**Response to the Reviewers**

**Paper No. P022:** Automating Hyperparameter Optimization for Deep Learning Applications

We would like to thank the reviewers and the technical committee for their invaluable insights, constructive criticism, and expert suggestions on our paper. We incorporated many of the suggestions and corrected the issues they pointed out to improve the paper.

**Reviewer #1:**

1. First, some easy (but necessary) corrections. There is a minor error in the notation – the first instance of <image002.png> uses a different “X” symbol than the rest of the paper.  In addition, several of the figure labels need to be corrected. It’s clear from the text that the figure at the bottom of the first column of page 3 is intended to be “Figure 2,” not “Figure 4.” The labels for the other two figures on page 3 are similarly incorrect due to the automatic numbering of the figures.

*Response: Thank you very much for catching these issues. We corrected these in the updated version.*

1. Second, I would suggest spending a little more time explaining the data in Table 3. You spend only a paragraph discussing these results, and only do so very generally. This is your novel contribution to the field – these results are the most important part of the paper, because they are what supports your conclusion. You should spend more time discussing each of the different data sets and what they mean. I am only passingly familiar with the paper’s topic, but at no point is it explained why three particular cells in the table are shaded. I can see that those MLP-M all have the lowest % error rate for the Eval set (except for the BER row, which adds some confusion while interpreting the table, so maybe it’s worth setting that apart somehow), and can sync that up with your statement about the manual models overperforming, but that’s something you want to draw attention to in the text as well to make sure the reader sees that result!

*Response: We agree with the reviewer and have incorporated changes to address the issues raised while staying within the page limit of the conference.*

1. A table discussing the time to reach optimal solution may also be worth including. You state that they all took 1-3 hours, but the difference between 1 hour and 3 hours is significant. If there’s a consistent speed gain by using a particular method, that’s worth reporting.

*Response: We modified the text to address this.*

1. Finally, three of the figures need to use larger font sizes. The axis labels and titles for the synthetic (Figure 4, which should be Figure 2), EEG data (Figure 3, which should be Figure 4), and decision surfaces (Figure 5) are too small. In the last two cases, they’re so small as to be incomprehensible to the reader. In general, fonts in a figure should be the same size or one font step smaller than the font size of the text (so for a 12-point font paper, the text in any figures should be at minimum 10 point). If that makes your figure unreadable, then it is too small and should take up more real estate on the page.

For example, Figure 4 (2) could span both columns so that it’s more readable, or the three images could be stacked vertically so that each can be wider such that the axis labels and titles can be larger. Figure 3 (4) may need to be changed in a similar fashion.

Basically, you should \***never**\* simply copy/paste an image from another source and resize it in your paper. It should either stay the same size, or you should remake the figure in another piece of software to make the text/axes/etc. readable at the smaller size you desire.

*Response: We edited figures 3 and 5 to make the axes more legible. Figure 2 could not be changed much as there is an issue with the page limit. We also would like to assure the reviewer that all the figures were generated by us for this paper.*

**Reviewer #2:**

1. - Please include a methodology section and write down the steps that you have done in your research. - Synthetic datasets have a limited number of features. Therefore, I think including more features will enhance the paper. Or you can show the comparisons with the benchmark datasets.

*Response: We added some more information in the Experimental Design section. Our future work will include more diverse datasets.*

1. - In Table 1, the number starts from 08. I think you can start from 01 through 05. - In Table 3, for data sets # 11 and #12, the best results need to be selected as BOLD.

*Response: We added an explanation for the dataset numbering in the paper. Sets #01-05 already exist for ECE 8527: Introduction to Machine Learning and Pattern Recognition course.*

*Thank you very much for catching the issue with Table 3. We highlighted the results and mentioned this in the text.*

1. - Adam optimizer was selected for synthetic datasets. I am curious to know if the selection is random. If so, the authors can try other optimizers or analyzers to select the best optimizer automatically.

*Response: Adam and SGD are two of the most effective optimizers, and we decided to limit some of the hyperparameters by picking them beforehand. We will incorporate an automatic selection of the optimizer in our future work.*

1. - Few typos need to be corrected. i.e., in the abstract (line 5, Line 17-18), there is unnecessary space. Also, in conclusion: line 4. Please check the similar typos in the entire paper.

