Assessing Predictive Performance and Gene Selection Consistency in Gene Expression Models

Group 2: Nick Treloar, Jiamin Zou, Yalei Zhao

Motivation

- Gene expression predicting disease status
    - Many studies have used gene expression data to predict disease status, progression, and survival

- Can basic, well-established machine learning models provide reasonable predictive performance on gene expression data?
    - LASSO, SVM, Random Forest, and Neural Network

Article | Open access | Published: 31 October 2025

Gene expression signatures from whole blood predict amyotrophic lateral sclerosis case status and survival

J Biom Biostat. Author manuscript; available in PMC: 2019 May 22.

Published in final edited form as: J Biom Biostat. 2018 Dec 11;9(5):417.

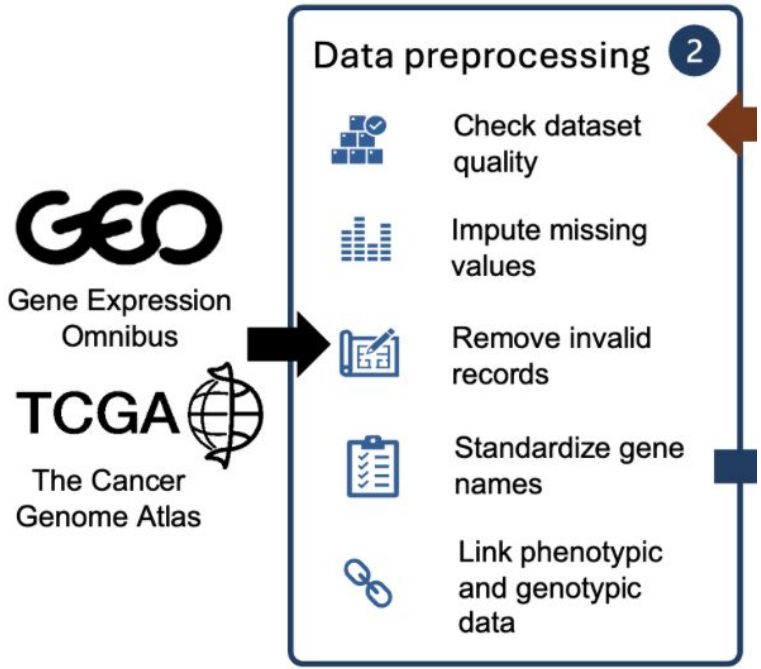
Deep Learning Methods for Predicting Disease Status Using Genomic Data

Toxicology. 2012 Nov 9;303:83–93. doi: 10.1016/j.tox.2012.10.014

Gene expression profiling to identify potentially relevant disease outcomes and support human health risk assessment for carbon black nanoparticle exposure

Data Source

- Kaggle: GenoTEX: LLM Agent Benchmark for Genomic Analysis https://www.kaggle.com/datasets/haoyangliu14/genotex-llm-agent-benchmark-for-genomic-analysis/data#dataset-structure
- 132 unconditional problems and **1,252 conditional problems**
- Average of **167** samples per dataset
- Average of **18,530 normalized gene features** per dataset



Dataset

```
output/                  # Analysis output data
├── preprocess/          # Preprocessed datasets
│   ├── Trait_1/
│   │   ├── clinical_data/
│   │   │   ├── GSE12345.csv
│   │   │   ├── Xena.csv
│   │   │   └── ...
│   │   ├── gene_data/
│   │   │   ├── GSE12345.csv
│   │   │   ├── Xena.csv
│   │   │   └── ...
│   │   ├── cohort_info.json
│   │   ├── GSE12345.csv
│   │   ├── Xena.csv
│   │   └── ...
│   └── ...
```

GSE: cohort ID
GSM: sample ID

```
/Users/yaleizhao/Documents/625/Alzheimers_Disease
├── clinical_data
│   ├── GSE109887.csv
│   ├── GSE117589.csv
│   ├── GSE122063.csv
│   ├── GSE132903.csv
│   ├── GSE137202.csv
│   ├── GSE139384.csv
│   ├── GSE185909.csv
│   ├── GSE214417.csv
│   └── GSE243243.csv
├── cohort_info.json
├── gene_data
│   ├── GSE109887.csv
│   ├── GSE117589.csv
│   ├── GSE122063.csv
│   ├── GSE132903.csv
│   ├── GSE137202.csv
│   ├── GSE139384.csv
│   ├── GSE185909.csv
│   ├── GSE214417.csv
│   └── GSE243243.csv
├── GSE109887.csv
├── GSE122063.csv
├── GSE132903.csv
├── GSE137202.csv
├── GSE139384.csv
├── GSE185909.csv
├── GSE214417.csv
└── GSE243243.csv
```

