Reviewer A:

Review of “Feasibility of Deep Learning in Shear Wave Splitting analysis using Synthetic-Data Training and Waveform Deconvolution” by Chakraborty et al., 2023.

This study presents an application of deep learning to automatically determine apparent splitting parameters, with application to teleseismic XKS waves. The study explores several versions of training datasets, and uses an effective deconvolution procedure to isolate robust signals. The study then successfully applies the deep learning procedure to both synthetic data and also real data measured by the USArray. The station-averaged results obtained on real data are particularly compelling, exhibiting a strong agreement with previous studies.

Overall, I find the paper well-written. The method is clearly laid out, the figures are clear, and the argument is persuasive. This study presents an opportunity to effectively automate and make a traditionally time-consuming measurement process far more replicable, and will be of interest to a wide group of seismologists, given the ubiquitous nature of XKS splitting studies.

However, since the method is applied to real data, in settings that likely have multi-layer anisotropy, I think some additional attention should be paid to the associated challenges with working with such settings, and I make some suggestions below for how this could be done. I also think a discussion/acknowledgement of measurement uncertainty could strengthen the paper.

Major Comments:

1. Regarding the comparison between the two sets of USArray station-average splitting parameters:

   According to the manuscript, the individual measurements that are used for the calculation of the station-average parameter are selected based on the 60% reduction threshold of the transverse component energy. My understanding is that the Liu et al., 2014 study does not use this criterion. Therefore, at a single station, the sets of individual measurements from this study and the Liu study might come from different events and therefore sample different back azimuths.

   Under the conditions of multi-layer seismic anisotropy, which result in variations in splitting parameters as a function of backazimuth, the two sets of station-averaged values might therefore represent a different sampling of the underlying variation of splitting parameters with backazimuth. This would impede an apples-to-apples comparison between the two sets of station-averaged values from the present study and the Liu study.
   *Note that in Liu et al., 2014, the ‘Complex Anisotropy Index’ (Figure 13 in that study) suggests that strong variation of splitting parameters with azimuth are observed at many locations in the Western/Central U.S, potentially indicative of layered anisotropy.

   Therefore, if this is not already the case, I think it would be good to ensure that the same events/backazimuths are used for both measurement studies when calculating the

station-averaged values, and thus also when calculating the statistics shown in Figure 2 c,d, as well as in Figure 5, Figure 7, S10, S11, and S12.

The authors thank Reviewer 1 for this comment; as such efforts were made to keep the method free of any requirements of prior knowledge and hence the threshold was applied for the selection of waveforms irrespective of whether or not they were used in the calculations of Liu et al., 2014. However, to get an idea about what such a comparison would look like, we performed the analysis for a subset of waveforms that were included both in the station-average calculations by Liu et al., and in the data used for SWSNet calculations. We recalculated the station averages by using just this data, and did a comparison similar to figure 5. The new camparison is added to the supplementary materials as Figure S12. As expected we find a closer alignment of the station averages in this case.

The related discussion was added to the Discussion section in lines 230-239.

2. I think that in addition to quantifying the difference between the station-averaged splitting parameters, I think that it would be instructive to quantify the differences between the -individual- measurements made for USArray station-event pairs that are common across the Liu et al., 2014 study and this study, for instance quoting the average difference between measured fast axes and measured delay times across common station-event pairs.

   We made this comparison for the waveforms in the subset mentioned above and found the mean absolute difference for φ and δt to be 11.08° and 0.239 s respectively. This information has been added to line 239-241 of the manuscript.

3. Given the common scenario of multi-layer anisotropy, I think the paper could benefit from a simple test showing that SWSnet can reproduce the apparent splitting parameters (i.e. variation with backazimuth of fast axis and delay time) for synthetic data generated by multiple layers of anisotropy. I recognize that the authors identify this as an avenue for future work, but I feel like a successful result here would strengthen the case for using SWSnet immensely, especially since the paper already applies it to real data.

   The authors appreciate this suggestion. We agree that the paper benefits greatly from such an analysis. We have added Figure 8 to the revised manuscript and discussed our results in lines 242-255 in the Discussion Section.

