Summer Project 2020

(Duration : Jan 13 - Feb 14)

Prediction of Glioblastoma from mRNA profiles of TEPs using transfer learning approach

- Abhishek Vijayan
  (Under guidance of Dr Fatemeh Vafaee)

Project Description

- Idea : Classify GBM(brain cancer) Vs Healthy, from mRNA profiles of TEPs (Tumour Educated Platelets) using transfer learning approach

- Blood platelets interact with tumour associated biomolecules and obtain a modified mRNA set - such platelets are called Tumour Educated Platelets (TEPs)
- Relevance :
  - Can be obtained from blood-based liquid biopsies
    - Min invasive molecular diagnostics
    - No tissue acquisition

Project Description (cont)

- GBM is rare, limited sample size
- Latest models like Deep Neural Networks require huge amounts of data
- Transfer Learning - apply knowledge learnt from one task to a different task
  - Eg : Medical image recognition task using large set of other common images
    - Learns general features
- In this project, attempt is to use relatively more abundant NSCLC (Non-Small Cell Lung Carcinoma) data to support prediction of GBM

Reference papers

- RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics (2015 Best et al)
  - 3 tasks :
    - Cancer Vs Non-Cancer prediction (96% acc)
    - MultiClass Classification of 6 types of Cancer
    - Determine tumour mutant type
- Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets (2017 Best et al)
  - Used particle swarm optimization to select optimal genes for NSCLC Vs Healthy classification
  - Used SVM model (Late stage : 84% acc, 0.94 AUC; Early Stage : 81% acc, 0.89 AUC)

Data

- Used data provided with previously mentioned papers : 2015, 2017
- Data from 2015 paper
  - mRNA read counts
  - 285 samples, 57736 genes
  - Contains 6 cancer types (one of which is GBM) and healthy control
- Data from 2017 paper
  - mRNA read counts
  - 779 samples, 4722 genes
  - Contains NSCLC and healthy control
- Both papers provide supplementary data on patient age, blood storage time, non-cancer diseases for healthy patients - currently not used

Pipeline

Preprocessing → Predictive Models

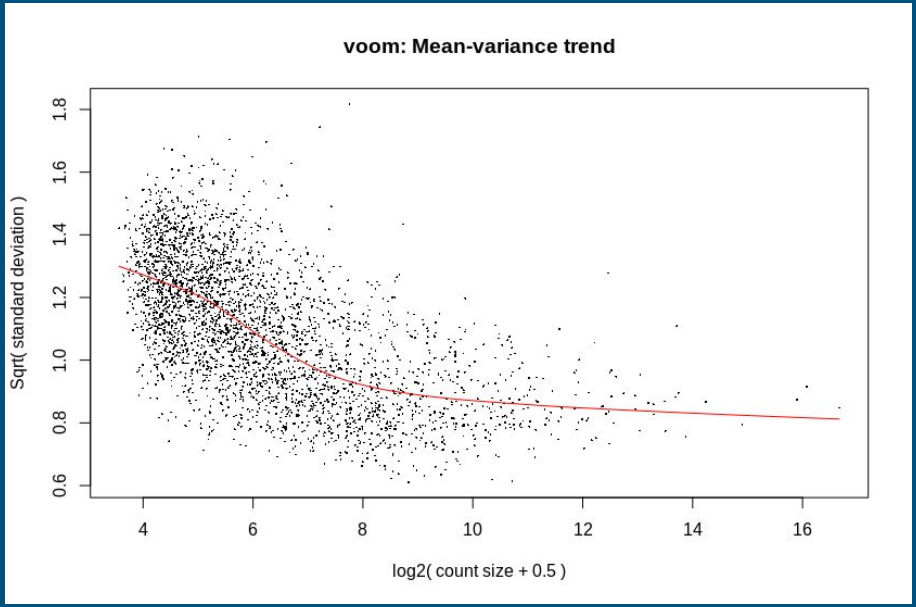
- Data filtering and normalization
- Used R Bioconductor EdgeR package

