《心理学报》审稿意见与作者回应

题目：基于事件相关电位（ERPs）和机器学习的考试焦虑诊断
作者：章文佩，沈群伦，宋锦涛，周仁来

第一轮

审稿人 1 意见：

意见 1：这一研究中，使用 TAS 得分作为学生高/低考试焦虑的指标，同时 TAS 又是 CNN 的训练的预测变量，亦即训练后的 CNN 能以一定的准确率通过 ERP 数据预测学生的 TAS 得分水平，这一使用实验任务和 ERP 数据进行预测的方法相比于直接使用 TAS 量表进行测量有何优势？

回应：这个问题涉及到研究的核心问题之一，即我们提出的基于事件相关电位（ERPs）和机器学习的考试焦虑诊断准确率的评估方法问题。我们将从三个方面来回答这个问题：

（1）TAS 的局限性：在传统心理学研究中，TAS 量表具有比较好的信效度，尤其是在测量群体时，它是一种可靠的区分高低考试焦虑的工具。但是同时单独使用 TAS 量表进行评估具有一定的局限性，即它的主观性比较强，很容易受到被试主观因素的干扰（见正文第 1 页第三段，“目前，国内外对考试焦虑的研究中……”），这种干扰尤其体现在对个体的诊断上，即个人当前的状态、意愿等会较大影响到对此个体进行考试焦虑程度的评估，使其结果变得不够稳定。而我们的诊断应用的目的在于对每个人达到精准评估，而不是仅限于传统研究中对群体的准确划分，因此才选用更加客观的 ERP 及机器学习方法减少这种不稳定性。

（2）使用 ERP 技术和机器学习的优势：首先，ERP 的测量可以有效减少主观因素的不稳定性（建正文第 2 页第二段，“为了降低这些限制，……”）；其次，机器学习对于数据的深度学习，尤其是采用 CNN 对全脑数据的深度学习可以减少信号的不稳定性对整体判断的影响，从而进一步减少个人因主观因素引起的不稳定性，增加对个体考试焦虑程度的准确度。

（3）最后，由于 TAS 既是学生高/低考试焦虑的指标，又是 CNN 的训练的预测变量，如果在筛选高/低考试焦虑时单独使用 TAS 量表，则无法避免之前提到的 TAS 的局限性，因此我们在筛选被试时参考前人研究(Lu, Jiang, & Liu, 2017)还进一步结合了两位心理学专家的专业评估（见正文 2.1 被试招募部分介绍），以减少 TAS 量表使用本身的局限性，提高我们在研究中筛选被试的准确性。

最终，我们通过结合 ERP 数据和机器学习进行预测就可以得到一个更为综合客观的评估方法，尤其是在遇到被试不想表达自己真实意愿（如不愿意让别人知道自己是考试焦虑，故意不认真作答问卷）以及不适合给被试做主观问卷的场景（如临近考试事件，不宜进行问卷调查以避免文字诱发更高的考试焦虑情绪）时具有更好的效力。

意见 2：（见论文自检报告第 1 项）使用 CNN 比传统方法有更高的预测准确率，是否有证据支持？文中只是指出了 CNN 方法的准确率，没有与传统方法进行比较；其次，对于具有独特认知特点的疾病（考试焦虑），ERP 比 EEG 的学习更好，是否有证据支持？这两项作为这一研究主要的理论意义所在，为何在文中没有详细说明？

回应：（1）CNN 比传统方法有更高的预测准确率问题：CNN 对于脑电类型的数据具有独特的分析优势，具体原因及证据来自于前人研究：

（a）CNN 作为一种强有力的非线性模型，两层神经网络理论上就可以拟合任意的数据

分布(Hornik, Stinchcombe, & White, 1989)，而传统方法往往做不到这一点；

（b）机器学习是一种表示学习，而特征就是表示的关键，这决定了模型的上限(Domingos, 2012)，CNN 通过构造特征图的实现了强有力的特征提取，除此之外它还有一系列传统方法所没有的特性（见正文 14 页第二段中“因为 CNN 模型拥有.....”）

（c）CNN 在图片分类(Krizhevsky, Sutskever, & Hinton, 2012)、分割(Long, Shelhamer, & Darrell, 2015)等一系列任务中已经取得了非常好的，远超于传统方法的表现，是我们认为它能在该任务上取得良好表现的理由。

（d）感谢您的建议，没有与传统方法的比较说明确实不能凸显出使用 CNN 的独特优势，因此我们已经在正文中前言关于 CNN 的部分进行了进一步的补充（见正文第 3 页最后一段蓝色字体部分）。

