Uncertainty-Aware Boosted Ensembling in Multi-Modal Settings

Utkarsh Sarawgi [*1] **Rishab Khincha** [*1 2] **Wazeer Zulfikar** [*3] **Satrajit Ghosh** [3] **Pattie Maes** [1]

Abstract

Reliability of machine learning (ML) systems is crucial in safety-critical applications such as healthcare, and uncertainty estimation is a widely researched method to highlight the confidence of ML systems in deployment. Sequential and parallel ensemble techniques have shown improved performance of ML systems in multi-modal settings by leveraging the feature sets together. We propose an uncertainty-aware boosting technique for multi-modal ensembling in order to focus on the data points with higher associated uncertainty estimates, rather than the ones with higher loss values. We evaluate this method on healthcare tasks related to Dementia and Parkinson's disease which involve real-world multi-modal speech and text data, wherein our method shows an improved performance. Additional analysis suggests that introducing uncertainty-awareness into the boosted ensembles decreases the overall entropy of the system, making it more robust to heteroscedasticity in the data, as well as better calibrating each of the modalities along with high quality prediction intervals. We open-source our entire codebase at https://github.com/usarawgi911/Uncertainty-aware-boosting.

1. Introduction

Rapid developments in machine learning (ML) across a variety of tasks have advanced its deployment in real-world settings (LeCun et al., 2015). However, recent works have shown how these models are usually overconfident at predicting probability estimates representative of the true likelihood, and can lead to confident incorrect predictions (Guo et al., 2017). This is particularly detrimental in real-world domains as the distribution of the observed data may shift and eventually be very different once a model is deployed in practice, leading to models exhibiting unexpectedly poor behaviour upon deployment (D'Amour et al., 2020).

Generating confidence intervals or uncertainty estimates along with the predictions is crucial for reliable and safe deployment of machine learning systems in safety-critical settings (such as healthcare) (Amodei et al., 2016; Varshney & Alemzadeh, 2017; Kumar et al., 2019; Thiagarajan et al., 2020). It helps mitigate possible risks and biases in decision making (Gal, 2016), and can also help in designing reliable human-assisted AI systems for improved and more transparent decision making as the human experts in the process can account for the confidence measures of the models for a final decision. Numerous works have proposed a variety of both Bayesian and non-Bayesian methods to model the heteroscedasticity introduced by the stochastic data generation process for predicting the uncertainty estimates along with the neural network predictions (Gal & Ghahramani, 2016; Hernández-Lobato & Adams, 2015; Wu et al., 2018; Lee et al., 2017; Pearce et al., 2020; Izmailov et al., 2020; Osband, 2016; Lakshminarayanan et al., 2017; Dusenberry et al., 2020; Jain et al., 2020; Sarawgi et al., 2020a; Kay et al., 1999; Welling & Teh, 2011; Kendall & Gal, 2017; Shridhar et al., 2018; Snoek et al., 2019; Qiu et al., 2020). Uncertainty estimation in trees and random forests has been studied in the past, with multiple methods being proposed for both classification and regression tasks. (Duan et al., 2020; Malinin et al., 2021; Wager et al., 2014; Coulston et al., 2016; Shaker & Hüllermeier, 2020)

Our experience of the world is multi-modal; data tend to exist with multiple modalities such as images, audio, text, and more in tandem. Interpreting these signals together by designing models that can process and relate information from multiple sources can help leverage different feature sets together for better understanding and decision making (Baltrušaitis et al., 2018). Parallel and sequential techniques are widely used to improve performance by ensembling weak learners trained with a single data modality (Freund & Schapire, 1997b;a; Friedman, 2000; Chen & Guestrin, 2016; Breiman, 1996; Sarawgi et al., 2020b; Zhang & Mahadevan, 2019; Nakamura et al., 2019). Similarly, base learners trained with different input modalities can be ensembled together for performance improvements (Baltrušaitis et al., 2018; Sarawgi et al., 2020b; Zhang & Mahadevan, 2019).

[*]Equal contribution [1]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, USA [2]Department of Computer Science, BITS Pilani, Goa, India [3]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Rishab Khincha <rkhincha@mit.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

Some works have briefly discussed uncertainty estimation in multi-modal settings and ensembles (Ashukha et al., 2020; Kendall et al., 2018; Chang et al., 2017; Sarawgi et al., 2020a; Oviatt et al., 2000; Sen & Stoffa, 1996; Hill et al., 1993).

We propose a notion of uncertainty-awareness with sequentially boosted ensembling in multi-modal settings. Particularly, we design an 'uncertainty-aware boosted ensemble' for a multi-modal system where each of the base learners correspond to the different modalities. The ensemble is trained in a way such that the base learners are sequentially boosted by weighing the loss with the corresponding data point's predictive uncertainty (Section 2). We evaluate our method on multi-modal speech and text datasets on healthcare tasks using different ML models and uncertainty estimation techniques (Section 3). We perform entropy, calibration, and prediction interval analyses to highlight the significance of introducing uncertainty-awareness into the ensemble (Section 3).

The motivation is to sequentially boost the data points for which a particular base learner is more uncertain about its prediction. Multi-modal data, in nature, can be more prone to noise in particular modalities due to various reasons (such as the stochastic data generation process at the source). With uncertainty estimation, the noisy modalities will exhibit high uncertainty with the predictions. In such situations, having the base learners pay more attention to such uncertain predictions can help design a more robust ensemble learner. This mechanism decreases the overall entropy of the multi-modal system while generating uncertainty estimates, thus making it more reliable.

