# 隐私保护下单细胞 RNA 测序数据细胞分类研究*

徐文嘉[1] 岑孟杰[2] 陈 亮[3]

([1] 上海交通大学医学院附属仁济医院 上海 200127 [2] 宁波市杭州湾医院 宁波 315336
[3] 上海锘崴信息科技有限公司 上海 200126)

〔摘要〕 **目的/意义** 研究安全的高维稀疏数据处理方法，提高分析精度并保障敏感信息安全，促进单细胞 RNA 测序技术的广泛应用。**方法/过程** 提出基于可信执行环境（trusted execution environment，TEE）的解决方案，将训练数据加密后传输至 TEE，在安全隔离环境中解密并训练，获得训练后的模型参数。对比分析 TEE 和传统明文环境下使用基于神经网络的自动细胞类型识别模型和支持向量机进行细胞分类的表现。**结果/结论** TEE 下两种分类模型的 *F*1 分别达到 0.904 和 0.879，与传统明文环境下性能相当；TEE 提供的安全执行环境对模型的准确性和效率影响极有限，可用于处理敏感或私有数据场景。

〔关键词〕 可信执行环境；单细胞 RNA 测序；单细胞分类；隐私保护
〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.10.016

**Study on Cell Classification of Single-cell RNA Sequencing Data under Privacy Protection**

*XU Wenjia*[1], *CEN Mengjie*[2], *CHEN Liang*[3]

[1] *Renji Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai 200127, China*; [2] *Ningbo Hangzhou Bay Hospital, Ningbo 315336, China*; [3] *Shanghai Nuowei Information Technology Co. Ltd., Shanghai 200126, China*

〔**Abstract**〕 **Purpose/Significance** To develop a secure single-cell RNA sequencing (scRNA-seq) classification method, which can enhance data analysis precision and ensure the security of sensitive information, and to promote the application of scRNA-seq technology in various fields. **Method/Process** The paper proposes a solution based on trusted execution environment (TEE). The training data is encrypted and transmitted to TEE. It is decrypted in a secure and isolated environment, while training the model to obtain the trained model parameters. Automated cell type identification using neural networks (ACTINN) and support vector machine (SVM) are used for cell classification in both TEE and traditional plaintext environments. The results are compared and analyzed. **Result/Conclusion** The results show that the *F*1 score of the two classification models in TEE environment reaches 0.904 and 0.879, respectively, which is comparable to the performance in traditional plaintext environment. The secure execution environment provided by TEE has extremely limited impact on the accuracy and efficiency of the models. This is of great significance for seeking both secure and efficient data processing solutions in scenarios where sensitive or private data needs to be processed.

〔**Keywords**〕 trusted execution environment (TEE); single-cell RNA sequencing (scRNA-seq); single cell classification; privacy protection

〔修回日期〕 2024-07-08
〔作者简介〕 徐文嘉，助理工程师。
〔基金项目〕 浙江省“尖兵”“领雁”研发攻关计划项目（项目编号：2024C01073）。

## 1 引言

单细胞 RNA 测序（single - cell RNA sequencing，scRNA - seq）技术是一种精细化的生物学工具，其提供了观察和分析生物体中单个细胞的能力，揭示了细胞间的微妙差异和复杂的细胞状态，相对于传统测序方法是质的飞跃，极大推进了研究人员对于生命科学的理解，尤其是在细胞分类和生物多样性研究方面。在细胞分类方面，scRNA - seq 技术使科学家能够以前所未有的分辨率识别和分类生物体内的细胞类型。通过分析单个细胞的基因表达模式，研究人员能够发现新的细胞类型，理解细胞如何在不同的生理和病理状态下响应，以及如何贡献于疾病的发展和进程。