（2）对于具有独特认知特点的疾病（考试焦虑），ERP 比 EEG 的学习更好，是否有证据支持：

首先，关于 ERP 比 EEG 的机器学习质量更好的直接证据我们暂时没有检索到，可能机器学习在对于人群分类的这类研究还比较新，所以可以参考的直接资料较少。

其次，我们将从两方面来回答我们提出用 ERP 而不是 EEG 来学习的原因：

（1）基于前人对与考试焦虑的传统方法研究，关于考试焦虑的注意及认知特点的相关研究结果都指向考试焦虑是一种情境性与诱发性较强的情绪障碍：（a）情境性：高考试焦虑者对考试的不合理认知及负性情绪可以由考试事件激发，在考试事件发生的前后期间，高考试焦虑者会有较为显著的焦虑情绪及认知(Lotz & Sparfeldt, 2017)，但是我们在本次研究中没有采用情境性的范式，因为我们的目的之一是为了避免考试评估方法的本身会诱发的考试焦虑等负面情绪（见正文第 2 页第一段中“诱导性”的介绍；（b）诱发性：在前人研究中，高考试焦虑者对考试的不合理认知及负性情绪可以由考试相关刺激（如考试相关词汇）诱发产生，一般而言，这类研究范式都是事件相关的设计(Liu, Zhang, & Zhou, 2015; Putwain, Langdale, Woods, & Nicholson, 2011)。为了减少范式诱发的负性情绪，本次研究中所采用的研究范式类型是基于注意的情绪 Stroop 范式，这样既可以有效测量个体对考试焦虑的认知又不会过多激发个体的负性情绪。

（2）EEG 信号在考试焦虑者区分中的不稳定性：在前人研究中，不同于 ERP 的诱发电位，EEG 属于自发电位，由于其噪音及其不稳定性，并没有较为可靠的、具有共识的可区分考试焦虑的特异性指标。同时，我们在未发表的研究中也对高、低考试焦虑的 EEG 信号进行了采集与分析，并没有得出显著差异。而机器学习的效率是基于数据本身的质量，如果数据本身没有很好的鉴别力，那机器学习的准确率可能不佳，因此我们才提出了使用指向性较强的 ERP 而不是 EEG 来学习。

（3）ERP 是在自发脑电 EEG 的基础上基于事件相关的诱发电位，这种诱发的脑电位变化是直接由当前的刺激产生的，并且这种电位变化反映的是个体对当前刺激的认知态度(Biedermann et al., 2016; O’Toole & Dennis, 2012)。在这样的基础上，由于刺激性质具有确定性，使得刺激背后反映的认知态度具有一定的稳定性，并且在前人的大量研究中得出了一些具有特定认知意义的特异性指标，如我们在本研究中选取的 P1，P2，N2，P3 和 LPP 成分，都具有较典型的认知含义，因此能够很好地反映出个体对当前刺激（考试相关威胁信息）的认知。

（4）本文在正文的第 2 页第三段以及第 3 页的第二段介绍了 ERP 的优势以及我们采用 ERP 的理论依据。但是由于 ERP 本身基于 EEG，且本文的重点不是比较机器学习在 ERP 与 EEG 数据上的差异，因此没有对 ERP 与 EEG 的学习效果进行深入的比较说明。

意见 3：（见论文自检报告第 6 项）此项作者只回答了实际样本量的问题，没有提及计划的样本量及样本量的选择依据。
回应：感谢您的问题，这一点确实没有在自检报告中说明，已经补上。

意见 4：（见正文 2.1）请报告被试的性别比例，高低考试焦虑组为何出现人数不均衡的问题？
回应：回应：（1）性别比例，我们已经在文中方法部分（见正文 2.1 被试招募部分）补上被试的性别比例；（2）高低考试焦虑组出现人数不均衡的问题：这一点之前没有进行说明确实不妥，现在已经在自检报告第 6 项中补充。

意见 5：（见正文 2.2）请报告本研究中量表的结构效度
回应：我们采用的相容效度法测量的 TAS 的结构效度，但是原文表述不够清晰，现在已经更正（见正文 2.2，“量表的结构效度采用与考试焦虑测验（TAI）的相关测得，……”）。

意见 6：（见正文 2.3）被试是否签署知情同意书，实验是否经过相关伦理审核？
回应：被试在实验前已经签署知情同意书，均为自愿参加实验，并且该实验也已通过相应机构的伦理审核及获得了相应的批号。这部分说明已经补充进正文 2.1 被试招募部分（“该实验已经通过伦理委员会的审查……”），由于批号上包含单位名称，故在此省去，如审稿人需要，将在文中补充伦理审核的批号。

意见 7：（见正文 2.5）数据预处理没有提及量表数据的预处理以及被试是如何被划分为高/低考试焦虑组；直接对所有数据除以 15 而不是使用 Z 分数进行正则化，是否真正解决了数据指标间可比性的问题，为何不使用争议较小的 min-max 法或 z 分数？
回应：（1）量表数据的预处理和划分依据已经补充在正文 2.2 的问卷说明中，感谢您的提醒；（2）对图片数据进行分类的时候（图片数据中每个元素的取值范围在 0~255），往往使用的是简单放缩的办法，将每个元素除以 255，使其取值范围变为 0~1，在这里也是用了这一种做法，将数据除以他们绝对值的最大值 15，使得数据取值范围变成-1~1。由于这组数据中并不存在量纲相差很大的情况，所以我们没有采用类似的归一化/标准化，而是选择在最大程度上保持数据原本的形态。并且，根据您的问题，我们对未进行正则化的数据进行了建模，得到了和原来一样的结果。

