Response to the reviewer's report

Reviewer's Comments:

In the manuscript entitled ”Automated classification of Chandra X-ray point source...” (MN-22-3602-MJ) by Shivam Kumaran et al the authors apply machine learning tools for the automated classification of X-ray point sources, from the Chandra Source Catalogue (CSC 2.0), into various types of astronomical sources like AGN, X-ray emitting stars, young stellar objects etc. The input data includes, in addition to the Chandra X-ray data, multiwavelength data from various surveys. The authors are able to identify a large number of sources at the 3σ and 4σ level. The paper is in general well written and the methods followed in the analysis have been described clearly in the earlier part of the paper. The later part needs some changes and improved explanations, which I suggest below. The results are useful, though some further considerations are needed. I should be able to recommend acceptance of the manuscript for publication after the changes are made.

We thank the reviewer for the valuable comments that have helped improve the scientific quality and readability of the manuscript. We have implemented all the suggestions in the revised version of the manuscript (main changes/additions are in red). Below we give our response after each of the reviewer's comments.

DETAILED COMMENTS:

Section 2.1:

1. The meaning and significance of the variability properties calculated using the three methods needs to be explained.

   We are using 7 variability properties given in the CSC, out of which 3 are inter-observation properties, and 4 are intra-observation properties. In the revised text, we have elaborated on the methods of calculating the variability properties separately for inter-observation and intra-observation. (Page 2, para 7: starting with 'The CSC...')

Section 3.3 :

2. In the context of missing values, what is the meaning of imputation? The statement “...we select to impute the missing values using column mode values” is not very meaningful without further explanation, please provide that.

   In the data table, rows represent different sources, and the column represents different features for these sources. There are missing values in many columns due to various reasons described in the manuscript. Some of the ML algorithms can not work with the missing values. Imputation refers to the method of filling in these missing values. One of the imputation methods is to fill the missing data column-wise. For each feature column, we calculate the mode of available values in that column and fill in all the missing values in the column using this mode. We have added this description in the modified manuscript (Page 6, para 2: starting with ‘Most ML..’)

3. From Figure 1 it is seen that the percentage of missing values is very high in the lower part of the table. Surely, imputation will not work very well in such cases. You would not be able to impute well if the percentage availability were zero. The availability is very small in several of the cases. This matter needs to be discussed well.

   For a high percentage of missing values, the imputation method may not give satisfactory results, and in some cases, it is counter-productive to the final output. The RF and GBDT require missing value imputation and, therefore, may not perform well, particularly for minority classes where the percentage of missing values is high. Therefore, we have tried to avoid imputation altogether using the model LightGBM as our final classifier. This model is capable of handling the missing values in the data table and also provides better results in every aspect compared to RF and GBDT. We have added this description in the revised manuscript (page 6, para started with ‘Most ML classification...’).

4. In Figure 1, there are only two variability related parameters are included, while three were defined in Section 2.1.

   Two types of variabilities, namely, inter-observation and intra-observation, are used in this work. In Section 2.1, we have described the three methods for computing intra-observation variability, and inter-observation variability is calculated using the distribution of the photon fluxes of the individual observations. In Figure 1, we have clubbed together the features based on their origin. For example, if Gaia has no observation for a source, then all the features related to Gaia will be unavailable. Similarly, three inter-observation variability features are clubbed together as ‘inter obs-var’, and four intra-observation variability features are clubbed together as 'intra obs-var' in Figure 1. We made the required modifications in the text to explain this (page 6, para 1).

Section 3.4 :

5. This Section should only contain the first two paragraphs which describe the upsampling and SMOTE. The other two paragraphs should be transferred to Section 3.5.

   In the revised text, we have transferred the two paragraphs to Section 3.5, with some modifications to maintain the flow of the text.

6. The same comments as those for Section 3.3 apply to the class imbalance problem. The authors say that there is a "vast imbalance" in the number of training sources. In such a case, does it make sense to use a technique such as SMOTE to tackle the problem?

