Ref: Submission ID 50eb9302-3e39-4ddc-8c2d-d7f39c7ae5cf

Dear Dr Wang,

Your manuscript entitled "PatchID: A Overfitting Patches Identification Method for Automated Program Repair" has now been assessed. If there are any reviewer comments on your manuscript, please find them below.

Regrettably, the above submission has been rejected for publication in Automated Software Engineering. . Although the reviewers find the topic important, unfortunately there are currently too many issues with the submission to be accepted at this stage. The main problems relate to 1) insufficient related work review, thus novelty of the proposed approach is unclear 2) unclear significance 3) unclear details of the approach and its evaluation, with artefact missing hindering verifiability. Presentation should also be much improved. Please refer to individual reviewers' comments for details. I would like to encourage the authors to take reviewers' comments into account for any future submission.

令人遗憾的是，上述提交文件已被拒绝在《自动化软件工程》上发表。尽管评审人员认为这个主题很重要，但不幸的是，目前提交的文件有太多问题，现阶段无法接受。主要问题与1）相关工作审查不足有关，因此拟议方法的新颖性不明确2）重要性不明确3）方法及其评估的细节不明确，人工制品缺失阻碍了可验证性。演讲也应该有很大的改进。有关详细信息，请参阅个别评审员的意见。我想鼓励作者在未来提交任何作品时考虑评审员的意见

I am sorry that we cannot be more positive on this occasion and hope you will not be deterred from submitting future work to Automated Software Engineering.

很抱歉，在这种情况下，我们再积极不过了，希望您不要被阻止向自动化软件工程提交未来的工作。

Kind regards,

Tim Menzies

Editor

Automated Software Engineering

While we are sorry we cannot publish your work in Automated Software Engineering, our colleagues at the Springer Nature Transfer Desk would be delighted to help you find the right journal for your submission. This is a free service that provides authors with a range of journal suggestions tailored for each article. The corresponding author will receive an email from the Transfer Desk with more information within the next 48 hours. To learn more about the Transfer Desk, please see https://www.springernature.com/gp/authors/transferdesk.

虽然我们很抱歉无法在自动化软件工程中发表您的工作，但我们在施普林格自然转移台的同事很乐意帮助您找到合适的期刊供您提交。这是一项免费服务，为作者提供一系列针对每篇文章量身定制的期刊建议。通讯作者将在接下来的48小时内收到一封来自转账台的电子邮件，其中包含更多信息。要了解有关中转台的更多信息，请参阅https://www.springernature.com/gp/authors/transferdesk.

Reviewer Comments:

Reviewer 1

Paper summary:

The paper proposes PatchID, an approach that identifies overfitting patches for APR. Specifically, PatchID is based on the idea that the dynamic behaviour of the passing tests between the buggy program and the correct patch is the same. By contrast, the dynamic behaviour of the failing tests between them is different. The paper evaluates PatchID on two datasets consisting of 157 Defects4J patches and 380 Java+JML patches. PatchID successfully identifies 63 overfitting patches and 15 correct patches on the first dataset and 169 overfitting patches on the second dataset.

本文提出了PatchID，这是一种为APR识别过拟合补丁的方法。具体来说，PatchID基于这样一种思想，即在有缺陷的程序和正确的补丁之间通过测试的动态行为是相同的。相比之下，它们之间失败测试的动态行为是不同的。本文在由157个Defects4J补丁和380个Java+JML补丁组成的两个数据集上评估了PatchID。PatchID在第一数据集上成功识别了63个过拟合补丁和15个正确补丁，在第二数据集上识别了169个过拟合修补。

Strengths:

- The research idea of the paper is very interesting, relevant to the critical problem of overfitting for APR.

- The paper conducts a large-scale evaluation of the PatchID.

- The paper is overall well-organised.

-本文的研究思路非常有趣，与APR的过拟合这一关键问题有关。

-本文对PatchID进行了大规模的评估。

-这篇论文总体上组织得很好

Weaknesses:

- The approach could be publicly available

- There are missing implementation details

- There are several typos

弱点：

-该方法可以公开使用

-缺少实现详细信息

-有几个打字错误

Significance:

Even though the paper investigates a crucial problem in APR that can save developers' time while evaluating the patches generated by APR,

there is not any discussion on the implications for researhers and practioners. I strongly suggest adding such a section to show the contribution of the paper.

