### **View Reviews**

**Paper ID**

2689

**Paper Title**

Bayesian Neural Networks for Uncertainty Estimation of Imaging Biomarkers

#### **REBUTTAL GUIDE (from MICCAI website)**

Your rebuttal is addressed to the Area Chairs only. Reviewers will not see it and will not be able to change their reviews.

The goal of the rebuttal is to inform the Area Chairs of major misunderstandings, in your opinion, in the reviewers' assessment, or of incorrect statements in the reviews. An effective rebuttal **focuses only on major critiques**. It is not helpful to try to address every minor point in the reviews. By prioritizing and focusing on the major concerns, and by **grouping multiple reviewer comments** that generally pertain to the same issue into a **few major categories,** you are demonstrating to the Area Chairs that you understand the high-level messages that were provided in the reviews.

Please **summarize or rephrase the criticism before you address it**, and clarify to which comment(s) you are responding. While the room for rebuttal is limited, if properly utilized by condensing the response down to the essentials, this is an effective way to let the Area Chairs know that you understood the reviewer's concerns and have valid answers to the questions raised in the reviews, or to establish that certain reviewer comments were false or unsubstantiated.

An effective rebuttal addresses reviewers' criticisms by explaining where in the paper you had provided the requisite information, perhaps further clarifying it.

Do not promise to expand your paper to address all the questions raised by the reviewers, as you will not be able to change your article substantially, and in all likelihood you don't have sufficient room to add to the paper. These promises are likely not to be taken seriously.

A good rebuttal is polite; being confrontational does not bring any added value to the paper. However, if you feel you have received a review that was not courteous, or made false or unsubstantiated arguments that you can succinctly refute, you should point this out.

Christian:

We thank the reviewers for the constructive critique. They consider our paper “a good idea of [..] integrating segmentation uncertainty in follow-up tasks” and that the “comparison of different uncertainty generation methods is applaudable”. Finally, our “results clearly demonstrate the effectiveness of the proposed method”. We address the points summarized by the AC in the following.

NOVELTY and RELATED WORK

R1 raises questions about the novelty of our work. Our article is about the integration of segmentation uncertainty in follow-up analyses. We do not claim that we are the first to consider this problem, but that it “has not yet been well studied”. In [14], uncertainty from MC-Dropout has been used for instance reweighting in group analyses. In our article, we take a more principled approach to this problem by:

- introducing the first approach for disease classification with uncertainty,

- proposing novel statistical models for the integration of uncertainty in group analyses,

- evaluating 4 different Bayesian neural networks for estimating the uncertainty.

We thank R1 for referring us to Mehta et al., MICCAI UNSURE, 2019. We would be happy to include it in a final version, but we would also like to emphasize that it does not impact the novelty of our work. First, they use uncertainty maps in a second neural network that again works in the image space, in contrast to the extraction of biomarkers in our work. Methods and application are therefore completely different. Second, the paper only uses MC dropout, while we compare 4 different uncertainty networks.

STATISTICAL SIGNIFICANCE

R1 mentions that our “results demonstrate clear usefulness of uncertainty in increasing diabetes classification accuracy”, but that we have not performed significance testing. Following the suggestion, we have performed the Friedman test with a pairwise Nemenyi post hoc test to assess statistical significance between disease prediction models at significance level 0.05. We note a significant improvement concerning the base model for all uncertainty methods with IoU in combination with Variable and Interaction. Further, MC-Dropout (Variable IoU) and probabilistic U-Net (Interaction IoU) significantly improve over the manual model. These results could be easily integrated into the final version of Table 2.

AVERAGING PREDICTIONS

R2 questions whether averaging the segmentation samples to obtain the final segmentation for computing the liver volume makes sense for the probabilistic and hierarchical U-Net. The reviewer is correct that both approaches were developed to generate diverse segmentations to model different raters. However, as discussed in the paper, we only have annotations from a single rater, so that the networks do not output segmentations that are structurally highly varying (different modes). Hence, computing the average is reasonable in this scenario.

