CyclOps: A unified framework for data extraction and rigorous evaluation of ML models for clinical use-cases

Vallijah Subasri [1,2,3 †], *Amrit Krishnan* [1 †], *Kaden McKeen* [1], *Yusuf Sheikh* [1,2],, *Maria Koshkina* [1], *Nadim Lalani* [1], *Azra Dhalla* [1], *Deval Pandya* [1], *Fahad Razak* [2,4], *Amol A. Verma* [2,4], *Elham Dolatabadi* [1,2]

[1] Vector Institute, [2] University of Toronto, [3] Hospital for Sick Children, [4] Unity Health Toronto

[†] These authors contributed equally

Background/Motivation

A fundamental hurdle in the deployment of machine learning (ML) models for healthcare is building an integrated ML system that is capable of being continuously operated and maintained throughout production. Data extraction and model evaluation are key components of this integrated system. Clinical data extraction is particularly convoluted due to a lack of standardization in Electronic Health Record (EHR) systems used across hospitals. Building robust clinical ML models has also proven to be difficult, attributed to dataset shifts that change feature distributions over time and lead to spurious predictions. Rigorous evaluation of ML models across time, hospital sites and diverse patient cohorts is critical for identifying model degradation and informing clinical end-users of changes. A unified software framework that can address these challenges would be a large step towards realizing the potential of ML for healthcare.

We introduce CyclOps, a framework designed to address these challenges and enable healthcare-oriented ML research and facilitate the rigorous evaluation of clinical ML models necessary for responsible deployment. Our framework design strongly integrates and leverages well-tested open-source components, targeted towards building a unified ML operations (MLOps) framework for healthcare, while providing APIs in the Python programming language.

Capabilities

The CyclOps framework provides 3 unified high-level APIs: **1) Data querying and processing** - EHR data querying and pre-processing; flexible and composable data processing for cleaning, aggregation, imputation, and vectorization; a visualization API to visualize patient timelines, features, metrics and other useful metadata. In our first release, we support the General Medicine Inpatient Initiative (GEMINI)[1,2], and MIMIC-IV[3] databases. In future releases, we plan to integrate the querying and processing APIs with datasets that are available in FHIR format[4] or the OMOP data model.[5] **2) Experimentation -** Strong integration with a workflow management tool to run orchestrated pipelines of tasks. We also provide baseline model implementations for benchmarking. In future releases, we plan to integrate the experimentation pipeline with popular open-source tools such as MLFlow to track experiments and register models. **3) Model Evaluation and Drift Monitoring** - A drift detection pipeline that provides state-of-the-art methods for dimensionality reduction and two-sample statistical testing to identify malignant shifts across source and target datasets. A suite of experiments to evaluate robustness to dataset drift which includes a) synthetic perturbations that span various covariates and label shifts and b) real-life experiments that evaluate dataset shift over time, across hospitals and paradigmatic events like the COVID-19 pandemic (Figure 1). In our next release, we plan to include an extensible evaluation API that can support the rigorous evaluation of ML models towards clinical decision support systems, tailored for deployment use-cases.

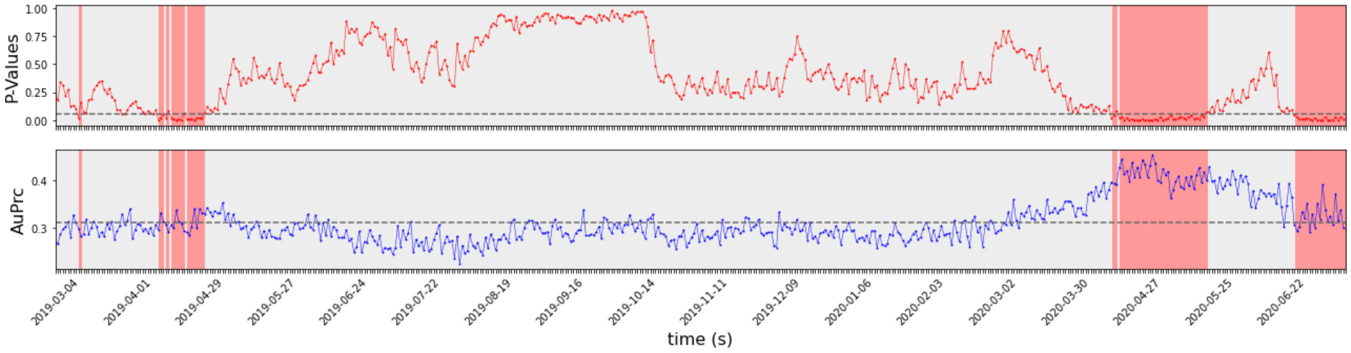

Figure 1. A LSTM model was trained on pre-COVID-19 (source) patient data from general internal medicine wards in Toronto, Canada-area hospitals to predict inpatient mortality over the next 2 weeks. Drift detection was performed using dimensionality reduction (Black Box Shift Estimation; BBSE) and multivariate two-sample testing (Maximum Mean Discrepancy; MMD) on data drawn from the same pre-COVID-19 (source) distribution that the model was trained on and each day from 03/2019 to 07/2020 using a 60-day rolling window. It can be seen that drift is detected from 04/2020 onwards (red), which corresponds to when Toronto declared the pandemic a state of emergency and entered a lockdown. Moreover, the drift is reflected in the AUPRC (blue) which increases during the first wave of the COVID-19 pandemic.

Link to Software: https://github.com/VectorInstitute/cyclops

Acknowledgements: This data is obtained from the GEMINI, for more details see https://www.geminimedicine.ca/

References:

1. Verma, A. A. *et al.* Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open* vol. 5 E842–E849 (2017).
2. Verma, A. A. *et al.* Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *J. Am. Med. Inform. Assoc.* **28**, 578–587 (2021).
3. Johnson, A. *et al.* MIMIC-IV. (2020) doi:10.13026/A3WN-HQ05.
4. Saripalle, R. K. Fast Health Interoperability Resources (FHIR). *International Journal of E-Health and Medical Communications* vol. 10 76–93 (2019).
5. Ohdsi. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. (OHDSI, 2019).