Clinical Data

	GSM7920782	GSM7920783	...	GSM7920825
Breast_Cancer	1.0	1.0	...	0.0
Age	49.0	44.0	...	74.0
Gender	0.0	0.0	...	1.0

Gene Expression Data

	GSM7920782	GSM7920783	GSM7920784	...	GSM7920825
A2M	13.210	13.238	14.729	...	7.359
ACVR1C	5.128	5.337	5.611	...	8.151
ADAM12	9.807	12.374	9.953	...	9.266
...	...	...	...	...	...
ZEB2	9.534	10.488	10.553	...	8.151

Linked Data

	Breast_Cancer	Age	Gender	A2M	ACVR1C	ADAM12	...
GSM7920782	1.0	49.0	0.0	13.210	5.128	9.807	...
GSM7920783	1.0	44.0	0.0	13.238	5.337	12.374	...
...	...	...	...	...	...	...	...
GSM7920825	0.0	74.0	1.0	7.359	8.151	9.266	...

Study Questions

Predictive Performance:

- How well do classical machine learning models (LASSO, SVM, Random Forest, Neural Network) predict disease status using gene expression data?

Gene Selection Consistency:

- Do these models identify similar important genes?
- How consistent are the genes identified as important across different cross-validation splits?

Biological Validity:

- Are some genes identified important matching existing literatures? If not, why?

Proposed Workflow

Merge multiple cohorts' linked data into one dataset

- Train/Test Split
- 5 folds

- LASSO, SVM, Random Forest, Neural Network
- Model-specific feature selection on the training fold
-Train model using selected features

-Prediction accuracy, RMSE, etc.
-Selected genes (stability across folds)

Compare the four models
-Average accuracy
-Average RMSE
-Jaccard Index (measure overall stability of feature selection)

Current Work - Models on Alzheimer's Disease

- Data: 8 cohorts; 567 total samples
- Models:
  - LASSO
  - Support Vector Machine
  - Random Forest
  - Neural Network

	file	rows
1	GSE109887.csv	78
2	GSE122063.csv	100
3	GSE132903.csv	195
4	GSE137202.csv	30
5	GSE139384.csv	12
6	GSE185909.csv	35
7	GSE214417.csv	24
8	GSE243243.csv	93

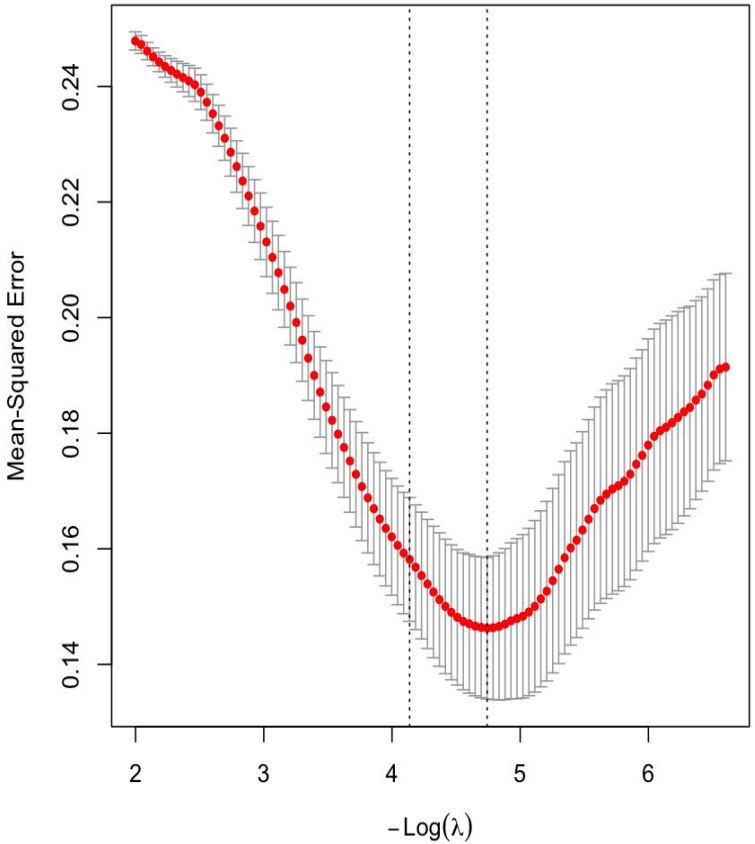
Lasso:

- Loss function:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- Adds a penalty based on the **sum of absolute coefficient values**
- The penalty strength is controlled by **λ**
- Some coefficients can be shrinked to **0** when the number of input variables is very large
- Enables **variable selection** and reduces model complexity

- Train data: original data
- The final model kept **56 variables**, achieved a accuracy **of 0.8012**, and completed in **0.2 seconds**



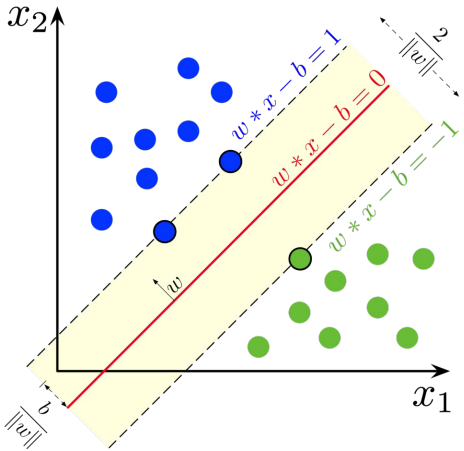
Support Vector Machine:

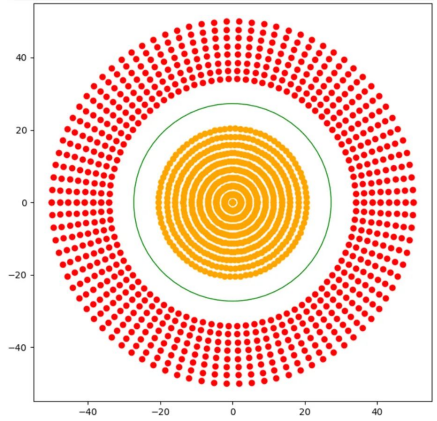
- Used for: **binary classification**
- Core idea: **map data into a higher-dimensional feature space** and find the **maximum-margin hyperplane** that separates two classes

- Loss function:

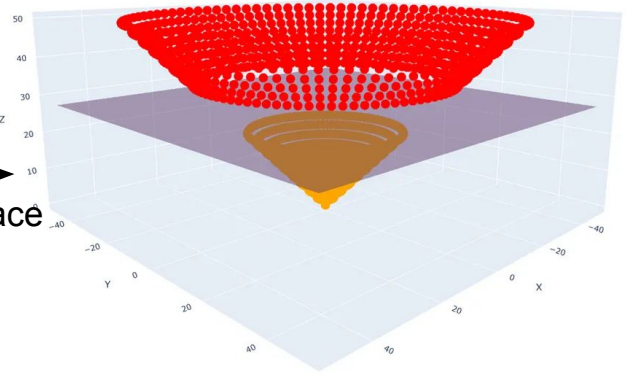
$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\zeta_i$$

- Train data: **56 variables** selected by Lasso
- achieved a accuracy **of 0.883**, and completed in **1.3 seconds**



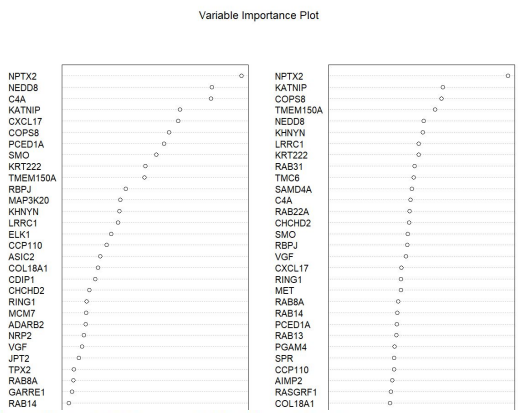


Map to high dimensional space



Random Forest

Random Forest selected 1472 features.

