Optimization of Biosensor Waveform Preprocessing

Siemens Healthineers

Capstone Proposal

Dylan Longert
Jinxin (Jessie) Wang
Nan Tang
Nayeli Montiel Rodríguez
Han (Eden) Chen

Introduction and Background

Biosensor technologies play a pivotal role in the medical and pharmaceutical sectors, offering swift and precise measurements of specific biomarkers. These instruments are indispensable for healthcare, facilitating real-time monitoring and diagnosis. Biosensors comprise a biological element that interacts with an analyte of interest, triggering a biochemical reaction. The biosensor transducer then converts this reaction into a measurable signal. Notably, biosensors excel at isolating analytes of interest from other substances during the biochemical process.

Among the diverse range of biosensors, our project will be focusing on potentiometric biosensors. In potentiometric biosensors, the transducer converts the biochemical interaction into an electric potential measurement, typically between a reference sample and the target analyte. These biosensors have already demonstrated their utility in various medical applications. Potentiometric sensors have demonstrated diverse applications across various industries. In the dairy sector, they have been employed to measure urea concentration in cow urine, as highlighted by Trivedi et al. [1]. Similarly, in the medical field, these sensors play a crucial role in the early detection of Alzheimer's disease by detecting amyloid β-42, as illustrated by Ribeiro et al. [2]. Our project involves evaluating the differences between two generations of epoc® Blood Analysis System produced by the Siemens Healthineers research and development department (further explained below).

The epoc® Blood Analysis System [3] is a portable blood analyzer comprising three components: Host, Reader, and Test Card. The epoc Host, as a mobile computer equipped with a well-designed software application, communicates with the epoc Reader that reads and measures electrical signals from the Test Card with the calibration fluid and the sensor module (multiple sensors built-in) to receive potentiometric sensor data and processes this data to calculate analytical values and then displays the test result.

Figure 1. The epoc® Blood Analysis System (System 2)

The calibration fluid rehydrates the sensors as a point value before every blood test, which is sealed in each test card for being released after insertion into the reader. Once the calibration is completed (three minutes after releasement), the sample fluid or blood sample will be injected into the test card and flow through the sensor to provide a new sensor value. The difference in the sensor values between these two windows can be employed to calculate the concentration for each sensor, which can be measured as biosensor waveforms via time in seconds. Our project will analyze the waveforms and compare the differences between the two systems across sensor A and sensor B, which will be detailed in the below section.

Objectives and Methodology

The primary research question that we are aiming to address is how the biosensor waveform characteristics differ between two types of systems of epoc® Blood Analysis built by Siemens Healthineers.

System 1 is the device currently applied in the market and System 2 is the next generation with a new epoc Host, a different Reader, and a different type of Test Card. Tests being performed on these two systems mainly involve three periods which are related to the sensor readings as shown in Figure 2.

Our response of interest is the electrical signal (in millivolts) measured from the potentiometric sensors. We will focus on the characteristics of the response within the calibration window and test sample window because they are considered representatives of the Calibration Fluid Present and Test Sample Fluid Present, respectively, and their differences and the degree of flatness are expected to provide stable sensor values for the Siemens team to calculate their analyte of interest.