意见 8：使用 F1 分数作为指标时，请在结果中报告 CNN 结果的查准率、查全率和 G-MEAN。
回应：感谢您的建议，查准率、查全率、G-MEAN 的数值已经补充在原文中（见正文 12 页表 2）。

意见 9：（见正文 5.1）五种成分在不同词汇下的波幅差异请报告差异检验的结果、显著性与效果量。
回应：（1）五种成分在不同词汇下的波幅差异统计结果：由于 Stroop 范式的设计理念在于通过比较条件（即目标条件-考试相关威胁词与参照条件-中性词）之间反应的差异来判断个体对目标条件付出资源多少，因此我们采用的是对五种成分在两种条件下的 ERP 波幅进行组别*条件的重复测量方差分析方法。并且，我们已经将对五种成分 ERP 波幅的统计方法（补充在正文 2.5 数据预处理部分的段末，见蓝色字体部分）和统计检验结果（补充在正文 5.1ERP 结果中，见蓝色字体部分）补充在正文相应的部分中。此外，由于涉及到统计检验，我们也对自检报告的问题 5 和 7 做了相应的补充回答。
（2）一开始没有在正文中报告的原因：起初，我们考虑到文章方法和结果的重点是机器学

习，所以一开始就没有呈现传统的心理学统计检验结果，想让读者只专注于对机器学习的关注。但是如果没有对这选取的五种成分进行一个基本的统计检验，则无法从客观角度证明这五种成分的选择是否正确、这五种成分在区分高、低考试焦虑这两类人群是否具有基本的鉴别力。感谢您的建议，这部分的补充让文章逻辑更完整。

意见 10：使用机器学习的方法进行预测时，是否对训练集进行了多次训练？训练的取样方式如何？是否进行多次预测并取平均准确率作为指标？这些问题没有在方法部分体现出来。
回应：感谢您的问题，我们依照审稿人的建议在新的稿件中采用多折交叉验证，对现有数据重新进行了机器学习，得到了更为稳定的计算结果，提升了模型的质量。非常感谢！

在初稿中，我们只是随机抽 80%的数据量作为训练集，剩下的 20%的作为测试集。每一种算法都是在训练集上的所有数据进行一次训练，再用训练得到的模型在测试集上进行预测。这样的做法可能会因为随机性而导致算法在测试集上表现很好，得到的结果不够有置信度。

而为了得到对模型表现更有信度的评价指标，我们将原有的方法更改为 k 折交叉验证，即：将样本分为 k 份，这样可以选出 k 种不同的 k-1 份作为训练集，将剩下的一份作为测试集，以训练的得到的模型在测试集上的预测结果进行评价指标的计算，最后将 k 次预测得到的指标进行平均作为最终的结果。由于样本只 82 个，为了保证每一折中的测试样本不太少，这里 k 取 3，同时为了对修改后的训练方法进行建模稿件中的 CNN 的架构也有所调整。

意见 11：在对 CNN 方法进行解释，尤其是对矩阵进行操作时，建议尽量采用标准规范的数学方程进行说明，图片可作为辅助说明的手段。
回应：感谢您的提醒，数学公式已在正文中补上（见 3.1 卷积层部分说明）。

参考文献：

Biedermann, B., de Lissa, P., Mahajan, Y., Polito, V., Badcock, N., Connors, M. H., … McArthur, G. (2016). Meditation and auditory attention: An ERP study of meditators and non-meditators. *International Journal of Psychophysiology*, *109*(September), 63–70. https://doi.org/10.1016/j.ijpsycho.2016.09.016

Domingos, P. M. (2012). A few useful things to know about machine learning. *Commun. acm*, 55(10), 78-87.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

Liu, Y., Zhang, W., & Zhou, R. (2015). Cognitive and Neural Basis of Attentional Bias in Test Anxiety Students. *Psychological Exploration*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Lotz, C., & Sparfeldt, J. R. (2017). Does test anxiety increase as the exam draws near? – Students' state test anxiety recorded over the course of one semester. *Personality and Individual Differences*, *104*, 397–400. https://doi.org/10.1016/j.paid.2016.08.032

Lu, Y., Jiang, H., & Liu, W. B. T.-I. C. on C. E. and N. (2017). Classification of EEG Signal by STFT-CNN Framework: Identification of Right-/left-hand Motor Imagination in BCI Systems.

O'Toole, L., & Dennis, T. A. (2012). Attention training and the threat bias: An ERP study. *Brain and Cognition*, *78*(1), 63–73. https://doi.org/10.1016/j.bandc.2011.10.007

Putwain, D. W., Langdale, H. C., Woods, K. A., & Nicholson, L. J. (2011). Developing and piloting a dot-probe measure of attentional bias for test anxiety. *Learning and Individual Differences*, *21*(4), 478–482. https://doi.org/10.1016/j.lindif.2011.02.00

审稿人 2 意见：

本文采集在 stroop 范式下用户的脑电数据，然后建立基于脑电数据的考试焦虑预测模型，从结果上看深度学习的预测性能最好。文章具有一定的创新性，但是也存在下面一些问题：