2. Uncertainty-Aware Boosted Ensemble

2.1. Notation and Setup

Let $\mathbf{x}$ represent the multi-modal input feature set and $y \in \mathbb{R}$ denote the real-valued label for regression. We let $\mathbf{x}^j \in \mathbb{R}^d$ represent a set of d-dimensional input features for the j^{th} modality, with $j = 1$ to k, where k is the total number of modalities.

Let $\{h^j\}_{j=1}^k$ represent the corresponding base learner for the j^{th} modality. The term base learner is just an abstraction for any learnt functions that maps an input to an output, for example, SVM, random forest, neural network, etc.

We subsequently have a training dataset $\{(\mathbf{x}_n^j, y_n)\}_{n=1}^N$ consisting of N i.i.d. samples for the j^{th} modality i.e.

$$h^j : \mathbf{x}_n^j \longrightarrow y_n \tag{1}$$

2.2. Defining Uncertainty-Aware Boosted Ensemble

We first define a 'vanilla ensemble' for a fair comparison with our proposed approach. We then define our 'uncertainty-aware ensemble' referred to as 'UA ensemble', and its variation referred to as 'UA ensemble (weighted)'. Fig. 3 (Appendix A) shows the process diagrams of vanilla ensemble, UA ensemble, and UA ensemble (weighted)).

2.2.1. Vanilla ensemble

This makes use of loss values, i.e. mean squared error (MSE) values for regression, to weight the loss function during training while sequentially boosting across the base learners. This means that the MSE values corresponding to the predictions from the j^{th} base learner are used to weight the loss function for the corresponding training samples while training the $(j+1)^{th}$ base learner. Then, the ensemble computes an average of the predictions $\{\hat{y}_{h^j}\}_{j=1}^k$ of all the (boosted) base learners for the final prediction $\hat{y}$.

2.2.2. UA ensemble

This makes use of predicted uncertainty estimates σ_{h^j} to weight the loss function during training while sequentially boosting across the base learners. This means that the uncertainty estimates σ_{h^j} corresponding to the predictions from the j^{th} base learner are used to weight the loss function for the corresponding training samples while training the $(j+1)^{th}$ base learner. Then, the ensemble computes an average of the predictions $\{\hat{y}_{h^j}\}_{j=1}^k$ of all the (boosted) base learners for the final prediction $\hat{y}$. We also experiment with another variation called UA ensemble (weighted), where we compute a weighted average of all the boosted base learners, where the weights used are the inverse of the respective predicted uncertaint estimates σ_{h^j}. (See Equation (2))

$$\hat{y}(\mathbf{x}_n) = \frac{\sum_{j=1}^k \frac{1}{\sigma_{h^j}(\mathbf{x}_n)} \hat{y}_{h^j}(\mathbf{x}_n)}{\sum_{j=1}^k \frac{1}{\sigma_{h^j}(\mathbf{x}_n)}} \tag{2}$$

Most of the previously proposed boosting methods sequentially boost across different base learners using the same set of total input features. However, UA ensembles sequentially boost through different base learners, with each base learner corresponding to a different input modality. This is because we want to best leverage each of the modality-wise features while deriving a strong multi-modal learner using individual modality-wise base-learners together. It is important to note that unlike other boosting techniques (Freund & Schapire, 1997a; Friedman, 2000; Chen & Guestrin, 2016), the base learners here need not be weak learners.

3. Experiments and Results

We test and evaluate our proposed methods on two speech and language-based multi-modal datasets in healthcare tasks related to Dementia and Parkinson's disease (refer to Appendix B for more details on the datasets). We make use of different types of machine learning models (Neural Net-

works and Random Forests) and uncertainty estimation techniques (Gaussian target distribution (Lakshminarayanan et al., 2017) and Infinitesimal Jackknife method (Wager et al., 2014)) for the two datasets.

3.1. Multi-modal Feature Extraction

For the Dementia dataset, we extract multi-modal acoustic, cognitive and linguistic features from the available speech samples and their corresponding transcripts, by using the feature engineering pipeline as developed by Sarawgi et al. (Sarawgi et al., 2020b), resulting in three input modalities, namely 'Disfluency', 'Interventions', and 'Acoustic'. (Refer to Appendix C.1 for more details).

In the Parkinson's dataset, we extract two input modalities from the available data, referring to them as the 'Amplitude' and 'Frequency' modalities. (Refer to Appendix C.2 for more details).

3.2. Model Architecture and Training

3.2.1. Dementia

For a fair comparison with the state-of-the-art, we use almost the same NN architecture as used by Sarawgi et al. (Sarawgi et al., 2020b) for each of the three input modalities. (Refer to Appendix D.1 for the exact model architecture). The target distribution is modelled as a Gaussian distribution $p_{h^j}(y_n|\mathbf{x}_n^j)$ parameterized by the mean μ_{h^j} and the standard deviation σ_{h^j}, predicted at the final layer of the models (Lakshminarayanan et al., 2017; Snoek et al., 2019; Sarawgi et al., 2020a). Each of the base learners is trained with their corresponding input modality features $\mathbf{x}^j$ and ground truth labels y using a proper scoring rule, optimizing for the negative log-likelihood (NLL) of the joint distribution. Each training run used a batch size of 32 and an Adam optimizer with a learning rate of 0.00125 to minimize the NLL.

