Uncertainty-Aware Boosted Ensembling in Multi-Modal Settings

Utkarsh Sarawgi[§]
MIT Media Lab
Massachusetts Institute of Technology
Cambridge, USA
utkarshs@mit.edu

Rishab Khincha[§]
MIT Media Lab
Massachusetts Institute of Technology
BITS Pilani, K K Birla Goa Campus
Bangalore, India
rkhincha@mit.edu

Wazeer Zulfikar[§]
McGovern Institute for Brain Research
Massachusetts Institute of Technology
Cambridge, USA
wazeer@mit.edu

Satrajit Ghosh
McGovern Institute for Brain Research
Massachusetts Institute of Technology
Harvard Medical School
Cambridge, USA
satra@mit.edu

Pattie Maes
MIT Media Lab
Massachusetts Institute of Technology
Cambridge, USA
pattie@mit.edu

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.

I. INTRODUCTION

Rapid developments in machine learning (ML) across a variety of tasks have advanced its deployment in real-world settings [1]. 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 [2]. 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 [3].

As such, generating confidence intervals or uncertainty estimates along with the predictions is crucial for reliable

[§]Equal contribution

and safe deployment of machine learning systems in safety-critical settings (such as healthcare) [4]–[7]. It is critical to understand what a model does not know when building and deploying machine learning systems to help mitigate possible risks and biases in decision making [8]. This 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.

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 [9]. Parallel and sequential techniques are widely used to improve performance by ensembling weak learners trained with a single data modality [10]–[14]. Similarly, base learners trained with different input modalities can be ensembled together for performance improvements [9], [15], [16].

Some works have briefly discussed uncertainty estimation in multi-modal settings [17]–[20]. 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 III).

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.

We evaluate our method on multi-modal speech and text datasets on healthcare tasks related to Dementia and Parkinson's disease using different machine learning models (Neural Networks and Random Forests) and uncertainty estimation techniques (Gaussian target distribution [21] and Infinitesimal Jackknife method [22]). Our analysis shows an increased reduction in the entropy as we sequentially move from the first base learner to the last base learner of the ensemble, further demonstrating the significance of introducing uncertainty-awareness into the ensemble. The model becomes more robust to heteroscedasticity in the data while also well calibrating each of the individual modalities along with high quality prediction intervals (Section IV-D). To the best of our knowledge, we are the first to explore such a boosting mechanism in a multi-modal ensemble using predictive uncertainty estimates.

A. Summary of Contributions:

- We first propose and formulate an 'uncertainty-aware boosted ensemble' (Section III), as also depicted in the process diagram in Fig. 1.
- We then evaluate our method on multi-modal speech and text datasets on healthcare tasks related to Dementia and Parkinson's disease using different ML models and uncertainty estimation techniques (Section IV).
- We also perform entropy, calibration, and prediction interval analyses to highlight the significance of introducing uncertainty-awareness into the ensemble (Section IV).

II. RELATED WORK

A. Uncertainty Estimation

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 a model's predictions. Bayesian approximation techniques such as dropout-based VI [23], [24], expectation propagation [25], variational inference (VI) [26], [27], deterministic VI [28], neural networks as Gaussian processes [29], approximate Bayesian ensembling [30], and Bayesian model averaging in low-dimensional parameter subspaces [31] have been shown to be quite useful in modelling the uncertainties in neural networks. Non-Bayesian approaches [17], [21], [32]–[34] that involve bootstrapping and ensembling multiple probabilistic neural networks have shown performances comparable to Bayesian methods with reduced computational costs and modifications to the training procedure. Additionally, there is a breadth of other theoretical, empirical, and review works on estimating predictive uncertainties with neural networks [35]–[41].