意见 1：被试只有 82 名，数量比较少。从脑电数据的特征提取看，特征的数量比较大，可能会远远高于样本数；

回应：首先，关于本研究的样本量问题。我们在研究之前参考了前人使用机器学习辅助疾病诊断/分类的相似研究(如：Halicek et al., 2017, N=50; Vandenberghe et al., 2017, N=71; 王艳娜 & 孙丙宇, 2017, N=13; 高军峰, 王沛, & 郑崇勋, 2010, N=30)，在这些文献中样本量的选取都不多，但是由于其样本质量可靠（如有专家综合诊断为依据），样本量并不会显著影响机器学习的效果。因此，参照前人研究，本研究的样本量的确定基于两点：（1）在样本量的选取上不少于同类文献（本研究中 N=82）；（2）保证样本本身的质量（即在对被试进行筛选时，通过问卷结合心理学专家的综合评估方法保证被试筛选的准确度，见正文 2.1 被试招募部分），从而保证在此样本量上机器学习的学习效果。

其次，关于本研究中特征量远大于样本量问题，我们将从三个方面来回答这个问题：（1）特征量较多的问题：在机器学习中，特征是对信息的抽取(LeCun, Bengio, & Hinton, 2015)，于是有意义的特征越多带来的有意义的信息就越多。而在我们研究中，我们选取的特征主要来自三方面：Stroop 任务中的条件、重要的脑电成分以及脑电信号的空间位置，这三大方面的特征背后都具有重要的心理学意义（见正文第 2 页最后一段，及 13 页第二段蓝色字体介绍），因此保留这些特征能够提供最大程度提供有意义的信息。（2）较多特征量对于机器学习的影响：特征量显著大于样本量潜在的问题在于当在机器学习中涉及到意义较弱的特征时，机器学习的效果可能会因此减弱，因此部分（非 CNN 算法的）研究会选择在机器学习之前进行特征筛选，通过留下较少的有较强意义的特征保证机器学习的效果(Qiu et al., 2017; Teixeira et al., 2019)。（3）CNN 算法对于特征的自动筛选能力可以达到特征选择的效果。CNN 算法可以通过卷积的操作自动提取并突出更有意义的特征，通过最大池化的操作丢弃意义较弱的特征，从而到达对特征的自动筛选(Giusti, Cires, Masci, & Gambardella, 2013; Krizhevsky, Sutskever, & Hinton, 2012)。因此较多的特征量提供丰富的意义信息，而 CNN 算法从中筛选出重要意义的特征，从而能够在最大程度上保留并准确地利用这些特征及背后的信息。

意见 2：训练和测试，一般建议做多折交叉验证，因为一次的训练测试数据的拆分，有可能会带来随机因素；

回应：我们依照审稿人的建议在新的稿件中采用多折交叉验证，对现有数据重新进行了机器学习，得到了更为稳定的计算结果，提升了模型的质量。非常感谢！具体做法为：我们将原有的方法更改为 k 折交叉验证，即：将样本分为 k 份，这样可以选出 k 种不同的 k-1 份作为训练集，将剩下的一份作为测试集，以训练得到的模型在测试集上的预测结果进行评价指标的计算，最后将 k 次在测试集上预测得到的指标进行平均作为最终的结果。由于样本只有 82 个，为了保证每一折中的测试样本不太少，这里 k 取 3，同时为了对修改后的训练方法进行建模，稿件中 CNN 的架构也有所调整。

意见 3：建议在特征使用方面，多做讨论，目前看主要只是一个机器学习的应用；
回应：我们依照审稿人的建议在新的稿件中增加了特征使用方面的讨论（见正文 13 页第二段蓝色字体部分）。

