Review of: "FLAML-Boosted XGBoost Model for Autism Diagnosis: A Comprehensive Performance Evaluation"

Friedemann Uschner[1]

[1] Technische Universität Dresden

Potential competing interests: No potential competing interests to declare.

The authors aim to predict autism spectrum disorder using an automatic optimization of an XGBoost model FLAML AutoML. On the employed dataset, a low error is achieved, indicating good classification results.

The paper is written fairly straightforward, nevertheless it misses scientific merit and has major problems in both writing and analysis. I will briefly summarize the most urgent and major points that need addressing. A scientific publication for the article at hand in its current state is not advisable.

Major concerns:

- The article has sections included that point to it being written with ChatGPT or another LLM. I refer to sections like, quote from the article: „By providing a rationale for the preprocessing steps, explaining the technique employed, and clarifying the purpose of dataset splitting, we offer a more comprehensive and informative description of the preprocessing stage. This enhanced version helps readers understand the reasoning behind the chosen techniques and the importance of each step in preparing the dataset for model development and evaluation." and even more obvious „In this improved version, the text is rephrased to enhance readability and maintain a consistent writing style. The use of "Furthermore" strengthens the connection between FLAML's ability to handle imbalanced classification tasks and the improved model's performance. The addition of the sentence regarding interpretability highlights the value of the combined FLAML-XGBoost approach in providing insights into the underlying features influencing the classification process."
  While using ChatGPT to help improve language (especially for non-native speakers), that fact that these sections have been left included in the upload is a concerning „oversight" to say the least. At minimum it bringst to question, if the rest of the article is original content.
- The brief and superficial workflow follows a published Kaggle contribution very closely. This is very concerning in the first place, but it especially emphasizes, that a worklflow like presented does not yet warrant a scientific publication, but would be better suited in a framework like Kaggle. As for the question of reproduction and copyright, I leave this for others to decide.
- In fact, the dataset itself has never been thoroughly cited or described. By the short section naming the column headers of the data, it can be assumed that this is the data in question https://www.kaggle.com/datasets/faizunnabi/autism-screening or https://www.kaggle.com/datasets/konikarani/autismdiagnosis , but the author never provided the much needed citation and attribution to his source, so it remains unclear where the data comes from. This can be either a

gross oversight or mal-intended practice or owed to the alleged use of ChatGPT. Please provide information on that.

- The citations are not informative or meaningful in the places they have been used. Already citations from the beginning of the article can be taken as examples for this problem: In the second sentence of the Introduction, Chawla et al. (2002) is cited in context of the class imbalance problem and development for models on that. Nevertheless, the work never even mentions any connection to imbalances. It is not clear what this citation is ment to show the reader. This is not a paper that should be cited here. The rest of the introduction has no mentionable citations. The second example can be found in the second paragraph: I fail to see, how the publication by Lord et al. (2000) contains any information on the context that is has been cited in. These examples are not the exception. Please elaborate on your choice of citations. I am fairly sceptic, because this is actually something that an LLM would do and after the earlier mentioned concerns, I am not convinced that this work is original. To me this is a red flag and it should be to anyone else who has read the paper carefully and is reviewing it.
- Oversampling: The article wants to comprehensively tackle the imbalanced nature inherent in the dataset. Nevertheless, it is using a rudimentary Oversampling technique without motivating its use or analysing its influence. The workflow is shallow and no scientific question regarding the use of Oversampling for this general problem is answered or even asked. No potential alternatives have been mentioned or tested, no gold standard (no oversampling) has been compared. The oversampling in such a case is in the opinion of many plain wrong. The author mentions in his „analysis“ that further „analysis and refinement may be necessary to enhance the model's sensitivity and reduce the number of false negatives“, but fails to mention that these false negatives are a bias introduced by the oversampling of the routine. This analysis and refinement should be the part of the article.
- Train/Test split and evaluation: The algorithm splits the dataset before applying the oversampling. This means that either the test-set has not been used for evaluation or the dataset was never imbalanced in the first place. The confusion matrix shows, that the test data is in fact balanced with 107 positive and 109 negative samples (the numbers are barely readable in the plot!). This problem further warrants explanation, since the only error that is mentioned is „0.0077“ (without any mentioning of the used error function to make it interpretable) and it seemingly is taken at training time at a certain timing, which means that it is either using the test-set to evaluate (and stop) training or the test-set has never been evaluated. In both cases, this is concerning and wrong. Please elaborate and clarify.
- ROC and calibration curve are merely shown, but never (as promised in the section) actually analysed. The „analysis“ of the confusion matrix is only a statement of what numbers appear in it. This is not an analysis!
- Comparison to other works. The author never puts the results into perspective of other workflows. This way, a novel finding can not be claimed and even calling an error „low“ is merely an unproven statement. Is it lower than others? And if so, why? This would be interesting to know.

I want to stop at this point. I have more major and some minor concerns, but it is already apparent that this work should be improved upon massively. In fact, I believe that the content as it is at the moment is not warranting the time invested by the many reviewers. The platform will additionally need to evaluate, if it is original and how it will deal with it, should it not be. But even so, it is only a superficial coding exercise and would better fit a format like Kaggle. In fact, that’s where close works to this have been published already, as mentioned.