Uncertainty estimation in trees and random forests has been studied in the past, with multiple methods being proposed for both classification and regression tasks. NGBoost generalizes gradient boosting to probabilistic regression by treating the parameters of the conditional distribution as targets for a multiparameter boosting algorithm [42]. Ensemble methods have been proposed to study the uncertainty in gradient boosted models [43]. Recently, various methods such as the infinitesimal jackknife [22], monte carlo based approaches [44], and using the decision tree as a probabilistic predictor [45] have been suggested to predict the uncertainty in random forests. Ashukha et al. [46] performed a broad study of ensembling techniques in context of uncertainty estimation.

B. Ensemble techniques

Parallel and sequential ensembling techniques have been widely studied and shown to improve performance of models in a variety of tasks [9]–[16]. Some works combine latent embeddings from different input modalities (feature fusion) and train the entire model together in a joint fashion [47]. Boosting algorithms such as Adaptive Boost [12], Gradient Boosting [13], and XG Boost [14] are common sequential ensembling techniques. These methods are often used to improve performance by ensembling weak base learners trained with a single data modality [10]–[14]. Similarly in multi-modal settings, base learners trained with different input data modalities can be ensembled together to help leverage different feature sets together [9], [15], [16].

There has been some work on incorporating uncertainty with ensembling methods. Kendall et al. [48] learns multiple tasks by using the uncertainty predicted as weights for the losses of each of the models, thus outperforming the individual models trained on each task. Chang et al. [49] uses uncertainty estimates and prefers to learn the data points predicted incorrectly with higher uncertainty in different mini-batches of SGD. Some other works have addressed the advantages of incorporating uncertainty estimates in multi-modal settings [17]–[20].

Our work particularly designs boosted ensemble techniques that are uncertainty-aware in multi-modal settings, explores how these can be really helpful in real-world settings, and motivates future research directions.

III. UNCERTAINTY-AWARE BOOSTED ENSEMBLE

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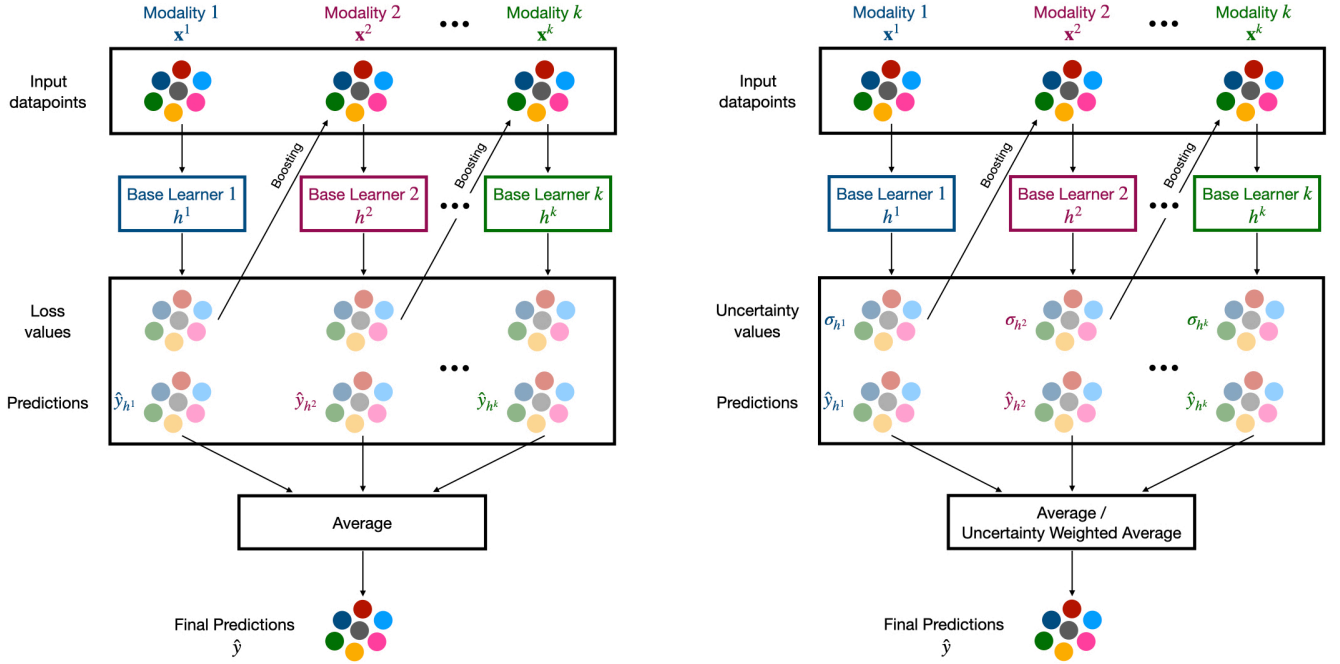
Fig. 1. Process diagrams of 1) 'vanilla ensemble' (left), and 2) 'uncertainty-aware (UA) ensemble' / 'UA ensemble (weighted)' (right). The symbols in the diagrams follow from the notation and setup in Section III-A.

