

From Source to Feature: Multimodal Data Preparation for Machine Learning

Dmitry Amelin¹, Matthias Jochem¹, Gizem Aydogan^{1,2}, Michael M. Gepp¹, Roman Ruff¹

¹Fraunhofer Institute for Biomedical Engineering IBMT, 66280 Sulzbach, Germany

²Saarland University, 66123 Saarbrücken, Germany

Abstract

Three-dimensional spheroid and organoid cultures are increasingly used in biomanufacturing due to their ability to mimic in-vivo tissue behavior, yet current analytical methods remain labor-intensive and insufficient for real-time process understanding. In this work, we present a multimodal data-ingestion and preparation framework that integrates sparse morphological features extracted from YOLO-based microfluidic imaging with dense electrochemical sensor measurements and associated metadata. Dedicated ingestion pipelines for imaging and biochemical data were developed, followed by structured preprocessing steps that harmonize these heterogeneous streams into a unified temporal representation. A human-in-the-loop labeling workflow ensures high-quality annotations for model development. This approach lays the groundwork for scalable, automated spheroid production by enabling early detection of abnormal growth, supporting rapid decision-making in the laboratory, and improving reproducibility and quality assurance in organoid manufacturing.

Keywords: bioprocess monitoring, organoids, multimodal data, data ingestion

Motivation

Three-dimensional (3D) spheroid and organoid cultures have become indispensable model systems in drug discovery, toxicology, and regenerative medicine due to their ability to replicate key aspects of in-vivo tissue physiology, including cell to cell interactions, nutrient gradients, and complex self-organized morphology [1], [2], [3]. As these systems move toward automated and scalable bioprocessing, the need for non-invasive, high-frequency monitoring technologies is increasing. Such approaches must capture dynamic growth behavior and morphological evolution without disturbing the culture.

Conventional analytical methods (e.g., manual microscopy, offline biochemical assays, or endpoint destructive measurements) are labor-intensive, disruptive, and poorly suited to real-time process understanding. This hampers optimization and limits the deployment of advanced 3D culture workflows at scale.

At the Fraunhofer Center for Sensor Intelligence (ZSI), integrated optical, microfluidic, and electrochemical sensor modules have been developed to enable continuous, AI-supported monitoring of spheroids in bioreactors [4]. Microfluidic sampling allows the extraction and imaging of spheroids without spheroids loss and culture conditions interruption. Parallel electrochemical sensors measure metabolic

analytes (glucose, lactate, oxygen, pH) at approximately 2 min intervals. Together, these systems generate rich multimodal data streams representing both morphological and metabolic states.

However, these data sources differ significantly in structure and sampling frequency. Their integration into predictive bioprocess models therefore remains a major challenge. Specifically, such a framework could enable:

- Early detection of aberrant growth trajectories,
- Scheduling of media exchanges,
- Real-time decision support,
- Ensuring reproducibility and quality in organoid- and spheroid-based manufacturing.

Because biological growth is driven by dynamic interactions between environmental, metabolic, and morphological variables, this becomes a time-series forecasting problem requiring the fusion of heterogeneous inputs. Modern machine learning models (e.g., RNNs, GRUs/LSTMs, TCNs, transformers) provide a promising solution.

To address these challenges, this work focuses on establishing coordinated data-ingestion workflows and a structured preparation framework that can incorporate YOLO-derived morphological features, dense electrochemical time-series measurements, and optional

bioprocess metadata. This structured approach creates the foundation on which forecasting models can be trained, enabling prediction of morphological characteristics of spheroids.

Data Acquisition

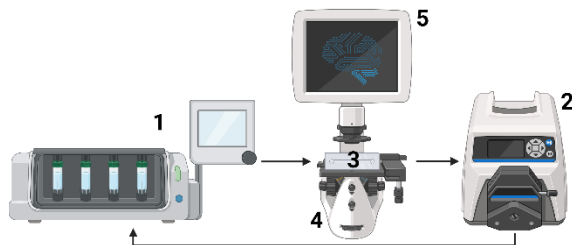


Figure 1 Schematics representation of the camera-based sensor. (The diagram was created with BioRender.com)

Figure 1 schematically depicts the experimental setup. Spheroid cultures were maintained in a bench-scale suspension bioreactor (1) equipped with an automated at-line sampling system. A peristaltic pump (2) transfers spheroids into a transparent channel (3), where they are imaged by a microscope (brightfield, phase-contrast, or fluorescence) (4). After acquisition, the spheroids are returned to the bioreactor. The images were processed on an Nvidia Jetson Nano single-board computer (5) in real-time during acquisition using a model based on You Only Look Once object segmentation architecture [5]. The model generated segmentation masks for spheroids, which can be seen on Figure 2. From each mask, morphological characteristics were computed, including area, diameter, perimeter, and circularity. Electrochemical sensors integrated into the bioreactor or alternatively integrated in-line with the imaging system provided continuous measurements of glucose, lactate, oxygen, and pH at around 2 min intervals. Figure 3 shows the compact measurement electronics for the electrochemical sensors with a diameter of 30 mm.