意见 4：在应用的可行性方面，让用户完成 stroop 任务并且同时戴上脑电仪，与直接完成问卷相比，是否更有优势？因为无论模型如何优化，最理想的结果就是达到问卷的结果。如果问卷的实施难度低于 stroop 任务加脑电，也可能直接问卷测量更合适。
回应：这个问题涉及到研究的核心问题之一，即我们提出的基于事件相关电位（ERPs）和机器学习的考试焦虑诊断方法的应用性。我们将从以下四点来解释我们应用此综合诊断方法的可行性：
（1）单独采用问卷测量的优势：在传统心理学研究中，考试焦虑量表（TAS）具有比较好的信效度(Sarason, 1978)，尤其是在测量群体时，它是一种可靠的区分高低考试焦虑的工具(Serrano Pintado, Delgado Sánchez-Mateos, & Escolar-Llamazares, 2016)，因此我们采用 TAS 量表作为我们综合测评的校标。
（2）单独采用问卷测量的局限性：正如正文中所提到的问卷本身的主观性比较强（见正文第 1 页最后一段，“具体限制在于……”），很容易受到被试主观因素的干扰，这种干扰尤其体现在对个体的诊断上，即个人当前的状态、意愿等会较大影响到对此个体进行考试焦虑程度的评估，使其结果变得不够稳定。而我们的诊断应用的目的在于对每个人达到精准评估，而不是仅限于传统研究中对群体的准确划分，因此才选用更加客观的 ERP 及机器学习方法减少这种不稳定性。
（3）使用脑电（ERP）技术和机器学习的优势：首先，ERP 的测量可以有效减少主观因素的不稳定性（见正文第 2 页第二段，“为了降低这些限制，……”）；其次，机器学习对于数据的深度学习，尤其是采用 CNN 对全脑数据的深度学习可以减少信号的不稳定性对整体判断的影响(Lu, Jiang, & Liu, 2017; Seijdel, Ramakrishnan, Losch, & Scholte, 2016)，从而进一步减少个人因主观因素引起的不稳定性，增加对个体考试焦虑程度的准确度。
（4）脑电结合 Stroop 任务测量相比于单独采用问卷测量的应用可行性：首先，本次研究中的被试质量较高，在使用问卷筛选被试时还进一步结合了两位心理学专家的专业评估（见正文 2.1 被试招募部分），使得我们可以在一批高质量的问卷之中建立模型，从而相比于实际应用中遇到质量不高的问卷时具有更大的准确度优势(Halicek et al., 2017; Vandenberghe et al., 2017)。其次，在实际诊断时，我们最终的目标是筛选出（潜在）考试焦虑的个体，并及时进行积极干预以降低考试焦虑的负面影响，因此个体的准确度在筛选方法中尤其重要。虽然脑电结合 Stroop 任务的方法相对于单独使用问卷较复杂，但是在对个体的考试焦虑判断上具有独特的优势，符合我们的需求。此方法更加客观、稳定、有效，受到主观因素的影响较少，尤其是在遇到被试不想表达自己真实意愿（如不愿意让别人知道自己是考试焦虑，故意不认真作答问卷）以及不适合给被试做主观问卷的场景（如临近考试事件，不宜进行问卷调查以避免文字诱发更高的考试焦虑情绪）时具有更好的效力。因此我们的脑电结合 Stroop 任务的综合筛选方法更具有应用优势。

参考文献：

Giusti, A., Cires, D. C., Masci, J., & Gambardella, L. M. (2013). Fast image scanning with deep max-pooling convolutional neural networks urgen Schmidhuber Lugano , Switzerland, 4034–4038.

Halicek, M., Lu, G., Little, J. V, Xu, W., Patel, M., Griffith, C. C., … Fei, B. (2017). Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *Journal of Biomedical Optics*, *22*(6), 60503.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, 1–9. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436. Retrieved from https://doi.org/10.1038/nature14539

Lu, Y., Jiang, H., & Liu, W. B. T.-I. C. on C. E. and N. (2017). Classification of EEG Signal by STFT-CNN Framework: Identification of Right-/left-hand Motor Imagination in BCI Systems.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, (august), 1–10. https://doi.org/10.1038/nmeth.4402

Sarason, I. G. (1978). The Test Anxiety Scale: Concept and Research. In *Stress and Anxiety* (pp. 193–216).

Seijdel, N., Ramakrishnan, K., Losch, M., & Scholte, S. (2016). Overlap in performance of CNN's, human behavior and EEG classification. *Journal of Vision*, *16*(12), 501.

Serrano Pintado, I., Delgado Sánchez-Mateos, J., & Escolar-Llamazares, M. C. (2016). A Stress Inoculation Program to Cope with Test Anxiety : Differential Efficacy as a Function of Worry or Emotionality. *Avances En Psicología Latinoamericana*, *34*(1), 3–18. https://doi.org/DOI: http://dx.doi.org/10.12804/apl34.1.2016.01

Teixeira, V. H., Pipinikas, C. P., Pennycuick, A., Lee-Six, H., Chandrasekharan, D., Beane, J., … Janes, S. M. (2019). Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nature Medicine*, *25*(3), 517–525. https://doi.org/10.1038/s41591-018-0323-0

Vandenberghe, M. E., Scott, M. L. J., Scorer, P. W., Söderberg, M., Balcerzak, D., & Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, (March), 1–11. https://doi.org/10.1038/srep45938

王艳娜, 孙丙宇. (2017). 基于卷积神经网络的烟瘾渴求脑电分类. 计算机系统应用, *26*(6), 254–258.

高军峰, 王沛, 郑崇勋. (2010). 基于P300和机器学习的测谎方法研究. 西安交通大学学报, *44*(10), 120–124.

……

审稿人 3 意见:

意见 1：本研究的主要创新在于"以卷积神经网络（CNN）算法为主对个体考试焦虑程度进行进一步的学习诊断"，方法和结果部分也主要是围绕 CNN 方法和结果展开，比较偏向"只是研究没有明确心理学问题的算法或技术的工作"，似乎不适合投稿《心理学报》。

回应：首先，本文的核心主题是对考试焦虑这种情绪（和认知）障碍人群的诊断，具体来说是希望通过一个主客观结合的方式提升对个体考试焦虑诊断的精确性，因此从主题上来说这属于心理学研究方法的范畴。其次，本研究的方法在于结合前沿的脑科学技术和人工智能技术对心理学疾病进行综合诊断，是一个跨学科的技术整合，即从多学科视角来解决心理学问题，最终落脚点是对心理疾病（即考试焦虑）诊断方法的探索和创新。因此，我们选择投递《心理学报》。但是最终本文主题是否适合《心理学报》，我们尊重审稿人和编辑部的决定。

