Addressal of Reviewer Feedback

Joe Zemmels, Susan VanderPlas, Heike Hofmann

Reviewer Comments #1

The manuscript describes the authors' development of an open-source implementation of the CMC method described by the NIST group. The authors contend that the NIST papers were insufficient for the advancement of science and for the implementation of their methods; however, the fact that the authors were able to create the described implementation disproves this assertion. The NIST CMC people are not listed as co-authors on this paper and thus the reader is left to assume that the authors obtained all the information they needed for CMC implementation from the published CMC papers.

We agree that we have been able to qualitatively reproduce the results of the original implementations. We argue that qualitative reproducibility is not sufficient to justify use in high-stakes applications like forensics. Our implementation, while providing qualitatively similar results to those in published papers on the Fadul and Weller data sets, is based on a large amount of guesswork and lengthy experimentation due to the ambiguity in the original CMC papers. We, for example, have no reason to believe that our implementation will behave the same on different data sets. As such, we assert that we were not able to obtain "all the information [we] needed for the CMC implementation from the published CMC papers." We apologize for not stating this point more clearly in the original submission and hope that clarifying what we mean by "reproducibility" helps make our point.

The authors are a bit harsh on the NIST authors, using phrases like "the comparison process is described with characteristic brevity" and highly criticizing the NIST authors' lack of detail in their papers.

We appreciate and agree with this feedback. Our original language was overly critical of the NIST authors. Our intention was to highlight and discuss the consequences of the lack of details in the original papers, not to attack the authors themselves. We have removed phrases like "characteristic brevity" that may be interpreted as criticism of the NIST authors.

CMC is described by the NIST developers as a family of methods, still under development, and not a single finalized approach (thus the series of CMC papers). Thus it is understandable that NIST papers not spell out a perfect step by step approach. Computational methods, like CMC, evolve and improve over time. Some preprocessing steps that might work for one approach do not work for another approach. Some steps will work for one type of data, or data from a specific type of firearm. For example, different preprocessing steps may be needed for data from different sources. It is therefore not clear why the authors appear confused that different CMC papers would have different processing steps "Complicating the issue of preprocessing the scans, the procedures used to process the surface matrices differ in each of the papers discussed here.".

We agree that authors should not be criticized for publishing methods that are still in-development. Our intended criticism here is not that the CMC pipeline wasn't fully-formed at conception, but rather that there is no discussion of why certain processing choices, especially manual processing choices, were made over others. As a specific example, the parameters and types of Gaussian filters used change between papers. It should be noted that all three of these papers share authors and use the 3D scans of the Fadul et al. (2011) cartridge cases for validation. Based on the changes between papers, it's clear that an internal validation has occurred. For the sake of transparency and reproducibility, we argue that a necessary component of any proposed pattern-matching algorithm is a discussion of the algorithm's sensitivity to different processing choices. Additionally, we argue that it is not acceptable in a

forensics/legal setting to only provide a conceptual description of an algorithm for which there clearly exists an actual implementation (i.e., a "step by step approach").

It is well known that differences in preprocessing, CMC gridding, thresholds, and parameters can affect the final CMC scores.

We agree that the final CMC scores change based on preprocessing, CMC gridding, thresholds, and parameters if one were to compare results across multiple papers. This is the case in any algorithm requiring parameter specification and why it is so important that an implementation is available. However, a discussion of the pipeline's sensitivity has never been discussed in a particular paper. Optimally, a sensitivity analysis in which parameters of the algorithm are individually changed to determine their impact on the final CMC scores would be provided. Instead, multiple changes are made to the pipeline simultaneously from paper to paper. The motivation behind these changes is not discussed. We are not aware of a published sensitivity analysis, but would be interested if the reviewer knows of such a study.

The optimal set of parameters and algorithmic steps may be dataset specific and those searching among these parameters (such as the authors) need to be careful about over-fitting. There did not appear to be anything in the paper about prevention of over-fitting.

We agree that avoiding over-fitting to a particular data set is critical for ensuring the generalizability of the CMC pipeline. However, there is no discussion of over-fitting to the Fadul data in the original CMC papers. Rather, authors appear to perform an internal parameter search on the entire Fadul data set and report the most-promising results. In this manuscript, our goal wasn't to optimize the performance of the CMC pipeline but instead show that our implementation could produce results that are qualitatively similar to published results. The cmcR package can and should be used to explore the problem of over-fitting in future applications.