- Data transformation, model training, model evaluation
- Used python & python packages - pytorch, scikit learn

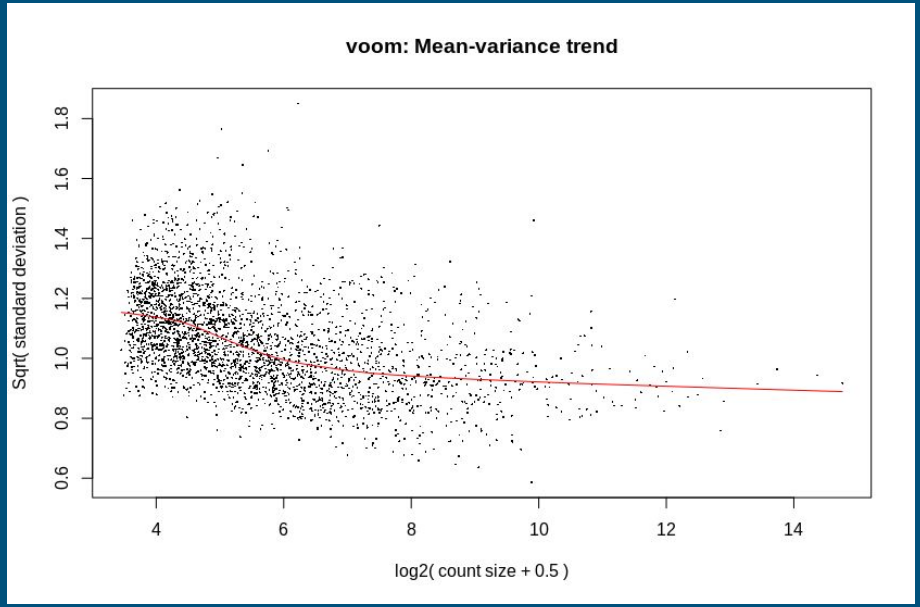
Preprocessing

- Generated 2 sets of filtered data
- Set 1
  - On 2015 dataset, filtered out only GBM and healthy control
  - On both datasets separately applied
    i. filterByExpr
    ii. calcNormFactors of DGEList
    iii. Using 2 groups - Cancer and Non-Cancer, applied voom
- Set 2
  - Same as previous till filterByExpr
  - filter out common genes from 2015 and 2017 data
  - Perform step 2, 3 on this common gene sets

