To the Recommender and Reviewers,

Please find my revised manuscript titled ‘Ammonoid Taxonomy with Supervised and Unsupervised Machine Learning Algorithms’. Thank you for the opportunity to revise my manuscript; I appreciate all of the feedback.

The insights from each reviewer were very helpful. I would like first to address the weakness of the dataset used being limited in both sample size and species count. Although I am very much interested in expanding the present study with more rich data in future, I am unfortunately limited for the time being to free, open data such as PBDB because as an amateur fossil enthusiast, I do not have access to the kind of data (especially images of ammonoids) which may be accessible to those employed at museums and universities. I am also limited in my literature searches to open-access articles because I lack the funds for pay-walled publications. (Should any of the reviewers be interested in continuing this line of research, please let me know!). This being said, I have tried to address this and other concerns. Listed below are my responses to the individual points raised.

Kenneth De Baets’s Comments

1. **Focus on conch parameters: It is ok to focus on conch parameters as these are easier to get and analyze in a biologically meaningful way but a bit more discussion on why this is the case as well as how adding additional parameters (e.g., suture line, ornamentation such as ribbing) might improve statistical power to separate species would be crucial to discuss. Many species are not just defined by conch parameters so it would be crucial to point out that you are working with only a subset of characters used to define species which are more readily available in the literature and easier to analyze quantitatively**

   My response: This is a very fair point, and although I did mention these additional features in the Introduction, I neglected discussing them elsewhere in the manuscript. I do believe that future studies which train models on images of ammonoid specimens would be a vastly improved way of going about this, however as an amateur with no academic affiliations to paleontological organizations I do not have access to such data. I have now added the following to the Limitations section 4.4:
   “In addition, the conch parameters used as fit data in these models are convenient to obtain and quantify but represent only a small subset of possible features. Future models may utilise the presence and characteristics of shell ornamentation and suture lines for greater statistical power. These were not directly available in the PBDB, though studies in the literature have explored quantitative analyses of suture lines (Manship, 2004; Wegerer et al., 2018).”

2. **Data: As pointed out in the manuscript, the dataset is limited to 11 species entered into the Paleobiology Database – the sample size of individual species are ok (> 50 – could still be better – some authors have suggested to have > 100 specimens available when including multiple ontogenetic stages, etc.). As an ammonoid worker which has worked on intraspecific variation – I can highlight that data for much more species would be available in the primary literature (a substantial part is still missing from the PDBD). I must admit that particularly in older literature measurements would need to be extracted from graphs and we still need to go some way before all paleontologists make this kind of data available as standard practice. Ideally, you should try to compile some additional data from the primary data to better understand what I mean and would help to broaden**

the scope of your analysis. As this is a pilot study, focusing on 11 species with samples > 50 could still ok, but it would be crucial to highlight which primary references yielded data for particular species. This also becomes crucial as for some species, data from multiple references are merged, presented data from multiple stratigraphic and geographic intervals (and likely also different degrees of preservation). This could for example explain the poorer performance for particular species like Owenites koeneni which derive from different localities and might also represent different preservations and ages.

My response: Thanks for pointing this out. I have included a new table (2) which highlights which primary references yielded data for particular species. I have also added the following to the Discussion:
"For example, multiple sources were used to provide *Owenites koeneni* data, and these sources differed in region, stratigraphic level, preservation, and authorship which may compromise their utility as a single homogeneous population related by conch proportions."
Please also see my response to 6.

3. **Please also, write species in italics as this is customary.**

   My response: All species throughout the manuscript are now written in italics.

