Active multi-fidelity Bayesian online changepoint detection
Supplementary material

Gregory W. Gundersen[1] **Diana Cai**[1] **Chuteng Zhou**[2] **Barbara E. Engelhardt**[1] **Ryan P. Adams**[1]

[1]Department of Computer Science, Princeton University
[2]Arm ML Research Lab

A MODEL DERIVATIONS

A.1 MF-POSTERIOR PREDICTIVE FOR EXPONENTIAL FAMILY MODELS

In multi-fidelity BOCD, we desire the posterior predictive distribution conditioned on the run length,

$$p(\mathbf{x}_t \,|\, r_t = \ell, \zeta_t, \mathbf{D}_{t-\ell:t-1}). \tag{1}$$

Assume this is an exponential family model with the following likelihood and and prior density functions:

$$\begin{aligned} p_{\boldsymbol{\theta}_t}(\mathbf{x}) &= h_1(\mathbf{x}) \exp\left\{\boldsymbol{\theta}_t^\top u(\mathbf{x}) - a_1(\boldsymbol{\theta}_t)\right\}, \quad (2)\\ \pi_{\boldsymbol{\chi},\nu}(\boldsymbol{\theta}_t) &= h_2(\boldsymbol{\theta}_t) \exp\left\{\boldsymbol{\theta}_t^\top \boldsymbol{\chi} - \nu a_1(\boldsymbol{\theta}_t) - a_2(\boldsymbol{\chi}, \nu)\right\}. \quad (3) \end{aligned}$$

See the main text for a description of these terms. We introduce the following notation to denote the data and parameter estimates for the previous ℓ observations, associated with the run length hypothesis $r_t = \ell$:

$$\mathbf{D}^{(\ell)} := \mathbf{D}_{t-\ell:t-1}, \quad \boldsymbol{\chi}_\ell := \boldsymbol{\chi} + \sum_{\tau=t-\ell}^{t-1} \zeta_\tau u(\mathbf{x}_\tau), \quad \nu_\ell := \nu + \sum_{\tau=t-\ell}^{t-1} \zeta_\tau. \tag{4}$$

Then the posterior predictive is

$$\begin{aligned} & p(\mathbf{x}_t \,|\, r_t = \ell, \zeta_t, \mathbf{D}^{(\ell)}) & (5)\\ &= \int_{\boldsymbol{\Theta}} p_{\boldsymbol{\theta}}(\mathbf{x}_t)^{\zeta_t} \pi_{\boldsymbol{\chi}_\ell,\nu_\ell}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} & (6)\\ &= \int_{\boldsymbol{\Theta}} [h_1(\mathbf{x}_t)]^{\zeta_t} \exp\left\{\boldsymbol{\theta}^\top \zeta_t u(\mathbf{x}_t) - \zeta_t a_1(\boldsymbol{\theta})\right\} & (7)\\ &\quad h_2(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\chi}_\ell - \nu_\ell a_1(\boldsymbol{\theta}) - a_2(\boldsymbol{\chi}_\ell, \nu_\ell)\right\} \mathrm{d}\boldsymbol{\theta} & (8)\\ &= [h_1(\mathbf{x}_t)]^{\zeta_t} \frac{\int_{\boldsymbol{\Theta}} h_2(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\top [\zeta_t u(\mathbf{x}_t) + \boldsymbol{\chi}_\ell] - a_1(\boldsymbol{\theta}) [\zeta_t + \nu_\ell]\right\} \mathrm{d}\boldsymbol{\theta}}{\exp\{a_2(\boldsymbol{\chi}_\ell, \nu_\ell)\}} & (9)\\ &\overset{\star}{=} [h_1(\mathbf{x}_t)]^{\zeta_t} \frac{\exp\{a_2(\zeta_t u(\mathbf{x}_t) + \boldsymbol{\chi}_\ell, \zeta_t + \nu_\ell)\}}{\exp\{a_2(\boldsymbol{\chi}_\ell, \nu_\ell)\}} & (10)\\ &= [h_1(\mathbf{x}_t)]^{\zeta_t} \exp\{a_2(\zeta_t u(\mathbf{x}_t) + \boldsymbol{\chi}_\ell, \zeta_t + \nu_\ell) - a_2(\boldsymbol{\chi}_\ell, \nu_\ell)\} & (11) \end{aligned}$$

Step $\star$ follows from the previous line because we know the normalizer for the integral. This result is similar to the result on power posteriors for the exponential family [Miller and Dunson, 2018]. However, our approach requires multiple values of powers, which represent data fidelities.

