The authors want to thank to the editor and the reviewers for the performed assessment of our work, as they gave us the opportunity to improve our contribution.

We have addressed all the comments received, as we are showing throughout the underneath specific comments.

Comments to the Author

I have read the manuscript entitled “Estimates of the atmospheric parameters of M-type stars: a machine learning perspective” by Sarro and collaborators. The authors provide a thorough description of the application of genetic algorithms and machine learning to the challenging problem of estimating M-type stellar parameters from spectroscopy. Although the results of their methods are not a significant improvement over the results of previous methods, the description of the application of machine learning is valuable to the community. I suspect with greater computational power and the inclusion of more features (the authors only use 10), their methods will eventually supersede current methods. The authors provide a road map for such a future application.

Thank you for the comment. Although the computational cost slightly increases by considering additional features, the main reason for restricting ourselves to ten is to be able to compare the ML approach with the one used in Cesetti et al.

I believe the general goals of the manuscript warrant publication in MNRAS. However, as currently written, the manuscript can be a bit hard to follow. I had trouble keeping track of the algorithmic vocabulary (predictor vs. predicted, mutation, elitism, gene, etc.), and how these terms related to the astrophysics. The flow of the manuscript and the choice of figures and tables should be improved before I can recommend it for publication. I list a number of comments below. I request that the authors address the comments in their revised version of the manuscript.

Thank you, we are glad of doing it.

Major Comments:

The authors make no mention of Mann et al. (2013), who approached the problem in a similar way: by searching for features that tracked M-type parameters. Mann et al. didn’t use a GA or machine learning, but the concept is similar. I was surprised the authors did not cite this paper:

http://adsabs.harvard.edu/abs/2013AJ....145...52M

We have include the proper citation as good settlement for metallicity estimations in such targeted stars. Thank you for the valuable contribution.

I was confused about the relationship between the genetic algorithm and the regression models.

We used the GA technique to propose the set of significant features. Therefore, regression models using such selected features are derived (as stated in second paragraph of section 2.2).

The authors state that the “fitness” in the genetic algorithm is how well the chosen features predict the physical values in the star. However, the authors did not provide an equation for the fitness. Could they add that equation?

More explanations have been given and the Eq 3 is now included. However, it was already explained in section 2.2 “Feature fitness was defined in terms of the RMSE of a linear regression model trained with the chromosome features. It is important to stress that the regression model used to evaluate the fitness of the feature sets (chromosomes) is not the same model that will be used in practice to predict physical parameters for observed spectra as described in Section 2.3 below. For fitness evaluation in the GA we used a simple multilinear model for the sake of speed, given the extreme size of the search space of all possible combinations of 10 spectral features.”

If the fitness of the genetic algorithm is how well the chosen features predict the physical values (Teff, M/H, logg), what is the need for the regression model? Or are the regression models run simultaneous to the genetic algorithm, and the regression models determining the fitness? This was not clear.

The fitness is implemented for selecting proper features, but as the number of evaluations are enormous when used by Gas, the fitness function is the simplest one.

After selecting the features, more sophisticated models are used, including non-linear models.

Another point of confusion was the nature of the regression models. Do they produce coefficients that relate the equivalent widths to physical parameters, or are they non-parametric? If they produce coefficients for, say, a polynomial relation, it would be useful to publish those.

As established in section 2.3 the different modelling techniques used are presented. Most of them are not suitable to be explicitly depicted under analytical expression. This is rather common in machine learning area. Trained models are, for sure, available if any practitioner is interested in RData format.

I was confused about what a “predicted” vs. “predictor” values were. These were never clearly defined in the text in relation to the spectra.

In our humble opinion, it was established that the feature set is composed by 10 genes (features).

Each feature involves two regions of the spectrum and the variable is created by the ratio of integrated flux (see section 2.2). Models want to define a relationship between the feature set and the predicted variables (physical parameters of stars). It is also described in section 2.2 and 2.3.

The authors also alternate between terms, using both “predictor” and “feature set” to describe the same thing. It would be easier for the reader I they consistently used one term for each concept.

Additional effort was carried out to avoid the word “predicted”. It can help to reduce confusion from readers.

Lines 45 to 65 on page 3 show the stages of the genetic algorithm, but the authors use generic GA terms (“chromosome”, “population”, “mutation”, “genes”). It would be helpful to the reader if they could state what those are in terms of the spectra. What is a gene in this context? Is it the feature wavelengths? What is a mutation in this context? Is it a change in the wavelengths chosen, or a change to the weighting of the feature, or both?

