者开展的涉及人类参与者或动物的研究。

参考文献

[1] Benke, K., Benke, G.等（2018年）。公共卫生中的人工智能与大数据。《International Journal of Environmental Research and Public Health》，15(12)，2796。

[2] Dash, S., Shakyawar, S.K., Sharma, M. 等。医疗保健领域的大数据：管理、分析与未来展望。《大数据杂志》6，54（2019年）。https://doi.org/10.1186/s40537-019-0217-0。[3] Glicksberg, B. S., Oskotsky, B., Thangaraj, P. M., Giangreco, N., Badgeley, M. A., Johnson, K. W., ... & Butte, A. J.（2019年）。PatientExploreR：一种可扩展的应用程序，用于动态可视化基于OMOP通用数据模型的电子健康记录中的患者临床病史。《生物信息学》35(21)，4515-4518。[4] Callahan, T. J., Tripodi, I. J., Pielke-Lombardo, H., & Hunter, L. E.（2020年）。基于知识的生物医学数据科学。《生物医学数据科学年度评论》3，23。[5] Tayal, A., Gupta, J., Solanki, A., Bisht, K., Nayyar, A., & Masud, M.（2022年）。采用图像处理技术诊断视网膜疾病的深度学习卷积神经网络方法。《多媒体系统》28(4)，1417-1438。[6] Konttila, J., Siira, H., Kyngäs, H., Lahtinen, M., Elo, S., Kääriäinen, M., ... & Mikkonen, K.（2019年）。医疗专业人员的数字化能力：系统综述。《临床护理杂志》28(5-6)，745-761。[7] Mintz, Yoav, and Ronit Brodie.《医学人工智能导论》。《Minimally Invasive Therapy Allied Technologies》28.2（2019年）：73-81。[8] Seol, K., Kim, Y. G., Lee, E., Seo, Y. D., & Baik, D. K.（2018年）。基于XML的电子健康记录系统的隐私保护属性基访问控制模型。《IEEE Access》6，9114-9128。[9] Hariri, R. H., Fredericks, E. M., & Bowers, K. M.（2019年）。大数据分析中的不确定性：综述、机遇与挑战。《大数据杂志》6(1)，1-16。[10] Yogish, D., Manjunath, T. N., & Hegadi, R. S.（2018年12月）。使用NLTK的自然语言处理趋势与技术综述。载于《图像处理与模式识别最新趋势国际会议》（第589-606页）。Springer，新加坡。[11] Jha, N. K.（2018年9月）。一种使用NLTK和WordNet实现文本与表情符号相互转换的方法。载于《2018年第二届数据科学与商业分析国际会议（ICDSBA）》（第161-166页）。IEEE。[12] Yao, J.（2019年4月）。使用NLTK对文本数据进行自动情感分析。载于《物理学杂志：会议系列》（第1187卷，第5期，第052020页）。IOP出版社。[13] Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., ... Lei, J.（2020年）。系统评估过去20年医学自然语言处理研究进展：基于PubMed的文献计量学研究。《医学互联网研究杂志》22(1)，e1681。[14] Assale, M., Dui, L. G. 等（2019年）。备注字段的复兴：利用电子健康记录中的非结构化内容。《医学前沿》6，66。[15] Carchiolo, V., Longheu, A., Reitano, G., & Zagarella, L.（2019年9月）。医疗处方分类：一种基于自然语言处理的方法。载于《2019年联合会议》