A.2 MULTI-FIDELITY GAUSSIAN MODEL

To simplify notation, we ignore the run length in this section, since it only specifies which data need to be accounted for in the MF-posterior distribution. Consider a univariate[1] Gaussian model with known variance.

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta_t, \sigma_x^2), \quad \theta_t \sim \mathcal{N}(\mu_0, \sigma_0^2). \tag{12}$$

The multi-fidelity likelihood is

$$\prod_{i=1}^{t} p_{\theta_t}(x_i)^{\zeta_i} = \prod_{i=1}^{t} \left[\frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{ -\frac{1}{2\sigma_x^2}(x_i - \theta_t)^2 \right\} \right]^{\zeta_i} \tag{13}$$

$$\propto \prod_{i=1}^{t} \exp\left\{ -\frac{\zeta_i}{2\sigma_x^2}(x_i - \theta_t)^2 \right\} \tag{14}$$

When $\zeta_i < 1$, the variance of $\mathcal{N}(x_i \mid \sigma_x^2/\zeta_i)$ increases, and the fidelity hyperparameter has the natural interpretation of increasing the variance of our model.

The multi-fidelity posterior is the product of $t+1$ independent Gaussian densities, which is itself Gaussian:

$$\pi(\theta_t \mid \mathbf{D}_{1:t}) \propto \mathcal{N}(\theta_t \mid \mu_0, \sigma_0^2) \prod_{i=1}^{t} \mathcal{N}(x_i \mid \theta_t, \sigma_x^2/\zeta_i) \tag{15}$$

$$\propto \mathcal{N}(\theta_t \mid \mu_t, \sigma_t^2), \tag{16}$$

where

$$\frac{1}{\sigma_t^2} = \frac{1}{\sigma_0^2} + \sum_{i=1}^{t} \frac{\zeta_i}{\sigma_x^2}, \quad \mu_t = \sigma_t^2 \left(\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{t} \frac{\zeta_i x_i}{\sigma_x^2} \right). \tag{17}$$

The MF-posterior predictive can be computed by integrating out θ_t. This is a convolution of two Gaussians, the posterior and the prior $\pi(\theta) = \mathcal{N}(\theta \mid \mu_0, \sigma_0^2)$, which is again Gaussian:

$$p(x_{t+1} \mid \zeta_{t+1}, \mathbf{D}_{1:t}) = \int_{\Theta} [\mathcal{N}(x_{t+1} \mid \theta_t, \sigma_x^2)]^{\zeta_{t+1}} \mathcal{N}(\theta_t \mid \mu_t, \sigma_t^2) \mathrm{d}\theta_t \tag{18}$$

$$= \mathcal{N}\left(x_{t+1} \mid \mu_t, \frac{\sigma_x^2}{\zeta_{t+1}} + \sigma_t^2 \right). \tag{19}$$

With a single fidelity and $\zeta = 1$, this results reduces to the standard result for Gaussian models with known variance [Murphy, 2007].

A.3 MULTI-FIDELITY BERNOULLI MODEL

To simplify notation, we ignore the run length in this section, since it only specifies which data need to be accounted for in the MF-posterior distribution. Consider a beta-Bernoulli model

$$x_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_t), \quad \theta_t \sim \text{Beta}(\alpha_0, \beta_0). \tag{20}$$

The multi-fidelity likelihood is

$$\prod_{i=1}^{t} p_{\theta_t}(x_i)^{\zeta_i} = \prod_{i=1}^{t} \left[\theta_t^{x_i} (1 - \theta_t)^{1 - x_i} \right]^{\zeta_i} \tag{21}$$

$$= \prod_{i=1}^{t} \theta_t^{\zeta_i x_i} (1 - \theta_t)^{\zeta_i (1 - x_i)}. \tag{22}$$

[1] This result straightforwardly extends to the multivariate Gaussian.