Chromosome and Gene are defined in section 2.2. “In this work we define a chromosome as a set of ten individual genes, and each gene codes a pair of non-overlapping spectral bands, the ratio of which is used as variable to model the physical parameters according to (1)”

A better explanation for mutation and population have been introduced in section 2.2 as well. Yes, they refer to change in the wavelengths defining the features.

And how is the fitness defined, mathematically? Consider that the audience for this article may not have any experience with genetic algorithms or machine learning.

Fitnes is explained in section 2.2 as well: “Feature fitness was defined in terms of the RMSE of a linear regression model trained with the chromosome features. It is important to stress that the regression model used to evaluate the fitness of the feature sets (chromosomes) is not the same model that will be used in practice to predict physical parameters for observed spectra as described in Section 2.3 below. For fitness evaluation in the GA we used a simple multilinear model for the sake of speed, given the extreme size of the search space of all possible combinations of 10 spectral features.”

In addition, the AIC criterion, which is the easiest implementation of the previous explanation, was also introduced.

Choice of figures and tables: The authors chose to include multiple appendices, each with a series of figures and tables. The figures and tables in the appendices more than double the length of the manuscript. Some of these figures and tables are important to understanding the goals of the manuscript and should be included in the main body of the text, not the appendices, in my opinion.

Other figures in the appendices do not add to the goals of the manuscript, and are only included for completeness. I worry that these will be distracting to readers and take away from the manuscript’s message. I wonder if they can be removed entirely, or included as a data supplement to the article online only.

Let’s follow your kind advises.

For example, Figures A1 through A9 are highly redundant with the tables in the appendix. I think one large (two column) figure showing the spectra is helpful and should be in the main body of the manuscript, but 9 are too many especially when the features are already identified in the tables.

We have grouped them in three figures and they were placed inside the text. This allow us to compare the bands in a visual way. The tables have been grouped also, in order to reduce their number and to keep them as much integrated as possible, for both datasets.

I don’t see the value of the table repeating the wavelengths used by Cesetti et al. (2013) in Table A3, since those are already in the literature, and as I said above. I think the other tables in Appendix A could be combined for brevity.

Such table was removed and the citation to the Cesetti et al. paper was kept.

It would be helpful if the tables included the fitness value (the “AIC”, although “AIC” is not defined when first mentioned), since they are sorted by fitness. Also, it would be better if the authors chose one term (either “fitness” or AIC) and used it consistently. I suggest using fitness, since it can be hard to remember what AIC stands for.

In this version of the paper the specific criterion used for fitness was introduced, including its mathematical formulation. Therefore, fitness was determined by calculating the Akaike Information Criterion (AIC). According to its definition (Eq 3), its specific value is meaningless, because k = 11 (10 coefficients plus the independent term) in all the cases and, only the –ln(Likelihood) makes sense to discriminate among features.

As usual, the lower AIC the better, but its current figures can distract the reader from the relevant items, that are the selected feature. Therefore, we have decided not to plot them.

We are including here the table including the AIC values as an example of the mentioned issue, for the IRTF dataset, and physical parameter Log(G) with SNR=infinity:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AIC** | **Lambda1** | **Lambda2** | **Lambda-C1** | **Lambda-C2** |
| -191.01222 | 10245.88 | 10304.02 | 11241.29 | 11328.54 |
| -190.78340 | 8415.91 | 8473.96 | 11511.51 | 11598.51 |
| -189.53985 | 12906.56 | 12993.61 | 13041.48 | 13133.82 |
| -188.78924 | 8716.00 | 8773.99 | 10425.90 | 10484.13 |
| -182.53653 | 8805.93 | 8863.97 | 12816.72 | 12903.73 |
| -181.11691 | 10126.02 | 10183.93 | 13086.46 | 13194.09 |
| -180.46070 | 8176.03 | 8234.13 | 10971.57 | 11058.46 |
| -174.85369 | 8626.02 | 8683.99 | 10746.43 | 10833.57 |
| -173.75193 | 8536.03 | 8594.06 | 10215.95 | 10274.10 |
| -172.74536 | 12951,62 | 13038.62 | 11196.56 | 11283.2 |