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

$$h^j : \mathbf{x}_n^j \longrightarrow y_n \qquad (1)$$

B. 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. 1 shows the process diagrams of vanilla ensemble, UA ensemble, and UA ensemble (weighted).

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 jth 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) 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 jth 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 $\hat{y}$.

3) UA ensemble (weighted): We experiment with a variation to the final averaging of the predictions in the UA

ensemble discussed above. Here, for the final prediction $\hat{y}$, the ensemble computes a weighted average of the predictions $\{\hat{y}_{h^j}\}_{j=1}^k$ of all the (boosted) base learners, where the weights used are the inverse of the respective predicted uncertainty estimates σ_{h^j}. Equation (2) mathematically formulates this, where $\hat{y}(\mathbf{x}_n)$ is the final prediction corresponding to n^{th} data point, and k is the total number of individual modalities as defined in Section III-A.

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

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 [12]–[14], the base learners here need not be weak learners.

IV. 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. We make use of different types of machine learning models (Neural Networks and Random Forests) and uncertainty estimation techniques

(Gaussian target distribution [21] and Infinitesimal Jackknife method [22]) for the two datasets.

A. Datasets

1) Dementia: We use the standardized and benchmark ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) dataset[1] [50]. 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 [51].

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.

2) Parkinson's Disease: We use the publicly available Parkinson's Telemonitoring dataset[2] [52]. 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 [53].

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.

B. Multi-modal Feature Extraction

1) Dementia: For a fair comparison with the state-of-the-art, 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. [15]. This results in three input modalities, namely 'Disfluency', 'Interventions', and 'Acoustic' (following the same terminology as Sarawgi et al. [15]). The details of features from each of the three modalities can be found in our Supplementary material[3].

2) Parkinson's Disease: 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 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

C. Model Architecture and Training

1) Dementia: Following from Section IV-B, 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. [15] 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 exact model architecture, along with regularizers for each of the three models (base learners) can be found in our Supplementary material[3].

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 [17], [21], [40]. 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}. The exact illustration of the implementation can be found in our Supplementary material[3].

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).

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

We use the boosting methods explained in Section III-B to 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.

