Project description

By: Jinyan Tao

Research summary

My research experience covers topics in chemoinformatic and biomedical informatics.

In all projects, I leveraged machine learning/deep learning to answer domain-specific research questions. My goal-driven persistence and problem-solving skills are highly recognized by project advisors and teammates who are talents and experts in their respective fields.

Meanwhile, during internship placements, I developed strong programming skills in Python, R, and Matlab and implemented a reproducible machine learning pipeline. Furthermore, I greatly enjoy designing web-based interactive applications for insightful data visualization to inform precision cancer target selection. As an active member in the local data science and machine learning community, I keep pace with the latest technology innovations by joining various seminars and keen on exchanging my research findings and opinions in thought-provoking discussions. I am in strong belief, that data science and machine learning can create real-world impacts in both science and society.

1. Ocular disease recognition with Convolutional Neural Networks

Department of Computer Science, ETH Zurich
GitHub: https://github.com/PinaColadast/ocular-disease-recognition-with-CNNs
Project description:

The goal of the project was to accurately diagnose ocular diseases such as diabetes, myopia, glaucoma and 5 other diseases with fundus images by traditional and novel computer vision deep neural networks.

Driven by a sheer interest in deep learning on image diagnosis, I joined a Kaggle challenge to implement Convolutional Neural Networks (CNNs) and novel ImageNets architecture such as Inception, VGG, ResNet to identify ocular diseases from fundus images with PyTorch.

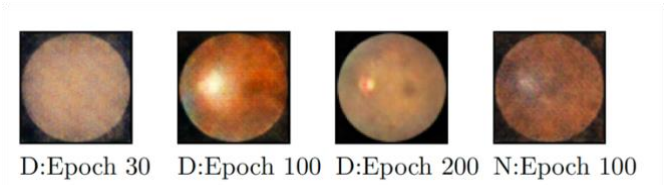

Figure 1 *Example of generated image using DCGAN for diabetes (D) and normal (N) eyes*

Given an imbalanced dataset of a multi-label prediction task, I applied classic methods such as contrast enhancement, noise adding and Generative Adversarial Networks for image augmentation to improve class balance and adversarial perturbation to strengthen the model robustness. The model accuracy achieved over 0.8 for glaucoma and diabetes disease recognition.

For implementation of AI in biomedicine, interpreting model decision is the key and often challenge to promote trustworthy solutions. Thus, I independently researched the literature and creatively implemented an innovative tool, the Gradient-Weighted Class Activation Mapping (Grad-CAM), to localize the discriminating hotspots on images for the classification, which typically traces back to a convolutional layer and provide region-weighted importance heatmap. The analysis on images from the Grad-CAM corresponds well to the clinical ophthalmic diagnosis rationales. This project greatly inspired me that the reliability of an AI solution should be addressed equally as its robustness. Therefore, not only do I wish to create robust data-driven solutions but also interpret them to facilitate a reliable usage of deep learning.

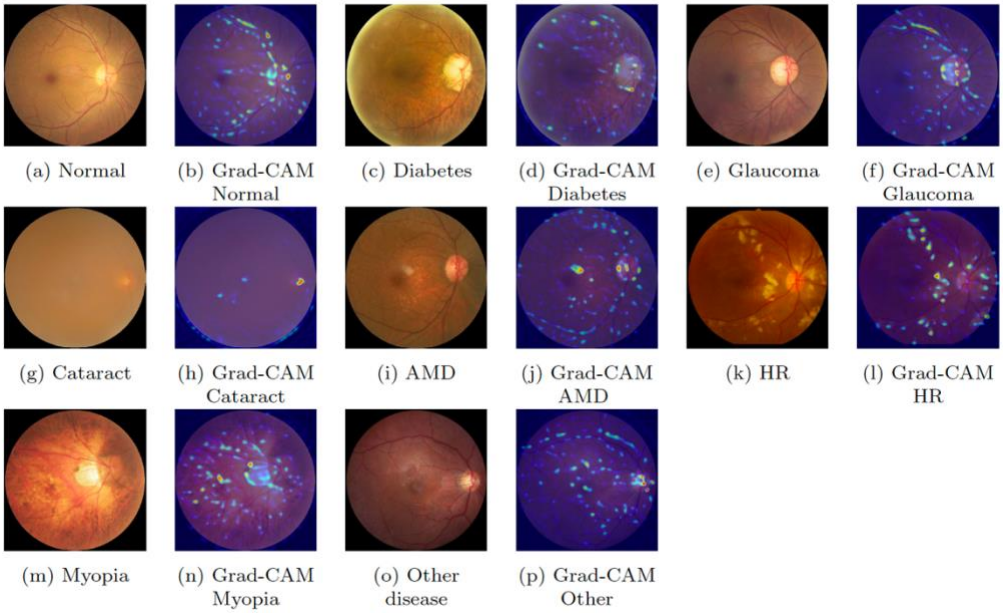

Figure 2 *Original images of 8 classes and their Grad-CAM localizations. Grad-CAM localization was visualized on the 4th convolutional layer in CNN. AMD refers to age-related macular degeneration and HR to hypertensive retinopathy. Regions of interest are colored by importance, from green to red.*