意见 2：有两篇正在投稿的文章，与本文数据有重合，有拆分数据发表之嫌，且并未附文章供审核。

回应：我们并没有对同一批数据进行拆分，因为这三篇文章研究的主题不同：

（1） 本研究的主题是结合脑电技术以及机器学习对考试焦虑程度进行综合诊断，属于心理学与人工智能的跨学科研究。本研究涉及到 82 名被试，进行的任务为情绪 Stroop。具体方法为先筛选出高、低考试焦虑者（共计 82 名），之后采用脑电技术采集情绪

Stroop 的数据，并通过机器学习算法对此脑电数据进行学习，最终获得考试焦虑程度的诊断方法，并对此方法的效果及应用进行讨论。

（2） 研究二的主题是考试焦虑者的注意控制能力缺陷，属于心理学研究中的认知神经研究（标题为 ERP Evidence for Inhibitory Control Deficits in Test-anxious Individuals，投递在 *Frontiers in Psychiatry*，在审）。研究二涉及到 46 名被试，进行的任务为情绪 Stroop 和数字 Stroop。具体方法为先筛选出高、低考试焦虑者（共计 46 名），之后采用脑电技术采集情绪 Stroop 和数字 Stroop 的数据，通过比较分析这两种任务下高、低考试焦虑者的注意控制能力受损程度，分析考试焦虑对个体注意控制能力的具体影响。

（3） 研究三的主题是正念冥想训练对考试焦虑者注意控制能力缺陷的干预效果，属于心理学研究中的干预研究（标题为 Effects of mindfulness-based cognitive therapy (MBCT) for test anxiety: Evidence from event-related brain potentials and behaviors，投递在 *Journal of Abnormal Psychology*，在审）。研究三涉及到 60 名被试，所有的被试均完成了两次各类相关的主观问卷，以及两次情绪 Stroop 和数字 Stroop 任务，在两次测试中间，有 20 名被试（训练组）进行了为期 8 周的正念冥想训练，剩下 40 名被试（控制组）不作任何操作。最终，通过比较分析训练组和控制组的差异来分析正念冥想训练是否能有效改善考试焦虑者受损的注意控制能力。

（4） 这三篇研究的主题相差较大：首先，研究二与研究三围绕的主题均为考试焦虑者的注意控制能力缺陷，与本研究（考试焦虑的综合诊断）的研究主题相去甚远，因此这三篇文章并不能整合成一篇文章；其次，研究二与研究三的研究方法、研究流程以及研究范式与本研究也不同；最后，从本研究机器学习的角度，由于提高样本量可以在一定程度上保证机器学习的效果(Halicek et al., 2017; Vandenberghe et al., 2017)，我们使用了部分研究二及研究三中的 Stroop 任务的实验数据（样本量为 22），并同时搜集了本研究中余下需要的 60 名数据。因此，本研究并不是拆分数据的产物，本研究的主要数据也是单独搜集的。若审稿人觉得需要，我们可以把研究二与研究三的文章附上以供内部审核。

意见 3：文章讨论缺乏对相关心理学问题的深入分析，有就事论事之嫌。

回应：由于本研究的主题是探讨结合脑电技术与机器学习对考试焦虑的诊断方法是否可以成为传统考试焦虑诊断方法的一个有效替代，因此我们在初稿中讨论的重点也是此综合诊断方法的有效性与可行性，为了突出此主题我们没有过多地从心理学角度对结果进行解释。但是我们研究的出发点是对情绪障碍的诊断，缺乏心理学角度的深入探讨确实不妥。因此，我们依照审稿人的建议加入了对结果从心理学角度的深入讨论（见正文 13 页第二段蓝色字体部分）。在本研究中，特征的选取是基于心理学意义，所以我们在讨论中加重了对特征意义方面的解释。

意见 4：本研究 ERP 的结果与文献中的发现是否一致，需要说明和讨论。

回应：我们依照审稿人的建议加入了对 ERP 结果的讨论，并且补充了相应文献（见正文 13 页第二段蓝色字体部分）。

意见 5：文章存在一些表达错误，如“主观评估过程中无法避免会一些与病症相关的信息”、“由于高考试焦虑者由于对考试相关威胁信息存在注意偏向”、“将预测的结果与已知的真是类别进行比较并计算两者之间的误差”、“我们认为这个概率可以反应被试个体的考试焦虑程度”等。

回应：我们依照审稿人的建议对文中的表达错误进行了更正（分别见正文第 2 页第一段蓝色字体部分，第 2 页最后一段，第 9 页第一段蓝色字体部分，第 14 页第三段蓝色字体部分）。

意见 6：文章也存在一些文字和标点符号的使用错误，如“被试作出反应”、“有效减少数据的位移，扰动和一些小的变化对数据稳定性和准确性的影响”。
回应：我们依照审稿人的建议对文中的文字和标点符号的错误进行了更正（见正文第 5 页第一段蓝色字体部分，第 3 页第三段蓝色字体部分）。

参考文献：

Halicek, M., Lu, G., Little, J. V, Xu, W., Patel, M., Griffith, C. C., … Fei, B. (2017). Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *Journal of Biomedical Optics*, *22*(6), 60503.