单细胞识别任务的实现已从无监督学习模型过渡到有监督学习模型，并采用大规模的单细胞基因表达数据集以提升识别效果。但使用数据集存在数据安全风险，一是包括 scRNA - seq 数据在内的生物信息学领域涉及众多伦理和隐私问题，二是个人信息具有通过遗传数据被识别的可能性，三是共享和转移个人遗传数据可能引发敏感健康信息泄漏。为了实现在保护数据安全的前提下，协同建立细胞分类识别模型，基于隐私计算的联合建模方法被提出，用以自动标注 scRNA - seq 实验中的细胞[1]。在隐私计算相关技术中，基于可信执行环境（trusted execution environment，TEE）[2]的机密计算技术可为生物信息学数据相关应用提供安全的数据处理和分析方案[3-4]。TEE 技术可与传统机器学习方法（如支持向量机（support vector machine，SVM）[5]和梯度提升树[6]等）结合用于细胞分类，也可处理复杂的细胞分类模型。

本文提出基于 TEE 的细胞分类方法，为研究人员提供一个机密计算环境进行 scRNA - seq 分析，确保数据隐私。采用 SVM 作为通用分类器，采用基于神经网络的自动细胞类型识别模型（automated cell type identification using neural networks，ACTINN）作为专门针对 scRNA - seq 数据的分类器。最后通过实验从准确性和计算时间两个维度评估该方法的性能。

## 2 基于 TEE 的细胞分类

本文提出的基于 TEE 的细胞分类方法，见图 1，提供了一个既安全又高效的机密计算环境，用于 scRNA - seq 数据分析，同时确保涉及的数据隐私得到充分保护。

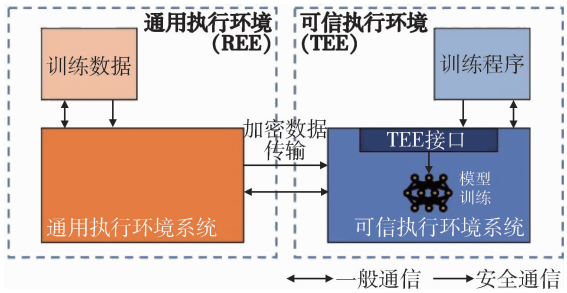


图 1 基于 TEE 的细胞分类方法

### 2.1 基于 TEE 的机密计算技术

2.1.1 概念与特点 TEE 主要通过在处理器或其他计算硬件中实现特定的安全扩展，使敏感数据和代码可以在一个隔离的环境中运行，以提高安全性和隐私保护水平。TEE 提供了一个安全的计算环境，用于保护应用程序的代码和数据免受外部攻击，通过硬件支持实现了计算过程的隔离和保护。TEE 在医疗、云计算、物联网、移动设备、金融和政府管理等领域提供关键安全保护：在医疗领域，保护患者数据和隐私，确保电子病历和远程医疗的安全；在云计算中，保护企业敏感数据；在物联网中，防止攻击；在移动设备中，保障支付和生物识别数据安全；在金融服务中，确保交易和客户数据安全；在政府管理中，提高电子政务和投票的安全性。TEE 广泛用于保护敏感任务和数据。相较于其他隐私保护计算方式（如多方安全计算、同态加密等），TEE 技术可以更加简易地实现隐私保护计算方案，且计算开销远低于基于密码学方案的隐私保护计算。TEE 不仅解决了高度敏感生物医学研究领域的数据安全和隐私保护问题，也推动了科学研究的进步，使研究人员能够在遵守法律和伦理标准的

同时，有效地聚合和分析不同来源的单细胞基因表达数据，有助于更安全、更有效地利用生物医学数据，为疾病的预防、诊断和治疗提供更加精准的科学依据。

2.1.2 主要技术方案 （1）英特尔软件防护扩展。一种针对 x86 架构处理器的安全扩展[7]，允许应用程序创建被称为飞地的安全区域。飞地中运行的代码和数据可以得到额外保护，确保其免受外部攻击，包括来自操作系统或超级用户的攻击。SGX 通过特定指令集支持数据和代码的安全调入、调出，实现与非保护代码的安全交互。（2）AMD 内存加密技术。AMD 提供的保护类似于 SGX 技术，但侧重于通过内存加密保护数据。使用硬件级别的加密功能，确保存储在内存中的数据在物理层面得到保护，防止数据被非授权访问或修改。（3）ARM TrustZone 技术。该技术在 ARM 架构的处理器中提供一种安全的执行环境[8]。通过在硬件级别划分出一个安全的虚拟处理器，TrustZone 支持运行安全敏感的应用程序，同时与常规应用程序环境分隔开。允许设备制造商在移动设备、嵌入式系统等产品上实现安全功能，如安全启动、移动支付和数字版权管理。（4）海光 TEE。基于海光 CPU 的可信计算平台充分整合内置安全处理器与国密算法硬件加速模块，构建高效且安全的可信平台控制模块，不仅能实现可信计算的基本信任根，还提供密码运算的硬件加速，大大提高运算效率并确保技术的国产化和自主可控，为计算机系统提供了强大的自我保护能力，从而有效地对抗恶意软件、提高系统整体安全性。