If now we consider for the same IRTF dataset, the safe physical parameter (log(G) but with SNR=50, we have:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AIC** | **Lambda1** | **Lambda2** | **Lambda-C1** | **Lambda-C2** |
| -66.461894 | 11151.63 | 11238.46 | 13086.46 | 13194.09 |
| -62.409394 | 8385.99 | 8443.94 | 13618.20 | 13734.14 |
| -61.200001 | 8176.03 | 8234.13 | 11241.29 | 11328.54 |
| -60.123256 | 8536.03 | 8594.06 | 13041.48 | 13133.82 |
| -60.093468 | 12771.70 | 12858.73 | 10306.03 | 10363.88 |
| -60.035034 | 13378.12 | 13494.13 | 10002.04 | 9999.92 |
| -59.554663 | 8626.02 | 8683.99 | 10926.46 | 11013.60 |
| -59.523491 | 9826.05 | 9883.91 | 10006.07 | 10064.01 |
| -59.410338 | 10521.56 | 10608.46 | 11736.71 | 11823.49 |
| -57.156750 | 8205.98 | 8263.96 | 9796.09 | 9853.94 |

As it becomes clear, AIC is not related to any astrophysical parameter, but to an information criterion for linear models, and it does not help as it introduces confusion because the values.

All in all, still differentiating fitness concept from the specific criterion used (AIC) matters. Therefore, after clarifying those elements in the text we have kept them.

Use of acronyms: The acronyms in the paper become very confusing about midway through the manuscript. As a reader, I had trouble keeping track of the different regression methods, and I never remembered what the acronyms stood for. I suspect there is a better way to distinguish the methods. I like the itemized list on page 4 (lines 45 to 54), but after that it gets confusing. Perhaps there isn’t a solution to this.

Well, we did on the best of our knowledge by analyzing different regression methods, but in order to report their performance some kind of acronyms are required as it does not make sense to repeat long sentences per table. Unfortunately, this is the best strategy we have found.

IPAC hosts the Dwarf Archives, but IPAC is much more than that, so I don’t think that is an appropriate acronym. I was confused by this choice. I suggest the authors use “DA” for dwarf archives, or just write it out completely. Writing out completely would help reduce the number of acronyms.

OK, Thank you for that clarification, which help to precise the scope. We have established, at the end of the Introduction that the specific section describes the same steps in the context of the Dwarf Archives (hereafter DA) collection of spectra. After that point we have used DA as the reference for such set of spectra.

I’m trying to understand the importance of including the RMDSE in the manuscript. Does using that metric result in different conclusions from using the RMSE alone? If not, then just use RMSE. The RMDSE confuses the picture. The authors could state that repeated the analysis with RMDSE and that the results were the same. If the results are different, the authors should try to explain why. Right now, it’s hard to follow why the authors used these two metrics, and it gets confusing since they are so similar.

The interest of authors was to provide not only the aggregated view of the error through the full dataset being modelled, but also its median, as it is less sensible to the outliers. As far as noise exists, we believe that provide the two figures helps the reader to understand when the model method has specific difficulties with a small set of spectra, which are responsible for increasing the total error. We have tried to make it clear with the added text just before of the subsection Effective Temperature Models.

They should be similar, yes. But this means that no specific outliers are responsible for increasing the error, which is relevant as well.

It’s important to point out that the results from Cesetti et al. also have errors. The RMSE between the new determinations and Cesetti et al. might be due to the errors in Cesetti et al. In which case, it’s not clear whether the RMSE is a good representation of which regression model is better.

We are not comparing or qualifying models but the quality of the features. Actually the general rule is that there are relevant information no only on specific features but in many regions of the spectrum. However, your comment is absolutely true and we have emphasized it in section 3.

The first item in the summary states that “the RMSE errors based on the training set of synthetic spectra are poor predictors of true performance.” I had trouble finding the evidence for this conclusion in the text. I did not catch this when I read through the paper. Since this is the first conclusion of the paper, perhaps it deserves its own section describing this result.

We have extended and better explaining our meaning in the conclusion section. However, the concept was explained in section 3 “… the cross-validation errors are far from informative with respect to the true performance when the models are applied to real data. In the case of the features defined by Cesetti et al. (2013), we find that the best RMSE/RMDSE (obtained not from cross-validation but from the comparison with the effective temperatures in Table 3 of Cesetti et al. (2013)) is attained by the CES-NNR model.”

The practical aspect is that errors when models are trained with part of the synthetic spectra and validated with the remaining part are small (RMSE ~ 40K) but there are not related to the ones found on real spectra. This means that neither the family of models nor the error values can be inferred from the training when synthetic library is used. (not informative).

Minor Comments:

Abstract:

M/H should be [M/H] in the abstract. Thank you, fixed.

I was surprised by the capitalization of I, C and A in independent component analysis. Similarly I didn’t think G and A should be capitalized in genetic algorithms. The authors should check with the MNRAS style guide to make sure they are using the right capitalization scheme.