Vandenberghe, M. E., Scott, M. L. J., Scorer, P. W., Söderberg, M., Balcerzak, D., & Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, (March), 1–11. https://doi.org/10.1038/srep45938

第二轮

审稿人 1 意见：

意见 1：（见自检报告审稿人 1 意见 1 及回应）作者虽然回答了 TAS 的局限性和 ERP 及机器学习技术的稳定性，但是仍不能回答所用的方法本质上是通过 CNN 预测 TAS 结果的问题，也就是说，训练 CNN 所使用的反馈是 TAS，这一反馈本身就具有 TAS 本身具有的局限，训练后的 CNN 能够以高准确率预测个体的 TAS，但不能超越 TAS 本身对焦虑测量的局限，缺乏对焦虑更为客观的训练反馈，在理论意义上的推进有限。

回应：我们将从以下几点回答此问题：

首先，为什么选择 TAS 量表：建立机器学习的诊断模型需要一个可靠的客观指标作为效标，TAS 量表可以量化成分数，并且可靠性高、普适性强(Sarason, 1978; Serrano Pintado, Delgado Sánchez-Mateos, & Escolar-Llamazares, 2016)，因此我们选择 TAS 量表作为效标（即机器学习的反馈）；

其次，如何解决 TAS 本身的局限，我们将根据前人文献做出回应（其中，我们对对应的概念进行了相同颜色的标注）：

（1） 机器学习所使用的反馈数据本身存在的局限普遍存在于大部分机器学习的文献中。机器学习是数据驱动的计算，在我们的研究中，反馈数据为 TAS 量表分数，而 TAS 本身的限制（如真实性）会影响反馈数据的质量，从而降低模型效果。而在目前发展较为成熟的对医学疾病进行机器学习诊断研究中，与本研究面临的问题类似，反馈数据一般采用医生的诊断结果，诊断结果的限制（如医生诊断的误差）会影响反馈数据的质量，从而降低模型效果(如：Halicek et al., 2017; Vandenberghe et al., 2017)；

（2） 但是，前人也明确认识到此类局限，并提出了相应的解决方法。即通过提高机器学习反馈数据本身的质量来减小反馈数据本身的局限，提高模型效果 (Lu, Jiang, & Liu, 2017)。如采用专家医生的诊断结果，降低诊断误差，提升反馈数据的质量；

（3） 因此，我们根据前人的解决方法，在本研究中进行了类似的措施。具体而言，TAS 量表涉及的局限（见正文第 2 页第 1 段）主要来自于单独采用 TAS 施测，无法保证其效度。因此，我们通过结合心理学专家的对被试的专业评估（见正文 2.1 被试招

募部分介绍）确保了 TAS 的效度。这样，降低了单独采用 TAS 量表的限制，提升了反馈数据的质量，提升了模型效果。

最后，该研究的理论意义：虽然现在采用机器学习方法深度分析心理疾病的研究逐步发展，但是由于心理状态本身受到个体主观态度等方面的影响极大，使得机器学习对心理数据的分析，尤其是在心理疾病的诊断效果方面不够稳定(Qin et al., 2014)。而本研究则创新地提出了在机器学习对心理疾病的诊断中引入事件相关电位（ERPs）数据，不仅降低了主观态度对心理状态的影响，提高了数据的客观性，还增加了数据的准确度（ERPs 结果可有效反映个体的认知模式，具有诊断的针对性）。目前鲜有文献通过对 ERPs 数据进行机器学习诊断心理疾病，不同的研究范式能够激发不同的心理反应，从而反映不同的认知模式特点，因此本研究能够为其他学者在应用机器学习分析心理数据时提供新颖的研究思路。

意见 2：在实际使用方面（见自检报告审稿人 2 意见 4 及回应）也无法回避需要被试戴着脑电帽做行为实验比之用量表测量焦虑缺乏可行性的问题。

回应：原先脑电实验由于设备及场地的限制，应用到诊断确实可行性较低。但是随着越来越多可靠的移动/便携脑电设备的发展和应用，使得根据脑电数据进行心理疾病诊断的可行性大大增加。为了避免偏离核心主题，我们没有在文中进行叙述，其实我们已经在别的研究中进行了采用便携式脑电设备进行集体施测的尝试，施测时间较短、群体测量可行度较高，信号质量较好，使得采用本研究提出的综合诊断方法的应用性得到保证。

意见 3：多折交叉建议进行更详细的阐述，另外按照文中叙述多折只重复了 3 次，即使是 K=3，也可以以 3 为倍数进行多次重复，取平均值，建议参考相关的已发表的文章重新检验