4. While the tests that compare the splitting parameter measurements with real data are clear, I wonder if there could be more exploration of the null measurements. In particular, I wonder if the paper would benefit from a simple exploration of the SWSnet predictions for the USArray measurements. For instance, what is the average non-null probability predicted by SWSNet of the fraction of the measurements that are designated as null in the Link or the Liu study, and how does this probability compare to the measurements designated as non-null in the Link or the Liu studies?

The predicted probability for non-null measurements varies between 0.53 to 1.0; it is however very skewed and has mean and median of ~1.0 each.

This is because of how the classes are distributed in the training data (there are only 4.45% null measurement), which makes a classification task very difficult for the deep learning model. We also do not apply any class-weights for this task as this classification task is not a goal of the model, it is merely used to provide more information to aid the learning.

5. The data and code availability section does not seem to show that the authors have made their code openly available. As is described on the Seismica website, the "Authors are expected to act in the spirit of Open Science and make their data/codes available when publishing with Seismica." If the authors do not wish to make their code available, I think a line justifying this would strengthen the manuscript for Seismica's audience.

   We are currently creating a github repository with the relevant codes, and we would include this in the manuscript once it is accepted for publication.

6. Unlike contemporary SWS methods that can obtain uncertainties using, for instance, the error ellipses from the grid search, I did not see any discussion of measurement uncertainty using the approach described here. I wonder if it is possible to provide some estimates of uncertainty.
   Perhaps one way this could be done for synthetic data could be to generate measurements on the same waveform but with multiple (say, hundreds) of realizations with different randomly generated noise, then explore the standard deviation of the resulting splitting parameters?
   Alternatively, since I recognize that uncertainty estimation in deep learning methods is challenging, I think that mentioning that uncertainty estimation could be an avenue for future work would strengthen the paper.
   We agree that this is a very challenging problem, and we would add this to the list of avenues we want to explore in future. We have mentioned this in our conclusion section in lines 282-284.

Minor Comments:

1. In lines 224-225, the authors comment that the grid search is on average 3-6x faster than SWSNet. However, I think that without information about the number of forward calculations, it is difficult for a reader to take much away from this point as is. I think that the point the authors are trying to make could be strengthened if they also mention the grid spacing, in fast axis and delay time, of the grid search, as well as the upper and lower bounds of the grid.

   Thank you, we have added this information in line 261-262 of the manuscript.

2. In the caption for figure S5 or somewhere else in the text, can the authors clarify what the exponential function that is alluded to in Fig. S5 is? That is, what is the argument of the exponent?

   The relevant equations have been added to lines 11-14 of the Supplementary Information.

3. Line 227: “In this study we introduce baseline a deep learning model” should be “In this study we introduce a baseline deep learning model”

   Thank you for pointing out this typo, we have corrected this (line 268).

4. I think the reference to “linketal” is not formatted correctly in line 252 of the manuscript.

   Thank you very much for pointing this out. We have rectified this error.

5. Only four thresholds for transverse energy reduction (20,40,60,80) are explored. I think the decision to pick a threshold of 60, and the analysis shown in Figure 5, could be more convincing if a finer range of thresholds are explored, e.g. every 5 or 10%.

   Thank you for the suggestion. We have updated figure 5 to explore a finer range of thresholds (every 10%) and once again we find 60% to be the most optimal threshold.

Reviewer B:

Summary:

The authors present a interesting and novel study, applying machine learning techniques to measure shear-wave splitting, which in turn can be used to quantify seismic anisotropy in the subsurface. The paper is generally well written. A particularly interesting result is that the authors find that a waveform deconvolution combined with a neural network is required to enable the method to work. Most of my suggestions relate to minor issues in order to provide greater clarity. For example, I think it would be useful to elaborate on why the authors chose the specific neural network architecture and provide a little more detail on the deconvolution method, which confused me a little. However, there is one larger issue that I would definitely like to see addressed. I very much enjoyed reading through the novel method and the well presented results, eagerly anticipating the discussion, which then is broadly absent. There is a very short discussion but that is almost an extension of the results. The discussion should be the most exciting part of the paper to write, so I’d really encourage the authors to revisit this and elaborate on the strengths and limitations of their method, and how it might be applicable/contribute to the field more widely. If/when a broader discussion exists, I would be very happy to see this work published.