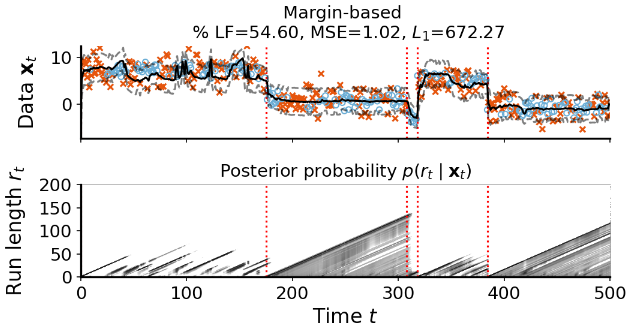

Figure 1: Orange x marks and blue circles denote low- and high-fidelity data respectively. A two-fidelity model that actively selects the lower fidelity if its information gain is close in value to the higher fidelity's information gain (Equation (31)).

Therefore the MF-posterior is

$$\pi(\theta_t)\prod_{i=1}^{t} p_{\theta_t}(x_i)^{\zeta_i} \propto \frac{1}{\mathrm{B}(\alpha_0,\beta_0)}\theta_t^{\alpha_0-1}(1-\theta_t)^{\beta_0-1}\prod_{i=1}^{t}\theta_t^{\zeta_i x_i}(1-\theta_t)^{\zeta_i(1-x_i)} \tag{23}$$

$$\propto \theta_t^{\alpha_0-1+\sum_t \zeta_i x_i}(1-\theta_t)^{\beta_0-1+\sum_t \zeta_i - x_i\zeta_i}. \tag{24}$$

So the MF-posterior is proportional to a beta distribution

$$\begin{aligned}\pi(\theta_t \mid \mathbf{D}_{1:t}) &= \mathrm{Beta}(\alpha_t,\beta_t),\\ \alpha_t &\coloneqq \alpha_0 + \sum_{i=1}^{t}\zeta_i x_i,\\ \beta_t &\coloneqq \beta_0 + \sum_{i=1}^{t}\zeta_i(1-x_i).\end{aligned} \tag{25}$$

The MF-posterior predictive is:

$$p(x_{t+1} \mid \zeta_{t+1}, \mathbf{D}_{1:t}) \tag{26}$$

$$= \int_0^1 p_{\theta_t}(x_{t+1})^{\zeta_{t+1}} p(\theta_t \mid \mathbf{D}_{1:t})\mathrm{d}\theta_t \tag{27}$$

$$= \int_0^1 \left(\theta_t^{x_{t+1}}(1-\theta_t)^{1-x_{t+1}}\right)^{\zeta_{t+1}}\left(\frac{1}{\mathrm{B}(\alpha_t,\beta_t)}\theta_t^{\alpha_t-1}(1-\theta_t)^{\beta_t-1}\right)\mathrm{d}\theta_t \tag{28}$$

$$= \frac{1}{\mathrm{B}(\alpha_t,\beta_t)}\int_0^1 \theta_t^{\zeta_{t+1}x_{t+1}+\alpha_t-1}(1-\theta_t)^{\zeta_{t+1}(1-x_{t+1})+\beta_t-1}\mathrm{d}\theta_t \tag{29}$$

$$= \frac{\mathrm{B}(\alpha_t+\zeta_{t+1}x_{t+1}, \beta_t+\zeta_{t+1}(1-x_{t+1}))}{\mathrm{B}(\alpha_t,\beta_t)}. \tag{30}$$

The last step as, as in the general case, depends on knowing the normalizer of the beta distribution. Notice that the base measure $h_1(x_t)$ of the Bernoulli distribution is one, and therefore $[h_1(x_t)]^{\zeta_t} = 1$.

B ALTERNATIVE DECISION RULE

Here, consider the scenario of just two fidelities, low fidelity ζ_{low} and high fidelity ζ_{high}. An alternative decision rule to the one in the main text would be to choose the lower fidelity when its utility or information gain is within some margin hyperparameter δ of the higher fidelity's utility:

$$\zeta_t^\star = \begin{cases}\zeta_{\mathsf{low}} & \text{if } |\mathcal{U}(\zeta_{\mathsf{low}}) - \mathcal{U}(\zeta_{\mathsf{high}})| < \delta,\\ \zeta_{\mathsf{high}} & \text{otherwise.}\end{cases} \tag{31}$$

However, we found that results on the Gaussian model in the main text were not promising (Figure 1). The model would frequently switch between fidelities because the utilties $\mathcal{U}(\zeta_{\mathsf{low}})$ and $\mathcal{U}(\zeta_{\mathsf{high}})$ were quite close in value. We found that information rate was more stable because it requires a more significant change in information gain to induce a switch.