RANKING

As this is the first comparison of 4 different techniques for uncertainty estimation, R1 asked about their ranking. For regression, we see the best performance for MC-Dropout with IoU and Variable. For classification, this setting has the second-best result, only topped by the probabilistic U-Net with Interaction. This difference can be attributed to the Interaction term in the classification setting, which cannot be used in regression as the volume is the dependent variable.

ALEATORIC - EPISTEMIC

We agree with R3 that the different Bayesian networks either focus on epistemic or aleatoric uncertainty. Our methods do not rely on a specific form of uncertainty, in fact, one of our main contributions is to investigate which type is better suited for follow-up analyses. Moreover, we agree that results might improve by calibrating the aleatoric uncertainty, but we only included methods in their original implementation, as we were interested in integrating the uncertainty of existing methods into follow-up analysis; not to introduce a new method for uncertainty estimation.

OLD:

This discrepancy can be explained by the inclusion of the Interaction term in the classification setting, which cannot be used in the regression setting as the volume is the dependent variable.

Following the suggestion, we have computed the McNemar test for the classification results. We have noted a significant improvement with respect to the base method for MC Dropout and Fully-Bayesian. These results could be easily integrated in the final version by adding (\*) to Table 2.

Currently, bold results indicate the best performing method significantly outperforming all others, except Manual, which we excluded to restrict the comparison to automated methods. Hence, Variable significantly outperformed Interaction for Hierarchical and MC Dropout, and Interaction outperformed Variable for Probabilistic. The performance of IoU variants was always significantly better than its CV-1 variant, except for Instance, which was not better than Base for Probabilistic and Hierarchical. Regarding Manual, Variable significantly outperformed Manual for all but Fully-Bayesian.

These results could be easily integrated in the final version by adding (\*) to Table 2.

* FB: Manual > Variable\_IoU = Interaction\_IoU > … > Instance > Base
* HQ: Variable\_IoU = Manual > Interaction\_IoU = Manual > everything else equal
* MC: Variable\_IoU > Interaction = Manual > … > Instance > Base
* PB: Interaction\_IoU > Variable\_IoU > Manual > … > Base

**Reviews:**

**Meta Reviewer:**

The reviewers agree that this paper lacks novelty and misses some essential references to previous work**.** The authors should comment on statistical significance, why averaging prediction is the right approach and how the different methods rank.

We are highly thankful to you for providing us your valuable feedback. We are also saying thanks to other reviewers who put their invaluable time to evaluate our paper.

We are thankful to all good words put by you and also trying to understand the limitations and short sightings in our work that have been brought to light by you.

**Lack of novelty and references:**

We are glad and thankful to you for providing this feedback. Though this is a good one, we are frightened to say that the reviewer has mistaken our points here.

Reviewer 1 pointed out that “ the uncertainty effect in group analysis is done in Bayesian QuickNAT paper [1], which authors cite but don't mention. Similarly, propagating uncertainty from one task to the follow up task has also been done before [2]”. We agree on this point, but In the paper, we mentioned “the integration of the segmentation uncertainty into follow-up analyses of extracted biomarkers, such as **group analyses** or **disease classification**, has not yet been **well studied**”

We analysed the effect of various uncertainty with other basic but important non-imaging biomarkers which can predict the target(Disease) by themselves with a good accuracy, we were willing to see how the inclusion of imaging biomarkers and uncertainty affect that particular setup, which in our knowledge was not tested before. \cite{Mehta et. al.} uses uncertainty maps in imaging setup and tries to improve the followup analysis. \cite{Roy et al.} shows the improvement in group analysis using predicted volume and respective uncertainty where he showed there is a significant improvement in the relation between organ volume only and autism if uncertainty is included.

**Statistical significance:**

We would like to convey our thanks for the feedback and happy to examine and explain the statistical significance of our results in the paper.