   Since the number of identified sources in the training sample belonging to the classes: LMXB, ULX, CV, and Pulsars is limited, we can not do much than oversampling. The oversampling also depends on the amount of data available and can be significantly effective only if the number is enough to represent the true distribution. Figure 2 in the manuscript shows that the oversampling improves the result significantly for the minority classes though it is still much lower than the accuracy of the majority classes. However, we keep the minority classes in our classification scheme; otherwise, these minority class objects will be classified as the majority class.

7. This fact needs to be clearly mentioned and discussed. The fourth paragraph of the current Section 3.4 does mention the problem, but it needs elaboration.

   We have discussed this issue in the revised text (page 7, paragraph starting with 'Nevertheless, any class balancing..').

Section 3.5 :

8. The new Section 3.5 will have the last two paragraphs of the current Section 3.4. The material needs to be rearranged so that the definitions and explanations are all done before the confusion matrix is introduced.

   We have modified the text as suggested by the reviewer.

9. What exactly are the numbers in the confusion matrix in Figure 2? Are these related to precision or recall? How are the numbers calculated? I find that these matters are explained when Figure 5 is described. The explanations should be transferred to the discussion of Figure 2, and need not be repeated later.

   The numbers in the confusion matrix show the percentage of sources truly belonging to the class given on the vertical axis, which is labelled as belonging to the class given on the horizontal axis. The cross-validation algorithm provides labels for each source from the model. Using this predicted label and the true label, the confusion matrix is computed. The diagonal elements essentially are the recall score for the individual classes. The explanations in the text are moved to section 3.5 (page 7, paragraph starting with "We use confusion matrix for....").

10. In Fig. 2, for the no upsampling case, the numbers for LMXB and later sources are not good. The improvement after using SMOTE does not take the numbers to satisfactory levels. Clearly the data is sufficient to give results at the acceptable level for AGN, Star and YSO and to a lesser extent HMXB, which constitute 10 percent of the training set. The other objects, LMXB, ULX, CV and pulsar together constitute less than 10 percent of the training set, and therefore the results obtained for them are poor. A catalogue of identifications of X-ray point sources made using the ML techniques can be trusted at better than 90 percent for the first three kinds of objects. But the entries for the other objects would have poor confidence and therefore would not be useful.

    As we have mentioned in response to point no 6, one of the reasons to include the minority class is to avoid misclassifying the minority class as the majority class. To test the effect of the inclusion of minority class, we have trained another model using only the majority classes: AGN, YSO, Stars and HMXB with and without MW features. Table 1 below shows the scores for the model trained only on the majority classes. The comparison between this table with Table 7 of the manuscript, where we considered all the classes, shows only a marginal improvement in performance. For the all-features case, the F1 scores of AGN, Star, and YSO show $\sim$ 1%, and HMXB shows $\sim$ 2% improvement. Considering only the X-ray features, the F1 scores of AGN, Star, YSO and HMXB have 3%, 3%, 1% and 2% improvement, respectively. In conclusion, when the MW features along with the X-ray features are considered, the inclusion of minority classes does not have a significant effect on the majority class scores. Therefore, we choose to keep all 8 classes in the final classifier. This table and the corresponding discussion are included in the revised manuscript as Table 8 on page 9, last paragraph.)

Table 1: Class wise precision, recall and F1 score for CCV using LightGBM model trained only on the majority classes: AGN, YSO, Star and HMXB.

Class	Sample type	Precision	Recall	F1 Score
AGN	all-features	98.3±0.1	98.4±0.1	98.3±0.1
	X-ray	95.5±0.2	96.5±0.2	96.0±0.1
STAR	all-features	97.3±0.1	96.3±0.1	96.8±0.1
	X-ray	92.7±0.2	89.6±0.2	91.1±0.2
YSO	all-features	93.7±0.2	96.5±0.2	95.1±0.2
	X-ray	84.4±0.3	91.2±0.4	87.7±0.3
HMXB	all-features	93.9±0.4	93.0±0.3	93.4±0.2
	X-ray	93.8±0.4	90.7±0.4	92.2±0.3

Section 4 :

11. It is said in the fourth paragraph of the section that "...steeper the plot towards unity, higher the predicted CMP and more confident the model. The figure reveals that the LightGBM is the most confident classifier...". But from the figure it is appears the curve for LightGBM is the least steep. Please clarify.

