**AI MONITORING**

*Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failures (Nushi et al., 2018)*

* Introduces *Pandora* which describes and explains failures in ML models
* It clusters input domain into topical clusters and uses the examples in each cluster to learn interpretable decision-tree classifiers for predicting and summarizing failure conditions
* Case study on an image captioning system with data from MSCOCO dataset
* Performance views contain content-based views (use ground truth or input data to learn situations associated with poor performance) and component-based views (use of metrics and confidence level to model how internal component dynamics lead to errors)
* Content and component feature data from crowdsourcing and internal system data flows
* Failure analysis report consists of evaluation metrics (from crowd workers), decision-tree performance predictors, and feature rankings (informative features for predicting failure)

*CheXbreak: Misclassiﬁcation Identiﬁcation for Deep Learning Models Interpreting Chest X-rays (Chen et al., 2021)*

* Introduces a corrective algorithm to selectively flip model predictions at inference time, which aims to solve misclassification problems
* Done by investigating patient subgroups (age, some radiographic features) more likely to be misclassified, developing misclassification predictors by model outputs and clinical features, finally introducing the corrective algorithm
* Data and model: CheXpert and top 10 performed models in the competition
* Misclassification predictors: best AUROC for “same label” (clinical + target output) and “all labels” (clinical + all outputs) approach, while using clinical features (age, sex, lateral views) only performed worse than a baseline using the negative absolute distance from prediction threshold to model output
* Corrective algorithm: flips more wrong model predictions than correct and increases true positive of disease predictions as much as possible. The algorithm gets significant F1 score improvement on predicting some diseases while not getting any significant F1 score drop (95% C.I.)

*The Clinician and Dataset Shift in Artificial Intelligence (Finlayson et al., 2021)*

* Dataset shift is a major driver of AI system malfunction, which occurs when a machine learning system underperforms because of a mismatch between the data set with which it was developed and the data on which it is deployed.
* Common causes of dataset shift are in terms of technology (software), changes in population and setting (demographics), and changes in behavior
* Recognition strategies includes checking differences, flagging changes / errors / poor performance, monitoring model accuracy and calibration, local validation before deployment
* Mitigation strategies include retraining / tuning / redesigning / recalibrating models, identifying updates for variable mappings

*Learning to Validate the Predictions of Black Box Machine Learning Models on Unseen Data (Redyuk et al., 2019)*

* Introduces an approach to automatically assess changes in prediction scores of black box ML models on unseen target data. Problem setting is on classification tasks
* Done by asking a domain expert to specify a set of dataset shift generators, then applying the shift generators to the testing dataset and compute new prediction scores and features
* By training a regression model on different shift generators, prediction validation for unlabelled target data is done by computing the output of black box and the features, then fit into the regression model to see if the prediction score is far from the prediction score of the testing data
* Evaluation on logistic regression and feed-forward neural network: three datasets and different types of dataset shift all show a small mean absolute error (MAE) between the performance predictor and the true target label score
* An updated version is *Learning to Validate the Predictions of Black Box Classifiers on Unseen Data (Schelter et al., 2020)*, with codes available at <https://github.com/schelterlabs/learning-to-validate-predictions> . The updated version tested more classifiers, including a gradient-boosted decision trees and a CNN

*MLOps Principles*

* The complete MLOps process includes three broad phases of “Designing the ML-powered application”, “ML Experimentation and Development”, and “ML Operations”. It includes continuous integration (CI), continuous delivery (CD), etc.
* MLOps Principles include versioning, testing, automation, reproducibility, deployment, and monitoring. For automation, the 3 steps are manual process (experimental nature, e.g. Jupyter notebooks), ML pipeline automation (continuous training of model, data and model validation steps), CI/CD pipeline automation (everything is automated)
* Change in a ML development pipeline includes data, model, code
* Best practice (documentation): data (data sources, how / where to get data, labelling methods), model (model selection criteria, experiment design, model pseudo-code), code (deployment process, method to run locally)
* Best practice (project structure): well-organized folder structure, with data (raw / processed data, data engineering pipeline, test for data engineering methods), model (trained model, notebooks, feature engineering, model engineering), code (bash / shell scripts, tests, deployment files (e.g. Docker files))

