《心理科学进展》审稿意见与作者回应

题目：机器学习方法在测验安全领域的应用
作者：高旭亮；李宁

第一轮

整体修改说明：

首先感谢两位审稿专家花费时间和精力审阅稿件，也感谢编辑部的辛勤付出。非常感谢审稿专家对于本文的认可和指导，专家的点评非常的富有针对性和实际意义，让本文有了“质”的提升。根据审稿专家们的意见，我们对文章进行了较大的修改和完善，为此，我们先简单描述了一个整体的修改说明，再对两位专家的意见进行一一回复。

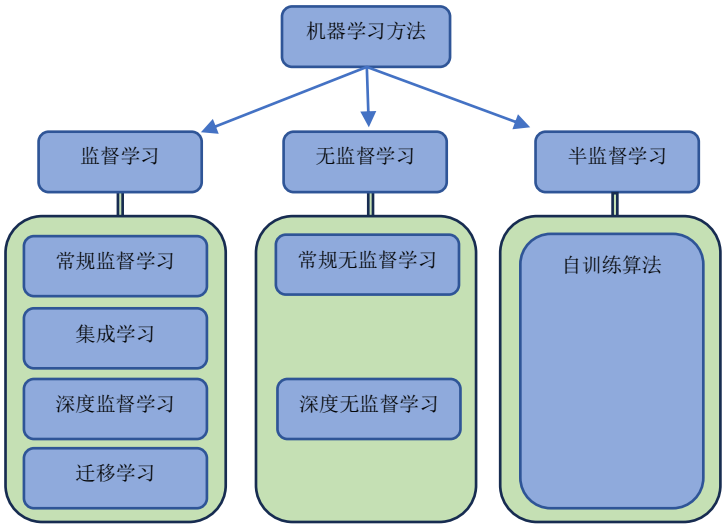
文中存在着最主要的问题是结构混乱、文献收集不齐全以及对方法理解不够清晰的问题，我们在此表示诚挚的歉意，我们对文章结构和内容进行了全面的调整和修改：

①由于文章的目的是讲述机器学习方法的应用思路，因此我们摒弃了先讲述完所有的方法原理再讲述方法应用的结构，理清了各类方法之间的并列和包含关系，从每种方法的角度重新梳理了整个文章，遵循方法原理与应用一起讲的原则，从而使读者更直观的看到每类方法的应用过程和适用场景。并且在文章中对机器学习方法在不同异常类型和测验类型下的选用进行了探讨（见 5.2 章节），我们还总结了如何获取已标记数据的方法以增强方法的实践意义（见 5.3 章节）。

②我们认真搜集了中英文献在测验安全领域的应用研究和理论性综述，引入了测验安全领域一些其他的统计量（如抄袭统计量、变化点分析法统计量），使文章中传统方法的概念不局限于个人拟合统计量。

③我们重新研读了机器学习在测验安全领域的研究，减少了对研究的罗列，将类似结构的研究进行了归纳和总结，缩减了文章篇幅的同时加入了更多自身的理解和探讨，使读者能从中获得更多信息。

文章结构：以方法角度开展，分为监督学习和无监督学习、半监督学习三大类，每大类下设相关小类（由于当前测验安全领域中半监督学习的研究仅有一种自训练算法，因此单列一小节）。



审稿人 1 意见：

这篇综述涉及机器学习方法在考试测验安全中的应用。话题选取既体现了在测验领域中的重要性，也反映了技术发展的前沿。综述本身收罗了本领域的很多重要研究，也进行了有逻辑的分类、总结与点评。因此，本文是一篇具有较高水平的综述论文，对推进本话题的进展具有一定意义。在此，本人给出较为正面的意见，但是请作者对下面问题进行思考和回应。

意见 1：第一，在 2.5 机器学习算法简评部分，没有包含“强化学习”与“半监督学习”。可能是因为这个话题尚未涉及这两种技术，但是需要在综述中对此进行简单说明；

回应：感谢审稿专家提出的问题，没有介绍“强化学习”确实是由于话题未涉及，我们已在文中引言部分进行了说明。当前话题中涉及到了“半监督学习”，而且该方法是解决监督学习缺少标记数据的一种重要方法，因此我们在原文中草率的将其纳入到了监督学习中。我们重新梳理了文章结构，半监督学习的介绍详见 4 章节。