    For LightGBM, the curve of cumulative count vs class membership probability is almost flat to the lowest level till ∼ 0.9 and then sharply rises as we approach 1 than other curves. This means that for LightGBM very less fraction of sources are below 0.9 CMP as compared to other models. For clarity, we have replaced the term '..steeper the plot..' with 'more sharply peaked the plot towards unity....' in the revised manuscript (page 8, para 6 starting with 'Apart from high ...').

12. It is seen from Table 7 that use of just the X-ray data already provides precision and recall values close to 90 percent. Use of the multiwavelength data further improves the levels. The results for LMXB and later are not as good. I believe that the results for the four classes with at least 10 percent data fraction (AGN, Star, YSO and HMXB) will improve significantly if the other four kinds of sources are not included in the training and validation. I would like the authors to provide a separate table with just the four top kinds of sources. If the results are good, then a catalogue of such source identified with the X-ray point sources would be very useful.

    As discussed above (point no. 10), this result does not bring significant improvement to our original findings and choose to keep all 8 classes in the final classifier. However, we present this result in Table 8 of the modified manuscript for a comparative study.

13. In the same spirit, the role of various data sets like GAIA, 2MASS etc. in the training needs to be investigated. From Fig 1 it seen that these catalogues have data only for a small percentage of the sources, leading to many missing values. These catalogues may therefore not be adding value to the training, and it could in fact be counterproductive to use them. Training with these data sets omitted needs to be considered.

    As suggested by the reviewer, we have removed various MW catalogues one by one from the data table and calculated the cross validation score to find the impact of individual MW catalogues. In LightGBM (which can work with missing values), we did not observe any negative impact of adding a catalogue with a large fraction of missing values. The result of this analysis is given in this response report in Table 2 below. Since we have used multiple catalogues for each band (optical/UV and IR), dropping off, one catalogue hardly makes any impact on cross-validation scores. However, we observe a significant decline in the scores (except HMXB and LMXB) when all the MW catalogues are dropped.

Table 2: Class-wise cross validation scores are presented for various scenarios. In the 'sample Type', 'all' means the case when all the MW features were used and 'no mw' indicates the case where only X-ray features were used. In other cases, for example, no WISE shows the case when all the MW features were used except the features obtained from the WISE catalogue and so on.