Reviewer 1 and 3 pointed out unavailability of statistical significance results in the paper. We are sorry for not able to put these results in the paper. There are 2 reasons.

1. From the table 1 below, everything looks significant, so we chose regression coefficient values to put in the paper instead.
2. We could have put both the results in, but that was not possible due to space constraint.

Now as in the table, albeit everything looks significant. We can compare that the relationship of manual volume with diabetes state is highly significant i.e. an upper limit. Where as predicted volumes hold imperfections and thus the significant degrades in base case(Comparison with predicted volume only). Now inclusion of various uncertainty improves the significant score and tries to attain the upper limit i.e. it improves the overall relationship between predicted volume and diabetes status which we tried in the paper by showing reg. Coefficients.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diabetes Status P-Values | | Variable | | Instance Weights | |  |
|  | Base | IoU | CV Inverse | IoU | CV Inverse | Manual |
| MC Dropout | 0.00048 | 0.00035 | 0.00050 | **0.00027** | 0.00181 |  |
| Bayesian | 0.00644 | 0.00381 | **0.00372** | 0.00430 | 0.00660 |  |
| Probabilistic | 0.00208 | **0.00120** | 0.00150 | 0.00209 | 0.03327 | 0.00015 |
| Hierarchical | 0.00089 | **0.00055** | 0.00112 | 0.00083 | 0.00294 |  |

**Averaging prediction:**

We are thankful to you to bring us this concern of yours, We would be glad to enlighten the topic here. As reviewer 2 and 3 says their concern about using averaging method for method #3 and #4 in classification tasks, which does not look good.

Yes, they are partially correct. Actually we did not calculate scores for method #3 and #4 considering average of N inferenced samples as final prediction, rather we use each sample to evaluate method #3 and #4 to get a spread of classification results and after N sample prediction classification results, we took an average.

* We did not find much variation evaluating either way.. So We did not mention it in the paper. Both ways the results are almost the same or differ very slightly.

**Ranking of methods:**

Ranking of different methods in the paper is an important virtue of our work and we are glad to convey our thanks to you for bringing this point to us. Reviewer one has concern about not having any clear winner, whereas reviewer 2 and 3 are not so sure about comparing aleatoric and epistemic methods together. For concrete evidence, it is better if we compare methods from different domains like aleatoric and epistemic. Regarding aleatoric uncertainty in probabilistic and hierarchical, we have segmentations from a single grader, hence should not have much aleatoric uncertainty added to the model(Epistemic) uncertainty. Additionally, at the end of uncertainty measurement, we had two different ranges of uncertainty whereas the distribution of uncertainties of MC Dropout and Full-bayesian volumes are more or less gaussian like and distribution of Probabilistic and Hierarchical are highly confident in nature. ~~We do calibrated final uncertainties into the range of 0-1.~~

* We compare different methods not based on the segmentation outcome, but the results from statistical analysis with uncertainty measures included.
* In Spite of having a **different spread in uncertainty**, the results in classification tasks are quite competitive in between MC Dropout and Probabilistic i.e. it is not biased or one sided. Which says irrespective of uncertainty measures from different domains, using them in followup analysis as per in the paper improves the overall score.
* Reviewer 3 also pointed out usage of CV\_inverse might have affected other features in group analysis task, which we did not see in our results.[See Table 2]. We found that the difference between IoU and CV\_inv is significant but for other features it is more or less the same.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MC Dropout |  | Probabilistic |  | Full Bayesian |  | Hierarchical |  |
| Reg.Coefficient(Beta\_5) | CV\_INV | IOU | CV\_INV | IOU | CV\_INV | IOU | CV\_INV | IOU |
| Intercept | -1.18E+00 | -2.01E+00 | -1.21E+00 | -2.15E+01 | -1.17E+00 | -2.17E+00 | -1.36E+00 | -9.97E+00 |
| Age | -3.60E-02 | -3.53E-02 | -3.47E-02 | -3.40E-02 | -3.80E-02 | -3.81E-02 | -3.56E-02 | -3.60E-02 |
| Sex | 6.33E-01 | 6.28E-01 | 6.69E-01 | 6.70E-01 | 5.35E-01 | 5.48E-01 | 5.98E-01 | 6.23E-01 |
| BMI | 9.40E-02 | 9.48E-02 | 8.62E-02 | 9.06E-02 | 9.32E-02 | 9.86E-02 | 9.66E-02 | 9.81E-02 |
| Diabetes\_status | 3.08E-01 | 3.18E-01 | 2.97E-01 | 3.02E-01 | 2.71E-01 | 2.69E-01 | 2.88E-01 | 3.06E-01 |
| IoU |  | 8.85E-01 |  | 2.05E+01 |  | 1.26E+00 |  | 8.79E+00 |
| CV\_Inv | 4.40E-05 |  | 5.47E-05 |  | 6.08E-04 |  | 2.82E-05 |  |