Overall, an interesting and novel study that was a pleasure to read, that in my opinion would be ready to publish after adding a brief but exciting discussion of the implications of the work.

Tom Hudson

Major comments:

L102: That is quite a large delay-time. Perhaps the authors can justify why they don't explore the parameter space of 0-0.2 s, where many real-world teleseismic splitting results lie?

Thanks for appreciating our work. The parameter space of 0-0.2s of delay-time is not specifically taken into account as the low transverse energy in these cases impedes the inference of a clear fast axis during training . Upon referring to figures S2 and S3, one can observe that the effect of splitting is barely noticeable at such small delay times.

Section 2.2: reference the actual detailed network diagram (in supplementary information) somewhere here too. Also in this section, for someone with only limited ML experience, it would be really interesting to have a brief justification for some of the key choices here (perhaps just focussing on the layers rather than the activiation functions)? Finally, what exact network structure do you decide upon? How is it decided upon? A little more detail here would be beneficial.

A reference to the detailed network diagram (Figure S7) has been added in line 121-122.

We have added some more detail about the final architecture and its formulation between lines 107-110.

L158-159: Sorry, I don't follow why the radial component would represent propagation without any anisotropy? Maybe I've missed something, so apologies if so.

Without an anisotropic layer, in a radially symmetric Earth, the transverse component vanishes on the receiver-side leg of the XKS raypath due to conversion from P to S at the CMB.

For a (thin) anisotropic layer with delay time dt, such that the XKS dominant period T >> dt, a transverse component appears, while the radial component remains (almost) unchanged. This is usually the case in XKS splitting analysis (e.g., T= 8s, dt =1.0 s). For much higher-frequency waveforms (smaller T) or larger dt, the effects on the radial component waveforms would be more significant, and our assumption would not be applicable.

Figure 4: Perhaps I'm wrong, but doesn't Figure 4 look worse than Figure 2? I guess the distribution is different, but it looks like there is a larger spread in phi, although dt looks perhaps better? Maybe just consider clarifying why the performance of the network that produces Figure 4 data is better than that of Figure 2?

The performance on the synthetic data in Figure 4 is indeed worse than in Figure 2, this is because a major difference in the deconvolution approach as compared to the method discussed in section 2.2 is that when we train the model on the deconvolved data, only the transverse component carries the relevant information and hence the model is only trained on this component as opposed to the previous method where both the radial and transverse components were used; using two components might help the model learn the noise characteristics in the data resulting in a smaller spread in the predicted parameters. However, despite this deterioration in the performance on the synthetic test data, the use of the deconvolution method, leads to a much better generalizability when applied to real world data, as one can see in the subsequent discussion and in Figure 7.

We have added this discussion to lines 213-220 in the revised manuscript.

Figure 5: Why does fast angle always improve, but dt doesn't? (assuming red and blue correspond to phi and dt)

While imposing a threshold on the reduction of energy in the transverse component ensures a good performance of the model, it might sometimes be the case that for certain contaminated waveforms, even though the energy reduction is above the cut-off given the recorded initial transverse component, the predicted parameters do not represent the ground truth. Imposing a higher threshold, however, always reduces the number of waveforms that satisfy the cut-off. As seen from figure S13 of the revised manuscript the robustness of the calculated station-averages increases with a higher number of measurements corresponding to a station. Due to this trade-off, increasing the threshold does not always guarantee an improvement in the station-averaged calculations of either splitting parameter. This effect can be seen more prominently in Figure 5 of the revised manuscript, where a finer binning is used for the threshold.

Discussion is rather short! I was really enjoying reading the paper and looking forward to an exciting discussion, but it feels rather absent! Even the discussion that is given reads more like results. There is some discussion in the results, but I'd like to see some elaboration on potential performance and impact to the field. This section should be the most exciting part of the paper to write, so I'd encourage the authors to add a little more.

