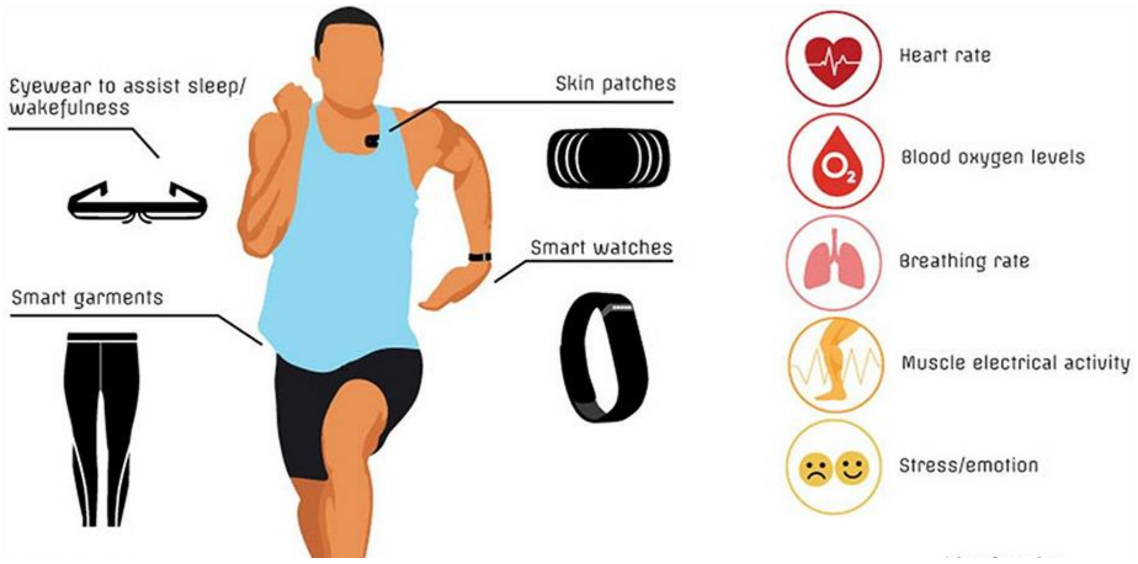
Benchmarking Uncertainty Quantification on Biosignal Classification Tasks under Dataset Shift

Tong Xia, Jing Han, Cecilia Mascolo.
University of Cambridge.



***Digital Health via Biosignals** -- the signal that can be continuously measured from human bodies, such as respiratory sounds, heart activity (ECG), brain waves (EEG), etc.*

Image from https://spie.org/news/spie-professional-magazine-archive/2019-january/wearables-move-beyond-the-consumer?SSO=1

Promise and Challenges for Biosignal Classifiers

✓ **Availability:** The growth of commercial wearables and the ubiquity of smartphones with numerous sensors have enabled multi-modal, affordable, non-invasive, round-the-clock biosignal collection.

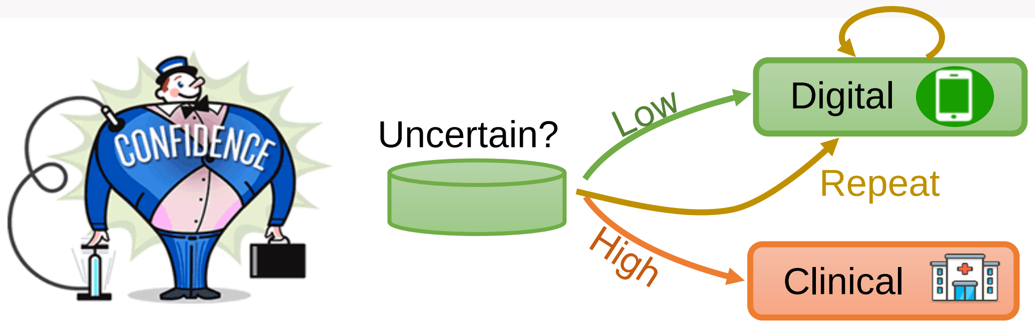
✓ **Promise:** A plural of machine learning models have been developed based on those biosignals with very promising performance for automatic disease detection and health status monitoring.

? **Reliability:** Dataset shift caused by user variability, device discrepancy, artifact, and other factors are ineluctable during in-the-wild biosignal acquisition.