意义：

尽管本文研究了APR中的一个关键问题，该问题可以在评估APR生成的补丁时节省开发人员的时间，

没有任何关于这对研究者和从业者的影响的讨论。我强烈建议增加这样一个部分来展示论文的贡献。

Novelty:

In the related work, Section 2, there is some missing related work regarding existing overfitting approaches [1,2]. Also, there should be a discussion on the delta between existing approaches and this paper.

新颖性：

在第2节的相关工作中，有一些关于现有过拟合方法的相关工作缺失[1，2]。此外，还应讨论现有方法与本文之间的差异。

Soundness:

Overall, the design and the theoretical background of PatchID is presented in detail. However, IMHO, there are missing implementation details of the approach.

How has the approach been developed? What are the tools and environment used?

Furthermore, has the paper used existing patches from previous papers or has rerun the tools? Please clarify in Section 5.1.

Additionally, it should be clear how the patches have been evaluated on their correctness. What criteria have been used? See for instance [4].

In Section 5.3.4, it is mentioned "We manually analyzed all the patches ...". How has this been done? What is the process followed? How many people have analysed the patches and how have you resolved any disagreements?

声音：

总体上，详细介绍了PatchID的设计和理论背景。然而，IMHO缺少该方法的实施细节。

该方法是如何制定的？使用了哪些工具和环境？

此外，该论文是否使用了以前论文中的现有补丁，或者是否重新运行了工具？请在第5.1节中说明。

此外，应该清楚补丁是如何评估其正确性的。使用了哪些标准？参见例如[4]。

在第5.3.4节中，提到“我们手动分析了所有补丁…”。这是如何做到的？接下来的过程是什么？有多少人分析过补丁，您是如何解决任何分歧的？

Verifiability:

Even though the paper conducts a large-scale empirical study to evaluate PatchID, I think there are some missing details.

First, in Section 5.2, why has JAID been selected over tools to build PatchID?

Second, in Table 3, what were the criteria for selecting these APR tools?

In Section 5.3.1, how have the correct paatches been identied?

In Section 5.3.5, it is mentioned that the study uses EvoSuite. How many new tests have been generated?

I would suggest that the study should mention threats to validity (internal, external, reliability)

可验证性：

尽管本文进行了大规模的实证研究来评估PatchID，但我认为仍有一些细节缺失。

首先，在第5.2节中，为什么选择JAID而不是构建PatchID的工具？

其次，在表3中，选择这些APR工具的标准是什么？

在第5.3.1节中，如何识别正确的paatches？

在第5.3.5节中，提到本研究使用EvoSuite。生成了多少个新测试？

我建议研究应该提到对有效性的威胁（内部、外部、可靠性）

Transparency:

Please consider making publicly available PatchID and the dataset used for a possible replication of the study. This would help future researchers use PatchID to create new approaches.

透明度：

请考虑公开PatchID和用于研究可能复制的数据集。这将有助于未来的研究人员使用PatchID来创建新的方法。

Presentation:

The paper is overall well-written and well-organised. However, there are several typos mentioned in the minor issues. Additionally, I would suggest adding a summary for the results of each RQ and a discussion subsection in the Results section.

演示：

这篇论文总体上写得很好，组织得很好。然而，在小问题中提到了几个拼写错误。此外，我建议在“结果”部分添加每个RQ结果的摘要和讨论小节。

Minor issues:

- However facing the large project, such as Defects4J ... -> IMHO, this sentence is unclear. It should be rephrased.

- patches.To sum up, -> There is a missing space before "To" (please fix any other similar issues in the paper.)

- 5-tuple for computing program execution ... -> This sentence does not read well. Please rewrite.

- 5.3 Experimental Result -> 5.3 Experimental Results

- 5.3.1 Result of RQ1 -> 5.3.1 Results of RQ1 (Please also fix all other Result sections.)

- These two reasons together lead to the wrong identification of PatchID. -> What do you mean?

小问题：

-然而，面对大型项目，如Defects4J…->IMHO，这句话不清楚。应该重新措辞。

-补丁。总之，->在“To”之前缺少一个空格（请修复论文中的任何其他类似问题）

-用于计算程序执行的5元组…->这句话读起来不好。请重写。

-5.3实验结果->5.3实验结果

-5.3.1 RQ1结果->5.3.1 RQ1的结果（请同时修复所有其他结果部分。）

-这两个原因共同导致PatchID的错误标识。->什么意思？

References:

[1] H. Ye, J. Gu, M. Martinez, T. Durieux and M. Monperrus, "Automated Classification of Overfitting Patches With Statically Extracted Code Features," in IEEE Transactions on Software Engineering, vol. 48, no. 8, pp. 2920-2938, 1 Aug. 2022, doi: 10.1109/TSE.2021.3071750.