Some of the NIST criticisms are over-stated. For example, the authors criticize NIST for not describing how scan rotation was performed and which type of interpolation was used during rotation. This is a minor point and not one that should not cause issue with any implementation. It is completely reasonable that the NIST authors left out this detail for including minutia at this level would cause all CMC papers to be unnecessarily long.

We agree that not including minor details in the prosaic description of the algorithm is admissible from a readability perspective. In this revised submission, we have tried to clarify that a conceptual description of an algorithm is incredibly useful for user comprehension, but is not sufficient to ensure reproducibility. It is a fact that rotating the contents of a 2D array by any angle other than a multiple of 90 degrees requires interpolation of, and therefore a change in, the values of the array. This requires the user to choose the algorithm used to perform the interpolation While any one of such decisions *may* not affect results, our purpose in bringing up such minutiae is to emphasize that en-masse, without access to the processed data or code, they compound and ultimately yield opaque, non-reproducible results.

Overall, the largest contribution of this paper is the open-source implementation provided by the authors. The authors were able to assemble a set of algorithmic steps and parameters that fully implement one (or two) of the CMC methods. The modularization described in their implementation is indeed beneficial for those interested in modifying or improving upon the methods. This is indeed a useful contribution to the community. My main criticism is the authors over-stating of the issues in the original CMC papers. As stated above, the fact that the authors were able to produce their implementation demonstrates that sufficient information was available.

As we have hopefully demonstrated in this revision, the fact that we were able to create an implementation does not mean that the original methods are reproducible. We agree that some of our language was overly harsh and misdirected. We have changed the language used in the paper to clarify our point that a conceptual description of an algorithm does not promote reproducibility.

In the conclusion the authors state "The results shared in thi manuscript indicate that the implementation of the CMC method in the cmcR package is qualitatively similar to those published in Song et al. (2014), Tong et al. (2015), Chen et al. (2017), and Song et al. (2018)." but I did not see any of these results? Perhaps there were missing tables in my manuscript?

We greatly appreciate this comment. We've addressed this comment by including Figure 14 showing variance ratio values across the various CMC papers.

One reference I wanted to examine was the Zemmels 2020 reference however it does not appear in the bibliography.

We recognized that this was a typo in the paper. The "Zemmels 2020" reference in Figures 4 and 8 was actually meant to refer to the current manuscript (which we began writing in 2020). We have clarified this by changing references in the paper to "Zemmels et al. (2020)" to "cmcR Implementation." We appreciate the reviewer pointing out this error.

The level of criticism should be toned down to reflect the reality of the CMC method as being a family of approaches still under active development and one that the authors were able to implement. Missing references (Zemmels) and results (comparison with the NIST published results as stated in the conclusion; missing table?) should be added.

We greatly appreciate the considerate comments.

Reviwer Comments #2

1. I don't think this paper works as a paper on reproducible research (RR), nor does it contribute to this field. However it is a good paper and the application is interesting. I would like you to remove the RR focus, and rather concentrate on the motivation for a parallel implementation (freedom of use, ease of extension, research platform, forensic transparency), and the implementation itself.

   We agree that this is not a standard contribution to the field of reproducible research and that the factors listed above are important motivations for our implementation. A critical component of a transparent research platform is results that can be reproduced by the wider scientific community, hence the focus on research reproducibility in the original submission. In this resubmission, we have shifted our focus towards demonstrating why and how the forensics community can promote more rigorous reproducibility standards. Due to the gravity of the application, it is dangerous to have parallel implementations of an algorithm that may lead to different results. We hope that our revised submission more clearly argues the point that the original CMC papers do not provide sufficient detail to yield reproducible results as defined in the manuscript.

2. It would be good to know if you actually tried to contact the original authors – especially since they appear to be NIST researchers, and NIST funds your research

   We appreciate and agree with this comment. Upon reaching out to the NIST authors, they were kind enough to give us access to the results that were used to construct Figure 14 in the revised manuscript.

3. Comparison of your package with published results is a must.

   We appreciate and agree with this point. We've addressed this comment by including Figure 14 showing variance ratio values across the various CMC papers.

We appreciate the constructive feedback from this reviewer.