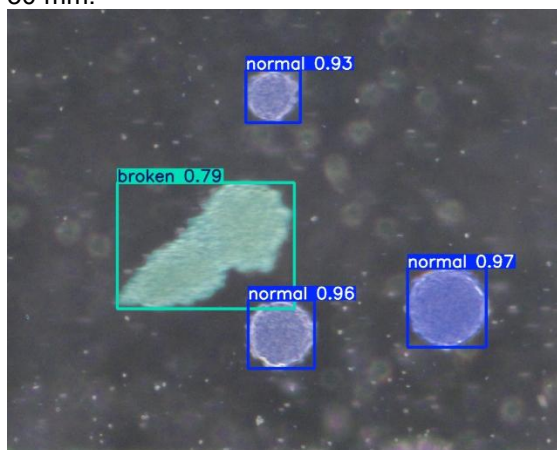


Figure 2 An example of spheroids segmentation

Because the electrochemical and imaging subsystems differ in throughput and metadata requirements, two separate data pipelines were implemented. A graphical user interface application for the electrochemical sensors receives data from the electronics via Bluetooth Low Energy and forwards it to the database. For microscopic pictures, due to the high volume of image data and additional metadata needs, a dedicated gRPC-based image transfer pipeline was developed. Each image and its metadata were serialized and sent to a middleware service running on the database server. The middleware validated metadata and orchestrated storage, separating raw images (stored in a dedicated high-volume repository) from structured metadata stored in PostgreSQL.

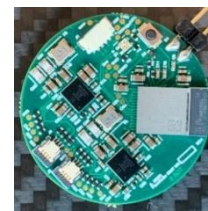


Figure 3 Electrochemical measurement electronics, 30 mm in diameter

Data Storage

All structured data, including sensor values, metadata, sampling times, and experimental parameters, were stored in PostgreSQL. Time-series measurements were organized in a dedicated measurement table, supplemented by a sensor metadata table for fast contextual lookup. Images were not stored as binary objects in PostgreSQL due to database bloat and slow retrieval associated with large-object storage [6]. Instead, only file paths and metadata were stored in the database, ensuring efficient and scalable data management. Additionally, all information about spheroids detected on images were stored in a separate table.

Both the electrochemical and image related measurement tables contain a foreign-key reference to a table with the experiment metadata. This way we can find related measurements for ML model training and validation.

Labeling

As previously stated, in addition to storing raw images, the segmentation output for each detected spheroid, including the contour, bounding-box, and class information, is recorded in a dedicated table within the database. To support systematic validation and correction of these annotations, a web-based interface was developed that allows biologists

to review segmentation and classification results. If the ML model produces incorrect or incomplete segmentations, users can flag the corresponding images directly through the interface. Flagged samples are then exported to dedicated labeling software for manual correction. After review, the corrected segmentation masks and class labels are re-imported and used to update the database records.

This workflow enables efficient human-in-the-loop curation of the dataset and ensures high-quality ground truth labels for both model training and subsequent validation cycles.

Outlook

The next major challenge is identifying the most suitable model architecture for combining sparse imaging data with dense electrochemical measurements to predict future morphology. We are going to focus on the following two approaches.

To synchronize the datasets, temporal windows of electrochemical data preceding each imaging event will be extracted (e.g., 12 h, 24 h, 48 h). Each window will be encoded into a compact representation using a dedicated time-series encoder such as one-dimensional convolutional neural network (1D CNN), Long short term memory (LSTM), gated recurrent units (GRUs), or Temporal Convolutional Network (TCN).

For each imaging timepoint, t , the morphology vector, $M(t)$, will be concatenated with the electrochemical embedding, $E(t)$, to form a joint feature vector:

$$X(t) = [M(t), E(t)]$$

As an alternative, a Cross-Attention Fusion strategy will be evaluated. Here, the full electrochemical time series is retained, sparse and dense measurements are encoded separately, and a cross-attention layer identifies relevant metabolic patterns conditioned on morphological features.

In both cases, the fused representation will be passed to a multi-layer perceptron trained to predict the spheroid morphology at the next imaging time point.

With this work, we aim to establish a foundation for scalable spheroid production by enabling early prediction of abnormal growth trajectories, supporting biologists in making faster and more informed decisions, and ultimately improving reproducibility and quality assurance in organoid manufacturing.

Acknowledgement

This work was supported by the Saarland Ministry of Economic Affairs, Innovation, Digital and Energy (MWIDE) and the Fraunhofer

Gesellschaft (High-Performance Center 'Sensor- Intelligence').

Bibliography

- [1] S. Chen, X. Chen, Z. Geng, and J. Su, "The horizon of bone organoid: A perspective on construction and application," *Bioact Mater*, vol. 18, pp. 15–25, Dec. 2022, doi: 10.1016/j.bioactmat.2022.01.048.
- [2] W. Hu and M. A. Lazar, "Modelling metabolic diseases and drug response using stem cells and organoids," *Nat Rev Endocrinol*, vol. 18, no. 12, pp. 744–759, Dec. 2022, doi: 10.1038/s41574-022-00733-z.
- [3] O. Kopper *et al.*, "An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity," *Nat Med*, vol. 25, no. 5, pp. 838–849, May 2019, doi: 10.1038/s41591-019-0422-6.
- [4] D. Amelin *et al.*, "Bioreactor Intelligence for Real-Time Monitoring and Automation of Stem Cell Cultivation," *Lectures*, pp. 137–138, May 2025, doi: 10.5162/SMSI2025/C2.2.
- [5] G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*. (2023). [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [6] R. Sears, C. Ingen, and J. Gray, "To BLOB or Not To BLOB: Large Object Storage in a Database or a Filesystem?," *Computing Research Repository - CORR*, Jan. 2007.