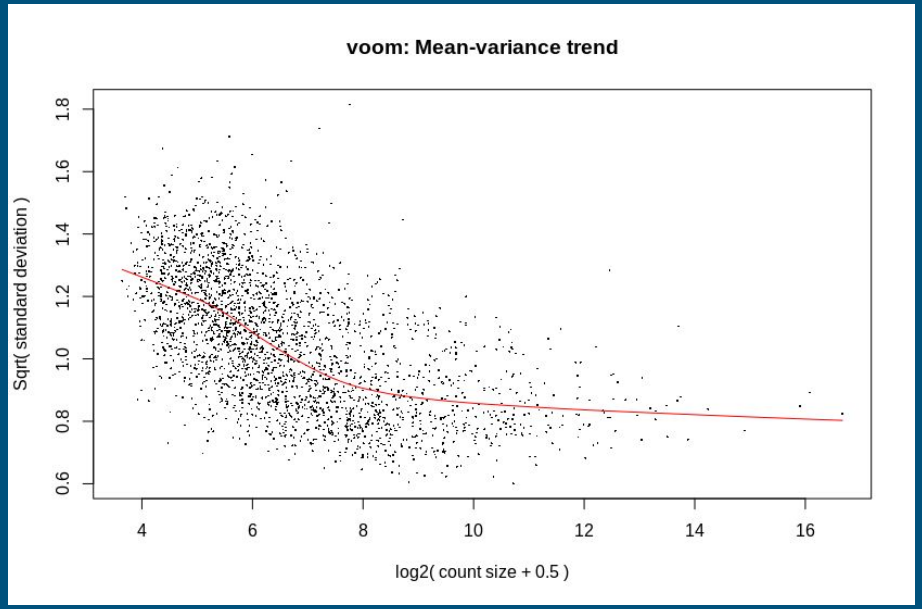
2015 data - all genes



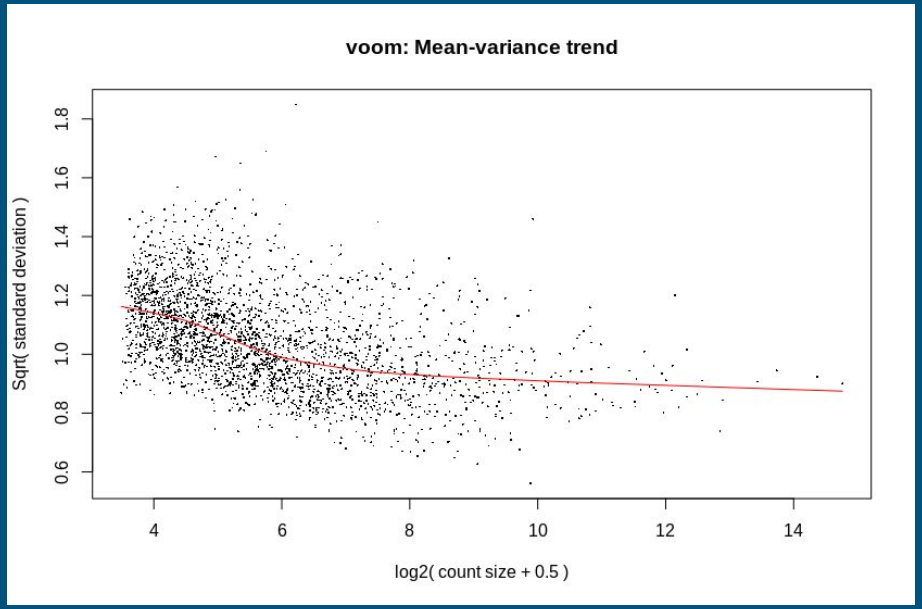
2017 data - all genes



2015 data - only common genes



2017 data - only common genes



Data Size

	Original Data	Set 1	Set 2 (found 2708 common genes)
2015 data (for GBM)	57736 x 285	57736 x 95 (only GBM and HC) 3368 x 95	57736 x 95 (only GBM and HC) 2708 x 95
2017 data (for NSCLC)	4722 x 779	3067 x 779	2708 x 779

	Cancer	Non Cancer
2015 data (for GBM)	40	55
2017 data (for NSCLC)	402	377

Models

- Baseline models
  - SVM
  - Logistic Regression
- 3 layer NN (3368-33-1)
- 5 layer NN (3368-1000-100-10-1)
- Using only common genes : 3 layer NN (2708-27-1)
- Using only common genes : 5 layer NN (2708-1000-100-10-1)

Models (cont)

- 2 basic transfer learning models

  using the common gene set, pretraining on NSCLC data and then using GBM data

  1. 3 layer NN (2708-27-1)
  2. 5 layer NN (2708-1000-100-10-1)