3.2.2. Parkinson's Disease

We model the probabilistic predictive distribution $p_{h^j}(y|\mathbf{x}^j)$ using a random forest (RF) regressor with parameters h^j and a mean squared error loss function. Each of the RFs makes use of 300 decision tree estimators. The uncertainty estimates σ_{h^j} of each of the data point is estimated as the confidence interval using the Infinitesimal Jackknife method (Wager et al., 2014; Wager, 2016).

3.3. Results

For robustness, we repeat every training and test-set evaluation 5 times and report the mean and variance of the root mean squared error (RMSE) results across the five runs. We first evaluate each of the modalities (i.e. base learners) individually and then compare them with the vanilla and uncertainty-aware ensembles. The order of sequential boosting for the propagation of the uncertainties is chosen in the order of the test set performance of the individual modalities. We observe that the uncertainty-aware ensembles perform better than the vanilla ensemble and the individual modalities. (Tables 1 and 2).

Table 1. Comparison of individual modalities i.e. base learners and ensemble methods on test set results of the ADReSS dataset.

Model	RMSE
Disfluency	5.71 ± 0.39
Interventions	6.41 ± 0.53
Acoustic	6.66 ± 0.30
Vanilla Ensemble	5.17 ± 0.27
UA Ensemble	$\mathbf{5.05 \pm 0.53}$
UA Ensemble (weighted)	$\mathbf{4.96 \pm 0.49}$

Table 2. Comparison of individual modalities i.e. base learners and ensemble methods on 5-fold cross validation results of the Parkinson's Telemonitoring dataset.

Model	RMSE
Amplitude	3.21 ± 0.06
Frequency	3.32 ± 0.10
Vanilla Ensemble	3.18 ± 0.05
UA Ensemble	$\mathbf{3.04 \pm 0.04}$
UA Ensemble (weighted)	$\mathbf{3.05 \pm 0.05}$

We further use the 65-95-99.7 rule (also called the empirical rule) to obtain calibration curves for a comprehensive analysis of calibration (Lakshminarayanan et al., 2017; Sarawgi et al., 2020a). To plot these curves, we first compute the $x\%$ prediction interval for each data point under evaluation based on Gaussian quantiles using the prediction value and variance. We then calculate the fraction of data points under evaluation with true values that fall within this prediction interval. For a well-calibrated model, the observed fraction should be close to the $x\%$ calculated earlier. To see how our models perform in this setting, we sweep from $x = 10\%$ to $x = 90\%$ in steps of 10. A line lying close to the line ($y = x$) would indicate a well-calibrated model. Fig. 1 shows the that calibration the Interventions and Acoustic modalities become better-calibrated compared to the Disfluency modality. However in case of vanilla ensemble, the calibration of the Interventions and Acoustic modalities become worse when compared to the Disfluency modality. Fig. 2 shows the that calibration the Frequency modality become better-calibrated compared to the Amplitude modality in the case of UA ensembles, whereas they became worse calibrated in the case of vanilla ensemble. This highlights the significance of introducing the notion of uncertainty-awareness in ensembles to obtain better-calibrated models.

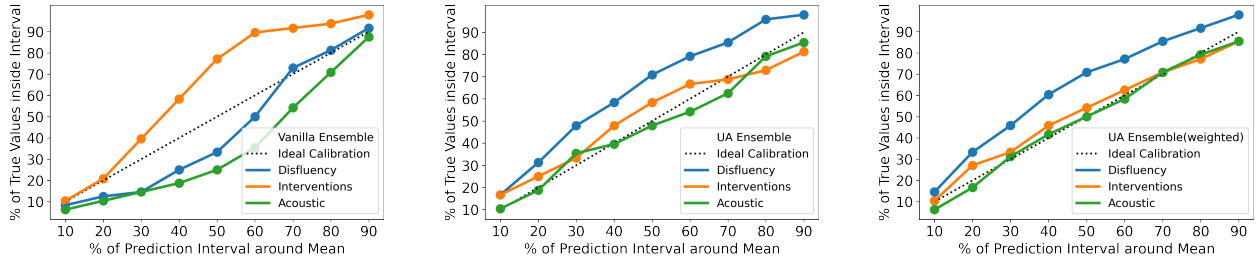

Figure 1. Calibration curves for the ensemble techniques on the ADReSS dataset.

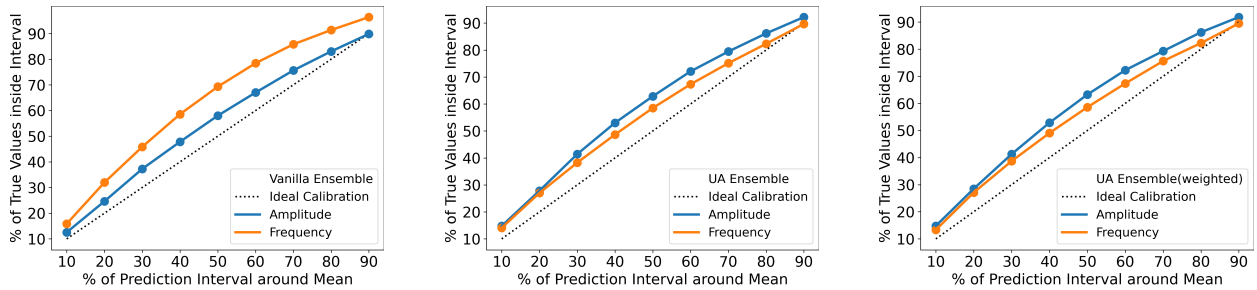

Figure 2. Calibration curves for the ensemble techniques on the Parkinson's Telemonitoring dataset.

