Replies for The Report:

-This paper describes a study of multi-class classification on skewed datasets. The example chosen is the Upsilon states using CMS open data and the e+e- final state. The methods are described clearly, but from what is written about the analysis and results, I cannot tell at all what is being done.

>ANSWER: The aim of this article was depicted in the abstract (lines between 4-11), introduction (lines 5-7), and conclusion (lines 2-5) sections. In summary, what is being done is to analyze the performance of machine learning techniques for highly skewed particle physics datasets. Upsilon states were selected as an example.

-As best I can tell from the text:

* The data are CMS open data for e+e- in the 8-12 GeV mass range.

* e+e- pairs within 0.05 GeV in mass of a known (PDG) Upsilon mass are labeled as signal of that Upsilon type, with everything else considered background.

This is how I read Sec 3.1; however, most candidates in any mass range in this data are clearly background, regardless of how close they are to an Upsilon mass. (The probability of being signal is higher near a known Upsilon mass, but it is never high.) Is this really what is being done? If so, why?

>ANSWER: In the analysis, the Upsilon data set that is specifically released for the Upsilon study from CMS [97] and the dielectron data set released for the dielectron study from CMS [96] were merged and used to make the dataset highly imbalanced as in particle analysis. Because in Upsilon analysis the states are not abundant meaning identification of states and background rejection is a challenge. Since the cross-section for bottomonium production is small, the experimental study of Y production is difficult [26]. Therefore, by merging the Upsilon dataset with the dielectron dataset the machine learning study became more realistic. Due to its nature, Upsilon particles are rarely produced in collisions. In the analysis most of the e+e- pairs are from the other sources resulting in suppression of the signal from a broad background.

Most of the real-world data is highly imbalanced like in particle physics. Therefore, it is aimed to develop machine learning techniques to extract rare states (Upsilon states in this case) that are suppressed by the background without considering systematics. CMS Collaboration released these datasets to the public also for software development. Therefore, the data is used to improve a machine learning method identifying multi states at once even if the signal is very low. Upsilon states were used as an example for this purpose. This paper is the first paper implementation of the multiclassification for the imbalanced dataset in particle identification especially for the particles few in the dataset.

-The choice of +-0.05 GeV is not motivated. Looking at a few CMS plots I’d estimate the mass resolution to be ~0.1-0.2 GeV. (This is a secondary issue, but still not explained.)

>ANSWER: In classic analysis, particle mass distribution has still residual background even if subtracting the background from raw mass distribution. This residual background is estimated by considering the sidebands near the particle signal distribution resulting in the selection of a broad mass spectrum range [35]. This is why in those studies the mass range selected as ~0.1-0.2 GeV/c2. With the machine learning approach since only the signal of the particle is identified so the spectrum range does not
need to be considered. These states have very narrow widths [33,34]. Therefore, in the analysis for each state particle signal was defined according to their PDG value which can be fitted by their signal fit function. The +/-0.05 GeV/c^2 signal mass range is selected according to their PDG and also considering references [29,30] in which multi Upsilon states were estimated.

- * Classifiers are trained using the various methods to identify the 4 Upsilon states and background. Given what I said above, that it reads like anything near an Upsilon is labeled Upsilon, I can make a perfect classifier by hand (namely a mass cut). The trained classifiers are not perfect even though mass is, in fact, included as an input feature. (That could just be due to limited sample sizes; however, one could ask what is the point of training a classifier that is not as good as a simple by-hand cut.)

> ANSWER: In the analysis machine learning method implementation was carried out in parallel with the classic particle analysis. In traditional particle analysis mass cuts were used as veto cuts already which can result in loss of signal. In this analysis, mass was used as a feature to estimate signal and background. Therefore, with good training of the classifier, the signals can be identified in high performance. As it can be seen from the article the machine learning algorithm was succeeded around 90% with high sensitivity.

In addition, as explained in Section 1.1 the classic analysis is composed of several steps including the implementation of veto cuts, estimation of raw background, and normalization processes. Alternatively, in machine learning, those steps were not used and the signals can be identified. The article proves that implementing artificial samples during the training process enhances the success of the algorithm even if the signal is very low. In the traditional case, these signals cannot be extracted. So, with this study, it is illustrated that the particle signal can be obtained even if from the bulk background.

- * No mass plots are provided showing the spectra classified as a given type. This makes it impossible to really assess what is happening.

> ANSWER: It is understood from Figure 1 (from the right side) dataset is not large enough (each state is less than 2% of the whole data) for showing mass plots after train/test set selection for the machine learning approach. As a result, they were not demonstrated. As it was stated in the article, ML multiclassification approach based on RFC was used to search whether it is an alternative way for determining the invariant mass spectrum of the states which can be a road map to extract rare states from a broad background. It was stated in the Conclusion part that is the results are promising the usage of artificial new samples to identify rare particles in the bulk dataset.

- There is some discussion early on about the "traditional" approach to such an analysis, estimating backgrounds using same-sign or other samples, extracting signal yields by fits to the mass spectrum using CB functions, etc. None of this is mentioned in Section 4 for how this data is analyzed. Given that the mass itself is used in the classifier, clearly classification will bias the mass spectrum making it unusable.

> ANSWER: In Section 1.1 traditional particle analysis was explained to show how much human efforts are needed with a high level of computational and physics knowledge. The data analysis with machine
learning technique was explained in Section 3 clarifying also model tuning in detail. In traditional analysis, mass is used as a cut parameter to define a spectrum. So, in this research, it is also used as a feature that is parallel to the classic way. If the classification is biased due to the mass information, then the models would not suffer to identify the states. Mass property is one of the discriminative features in the models. Also, as in Ref [7-9], mass information was used in a similar way.

As stated in the Abstract, end of the Introduction part, and Conclusion, this paper is a guide for the application of multiclassification in particle dataset and using artificial samples in training for identifying rare states by machine learning which is hard to be found by traditional techniques. In this study, the performance of different widely used machine learning techniques was demonstrated. If the dataset is large enough the identified signals from machine learning implementation can be used to extract mass, width, and yield information of the states to get the physics results.

-It is entirely possible that my understanding of what is being done here, as described above, is completely wrong, but I reread this twice and I am convinced that what is written above is a reasonable interpretation of what is written in the paper, and that there are no hints in the text of several alternative paths that would make more sense. It is also possible that the example is not meant to be realistic or relevant for an actual particle/nuclear-physic analysis, but that is not how this paper seems to be sold.

>ANSWER: The purpose of the paper is to implement machine learning techniques to handle highly imbalanced particle physics data and present their performance. As case study Upsilon states were selected. In classic analysis to have the rare particle signals a large number of events are required to overcome the background which may result in loss of the signal. With this study, it is revealed that even if the particle signal is very few, it can be identified by machine learning techniques without any background estimation. This implies that with machine learning applications background estimation methods, spectrum range selection, normalization processes will not be required even if for multi-particle states and for highly skewed data structures.

It is believed that the paper has enough evidence including detailed information about how the models implemented and hyperparameters were tuned. The dataset is enough for software development as stated in CMS open-source documents [95-97]. This paper is also a new example for multiclassification of high imbalanced particle physics data. Recently similar studies have been started and published as Ref [7-9].

-Therefore, I conclude that at a minimum a substantial revision is required to better explain what is being done.

>ANSWER: This article is the first attempt of the application of RFC-based models to highly skewed particle data set for multiclassification. In addition success of the models was evaluated by the metrics specifically developed for understanding the performance of the models for highly skewed datasets. For these reasons it is believed that this study is an original research article and brings a new aspect on implementation of machine learning techniques in particle physics analysis.