2) Parkinson's Disease: Following from Section IV-B, 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.
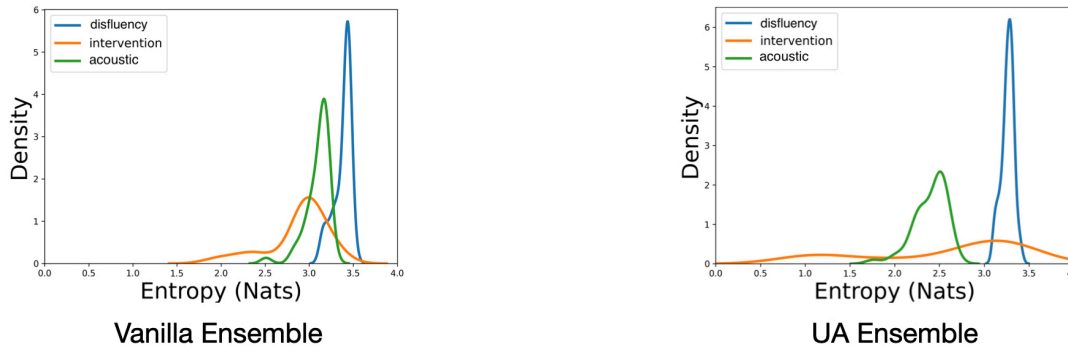
Fig. 2. Entropy analysis, using kernel density estimation plots, of the base learners in a vanilla ensemble (left) and UA ensemble (right). UA ensemble shows a decrease in the overall entropy of the system. The increased reduction in the entropy as we sequentially move from the first base learner to the last base learner of the ensemble further indicates the significance of introducing uncertainty-awareness into the ensemble (Section IV-D1).

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 [22], [54].

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

D. Results

1) Dementia: 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 (Table I).

TABLE I
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	**5.05 ± 0.53**
UA Ensemble (weighted)	**4.96 ± 0.49**

We also compare our uncertainty-aware ensemble methods with current state-of-the-art results on the ADReSS test set. Table II shows that the best of 5 runs of UA Ensemble is competitive, and that of the UA Ensemble (weighted) outperforms other methods.

The use of uncertainty awareness can improve the robustness of the ensemble with uncertain data points (i.e subjects). To highlight this, we evaluate the entropy of the base learners

TABLE II
COMPARISON OF UNCERTAINTY-AWARE ENSEMBLE METHODS WITH
STATE-OF-THE-ART RESULTS ON THE ADReSS TEST SET.

Model	RMSE
Pappagari et al. [55]	5.37
Luz et al. [50]	5.20
Sarawgi et al. [15]	4.60
Searle et al. [56]	4.58
Balagopalan et al. [57]	4.56
Rohanian et al. [58]	4.54
Sarawgi et al. [17]	4.37
UA Ensemble	**4.35**
UA Ensemble (weighted)	**3.93**

in the ensemble methods while sequentially boosting in the vanilla ensemble and the UA ensemble. Upon comparison, we see a decrease in the overall entropy of the system when the ensemble is uncertainty-aware (Fig. 2). 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.

We also analyse our approach on the Mean Prediction Interval Width (MPIW) and Prediction Interval Coverage Probability (PICP), two widely used metrics for evaluating uncertainty in regression. PICP is the percentage of the times the prediction interval contains the actual regression value, while MPIW is the average size of all prediction intervals. Pearce et al. [59] discusses that high-quality prediction intervals should be as narrow as possible, whilst capturing some specified proportion of data points. Accordingly, it is desirable to have low MPIW values and PICP $\geq (1 - \alpha)$, a common choice of α being 0.05.

Table III shows the 5-times repeated test set results on MPIW and PICP metrics with a comparison between the vanilla ensemble and the uncertainty-aware ensembles. It is important to note that the comparison holds value for the results corresponding to the modalities that have been boosted (in this case, the Interventions and Acoustic modalities). UA ensemble and its variation UA ensemble (weighted) observe