We also compare our approach on other state-of-the-art methods on the ADReSS test set and find that it outperforms the current best approaches (Table 3 in Appendix E.1). Further analysis shows that the entropy of each modality as we sequentially boost shows a drastic reduction in the case of UA ensembles (Fig. 5 in Appendix E.1). An analysis of the Mean Prediction Interval Width (MPIW) and the Prediction Interval Coverage Probability (PICP) (see Table 4 in Appendix E.1 and Table 5 in Appendix E.2) shows that uncertainty-aware boosting results in tighter bounds for the confidence intervals along with higher PICP values, and high quality prediction intervals as desired (Pearce et al., 2018).

4. Discussion and Future Work

We proposed an uncertainty-aware boosted ensembling method in multi-modal settings, improving the performance when compared to individual modalities and boosting using loss values. By focusing more on data points with higher uncertainty, through uncertainty-weighting of the loss function (UA Ensemble) and the predictions as well (UA Ensemble (weighted)), we showed how our ensemble outperforms the results of state-of-the-art methods. Our experiments showed that the propagation of the uncertainty sequentially through the base learners of every modality aids the multi-modal system to decrease the overall entropy in the system, making it more reliable when compared to vanilla ensembles. Additionally, the modalities indeed become well calibrated along with high quality prediction intervals when boosted using uncertainty values, rather than loss values.

Such characteristics are significantly desired in real-world settings where data tends to exist in multiple modalities together. Understanding what a machine learning model does not know is crucial in safety-critical applications. Access to such information helps with designing a more reliable and aware decision-making system (Amodei et al., 2016; Varshney & Alemzadeh, 2017; Kumar et al., 2019; Thiagarajan et al., 2020; Gal, 2016). Furthermore, the availability of predictive uncertainties corresponding to each modality adds a level of transparency to the machine learning system. This can assist the user in making more informed decisions, thereby nurturing the synergy between humans and AI.

There are a lot of interesting possible future research directions to this work. One could definitely expand the proposed method itself to account for uncertainty values, as well as loss values, while boosting the base learners. Our current experiments make use of speech and text data with neural networks and random forests as the base learners. This can be extended to other forms of machine learning systems, making use of other Bayesian and non-Bayesian uncertainty estimation techniques and data modalities. Additionally, we encourage the community to further evaluate such techniques in other safety-critical tasks and applications, as well as assess the longitudinal performance and attributes of these systems, especially in the presence of noisy data and/or when the observed data distribution tends to shift over time and eventually becomes very different. This also opens up avenues to potentially design adaptive systems which could actively learn from the uncertainty estimates at deployment time.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*, 2020.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Breiman, L. Bagging predictors. *Machine learning*, 24(2): 123–140, 1996.

Chang, H.-S., Learned-Miller, E., and Mccallum, A. Active bias: Training a more accurate neural network by emphasizing high variance samples. 04 2017.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. URL https://doi.org/10.1145/2939672.2939785.

Coulston, J., Blinn, C., Thomas, V., and Wynne, R. Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing*, 82:189–197, 03 2016. doi: 10.14358/PERS.82.3.189.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y., and Schuler, A. Ngboost: Natural gradient boosting for probabilistic prediction, 2020.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., and Dai, A. M. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 204–213, 2020.

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835–838, 2013.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997a. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997b.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

Gal, Y. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015.

Hill, D. L., Hawkes, D. J., Harrison, N. A., and Ruff, C. F. A strategy for automated multimodality image registration incorporating anatomical knowledge and imager characteristics. In *Biennial International Conference on Information Processing in Medical Imaging*, pp. 182–196. Springer, 1993.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pp. 1169–1179. PMLR, 2020.

Jain, S., Liu, G., Mueller, J., and Gifford, D. Maximizing overall diversity for improved uncertainty estimates in deep ensembles. In *AAAI*, pp. 4264–4271, 2020.

Kay, J. W., Titterington, D. M., et al. *Statistics and neural networks: advances at the interface*. Oxford University Press on Demand, 1999.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3787–3798, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Lopez-de Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Faundez-Zanuy, M., Ecay-Torres, M., Travieso, C. M., Ezeiza, A., Estanga, A., et al. Alzheimer disease diagnosis based on automatic spontaneous speech analysis. 2012.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*, 2020.

Malinin, A., Prokhorenkova, L., and Ustimenko, A. Uncertainty in gradient boosting via ensembles. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1Jv6b0Zq3qi.

Nakamura, K., Levy, S., and Wang, W. Y. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

on Rating Scales for Parkinson's Disease, M. D. S. T. F. The unified parkinson's disease rating scale (updrs): status and recommendations. *Movement Disorders*, 18(7):738–750, 2003.

Osband, I. Risk versus uncertainty in deep learning : Bayes , bootstrap and the dangers of dropout. 2016.

Oviatt, S., Cohen, P., Wu, L., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., et al. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human-computer interaction*, 15(4):263–322, 2000.

Pappagari, R., Cho, J., Moro-Velazquez, L., and Dehak, N. Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. 2020.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/pearce18a.html.