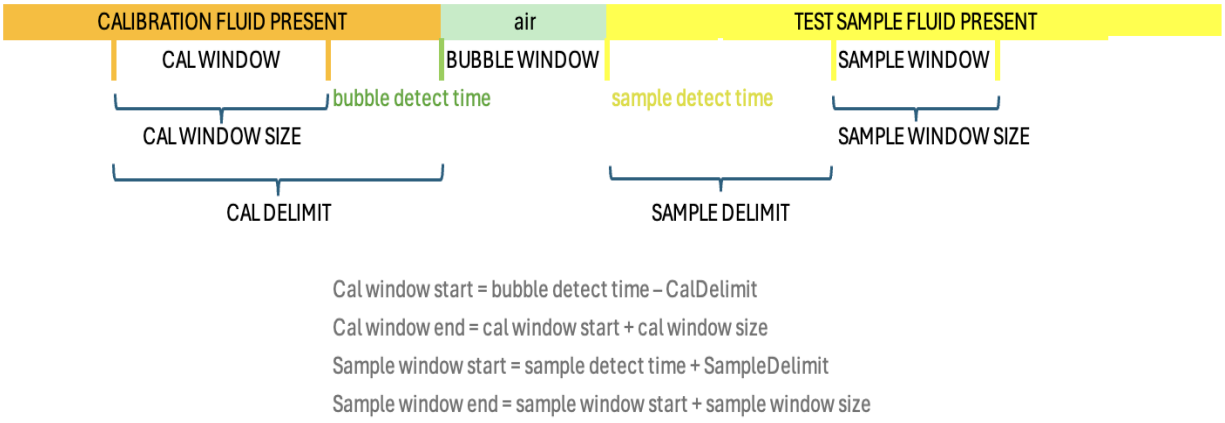

Figure 2. Current window partitioning. “Current Fluid Present” corresponds to the period when the calibration fluid flows through the sensor, “Air” corresponds to the period when the bubble is being detected after sample injection, and “Test Sample Fluid Present” corresponds to the period when the sample is being detected.

We will use Python to write a program to summarize and visualize the trend differences of the waveform between System 1 and System 2. Our proposed methodology is explained in detail below.

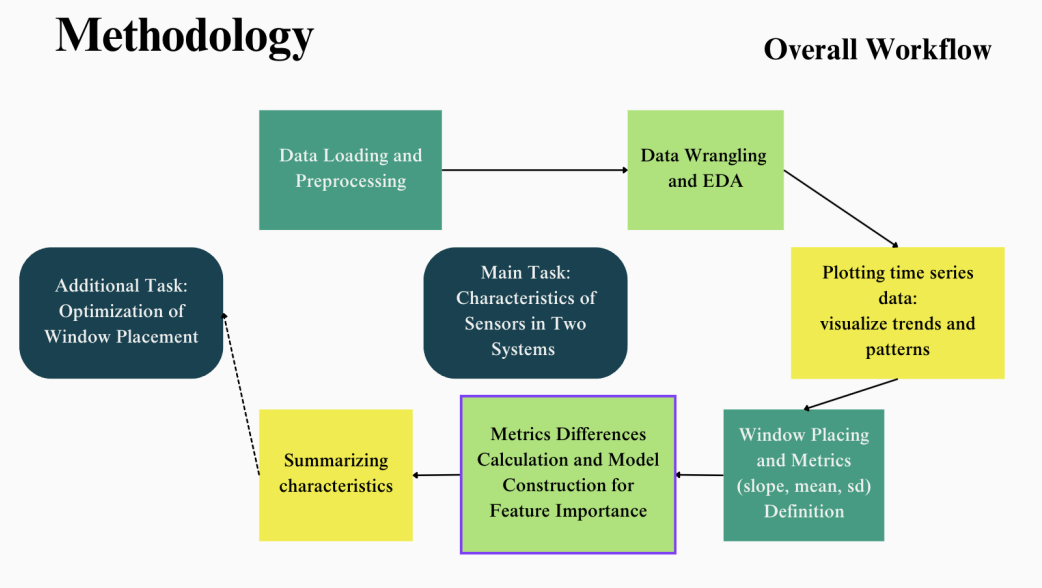

Figure 3. Preliminary Pipeline

First, we will preprocess the data to ensure its quality and prepare it for analysis. This includes handling missing values, removing outliers, and transforming the data as needed. Once the data is prepared, we will perform Exploratory Descriptive Analysis (EDA) to gain a deeper understanding of the sensor data. This will involve calculating statistical summaries (e.g. measures of central tendency) and visualizing the data using tools like histograms.

We will proceed by plotting the average across each TestID. using the sensor time series data from both sensor A and sensor B across systems 1 and 2 [4]. Our dataset includes various features such as fluid types (including blood samples and Eurotrol samples), fluid temperature, and card age, among others.

Then we will put windows in both the calibration period and the sample period by employment of the current window limits information (window size and the respective delimitation) determined by the Siemens team.

