PyHealth: A Deep Learning Toolkit for Healthcare Applications

Citation for published version:
Yang, C, Wu, Z, Jiang, P, Lin, Z, Gao, J, Danek, BP & Sun, J 2023, PyHealth: A Deep Learning Toolkit for Healthcare Applications. in *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* ACM Association for Computing Machinery, pp. 5788–5789. https://doi.org/10.1145/3580305.3599178

Digital Object Identifier (DOI):
10.1145/3580305.3599178

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

PyHealth: A Deep Learning Toolkit For Healthcare Applications

Chaoqi Yang
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
chaoqiy2@illinois.edu

Zhenbang Wu*
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
zw12@illinois.edu

Patrick Jiang
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
pj20@illinois.edu

Zhen Lin
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
zhenlin4@illinois.edu

Junyi Gao
University of Edinburgh
Health Data Research UK
Edinburgh, Scotland, UK
junyi.gao@ed.ac.uk

Benjamin P. Danek
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
bdanek2@illinois.edu

Jimeng Sun
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
jimeng@illinois.edu

ABSTRACT

Deep learning (DL) has emerged as a promising tool in healthcare applications. However, the reproducibility of many studies in this field is limited by the lack of accessible code implementations and standard benchmarks. To address the issue, we create PyHealth, a comprehensive library to build, deploy, and validate DL pipelines for healthcare applications. PyHealth supports various data modalities, including electronic health records (EHRs), physiological signals, medical images, and clinical text. It offers various advanced DL models and maintains comprehensive medical knowledge systems. The library is designed to support both DL researchers and clinical data scientists. Upon the time of writing, PyHealth has received 633 stars, 130 forks, and 15k+ downloads in total on GitHub.

This tutorial will provide an overview of PyHealth, present different modules, and showcase their functionality through hands-on demos. Participants can follow along and gain hands-on experience on the Google Colab platform during the session.

ACM Reference Format:
Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin P. Danek, and Jimeng Sun. 2023. PyHealth: A Deep Learning Toolkit For Healthcare Applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599178

1 TARGET AUDIENCE AND PREREQUISITES

This tutorial is **hands-on** and will last for **3 hours**. It is designed for audiences interested in deep learning and health informatics, including both deep learning researchers with experience in data science and Python/PyTorch programming, and clinical informaticians with clinical expertise and some exposure to data science.

Prerequisites for this tutorial include basic knowledge of deep learning and Python programming. No prior knowledge of healthcare is required. Throughout the tutorial, we will alternate between lectures and hands-on practice to encourage audience participation. Attendees can access the same Colab notebook on our website and follow along step-by-step. After the tutorial, we will make the tutorial materials (e.g., tutorial summary, presentation slides, code, and recordings) publicly available for wider dissemination.

2 TUTORIAL OUTLINES

The outline of the tutorial is listed below. Resources, including the GitHub repository, documentation, YouTube playlist, slides, and Colab notebooks, can be found on our website[1].

Overview of PyHealth. This session introduces the background and motivations behind PyHealth and showcases its main features with quickstart examples to motivate the audience.

Clinical Predictive Modeling with EHRs. This session provides a detailed explanation of the five-stage pipeline with EHR data. We will cover data loading[2] (e.g., MIMIC [6, 7], eICU [13]), task definition (e.g., mortality prediction), model initialization (e.g., RETAIN [1], SafeDrug [18]), model training, and evaluation.

Deep learning for Physiological Signals. This session demonstrates how to utilize PyHealth for processing physiological signal data. We will introduce the biosignal datasets (e.g., ISRUC [9], Sleep-EDF [8]) and existing biosignal models (e.g., ContraWR [19]) supported by PyHealth. We finally show a demo: sleep staging with SPaRCNet [5] on the Sleep-EDF dataset.

Medical Imaging Analysis. This section demonstrates how to utilize PyHealth for medical image data. We will introduce medical image datasets (e.g., CheXpert [4], COVID [14]), relevant tasks (e.g., disease classification, segmentation), and existing models (e.g., ResNet [3]) in PyHealth. We finally show a demo: chest disease classification with ResNet on the COVID dataset.

Natural Language Processing for Clinical Text. This section demonstrates how to utilize PyHealth for medical text data. We will introduce medical text datasets (e.g., MIMIC-III clinical notes [7]), relevant tasks (e.g., medical report generation), and existing models

*C. Yang and Z. Wu contributed equally.

[1] https://sunlabuiuc.github.io/PyHealth/
[2] In compliance with dataset policies, we will utilize our synthetic version of the datasets as a substitute in the tutorial.

KDD '23, August 6–10, 2023, Long Beach, CA, USA.

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599178

(e.g., CAML [12]) in PyHealth. We finally show a demo: assigning medical billing codes to patient discharge summaries with CAML.

Medical Knowledge Graph. This session demonstrates how to utilize PyHealth's comprehensive medical knowledge base. We will introduce different medical coding systems (e.g., ICD-9/10, ATC codes), tools for concept lookup and mapping cross systems (e.g., rule-based mapping, AutoMap [17]), and the pre-trained medical concept embeddings in PyHealth. We finally show a demo: utilizing the Unified Medical Language System (UMLS) knowledge graph embeddings to improve the drug recommendation task.