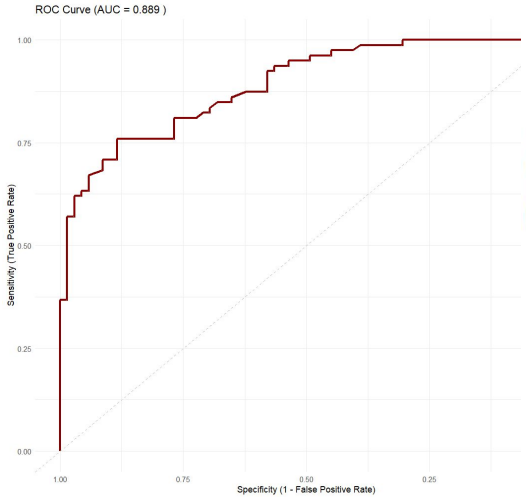


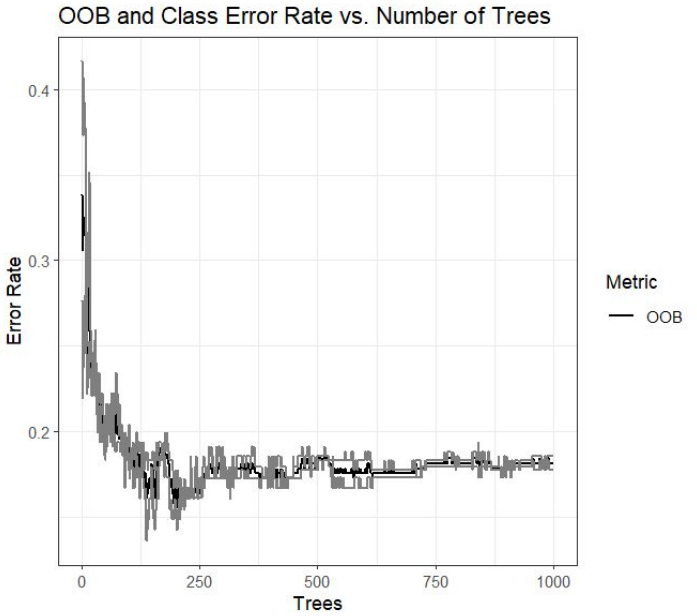
Training (OOB) Accuracy: 0.7902

Test Set Accuracy: 0.8311

```
          Reference
Prediction X0 X1
        X0 60 16
        X1  9 63
```

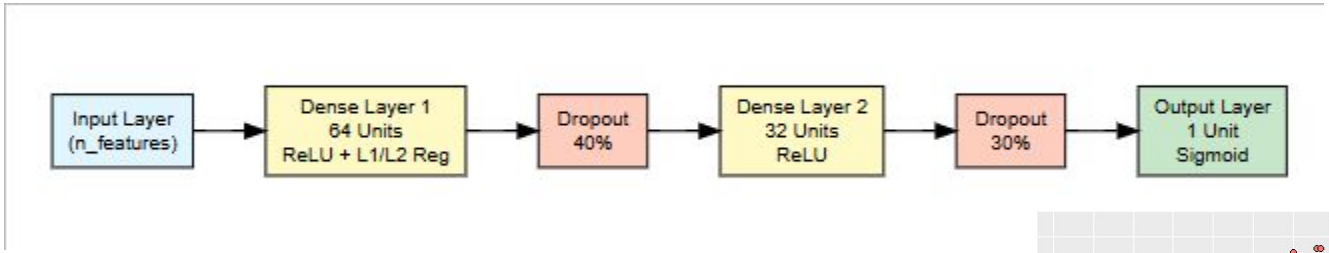
Achieved an accuracy of 0.83 in 8.30 seconds