Pearce, T., Leibfried, F., and Brintrup, A. Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, pp. 234–244. PMLR, 2020.

Qiu, X., Meyerson, E., and Miikkulainen, R. Quantifying point-prediction uncertainty in neural networks via residual estimation with an i/o kernel. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxNh1Stvr.

Rohanian, M., Hough, J., and Purver, M. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. 2020.

Sarawgi, U., Zulfikar, W., Khincha, R., and Maes, P. Why have a unified predictive uncertainty? disentangling it using deep split ensembles. *arXiv preprint arXiv:2009.12406*, 2020a.

Sarawgi, U., Zulfikar, W., Soliman, N., and Maes, P. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. *arXiv preprint arXiv:2009.00700*, 2020b.

Searle, T., Ibrahim, Z., and Dobson, R. Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv:2006.07358*, 2020.

Sen, M. K. and Stoffa, P. L. Bayesian inference, gibbs' sampler and uncertainty estimation in geophysical inversion 1. *Geophysical Prospecting*, 44(2):313–350, 1996.

Shaker, M. H. and Hüllermeier, E. Aleatoric and epistemic uncertainty with random forests, 2020.

Shridhar, K., Laumann, F., and Liwicki, M. Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference. *arXiv preprint arXiv:1806.05978*, 2018.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., and Nado, Z. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.

Thiagarajan, J. J., Venkatesh, B., Sattigeri, P., and Bremer, P.-T. Building calibrated deep models via uncertainty matching with auxiliary interval predictors. In *AAAI*, pp. 6005–6012, 2020.

Tombaugh, T. N. and McIntyre, N. J. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.

Tsanas, A., Little, M., Mcsharry, P., and Ramig, L. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE transactions on bio-medical engineering*, 57:884–93, 11 2009. doi: 10.1109/TBME.2009.2036000.

Varshney, K. R. and Alemzadeh, H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

Wager, S. randomForestCI, September 2016. URL https://github.com/swager/randomForestCI.

Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of machine learning research : JMLR*, 15:1625–1651, 05 2014.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. Deterministic variational inference for robust bayesian neural networks. *arXiv preprint arXiv:1810.03958*, 2018.

Zhang, X. and Mahadevan, S. Ensemble machine learning models for aviation incident risk prediction. *Decision Support Systems*, 116:48–63, 2019.

A. Process Diagrams

The process diagrams for Vanilla Ensemble and Uncertainty-aware ensemble is shown in Fig. 3.

B. Datasets

B.0.1. Dementia

We use the standardized and benchmark ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) dataset[1] (Luz et al., 2020). This dataset consists of speech samples (WAV format) and transcripts (CHA format), and their corresponding 'MMSE' (Mini-Mental State Examination) scores as labels for regression. MMSE scores (ranging from 0 to 30 and widely used in clinical practice) offer a way to quantify cognitive function, as well as screening for cognitive loss by testing the individuals' attention, recall, language, and motor skills (Tombaugh & McIntyre, 1992).

The dataset consists of 156 data points, each from a unique subject, matched for age and gender. A standardized train-test split of around 70%-30% (108 and 48 subjects) is provided by the dataset. We further split the train set into 80%-20% train-val sets. The test set was held out for all experimentation until final evaluation.

B.0.2. Parkinson's Disease

We use the publicly available Parkinson's Telemonitoring dataset[2] (Tsanas et al., 2009). This dataset consists of a range of 16 biomedical voice measurements and their corresponding 'Total UPDRS' (Unified Parkinson Disease Rating Scale) scores as labels for regression. Total UPDRS scores (ranging from 0 to 199 and widely used as a measure of severity of the Parkinson's disease (PD)) offer a way to quantify the course of PD in patients by testing the individuals' mentation, behaviour, mood, daily-life activities, and motor examination (on Rating Scales for Parkinson's Disease, 2003).

The dataset consists of 5,875 data points from 42 subjects with early-stage PD recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. Since a standardized train-test split is not provided by the dataset, we use a 5-fold cross validation with consistent folds across the different methods for a fair evaluation.

[1]ADReSS dataset can be downloaded from https://dementia.talkbank.org/ along with an email to obtain the password for access.

[2]Parkinson's Telemonitoring dataset can be downloaded from https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring.

C. Multi-modal Feature Extraction

C.1. Dementia

we extract multi-modal acoustic, cognitive and linguistic features from the available speech samples and their corresponding transcripts, by using the feature engineering pipeline as developed by Sarawgi et al. (Sarawgi et al., 2020b). This results in three input modalities, namely 'Disfluency', 'Interventions', and 'Acoustic' (following the same terminology as Sarawgi et al. (Sarawgi et al., 2020b)). The three feature sets - namely 'Disfluency', 'Acoustic', and 'Interventions' mentioned in Section 3.1 are explained below:

C.1.1. Disfluency

A set of 11 distinct and carefully curated features from the transcripts, such as word rate, intervention rate, and different kinds of pause rates reflecting upon speech impediments such as slurring and stuttering. These are normalized by the respective audio lengths and scaled thereafter.