| 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	$\mathbf{75.83 \pm 10.59}$	$\mathbf{95.00 \pm 3.86}$	$\mathbf{99.16 \pm 1.02}$
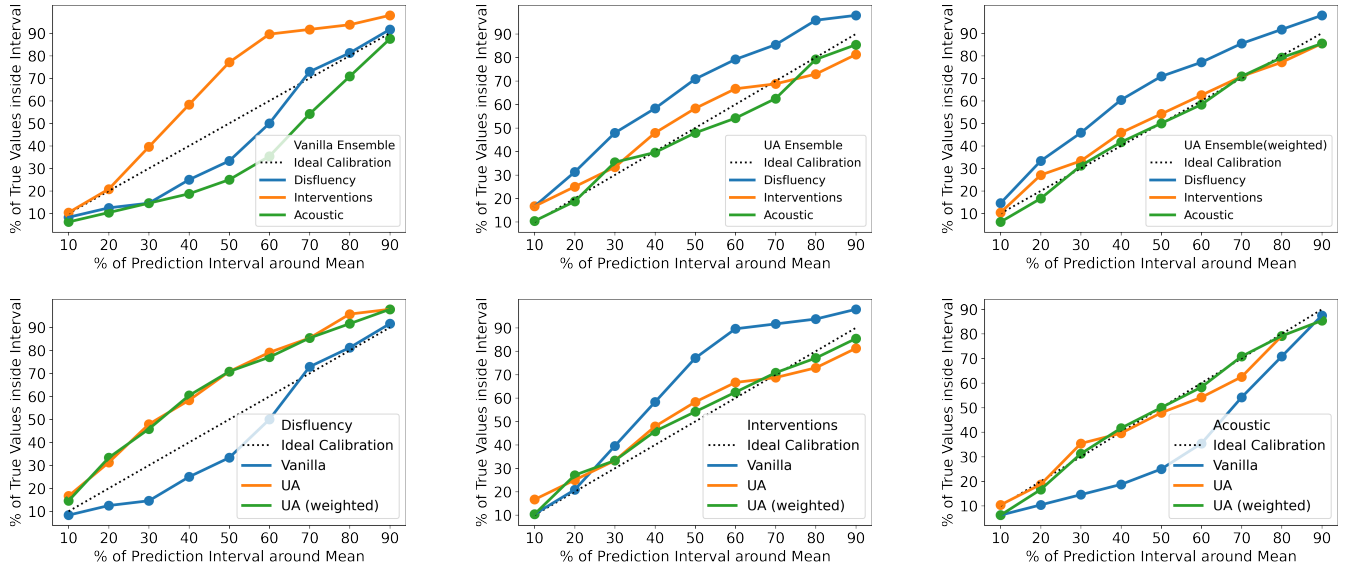
Fig. 3. Calibration curves for the ensemble techniques on the ADReSS dataset. The two rows of plots use the same data to just visualize the comparison differently (ensemble-wise and modality-wise respectively). The plots show that the boosted modalities are better-calibrated in case of uncertainty-aware boosting (Section IV-D1).

reduced MPIW compared to vanilla ensemble for the Interventions modality. While the vanilla ensemble observes a reduced MPIW for the Acoustic modality, the uncertainty-aware ensembles observe a gradual decrease in the MPIW values while sequentially boosting across the modalities. The uncertainty-aware boosting thus results in tighter bounds for the confidence intervals, along with higher PICP values, and high quality prediction intervals as desired [59].