[2] Q. Xin and S. P. Reiss, “Identifying test-suite-overfitted patches through test case generation,” in Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis, 2017, pp. 226–236

[3] Ali Ghanbari. 2020. ObjSim: lightweight automatic patch prioritization via object similarity. In Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2020). Association for Computing Machinery, New York, NY, USA, 541–544. https://doi.org/10.1145/3395363.3404362

[4] Kui Liu, Shangwen Wang, Anil Koyuncu, Kisub Kim, Tegawendé F. Bissyandé, Dongsun Kim, Peng Wu, Jacques Klein, Xiaoguang Mao, and Yves Le Traon. 2020. On the efficiency of test suite based program repair: A Systematic Assessment of 16 Automated Repair Systems for Java Programs. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20). Association for Computing Machinery, New York, NY, USA, 615–627. https://doi.org/10.1145/3377811.3380338

Reviewer 2

# Summary #

In this paper, the authors present PatchID, a tool to classify patches generated by automated repair tools as correct or overfitting and, if overfitting, their sub-types. While the topic of this paper is very relevant, the paper is far from being ready for publication. I recommend the authors use this review to improve their paper for future submission.

评审员2

#摘要#

在本文中，作者介绍了PatchID，这是一种将自动修复工具生成的补丁分类为正确补丁或过拟合补丁的工具，如果过拟合，则分类为其子类型。虽然这篇论文的主题非常相关，但这篇论文还远远没有准备好发表。我建议作者利用这篇综述来改进他们的论文，以便将来提交。

# Strengths #

- Relevant topic (overfitting patch identification).

# Weaknesses #

- The novelty of the work is unclear.

- The review of related work is very limited.

- The evaluation has major problems, i.e., the research questions are not clear, the core information of the datasets is not explained, some results sections present unclear information, and there are no takeaway messages.

- There is no discussion of the results.

- Comparison with existing works is almost non-existent.

- The paper is not well-written.

#弱点#

-这部作品的新颖性尚不清楚。

-对相关工作的审查非常有限。

-该评估存在主要问题，即研究问题不清楚，数据集的核心信息没有得到解释，一些结果部分的信息不清楚，没有外卖信息。

-没有对结果进行讨论。

-与现有作品的比较几乎不存在。

-这篇论文写得不好。

# Detailed comments #

In the abstract, the authors do not present a real problem to be solved. The authors claim that they are proposing an approach to "reduce the number of overfitting patches generated by APR". This is not the problem the authors are trying to solve. I suggest the authors reformulate their motivation to the fact that, because there are so many overfitting patches, developers have to spend a long time analyzing the patches until finding the correct one. What the authors propose is to solve that problem, not to reduce the number of overfitting patches generated by APR.

#详细意见#

在摘要中，作者并没有提出一个真正需要解决的问题。作者声称，他们正在提出一种方法来“减少APR产生的过度拟合补丁的数量”。这不是作者试图解决的问题。我建议作者重新表述他们的动机，因为有太多过拟合的补丁，开发人员必须花很长时间分析补丁，直到找到正确的补丁。作者提出的是解决这个问题，而不是减少APR生成的过拟合补丁的数量。

The extent to which PatchID is novel is unknown. I think the novelty is in classifying overfitting patches into three categories, but that is not claimed in the paper (as the usefulness of doing so). This problem of unclear novelty is also due to the fact the authors do not present a proper literature review of the existing approaches. The related work section is very weak, especially the overfitting patch identification part. The authors should use http://program-repair.org/ and http://bit.ly/2CehUt5 to improve it.

PatchID的新颖程度尚不清楚。我认为新颖之处在于将过拟合补丁分为三类，但这并没有在论文中得到证实（因为这样做很有用）。这种新颖性不明确的问题也是由于作者没有对现有方法进行适当的文献综述。相关工作部分非常薄弱，尤其是过拟合补丁识别部分。作者应使用http://program-repair.org/和http://bit.ly/2CehUt5以改进它。

The flow of the introduction makes it hard to be understood. The flow is overfitting problem -> automatic test generation as a solution -> one existing automatic test generation approach -> patch overfitting categories -> overfitting identification methods -> PatchID. The automatic test generation part is breaking the flow of the introduction. The authors should present the overfitting problem and just after it the existing works on that, not only the ones that are about test generation. Then, they should motivate their choice of a new solution that is based on test generation.