C.1.2. Acoustic

The ComParE 2013 feature set (Eyben et al., 2013) was extracted from the audio samples using the open-sourced openSMILE v2.1 toolkit, widely used for affect analyses in speech (Eyben et al., 2010). This provides a total of 6,373 features that include energy, MFCC, and voicing related low-level descriptors (LLDs), and other statistical functionals. This feature set encodes changes in speech of a person and has been used as an important noninvasive marker for AD detection (Lopez-de Ipiña et al., 2012; Luz et al., 2020). The system standardizes this set of features using z-score normalization, and uses principal component analysis (PCA) to project the 6,373 features onto a low-dimensional space of 21 orthogonal features with highest variance. The number of orthogonal features was selected by analyzing the percentage of variance explained by each of the components.

C.1.3. Interventions

Cognitive features reflect upon potential loss of train of thoughts and context. The system extracts the sequence of speakers from the transcripts, categorizing it as subject or the interviewer. To accommodate for the variable length of these sequences, they are padded or truncated to length of 32 steps, found upon analysis and tuning of sequence lengths.

C.2. Parkinsons

The dataset consists of features related to amplitude and frequency. Accordingly, we extract two input modalities from the available data, referring to them as the 'Amplitude' and 'Frequency' modalities. Consequently, the list of features in

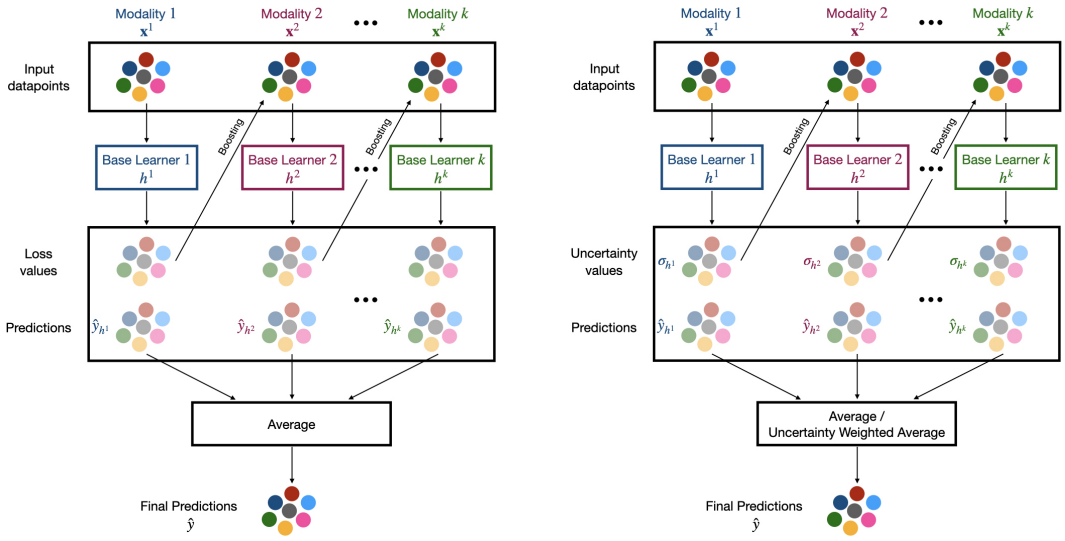

Figure 3. Process diagrams of 1) 'vanilla ensemble' (left), and 2) 'uncertainty-aware (UA) ensemble' / 'UA ensemble (weighted)' (right). The symbols in the diagrams follow from the notation and setup in Section 2.1.

the two input modalities are as below:

- 'Amplitude': Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA, NHR, HNR, RPDE, DFA
- 'Frequency': Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP, PPE

D. Model Architecture

D.1. Dementia

Following from Section 3.1, we have three input modalities here i.e. $j = 1, 2, 3$ and $k = 3$. Now, with a training dataset $\{(\mathbf{x}_n^j, y_n)\}_{n=1}^N$ consisting of N i.i.d. samples for each of the three modalities, we model the probabilistic predictive distribution $p_{h^j}(y|\mathbf{x}^j)$ using a neural network (NN) with parameters h^j.

For a fair comparison with the state-of-the-art, we use almost the same NN architecture as used by Sarawgi et al. (Sarawgi et al., 2020b) for each of the three input modalities. The Disfluency and Acoustic models make use of multi-layer perceptrons (MLPs), while the Interventions model makes use of LSTM, along with regularizers.

The individual model architecture of the base learners used for each of the three modalities (feature sets) are shown in Fig. 4.

D.1.1. Disfluency

The disfluency model is a multi-layer perceptron (MLP) that projects the 11-feature input to a higher dimensional space for better separability of the features.

D.1.2. Acoustic

The acoustic model is an MLP with a single hidden layer that adds non-linearity and regularizes the PCA decomposed feature space.

D.1.3. Interventions

The interventions model uses a recurrent architecture to learn the temporal relations from the sequence of interventions.

For uncertainty estimation, each of the models predicts a target distribution instead of a point estimate to account for the heteroscedasticity in data and yields predictive uncertainties along with the predicted mean value (Lakshminarayanan et al., 2017; Snoek et al., 2019; Sarawgi et al., 2020a). The target distribution is modelled as a Gaussian distribution $p_{h^j}(y_n|\mathbf{x}_n^j)$ parameterized by the mean μ_{h^j} and the standard deviation σ_{h^j}, predicted at the final layer of the models i.e. $y_n \sim \mathcal{N}\left(\mu_{h^j}, \sigma_{h^j}^2\right)$. It is important to note here that the prediction $\hat{y}_{h^j}$ is the predicted mean μ_{h^j}, and the predicted uncertainty estimate is the predicted standard deviation σ_{h^j}.