2.1.3 流程步骤 上述主要技术方案通过硬件提供可信执行环境，创建安全空间，保护代码和数据，比纯软件方案更安全简洁，并能防御部分硬件威胁。一般性的 TEE 下机密计算流程，见前文图 1。

### 2.2 单细胞分类方法

2.2.1 ACTINN 深度学习网络无须特征工程，可以直接在数据中训练高维度特征，对于大量细胞数据的分类任务非常有效。近年来已经有多种深度学习模型在细胞分类任务上取得显著成效。其中 ACTINN[9] 是较新的细胞分类方法，采用全连接神经网络进行细胞类型分类。

2.2.2 SVM SVM 由于理论基础坚实、准确率高，对过拟合具有鲁棒性[10]，成为近年来流行的分类方法，已广泛应用于对 scRNA - seq 数据进行基于基因表达量的细胞分类[11]，可以有效解决这类数据固有的高维度、稀疏性和噪声问题。

## 3 实验结果与分析

### 3.1 实验设计与结果

分别在 TEE 和标准环境（传统明文环境）下，使用 Zhengsorted 数据集[12] 进行细胞类型识别实验。采用 ACTINN 和 SVM 两种模型，从 *F*1 得分、精确度及运行时间 3 个维度分析与比较实验结果，见表 1，其中每项数据为 5 次测试的平均值 ± 标准差。

**表 1 不同环境下单细胞测序模型分类性能**

| 环境 | 模型 | *F*1 | 精确度 | 运行时间（秒） |
|---|---|---|---|---|
| TEE | ACTINN | 0.904 ± 0.002 | 0.904 ± 0.002 | 445.655 ± 8.776 |
| | SVM | 0.879 ± 0.002 | 0.881 ± 0.001 | 133.133 ± 4.177 |
| 标准环境 | ACTINN | 0.901 ± 0.003 | 0.901 ± 0.003 | 459.975 ± 25.342 |
| | SVM | 0.879 ± 0.002 | 0.881 ± 0.001 | 137.695 ± 2.811 |

### 3.2 结果分析

ACTINN 模型在 TEE 环境中的 *F*1 得分为 0.904 ± 0.002，精确度为 0.904 ± 0.002，运行时间为 445.655 ± 8.776 秒，与标准环境相比，*F*1 得分和精确度略微提高，但运行时间略有下降。表明 TEE 环境对 ACTINN 模型性能影响不大，对计算效率的影响较有限。

SVM 模型在 TEE 环境中的 *F*1 得分为 0.879 ± 0.002，精确度为 0.881 ± 0.001，运行时间为 133.133 ± 4.177。与标准环境相比，*F*1 得分和精确度保持稳定，运行时间略有下降。结果同样表明，TEE 环境下虽然存在额外负担，但是 SVM 模型能够维持其精确度，且对处理速度影响有限，TEE 环境对计算效率的影响较有限。

综合可知，TEE 环境和标准环境下基于 ACT-

INN 和 SVM 的 scRNA - seq 分类模型性能和运行时间均基本保持一致，说明 TEE 提供的安全执行环境对于模型的准确性和效率影响微乎其微，能够在不影响性能效率的情况下，进行安全保护的单细胞测序分析，为敏感或私有数据处理场景提供了既安全又高效的解决方案。

