Referee #2 – Anonymous Referee

Comment 1

The abstract refers multiple times to the “two datasets” but it is never stated, in the abstract, what those two datasets are. I would recommend adding that information (i.e. mixed PSLs and a mixture of laboratory generated bioaerosol) somewhere in the abstract. Also, it seems like it would be worthwhile to include something in the abstract about the findings regarding the utility of additional data like the shape information.

We updated the abstract to include the suggestions provided by the reviewer.

Comment 2

One p4, line 31, you say you have “used the z-score to standardize the data” but I don’t really understand what that means. Does that mean that each channel in each dataset has been individually standardized using the z-score? Why does that make sense in populations of laboratory particles for which they need not be (and likely aren’t) normally distributed? Also, was this only done for the HCA case or was this done for all of the data?

We agree that this information is not very clear and therefore have added a paragraph at the start of the Section 2 in the updated paper that details this information.

> *“For some of the methods it is necessary to standardise the data before conducting the analysis. This is the case for Cluster Analysis (2.1), Support Vector Machines (2.5) and Neural Networks (2.6). The purpose of standardisation is to consider each of the variables with equal weight. For example the fluorescent measurements are much larger than the size measurements and for the aforementioned methods this would cause the fluorescent measurements to have more of an influence on the classification than the other variables which in turn leads to a significant drop in performance. For Decision Trees and ensemble methods (2.2) we do not standardise as each of the variables are considered in isolation so standardisation is not necessary. For the Gaussian methods (2.3), standardisation is conducted implicitly when the models are fitted so standardisation again is not necessary. For K-Nearest Neighbours (2.4), standardisation is usually recommended but with our initial tests we found that it hindered performance for our data, so for our results this method are produced from unstandardised data. In order to standardise the data we apply the z-score to each of the variables since this is the method of standardisation that is most commonly used in the literature (e.g. Crawford 2015).“*

The referee is correct that each variable in the dataset has been individually standardised by using the z-score. This is not done with any implicit assumption on the distribution of the data, but instead is done to give each of the variables equal weight. We have now detailed which methods we have used standardisation for and where we haven't have given an explanation as to why it was not necessary.

Comment 3

On pg 6, lines 28-30, this is not the most informative example as it is a value that has a much higher probability of being assigned to a particular category. What if the value was 0 instead? Would it then classify the particle as a blue particle because that is the most likely answer even though the particle is nearly as likely to be a red or a green?

We agree with the referee that this is not the best example so we have changed the example given to the one suggested. It would indeed classify the particle as blue since it is the most likely answer even though the particle is nearly as likely to be red or green.

Comment 4

Section 3, is there no reference specific to the MBS?

There is no reference for the MBS since this is the first paper on the MBS, those who built the instrument have written the instrumentation section and have been included as authors on the paper.

Comment 5

For Figure 5, please add error bars in the y-direction to show the variability of how a particular kind of well-known particle appears to the instrument. Also, it would be interesting to see size distributions, just as a metric for how the instrument does there.

We have made an attempt to provide error bars on such a plot, but the variation in the data is simply too large to provide a graph that would be legible. We could alternatively provide size and fluorescence histograms, but this would involve creating many more graphs which would be too cumbersome for a manuscript. We do nonetheless agree that the size information is of value and have hence added this to the plots. We also, now mention the inclusion of this information in the text.

Comment 6

For figure 6, so are you lumping all bacteria, fungi and pollens together for each of these traces? I would think there could be some important differences. For example, between washed and unwashed bacterial populations, bacterial vs fungal spores as well as different pollen. Are the bacterial spores lumped in with the bacteria or the spores? This graph needs much more explanation and the laboratory methods used to generate it probably also deserve a bit more discussion. Also, again, vertical error bars and an accompanying size distribution would be informative.

We realise we have not been clear in our production of this graph, but the paper mulberry samples were highly fluorescent compared to the other samples so have not been included in the graph because the prevent the reader from seeing any differences between the groups (other than the paper mulberry is very fluorescent and everything else is not as fluorescent). As the referee has suggested we have split the groups into their subspecies samples to show the differences between the aspen pollen and poplar pollen and the bacterial samples. As is the case for Comment 5 we are also unable to reasonably include error bars due to the high variability in the data.

Comment 7

Pg 13, line 12: The authors state that the division of data into training and testing sets has been described in section 2 but I can't find that text.

We have not been precise in Section 2 on the division of the data so we have added a more detailed paragraph at the beginning of this section.

> *"We split the data using 5-fold cross validation. Here the data is randomly split into five groups. We then progressively take each group to be the test set and use the remaining four groups to train each of the methods and then record the percentage of the test group that was correctly classified. Finally we average our results over the five tests."*

Comment 8

For table 3, what are the units of "Performance"? Is this the percentage of particles that are correctly identified? For the lab data is that identification as a broad class (bacteria, fungi and pollen) or is it on a sample-by-sample basis? It seems like that number is pretty dependent on the kinds of particles you used for your laboratory component. For example, you only have one sample of spores and they are a small fraction of the total data (by number) such that if they were getting lumped in with bacteria or pollen it wouldn't make the overall fraction of correctly identified particles that different. Similarly, do you classify all the non-fluorescent particles correctly (by number nearly half of your laboratory particles are non-fluorescent if I read your methods correctly) but then get the various classes of bioparticles confused with each other? It would be interesting if you could break down the correct identifications as a function of particle type. This would be complicated to show for all of the methods tested but, perhaps just for the top performers or something. This speaks to the potential of translating these methods to more diverse data sets.