4. **Performance of particular methods and species. The original authors might have assigned all their specimens to a particular species (e.g., Owenites koeneni) but mostly did not statistically evaluate how the conch parameters of their specimens compared with those of other localities and some even highlight qualitative differences with material from other localities. The homogeneity of conch parameters and there use to define species might therefore be to some degree compromised even before applying machine learning approaches.**

   My response: This is an excellent point, and definitely worth mentioning in the manuscript. I have added the following to the Limitations section 4.4:
   "A possible source of error in supervised models is the assumption that the species attributed to each specimen is itself suitable. It is not possible for all authors in the literature from which data are drawn to conduct their own statistical evaluations of specimens for verification. Indeed, some literature describe qualitative differences between their material and that from other localities. For example, multiple sources were used to provide *Owenites koeneni* data, and these sources differed in region, stratigraphic level, preservation, and authorship which may compromise their utility as a single homogeneous population related by conch proportions. A model trained on unreliable data produces unreliable results. This emphasises a need for future supervised models to be particularly critical of data sources. No such review was conducted for this pilot study. This also highlights the role of unsupervised models in future research."

5. **To place the performance of the methods into context for particular species, it would be crucial to add at least the primary reference providing data, their age range (single bed, biozone, etc.) as well the geographic scope (same locality, continent, etc.), so such potential issues could be glanced more transparently.**

   My response: Please see my response to 2. Table 2 now indicates the primary reference, period, and region for each source for each species. I hope this is now clear.

6. **In the discussion you focus on the performance of methods, but I would also be crucial to highlights which species are consistently picked up and which ones are not to better understand the impact of the issue of species definition. Which ones are often/sometimes merged and which ones are sometimes/often oversplit. This would allow a better discussion and understanding of how species definition and homogeneity of conch parameters might impact on the performance of the methods. At first glance, particularly Owenites koeneni seems to perform peculiarly and it is also one of the species which measurements deriving from several continents and publications. So it would be crucial to discuss this at greater length in the discussion**

   My response: This is certainly a good question. I have updated the manuscript to explain that nested cross-validation was used to determine which algorithm was best for supervised modeling, and I have then provided a detailed classification report from non-nested cross-validation to highlight which species were consistently picked up and which were not.
   I have updated section 2.2.1 of the Methods:
   "First, a range of supervised algorithms were compared with a 5 × 5 nested cross-validation approach to determine which is most suitable for ammonoid taxonomy. Nested cross-validation was implemented to avoid potential bias in performance evaluation due to over-fitting in model selection (Cawley and Talbot, 2010). This way, estimates for the unbiased generalization performance of each classifier were obtained through test and train accuracies averaged across the outer folds. For the inner folds, a grid search was utilized to select the model parameters which resulted in the best test accuracy. The classifiers implemented, as well as the range of parameters used in cross-validation, are summarized in Table 3. Next, the model with the best test accuracy (and with similar train accuracy to prevent over-fitting) was implemented with a five-fold cross-validation approach to identify how performance varied for each species."
   I have also updated the results with Table 6 which includes a detailed classification report for each species as requested. I have similarly updated the Discussion as appropriate.

7. **Code availability and reproducibility: It has been become standard practice to share the code at least upon publication (see Reviewer 2). Ideally, this should even be done during the review process as it would allow reviewers to verify the results, but I can to some degree understand the reluctance to do so before publication. Special repositories are however available for this purpose (GitHub) which allow to put embargos and restrictions on the availability of the data**.

   My response: I have uploaded my complete code to the online supplemental materials which is now hyperlinked at the end of my Methods section.

   Kenneth, I have also read through your annotated PDF comments and made appropriate modifications to the manuscript. Thanks for those.

Anonymous Reviewer's Comments

8. **My main concern with the manuscript is the reporting of the work. It has become routine in this field to publish code in a public repository like GitHub and to at least mention the platform used for the analysis (e.g. PyTorch) so that other can replicate, corroborate, and build on the work. The detail in this manuscript is very minimal. Detailed procedures**

would be especially helpful for the paleontological community, as it would give others guidance on how to apply machine learning to their own morphological datasets.

My response: Thank you for this suggestion. Please see my response to 7. Code is now available in the online supplemental materials.