*Dissecting racial bias in an algorithm used to manage the health of populations (Obermeyer et al., 2019)*

* Label choice in algorithms affects both predictive performance and racial bias
* In this specific study, less-healthy (in terms of chronic illness) Blacks scored at similar risk scores to more-healthy Whites. If there is no gap in health conditional on risk, the fraction of Black patients auto-identified for the program will increase from 17.7% to 46.5%
* The bias is led by Black patients generating lesser medical expenses conditional on health, due to substantial barriers to health care access and direct discrimination, changes to the doctor-patient relationship, etc.
* Meanwhile, the algorithm measures risk score, and the risk score is related to medical cost. So, Blacks with similar health got lower medical cost than Whites counterparts, meaning that their risk score is lower
* Experiments on label choice (predicted cost / avoidable cost / health) shows that the fraction of Black patients in the highest risk group differs a lot although all label choices perform similarly in terms of the concentration of all realized outcomes in the highest percentile

*The What-If Tool: Interactive Probing of Machine Learning Models (Wexler et al., 2020)*

* What-If Tool (WIT) is a model-agnostic interactive visual tool for model understanding. It is part of Tensorboard but also available as an extension for Jupyter / Colab notebooks
* It solves 5 user needs: test multiple hypotheses with minimal code, use visualizations as a medium for model understanding, test hypotheticals without having access to the inner workings of a model, conduct exploratory intersectional analysis of model performance, evaluate potential performance improvements for multiple models
* This can be achieved by having several visualization options. Further model understanding can be done by having a datapoint editor, being able to show closest counterfactual value to a selected data point, having partial dependence (PD) plots, etc.
* The tool is also capable of uncovering nontrivial, long-lived bugs and provide practitioners with new insights into their models. Useful for ML novices / experienced engineers

*Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead (Rudin, 2019)*

* A review suggesting that interpretable ML models should be used whenever possible instead of explanatory models for black-box
* The explanatory ML model of a black-box model introduces an extra error term to the black-box, making the explanation not trustworthy
* Explanations may also do not provide enough detail to understand what the black-box does, e.g. attention maps in CV highlight main decision regions without explaining how the decision is made by using that specific region
* Challenges in interpretable ML includes constructing optimal logical models, constructing optimal sparse scoring systems, difficulty in defining interpretability for specific domains

*Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks (Nestor et al., 2019)*

* State-of-the-art models decay in prediction quality when trained on historical data and tested on future data, particularly in response to a system-wide record-keeping change
* Dataset: MIMIC-III, but with Agreement to perform non year-agnostic training regimes
* Prediction on mortality task and long length of stay task from data transformation to 24-hours time-series by logistic regression (LR), RF, LSTM, GRU-D models
* Solution proposed: data representation using clinical aggregation. It aggregating raw features into expert-defined clinical concepts, then apply imputation on 78.25% missingness
* Year-agnostic (public dataset) VS non year-agnostic (with Agreement): year-agnostic methods creates unrealistic upper bounds to model performance, especially on raw data
* Clinical aggregation increased robustness to data shift (increased AUROC, decreased SD and maximum AUROC drop with respect to using raw data and other pre-processing methods)
* Model complexity has no advantage in a non-stationary data setting (RF, LR preferred). Also, performance on unseen future data does not dramatically change even as the training dataset grows. Future: more difficult task / find alternative not-measured signals to improve
* Future research: automatically generating mappings between multiple electronic health record (EHR) systems, since clinical aggregation involves burden of manual concept definition

*Continual learning in medical devices: FDA’s action plan and beyond (Vokinger et al., 2021)*

* Continual learning: decision logic of models is updated through new data while retaining previously learned knowledge. It has been introduced in other sectors like Tesla cars
* Risks of continual learning includes new errors introduced by new data, performance deterioration if new data is biased, possible catastrophic forgetting of previous knowledge
* FDA issued an action plan containing a predetermined change control plan, which should include which aspects is intended to be changed through learning (prespecifications) and the associated methodology being used to implement changes (algorithm change protocol)
* Prespecifications should determine which post approval data is relevant and to be selected
* Standardised testing routines are recommended for methods in algorithm change protocol