计算机科学与信息系统（FedCSIS）（第 605-609 页）。IEEE。

[16] Pooja, H., & MP, P. J.（2019年3月）。《电子健康记录》中可获取的大型健康数据的数据挖掘技术综合研究。载于2019年智能控制、环境、计算与通信工程先进技术第一届国际会议论文集（ICATIECE）（第51-55页）。IEEE[17] Reddy, R. P., Mandakini, C., & Radhika, C.（2020年）。医疗领域数据挖掘技术及挑战综述。《国际工程研究与技术杂志》（IJERT），9(8)，329-333。[18] Ganino, G., Lembo, D., Mecella, M., & Scafoglieri, F.（2018年）。面向开源情报的本体填充：一种基于GATE的解决方案。《软件：实践与经验》，48(12)，2302-2330。[19] Amjad, A., & Qamar, U.（2019年9月）。UAMSA：使用GATE进行多语言情感分析的统一方法。载于第六届计算机系统工程会议论文集（第1-5页）。[20] Pezoulas, V. C., Kourou, K. D. 等（2019年）。医疗数据质量评估：医疗数据整理自动化框架的开发。《生物与医学中的计算机》，107，270-283。[21] Goldberg, S. B., Flemotomos, N. 等（2020年）。心理治疗研究中的机器学习与自然语言处理：以联盟为例。《咨询心理学杂志》，67(4)，438。[22] Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H.（2019年）。从临床记录中挖掘跌倒相关信息：基于规则的方法与新型词嵌入机器学习方法的比较。《生物医学信息学杂志》，90，103103。[23] Al-Hroob, A., Imam, A. T., & Al-Heisa, R.（2018年）。利用人工神经网络从需求文档中提取动作与参与者。《信息与软件技术》，101，1-15。[24] Wang, Y., Wang, L. 等（2018年）。临床信息抽取应用：文献综述。《生物医学信息学杂志》，77，34-49。[25] Kim, M. C., Nam, S., Wang, F., & Zhu, Y.（2020年）。绘制UMLS研究的科学领域：一项科学计量学综述。《美国医学信息学会杂志》，27(10)，1612-1624。[26] GORRELL, Genevieve, SONG, Xingyi 等。Bio-YODIE：一种面向生物医学文本的命名实体链接系统。arXiv预印本 arXiv:1811.04860，2018年。[27] Abbas, A., Hussain, J. 等（2022年1月）。从临床出院摘要中显式和隐式识别章节。载于2022年第16届泛在信息管理与通信国际会议（IMCOM）论文集（第1-8页）。IEEE。

[28] Taoussi, C., Hafidi, I., Metrane, A., Lasbahani, A.（2021年4月）。根据电子病历预测心理病理。载于人机交互与新兴技术国际会议论文集（第493-500页）。Springer, Cham。[29] Alawad, M., Hasan, S. S., Christian, J. B., & Tourassi, G.（2018年12月）。使用UMLS元词库对词嵌入进行改造，以用于临床信息抽取。载于2018年IEEE大数据国际会议（Big Data）论文集（第2838-2846页）。IEEE。[30] Maurits, M. P., Korsunsky, I., Raychaudhuri, S., Murphy, S. N., Smoller, J. W., Weiss, S. T., ... & Knevel, R.（2022年）。利用纵向收集的多中心电子健康记录，依据疾病史对异质性患者群体进行分层的框架。《美国医学信息学会杂志》，29(5)，761-769。[31] Ricciardi, C., Amboni, M., Santis, C. D., Ricciardelli, G., Improta, G., Iuppariello, L., ... & Cesarelli, M.（2019年9月）。通过随机森林对帕金森病不同阶段进行分类。载于地中海医学与生物工程及计算会议论文集（第1155-1162页）。Springer，Cham。[32] Wang, Y., Zhao, Y., Therneau, T. M., Atkinson, E. J., Tafti, A. P., Zhang, N., ... & Liu, H.（2020年）。利用电子健康记录，通过无监督机器学习发现潜在疾病簇和患者亚组。《生物医学信息学杂志》，102，103364。[33] Kadhim, A. I.（2018年）。文本分类预处理技术评估。《国际计算机科学与信息安全杂志》（IJCSIS），16(6)，22-32。[34] Kashina, M., Lenivtceva, I. D., & Kopanitsa, G. D.（2020 年）。非结构化医疗数据的预处理：各预处理阶段对分类的影响。Procedia计算机科学，178，284-290。[35] Jerlin Rubini, L., & Perumal, E.（2020年）。使用多核支持向量机和果蝇优化算法对慢性肾病进行高效分类。《国际成像系统与技术杂志》，30(3)，660-673。[36] Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D.（2019年）。患者聚类可提高联邦机器学习的效率，从而利用分布式电子病历预测死亡率和住院时间。《生物医学信息学杂志》，99，103291。[37] Desai, H.（2020年）。采用改进的深度学习架构进行生物医学数据分类。[38] Wood, A., Shpilrain, V., Najarian, K., & Kahrobaei, D.（2019年）。个人生物医学数据的隐私保护朴素贝叶斯分类：在癌症数据分析中的应用。《生物与医学中的计算机》，105，144-150。[39] MacEachern, S. J., & Forkert, N. D.（2021年）。精准医学中的机器学习。《Genome》，64(4)，416-425]。