So Considering all above points, we understand that our way of comparing different methods is ~~properly calibrated and~~ not biased or ill advised.

And hence, in group analysis, MC\_Dropout wins and from classification tasks, MC Dropout and Probabilistic are close to each other. But overly We would like to recommend MC Dropout for segmentation task with IoU as uncertainty measures to include in the followup analysis

##### **Reviewer #1**

#### **Questions**

* **2. Please provide a summary of the paper (a few lines)**
  + The paper investigates four different segmentation uncertainty generation methods for the follow up group analysis and disease classification. The results show that integrating segmentation uncertainty can lead to higher classification accuracy.
* **3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)**
  + + A good idea of investigating the effect of integrating segmentation uncertainty in follow-up tasks.  
    + Comparison of different uncertainty generation methods is applaudable.  
    + Results clearly demonstrate the effectiveness of the proposed method.
* **4. Please list the major weaknesses of the paper (bulleted list).**
  + - Authors state that the integration of uncertainty in group analysis and follow up tasks has not been done before. This is clearly false as the uncertainty effect in group analysis is done in Bayesian QuickNAT paper [1], which authors cite but don't mention. Similarly, propagating uncertainty from one task to the follow up task has also been done before [2]. This makes the novelty of the proposed method questionable.
  + Ans: In the paper, we mentioned “the integration of the segmentation uncertainty into follow-up analyses of extracted biomarkers, such as group analyses or disease classification, has not yet been **well studied**”
  + We analysed the effect of various uncertainty with other basic but strong non-imaging biomarkers which can predict the target(Disease) by themselves with a good accuracy, we were willing to see if inclusion of uncertainty improves that particular setup which in our knowledge was not tested before. \cite{Mehta et. al.} uses uncertainty maps in imaging setup and tries to improve the followup analysis. \cite{Roy et al.} shows the improvement in group analysis using predicted volume and respective uncertainty where he showed there is a significant improvement in the relation between organ volume and autism if uncertainty is included.
  + - It is not clear, why coefficient values of Beta\_4, which is related to diabetes status, if reported in the Table:1 instead of coefficient values related to uncertainty measures like IoU or CV^-1, which would reflect the usefulness of them in the group analysis.
  + Ans: ~~Coefficient values of uncertainty measure for group analysis is not available for all the methods that we have discussed like instance weighting. So we rather took diabetes coefficient value for a better comparison where we will have coefficients for all methods we proposed.~~
  + we have diabetes status as ground truth available to compare with and it is what we actually care about.
  + The reg. coefficient of uncertainty measures would have also been interesting, but due to page limitations we couldn't include it.  
      