? **Uncertainty:** Predictive uncertainty is recognised to be helpful by representing the prediction confidence, yet it is under-explored for biosingal classifiers.

Our Main Contribution

In this paper, we conduct a comprehensive evaluation across five uncertainty quantification methods on three representative biosignal classification tasks under a controlled dataset shift. Without requiring the collection of new datasets, the key mechanism is to empirically synthesise signal-specific distributional shift according to real signal data collection scenarios, so that both the shift type and the degree can be controlled and the evaluation framework can be generalised to any biosignal tasks.



Uncertain?

Low → Digital

Repeat

High → Clinical

Uncertainty can be utilised to flag unconfident prediction.

Baselines and Benchmark Tasks

Baselines. We select five methods for uncertainty estimation considering their prevalence, scalability, and practical applicability.

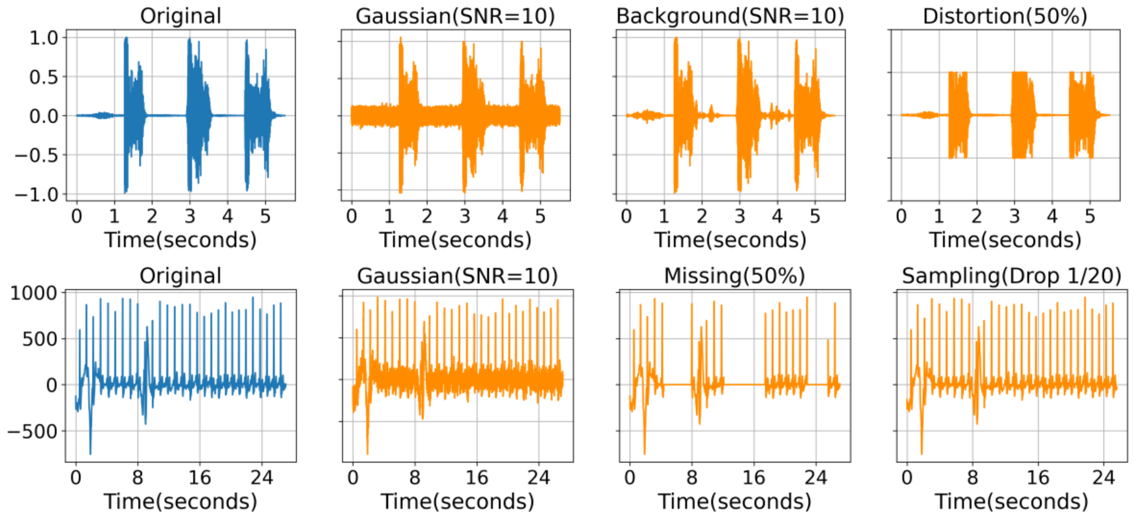
- **Vanilla**: Maximum softmax probability from the deterministic model.
- **Scaling**: Post-hoc calibration from vanilla probabilities by temperature scaling parameterized by value T.
- **MCDropout**: Monte-Carlo Dropout with a dropout rate of p during inferenceA sample will be fed into the model M times to quantify the model uncertainty.
- **Bayesian**: Stochastic variational Bayesian inference with Gaussian priors.
- **Ensemble**: Ensembles of several networks with identical structures, which are trained with random initialisation.

Tasks. Three representative tasks with different biosignals to investigate whether the estimated uncertainty works when dataset shift occurs.

- **COVID-19 prediction**. It is a binary classification task, where cough, breathing, and voice sounds are transferred into spectrograms to distinguish the positive from negative participants.
- **Respiratory abnormality detection**. Auscultation of the lung is a part of the clinical examination. A binary classification task is formulated to detect whether a breathing sound segment contains abnormalities, including crackle and wheeze.
- **Heart arrhythmia detection**. ECG is the recording of electrical impulses generated by the heart muscle during beating activity. Through ECG, arrhythmia (irregular beat) can be identified.