[40] Dahiwade, D., Patle, G., & Meshram, E.（2019年3月）。采用机器学习方法设计疾病预测模型。载于《2019年第3届计算方法与通信国际会议（ICCMC）论文集》（第1211-1215页）。IEEE。[41] Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L.（2019）。使用逻辑回归开发临床应用预测模型：概述。《胸部疾病杂志》，11（增刊4），S574。[42] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y.（2020）。在预测主要慢性疾病方面，逻辑回归与机器学习表现同样出色。《临床流行病学杂志》，122，56-69。[43] Ramzan, F., Khan, M. U. G. 等（2020）。一种深度学习方法：利用静息态功能磁共振成像和残差神经网络，对阿尔茨海默病各阶段进行自动诊断和多分类。《医学系统杂志》，44（2），1-16。[44] Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y.（2020）。使用机器学习和临床记录预测糖尿病危重患者的死亡率。《BMC医学信息学与决策制定》，20（11），1-7。[45] Alawad, M., Hasan, S. S.等（2018年12月）。利用UMLS元词库改进词嵌入，以进行临床信息抽取。载于《2018年IEEE大数据国际会议（大数据）论文集》（第2838-2846页）。IEEE。[46] Wang, S., Chen, A. 等（2018）。全面分析肺癌病理图像，发现能够预测生存结局的肿瘤形状和边界特征。《科学报告》，8（1），1-9。[47] Cho, C. H., Lee, T., Kim, M. G., In, H. P., Kim, L., Lee, H. J.（2019）。基于昼夜节律相关的被动数字表型，使用机器学习预测情绪障碍患者的情绪：前瞻性观察性队列研究。《医学互联网研究杂志》，21（4），e11029。[48] Alonso, S. G., de la Torre-Dˊıez, I.等（2018）。心理健康领域的数据挖掘算法与技术：系统综述。《医学系统杂志》，42（9），1-15。[49] Ljubic, B., Hai, A. A.等（2020）。使用高级机器学习算法预测糖尿病并发症。《美国医学信息学会杂志》，27（9），1343-1351。[50] Rasmy, L., Tiryaki, F. 等（2020）。用于预测建模的电子健康记录数据表示：UMLS与其他术语体系的比较。《美国医学信息学会杂志》，27（10），1593-1599。[51] Bird, S.（2006年7月）。NLTK：自然语言工具包。载于《COLING/ACL 2006互动展示会议论文集》（第69-72页）。[52] Bharadwaj, H. K., Agarwal, A., Chamola, V., Lakkaniga, N. R., Hassija, V., Guizani, M., & Sikdar, B.（2021）。机器学习在实现基于物联网的医疗保健应用中的作用综述。《IEEE Access》，9，38859-38890。

[53] Pramod, A., Naicker, H. S., & Tyagi, A. K.（2021）。机器学习与深度学习：未来10年的开放问题与研究方向。《医疗护理的计算分析与深度学习：原理、方法与应用》，463-490。

作者信息

简介：

- Chaimae Taoussi 是苏丹穆莱·斯利曼大学的博士生。她最新发表的论文是《基于电子病历预测心理病理》。

- Imad Hafidi 是苏丹穆莱·斯利曼大学的研究教授。他最新发表的论文是《一种基于半同态加密的无线传感器网络新型安全数据聚合方案》。

- Abdelmoutalib Metrane 是苏丹穆莱·斯利曼大学的研究教授。他最新发表的论文是《独立源与相关源含噪混合信号盲分离的稳健方法》。

作者单位：

- Chaimae Taoussi，过程工程、计算机科学与数学实验室，苏丹穆莱·斯利曼大学，贝尼迈拉勒，摩洛哥。- Imad Hafidi，过程工程、计算机科学与数学实验室，苏丹穆莱·斯利曼大学，贝尼迈拉勒，摩洛哥。- Abdelmoutalib Metrane，过程工程、计算机科学与数学实验室，苏丹穆莱·斯利曼大学，贝尼迈拉勒，摩洛哥。

作者贡献：

- 作者 Chaimae Taoussi、Imad Hafidi 和 Abdelmoutalib Metrane 对本研究作出了同等贡献。所有作者均已阅读并批准最终稿件。

通讯作者：

- 通讯请联系 Chaimae Taoussi。