回应：首先，感谢审稿人的意见，多折交叉验证的方法已经详细补充在正文 2.6 多折交叉验证部分中，并附图加以解释（见正文第 5 页）。

其次，关于具体多折交叉验证的方式，我们根据审稿人的建议整理了相关文献，发现使用 CNN 算法的研究大多不采用多折交叉验证（如： Abdel-Hamid et al., 2014; Cruz-Roa et al., 2017; Targ, Almeida, & Lyman, 2016），少数文献采用单次 k 折交叉验证（如：Cha et al., 2017, k=2）；同时，传统算法的研究中部分采用单次 k 折交叉验证（如：Hahn, Ritchie, & Moore, 2003, k=10; Spetsieris, Dhawan, & Eidelberg, 2010, k=3），部分采用 k 折 k 次交叉验证（如：Pfurtscheller et al., 2000, k=10; Zhou, , Sun, & Li, 2009, k=10）。

最后，由于在本研究中我们同时对包含 CNN 算法的多种机器学习算法进行了计算和比较，需要统一进行数据处理，并且我们的重点为 CNN 算法，因此我们结合 CNN 文献及审稿人的建议采用单次 k 折交叉验证来进行计算。并且，我们根据审稿人的建议，通过查阅前人文献 k 值的计算方法计算出了更为可靠的 k 值（即 k=5，详见正文 2.6 部分），并根据新的 k 值对所有的数据重新进行了机器学习，最新结果已经更新在正文的结果比较中（相见正文 5.2）。

意见 4：建议在几张图片的规范性上多加努力，例如附件中各种方法的 F-1 和准确率的柱状图，就与 APA 柱状图格式相去甚远。

回应：根据审稿人的意见，首先，我们已经将原来展示各个机器学习算法结果的柱状图改成了表格（见正文表 2），并将作为重点的 CNN 算法结果进行了加粗，这样能够呈现更多数据，方便读者进行各个机器学习算法的比较。其次，我们也对其他的图进行了进一步的修正，使之更加规范化。

意见 5：ANOVA 结果过于罗列，可以只讲有显著差异的，用 Fs 和 ps 进行表达。

回应：根据审稿人的意见，已经将 5.1ERP 结果中的 ANOVA 结果进行了整理.

意见 6：文章中存在混用“机器学习”和“人工神经网络”两个概念的问题。
回应：根据审稿人的意见，我们对“机器学习”和“人工神经网络”（CNN）的表述在文中进行了区分，在修改的稿件中，我们用“CNN”专门指代本文的核心算法“卷积神经网络（CNN）”，而用“机器学习”指代文中提及的各类机器学习算法的整体（效果）。

意见 7：对那些不足之处（即审稿人提出的意见），作为本文的 limitation。
回应：感谢审稿人的意见，已根据审稿人的意见在正文末尾处本文的局限部分中进行了补充（见正文 14 页）。

参考文献：

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, *22*(10), 1533-1545.

Cha, K. H., Hadjiiski, L. M., Chan, H. P., Samala, R. K., Cohan, R. H., Caoili, E. M., ... & Weizer, A. Z. (2017, March). Bladder cancer treatment response assessment using deep learning in CT with transfer learning. In *Medical Imaging 2017: Computer-Aided Diagnosis* (Vol. 10134, p. 1013404). International Society for Optics and Photonics.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., ... & Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific reports*, 7, 46450.

Lu, Y. , Jiang, H. , & Liu, W. . (2017). Classification of EEG Signal by STFT-CNN Framework: Identification of Right-/left-hand Motor Imagination in BCI Systems. *International Conference on Computer Engineering & Networks.* 7th International Conference on Computer Engineering and Networks.

Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, 19(3), 376-382.

Halicek, M., Lu, G., Little, J. V, Xu, W., Patel, M., Griffith, C. C., … Fei, B. (2017). Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *Journal of Biomedical Optics*, *22*(6), 60503.

Pfurtscheller, G., Neuper, C., Guger, C., Harkam, W. A. H. W., Ramoser, H., Schlogl, A., ... & Pregenzer, M. A. P. M. (2000). Current trends in Graz brain-computer interface (BCI) research. *IEEE transactions on rehabilitation engineering*, *8*(2), 216-219.

Qin, S., Young, C. B., Duan, X., Chen, T., Supekar, K., & Menon, V. (2014). Amygdala subregional structure and intrinsic functional connectivity predicts individual differences in anxiety during early childhood. *Biological psychiatry*, 75(11), 892-900.

Sarason, I. G. (1978). The Test Anxiety Scale: Concept and Research. In *Stress and Anxiety* (pp. 193–216).

Serrano Pintado, I., Delgado Sánchez-Mateos, J., & Escolar-Llamazares, M. C. (2016). A Stress Inoculation Program to Cope with Test Anxiety : Differential Efficacy as a Function of Worry or Emotionality. *Avances En Psicología Latinoamericana*, *34*(1), 3–18. https://doi.org/DOI: http://dx.doi.org/10.12804/apl34.1.2016.01

Spetsieris, P. G., Dhawan, V., & Eidelberg, D. (2010, August). Three-fold cross-validation of parkinsonian brain patterns. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 2906-2909). IEEE.

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Vandenberghe, M. E., Scott, M. L. J., Scorer, P. W., Söderberg, M., Balcerzak, D., & Barker, C. (2017). Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, (March), 1–11. https://doi.org/10.1038/srep45938

Zhou, Z. H., Sun, Y. Y., & Li, Y. F. (2009, June). Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1249-1256). ACM.