C MF-BOCD ALGORITHM IN DIDACTIC CODE

This Python code is a didactic example of the MF-BOCD algorithm. At each time step, the algorithm (1) chooses a data fidelity using maximal information rate; (2) observes a datum of the chosen fidelity; (3-4) computes the posterior predictive and run-length posterior distributions; (5) updates the model parameters; and (6) makes a prediction. Please see the code repository[2] for a complete example.

Note that in practice, each datum will be observed by evaluating an observation model in real-time. Here, for clarity, we simply index into a pre-initialized data array.

```python
import numpy as np
from   scipy.special import logsumexp

def mf_bocd(data, model, hazard, costs):
    J, T        = data.shape
    log_message = np.array([1])
    log_R       = np.ones((T+1, T+1))
    log_R[0, 0] = 1
    pmean       = np.zeros(T)
    igs         = np.empty(J)
    choices     = np.empty(T)

    for t in range(1, T+1):

        # 1. Choose fidelity.
        rl_post = np.exp(log_R[t-1, :t])
        for j in range(J):
            igs[j] = compute_info_gain(t, model, rl_post, log_message, hazard, j)
        j_star = np.argmax(igs / costs)
        choices[t-1] = j_star

        # 2. Observe new datum.
        x = data[j_star, t-1]

        # 3. Compute predictive probabilities.
        log_pis = model.log_pred_prob(t, x, j_star)

        # 4. Estimate run length distribution.
        log_growth_probs = log_pis + log_message + np.log(1 - hazard)
        log_cp_prob      = logsumexp(log_pis + log_message + np.log(hazard))
        new_log_joint    = np.append(log_cp_prob, log_growth_probs)
        log_R[t, :t+1]   = new_log_joint
        log_R[t, :t+1]  -= logsumexp(new_log_joint)

        # 5. Update model parameters and message pass.
        model.update_params(t, x, j_star)
        log_message = new_log_joint

        # 6. Predict.
        pmean[t-1] = np.sum(model.mean_params[:t] * rl_post)

    return choices, np.exp(log_R), pmean
```

[2] https://github.com/princetonlips/mf-bocd

D ABLATION STUDIES

Here, we report the results of an ablation study for the multi-fidelity Gaussian and multi-fidelity Bernoulli models. For varying costs, a multi-fidelity model using information gain-based switching was run on data generated from their respective data generating proceses. The percentage of low-fidelity observations was recorded; call this P_{low}. Then a randomized multi-fidelity model was run on the same dataset. At each time step, the randomized model chose low-fidelity data based on a Bernoulli random variable with bias P_{low}. The goal of this experiment is to demonstrate that when the model switches to high-fidelity data is important to model performance, not just the fact that some percentage of high-fidelity data are used. We found that for both Gaussian (Table 1) and Bernoulli data (Table 2), choosing when to switch fidelities was often useful.

Table 1: Ablation study for multi-fidelity Gaussian models. “LF only” is BOCD using only low-fidelity data. Mean and two standard errors, representing 95% confidence intervals, are reported over 200 trials. Bold numbers indicate statistically significant using 95% confidence intervals.

	MSE			L_1		
LF (%)	LF only	Random	Info-based	LF only	Random	Info-based
1	0.879 (0.034)	0.046 (0.056)	0.003 (0.001)	270.87 (8.35)	**5.98** (3.06)	73.92 (9.61)
2		0.125 (0.073)	0.111 (0.046)		**18.18** (7.37)	77.68 (9.79)
38		0.680 (0.118)	**0.494** (0.059)		162.31 (12.97)	161.05 (11.08)
53		0.702 (0.091)	**0.483** (0.066)		183.40 (11.36)	173.01 (10.46)
60		0.752 (0.140)	**0.452** (0.037)		186.11 (10.89)	174.95 (10.13)
67		0.665 (0.075)	**0.466** (0.036)		187.72 (10.01)	173.41 (9.91)
74		0.643 (0.064)	**0.480** (0.043)		182.18 (9.48)	175.88 (9.36)
80		0.656 (0.087)	**0.492** (0.044)		184.66 (9.20)	175.70 (9.20)
97		0.547 (0.028)	0.537 (0.028)		176.76 (9.40)	175.34 (9.33)

Table 2: Ablation study for multi-fidelity Bernoulli models. “LF only” is BOCD using only low-fidelity data. Mean and two standard errors, representing 95% confidence intervals, are reported over 200 trials. Bold numbers indicate statistically significant using 95% confidence intervals.