9. **Additionally, while it appears that the results are promising, higher accuracies may be possible with new machine learning approaches, e.g. convolutional neural networks, and with a morphological dataset that perhaps includes more morphological features. Many of the morphological traits that we use in taxon identifications are chosen as much for their ease of measurement and reproducibility, as their diagnostic strength. Computer vision has the potential to expand the morphological traits we can use in taxonomic determinations.**

   My response: Thank you for raising this point, and it is a good one. I think this also ties in with what other comments were suggesting about the use of additional features such as suture lines achieving higher accuracy. Indeed, images of specimens which capture there features fed into a convolutional neural network would presumably present a vast improvement over this study. However as previously noted I lack access to a vast array of ammonoid images. That said, I wanted to investigate your suggestion with the tools and data available to me and so I undertook a supervised neural network approach to these data implemented as a multilayer perceptron. I have added this model and its results to the manuscript (see Tables 2 and 4) as well as a discussion of other neural network approaches under section 4.1:
   "For these particular fit data, the neural network approach of the Multilayer Perceptron did not offer a performance benefit over other classifiers. The Multilayer Perceptron is however a more dated neural network (Hastie et al., 2009), and a more recent paradigm such as the convolutional neural network popularized by Krizhevsky et al. (2012), which improves upon the Multilayer Perceptron by utilizing regularization, is likely to improve vastly upon the present models for richer sources of data, for example images of ammonoids which include additional features such as ornamentation and suture lines. However, these models are beyond the scope of the present pilot study and these data are unavailable in PBDB."

Jérémie Bardin's Comments

10. Indeed, my main concern is that the somewhat good results in this paper come from the very low number of species treated. The author mentions as a limitation the low number of specimens for training but, to me, the real weakness is the number of species. Using machine learning methods for species identification of a higher number of species will require way more descriptors. The parameters used in this paper are the most common and the corresponding measurements are provided in the majority of papers describing ammonoids. Given that these measures are extremely available and that many people have

important databases of such measures, I would have expected more than 11 species to demonstrate that these parameters with machine learnings methods could provide useful tools to identify ammonites and build a robust taxonomy.

My response: Yes, it is unfortunate that I do not have direct access to richer data. I have therefore included a discussion of the limited number of species in the Limitations: "Relatedly, the number of species used is very limited. This is also a consequence of the available PBDB data and the inclusion criteria of a species having at least 50 specimens. Future studies should seek to replicate these findings which richer data sources in both sample size and species count."

11. **Moreover, I would like to see more insights on the very general use of supervised vs unsupervised methods for ammonoids taxonomy. Supervised methods make the assumption that the target variable is true. These methods are thus suited to identify specimens given a robust taxonomy. Unsupervised methods, if performed on already settled taxonomy, quantify the congruence/difference between, on the one side, species definitions and taxonomic attributions of specimens and, on the other side, their morphological clustering. They should also be used to create taxonomy but there is much to do to properly include stratigraphical time, ontogeny and any kind of morphological features. All those points could be addressed in the discussion.**

    My response: Thank you for these suggestions, I think they will be helpful for researchers who do not understand the value of quantitative taxonomy. I have added the following to section 4.1: "Supervised models operate on the assumption that the labels (here species) in the training data are true. Therefore, given a robust taxonomy, supervised models such as those in the present study may enable researchers to identify specimens in an unambiguous and objective way. With larger datasets and more features beyond conch proportions, the methods demonstrated here may provide a rigorous quantitative determination of ammonoid taxa."
    And the following to section 4.2:
    "In contrast to supervised models, unsupervised models operate on the assumption that the labels (here species) in the data are unknown. These methods then allow for the construction of a robust and quantitative taxonomy. When applied to a settled taxonomy, unsupervised models quantify the congruence between species definitions and taxonomic attributions of specimens, and their actual morphological clustering. Unsupervised models such as those in the present study, improved with a greater range of ammonoid features, should be implemented extensively by researchers to suggest improvements upon the current taxonomy."

12. **Abstract - The verb "taxonomize" is rarley used. I am not sure of its meaning, I would recommend being more specific.**

    My response: Sorry, I was unaware of this. I have replaced instances of "taxonomize" in the manuscript with "determine the taxa of". I hope this is more specific.