    [1] Roy et al. "Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control." NeuroImage 2019  
    [2] Mehta et al. "Propagating Uncertainty Across Cascaded Medical Imaging Tasks for Improved Deep Learning Inference." UNSURE, MICCAI 2019.
  + <http://cim.mcgill.ca/~raghav/papers/UNSURE_2019.pdf>
* **5. Please rate the clarity and organization of this paper**
  + Good
* **6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html**
  + -- It would be good, if authors can clarify if segmentation network was also re-trained 1000 times, similar to the process mentioned in sec:3.4
  + Ans: No, Segmentation networks were trained once with 50-60 epochs each. Then we inference 10 samples and calculate uncertainty measures.  
    -- Although, results demonstrate clear usefulness of uncertainty in increasing diabetes classification accuracy, there is no statistical significance analysis done for the same.

**TABLE 1:** p-Values of McNemar test with Base Method (Diabetes = a\_1 + b\_1\*Predicted\_Volume) (IoU only)

variable instance interactions

MC\_Dropout 0.522397 0.031250 0.555998 # Comparatively Uncertain

FullBayesian 0.001496 0.015625 0.049800 # Highly Uncertain

Probabilistic 0.111161 1.000000 0.542610 # Less Uncertain

Hierarchical 1.000000 1.000000 0.271679 # Less Uncertain

**TABLE 2:** P-Values of McNemar test referenced to Manual Model (Diabetes = a\_1 + b\_1\*Manual\_Volume) (IoU only)

[ For reference: In McNemar test, the Null Hypothesis is that there is no significance difference between compared methods. ]

Models base variable instance interactions

MC\_Dropout 0.061428 0.440799 1.430906e-03 0.050452

FullBayesian 0.000001 0.245061 4.461782e-10 0.063950

Probabilistic 0.033552 1.000000 3.355244e-02 0.529602

Hierarchical 0.042774 0.053252 2.944937e-02 0.009853

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diabetes Status P-Values | | Variable | | Instance Weights | |  |
|  | Base | IoU | CV Inverse | IoU | CV Inverse | Manual |
| MC Dropout | 0.00048 | 0.00035 | 0.00050 | **0.00027** | 0.00181 |  |
| Bayesian | 0.00644 | 0.00381 | **0.00372** | 0.00430 | 0.00660 |  |
| Probabilistic | 0.00208 | **0.00120** | 0.00150 | 0.00209 | 0.03327 | 0.00015 |
| Hierarchical | 0.00089 | **0.00055** | 0.00112 | 0.00083 | 0.00294 |  |

* + The table shows p-value score of diabetes status in a group analysis task.
  + Manual p-values show the upper limit. Whereas p-values related to inferenced segmentation degrades the significance. inclusion of uncertainty measures helps improve the relation and attain the upper limit .  
    -- There is no clear winner among different uncertainty generation methods. Some explanation for the same would be appreciated.
  + In group analysis tasks, MC dropout is a clear winner as it is much closer to the upper limit set by group analysis with Manual segmentations.
  + In the Classification task, MC dropout and Probabilistic U-Net are doing good compared to the other two. Overally, MC dropout performs best as per our result.

##### **Reviewer #2**

#### **Questions**

* **2. Please provide a summary of the paper (a few lines)**
  + This paper proposes the use of uncertainty estimates for the biomarker analysis task. They specifically use four probabilistic methods to estimate the segmentation uncertainty and use the associated confidence measure for the group analysis and disease classification.
* **3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)**
  + Trying to make use of the prediction uncertainty to improve the disease classification task is the only strength of the paper to me.
* **4. Please list the major weaknesses of the paper (bulleted list).**
  + - Methods #3 and #4 don't look suitable for the task in hand  
    - It is not clear how the uncertainty maps are generated
  + Ans: We have 10 samples per volume inferenced from each model. We then calculated the sum of pixel entropy across samples to get an uncertainty map.
  + Uncertainty Map: -)  
    - The method employed for making use of uncertainty estimations in the disease classification is basic
* **5. Please rate the clarity and organization of this paper**
  + Good
* **6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html**
  + Comparing the first two methods (MC dropout and Full Bayesian) with the last two methods in generating reliable uncertainty maps and segmentation predictions are not valid as the last two methods are not designed for that purpose (as the authors also admit in page 7). Instead, they could have used methods such as Model Ensembles or MC-Dropconnect or M-heads for that purpose.
  + -On the contrary, it is good to know how models from epistemic and aleatory domain response with their set of uncertainty measures in the follow up analysis.
  + - From comparing model perspective, We are comparing results from the followup analysis where segmentation models from different domains like aleatoric and epistemic with separate **range of uncertainty** outputs included in the proposed methods to give the best estimate. From the final outcome, we are comparing methods where MC dropout from epistemic and Probabilistic from aleatoric works better in group analysis and disease classification respectively.