1 章节引言相关内容：“最终参考了 Alpaydin(2020)和 Géron(2022)的著作，结合当前测验安全领域的应用研究和机器学习算法的学习方式分为监督学习、无监督学习、半监督学习三大类(Alpaydin, 2020; Géron, 2022)进行述评，由于当前研究中涉及强化学习极少，遂未对此大类作介绍。除此之外还介绍了集成学习、深度学习和迁移学习，这些方法与三大类方法并不是并列关系，集成学习可以通过监督、无监督或者半监督方法完成(Dong et al., 2020)，深度学习中既有监督学习、无监督学习也有半监督学习方法(Goodfellow et al.,2016)，迁移学习可以实现对监督或者无监督模型效果的迁移(Weiss et al., 2016)。鉴于方法种类较多且各类方法之间的关系较为复杂，我们根据当前研究中的实际情况对方法进行了归纳：当前研究中的集成学习基础模型都是监督学习模型，因此纳入监督学习板块进行介绍；迁移学习使用的基础模型是监督学习模型，因此一同纳入监督学习板块；将深度学习中涉及监督、无监督和半监督学习的内容分别纳入对应的板块；有些研究中包含了各种不同的方法，我们尽可能按照其核心方法对其进行分类”。

意见 2：3.1，3.2 与 3.3——3.8 关系不明。请明确“项目预知与题目泄露检测”与后面 5 种检测方式是什么关系？ 后面这五种方法是从什么角度定义的？

回应：感谢审稿专家提出的宝贵意见。由于机器学习在测验安全领域中对于项目预知的研究比较多，我们原意是想单独为其开设一个章节进行讲述，但是除项目预知外的测验类型和机器学习方法都比较零散，因此我们对于讲述的重点有所混淆，造成了定义角度不明的情况。根据审稿专家的意见，我们已经将其修改为从监督学习、无监督学习、半监督学习三大类进行归纳和述评，将原文中的 3.1-3.8 章节进行拆分归纳并置入了对应的方法类别之中，这样可以清楚的了解每类方法的应用。

意见 3：5“开放性问题试题”与前面的教育测量、心理测量什么关系？ 从使用“试题”一词还有表 2 的内容来看，归属于教育测量，但是为什么把心理测量放在中间？ 这样的文章结构令人费解。

回应：感谢审稿专家的意见，我们对结构的混乱表示诚挚的歉意，由于开放性文本试题是对文字进行识别和转换，与绝大多数使用数值型数据的教育测量和心理测量研究不同，我们原意是单独为其设一章节，但是只顾及了开放性文本试题的特殊性，却未周全的考虑到文本试题也是属于教育测量的一部分，因此造成了这种结构的混乱。考虑到教育测验与心理测验的交叉性较强，我们更改了这样的结构，将两种测验统一纳入了方法大类之中，在修改的文章中我们将开放性文本试题的研究纳入到了监督学习类型中进行讲述，详见 2.1.2 章节尾段。

意见 4：“开放性问题试题”是否能当综述作文抄袭有关的研究，或者由生成式人工智能生成的作答？目前这两题问题显然属于这个大话题下最炙手可热的子领域。

回应：感谢审稿专家的指导，这条意见具有十足的前瞻性和延展性，我们对其做了一定的解释（见 2.1.2 章节尾段）并对其进行了展望（6 章节）。

2.1.2 章节中的相关内容：“当前将机器学习文本挖掘技术应用于测验作弊的研究极少，相关领域大多研究都集中在学术剽窃，而测验中的开放性文本试题有字数少、回答零散等特点，且主要是为了检测同一考场中的考生是否存在互相抄袭行为，因此与学术论文的抄袭研究的重点有所差别，学术抄袭的研究更集中于对于大型段落的再译、近义式抄袭，目的是通过机器学习识别语义特征等”。

6 章节中的相关内容：“此外，当前机器学习的文本挖掘技术已经充分运用到检测学术上的剽窃抄袭(Foltýnek et al., 2019)以及检测学术论文或大型文字任务中的人工智能生成内容(Taloni et al., 2024)，由此看来，使用文本挖掘技术检测异常受试者是非常有潜力的”。

审稿人 2 意见：

本文聚焦于机器学习方法在测验安全领域的应用进行综述，针对测验安全中的机器学习方法应用的介绍是具有一定意义的，有助于梳理机器学习方法在测验安全领域应用的发展脉络。

意见 1：但是文章中的一些观点值得推敲，比如摘要中对机器学习优点，和 PFS 方法缺点的表述存在不准确之处，机器学习方法在实际测验数据中检测异常考生或异常题目时会受到更大的局限，比如受到样本量、标签等的限制，可能还没有 PFS 好用。因此，作者不能因为推崇某一方法而过于否认其它方法。