2. Matching Mutational Signature on Gene Panel Assays

Biomedical informatics, Department of Computer Science, ETH Zurich
Supervisor: Prof. Gunnar Raetsch
GitHub: https://github.com/PinaColadast/mutsig
Project description:

Mutational signatures are patterns of somatic nucleotide changes observed in cancer patients. Depending on the etiology of cancer, the types of mutations observed can vary largely. Determining the mutational signatures' presence in an individual patient can help provide clinically relevant information. Our goal is to identify and recognize the signatures present in patients based on panel gene data, with machine learning approaches, to develop a lightweight yet robust method to predict the signatures.

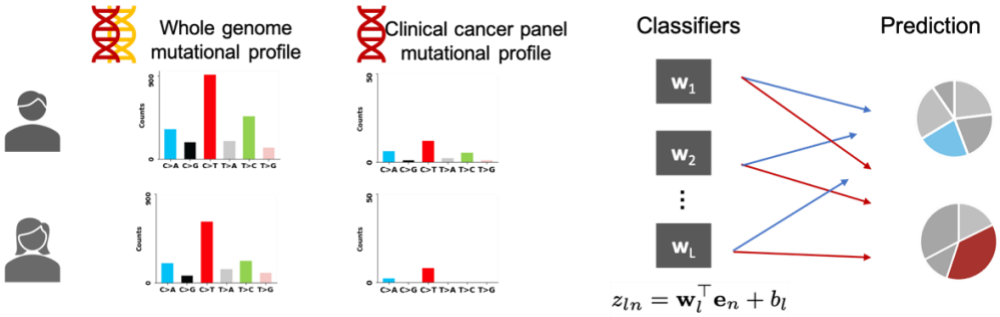

Figure 3 *Concept of matching mutational signatures on cancer gene panels.*

To begin with, based on whole-genome data of 2703 patients from the PCAWG dataset, I carried out data integration and exploratory data analysis. Individual mutational profile in WGS and panel gene context were constructed respectively. To effectively compare mutational profiles, I benchmarked several distance computing methods (Euclidean, entropy, cosine distance, etc.) on noised/clean data for sensitivity testing. Following the comparison, the unique signature exposure profile of individuals was decomposed from the mutational profiles by Non-negative matrix factorization (NMF) and matching pursuit algorithms. Further analysis using clustering based on signature exposures was performed to summarize the diversity and homogeneity of signature compositions across samples, the optimal number of clusters was recommended via computing silhouette and purity score. The information from clustering then served as a critical guideline for sample labeling.

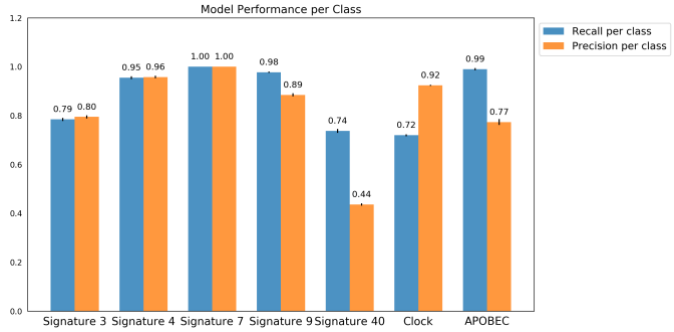

Figure 4 *Final model performance on multi-class classification on 7 signatures. Overall performance accuracy achieves 0.78*

In the end, after benchmarking multiple machine learning algorithms and kernel-based model selection, I managed to implement a Gaussian SVM classifier that accurately predicts seven signature labels that are prominent across multiple aggressive cancer types. Having solved data imbalance by bootstrapping, overall prediction accuracy reaches 0.78 and 0.73 in whole-genome and panel gene context respectively. To promote the model interpretability, I visualized the kernel similarity metrics as a heatmap to understand the model classification decision. This model will be further implemented for non-computational professionals to facilitate cancer diagnosis and personalized medicine.

3. Lipophilicity modeling for chemical entities

Department of Computational Chemistry & Biology, Merck KGaA, Germany
Supervisor: Dr. Michael Krug, Prof. Dr. Gisbert Schneider
Project description:

Due the challenges in measuring molecular log P values experimentally, we came up with the idea of computing log P values by deriving mathematical correlations between pKa and experimental log D values of molecules. Serving the computed log P as the ground truth, we aimed to develop regression model to predict molecular log P value using machine learning approaches.