To evaluate differences in waveforms within these windows, we will define metrics such as mean, slope, and standard deviation of the response. These metrics serve as indicators of the data's behavior and variability within each window. Then we can start from the calibration window to see if the difference exists; and based on the results we can continue the analysis with the sample window, which may involve fitting curves to find sensor characteristics within the two systems or modeling the difference of the sensor values between two systems by selected features.

Due to the complexity of visualizing every possible combination of these features for each sensor and comparing them, specific time series clustering techniques, like K-means based on Dynamic Time Warping (DTW) can be used here. Traditional K-means uses Euclidean distance, which might not be suitable for time series data. DTW provides time-series-related distance metrics and thus can better capture the time shifts. [5] This clustering method will help us group similar tests into a smaller number of clusters, making tests within each cluster more alike than those in different clusters. This approach enables us to efficiently capture the unique characteristics of each cluster. Subsequently, we will visualize the pattern differences in responses for each cluster. For example, if we identify six distinct clusters for sensor A after clustering, we can generate six plots, each showing the data trend for sensor A while comparing responses between system 1 and system 2.

In terms of modeling, our initial step should involve the utilization of statistical hypotheses to assess the degree of autocorrelation in the time series data. Following this, we will decide the specific modeling approaches to take, such as random forest, boosting or other machine learning algorithms, and ARIMA or other time series models.

If observations across timestamps are independent and uncorrelated, we can utilize supervised machine learning algorithms such as random forests and boosting. These algorithms are suitable for fitting uncorrelated time series models and extracting feature importance, helping us understand each feature's contribution to the model.

When autocorrelation is evident, it is crucial to consider specific time series models. A commonly used model in such cases is ARIMA (Autoregressive Integrated Moving Average). ARIMA is well-suited for modeling linear relationships and stationary time series data exhibiting trends and seasonality.

We also propose functional data analysis, which is a set of methods designed for analyzing data presented as functions and shapes that vary continuously. Techniques such as Functional Principal Component Analysis can be used to investigate and characterize the primary modes of variation in functional data [6].

First, we extract and visualize the Functional Principal Components that contain the most variance. We compare them in different systems and the same sensor by eyeball and perform hypothesis testing using One-way functional ANOVA, to assess if there are significant differences among the means of the two systems [7].

Following these, we include the external variables (e.g. Card Age and Fluid Type) as predictors and run functional regression analysis models to the time-series data [8]. The final step involves summarizing the regression result to obtain the coefficient plot with confidence intervals (CI). We compare the differences in the curve shapes as well as the width of CI in different systems against all predictors to draw a comprehensive conclusion [9].

Eventually, we will summarize the overall workflow or pipeline to efficiently identify waveform differences across different sensors, and evaluate its applicability using the provided data of sensor A and sensor B.

If time allows, we will also investigate whether enhancements can be made to the current windows, with an emphasis on improving their effectiveness in terms of flatness. This evaluation will involve exploring adjustments to both window size and placement, shedding light on the possible window placement optimization.

Dataset

The data we will analyze consists of four identically structured CSV files containing multiple time-series signals, also known as waveforms or readings. The first column in each file represents the time in seconds, while the remaining columns contain the values of the electrical signals from the potentiometric sensors in millivolts (mV) for every successful test run. Two of the files are from System 1, and two from System 2, corresponding to sensors A and B, respectively. There are a total of 7,039 and 15,561 time series from tests run on System 1 and System 2, respectively.

Furthermore, we were provided with a file consisting of additional information for each unique test run. This includes the test identifier, the type of fluid injected into the card (Table 1), the system that ran the test (either 1=old or 2=new), the type of sensor (A or B), the age of the card at the time of the test (in days), the ambient temperature (the temperature of the entire system: the card, reader, host, and fluid), the time when the sample fluid front is detected, and the time when the bubble between calibration fluid and sample fluid is detected. These characteristics can be linked to the waveforms by their test identifier.