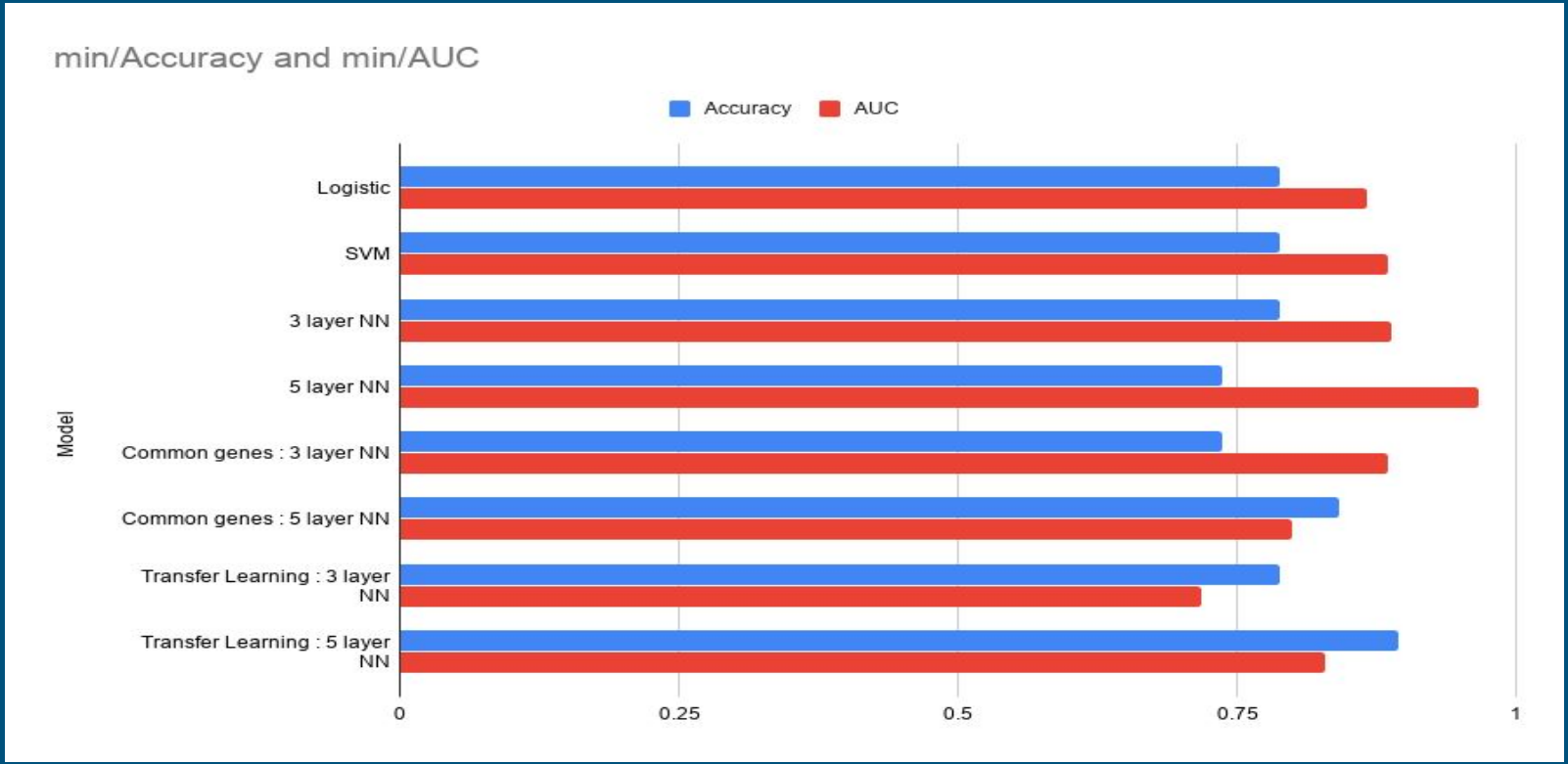
Metrics

- Accuracy and AUC(Area Under ROC Curve : TPR Vs FPR) used as metrics
- 5 different random train and test subsets in 80:20 ratio selected, model run on each
- Reported :
  - Mean accuracy, AUC of 5 test subsets
  - Accuracy, AUC combination corresponding to minimum accuracy + AUC sum