回应：非常感谢专家的宝贵意见，我们深刻认识到对机器学习和统计量方法的优缺点描述有失偏颇，机器学习方法最主要的特点是灵活性强，更容易纳入一些统计量方法难以纳入的变量，受理论框架和测验类型的限制也少一些，但是最其致命的缺点是受数据影响严重，数据有无标签、数据质量高低都会对方法的选用和效果产生很大影响。我们重新审视了文章中对于两种方法的整体态度，这篇文章的出发点是：机器学习方法的出现可以弥补一些统计量方法的局限性，两种方法应该是相互补充、相得益彰的关系，在文中我们也强调了这一点。当前研究中有大量机器学习方法与统计量方法混用的例子，例如将统计量作为机器学习输入特征、使用统计量先行标记异常被试从而获取训练集中的标记数据等等，因此，这是一个传统与现代相融合的方法，目的都是为了检验异常被试。

1 章节中的相关内容：“统计量通过被试对项目反应的统计分布与特定的理论进行统计方法的研究，其理论研究不断趋于完善，在实证研究中也取得了不错的效果，算法简单易用，广受研究者推崇，但其也存在一定的局限性”；“机器学习在测验安全领域不失为一种与统计量互补的友好方式，本领域的很多研究是机器学习与统计量相结合，因此两种方法可以做到相互补充、相得益彰”。

2.1.2 章节中的相关内容：“Meng 和 Ma (2023)通过高检验力和不受测验条件限制的统计量对作弊考生进行标记，然后找出作弊考生在各方面的具体特征放入机器学习模型进行学习，这样模型可以对数据中更多接近这个特征的考生进行标记，充分利用了统计量的优势和机器学习方法的优势”。

2.2.2 章节中的相关内容：“Jiao 等人(2023)还尝试了使用基于项目反应和反应时的个人拟合统计量和异常值检测算法的结果作为输入特征来提升集成学习的效果”。

5.3 章节与 6.1 章节中也有相当篇幅的相关内容，感谢专家关注！

意见 2：本文作者对于测验安全领域一些重要文献并没有关注到，一方面，文章列出的参考文献中完全没有中文文献，其实近年来，已有一些中文文献在讨论测验安全的主题；另一方面，即使是英文文献，也还有一些值得纳入的重要文献，这些都不能被忽视。

回应：感谢专家的意见，我们对这种疏忽大意感到抱歉，我们重新梳理并纳入了测验安全领域的中英文献，尽量囊括了个人拟合统计量、抄袭统计量、以变点分析法构建的统计量以及除统计量外的混合模型等方法，并尽量纳入了一些详尽描述各种机器学习方法的书籍和文献，以确保文章的完整性。还纳入了一些将机器学习应用于心理学领域的典型中文文献，如有疏漏还烦请专家不吝赐教。

加入的文献如下：

韩丹，郭庆科，王昭，陈雪霞. (2008). 考试抄袭识别的心理测量学研究回顾. *心理科学进展* , *16*(1), 175-183.

胡佳琪，黄美薇，骆方 .(2020). 考试作弊甄别技术的研究进展：个体作弊的甄别.*中国考试*(11),32-36.

黄美薇，潘逸沁，骆方.（2020）.结合选择题与主观题信息的两阶段作弊甄别方法, *心理科学*, (1): 75-80.

李亚.(2022). *基于机器学习的考生异常行为识别研究*（硕士学位论文）. 东北林业大学.

刘冬予，骆方，屠焯然，饶思敬，沈阳 .(2024). 人工智能技术赋能心理学发展的现状与挑战.*北京师范大学学报(自然科学版)*(01),30-37.

刘玥, 刘红云. (2021). 心理与教育测验中异常作答处理的新技术: 混合模型方法. *心理科学进展* , *29*(9), 1696-1710.

骆方，田雪涛，屠焯然，姜力铭 .(2021). 教育评价新趋向：智能化测评研究综述.*现代远程教育研究*(05),42-52.

骆方，王欣夷，徐永泽，封慰.(2020). 考试作弊甄别技术的研究进展：团体作弊的甄别.*中国考试*(11),37-41.

彭恒利, 孔祥 .(2015). *标准化考试作弊甄别的理论与方法*. 北京: 北京语言大学出版社.

童昊, 喻晓锋, 秦春影, 彭亚风, 钟小缘. (2022). 多级计分测验中基于残差统计量的被试拟合研究. *心理学报*, *54*(9), 1122-1136.

