

# Machine Learning for longitudinal data analysis in gas sensors: A case study in early subclinical ketosis in dairy COWS

Albert Belenguer-Llorens<sup>1,2</sup>, Jonas Hendricks<sup>3</sup>, Jaehyun Ko<sup>4</sup>, Anna-Maria Reiche<sup>3</sup>, Frigga Dohme-Meier<sup>3</sup>, Andreas T. Güntner<sup>4</sup>, Catherine R. Jutzeler<sup>1,2</sup>

<sup>1</sup> *Department of Health Sciences and Technology (D-HEST), ETH Zurich, Switzerland*

<sup>2</sup> *SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*

<sup>3</sup> *Ruminant Nutrition and Emissions, Agroscope, Posieux, Switzerland*

<sup>4</sup> *Human-centered Sensing Laboratory, Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland*  
jutzelec@ethz.ch

## Summary:

Machine learning offers powerful opportunities to transform continuous sensor measurements into clinically or biologically predictions. However, translating raw sensor signals into robust diagnostic models requires careful consideration of temporal dynamics, feature engineering, latent-state modeling, and interpretability. In this tutorial, we present a generalizable pipeline for developing predictive diagnostic models from sensor measurements, illustrated through a real-world case study: the early detection of subclinical ketosis in dairy cows using breath acetone measured by gas sensors.

**Keywords:** Gas Sensors, Machine Learning, Dairy science, Sensor Data Analysis, XGBoost

## Background, Motivation and Objective

Modern sensor systems generate increasingly rich time-series data, yet many sensing applications still rely on threshold-based or linear analyses that underutilize the available information. Machine learning (ML) enables the identification of dynamic patterns that can improve predictive performance while providing insight into the underlying system dynamics.

This tutorial presents a practical workflow for converting longitudinal sensor measurements into predictive models. The proposed framework addresses common challenges encountered in data analysis for sensing applications, including temporal windowing, feature engineering, latent-state inference, validation under limited sample sizes, and model interpretability.

To illustrate the methodology, we use a case study on early diagnosis of subclinical ketosis (SCK) in dairy cows. SCK is a prevalent metabolic disorder that significantly affects milk production, animal welfare, and farm economics [1]. Current diagnosis relies on invasive and expensive blood  $\beta$ -hydroxybutyrate (BHB) measurements, whereas breath acetone provides a

promising non-invasive biomarker measurable by gas sensors [2]. The example demonstrates how sensor measurements can be transformed into actionable predictions several days before disease onset using a ML-based pipeline.

## Materials and Data Collection

The proposed pipeline consists of five consecutive stages:

- **Sensor Data Acquisition:** Collection of longitudinal gas sensor measurements together with reference clinical labels (SCK or healthy) based on reference BHB measurements.
- **Temporal Representation:** Construction of rolling temporal windows that preserve the evolution of the sensor signal and define prediction horizons relevant for prospective diagnosis.
- **Feature Engineering:** Extraction of dynamic descriptors including slopes, rates of change, acceleration, volatility, cumulative exposure (KAUC), and

trend characteristics that summarize acetone behavior over time.

- **Latent State Modeling:** Use of Hidden Markov Models (HMM) to identify latent physiological states linking sensor measurements with the underlying biological process. During training, both sensor (acetone) and reference measurements (BHB) calibrate the latent states, while inference relies only on sensor measurements.
- **Predictive ML:** Integration of temporal measurements, engineered features, and latent-state transition probabilities into an interpretable XGBoost classifier validated through nested cross-validation and explained using SHapley Additive exPlanations (SHAP) values [4].

Although demonstrated using breath acetone sensing, this methodology is applicable to many longitudinal sensor systems where complex temporal dynamics encode predictive information.

### Case Study: Early Detection of Subclinical Ketosis

The case study uses breath acetone measurements collected with a novel sampling mask developed by ETH Zurich and Agroscope [3] together with blood BHB measurements from 31 dairy cows monitored before and after calving.

Using seven-day temporal windows and prediction horizons of 7–9 days, the proposed pipeline learns dynamic signatures preceding disease onset. The resulting predictive model achieved balanced and competitive performance across validation folds in both Area Under the Curve (AUC) and balanced accuracy, demonstrating that temporal sensor patterns contain substantial predictive information beyond instantaneous measurements (Table 1).

Tab. 1: Classification performance (mean  $\pm$  std) of our proposed algorithm stratified by group. AUC: Area under the curve; ACC: Accuracy

| Cohort | AUC               | Balanced Acc      |
|--------|-------------------|-------------------|
| ALL    | 0.785 $\pm$ 0.071 | 0.778 $\pm$ 0.027 |
| CON    | 0.832 $\pm$ 0.043 | 0.801 $\pm$ 0.034 |
| SCK    | 0.744 $\pm$ 0.027 | 0.734 $\pm$ 0.124 |

Interpretability analysis using SHAP identified cumulative exposure (KAUC), acceleration, and temporal lag measurements among the most relevant predictors, illustrating how ML can simultaneously improve prediction and generate physiological hypotheses (Figure 1).

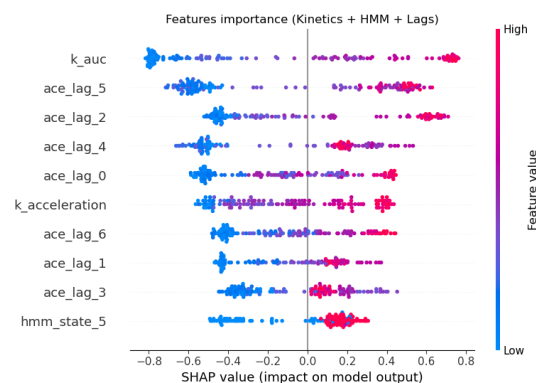


Fig. 1. SHAP values of the top 10 most relevant features found by XGBoost.

The proposed pipeline is computationally efficient and lightweight, making it suitable for deployment on resource-constrained embedded devices. Its transparent design and interpretability ensure compliance with the EU AI Act for high-risk AI support systems, supporting its potential translation to a broad range of medical and diagnostic applications.

### Conclusions and Future Lines

This work presents a generalizable ML-based framework to extract predictive patterns from sensor measurements. By combining temporal feature engineering, probabilistic latent-state modeling, and interpretable ML algorithms, the proposed pipeline provides a practical blueprint for sensor scientists interested in developing predictive algorithms from longitudinal measurement data.

The dairy cow application serves as an illustrative example, but the methodology is broadly transferable to environmental sensing, wearable technologies, industrial monitoring, breath analysis, and other sensor-driven diagnostic applications. Future work will focus on extending the framework to larger datasets, additional ML architectures, and multimodal sensing platforms.

### References

- [1] Mostert, P. F., et al. Estimating the economic impact of subclinical ketosis in dairy cattle using a dynamic stochastic simulation model. *Animal* 12.1 (2018): 145-154; doi: 10.1017
- [2] Rico, J. Eduardo, et al. Invited review: Ketone biology—The shifting paradigm of ketones and ketosis in the dairy cow. *Journal of Dairy Science* 107.6 (2024): 3367-3388; doi: 10.3168
- [3] Ko, Jaehyun, et al. Mask for Individual Sampling of Tidal and Ruminal Gas Contaminated Breath in Dairy Cows. (2026); chemrxiv; doi: 10.26434
- [4] Lundberg, Scott M., and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).