*A unifying view on dataset shift in classification (Moreno-Torres et al., 2012)*

* Types of dataset shift includes covariant shift (population distribution (x) change, X->Y problems), prior probability shift (target distribution (y) change, Y->X problems), concept shift (x,y relationship change), and general dataset shift
* Sample selection bias is a main reason for dataset shift, and it could lead to dataset shifts (except concept shift)
* Adversarial environments could lead to general dataset shift
* Non-stationary (time or space) environment could generate covariant / prior / concept shift
* Solutions include shift detections and adapting to the shift once it occurs

(New article to be added)

**PHARMACOEPIDEMIOLOGY**

*External validity of randomised controlled trials: To whom do the results of this trial apply? (Rothwell, 2006)*

* Randomised controlled trials (RCTs) must be internally valid, but its external validity is not considered enough by researchers, governmental regulators, etc.
* RCTs should at least be designed and reported to let clinicians to judge to whom they can reasonably applied
* Factors undermining external validity includes setting of the trial, selection and exclusion of patients, characteristics of randomised patients, differences between trial protocol and routine practice, outcome measures and follow-up, adverse effects of treatment
* The design and reporting of trials should be improved, as well as the reporting of trial eligibility criteria

*Preventing adverse drug events (Leape, 1995)*

* Humans errors are a reason for preventable adverse events, and it is caused by system failures or characteristics of work or the workplace that makes errors more likely and difficult to be detected
* Powerful system changes include computerized physician drug order entry and redefinition of the role of pharmacists. Computer system reduces the likelihood of errors by providing information and orders to eliminate transcription errors
* Redesigning systems aim to prevent and detect errors. Information from physicians, nurses, and pharmacists are useful to achieve the aim

*Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports (Evans et al., 2001)*

* PRR is used to generate signals for adverse drug reaction (ADR), calculated by the ratio of the proportion of specified reaction(s) of interest for the drug of interest and the proportion of that for all other drugs in the database
* Signal is defined with PRR > 2, chi-squared > 4 and > 3 cases
* Using PRR with the UK Yellow Card database, 70% of the evaluable potential signals were recognized adverse reactions
* Advantage of PRR : derive solely from spontaneous ADR data, easy to calculate and interpret, using proportions avoids biases caused by variable degrees of reporting
* Limitation of PRR : strong signals of a certain drug will reduce the magnitude of other PRR signals with that particular drug

*Methods for safety signal detection in healthcare databases: a literature review (Arnaud et al., 2017)*

* Reviewed existing approaches for safety signal detection
* Performance-wise, the self-controlled designs, the sequence symmetry analysis (SSA), and the supervised machine learning (SML) approach are interesting for health-care databases
* Self-controlled: only one cohort of patients is considered and each patient is his/her own control, control for all time-invariant (e.g. chronic comorbidities) and patient-invariant confounders, but with settings tuned to knowledge about association (a priori assumption)
* SSA: compare chronological order between 2 drugs in terms of crude/null-effect sequential ratio, detect a drug as a prescription after possible ADR of another drug, dedicated to longitudinal data , developed for large-scale and standardized applications, understandable
* SML: high performance, but not assessed as intensive as self-controlled/SSA, black-box not understandable by clinicians, providing probabilities of drug-event association instead of risk estimates is not ideal for highlighting worthy-to-investigate signals from thousands of signals

*The medical algorithmic audit (Liu et al., 2022)*

* Algorithmic audit approach is based on the SMACTR framework (scoping, mapping, artifact collection, testing, reflection)

Graphical user interface

Description automatically generated

* Testing methods include exploratory error analysis, subgroup testing, adversarial testing
* Person conducting the audit should possess various skills

*Integrating AI Algorithms into the Clinical Workflow (Juluru et al., 2021)*

* System should commence image analysis immediately after image acquisition, identify and report examination quality problems, generate and send preliminary reports for review / signature, allow users to correct AI results, dashboard for users to monitoring performance

Diagram

Description automatically generated

(New article to be added)