Digital ICU : HiWi [12-16h/week] - ICU dataset recording, cleaning and preprocessing

General Info

Contact Person: Kai Wu
Contact Email: k.wu@tum.de

Project Abstract

Digitalization in healthcare has led to the increasing use of digital medical systems in the Intensive Care Unit (ICU). They generate a large amount of data, such as the vital signs of patients, the blood gas analysis results, and the medication that a patient receives. This data can be analyzed using machine learning and data analytics techniques to help clinicians identify clinical deterioration in patients earlier and determine if a patient's treatment is working.

However, existing datasets are not fully harnessing the available density of raw medical data. In our project, we are recording a dense ICU patient dataset with additional patients' activity information to enable studies on transient changes in patients' vital signs. We are collecting data from (1)three realsense cameras, (2)bedside patient monitor, (3)ventilator and (4)perfusor. This project aims to develop a pipeline including data cleaning, formatting, labeling, visualizing, and therefore provide a structured dataset for further research (mostly with machine learning models[10,11]). Moreover, the project also strives to implement a visualization tool capable of visualizing both the raw unstructured data and the final structured data to check if the data is correctly cleaned and aligned, and also better inform the clinicians on how the data is processed.

Task Description

- Literature review on SOTA works on ICU datasets, with a focus on the strategies used to preprocess raw ICU data.
- Develop and implement a pipeline to process unstructured raw ICU data [1,2,3] into structured ones. The works in [6,7,12,13] can be used as references.
- Implement an annotation tool for patient activity labeling, where input is depth image sequences and extracted skeleton poses.
- Implement a visualization tool to intuitively visualize the processed data.
- Write technical reports/documentations.

Technical Prerequisites

- advanced programming experience with Python3, and optimal if you are familiar with C++.
- Basic knowledge on SQL.
- (Optional) Experience with the following python libraries: Pandas, Matplotlib, Plotly, OpenCV, scikit-learn.

References

[1] Johnson, A., Pollard, T., Badawi, O., & Raffa, J. (2021). eICU Collaborative Research Database Demo (version 2.0.1). PhysioNet. https://doi.org/10.13026/4mxk-na84.

[2] Kallfelz, M., Tsvetkova, A., Pollard, T., Kwong, M., Lipori, G., Huser, V., Osborn, J., Hao, S., & Williams, A. (2021). MIMIC-IV demo data in the OMOP Common Data Model (version 0.9). PhysioNet. https://doi.org/10.13026/p1f5-7x35.
[3] Yèche, H., Kuznetsova, R., Zimmermann, M., Hüser, M., Lyu, X., Faltys, M., & Ratsch, G. (2021). HiRID-ICU-Benchmark---A Comprehensive Machine Learning Benchmark on High-resolution ICU Data.
[6] https://ohdsi.github.io/CommonDataModel/index.html
[7] Jarrett, D., Yoon, J., Bica, I., Qian, Z., Ercole, A., & van der Schaar, M. (2020, September). Clairvoyance: A pipeline toolkit for medical time series. In International Conference on Learning Representations.
[10] Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. Journal of biomedical informatics, 83, 112-134.
[11] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific data, 6(1), 1-18.
[12] Tang, S., Davarmanesh, P., Song, Y., Koutra, D., Sjoding, M. W., & Wiens, J. (2020). Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. Journal of the American Medical Informatics Association, 27(12), 1921-1934.
[13] Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., & Naumann, T. (2020, April). Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In Proceedings of the ACM Conference on Health, Inference, and Learning (pp. 222-235).