王昭;郭庆科;岳艳. (2007). 心理测验中个人拟合研究的回顾与展望. *心理科学进展* , *15*(3), 559-566.

徐静, 骆方, 马彦珍, 胡路明, 田雪涛. (2024). 开放式情境判断测验的自动化评分. *心理学报*, *56*(6), 831-844.

张龙飞, 王晓雯, 蔡艳, 涂冬波. (2020). 心理与教育测验中异常反应侦查新技术：变点分析法. *心理科学进展* , *28*(9), 1462-1477.

钟小缘, 喻晓锋, 苗莹, 秦春影, 彭亚风, 童昊. (2022). 基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数. 心理学报, 54(10), 1277-1292.

钟晓钰, 李铭尧, 李凌艳. (2021). 问卷调查中被试不认真作答的控制与识别. *心理科学进展* , *29*(2), 225-237.

Alsabhan, W. (2023). Student cheating detection in higher education by implementing machine learning and LSTM techniques. *Sensors*, *23*(8), 4149.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241-258.

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, *52*(6), 1-42.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied measurement in education, 16*(4), 277-298.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer science & business media.

Ranger, J., Schmidt, N., & Wolgast, A. (2020). The detection of cheating on E-exams in higher education—the performance of several old and some new indicators. *Frontiers in psychology, 11*, 568825.

Taloni, A., Scorcia, V., & Giannaccare, G. (2024). Modern threats in academia: Evaluating plagiarism and artificial intelligence detection scores of ChatGPT. *Eye, 38*(2), 397-400.

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577-596.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1-40.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1-130.

意见 3：文章中对于机器学习的一些方法的总结和归纳还存在概念上的混淆，比如将监督学习、无监督学习、集成学习和迁移学习进行并列，这四个概念之间并不是完全并列的关系。

回应：感谢专家的指导，这些方法确实不完全是并列关系，我们认真的研究了几种方法的包含、并列关系，最终参考了 Alpaydin(2020)和 Géron(2022)的著作，结合当前测验安全领域的应用研究和机器学习算法的学习方式分为监督学习、无监督学习、半监督学习三大类(Alpaydin, 2020; Géron, 2022)进行述评，由于当前研究中涉及强化学习极少，遂未对此大类作介绍。除此之外还介绍了集成学习、深度学习和迁移学习(这些方法与三大类方法并不是并列关系，集成学习可以通过监督、无监督或者半监督方法完成(Dong et al., 2020)，深度学习中既有监督学习、无监督学习也有半监督学习方法(Goodfellow et al.,2016)，迁移学习可以实现对监督或者无监督模型效果的迁移(Weiss et al., 2016))。鉴于方法种类较多且各类方法之间的关系较为复杂，我们根据当前研究中的实际情况对方法进行了归纳：当前研究中的集成学习基础模型都是监督学习模型，因此纳入监督学习板块进行介绍；迁移学习使用的基础模型是监督学习模型，因此一同纳入监督学习板块；将深度学习中涉及监督、无监督和半监督学习的内容分别纳入对应的板块；有些研究中包含了各种不同的方法，我们尽可能按照其核心方法对其进行分类。

意见 4：从整个论文来看，作者更多是将三种类型的测验（教育测量、调查问卷和开放性文本测验）下的机器学习研究进行的罗列，并且对每类方法只进行了很简单和表面的介绍，而没有进行更深入的理解和探讨。

回应：感谢专家批评指正，针对这个问题，从结构上我们基于讲清楚每种方法应该如何应用到测验安全中去的思路，从机器学习方法的角度重新梳理了整个文章，对每类方法都下设“方法介绍”和“应用研究”，丰富了原文中方法介绍的部分，并对深度学习、半监督学习方法做了补充（详见 2.3 章节、3.2 章节、4 章节）。在每类方法下的“应用研究”部分加入了更多理解和探讨性的内容，将结构相似的研究进行归纳，总结研究的特点，探讨研究的优势和局限性，减少了罗列堆砌感，整体上使读者能更直观的看到每类方法的应用过程和适用场景。我们还在 5 章节对方法进行了综合的分析，针对不同测验类型和异常类型提出了选用建议。

我们期望读者能够从文章中了解不同方法应用思路后再去选择和学习对应的机器学习方法。由于文献中使用到的每个大类下的具体方法十分繁多，我们深知自身知识浅薄，无法对每种方法进行算法层面的细致推导，于是尽可能的加入了讲解各种机器学习方法的高引用

量书籍和比较前沿的文献以便读者查阅。

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4, p. 738). New York: springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *JAIR Journal of Artificial Intelligence Research*, 16,321–357.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241-258.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*, 85-126.