The referee is correct the units of performance are the percentage of particles correctly classified. We have added this into the caption for the table to make it clearer. Identification is conducted on broad class basis initially but we have now updated the manuscript with further analysis using the Gradient Boosting Classifier on a sample by sample basis.

The referee makes a very good point regarding the lack of fungal spores in our analysis and we do indeed see that a significant proportion of the fungal spores were being misclassified. We therefore have placed two samples of puffballs that were originally removed from the analysis due to their lacking fluorescence. This has reduced the overall performance for the dataset across all methods as a large proportion of the fungal spores are being misclassified as non-fluorescent, but for our best performing methods, we now see a good proportion of the fungal spores which are classified as fluorescent, being correctly classified as fungal spores, which is reflected in our break down in Section 5.2.2

The way in which the results are presented may be misleading so we have broken down the errors for all the methods which are now presented in the form of figures. We have made minor updates in the text to refer to the new figures instead of the old table. In the Laboratory data we produce a plot that includes the misclassification error due to fluorescent material being misclassified as non-fluorescent, the error due to non-fluorescent material being misclassified as fluorescent and finally the error between the fluorescent classes. Unfortunately a large proportion of the error is due to fluorescent material being misclassified as non-fluorescent, but we do see that of the material that is classified as fluorescent, a good proportion is attributed to the correct broad biological class (bacteria, fungal spores or pollen).

We have also updated Table 2 to reflect the addition of the extra fungal spore samples. In addition to this we noticed that Table 1 is incorrect. The figures given were for an initial analysis, earlier in the construction of the paper; we have now updated the table with the correct number of particles for which the analysis has taken place.

Comment 9

Also for table 3, there are several things that don't make sense to me and which seem to be inconsistent with the text. First, for all of the HCA cases, it looks like the algorithm does worse for PSLs than for the lab data. Is this a type-o or are the columns mislabeled? Also, the text says that omitting the shape information leads to improved identification for PSLs but degrades identification for the laboratory data whereas that is not how I interpret the table unless the authors are only referring to the HCA cases. For 6 of the 10 supervised methods the reduced data leads to better identification of laboratory data unless, again, columns are mislabeled.

It is correct for all of the HCA cases, that the algorithm does worse for the mixed PSLs compared to the laboratory generated aerosol in the case of the full dataset. In both cases the method performs poorly but in the case of the laboratory data we are able to aid classification by placing the non-fluorescent material and the saturated material into separate clusters prior to analysis, which is why we see better performance for the laboratory aerosol. In general the reader should conclude, however, that the method is ineffective for large dimensional data.

This information is presented in the following paragraphs which replace (or update) the current paragraphs on HCA.

> *"In Figure 2 we see that the dye-doped PSLs should be highly separable by eye whereas in Figure 3 it appears the laboratory data would present more of a challenge to the different algorithms. This is demonstrated also in our results where the percentage of data correctly classified for the laboratory data is in general much lower than that of the PSLs.*
>
> *The exception to this is with HCA. For the Full data the algorithm performs relatively poorly for both the PSLs and the Laboratory generated aerosol. However, since we have already placed the non-fluorescent material and saturated material into group on their own prior to analysis for the laboratory generated aerosol we see better performance for the laboratory data compared to the PSLs in this case. For the reduced data however we can yield generally*

> *good performance for the PSLs using the Ward linkage, but for the laboratory data the performance in generally poorer compared with the supervised methods regardless of whether the shape information is included."*

We agree with the referee that our text that says that omitting the shape information leads to improved identification for the PSLs but degrades identification for the laboratory data is not consistent with our results and have hence removed this text. The problem regarding the inclusion of the higher dimensions is a topic we feel that needed to be covered in more detail so we have replaced the paragraph starting *"The problem with the dimensionality is also clear…"* with the following paragraph that details the issue more clearly.

> *"Some of the methods, in particular the Cluster Analysis and the Quadratic Discriminant Analysis perform poorer or equally as poor when the shape information is included.*
> *This is a reflection of the methods' ability to utilise large dimensional data effectively rather on the instrument itself. In the case of HCA we would suggest that further research need to be conducted in order to reduce the dimensionality of the data without losing information from the shape channels before using this method for the MBS. For Quadratic discriminant analysis, the difficulty is in approximating the covariance matrix when the number of samples is less or similar to the number of dimensions. We would hence expect this method to perform better as larger samples are collected."*

Comment 10

The other thing that would be helpful in interpreting Table 3 would be if you could label the different methods according to the section headings from section 2.

This is a great idea. Since we no longer have space for the full names of the methods in the figures, we have created a separate table that details this information. We have also reorganised the results so they are presented in the same order as the Section 2 subsections.