Study of Datasets in AI Based Medical Informatics

Samaneh Shamshiri Insoo Sohn

Division of Electronics and Electrical Engineering
Dongguk University

samaneh.shamshiri@gmail.com isohn@dongguk.edu

Abstract

In recent years, Artificial intelligence (AI) has played a significant role in health informatics and medical applications to promote early detections, disease diagnosis, and referral managements. In terms of classification, labeling, training process, dataset size, and algorithm validation of AI, uncourtly, data is the first step to developing any treatment or detecting tools. This paper reviews some dataset analytic role and data driven models in medical informatics and investigated the two first datasets from CT scan images and chest X-ray images for detecting COVID-19 by AI models.

1. Introduction

From 1970's artificial intelligence and deep learning algorithms have applied to solve wide range of problems in various domains of health informatics ranging from bioinformatics, medical informatics, medical imaging and public health [1]. These include clinical diagnosis of acute and chronic diseases such as Alzheimer disease that have been assisted via AI technologies such as Support Vector Machines, classification trees, and artificial neural networks. AI algorithms provide detecting of malignant cells with high diagnostic accuracy. Furthermore, predicting of different kind of cancers such as breast cancer, lung cancer etc. can contributed by AI techniques. Health informatics and its subfields have improved diagnosis in various tasks including classification of lung diseases, detection of nodules, brain tumor segmentation and body organs recognition. Employing AI technologies to help the detection of COVId-19 from medical images, especially chest X-ray images and CT- scans, are strong evidence for the significant role of AI algorithms in healthcare informatics. Therefore, physicians try continually to improve their knowledge about applying AI to provide their patients with the best care services. To achieve these purposes, they need more targeted and precise data from their patients. Medical informatics is their tool for delivering the data. Medical informatics is a sub-branch of medical health that directly relates physician to their patients. The main goal of medical informatics is gathering data by applying the state-of-the-art technology to develop medical knowledge and to facilitate delivery of patient medical care. Another goal of medical informatics is concentrating on the management of medical data from research and education. Therefore, health care data is a wide variety of information including patients, doctors and healthcare systems [2].

The process of collecting data can improve healthcare systems through the precise analysis of large datasets to draw conclusions and optimize efficiency. Since healthcare data is a highly technique and emerging field data analysts employ a specific skill set to success.

Although, the diversity and amount of data have increased rapidly, there are still many challenges about processes for data sharing and access. In this paper, we survey existing researches in medical informatics to review current trends in this area and investigate the datasets and AI algorithms, which have applied in the field of health informatics.

2. Medical informatics datasets

The combination of sufficient and distributed datasets in medical informatics and the ability of AI algorithms to learn new features and patterns has resulted in a number of significant advances in healthcare systems. Therefore deep learning as a family of AI algorithms has quickly been applied in medical informatics researches, for example, Shin et al [3] to gain the most comprehensive interpretation of diagnostic semantics, use radiology reports of around 780k imaging examination stored in PACS of National Institute of Health clinical center and present a combined text-image CNN to identify semantic information that links radiology images. Liang et al. [4] used large-scale datasets for effecting training methods for hypertension. Pautin et al.[5] to train the DNNs for identifying markers (aging biomarkers) used over 60000 samples from common blood biochemistry and cell count tests to predict human chronological age. Nie et al. [6] propose a new deep learning model, their datasets including more than 900 popular dieses concepts from EveryoneHealthy, WebMD and Medline plus. They cover a wide range of disease, including endocrine, urinary, neurological and other aspects, which consist of manual gathering the key symptoms or questions related to the disease. Miotto et al. [7] provide a comprehensive review of the key deep learning architecture that have applied to Electronic Health records (EHR) datasets. They also introduced CPRD, which is one of the world's largest primary care datasets and showed how data from primary care can provide predictions, the probability of a patient developing specific disease, such as diabetes, schizophrenia and cancer. Futoma et al. [8] describe and compare different DNN models based on EHR datasets. Their purpose is predicting

hospital readmissions. In the case of penalized logistic regression, DNN have significantly higher prediction accuracy.

