Variational Autoencoders: Wearable Respiration Monitoring Example

ELG 5218 - Uncertainty Evaluation in Engineering Measurements and Machine Learning

Miodrag Bolic

University of Ottawa

February 25, 2026

Motivation (why unsupervised?)

Wearable respiration signals are messy

- Real-world recordings contain motion artifacts, drift, sensor slippage, and changing posture.
- Labels for apnea/irregular breathing are expensive (clinical scoring) and often incomplete.

Approach

Train a VAE on **normal** breathing windows only; use an **anomaly score** at test time.

Data: what does x look like?

Typical sources

- Chest belt (inductive plethysmography): expansion/contraction waveform
- IMU on torso: low-frequency motion correlated with respiration
- PPG-derived respiration: amplitude/frequency modulation proxies

Windowing

Segment into fixed-length windows (e.g., 10–20s):

$$\boldsymbol{x} \in \mathbb{R}^T, \quad T = \text{samples per window.}$$

Preprocessing (minimal but practical)

A simple pipeline

1. Detrend / remove slow drift (high-pass or polynomial)
2. Optional bandpass around respiration (e.g., 0.1–0.7 Hz)
3. Normalise per window (z-score) to reduce amplitude scale issues

In the notebook we simulate belt-like signals and apply light normalisation.

Generative model and encoder

Latent-variable model

$$p_{\theta}(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})\, p_{\theta}(\boldsymbol{x} \mid \boldsymbol{z}), \qquad p(\boldsymbol{z}) = \mathcal{N}(0, I).$$

Variational encoder

$$q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\mu_{\phi}(\boldsymbol{x}), \operatorname{diag}(\sigma^2_{\phi}(\boldsymbol{x}))\right).$$

Training objective (normal windows only)

ELBO

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\big] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z}))\,.$$

- For signals, a Gaussian decoder implies reconstruction loss $\propto \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$.
- KL is closed-form for diagonal Gaussians.

How the encoder learns: reparameterisation

Pathwise trick

$$\epsilon \sim \mathcal{N}(0, I), \qquad \boldsymbol{z} = \mu_\phi(\boldsymbol{x}) + \sigma_\phi(\boldsymbol{x}) \odot \epsilon.$$

- Sampling becomes a differentiable transformation.
- Enables low-variance gradients for ϕ.

Abnormal patterns we want to catch

Clinically relevant

- **Apnea-like**: near-flat respiration amplitude for multiple seconds
- **Erratic**: irregular rate/amplitude (stress, obstruction, arousal)
- **Tachypnea / bradypnea**: sustained too fast/slow breathing

Wearable nuisance

- Motion bursts (turning in bed, walking)
- Belt slippage / baseline shift
- Low SNR periods (loose sensor contact)

Anomaly score (window-level)

Score

$$S(\boldsymbol{x}) = \mathbb{E}_{q_\phi}[-\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] + \beta\, \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z}))\,.$$

- High $S(\boldsymbol{x})$ indicates "hard to explain" under the normal-breathing model.
- Choose threshold via validation, or via a percentile of scores on normal training data.

Architecture sketch (small 1D Conv VAE)

Encoder

Conv1D layers $\rightarrow$ pooling/flatten $\rightarrow$ linear heads:

$$\boldsymbol{x} \mapsto (\mu_{\phi}(\boldsymbol{x}), \log \sigma^2_{\phi}(\boldsymbol{x})).$$

Decoder

Latent $\boldsymbol{z} \rightarrow$ upsampling/deconvs $\rightarrow \hat{\boldsymbol{x}}$.

Notebook demo (what you can show live)

In `vae_wearable_breathing_case_study.ipynb`

1. Simulate wearable respiration windows: **normal** + **apnea** + **erratic** + **motion** + **low SNR**
2. Train VAE on normal windows only
3. Visualise reconstructions (normal vs abnormal)
4. Score histograms + ROC/AUC for anomaly detection

Results at a glance

Operating point (calibrated on held-out normal)

- ROC AUC (score): **1.000**
- Threshold τ (99th percentile of normal): **14.597**
- Empirical FPR on held-out normal: **0.010**
- Empirical TPR on abnormal set: **0.995**

Per-class detection rate at τ

Anomaly type	**Detection rate**	**n**
Apnea-like suppression	1.000	800
Erratic breathing	0.980	800
Motion artifact	1.000	800
Low SNR	1.000	800

Note: AUC measures ranking across thresholds; the chosen τ is a conservative operating point to control false alarms.