	MSE			L_1		
LF (%)	LF only	Random	Info-based	LF only	Random	Info-based
9	0.123 (0.009)	0.003 (0.001)	**0.002** (0.000)	186.27 (7.02)	45.55 (6.04)	40.31 (5.43)
21		0.008 (0.001)	0.009 (0.002)		76.42 (7.68)	71.88 (7.54)
25		0.011 (0.002)	0.011 (0.002)		84.47 (8.11)	80.61 (7.89)
46		0.025 (0.003)	0.021 (0.003)		124.80 (7.07)	117.34 (7.31)
61		0.040 (0.004)	0.034 (0.005)		143.87 (6.34)	139.88 (7.23)
68		0.050 (0.005)	**0.040** (0.005)		158.48 (6.34)	149.79 (7.02)
73		0.057 (0.005)	0.048 (0.006)		163.68 (6.39)	158.08 (6.66)
83		0.077 (0.006)	**0.064** (0.007)		174.01 (6.21)	170.26 (6.49)
90		0.098 (0.007)	**0.082** (0.007)		184.46 (6.11)	178.86 (6.28)

E EXPERIMENTAL DETAILS

E.1 CAMVID EXPERIMENTS

The pretrained MobileNets were downloaded from the Fastseg Python library.[3]

We can estimate the computational cost of MF-BOCD (λ_{MF}) relative to BOCD using only high- (λ_{HF}) and low- (λ_{LF}) fidelity data. We used 85 low- and 86 high- fidelity observations. The low- (high-) fidelity observation model required 19.48 (36.89) billion flops (Table 3). Computing the information gain required 465,291 flops. The total cost of our algorithm in billions of flops is

$$\begin{aligned}
\lambda_{\mathsf{LF}} &= 171 \times 19.5 \approx 3333, \\
\lambda_{\mathsf{HF}} &= 171 \times 36.9 \approx 6303, \\
\lambda_{\mathsf{MF}} &= 0.00046 + (85 \times 19.5) + (86 \times 36.7) \approx 4827.
\end{aligned}$$

As we can see, decision-making has a marginal cost.

Table 3: Observation model details for CamVid and MIMII experiments. (CamVid) The high-fidelity model has roughly twice times the number of flops and higher accuracy as measured by intersection-over-union (IoU) on the Cityscapes dataset. (MIMII) The high-fidelity model requires roughly 250 times as many floating point operations (ops). “FC”, “M”, and “B” mean fully-connected, millions, and billions respectively.

	Fidelity	Model	Ops	Accuracy
CamVid	HF	V3-large	36.86 B	72.3 (IoU%)
	LF	V3-small	19.48 B	67.4 (IoU%)
MIMII	HF	MicroNet-AD(M)	124.7 M	96.15 (AUC%)
	LF	Two-layer FC	0.5 M	86.7 (AUC%)

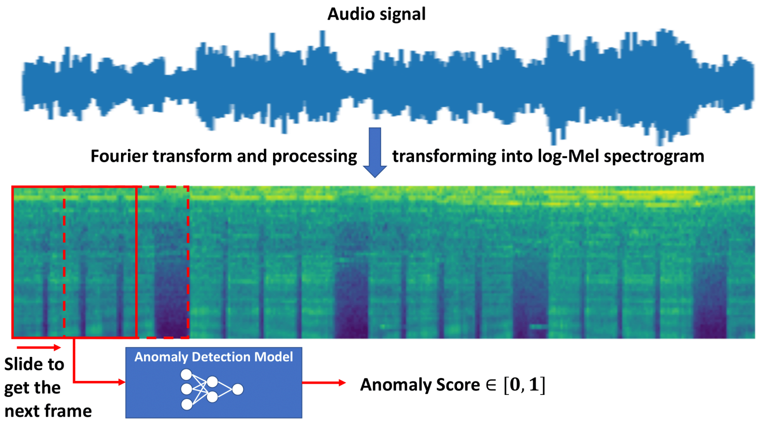
Audio signal
Fourier transform and processing
transforming into log-Mel spectrogram
Slide to get the next frame
Anomaly Detection Model
Anomaly Score ∈ [0, 1]

Figure 2: Illustration of pipeline to generate anomaly scores from log-Mel spectrograms using deep neural networks.