Table 1. Number of tests run by fluid type

Fluid Type	Fluid Name	System 1		System 2		Total
		Sensor A	Sensor B	Sensor A	Sensor B	
Aqueous	Eurotrol L1	617	617	1,495	1,494	4,223
	Eurotrol L3	602	602	1,629	1,629	4,462
	Eurotrol L4	317	312	525	525	1,679
	Eurotrol L5	423	423	1,263	1,263	3,372
Blood	NB	594	593	1,008	1,008	3,203
	AB	232	232	728	728	1,920
	HNB	189	189	402	402	1,182
	TB11	175	174	225	225	799
	DB	117	117	210	210	654
	SB	111	109	260	260	740
	SB-3	30	30	36	36	132
	Total	3,407	3,398	7,781	7,780	22,366

Lastly, there is a file with the values to specific intervals of time (aka windows) within the signal data. These represent periods of interest such as the calibration window and sample window.

The analysis of the multiple time-series signals from System 1 and System 2 will require the following pre-processing steps:

- Discard the first 50 seconds of the waveform since the sensors are not yet wet up and cannot provide reliable readings.
- Remove missing values for System 1.
- Normalization for distance-based methods to scale the sensor data so that they lie between 0 and 1.

Deliverables and Timeline

Responsibility

Our team consists of Dylan, Eden, Jessie, Nan, and Nayeli. For the wavelength characterization task, we will divide into two groups: three people will be responsible for comparing sensor A for systems 1 and 2, and 2 people will be responsible for sensor B. Following this, each of us will choose one modelling approach to analyze the result. Finally, we

will integrate the results. Throughout the process, we will discuss our progress and integrate our results, with each of us at some point acting as recorders during meetings.

Timeline

The duration of the project will be from 29 April to 25 June. We propose to meet with our client at least once a week to share progress updates. The high-level schedule of our weekly goals is as follows(we may adapt it flexibly based on the results and the useful references we have found):

Table 2. Preliminary Plan

Time	Theme	Date Goals
2 weeks	Initiation	Define goals/Data Preparation/ Research Retrieval
2 weeks	Characterization I	Visualize waveform/ Implement FDA & Clustering
2 weeks	Characterization II	Mid-Point Presentation/ Testing & Validation/ (new methods)
1 week	Window Optimization	Improve window effectiveness
1 week	Conclusion	Further improvements/ Final Report and Presentation

Deliverables

The deliverables for this project encompass a thorough characterization of biosensor waveforms and suggestions for improved window placements to enhance accuracy and precision in calculations. These deliverables consist of (1) a final presentation, (2) a comprehensive report detailing the analysis process and findings, and (3) a Python script for the Siemens team to replicate the pipeline on all the other sensors according to their demands. Additionally, consultation and advice from the data science and R&D teams will be integrated into the project's progress and reflected in the final report.

References

[1] U. B. Trivedi et al., "Potentiometric biosensor for urea determination in milk," Sensors and Actuators B: Chemical, vol. 140, no. 1, pp. 260-266, 2009.

[2] S. C. Ribeiro et al., "Potentiometric biosensor based on artificial antibodies for an Alzheimer biomarker detection," Applied Sciences, vol. 12, no. 7, art. 3625, 2022.

[3] Epocal Inc., "System Manual with Epoc NXS Host," Siemens Healthcare Diagnostics Inc., Ottawa, ON, Canada, 2023. [Online]. Available: https://siemens-healthineers.com/epoc

[4] C. Richter, "Visualizing Sensor Data," 2009. [Online]. Available: https://www.medien.ifi.lmu.de/lehre/ws0809/hs/docs/richter.pdf

[5] R. Tavenard, "Time Series Clustering," tslearn's documentation, 2017. [Online]. Available: https://tslearn.readthedocs.io/en/stable/user_guide/clustering.html#time-series-clustering

[6] "Functional Principal Component Analysis Models." KI Global Health Advocates. [Online]. Available: https://www.kiglobalhealth.org/resources/functional-principal-component-analysis-models/

[7] Grupo de Aprendizaje Automático - Universidad Autónoma de Madrid: One-way functional ANOVA with real data.[Online]. Available: https://fda.readthedocs.io/en/latest/auto_examples/plot_oneway.html

[8] Morris, Jeffrey S. "Functional regression." Annual Review of Statistics and Its Application 2 (2015): 321-359.

[9] "What is functional data analysis?", YouTube. Available: https://www.youtube.com/watch?v=U2TvHLA18lo