- Moreover, we have single grader segmentation data which has been used to train models. That eradicates the possibility of data randomness incorporated into aleatoric models like probabilistic and hierarchical and left with least amount of model uncertainty(epistemic) as trained which can be seen in Fig 3. That also gives high confidence and less uncertainty for both hierarchical and probabilistic models. Whereas the range of confidence measured in MC dropout and Full Bayesian is different. But when compared in followup analysis, Not a particular group of models come victorious, but MC dropout and Probabilistic. So Our method suggests that irrespective of methods, inclusion of segmentation uncertainty improves the overall performance in follow up analysis, in our case, group analysis and disease classification.

-----ANSWER VVI

----- TODO

* + It is also unclear how they generated the uncertainty maps. Is it by computing the standard deviation over the predicted samples? How many samples were generated in each case? For example, for the MC Dropout method, how many rounds of MC simulations were used?
  + -10 samples per volume per segmentation method.
  + - Uncertainty Map: -)  
      
      
    In any case, averaging the predictions for one sample and considering it as the prediction and use it for the next steps does not sound right to me for methods #3 and #4 as they initially were developed and proposed to generate a set of diverse but plausible segmentations.
  + - Yes, you are correct. Actually we did not calculate scores for method #3 and #4 considering mean of N inferenced samples as final prediction, rather we use each sample to evaluate method #3 and #4 to get a spread of classification results and after N sample prediction, we took an average.
  + We did not find much variation from taking N sample average as prediction and classifying compared to N time classification and averaging. So We did not mention it in the paper.
  + --- TODO--   
      
    Performance of the probabilistic methods are evaluated using dice score. While dice score is commonly used in the literature for comparing a deterministic prediction with a unique ground truth, it does not tell much about the distribution of segmentations and the quality of the generated uncertainty maps.
  + ------   
      
    Replace Figure 2 with tables as the numbers are very close and not readable. Adding statistical analysis can illustrate more about the significance of the differences.
  + ---- Not Doing--  
      
    Equation (6) is unclear to me. It is not clear how this formulation gives higher importance to the samples with more confident predictions. Adding a couple of sentences on that or correcting the possible typos can help to clarify it.
  + -- While training a statistical model with instance weighting(Eq 6), Higher confidence prediction(C\_i) will give rise to high value to loss function and that would be back propagated to optimize the model. The high confidence score will enhance the loss score. Whereas the low confidence score will minimise the loss score and hence contribute less for the model optimization, which is desirable as we don’t want some faulty segmentations to take a major part in follow up model training and optimization..

