AI for Medical Diagnosis and Treatment Use Case - Digital Healthcare

Sourcing Data

Overview

Identifying and acquiring the right data is a corner stone for any data-driven solutions atop of them those using AI and ML for building core models like in predictive analytics, optimization, object detection, computer vision, emotions detection and even simply clustering and segmentation. There are numerous ways to acquire and integrate data into the target modeling platforms depends on the source itself. Following are common approaches to accomplish such step. Open-source data is one key asset when working on such projects like the safe transportation in Smart cities where many countries are contributing their real-life data over many years including statistical and unstructured data.

Batch Loading

Typically take place by uploading or ingesting relatively large portion of the data at once where data is already stacked somewhere like on server or enterprise storage and could eb accessible though file-based protocols like SFTP or SHTTP suitable for file uploads. In many cases data already exists on a database or NOSQL data platform so that direct connection can be stablished an and ELT, CDC or JDBC connection can be used to load data from the source to the target modeling platform.

RESTful APIs

REST is an acronym for REpresentational State Transfer and an architectural style for distributed hypermedia systems. Roy Fielding first presented it in 2000 in his famous dissertation. Like other architectural styles, REST has its guiding principles and constraints. These principles must be satisfied if a service interface needs to be referred to as RESTful. A Web API (or Web Service) conforming to the REST architectural style is a REST API [3].

REST APIs are wieldy supported as means of connecting and consuming source data from applications, systems, devices and open data portals. Such support enables developers to integrate their applications directly with those sources thus override the need for batch loading which involves human intervention or requires unreliable hard-to-implement automation. Using direct APIs connection enables getting data updates immediately for real-time and applications.

OData

OData (Open Data Protocol) is an OASIS standard that defines the best practice for building and consuming RESTful APIs. OData helps you focus on your business logic while building RESTful APIs without having to worry about the approaches to define request and response headers, status codes, HTTP methods, URL conventions, media types, payload formats and query options etc. OData also guides you about tracking changes, defining functions/actions for reusable procedures and sending asynchronous/batch requests etc. Additionally, OData provides facility for extension to fulfil any custom needs of your RESTful APIs [4].

Data Sources for AI Topics

In the SoW activity 1 in this project, different AI potential topics have been elaborated with sub problems those need to be addressed to support human wellbeing in digital healthcare transformation. The following table provides example datasets those can be used to source data for usage in the budling AI models for the aforementioned topics and their associated problems. Ideally, actual data from an agriculture ecosystem deployed digital applications should be used and later get integrated into a working solution to address real-time, near-real-time and batch AI application patterns:

AI Topic	Description	Example Data Sources
Topic #1	Accurate Diseases Detection	https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD https://www.broadinstitute.org/data-software-and-tools?field_data_broad_tags%5B%5D=612&type=All&search_api_views_fulltext=&items_per_page=50 https://www.cs.rug.nl/~imaging/databases/melanoma_naevi/ https://github.com/v7labs/covid-19-xray-dataset https://www.iccr-cancer.org/datasets/published-datasets/ https://challenge.isic-archive.com/data/#2020 https://www.mortality.org/Data/DataAvailability

		https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection https://www.oasis-brains.org/ https://nihcc.app.box.com/v/DeepLesion https://adni.loni.usc.edu/data-samples/adni-data-inventory/
Topic #2	Early Diagnosis of Critical Diseases	https://archive.ics.uci.edu/dataset/174/parkinsons https://dbarchive.biosciencedbc.jp/index-e.html https://adni.loni.usc.edu/data-samples/adni-data-inventory/ https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/ https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html https://www.ukbiobank.ac.uk/enable-your-research/about-our-data https://seer.cancer.gov/statistics-network/
Topic #3	Clinical Genomics	https://www.dgidb.org/downloads https://dbarchive.biosciencedbc.jp/index-e.html https://leo.ugr.es/elvira/DBCRepository/ https://adni.loni.usc.edu/data-samples/adni-data-inventory/ https://registry.opendata.aws/1000-genomes/

		https://www.broadinstitute.org/data-software-and-tools?field_data_broad_tags%5B%5D=612&type=All&search_api_views_fulltext=&items_per_page=50 https://www.cancer.gov/ccg/access-data https://www.genome.jp/tools-bin/dinies?mode=data&id=example&pa=0&thval=0.3
Topic #4	Electronic Health Records (EHR)	https://idr.ufhealth.org/wordpress/files/2021/09/DataGuide_2021_August.pdf https://vitaldb.net/dataset/ https://inspire.or.kr/ https://pcornet.org/data/ https://physionet.org/content/mimiciv/0.4/
Topic #5	Smart Treatment	https://vitaldb.net/dataset/ https://zenodo.org/records/7622128 https://physionet.org/content/mimiciv/0.4/ https://eicu-crd.mit.edu/gettingstarted/access/ https://github.com/AmsterdamUMC/AmsterdamUMCdb/wiki https://hirid.intensivecare.ai/
Topic #6	Virtual Health Assistants	https://www.nhs.uk/conditions/ https://drive.google.com/file/d/1ImYUSLk9JbgHXOemfvyiDiirluZHPeQw/view https://www.digitisation.eu/impact-dataset/

		https://zenodo.org/records/7622128 https://eicu-crd.mit.edu/gettingstarted/access/ http://bci.med.tsinghua.edu.cn/download.html
Topic #7	Statistical Modeling and Simulation of Events	https://apps.who.int/gho/data/node.resources https://healthdata.gov/browse https://dhsprogram.com/data/available-datasets.cfm https://dbarchive.biosciencedbc.jp/index-e.html https://data.gov.au/search?q=healthcare https://wonder.cdc.gov/DataSets.html https://data.cms.gov/provider-data/?redirect=true https://seer.cancer.gov/statistics-network/ https://open.fda.gov/data/datadictionary

References & Resources

#	Topic	Source
1	Critical Challenges of Artificial Intelligence and Data in Open Source	https://linuxfoundation.org/wp-content/uploads/LFR_LFAID_Guide_to_Enterprise_Open_Source_letter_082222.pdf
2	Batch Loading ETL and ELT as a Code	https://datalakehouse.org/what-is-etl-as-code/
3	Introduction and resources for REST APIs	https://restfulapi.net/
4	Understanding oData in 6 Steps	https://www.odata.org/getting-started/understand-odata-in-6-steps/

5	An Example of Using APIs to Access Open Data – Widely Supported	https://dev.socrata.com/consumers/getting-started.html