Each of the three base learners is trained with their corresponding input modality features $\mathbf{x}^j$ and ground truth labels y using a proper scoring rule. We optimize for the negative log-likelihood (NLL) of the joint distribution $p_{h^j}(y_n|\mathbf{x}_n^j)$ according to the equation below (3).

$$-\log\left(p_{h^j}(y_n|\mathbf{x}_n^j)\right) = \frac{\log\left(\sigma_{h^j}^2\right)}{2} + \frac{(y-\mu_{h^j})^2}{2\sigma_{h^j}^2} + \text{constant} \tag{3}$$

We use the boosting methods explained in Section 2.2 to

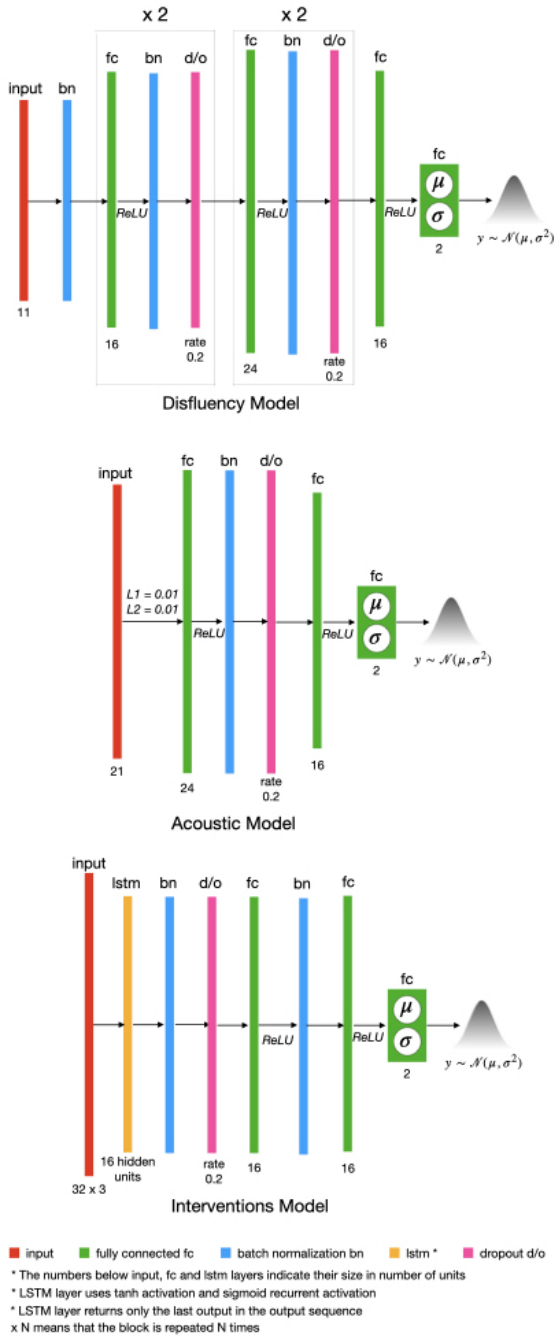

Figure 4. Model architecture of the (1) Disfluency (2) Acoustic and (3) Intervention models (base learners).

train an ensemble with the three base learners (Disfluency, Acoustic, and Interventions). Each training run used a batch size of 32 and an Adam optimizer with a learning rate of 0.00125 to minimize the NLL.

D.2. Parkinsons

Following from Section 3.1, we have two input modalities here i.e. $j = 1, 2$ and $k = 2$. Now, with a training dataset $\{(\mathbf{x}_n^j, y_n)\}_{n=1}^N$ consisting of N i.i.d. samples for each of the two modalities, we model the probabilistic predictive distribution $p_{h^j}(y|\mathbf{x}^j)$ using a random forest (RF) regressor with parameters h^j. Each of the RFs makes use of 300 decision tree estimators. This was decided upon sweeping the number of decision trees, from 100 to 1000, as a hyperparameter.

The two base learners are trained with their corresponding input modality features $\mathbf{x}^j$ and ground truth labels y using a mean squared error (MSE) loss. The uncertainty estimates σ_{h^j} of each of the data point is estimated as the confidence interval using the Infinitesimal Jackknife method (Wager et al., 2014; Wager, 2016).

We use the boosting methods explained in Section 2.2 to train an ensemble with the two base learners (Amplitude and Frequency).

E. Additional Results

E.1. Dementia

Table 3 compares our uncertainty-aware ensembling approach with other state-of-the-art approaches on the ADReSS test set. The best run out of the five discussed in Section 3.3 is used for reporting our RMSE scores.

Table 3. Comparison of uncertainty-aware ensemble methods with state-of-the-art results on the ADReSS test set.

Model	**RMSE**
Pappagari et al. (Pappagari et al., 2020)	5.37
Luz et al. (Luz et al., 2020)	5.20
Sarawgi et al. (Sarawgi et al., 2020b)	4.60
Searle et al. (Searle et al., 2020)	4.58
Balagopalan et al. (Balagopalan et al., 2020)	4.56
Rohanian et al. (Rohanian et al., 2020)	4.54
Sarawgi et al. (Sarawgi et al., 2020a)	4.37
UA Ensemble	**4.35**
UA Ensemble (weighted)	**3.93**

Fig. 5 shows how the entropy of the modalities change as we move across the base learners in both vanilla ensemble as well as UA ensemble. It is clear that as we move sequentially across the base learners, the entropy of the acoustic and intervention modalities in UA ensembles reduce drastically when compared to the vanilla ensemble.