Thank you, we have revisiting https://academic.oup.com/mnras/pages/General\_Instructions#6 Style guide and we did not find any specific rule. Therefore, and because they are commonly used as acronyms in capital letters, we suggest to keep them as they are. For sure, if it is mandatory, we will proceed immediately.

Introduction:

The authors focus on M dwarfs in the abstract, but their paper concerns M type stars generally, including M giants. This was a bit confusing.

To the best of our knowledge, what it is stated in the abstract is to be focus on M stars, as well as that results are validated with two different sets of M-stars. The proposed method is not restricted to dwarf stars at all.

The authors focus on Super-Earths, but M dwarfs host many more Earth-sized planets than sun-like stars as well. See Dressing & Charbonneau. Another advantage of the short period of the HZ is that there are more transits to see within a set time. The authors mention deeper transits and a higher probability, but didn’t mention that there are more transits. All of these contribute to the “M dwarf advantage.” => ESTO NO SE ARGUMENTARLO

The authors could cite Adams, Laughlin & Bodenheimer when discussing the lifetimes of M dwarfs. They were the first to investigate this. => ESTO NO SE ARGUMENTARLO

“heavily veil their spectra and H2” This was confusing. Do the authors mean the continuum is heavily veiled? H2 is itself a molecule, so to say molecules veil H2 is odd. => ESTO NO SE ARGUMENTARLO

I would avoid words like “extraordinarily.” Unless qualified numerically, they do not add scientific value to the paper. Fixed

“They predict…” Should be “predicted” in past tense. Fixed

“Newton et al. (2015) develop” should be “developed” in past tense. Fixed

“They quote…” should be “quoted” in past tense. Fixed

“Uncertainties of Temperatures…” temperatures should not be capitalized. Fixed

“proposed the sensitivity maps” remove “the” Fixed

“Machine Learning” should M and L be capitalized? I don’t think so. Fixed. We have implemented its common acronym (ML)

Section 2

“extremely difficult task” remove “extremely.” It’s not scientific unless numerically qualified. Fixed

“two bands is hoped to be a good indicator” I don’t like the term “hope” here. Hope is not scientific. “assumed to be” would be better. Fixed

“in-silico evolution” I don’t know what this term means. This term refers to the implementation of GA based optimizers by using computers

“very popular” remove very. Fixed.

“see De Geyter et al. 2013, for the last application of GAs in astronomy” There are thousands and thousands of astronomy papers published each year. How can the authors be sure this was the last use of GAs in astronomy?

Yes, sorry, it was a clear mistake. It was modified to say that such paper is a significant example of the use of GA in the field.

“The concept of fittest is context dependent…” The authors could add an equation for their definition of “fitness” in this application.

Yes, as far as we use AIC criterion, such mathematical formulation has been introduced.

“Elitism” I don’t know what this term means here.

It’s a concept related to GA, it means that in the evolution of population, new generations mostly relay on better parents. That is, the implementation natural selection is elitist in this sense.

“per physical parameter” add period at the end of this sentence. Fixed.

“can not” should be cannot. Fixed

“(often analytical)” I don’t know what non-analytical means here. Parameter estimation can be non-analytical? Fixed.

“Projection Pursuit” should this be capitalized? Fixed

“the Independent Components” should this be capitalized? Fixed

I am not used to pseudocode, so I have to trust the authors followed convention in their algorithm 2.1. Thanks for trusting

Section 3:

“During the preprocessing stage…” This would be better in section 2. It’s odd here in section 3.

Moved to Section 2

“The application of GAs to the selection of features…” This sentence was long and hard to follow. Break into two sentences? Yes, it was done.

Section 3.1 references figures and tables in the appendix. I think it would be better to have them here in the text, but with fewer figures and tables. There are several one sentence paragraphs here that could be combined to make full paragraphs. Otherwise, it reads like it was written very quickly.

Fixed.

“we conclude the two approaches are equivalent” It would be good to clarify which approaches the authors are talking about. I think they are talking about the different ways to measure equivalent widths, but there are so many “approaches” in the paper I’m not sure.

We explicitly described the two approaches

“in Table 3 ofCesseti et al.” typo Fixed.

Section 4:

“metallicities as low as -2.1” change to “metallicities as low as [M/H] = -2.1” Fixed.

Figure 6: Please indicate which Teff is on the x-axis and which Teff is on the y-axis in the caption. Fixed.

Table 1 and References: Lepine is spelled incorrectly. Fixed.

Figure 9: It would be clearer if the authors chose a smaller range for the x and y axis, and if they made the circles unfilled. Right now it’s just a cluster in the center and I can’t distinguish the circles because they are filled and overlap. No tengo (y no están en el github los fuentes de esta imagen para tocarla.