介绍的流程让人很难理解。流程是过拟合问题->自动测试生成作为解决方案->一种现有的自动测试生成方法->补丁过拟合类别->过拟合识别方法->PatchID。自动测试生成部分正在打破介绍的流程。作者应该提出过拟合问题，并在此之后介绍现有的相关工作，而不仅仅是关于测试生成的工作。然后，他们应该激励他们选择基于测试生成的新解决方案。

Unclear aspects of the proposed approach:

- When introducing the first main part of PatchID, the authors mention they want to find "the most suspicious snapshot". However, they do not explain what that means. What does "suspicious" mean in this context?

- In Section 4.4.2, the authors should relate the detection formula with the overfitting patch categories, to explain why the formula was constructed that way.

拟议方法的不明确方面：

-在介绍PatchID的第一个主要部分时，作者提到他们想找到“最可疑的快照”。然而，他们并没有解释这意味着什么。在这种情况下，“可疑”是什么意思？

-在第4.4.2节中，作者应将检测公式与过拟合补丁类别联系起来，以解释为什么以这种方式构建公式。

Research questions:

- When presenting the research questions for the evaluation of PatchID, the authors should elaborate on them. This is necessary for all research questions, especially because some of them are not intuitive. It is not possible to understand the rationale behind RQ3, RQ5, and RQ6. Moreover, RQ1 has two questions, and this should not happen.

研究问题：

-在提出用于评估PatchID的研究问题时，作者应详细说明这些问题。这对于所有的研究问题都是必要的，尤其是因为其中一些问题不是直观的。不可能理解RQ3、RQ5和RQ6背后的基本原理。此外，RQ1有两个问题，这不应该发生。

Comparison with existing tools:

- The authors do not establish which tools they will compare their tool against. Later, in the results section, they present some comparing sentences with Xiong et al.'s work, but they do not explain why they selected that tool and not others. Many recent tools perform overfitting patch identification.

与现有工具的比较：

-作者没有确定他们将与哪些工具进行比较。后来，在结果部分，他们提出了一些与Xiong等人的工作进行比较的句子，但他们没有解释为什么选择该工具而不是其他工具。许多最近的工具执行过拟合补丁识别。

Datasets:

- The datasets used for the evaluation are unclear. The authors first state that they used two datasets: 1) Xiong et al. (2018)'s dataset and 2) Nilizadeh et al.'s dataset. However, they detail Defects4J and mention that "In this paper, six existing repair tools are used to repair the Defects4J dataset, and candidate patches are obtained", implying that they executed repair tools on Defects4J to collect patches, which does not seem to be true.

- The authors explain the repair tools that were used to generate patches, but this is content that adds no information to the paper.

- There is no insight into the actual datasets. To evaluate PatchID, the authors would need datasets that consist of patches classified as overfitting and non-overfitting, and for the overfitting ones, they would also need the overfitting category. However, there is no information about that.

数据集：

-用于评估的数据集尚不清楚。作者首先表示，他们使用了两个数据集：1）Xiong等人（2018）的数据集和2）Nilizadeh等人的数据集。然而，他们详细介绍了Defects4J，并提到“在本文中，使用六个现有的修复工具来修复Defects4J.数据集，并获得了候选补丁”，这意味着他们在Defects4J上执行修复工具来收集补丁，这似乎不是真的。

-作者解释了用于生成补丁的修复工具，但这些内容没有为论文添加任何信息。

-没有对实际数据集的深入了解。为了评估PatchID，作者需要由分类为过拟合和非过拟合的补丁组成的数据集，对于过拟合的，他们还需要过拟合类别。然而，目前还没有关于这方面的信息。

Evaluation results:

- Why does Table 2 exist? The authors only mention it and this is it.

- For RQ1, how do the authors know if the classifications were correct or not? There is no explanation for that in the paper, neither in the datasets section nor in the experiment setup.

- Why is it important to show the results in terms of repair tools and projects? A much more important table is missing, which is with the results of PatchID in comparison to related work.

- It is not possible to follow the section that presents the results of RQ3.

- What is presented as results for RQ5 say nothing about the authors' approach. Maybe the research question itself is not sound, which is not possible to know since there is no elaboration on it.

