SSBSE Summary of EvoCLINICAL: Evolving Cyber-cyber Digital Twin with Active Transfer Learning for Automated Cancer Registry System *

Chengjie Lu[1,2 (✉)] [0000−0002−5818−7547], Qinghua Xu[1,2] [0000−0001−8104−1645], Tao Yue[1] [0000−0003−3262−5577], Shaukat Ali[1] [0000−0002−9979−3519], Thomas Schwitalla[3] [0000−0002−0286−1686], and Jan F. Nygård[3,4] [0000−0001−9655−7003]

[1] Simula Research Laboratory, Oslo, Norway
[2] University of Oslo, Oslo, Norway
[3] Cancer Registry of Norway, Oslo, Norway
[4] UiT The Arctic University of Norway, Tromsø, Norway
chengjielu@simula.no, et.qinghua@gmail.com,
{tao,shaukat}@simula.no, {thsc,jfn}@kreftregisteret.no

Abstract. EvoCLINICAL evolves the cyber-cyber digital twin (CCDT) of a key component of the cancer registration and support system (CRSS) called GURI. CRSS is built and operated by the Cancer Registry of Norway, which collects cancer-related data from medical entities to produce cancer-related statistics for its end users. The CCDT is constructed to facilitate various research activities without intervening in the real system. GURI changes like any other software system, thus requiring evolving CCDT as well. EvoCLINICAL was specifically designed to evolve CCDT to synchronize with the evolving GURI. To this end, EvoCLINICAL employs a search algorithm to select the most valuable cancer messages and adopts active transfer learning to evolve CCDT for a new version of GURI. The evaluation of EvoCLINICAL on three different evolution processes demonstrated its effectiveness and practicality in a real-world context. This paper summarizes the EvoCLINICAL paper [1].

Keywords: Search-based Selection · Cyber-cyber Digital Twin · Transfer Learning · Validation System

Summary

The Cancer Registry of Norway (CRN) collects cancer data and provides statistics on cancer cases in Norway. To ensure data quality, CRN developed an automated cancer registry system with a core component called GURI that checks the validity of cancer messages against a set of medical rules. A cancer message is a JSON-like file, encompassing multiple fields (e.g., chemotherapy) related to

* This work is supported by the AIT4CR (No. 309642) project funded by the Research Council of Norway.

the cancer registry. Together with CRN, we built a neural network-based cyber-cyber digital twin (CCDT) for GURI, which simulates a specific GURI version by predicting the validation result of a given cancer message. CCDT is built to enable large-scale simulations and analyses without interfering with the real operation of GURI while keeping the CCDT in synchronization with the real system. However, GURI undergoes continuous evolution as new rules are added, modified, or removed due to new treatments, improved diagnostics, or medical findings. Therefore, to synchronize with the evolving GURI, CCDT must continue to evolve. To this end, we proposed EvoCLINICAL, which integrates the search-based selection technique to build datasets with the most valuable data and adopts active transfer learning (TL) to support the evolution of CCDT.

Considering GURI's evolution from a source version (GURI-S) to a target version (GURI-T), EvoCLINICAL constructs a CCDT-S for GURI-S by training CCDT-S from scratch with abundant labeled data generated by the continuous operations of GURI-S. However, the number of labeled data from GURI-T is not guaranteed because of its short operation time. Therefore, we utilized active TL to transfer knowledge from CCDT-S to CCDT-T since it required less data. To construct the dataset for active TL, we adopted the search-based selection technique to actively select the most valuable cancer messages from a candidate dataset. We guide the dataset selection process with five optimization objectives: minimizing the number of cancer messages, maximizing the cancer message diversity, maximizing the prediction result diversity of the cancer messages, maximizing the proportion of cancer messages with false predictions, and maximizing the predictive uncertainty of cancer messages. These objectives ensure that the search selects the optimal datasets with high coverage of cancer messages and prediction results. We employed the indicator-based evolutionary algorithm, IBEA, as the search algorithm, given its effectiveness in solving multi-objective problems. After obtaining the selected dataset, we performed dataset augmentation to balance the dataset and then adopted this dataset to fine-tune CCDT-S to obtain CCDT-T.

We evaluated EvoCLINICAL with three evolution processes and compared EvoCLINICAL with three baselines: utilizing CCDT-S as CCDT-T without fine-tuning (OTS), training CCDT-T from scratch (TFS), and fine-tuning CCDT-S to CCDT-T using randomly selected datasets (EvoCLINICAL-RS). The evaluation results showed the effectiveness of EvoCLINICAL with precision, recall, and F1 score of at least 0.9200, 0.9105, and 0.9137. Besides, the comparison with EvoCLINICAL-RS demonstrated that actively selecting datasets with search can significantly improve the performance of EvoCLINICAL.

References

1. Lu, C., Xu, Q., Yue, T., Ali, S., Schwitalla, T., Nygård, J.: Evoclinical: Evolving cyber-cyber digital twin with active transfer learning for automated cancer registry system. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 1973–1984 (2023)