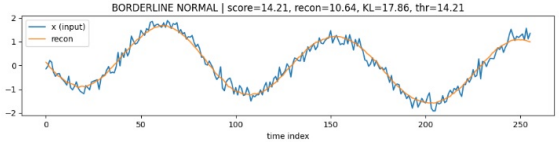
Training diagnostics (objective, recon, KL)

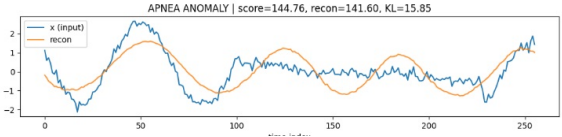
What we want to see

- Objective decreases / ELBO-like improves
- Recon improves on train; val tracks with noise
- KL stays stable (no posterior collapse or explosion)

Observed

- Train recon improves; val is noisier but reasonable
- KL gradually decreases and train/val track closely
- Overall: stable convergence for scoring





What drives detection? (score decomposition)

Score

$$S(x) = \underbrace{\text{ReconError}(x)}_{\text{fit to normal waveform}} + \beta \underbrace{\mathrm{KL}(q_\phi(z \mid x) \,\|\, p(z))}_{\text{latent "unusualness"}}.$$

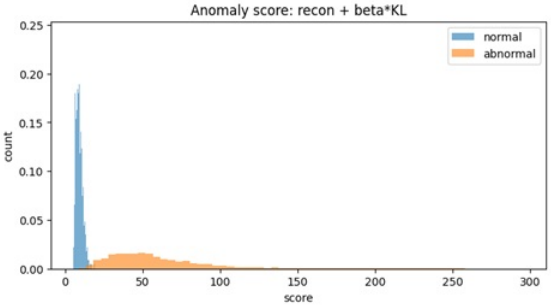
Key observation

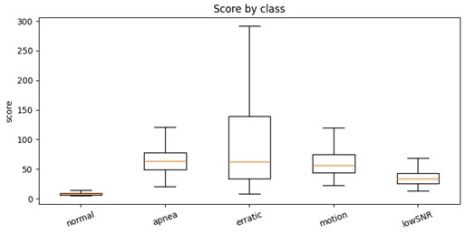
In these results, **reconstruction error dominates separation**:

- Score and recon distributions are near-disjoint (normal vs abnormal).
- KL overlaps more across classes; it helps regularize and adds a small signal.

Practical takeaway: interpret a high score as "hard to reconstruct under the learned normal manifold," not necessarily "far in latent space."

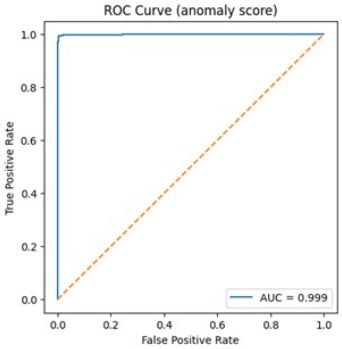
Score distributions and per-class behavior





Erratic breathing shows the widest spread; low-SNR tends to be closer to normal than motion/apnea, but still elevated.

ROC and threshold calibration



Why AUC can be 1.0 while TPR at τ is < 1

AUC reflects *ranking* across all thresholds; τ is chosen to enforce a target false-alarm rate on normal windows.

Discussion: what a high score means & how to use it

Meaning of high score

A high $S(x)$ means the window is **poorly explained by the normal-breathing model**: either reconstruction fails, or the encoder posterior becomes unusual (higher KL), or both.

How to detect events in practice

- Window-level: flag if $S(x) > \tau$ (threshold from held-out normals or desired FPR).
- Event-level: smooth scores and require **K consecutive flagged windows** (reduces false alarms).
- Inspect recon/residuals to distinguish physiology vs sensor artifact.

Limitations and next steps

- Synthetic/demo data can be optimistic; validate with subject-wise splits and device/session shifts.
- Add PR curves (rare anomalies), robustness tests (drift, clipping, missing data), and a β sweep.
- If long-range dynamics matter, consider sequential VAEs (e.g., LSTM-VAE / VRNN).

Takeaways

Key message

A VAE provides a principled way to learn a **normal-breathing manifold** via the encoder $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, then detect deviations with a simple, interpretable score.

Extensions: multichannel inputs (belt+IMU), spectral losses (STFT), conditional VAE with posture/activity covariates, patient-specific fine-tuning.