E.2 MIMII EXPERIMENT

In the MIMII experiment, the output of the observation models (Table 3) is a scalar anomaly score in the range $[0, 1]$, where 0 indicates normal machine operation. An illustration of how these scores are obtained for an audio clip is shown in Figure 2. To convert these anomaly scores to binary numbers for a Bernoulli multi-fidelity posterior predictive model, we thresholded the scores to integers in $\{0, 1\}$. The quality of the observation models depends on the choice of threshold. For examples of these data, see Figure 3. To select the appropriate threshold, we used the intersection of the false negative and false positive

[3] https://github.com/ekzhang/fastseg

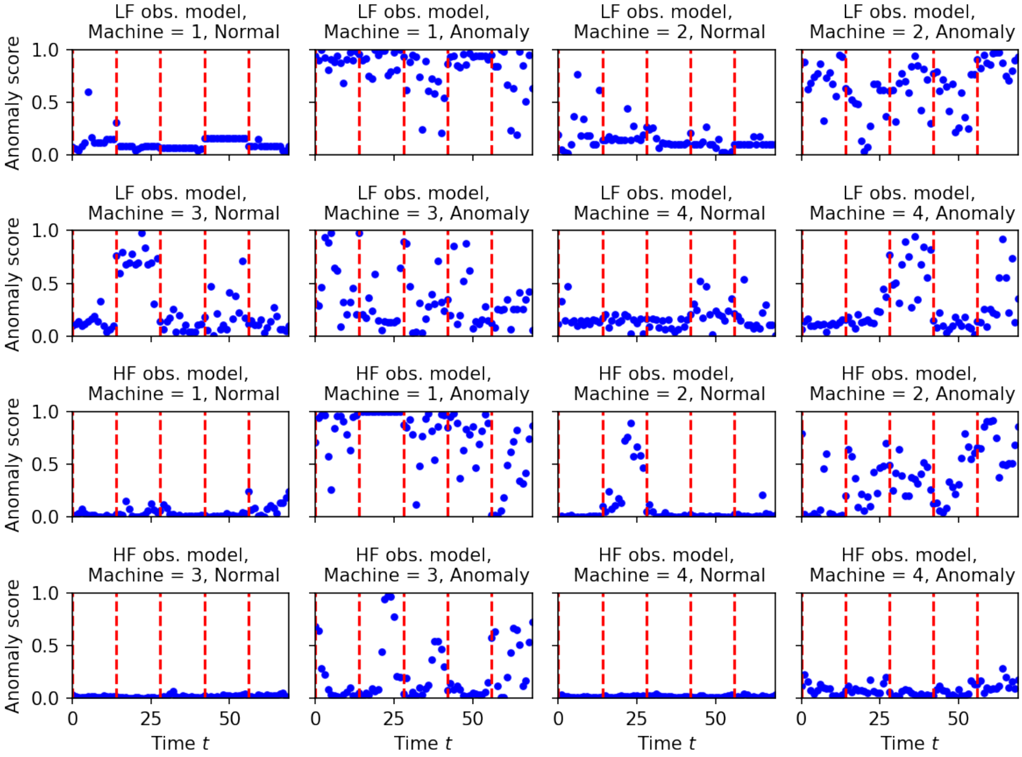

Figure 3: Examples of MIMII anomaly scores, five audio clips for each machine. Dashed red lines separate audio clips.

rate curves, which corresponds to the top-left corner of the receiver operating characteristic (ROC) curves for each machine and each observation model (Figure 4).