In the revised version of the manuscript the Discussion section has been made much more elaborate by incorporating the suggestions of the reviewers regarding further analysis of our method.

Minor comments:

L49: Feel free to ignore this comment, but we recently developed a fully automated SWS method implemented in python (Hudson et al., 2023). I only mention this for completeness in your list, but as a reviewer I certainly would not insist on you citing it too, but if you were tempted to then great.

Thank you for the reference, we have added this to lines 47-48 of the revised manuscript.

L62: Grammar issue, perhaps remove "amount of"?

Thank you. We have incorporated this suggestion in the manuscript.

L76: mention that phi is clockwise from North (if it is).

Thank you. This suggestion has been incorporated (line 76).

L82: I've never thought about rotations in the frequency domain, but I don't quite follow how it works? It would be nice to explain why you work in the frequency rather than time domain. Sorry for my ignorance. Also, do you then move into the time-domain for analysis?

The equations are given in the frequency domain as the application of the convolution step becomes much easier to describe (avoiding convolution symbols). However, the switch from the time domain to the frequency domain only applies to the waveform (which is time dependent) but not to the coordinate directions which are defined by (e.g.) the fast and the slow axis within the anisotropic layer. In the code, the deconvolution step is also performed in the frequency domain. The splitting analysis is performed afterwards, after back-transform into the time domain.

I really like Figures S2,S3!

Thank you.

A perhaps unconventional structure to have methods and results sections together. I'd prefer them split for clarity, but if the authors are really wed to their current structure then I am happy for them to ignore this suggestion.

We intentionally went for this structure as, for this study, the method and results are very much intertwined, the method has been developed continuously while referring to our observations with the results. This led us to adopt this particular structure for our paper.

L124-130: How many samples are in the training dataset? How many in a verification dataset?

A total of 1,000,000 waveforms are used for the training process; this dataset is split in a ratio of 80:20 for training and validation purposes. This information has been included in lines 103-104 of the revised manuscript.

L149: Is it impossible, or computationally not feasible? Theoretically, one could run numerical propagation models with varying source mechanism, attenuation and velocity structure and train from there.

We agree that "impossible" is a very strong word, so we have replaced it with "computationally not feasible". Please check lines 153-154 of the revised manuscript.

L153: Why? A brief justification of why both components are deconvolved by the radial here would be useful. Also, if the radial component is deconvolved by the radial component, doesn't the signal become 1?

Discussed in the next comment.

Eq. 6,7: Ah, this clarifies my last point. Small mistake in Eq. 6 though? Shouldn't numerator be u_0 not u_1? If so, then "=1" should be removed. Also, in Eq. 7, shouldn't it be u_0^(t) rather than u_0^(r)? If so, then please also correct in L158.

It should be u_1 because, this is the information that is available to us, i.e. the components after splitting; this is the waveform that is finally recorded at the station. Please also refer to eq. 5 for derivation of eq 7.

L160: Avoid words like relatively unless this is quantified (e.g. relative to what? What is a long period here?).

Long here refers to periods much longer than the delay time (T >> dt). This is added to line 165 of the revised manuscript.

Figure order looks like it doesn't correspond to order of figures referenced in the text. Consider checking this and reordering appropriately (e.g. Figure 6 appears to be referenced before 5?)

The description of the figure is only given later, while in lines 169 and 180 we see a secondary reference. We have modified the reference to reflect this.

L199: Consider a different word to "seems". Perhaps "is defined as optimum here" (as it corresponds to minimum, see Figure 5)?

As per this recommendation we have made some changes in lines 204-206 and also in the caption of figure 5.

L204: Make it clear that SplitRacer is the method used in Link et al. (2022).

We have made a modification in line 210 of the revised manuscript.

Figure 5: Needs some more labelling. I don't follow what red and blue data correspond to.

We thank you for pointing this out. A legend has been added to figure 5.

L223: "quite close"? How close? Please quantify.

This has been already visualized by Figure S13.