然而，此方法的潜在局限性也值得注意。一是在处理更大规模的数据集或更复杂的模型时，TEE 在资源配置（如内存大小、CPU 种类、容器数量和异构加速）方面通常受到限制，可能影响大数据集或计算密集型模型的处理效率。二是 TEE 技术的安全措施如数据加解密，虽然增强了安全性，但也引入了额外的计算开销，在资源受限的环境中可能成为性能瓶颈。因此，未来研究需要探索优化算法和技术，以在不牺牲安全性的前提下，提高 TEE 处理更复杂或更大规模数据的性能表现。

## 4 结语

本文提出基于 TEE 的 scRNA - seq 细胞分类方法，以应对处理高维稀疏 scRNA - seq 数据及满足新兴隐私保护法规的挑战。利用公开数据集，采用通用分类器 SVM 和单细胞测序专用分类器 ACTINN 在 TEE 中执行细胞分类任务，准确率与在标准环境下运行的模型相当，证实 TEE 在确保数据处理安全性的同时，并不会降低模型性能，为细胞类型识别提供了一个既安全又高效的解决方案，也显示出 TEE 技术促进生物医学研究领域在隐私保护下进行协作的巨大潜力。

本文提出的方法可帮助研究者安全地处理个人健康数据，在确保患者隐私的前提下，进行深入的数据分析，支持基于数据的个性化诊断和治疗策略。应用该方法进行大规模基因组数据分析，对于识别疾病机制、开发靶向药物具有重要意义，有助于药物研发过程的加速和优化。未来工作将专注于优化算法和模型、开发跨平台解决方案、探索实时数据处理，构建合作和数据共享框架，以支持精准医疗、加速药物发现，并提高临床试验的安全性和效率。

**作者贡献**：徐文嘉负责数据分析、论文撰写；岑孟杰负责数据收集与预处理；陈亮负责算法功能实现。
**利益声明**：所有作者均声明不存在利益冲突。

## 参考文献

1 WANG S, SHEN B C, GUO L T, et al. ScFed: federated learning for cell type classification with scRNA - seq [J]. Briefings in bioinformatics, 2023, 25 (1): bbad507.
2 张锋巍, 周雷, 张一鸣, 等. 可信执行环境：现状与展望 [J]. 计算机研究与发展, 2024, 61 (1): 243 - 260.
3 DONG X, LU Y, GUO L, et al. PICOTEES: a privacy - preserving online service of phenotype exploration for genetic - diagnostic variants from Chinese children cohorts [J]. Journal of genetics and genomics, 2024, 51 (2): 243 - 251.
4 CHEN F, WANG S, JIANG X Q, et al. PRINCESS: privacy - protecting rare disease international network collaboration via encryption through software guard extensions [J]. Bioinformatics, 2017, 33 (6): 871 - 878.
5 吴青, 付彦琳. 支持向量机特征选择方法综述 [J]. 西安邮电大学学报, 2020, 25 (5): 16 - 21.
6 欧明望, 叶春杨. 基于神经网络的医疗诊断研究 [J]. 海南大学学报 (自然科学版), 2019, 37 (3): 219 - 226.
7 崔津华, 蔡志平, 刘柯江. SGX 隔离技术研究综述 [J]. 华中科技大学学报 (自然科学版), 2024, 52 (2): 1 - 15.
8 曾凡浪, 常瑞, 许浩, 等. 基于精化的 TrustZone 多安全分区建模与形式化验证 [J]. 软件学报, 2023, 34 (8): 3507 - 3526.
9 MA F Y, PELLEGRINI M. ACTINN: automated identification of cell types in single cell RNA sequencing [J]. Bioinformatics, 2020, 36 (2): 533 - 538.
10 ABDELAAL T, MICHIELSEN L, CATS D, et al. A comparison of automatic cell identification methods for single - cell RNA sequencing data [J]. Genome biology, 2019, 20 (1): 194.
11 DONG X, CHOWDHURY U, LI V X, et al. Semi - supervised deep learning for cell type identification from single - cell transcriptomic data [J]. IEEE/ACM transactions on computational biology and bioinformatics, 2023, 20 (2): 1492 - 1505.
12 ZHENG G X Y, TERRY J M, BELGRADER P, et al. Massively parallel digital transcriptional profiling of single cells [J]. Nature communications, 2017, 8 (1): 14049.