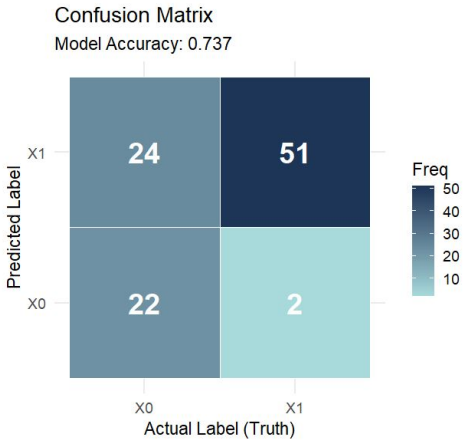
Neural Network

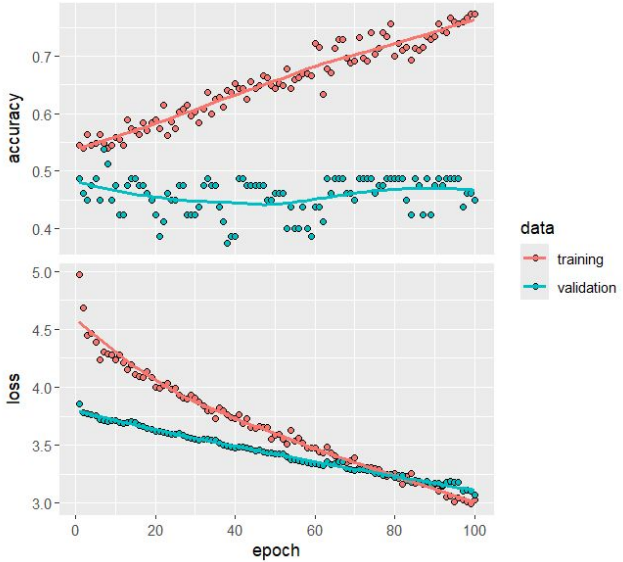
Training time was 5.22 seconds



Prediction Accuracy:
- Training Set: 80.13%
- Testing Set: 65.66%

RMSE (Root Mean Squared Error):
- Training Set: 0.3988
- Testing Set: 0.4671





How to Compare the 4 Models

	Number of gene selected	Prediction accuracy	Runtime	RMSE
Lasso	56	0.8012	0.2030208 secs	0.3849663
SVM	/	0.883	1.341278 secs	0.3746263
Random Forest	1472	0.8311	8.30 secs	0.3386
Neural Network	/	0.6566	5.22 secs	0.4671

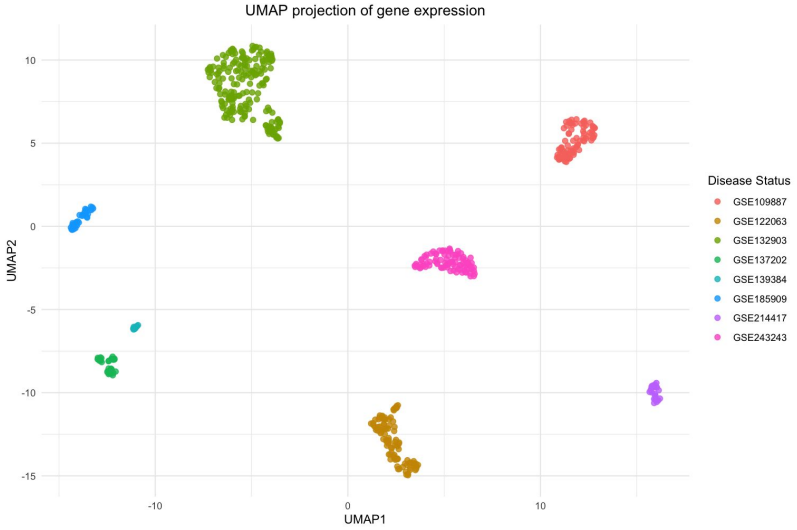
Next steps

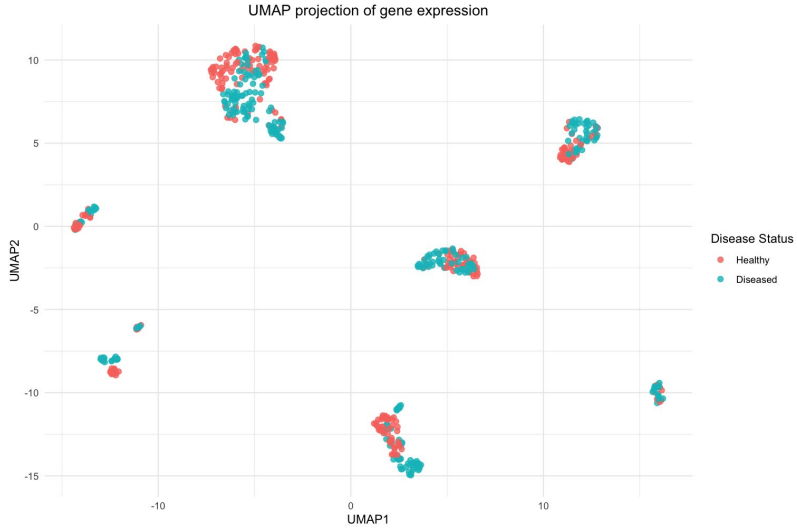
- Cross Validation
- Refined feature selection

Optimizations

- Code optimization:
  - Vectorize operations
  - Avoid repeatedly computing large matrices
  - Avoid to read big data by read.csv (vroom for data loading)
- Parallel computing:
  - Speed up operations with matrices (torch)
  - Parallelize cross-validation loops (future.apply)
  - Run different diseases in parallel (separate jobs or cores)

Limitations





- UMAP colored by cohort reveals strong batch effect, whereas disease labels show no clear grouping.
- Model performance may reflect **batch-related information** rather than **disease-related biology**.
- Requires batch correction or cohort-aware modeling before prediction.