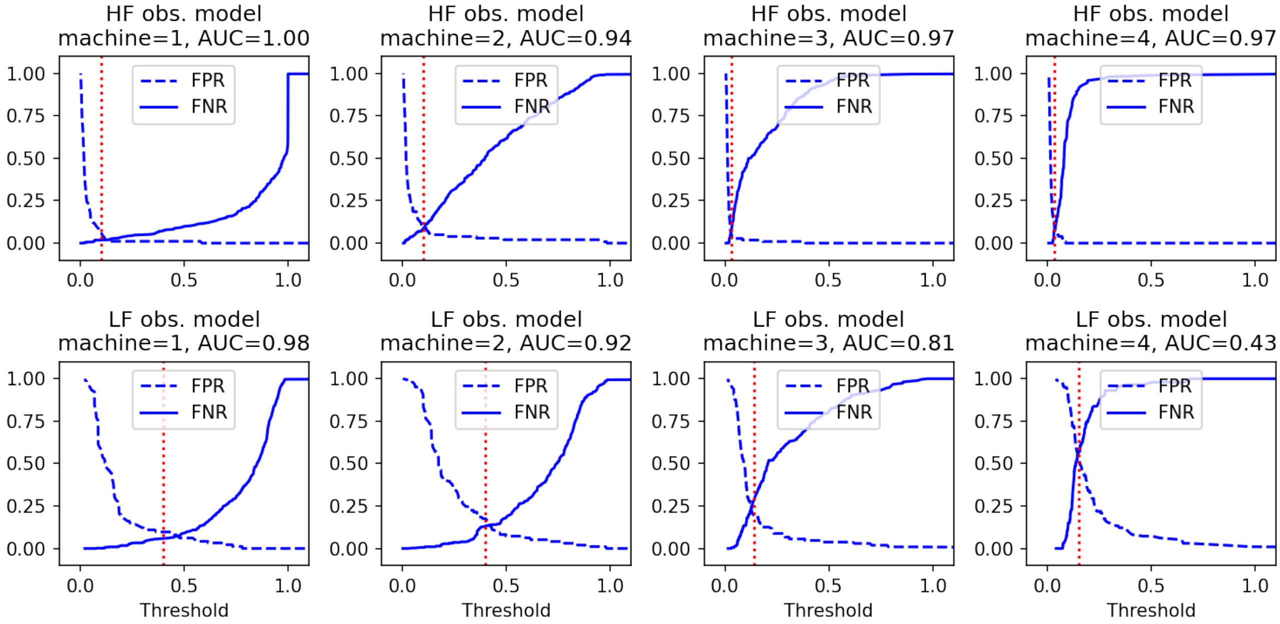

Figure 4: False positive (FPR) and false negative rates (FNR) for high- (HF) and low- (LF) fidelity observation models on MIMII cross-validation data. Vertical dashed lines indicated the chosen threshold

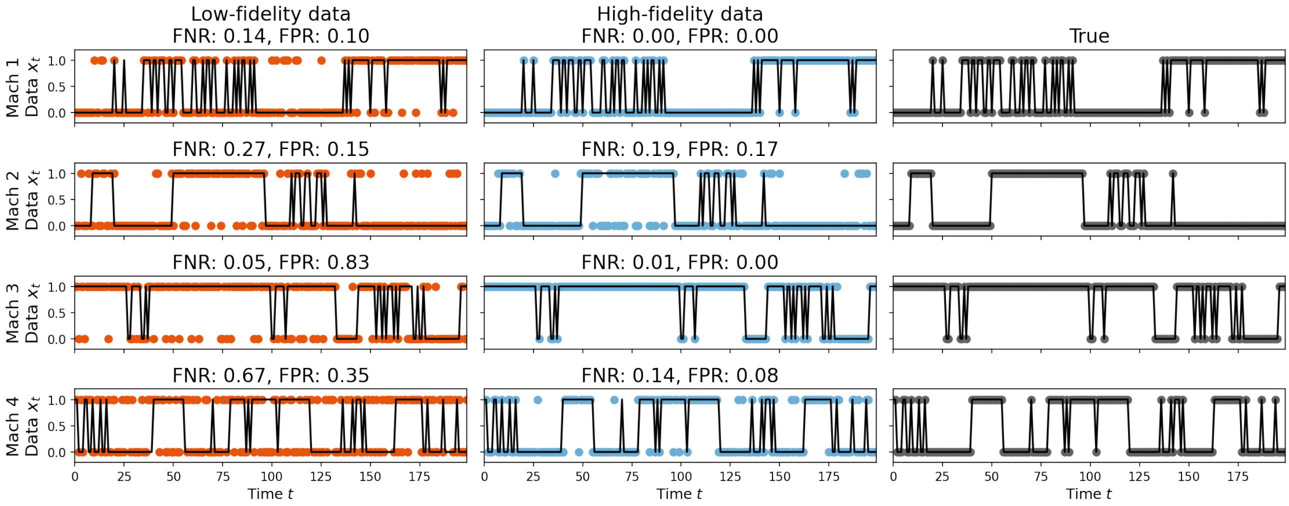

Figure 5: Illustration of MIMII data after converting log-Mel spectrograms to binary numbers with machine- and observation model-specific thresholds. The true binary value is denoted with a black line.

References

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1($2\sigma 2$):16, 2007.