Machine Learning Odyssey
Final Remark

Oct 31, 2024

Seungeun Lee

So far…

We so far have covered various topics for Tabular data, including

- Data Preprocessing (Scaling, dealing with Class Imbalance, …)
- Classical Machine Learning based methods, including XGBoost and LightGBM
- Deep Learning for Tabular data (TabNet, …)

In this section,

I would like to talk about

1. **My vision for the future of tabular data**,

and in the long term,

2. **The vision of the whole process of data analysis, which I call “Data Ecosystem”**

0. Personal Motivation

- My first research,

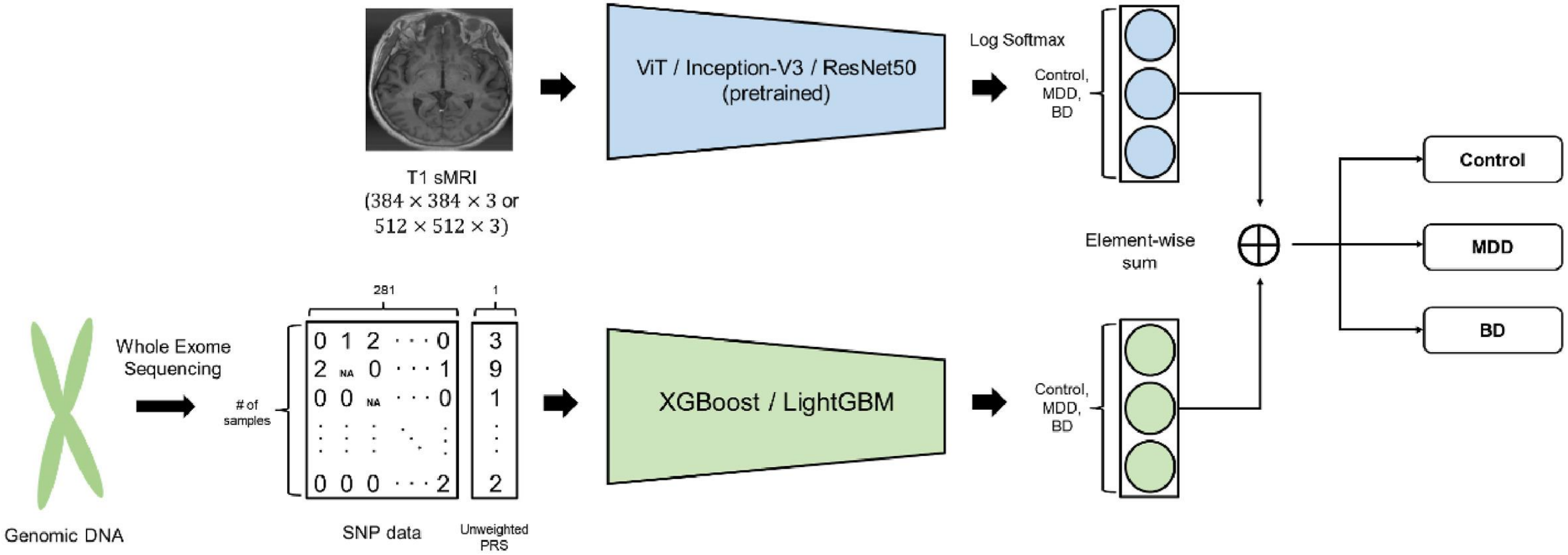
S. Lee et al., Multimodal integration of neuroimaging and genetic data for the diagnosis of mood disorders based on computer vision models (https://doi.org/10.1016/j.jpsychires.2024.02.036)

designed a multimodal fusion approach for classifying mood disorders

by integrating patient-specific brain structural MRI (sMRI) scans with DNA whole-exome sequencing (WES) data and the corresponding unweighted polygenic risk scores (PRS)

0. Personal Motivation

While the brain structural MRI (sMRI) scans could be analyzed with Deep Learning-based image models, DNA WES & PRS data (tabular data) had no choice but to model with Machine Learning methods, because of the inferior performance of Deep Learning models for tabular data, especially with the real-world dataset.