Class	**Sample type**	**Precision**	**Recall**	**F1 Score**
AGN	all	96.8±0.1	97.6±0.2	97.2±0.1
	no WISE	95.9±0.1	97.4±0.1	96.6±0.1
	no 2MASS	96.7±0.2	97.3±0.2	97.0±0.1
	no SDSS	96.6±0.1	97.5±0.1	97.0±0.1
	no GAIA	96.4±0.1	97.0±0.1	96.7±0.1
	no GALEX	96.6±0.2	97.5±0.1	97.1±0.1
	no MIPS	96.7±0.2	97.4±0.1	97.0±0.1
	no mw	91.1±0.1	95.1±0.1	93.0±0.1
STAR	all	96.1±0.1	95.7±0.1	95.9±0.1
	no WISE	95.9±0.2	95.5±0.1	95.7±0.1
	no 2MASS	95.0±0.3	94.2±0.1	94.6±0.1
	no SDSS	95.8±0.1	95.4±0.1	95.6±0.1
	no GAIA	95.5±0.3	94.7±0.1	95.1±0.1
	no GALEX	96.0±0.2	95.7±0.1	95.9±0.1
	no MIPS	95.9±0.2	95.7±0.1	95.8±0.0
	no mw	89.2±0.2	88.2±0.1	88.7±0.1
YSO	all	92.8±0.2	95.4±0.3	94.1±0.2
	no WISE	92.7±0.2	95.5±0.3	94.1±0.3
	no 2MASS	90.6±0.4	94.6±0.6	92.5±0.2
	no SDSS	92.3±0.0	95.4±0.1	93.9±0.0
	no GAIA	92.6±0.1	95.1±0.2	93.9±0.2
	no GALEX	92.7±0.2	95.7±0.3	94.2±0.2
	no MIPS	92.2±0.4	95.1±0.4	93.6±0.3
	no mw	82.9±0.1	89.4±0.3	86.0±0.1
HMXB	all	91.8±0.2	90.7±0.7	91.2±0.4
	no WISE	91.7±0.1	90.5±0.6	91.1±0.4
	no 2MASS	91.6±0.2	90.7±0.3	91.1±0.2
	no SDSS	91.9±0.2	90.6±0.4	91.3±0.3
	no GAIA	91.5±0.5	91.0±0.3	91.2±0.3
	no GALEX	91.7±0.3	90.7±0.6	91.2±0.3
	no MIPS	92.0±0.4	90.5±0.3	91.3±0.3
	no mw	92.3±0.5	90.1±0.4	91.2±0.4
LMXB	all	94.8±1.3	80.7±0.8	87.2±0.6
	no WISE	95.6±0.8	81.0±0.5	87.7±0.3
	no 2MASS	94.9±0.6	81.0±0.8	87.4±0.4
	no SDSS	93.8±1.0	80.8±0.3	86.9±0.4
	no GAIA	95.2±0.3	81.1±0.0	87.6±0.1
	no GALEX	94.8±0.6	81.1±0.4	87.4±0.4
	no MIPS	95.5±1.1	80.8±1.3	87.6±1.0
	no mw	94.5±1.0	81.7±0.5	87.6±0.5
ULX	all	71.7±1.2	71.2±1.1	71.5±1.0
	no WISE	69.5±1.0	65.9±0.9	67.6±0.7
	no 2MASS	71.3±0.5	70.7±1.1	71.0±0.7
	no SDSS	72.2±1.2	69.3±1.0	70.7±0.7
	no GAIA	66.7±1.4	66.2±1.5	66.4±1.3
	no GALEX	72.0±1.3	68.5±1.3	70.2±0.7
	no MIPS	70.3±1.8	69.9±1.3	70.1±1.5
	no mw	61.9±1.3	43.6±1.3	51.2±1.4
CV	all	60.6±1.2	54.5±1.5	57.4±1.3
	no WISE	59.8±1.0	52.9±0.8	56.1±0.8
	no 2MASS	63.5±0.6	56.7±1.8	59.9±1.1
	no SDSS	59.5±1.1	54.2±1.1	56.7±0.9
	no GAIA	59.9±1.7	54.8±1.6	57.2±1.5
	no GALEX	60.9±1.3	55.5±1.4	58.1±1.3
	no MIPS	59.5±0.9	53.5±1.6	56.3±0.9
	no mw	56.1±2.8	45.1±1.0	50.0±1.7
PULSAR	all	42.1±1.9	45.3±1.9	43.7±1.9
	no WISE	41.8±1.4	42.0±0.5	41.9±0.5
	no 2MASS	34.4±2.1	35.4±3.2	34.9±2.6
	no SDSS	41.8±2.6	43.4±3.0	42.6±2.7
	no GAIA	34.7±1.4	41.0±1.6	37.6±1.4
	no GALEX	42.1±1.2	44.2±1.5	43.1±1.3
	no MIPS	41.1±2.6	42.6±3.1	41.8±2.7
	no mw	28.2±1.9	18.6±0.7	22.4±0.9

We have also tested the importance of MW catalogue band-wise and calculated the cross validation scores for four combinations of data-sets: All MW features, X-ray+optical/UV features, X-ray+IR features and only X-ray features. For majority classes, we observe only about 1-3% reduction in the F1 scores when an individual band of MW features are not used. As stated above, MW features are unimportant for HMXB and LMXB. For ULX and pulsar, each additional MW band provide 7-10% improvement in the F1 scores. We modified the discussion on the impact of various MW catalogues in the revised manuscript (page no 8, para 9: starting with 'For the ULX').

Table 3: Class-wise comparison of cross-validation scores for band-wise addition of catalogue (all, X-ray+optical/UV, X-ray+IR, no MW) in the data table.