13. **Introduction - Ammonoids is not the name of the subclass, it is Ammonoidea.**

    My response: Sorry. I have replaced ‘Ammonoids’ with ‘Ammonoidea’.

14. **Using Linnean nomenclature is very controversial now (even if still correct), I would replace “subclass” by “clade” or simply “group”**

    My response: Thank you for catching all of these nomenclature errors. I have replaced ‘subclass’ with ‘group’. I appreciate your attention to detail.

15. **The two parts of this sentence are somewhat redundant “ammonoids are crucial index fossils for biostratigraphy (Cox, 1995), therefore ammonoid taxonomy is useful for the study of stratigraphic subdivision”.**

    My response: You’re right. I have deleted the latter half of this sentence.

16. **“conch morphology, coiling, and aperture shape.” Actually, coiling and aperture shape are parts of conch morphology. Better say: “conch morphology such as coiling, and aperture shape”**

    My response: Thanks, I have changed this sentence to “conch morphology such as coiling and aperture shape” as suggested.

17. **“ribs (their direction, spacing, and type) may be used for family classification”. I would remove “family”, ribs are useful at every level of taxonomy. By the way, I would also recommend removing Linnean ranks as much as possible.**

    My response: As suggested I have removed “family” from this sentence.

18. **Despite all the great things Dieter Korn did and does on ammonoids description, I am not sure he defined the numerous parameters in its 2010 paper as written. It seems to me that its contribution is more a summary and formalization.**

    My response: I have replaced the word “defines” with “standardized” in this sentence.

19. **“Since ammonoids exhibit intraspecific variation (De Baets et al., 2015), it follows that each species has a typical range of conch proportions which are diagnostic of taxonomy.” This is one of the main problems. Ammonite species are usually not built on advanced quantitative diagnoses. Most of the time, several species (usually a lot) will have overlapping morphologies. Moreover, many species are partly defined on stratigraphy itself. I think that the way ammonites’ species are built and the variability in this practice are of prime importance to properly use machine learnings algorithms.**

    My response: You’re right, and this point was also made by another reviewer. Please see the

following addition to the Limitations:
"A possible source of error in supervised models is the assumption that the species attributed to each specimen is itself suitable. It is not possible for all authors in the literature from which data are drawn to conduct their own statistical evaluations of specimens for verification. Indeed, some literature describe qualitative differences between their material and that from other localities. For example, multiple sources were used to provide *Owenites koeneni* data, and these sources differed in region, stratigraphic level, preservation, and authorship which may compromise their utility as a single homogeneous population related by conch proportions. A model trained on unreliable data produces unreliable results. This emphasises a need for future supervised models to be particularly critical of data sources. No such review was conducted for this pilot study. This also highlights the role of unsupervised models in future research."

20. **Supervised (eg discriminant analyses) and unsupervised (clustering) methods have already been used on ammonoids, I expect the introduction and/or the discussion to review what has already been done on the topic (e.g. Hohenegger and Tatzreiter 1992, Meister et al. 2011, Bardin et al. 2015).**

    My response: Sorry for missing these. I can honestly say that I did conduct a Google Scholar search for machine learning ammonoid articles but I must have either missed these or they weren't returned in my searches. Unfortunately I'm not an expert on ammonoid literature so I wasn't aware of these studies. Thanks for bringing them to my attention. I have added the following to the introduction of the manuscript:
    "Since ammonoids exhibit intraspecific variation (De Baets et al., 2015), it follows that each species has a typical range of conch proportions which are diagnostic of taxonomy. Hohenegger and Tatzreiter (1992) attempted multivariate classification of Balatonites by their morphological parameters with discriminant analysis. This work identified two well-separated groups in data from 19 reported species, suggesting these really represent only two distinct biological species. In contrast, Bardin et al. (2015) more recently demonstrated no overlapping morphology in Dactylioceras specimens, suggesting these are well classified. Taken together, these studies imply that classification algorithms are useful tools for assessing the validity of taxonomy in the literature. The aim of this study is to determine the taxa of ammonoids by their conch proportions, and expand upon the work of Hohenegger and Tatzreiter (1992) and Bardin et al. (2015) by implementing a range of supervised and unsupervised machine learning algorithms not previously applied to ammonoid taxonomy. This presents novel methodological research and lays the groundwork for future methods in biostratigraphy, systematic palaeontology, and evolutionary biology."
    As well as the following to the discussion:
    "Most relevantly, Meister et al. (2011) achieved similar accuracy (≥ 73.7%) classifying Fuciniceras specimens with data from Fourier transforms of ribs."
    I hope that my interpretation of these works is correct. I have also cited Manship (2004) and Wegerer et al. (2018) at the suggestion of another reviewer.