We further use the 65-95-99.7 rule (also called the empirical rule) to obtain calibration curves for a comprehensive analysis of calibration [17], [21]. 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. 3 shows the calibration curves of the three modalities in case of vanilla ensemble, UA ensemble, and UA ensemble (weighted). We observe that in the case of UA ensemble and UA ensemble (weighted), 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. Consequently, it follows that the boosted modalities are better-calibrated in case of uncertainty-aware boosting, as desired in real-world settings. This highlights the significance of introducing the notion of uncertainty-awareness in ensembles to obtain better-calibrated models.
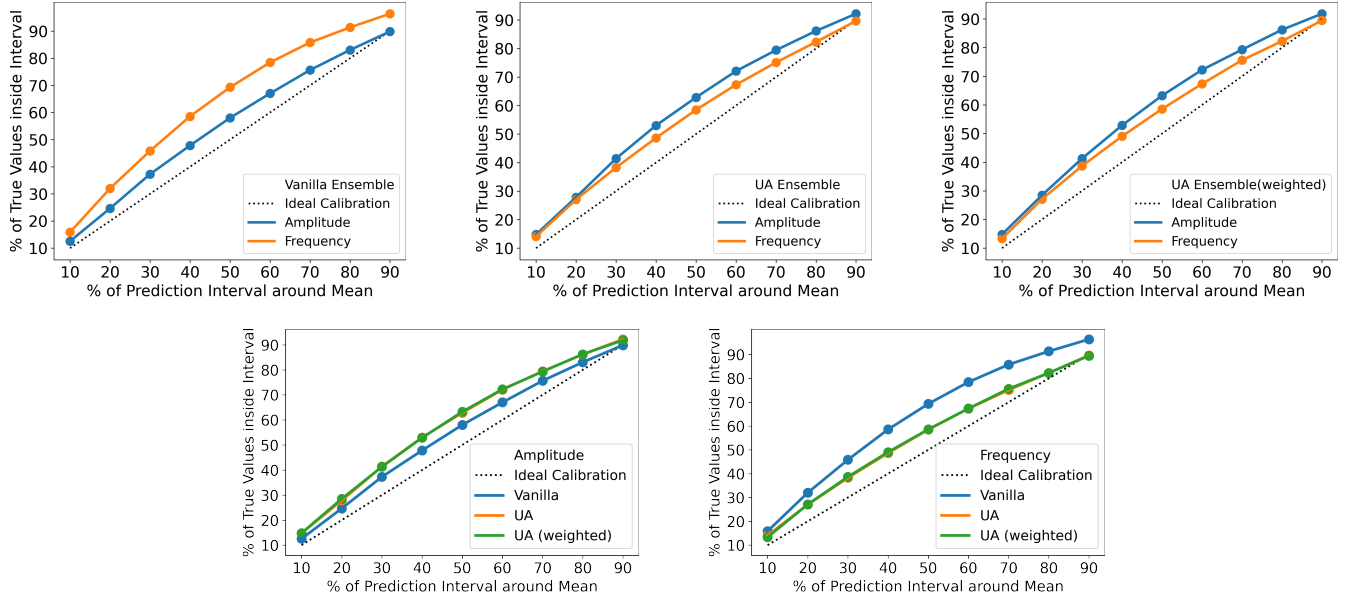
Fig. 4. Calibration curves for the ensemble techniques on the Parkinson's Telemonitoring dataset. The two rows of plots use the same data to just visualize the comparison differently (ensemble-wise and modality-wise respectively). Also in the second row, the green and orange plots almost overlap (and hence, might not be clearly visible). The plots show that the boosted modalities are better-calibrated in case of uncertainty-aware boosting (Section IV-D2).

2) Parkinson's Disease: Since there is no standardised train-test split available, we perform a 5-fold cross validation and report the mean and variance of the RMSE results across the five folds. We first evaluate each of the modalities (i.e. base learners) individually and then compare them with the vanilla and uncertainty-aware ensembles. Again, the order of sequential boosting for the propagation of the uncertainties is chosen in the order of the cross-validation performance of the individual modalities. We observe that the uncertainty-aware ensembles perform better than the vanilla ensemble and the individual modalities (Table IV).

TABLE IV
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	**3.04 ± 0.04**
UA Ensemble (weighted)	**3.05 ± 0.05**

Table V shows the 5-fold cross-validation results on MPIW and PICP metrics with a comparison between the vanilla ensemble and the uncertainty-aware ensembles. It is important to note that the comparison holds value for the results corresponding to the modalities that have been boosted (in this case, the Frequency modality). UA ensemble and its variation UA ensemble (weighted) observe reduced MPIW compared to vanilla ensemble, indicating that the uncertainty-aware boosting results in tighter bounds for the confidence intervals, along with higher PICP values, and high quality

prediction intervals as desired [59].