*Response: We revised the paper to fix the grammatical issues and typos. However, this specific issue was caused by MS Word’s justified text option.*

**Reviewer #3:**

1. Please change the title to "On the Automation of Hyperparameter Optimization for Deep Learning Applications". The current title is misleading. It implies that you are proposing a new method for parameters tuning, which is not the case. You are discussing and comparing autotuning vs manual tuning. The title should clearly reflect this.

*Response: Thank you very much for your suggestion. We modified the title to reflect this.*

In the abstract:

1. Authors repeatedly use sentences like "auto-tuning produces errors that are tedious to  
   resolve for those with limited experience in machine learning." "Auto-tuning tools are excellent for creating baseline models on new datasets, but they need more attention to formulate optimal solutions for end-users with less background in deep learning."  
   These sentences need to be removed or reformulated. Who cares about the problems met by inexperienced users? It is natural that you should not use a tool in a real-life application unless you are experienced with it.

*Response: We added a sentence in the abstract to address this issue. Engineering tools are created so that everyone can benefit from them without diving deep into the intricacies posed by the inner mechanism. Autotuning tools can be categorized as such.*

1. The sentence "Because of this, manual tuning based on domain knowledge and experience is still preferred in machine learning because it produces better performance, even though it requires extensive machine learning expertise." should be either removed or reformulated. It is quite misleading. Using experience and domain knowledge to tune neural networks parameters is only possible for a small class of problems, where you can map these parameters to quantities related to the problem that has physical tangible meaning. And this is only possible for small-sized neural nets (limited number of layers/neurons per layer).

*Response: We would like to clarify that the paper discusses hyperparameters of deep learning algorithms. The parameters of a model were not within the scope of this paper.*

In the Introduction:

1. The sentence "This has led to a widespread perception that what used to be known as knowledge engineering is no longer needed because the entire technology development process can be automated using inexperienced technologists." Please add a reference to justify such a strong statement OR replace "widespread" by less strong word like "This led some researchers to believe"

*Response: We added a few contemporary references to bolster our view on this issue.*

1. The sentence "since it requires a detailed understanding of the algorithms." should be rephrased. You should not use an algorithm unless you understand it. Once again, we should not concern ourselves with people using tools that they do not understand. This is not right.

*Response: As mentioned before, we addressed this issue in the abstract and conclusion.*

In section IV "Experimental Design"

1. The sentence "For manual tuning, we preprocessed the data using a bandpass filter in the  
   range 0.5 − 35 Hz and then reduced the sampling rate to 50 Hz. We then applied framing and windowing to extract 0.1-second frames with 0.3-second windows." Does this imply you perform this preprocessing only for manual tuning case? Why not use it with auto-tuning, too? I don't believe the comparison is fair when you pre-process in cases and not others. Preprocessing of inputs can be used with autotuning, too.

*Response: Thank you very much for catching it. We added the information to clarify that both autotuned and manually tuned models were fed the same preprocessed data.*

In short, I believe the paper requires major revisions and considerable rewriting:

1. a. It should be stated clearly that the purpose of the paper is showing that manual tuning outperforms autotuning especially with real datasets and omit all statements describing the problems of users who have no experience or do not understand algorithms. If you want to make a point of this, you should state it as "**Many researchers from different fields, who are inexperienced with machine learning, think they can use deep neural nets as black boxes using model auto-tuning tools. However, this paper shows that this is not guaranteed for real datasets and that experts' knowledge is still needed to obtain acceptable results and achieve model reliability**".

*Response: We believe that our abstract and conclusion illustrate this exact point.*

1. b. You should emphasize in results discussion what aspects of experts' knowledge helped in parameters tuning and why did the autotuning tools fail to capture these aspects (why wasn't it possible to merge some of experts' preprocessing with autotuning to make the task easier for auto-tuning tools? In other words, use experts’ knowledge to prepare and choose a suitable form of data inputs/features and then use auto-tuning tools to do the rest of the job). Unless, this is clearly discussed in the paper, it may appear to the reader that autotuning was not used efficiently by the author.

*Response: Both the manual tuned and autotuned models began with the same preprocessed data. We decided to only work with a few of the hyperparameters for both of the tuning mechanisms (such as MLP/CNN/LSTM layers, dropout, activation functions, loss functions, etc.) and ignored some hyperparameters (such as regularizes, learning rate scheduler, etc.) so that there could be a fair comparison. We also intentionally ignored dropout for one experiment for set #10 as reported, where we showed that the auto-tuner overfitted to the dev set.*