21. **Data - The number of specimens and species may be a weakness of this work. I am not surprise to see such results for 11 species.**

    My response: Yes, like the sample size, the number of species is also a limiting factor. I have updated the sample size limitation paragraph to also state "Relatedly, the number of species used is very limited. This is also a consequence of the available PBDB data and the inclusion criteria of a species having at least 50 specimens. Future studies should seek to replicate these findings which richer data sources in both sample size and species count."

22. **The author mentions in the "Limitations" section that he was not able to differentiate juveniles and adults but given the fact the diameter is used, there are ways to infer it.**

    My response: This is a very fair point, but making these distinctions is beyond the scope of the present study. To reflect this, I have reworded the sentence in question to "Furthermore, in the present study no distinction is made between mature and juvenile specimens, nor males and females."

23. **Discussion - "A nearest neighbours algorithm was then implemented to calculate the average distance between each point and its m nearest neighbours". Use m as defined before.**

    My response: Good suggestion, thanks. I have made this change in the manuscript.

24. **I don't understand the comparison of test accuracies to the accuracy of majority class prediction. Could you be more specific and explain why it is your baseline hypothesis?**

    My response: Sorry for the confusion. The purpose of this comparison is simply a 'sanity test' to determine whether the models are at least better than a very simple classifier. In my experience, this is common in evaluating the performance of complex classification models. I have now prefaced this comparison with the sentence:
    "As a sanity test, these results can be compared to a very simple classifier."

25. **A naïve question: is the test accuracy sufficient to choose between methods or do we need to consider the difference between test and train accuracies. In other words, do we care about an important over-fitting is the test accuracy is really good?**

    My response: Thanks for this question. My understanding is that if there is very little difference between the test and train accuracies, then there is no evidence of over-fitting. However, if there is a significant difference between the test and train accuracies, then this may be evidence of over-fitting in the model (though this is not a guarantee). Because of this risk, if we have two models with test accuracies which are not significantly different (looking at the

SDs) but one has a test-train difference significantly higher than the other, then it is generally safer to select the model with the lower difference to avoid the possibility of poor out-of-sample performance. This was my reasoning behind the paragraph beginning “Another result to consider when evaluating which model performed best is the average train accuracy…”

26. **Additional references:**
Bardin J, Rouget I, Benzaggagh M, Fürsich FT and Cecca F. 2015. Lower Toarcian (Jurassic) ammonites of the South Riffian ridges (Morocco): systematics and biostratigraphy. Journal of systematic paleontology. 13 (6). 471-501.
Hohenegger J and Tatzreiter F. 1992. Morphometric methods in determination of ammonite species, exemplified through Balatonites shells (Middle Triassic). Journal of Paleontology 66(5): 801-816.
Meister C, Dommergues JL, Dommergues C, Lachkar N and El Hariri K. 2011. Les ammonites du Pliensbachien du jebel Bou Rharraf(Haut Atlas oriental, Maroc). Geobios. 44. 117.e1–117.e60

    My response: Thanks again for bringing these to my attention. Please see my response to 20.

I hope the changes I have made are agreeable, and I am very grateful for your continued consideration of this manuscript. I have included the names of the reviewers in the Acknowledgments section of the manuscript to express my appreciation for helping me better my research.

Sincerely,

Floe Foxon
ORCiD: https://orcid.org/0000-0002-4893-9178
Twitter: https://twitter.com/FloeFoxon