Synthetic Data Generation. In this session, we will demonstrate PyHealth's synthetic data generation capability. We will introduce HALO [16], a method capable of generating synthetic longitudinal healthcare records which have the training utility of real patient records, without privacy and regulatory concerns.

Post-Hoc Uncertainty Quantification. This session introduces PyHealth's uncertainty quantification module, covering important tasks such as model calibration and prediction set construction. We will provide a demo applying calibration methods [2, 10] and prediction set construction methods [11, 15] on a trained sleep-staging SPaRCNet [5] classifier on the ISRUC [9] dataset.

In the end, we summarize the tutorial and provide links to other PyHealth resources to our users and potential collaborators.

3 BRIEF BIOGRAPHICS OF TUTORS

Chaoqi Yang is a Ph.D. student in Computer Science at the University of Illinois Urbana-Champaign. His research interests include clinical predictive modeling, biosignal modeling, tensor decomposition, and self-supervised learning.

Zhenbang Wu is a Ph.D. student in Computer Science at the University of Illinois Urbana-Champaign. His research interest is developing generalizable and adaptable deep learning algorithms to solve important healthcare problems.

Patrick Jiang is an M.S. student in Computer Science at the University of Illinois Urbana-Champaign. His research interests are healthcare natural language processing and graph learning.

Zhen Lin is a Ph.D. student in Computer Science at the University of Illinois Urbana-Champaign. His research interests include uncertainty quantification in healthcare and biosignal modeling.

Junyi Gao is a Ph.D. student at the University of Edinburgh funded by the HDR UK-Turing Welcome Ph.D. Program. His research interests include spatio-temporal epidemiology prediction and individual-level clinical predictive modeling.

Benjamin Danek is an MCS student in Computer Science. His interests are in federated learning and fairness, and synthetic data generation.

Jimeng Sun is a Professor at Computer Science Department and Carle's Illinois College of Medicine at University of Illinois Urbana-Champaign. His research focuses on data mining for healthcare, especially in developing tensor factorization, deep learning methods, and large-scale predictive modeling systems.

ACKNOWLEDGEMENTS

This work was supported by NSF awards SCH-2205289, SCH-2014438, IIS-1838042, NIH award R01 1R01NS107291-01. Junyi Gao acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z).

REFERENCES

[1] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *NeurIPS* 29 (2016).

[2] Chirag Gupta and Aaditya Ramdas. 2021. Distribution-Free Calibration Guarantees for Histogram Binning without Sample Splitting. In *ICML*. 3942–3952.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *CoRR* abs/1901.07031 (2019). arXiv:1901.07031 http://arxiv.org/abs/1901.07031

[5] Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. 2023. Development of Expert-Level Classification of Seizures and Rhythmic and Periodic Patterns During EEG Interpretation. *Neurology* (2023).

[6] A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. 2020. MIMIC-IV (version 1.0).

[7] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (Dec. 2016), 160035. https://doi.org/10.1038/sdata.2016.35

[8] Bastiaan Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A.C. Kamphuisen, and Josefien J.L. Oberyé. 2000. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* (2000). https://doi.org/10.1109/10.867928

[9] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. 2016. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine* (2016). https://doi.org/10.1016/j.cmpb.2015.10.013

[10] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Taking a Step Back with KCal: Multi-Class Kernel-Based Calibration for Deep Neural Networks. In *ICLR*.

[11] Zhen Lin, Shubhendu Trivedi, Cao Xiao, and Jimeng Sun. 2023. Fast Online Value-Maximizing Prediction Sets with Conformal Cost Control. arXiv:2302.00839 [cs.LG]

[12] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1101–1111. https://doi.org/10.18653/v1/N18-1100

[13] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data* 5, 1 (Dec. 2018), 180178. https://doi.org/10.1038/sdata.2018.178

[14] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in biology and medicine* 132 (2021), 104319.

[15] Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least Ambiguous Set-Valued Classifiers With Bounded Error Levels. *J. Amer. Statist. Assoc.* 114, 525 (2019). https://doi.org/10.1080/01621459.2017.1395341

[16] Brandon Theodorou, Cao Xiao, and Jimeng Sun. 2023. Synthesize Extremely High-dimensional Longitudinal Electronic Health Records via Hierarchical Autoregressive Language Model. *arXiv preprint arXiv:2304.02169* (2023).

[17] Zhenbang Wu, Cao Xiao, Lucas M. Glass, David M. Liebovitz, and Jimeng Sun. 2023. AutoMap: Automatic Medical Code Mapping For Clinical Prediction Model Deployment. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II* (Grenoble, France). Springer-Verlag, Berlin, Heidelberg, 505–520. https://doi.org/10.1007/978-3-031-26390-3_29

[18] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Safe Drug Recommendations. In *IJCAI*.

[19] Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. 2021. Self-supervised EEG Representation Learning for Automatic Sleep Staging. *arXiv preprint arXiv:2110.15278* (2021).