Table 4 shows the MPIW and PICP metrics on a 5-times repeated test set evaluation.

E.2. Parkinsons

Table 5 shows the MPIW and PICP metrics on a 5-times repeated test set evaluation.

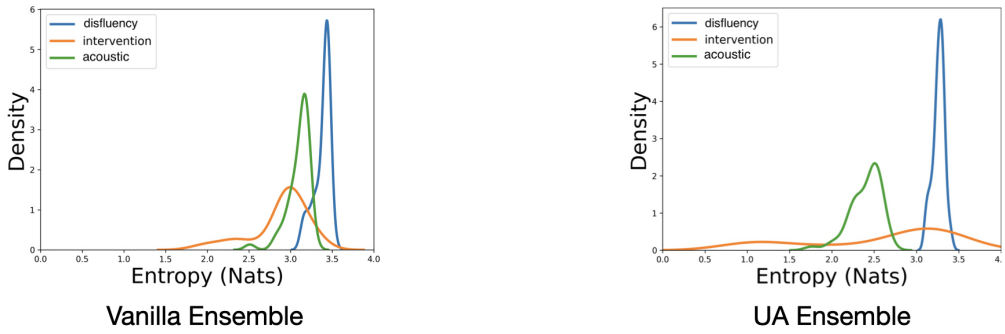

Figure 5. Entropy analysis, using kernel density estimation plots, of the base learners in a vanilla ensemble (left) and UA ensemble (right). UA ensemble shows a decrease in the overall entropy of the system. The increased reduction in the entropy as we sequentially move from the first base learner to the last base learner of the ensemble further indicates the significance of introducing uncertainty-awareness into the ensemble.

Table 4. 5-times repeated test set results of Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP) for the ensemble techniques on the ADReSS dataset. We report PICP results with the prediction interval (Δ) equal to 1, 2, and 3 times the standard deviation (i.e. 1σ, 2σ, and 3σ). The uncertainty-aware boosting results in tighter bounds for the confidence intervals, along with higher PICP values, and high quality prediction intervals as desired.

Model	Modality	MPIW	PICP (%)		
			$\Delta = 1\sigma$	$\Delta = 2\sigma$	$\Delta = 3\sigma$
Vanilla Ensemble	Disfluency	4.47 ± 0.39	61.66 ± 8.29	95.83 ± 2.63	97.50 ± 0.83
	Interventions	7.27 ± 0.58	$\mathbf{87.50 \pm 5.43}$	$\mathbf{99.17 \pm 1.02}$	$\mathbf{100.00 \pm 1.18}$
	Acoustic	$\mathbf{4.50 \pm 0.73}$	59.58 ± 12.54	94.58 ± 2.12	98.75 ± 1.02
UA Ensemble	Disfluency	6.29 ± 0.81	82.91 ± 6.37	97.91 ± 1.31	100.00 ± 0.00
	Interventions	$\mathbf{5.46 \pm 1.57}$	73.75 ± 14.47	93.33 ± 5.17	97.91 ± 1.86
	Acoustic	5.31 ± 1.30	$\mathbf{75.41 \pm 11.21}$	$\mathbf{96.25 \pm 3.06}$	$\mathbf{99.16 \pm 1.02}$
UA Ensemble (weighted)	Disfluency	6.29 ± 0.81	83.33 ± 6.58	97.91 ± 1.31	100.00 ± 0.00
	Interventions	$\mathbf{5.46 \pm 1.57}$	76.25 ± 13.85	92.50 ± 5.98	96.66 ± 3.11
	Acoustic	5.31 ± 1.30	$\mathbf{75.83 \pm 10.59}$	$\mathbf{95.00 \pm 3.86}$	$\mathbf{99.16 \pm 1.02}$

Table 5. 5-fold cross-validation results of Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP) for the ensemble techniques on the Parkinson's Telemonitoring dataset. We report PICP results with the prediction interval (Δ) equal to 1, 2, and 3 times the standard deviation (i.e. 1σ, 2σ, and 3σ). The uncertainty-aware boosting results in tighter bounds for the confidence intervals, along with higher PICP values, and high quality prediction intervals as desired.

Model	Modality	MPIW	PICP (%)		
			$\Delta = 1\sigma$	$\Delta = 2\sigma$	$\Delta = 3\sigma$
Vanilla Ensemble	Amplitude	6.79 ± 1.28	84.56 ± 1.46	98.51 ± 0.58	99.89 ± 0.12
	Frequency	8.69 ± 0.59	74.17 ± 8.25	94.28 ± 3.37	98.60 ± 1.18
UA Ensemble	Amplitude	6.50 ± 1.76	74.09 ± 9.15	93.70 ± 4.11	98.23 ± 1.47
	Frequency	$\mathbf{6.91 \pm 0.85}$	$\mathbf{77.90 \pm 5.28}$	$\mathbf{95.64 \pm 2.40}$	$\mathbf{99.33 \pm 0.51}$
UA Ensemble (weighted)	Amplitude	6.50 ± 1.76	74.24 ± 8.59	93.71 ± 4.13	97.97 ± 1.66
	Frequency	$\mathbf{6.91 \pm 0.85}$	$\mathbf{77.65 \pm 5.67}$	$\mathbf{95.45 \pm 2.56}$	$\mathbf{99.18 \pm 0.70}$