Lipton et al. [9] evaluate the ability of long short-term memory (LSTMs) to recognize patterns in multivariate time series of clinical measurements. They consider multipliable classification of diagnosis, training a model to classify 128 diagnosis given 13 diagnosis frequently but irregularly sampled clinical measurements. Their data extracted from the EHR system consist of 10401 PICU episode. Mehrabi et al. [10] use DBM model to find common temporal patterns among the diagnosis metrics. They developed this deep learning algorithm for temporal pattern discovery over Rochester epidemiology Project data. In recent years, ML applications on Chest X-ray data have been rising interest, such as lung segmentation, tuberculosis and cancer analysis, abnormality detection and multi-modality predictions. The most important datasets are large-scale public CXR datasets (e.g. CheXpert), chest-xray8, Padchest or MIMIC-CXR. In the terms of rapidly emerging coronavirus disease 2019 (COVID-19), the need to streamline patient diagnosis and management has become more pressing than over. Many measures have been taken to predict COVID-19 through medical imaging. The following datasets are one of the first datasets, which created for detecting COVID-19: Wang et al. [11] present COVID-Net, a deep convolutional network for COVID-19 diagnosis based on chest X-ray images. The dataset called COVIDx, which consists of 13800 chest radiography images from 13725 patient cases from three open access data repositories. Cohen et al. [12] describe the public COVID-19 image collection consisting of X-ray and CT scans. This dataset consists of more than 125 images. Which is one of the references for many researchers.

3. Conclusion

This paper has presented a survey of researches on the datasets in medical informatics. Medical informatics role in diagnosis and recognition is starting to be automated by DNNs with a clear advantage of being very efficient in diagnosing outcomes. Datasets are urgently needed to support the diagnosis and treatment of diseases with machine learning or artificial intelligence technologies. In most of the papers studied, the researcher's attempts were improving clinical decision making and optimizing clinical pathways by accurate prediction of length of stay, future illness, readmission, mortality. This goal can achieved by discovering meaningful data-driven features and disease characteristics. Although recently healthcare datasets are available, some specific data is often limited especially for rare diseases. So during the training of a DNN (in the case of small datasets) overfitting is arisen the makes the generalization with problem. Nevertheless, recently data from EHRs, which provide a detailed picture of the patients history, pathology, treatment, diagnosis, outcome, have been provided equally and therefore DL algorithms have mostly been employed in applications where the dataset is adequate amount.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRFK) funded by the Ministry of Education (2018R1D1A1B07041981).

References

[1] Kate Fultz Hollis, Lina F. Soualmia, Brigitte Séroussi" Artificial Intelligence in Health Informatics: Hype or Reality? IMIA and Georg Thieme Verlag KG.2019.

[2] Jiayi Shen1, Casper J P Zhang, Artificial Intelligence versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med Inform 2019.

[3] H. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation," CoRR, vol. abs/1505.00670, 2015.

[4] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, "Deep learning for healthcare decision making with emirs," in Proc. Int. Conf. Bioinformatics. Biomed. pp. 556–559. Nov 2014

[5] E. Putin et al., "Deep biomarkers of human aging: Application of deep neural networks to biomarker development," Aging, vol. 8, no. 5,

[6] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua, "Disease inference from health-related questions via sparse deep learning," IEEE Trans. Knowl. Data Eng, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[7] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Sci. Rep., vol. 6, pp. 1–10, 2016.

[8] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," J. Biomed. Informat. vol. 56, pp. 229–238, 2015.

[9] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnosewithLSTMrecurrent neural networks," CoRR, vol. abs/1511.03677, 2015.

[10] S. Mehrabi et al., "Temporal pattern and association discovery of diagnosis codes using deep learning," in Proc. Int. Conf. Healthcare Informat., Oct. 2015, pp. 408–416.

[11] Wang L, Wong A. "Covid-net: A tailored deep convolutional neural network design for detecton of COVID-19 cases from chest radiography images." arXive: 2003.09871, 2020.

[12] Cohen JP,Morrison P, Dao L, Roth K, dunong TQ, Ghassemi M" COVID-19 image data collection:, Prospective Prediction are the future". Arxive: 2006.11988, 2020.