Khyani, D., Jakkula, S., Gowda, S., KJ, A., & KR, S. (2021). An interpretation of stacking and blending approach in machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *8*(07).

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Liu, H., & Motoda, H. (2012). *Feature selection for knowledge discovery and data mining* (Vol. 454). Springer science & business media.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1-40.

意见 5：从 3.2 节开始，论文的逻辑就比较混乱，更多的像是将相关研究的堆砌，难于判断作者的行文逻辑。

回应：感谢专家的意见，根据审稿专家这条宝贵意见，我们再次根据文章的核心思路修正了这一点，从监督学习、无监督学习、半监督学习三大类方法出发，每类方法都下设“方法介绍”和“应用研究”，最后总结三类方法并提出使用建议。由此完善了文章结构，使文章逻辑更加清楚（可见整体修改说明中的结构图）。为了减少文章的堆砌感，我们在丰富内容的基础上缩减了文章的篇幅，对类似研究进行了归纳，加入了更多自身的理解和探讨。

意见 6：本文当前存在重点不突出，文章前后缺乏逻辑性，未找到明确的行文脉络，方法介绍与方法总结不具有针对性。如果能结合测验安全领域中各测验类型下的不同异常类型具体特征与机器学习中各方法适用的场景进行综合讨论，会更具有实践意义。

回应：感谢专家的综合性意见，我们的重点是“讲清楚三类方法的应用思路、如何选用方法以及方法使用过程中的建议”。

在逻辑和行文脉络上我们首先在引言中讲解了测验安全的威胁和机器学习在处理测验安全的优势，然后在监督学习、无监督学习、半监督学习三大类下设小类，并在小类方法下设“方法介绍”与“应用研究”，讲完一类方法后紧接着讲一类方法的应用，从结构上增强方法介绍的针对性，从而使读者更直观的看到每类方法的应用过程和适用场景，最后对方法进行综合分析并提出使用建议。

除了在结构上增强方法介绍的针对性，在内容上我们首先在引言中阐明了方法选用的前提：“在测验安全领域中大部分情况都是处理数值型数据的异常分类任务，在根据数据情况和任务目标正确选用方法的前提下，每种方法都可以处理大部分异常类型”，并且在 5.2 章节结合各测验类型下的不同异常类型具体特征与机器学习中各方法适用的场景进行了方法选用的讨论：“当前的研究中的异常类型总共可以分为以下两种：①教育测验中的异常反应，

如作弊、随机作答、疲劳作答等；②调查问卷测验中的粗心作答(受试者由于动机低下而随机作答、直线作答或者规律作答等)。在基于计算机的测验中我们多数时候能获得的数据都包含最基础的项目反应、反应时，部分测验还会提供诸如考生的修改答案次数、情绪、点击流等更丰富的过程数据，我们往往是根据数据的情况选用三种方法(有无已标记数据？已标记数据的数量？)，由于机器学习是学习数据规律的方法，因此这三种方法在多数的异常类型下都是适用的，只是我们选择的输入特征会有一些不同的侧重点，例如我们想检测项目预知，我们就会重点关注考生极速答对的项目反应时并作为输入特征，如果我们想检测被试在测验尾部的疲劳作答，我们就会将尾部的项目反应等变量作为输入特征。

有一部分异常类型有很强的随机性，教育测验中的受试者在作答动机不强时可能会对任意题目进行随机作答，这就导致项目反应和反应时都会比较随机；另一种是调查问卷中的粗心作答，调查问卷与教育测验的过程数据有着显著的不同，在教育测验中考生在项目上的项目反应和反应时遵循着随着题目难度上升则分数降低、反应时增加的基本规律，而在调查问卷中除了明显异常的连续一致作答和规律作答，我们很难去判断其是否认真作答。因此面对这些随机性强、过程数据无明显规律的异常类型，常用的监督学习对其效果并不明显，这一点也经过了研究者的佐证(Schroeders et al., 2022)。目前比较有效的方法是无监督学习中的异常检测方法，尤其是深度无监督学习中的自编码器，在异常检测中，自编码器通常被用来学习正常数据的表示，一旦训练完成，自编码器可以用来重构新的输入数据。如果重构误差(即重构的数据与原始数据之间的差异)超过了某个阈值，就可以将该输入数据标记为异常，而随机性强的异常反应往往结构十分混乱、重构误差较大，因此可以较好的被识别出来”。