Synthetic Dataset Shift and Evaluation Protocol

To assess the quality of the predictive uncertainty yielded by different methods, we propose and apply the following perturbations covering major potential shifts in practice:



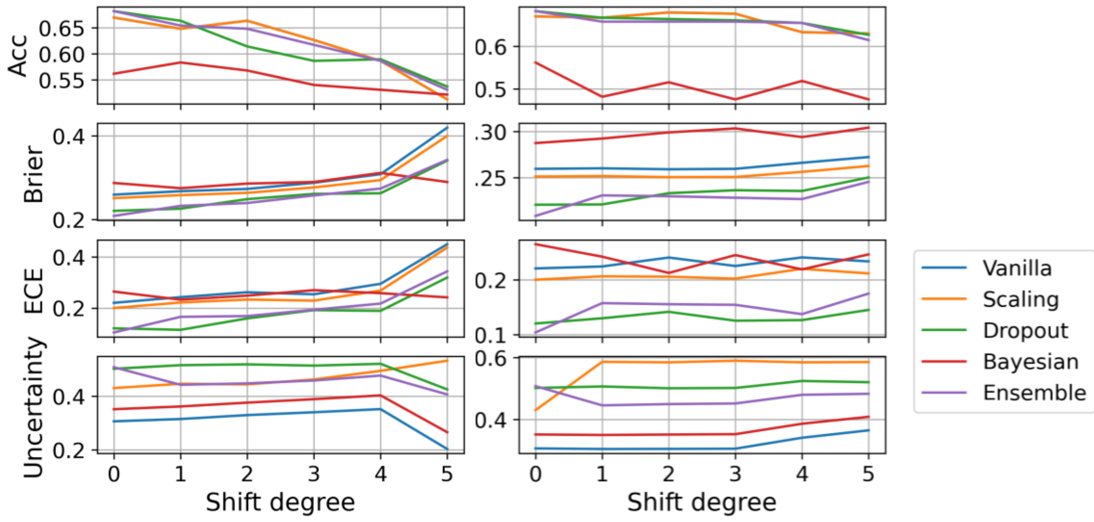
Illustrative examples of generating dataset shift from original signals.

Results and Analysis

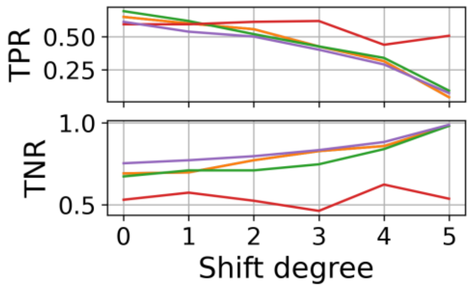
Observations for COVID-19 prediction task. All methods achieve worse performance with the increased shift degree, as Accuracy decays significantly with Brier and ECE showing an upward trend. Yet, Uncertainty does not perform as expected for all the cases. Ensemble is relatively the best regarding Accuracy and Uncertainty. The post-hoc calibration method cannot keep ECE.

Observations for other two tasks. For all methods, Accuracy declines, Brier becomes larger, and for most cases, Uncertainty goes up, indicating that the models are getting more uncertain. Bayesian and Ensemble methods can keep relatively good Accuracy compared to other methods, the slight increase in ECE and the small reduction in Uncertainty suggest that the uncertainty might be not fully reliable.

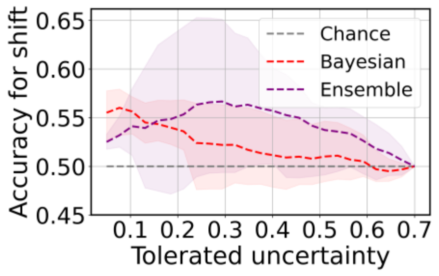
Some empirical analysis and comparison:



Results with Gaussian and distortion shift for COVID-19 task.



Model is biased with shift.



Unable to flag shift.

Summary and Takeaways

- With increasing dataset shift, all uncertainty estimation approaches we evaluated fail to report a reasonable increasing uncertainty to notify the changes in distribution.
- Ensemble can achieve a slightly better uncertainty estimation than the other methods, although it needs relatively heavier computing cost and memory consumption. Bayesian method can obtain similar performance when training data is sufficient.
- Classifiers trained on non-shifted data might be biased on a specific dataset shift during inference. Thus, the measure of prediction uncertainty is as important as the prediction itself, particularly in safety-critical healthcare applications.
- Models may become more and more over-confident as the shift gets severer. None of the existing methods is perfect in capturing distributional shifts and calibrating the deep neural networks. New approaches are needed.

https://arxiv.org/abs/2112.09196

THE 6TH INTERNATIONAL WORKSHOP ON HEALTH INTELLIGENCE (W3PHIAI-22)

tx229@cam.ac.uk