0. Personal Motivation

- Such an ensemble of DL and ML models eventually led to the incapability of **“whole gradient update”**, possibly bringing about the suboptimal performance of the entire ensemble model.
- After conducting this research, I truly wanted to design a Deep Learning-based architecture, especially designed for tabular data that is

(1) Fast & Lightweight & Easy to implement

(2) Showing modest performance not only with the benchmark but also with the real-world dataset

(3) Interpretable

even when compared with XGBoost or LightGBM, the game-changers of tabular data analysis.

1. Deep Learning for Tabular data

(1) Merits of classical ML-based methods for tabular data

- Fast & Lightweight & Easy implementation
- Guaranteed Performance (Especially GBDT-based methods)
- Interpretable (Internal Feature Importance Mechanism; Recall XGBoost & LightGBM)

(2) Demerits of classical ML-based methods for tabular data

- Poor Compatibility with Deep Learning Models – possibility of being eliminated in this Multimodal Era

1. Deep Learning for Tabular data

- Recent advances in DL for Tabular data

(1) TabNet

(2) FT-Transformer

(3) TabPFN

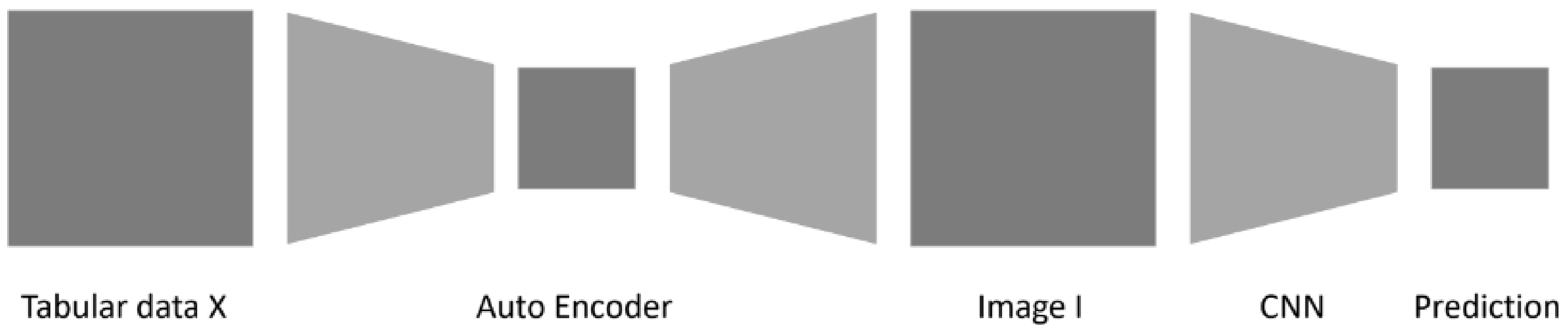
(4) Tunetables

1. Deep Learning for Tabular data

So, I have designed an interpretable tabular data classification framework that

(1) transforms tabular data into realistic images

(2) and utilizes Bayesian methods to incorporate latent variables (Interpretation)



1. Deep Learning for Tabular data

1. Fast & Lightweight & Easy to Implement

- Thanks to the MLP-based Autoencoder architecture, combined with Simple CNN at the backend

2. Performance

- shows a modest performance when compared with the existing SOTA models & currently validating our model with the real-world dataset (granulation process device parameter dataset), attained from Handok Pharmaceuticals

3. Interpretable

- Through the Bayesian methods that incorporate latent variables (before transforming tabular data to images)

2. Data Ecosystem

- I have keen interests not only for Deep Learning for **Tabular data** and **Interpretable** methods, but also for **Responsible Machine Learning Operations (MLOps)**.

- Those three fields may seem different at a first glance, but they ultimately converge – I call this as a ***"Data Ecosystem"***

- The goal is to analyze **ubiquitous tabular data interpretably** with good performance and **safely** deploying it through **MLOps**, while **responsibly** and **reliably** addressing the process using a **mathematical** lens.

2. Data Ecosystem

- Please also check my paper

S. Lee et al., "BCCP: An MLOps Framework for Self-cleansing Real-Time Data Noise via Bayesian Cut-off-based Closest Pair Sampling" *(to appear, https://duneag2.github.io/publications/)*

- An MLOps pipeline for real-time noise control, enabling self-cleansing and maintaining performance in data corruption scenarios using Bayesian methods
- To devise a reliable MLOps pipeline, we have leveraged the mathematical tools