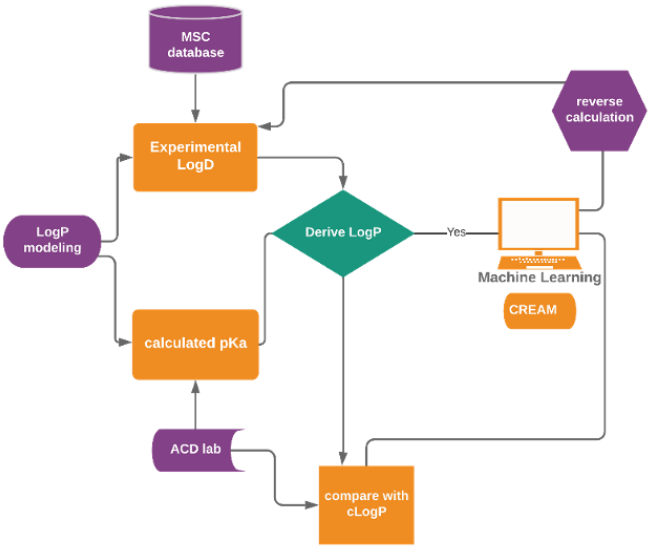

Figure 5 *project concept of Log P modeling*

In charge of lipophilicity modeling based on ~30000 small molecules, I independently carried out literature research to derive mathematical correlations of log D, log P and pKa values and performed data analysis and cleaning. After rational selection of molecular descriptors, ML algorithms (SVM, Neural Nets, SGD, Adaboost etc.) and validation methods, I performed grid-search for hyperparameter tuning then successfully implemented a predictive log P regression model with R^2 and RMSE value of 0.864 and 0.35 respectively.

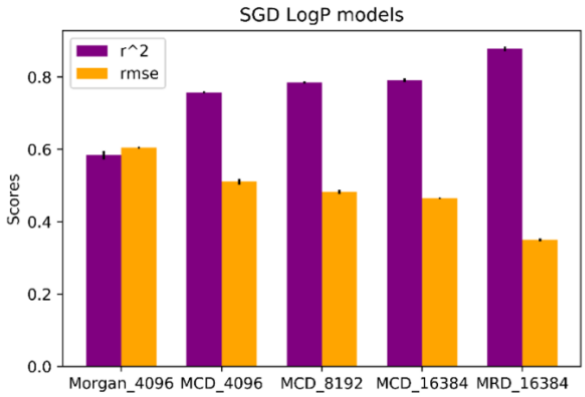

Figure 6 *Benchmarking of molecular descriptors with Stochastic Gradient Descent regression model*

In addition, I benchmarked novel and conventional substructure analysis methods based on ~200,000 small molecules in Python and implemented on the HPC, in order to optimize molecule characterization. By never stop questioning, I discovered abnormalities in the data distributions, which led to an identification and correction of data processing errors in the company database and saved millions of chemical compounds from being erroneously reported.

4. Machine learning assisted ligand-based drug design

Institute of Pharmaceutical Sciences, ETH Zurich
Supervisor: Prof. Dr. Gisbert Schneider
Project description:

In my project, with the assistance of modern machine learning algorithms, we strived to develop an affinity prediction model to identify potent ligands against protein peroxisome proliferator-activated receptor - δ, that governs multiple critical metabolic pathways.

With ligand information available on ChEMBl, I computed molecular fingerprints to describe structural features of ligands, first to conduct clustering analysis to visualize intra-set relationships between ligands, then benchmarked various molecular descriptors to select the optimal molecular representations. With fingerprints as feature inputs and IC50 as predicted value, I generated regression and classification models by implementing machine learning algorithms such as random forest, support vector machines, Gaussian Naive Bayes, and linear regression.

5. Data-driven prediction of the age of NGO's donors

Hack4good Hackathon, Analytics Club, ETH Zurich
GitLab: https://gitlab.com/analytics-club/hack4good/hack4good-fall-2020-public/helvetas
Report : https://analytics-club.org/wordpress/wp-content/uploads/2021/03/Helvetas.pdf

Project description:

A big challenge for a fundraising organization is the generational change of its donor base. The goal of the project was to enhance demographic understandings of the NGO's donor base, in particular, to impute the missing birth years of 2/3 donors in the NGO's database, which helps to understand the generational shifts across decades and to promote targeted fundraising campaigns.

I had been actively involved in the hackathon to offer impactful data-driven solutions to a Swiss NGO Helvetas by inferencing the age distribution of their donors. Collaborating with three ETH data science talents, within eight weeks, we were able to integrate and implement a data engineering and machine learning pipeline as a ready-for-use Python package that is suitable for Windows and Unix users. During the hackathon, I was responsible for building data analysis, data preprocessing and model training and testing modules in Python scripts, which allows the usage of classic and robust ML algorithms such as random forest and neural nets. I independently improved the accuracy of the random forest by 20% by performing a grid search of hyper-parameters. In the end, we delivered a reproducible machine learning solution that helps the Swiss NGO easily comprehend their donor information and prepare targeted fundraising campaigns.