意见 7：综上，因此无论是从内容，还是从结构上本文还需要进行大的修改和调整。
回应：感谢专家的整体建议，我们对文章进行了大规模的修改、调整与完善，希望能使读者从中获益，也希望修改能符合您的预期！

第二轮

审稿人 1 意见：
作者已经较好的回复了我的意见，没有进一步的意见。

审稿人 2 意见：
论文在内容和结构修改后质量有了明显的提高，看得出来作者做出了努力。还存在一些问题。

意见 1：展望部分需要结合一些最新的研究展开，并且需要有实质性的内容，而不能只是一般性的介绍
回应：感谢专家的意见，我们的展望部分确实存在实质性内容不够的问题，我们研读了当前该领域的前沿文献，极大的丰富了展望部分的内容，主要丰富的内容在于：6.1 如何进行基于机器学习的个人拟合研究；6.2 基于多模态数据的机器学习测验安全研究，介绍了机器学习领域不同类型数据的测验安全研究，提出融合多模态数据进行更综合性、更有可信度研究的展望；6.3 如何使用生成对抗网络(GAN)进行异常数据生成和检测；6.4 关于增强研究结果可解释性的建议。重写后的内容丰富详实，希望能为读者提供有益的见解！详细内容请参阅 6 章节。

意见 2：6.2 节提到“...在过程中尽量要求其认真作答..."，这种表述不应该出现，而且也没有操作性
回应：感谢专家的细致意见，此处确实表达不当，我们已在文中将该表述删除。

意见 3：作者在论文中传递了一种误解，即机器学习之外的方法需要较大的样本量的缺点，机器学习似乎没有这样的缺点，这一点需要澄清，否则容易造成对读者的误导。统计量方法对于样本量的依赖并不会强于机器学习方法，相反，机器学习方法可能对样本的依赖更严重。
回应：感谢专家的点评，我们认识到了这个问题并进行了补充和澄清，在正文第 3 页第 2 段的引言中我们在描述了统计量方法的局限性之后和描述机器学习方法优势前补充了如下内容：“虽然机器学习方法存在对样本数据质量和数量要求高、模型可解释性差、实验重复性差等争议性问题，但相比统计量方法来说，机器学习仍有一些优势……”，以此避免读者误解。专家的建议让我们进一步想到机器学习对样本质量和数量的高需求，于是在 5.1 章节三种方法的综合分析中我们补充了如下内容：“机器学习方法整体上最大的局限性就是受数据的数量和质量影响，数据是模型的营养，如果数据质量低下或数量较少，任何一种机器学习方法的效果都不会太好”。

意见 4：一些写作方面的问题如
（1）引用文献时，如果是放在括号里，三个或以上作者时，只需写出第 1 作者，第 1 作者后的“等”字需要与前面留一空格，这个问题在论文中很多处都存在，比如引言的第 3 段，至少有 5 处以上类似的不规范；
回应：感谢专家耐心指出格式问题，我们认真检查了格式问题，已经全部修改！
（2）2.2.2 的第 1 句话，开头的文献引用应该为“Zhou 和 Jiao (2022a, 2022b)”。
回应：感谢专家耐心指出格式问题，此处已经修改！

第三轮

审稿人 2 意见：
本次修改基本上解决了我提出的问题，还存在两个小的问题：

意见 1：作者从 5.3 节又重新开始了新的页码。
回应：感谢专家认真细致的审稿，该问题已经修改！

意见 2：6.1 下面的第 7 行的引用 zhu 应该改为 Zhu。
回应：感谢审稿专家，该问题已经修改！

第四轮

编委 1 意见：经过两轮审稿与修改，该文有较大的改进。但行文和文句还有改进空间，尤其是引言部分。引言要在背景介绍基础上尽快切入本文要做什么，摘要缺乏实质性内容，已经用修订模式做了些具体修改，供作者参考。
回应：感谢编委的意见和批改，摘要和引言确实存在这样的问题。我们对摘要和引言部分进行了精简和完善：摘要部分增添了关于文章核心的实质性内容，引言部分我们删减了冗杂的表述，使其更加精炼易读且直入主题，引言逻辑如下，测验安全问题的危害---当前解决测

验安全问题的主流方法---主流方法面临的挑战---新方法的优势---文章结构与核心内容。同时我们对全文进行了进一步的细致检查，使得行文更加流畅。希望达到读者的预期！

编委 2 意见：同意两位审稿人的意见，建议发表。

第五轮

主编意见：有新内容，值得发表