- The results of RQ6 are more like what is expected for RQ1.

- At the end of each section that presents results, the authors should add a box containing a summary of the findings.

评估结果：

-为什么存在表2？作者只提到了这一点，就这样。

-对于RQ1，作者如何知道分类是否正确？论文中没有对此进行解释，无论是在数据集部分还是在实验设置中。

-为什么在维修工具和项目方面展示结果很重要？缺少了一个更重要的表，与相关工作相比，该表包含PatchID的结果。

-不可能遵循RQ3结果的章节。

-RQ5的结果并没有说明作者的方法。也许这个研究问题本身并不合理，这是不可能知道的，因为没有对此进行详细阐述。

-RQ6的结果更像是对RQ1的预期。

-在介绍结果的每一节的末尾，作者应该添加一个方框，其中包含研究结果的摘要

Just after the presentation of the results end, the paper ends. There is no discussion of the results or real comparison with related work.

结果刚刚结束，论文就结束了。没有对结果进行讨论，也没有与相关工作进行实际比较。

Other comments:

- The authors should explain what "dynamic behavior expressions" is when first mentioning it.

- In both the abstract and introduction, there is no clue about the comparison with related works. After reading the whole paper, I understand why: there is not, really, a comparison, but there should be.

- At the end of the introduction, the authors list their contributions. "5-tuple for computing program execution similarity is proposed" comes from nowhere.

- Algorithm 1 is never referred to in the text. The authors should refer to it. Moreover, when explaining the algorithm at the end of page 9, the authors should refer to lines in the algorithm.

- Throughout the paper, the authors use the term "patch identification" when they want to say "overfitting patch identification". This should be fixed.

- The authors refer to the wrong tables and figures. Examples: "From Table 4, we can find that PatchID works well on Nopol2015" -> In that sentence, the authors want to say Table 3. "Figure 7 that the number of patches with Nf = 1 is the largest" -> In that sentence, the authors want to say Figure 9.

其他意见：

-作者在第一次提到“动态行为表达”时应该解释它是什么。

-在摘要和引言中，都没有与相关作品进行比较的线索。看完整篇论文后，我明白了为什么：真的没有比较，但应该有。

-在引言的最后，作者列出了他们的贡献。“提出了用于计算程序执行相似性的五元组”。

-文本中从未提及算法1。作者应该参考它。此外，在第9页末尾解释算法时，作者应该参考算法中的行。

-在整篇论文中，作者在想说“过度拟合补丁识别”时使用了“补丁识别”一词。这个问题应该解决。

-作者引用了错误的表格和数字。示例：“从表4中，我们可以发现PatchID在Nopol2015上运行良好”->在这句话中，作者想说的是表3。“图7显示Nf=1的补丁数量是最大的”->在这句话中，作者想说的是图9。

Minor:

- The paper is full of typos (e.g., "themis", "identifys", "doe") and ill-formed sentences ("(Xuan et al, 2016).The", "list1.However", "follows: The", "list2, When"). I was noting them down, but I stopped at some point because it is just too much. The authors should review their whole paper.

- Citations should be fixed. There should be a space before the opening parenthesis.

- When mentioning tools for the first time, the authors should provide a citation. For instance, in the following sentence "According to Xin (2017a), the majority of patches generated by GenProg, AE and RSRepair are incorrect.", the authors should provide a reference for AE and RSRepair. The same applies to the sentence "The six APRs are jGenProg, Nopol 2015, Nopol 2017, ACS, HDRepair and jKali, respectively".

- The sentences that start as follows are too long and difficult to follow: "The algorithm first [...]", "Because the test suite is not complete [...]".

- "The patch generated by automatic program repair (APR) needs to be verified" -> This sentence is missing the why.

- "The paper is evaluated on two datasets" -> "PatchID is evaluated on two datasets".

- "their repair performance is still lower" -> "their repair performance is still low"

- "However, automatic test generation tools can only generate test inputs, and the appropriate test outputs still need to be determined manually." -> Provide references.

- "Even so, these approaches still fail to express a complete oracle" -> This claim should be supported.

- "B-Overfitting Patch: The patch that fixes" -> B-Overfitting Patch: The patch fixes"

- "So PatchID identifies a patch based" -> ?

- "Patch is evaluated" -> "PatchID is evaluated"

- "Defecets4j" -> "Defects4J"