##### **Reviewer #3**

#### **Questions**

* **2. Please provide a summary of the paper (a few lines)**
  + The paper proposes the use of image segmentation to extract metrics that can serve as biomarkers for predicting diseases. However, to combat uncertainties within predictions that may negatively affect biomarker analysis, the paper propagates segmentation uncertainty to aid in better group analysis and predictions. Four varieties of BNNs are used for segmentation and experiments show the effect of uncertainties in statistical inference.
* **3. Please list the major strengths of the paper; you should write about a novel formulation, an original way to use data, demonstration of clinical feasibility, a novel application, or anything else that is a strong aspect of this work (bulleted list)**
  + The paper compares a variety of methods for segmentation and shows that uncertainty can be used to improve accuracy in the subsequent statistical analysis.
* **4. Please list the major weaknesses of the paper (bulleted list).**
  + The differences between aleatoric and epistemic uncertainties are seriously misunderstood in the paper, and no justification has been shown for comparing both the uncertainties (more details in comments).
  + ----LOOK ANSWER VVI
* **5. Please rate the clarity and organization of this paper**
  + Good
* **6. Please provide detailed and constructive comments for the authors. Please also refer to our Reviewer's guide on what makes a good review: https://miccai2020.org/en/REVIEWER-GUIDELINES.html**
  + “image segmentation [7,10,11] have been developed that do not only provide the mode” - A BNN usually predicts the parameters of the posterior distribution. In most cases like the Gaussian one of the parameters happens to be the mean (and the mode) but in general the NN outputs the parameters of the distribution and not necessarily the mode. Please change the mode to parameters of an underlying distribution.
  + -- Parameters of distribution contain a mean, we refer to the mean of the distribution which is taken as predicted output.  
      
    Out of four of the networks, the first two use epistemic uncertainty (stochasticity in model parameters) and the other two use aleatoric uncertainty (output is a probability distribution). Is the comparison of models with these kinds of uncertainty justified? In fact, Figure 3 shows that the 3rd and 4th models are overconfident in their uncertainty estimation. This is indeed a problem that occurs due to miscalibration in neural networks and their uncertainty [1] [2] and comparing epistemic with aleatoric uncertainties is not justified since they have semantically different meanings. How do these uncertainties affect the final prediction? Fig 2(a) seems to show that MC Dropout performs the best in terms of segmentation, but all four methods have similar results on prediction of liver volume, suggesting that the confidence value predicted by the NN is not informative.
  + - We calculate uncertainty from N samples and not from Volumes!  
    The paper has completely missed out on the point that aleatory uncertainties need to be calibrated, especially in the context of classification (segmentation) where the outputs are parameters of a categorical distribution.
  + - ~~We have uncertainty scores from all the models calibrated and within the range of 0-1 always.~~
  + **Abhijit:** We thank the reviewer for the comment. We agree to the fact that the different methods used estimate the epistemic and aleatoric uncertainty. It must be noted that the aim of this paper is not to improve or propose a new uncertainty estimation strategy. Rather to explore how these approaches which are already proposed in the literature translate for followup group analysis. We agree that results might improve by calibrating the aleatoric uncertainty by using standard techniques like ensembles or Platt's scaling but we restricted the implementations consistent to the original literature. We show how the group analysis changes if we translate either the epistemic uncertainty or the aleatoric uncertainty and present these interesting findings in the manuscript.   
      
    “The probabilistic and hierarchical models were designed to learn annotations from  
    multiple raters, while we only have annotations from a single rater, which may explain the lower stochasticity of these models in our experiments.” This is partially correct, but misleading. If that is the case, one might ask how do MC dropout and Full Bayesian output more meaningful uncertainties. The authors have misinterpreted the sources of uncertainty and how they should (not) be compared.
  + -- I think, as higher stochasticity is not available in aleatoric models the uncertainty is less compared to epistemic models. Whereas data randomness is equal for all and hence does not add up much into aleatoric model uncertainty. Then We have a different range of uncertainty where aleatoric models are highly confident and epistemics are not. After follow up statistical analysis these variations in uncertainty range get normalised and suggested methods incorporate them well which allows probabilistic and MC dropout quite competitive to each other in disease classification tasks.   
      