Fig. 4 shows the calibration curves of the two modalities in case of vanilla ensemble, UA ensemble, and UA ensemble (weighted). We observe that in the case of UA ensemble and UA ensemble (weighted), the Frequency modality becomes better-calibrated compared to the Amplitude modality. However in case of vanilla ensemble, the calibration of the Frequency modality becomes worse when compared to the Amplitude modality. Consequently, as also observed with the ADReSS dataset, it follows that the boosted modality is better-calibrated in case of uncertainty-aware boosting, as desired in real-world settings. This again highlights the significance of introducing the notion of uncertainty-awareness in ensembles to obtain better-calibrated models.

V. DISCUSSION AND FUTURE WORK

We proposed an uncertainty-aware boosted ensembling method in multi-modal settings. Such an ensemble improves the performance when compared to individual modalities and ensembles boosted 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. More importantly, our discussion in Section I highlight how such an ensemble system can help design a more robust learner by having the base learners pay more attention to uncertain prediction corresponding to noisy data modalities. 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

TABLE V

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 (SECTION IV-D2).

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	**6.91 ± 0.85**	**77.90 ± 5.28**	**95.64 ± 2.40**	**99.33 ± 0.51**
UA Ensemble (weighted)	Amplitude	6.50 ± 1.76	74.24 ± 8.59	93.71 ± 4.13	97.97 ± 1.66
	Frequency	**6.91 ± 0.85**	**77.65 ± 5.67**	**95.45 ± 2.56**	**99.18 ± 0.70**

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 [4]–[8]. 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

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

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

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

[5] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big data*, vol. 5, no. 3, pp. 246–255, 2017.

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

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

[8] Y. Gal, "Uncertainty in deep learning," *University of Cambridge*, vol. 1, p. 3, 2016.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[10] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[12] ——, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, p. 119–139, Aug. 1997. [Online]. Available: https://doi.org/10.1006/jcss.1997.1504

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

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

[16] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, vol. 116, pp. 48–63, 2019.

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

[18] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson *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*, vol. 15, no. 4, pp. 263–322, 2000.

[19] M. K. Sen and P. L. Stoffa, "Bayesian inference, gibbs' sampler and uncertainty estimation in geophysical inversion 1," *Geophysical Prospecting*, vol. 44, no. 2, pp. 313–350, 1996.

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

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

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

[23] D. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," 06 2015.

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

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

[26] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2348–2356.

[27] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1613–1622. [Online]. Available: http://proceedings.mlr.press/v37/blundell15.html

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

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

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

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

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

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

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

[35] D. J. C. MacKay, "A practical bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, p. 448–472, May 1992. [Online]. Available: https://doi.org/10.1162/neco.1992.4.3.448

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

[37] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. Madison, WI, USA: Omnipress, 2011, p. 681–688.

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

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

[40] J. Snoek, Y. Ovadia, E. Fertig, B. Lakshminarayanan, S. Nowozin, D. Sculley, J. Dillon, J. Ren, and Z. Nado, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 969–13 980.

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

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

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

[44] J. Coulston, C. Blinn, V. Thomas, and R. Wynne, "Approximating prediction uncertainty for random forest regression models," *Photogrammetric Engineering & Remote Sensing*, vol. 82, pp. 189–197, 03 2016.

[45] M. H. Shaker and E. Hüllermeier, "Aleatoric and epistemic uncertainty with random forests," 2020.

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

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

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

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

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

[51] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[52] A. Tsanas, M. Little, P. Mcsharry, and L. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE transactions on bio-medical engineering*, vol. 57, pp. 884–93, 11 2009.

[53] M. D. S. T. F. on Rating Scales for Parkinson's Disease, "The unified parkinson's disease rating scale (updrs): status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[54] S. Wager, "randomForestCI," Sep. 2016. [Online]. Available: https://github.com/swager/randomForestCI

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

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

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

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

[59] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4075–4084. [Online]. Available: http://proceedings.mlr.press/v80/pearce18a.html

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

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

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