Class	Sample type	Precision	Recall	F1 Score
AGN	all	96.8±0.1	97.6±0.2	97.2±0.1
	X-ray+optical/UV	94.9±0.2	97.0±0.1	96.0±0.1
	X-ray+IR	95.8±0.1	96.4±0.2	96.1±0.2
	no mw	91.1±0.1	95.1±0.1	93.0±0.1
STAR	all	96.1±0.1	95.7±0.1	95.9±0.1
	X-ray+optical/UV	94.1±0.3	93.2±0.1	93.7±0.2
	X-ray+IR	95.2±0.2	94.4±0.2	94.8±0.2
	no mw	89.2±0.2	88.2±0.1	88.7±0.1
YSO	all	92.8±0.2	95.4±0.3	94.1±0.2
	X-ray+optical/UV	88.6±0.3	93.9±0.4	91.2±0.3
	X-ray+IR	91.7±0.7	95.0±0.2	93.3±0.4
	no mw	82.9±0.1	89.4±0.3	86.0±0.1
HMXB	all	91.8±0.2	90.7±0.7	91.2±0.4
	X-ray+optical/UV	91.5±0.1	90.4±0.4	90.9±0.2
	X-ray+IR	91.9±0.2	90.8±0.3	91.3±0.2
	no mw	92.3±0.5	90.1±0.4	91.2±0.4
LMXB	all	94.8±1.3	80.7±0.8	87.2±0.6
	X-ray+optical/UV	95.9±1.6	81.1±0.4	87.9±0.9
	X-ray+IR	94.4±1.7	81.5±0.3	87.5±0.8
	no mw	94.5±1.0	81.7±0.5	87.6±0.5
ULX	all	71.7±1.2	71.2±1.1	71.5±1.0
	X-ray+optical/UV	69.9±2.0	61.8±1.7	65.6±1.6
	X-ray+IR	64.6±2.0	63.6±2.0	64.1±2.0
	no mw	61.9±1.3	43.6±1.3	51.2±1.4
CV	all	60.6±1.2	54.5±1.5	57.4±1.3
	X-ray+optical/UV	59.8±1.5	51.1±1.0	55.1±1.2
	X-ray+IR	58.5±0.6	54.8±1.3	56.6±0.7
	no mw	56.1±2.8	45.1±1.0	50.0±1.7
PULSAR	all	42.1±1.9	45.3±1.9	43.7±1.9
	X-ray+optical/UV	33.2±2.1	27.7±1.5	30.2±1.3
	X-ray+IR	35.3±2.1	40.0±2.6	37.5±2.2
	no mw	28.2±1.9	18.6±0.7	22.4±0.9

14. In Figure 6, some remarks, if possible, on the different shapes of pdf for different class of sources will help.

    The PDF curves show the overall classification confidence for each class. The PDFs are more sharply peaked close to 1.0 for the majority classes (e.g., AGN, STAR, YSO). It means the LightGBM model can predict for these classes with the high confidence. In contrast, for some other classes (e.g., ULX, CV), the peak appears at lower CMP values, implying a good fraction of sources with lower classification confidence. We find double-peak profiles for some classes (e.g., LMXB). This implies that the nature of some sources of such a class is predicted with low confidence values, while that of other sources of the same class is predicted with high confidence values. The reason could be the availability or unavailability of one or more observational features for certain sources of that class. For example, we find that the inter-observation variability feature is not available for a larger fraction of low confidence LMXBs. (page no 10).

Section 5 :

15. The meaning of the statement "...we will identify those sources which could be assigned to various classes...and thus claim the discovery of many new sources..." is not clear. Are the sources not already identified in the present paper?

The sources are identified in this paper. We will list the sources and their coordinates along with the entire MW data table in a subsequent publication discussing the properties of the identified objects in detail. We have modified the above statement in the revised manuscript (3rd paragraph in 'summary and conclusion section').

General Comment:

16. There are several minor grammatic errors spread through the manuscript which need to be attend to by the authors and/or the editorial team.

    We have carefully gone through the manuscript and tried to address grammatical errors in it.