    The use of CV-1 as a confidence metric is fine, but using it in group analysis sounds a bit problematic to me. This is because CV-1 is nothing but (mean volume from different outputs) / (std. volume by outputs). Neural nets can make systematic mistakes, and if the std is more or less similar for different images, then the regression coefficient \beta\_5 can ideally be equal to std and all other coefficients can be zero leading to a good fit in terms of MSE error but zeroes out all other \beta which provides no information at all.[Not happening as per the table data below] However, in a more realistic scenario, the use of C = CV-1 can still interfere with the actual values of other \beta because this input feature is equal to the volume of the liver multiplied with some value and the output variable is also the volume.[However, I see more or less equal beta values for other features in IOU and CV\_INV cases. Whereas IOU is having a distinctively higher reg coefficient than CV\_INV which is neutralised by similarly high but opposite polarity value in intercept. But it does not affect reg. Coefficient of other features.]  
    This may also explain why IoU performs better than CV-1 in all experiments. Overall, it seems like CV-1 is a bad design choice given the inputs and outputs of the group analysis.
  + - We will try to incorporate other uncertainty methods in future work.
  + - However, in the below table, we have regression coefficients for CV\_INV and IOU from the Variable method.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MC Dropout |  | Probabilistic |  | Full Bayesian |  | Hierarchical |  |
| Reg.Coefficient(Beta\_5) | CV\_INV | IOU | CV\_INV | IOU | CV\_INV | IOU | CV\_INV | IOU |
| Intercept | -1.18E+00 | -2.01E+00 | -1.21E+00 | -2.15E+01 | -1.17E+00 | -2.17E+00 | -1.36E+00 | -9.97E+00 |
| Age | -3.60E-02 | -3.53E-02 | -3.47E-02 | -3.40E-02 | -3.80E-02 | -3.81E-02 | -3.56E-02 | -3.60E-02 |
| Sex | 6.33E-01 | 6.28E-01 | 6.69E-01 | 6.70E-01 | 5.35E-01 | 5.48E-01 | 5.98E-01 | 6.23E-01 |
| BMI | 9.40E-02 | 9.48E-02 | 8.62E-02 | 9.06E-02 | 9.32E-02 | 9.86E-02 | 9.66E-02 | 9.81E-02 |
| Diabetes\_status | 3.08E-01 | 3.18E-01 | 2.97E-01 | 3.02E-01 | 2.71E-01 | 2.69E-01 | 2.88E-01 | 3.06E-01 |
| IoU |  | 8.85E-01 |  | 2.05E+01 |  | 1.26E+00 |  | 8.79E+00 |
| CV\_Inv | 4.40E-05 |  | 5.47E-05 |  | 6.08E-04 |  | 2.82E-05 |  |

* + To check this, the paper can also show the coefficient \beta\_5 for the “Variable” method and analyse if there are significant differences in the values assigned to them for both confidences. [Yes, there are significant differences in value assigned to \beta\_5 of different uncertainty measures, but that does not affect other features. It affects the intercept. Which is okay?? ]  
      
    Also, is volume supposed to linearly depend with the confidence value? I expected the “Instance” method to work better but it does not. Is there more analysis as to why this happens?[Instance methods impose imaging confidence measures on all other features and if the confidence is low, it hampers the predictability of raw non-imaging features as well. Hence, Instance methods did not work well. However we thought of Interaction methods in classification task to avoid this very issue. ] If there are definite trends (like bigger volumes have more uncertainties)[Often, bigger volumes have high certainty/confidence scores.] then the performance of “Variable” can be justified.  
      
    What is the “ground truth” for the volume used in group analysis? Is it computed from the expert segmentation or is present independently? This is not clear from the paper.
  + -- There are base case scores where we used manual segmentation volumes as ground truth for group analysis. For other cases like base, instance, variable, we used inferenced segmentation from the segmentation model as ground truth / target.  
      
      
    [1] Guo, Chuan, et al. "On calibration of modern neural networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.  
    <https://arxiv.org/pdf/1706.04599.pdf>  
    [2] Jena, R., & Awate, S. P. (2019, June). A Bayesian Neural Net to Segment Images with Uncertainty Estimates and Good Calibration. In International Conference on Information Processing in Medical Imaging (pp. 3-15). Springer, Cham.
  + <https://rohitrango.github.io/data/ipmi2019.pdf>
* TODO: <https://machinelearningmastery.com/mcnemars-test-for-machine-learning/>