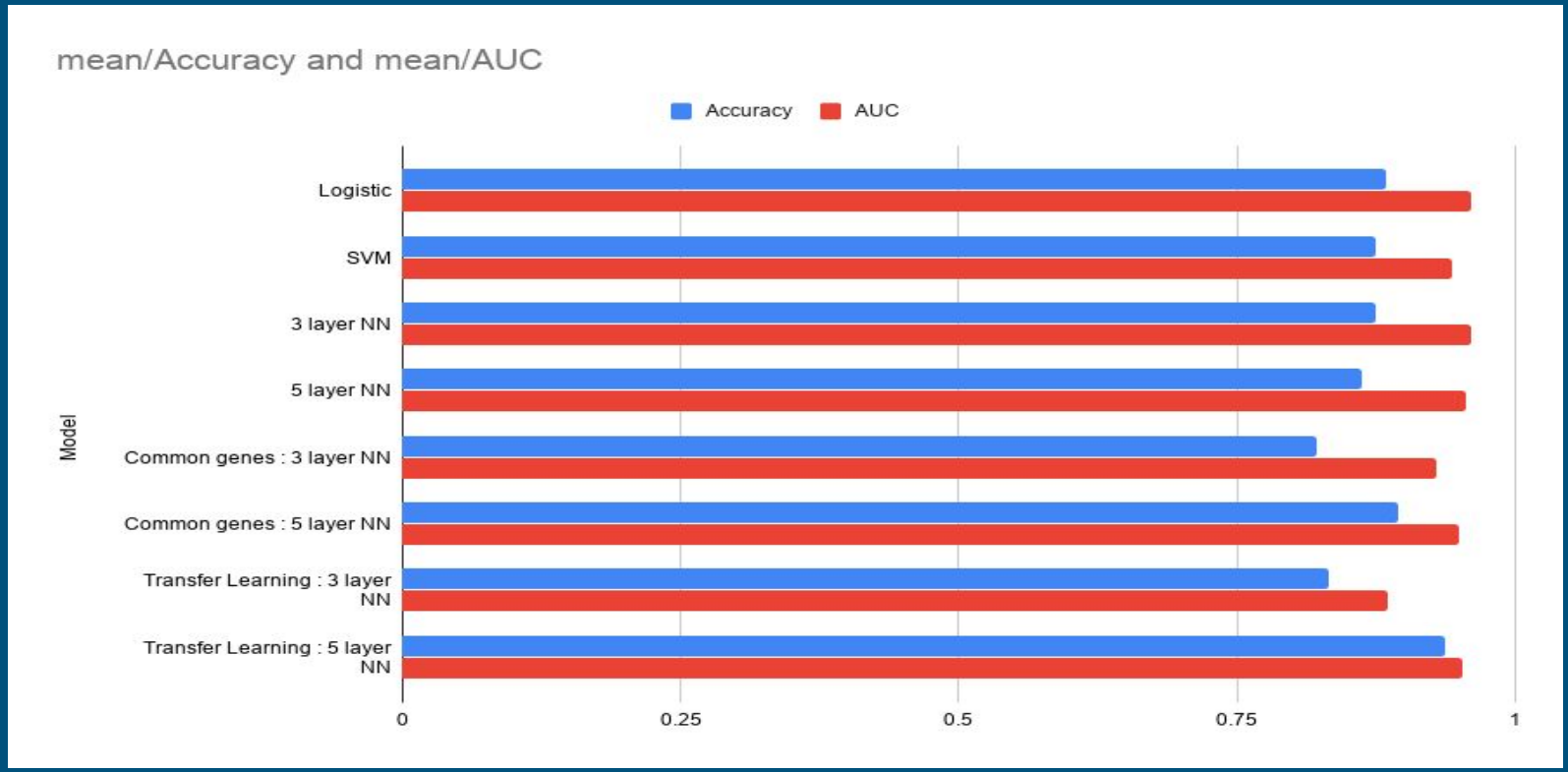
Results

Model	Details	Min Acc	Min AUC	Mean Acc	Mean AUC
Logistic		0.789	0.867	0.884	0.96
SVM		0.789	0.885	0.874	0.943
3 layer NN	3368-33-1	0.789	0.889	0.874	0.961
5 layer NN	3368-1000-100-10-1	0.737	0.966	0.863	0.956
Common genes : 3 layer NN	2708-27-1	0.737	0.885	0.821	0.93
Common genes : 5 layer NN	2708-1000-100-10-1	0.842	0.8	0.895	0.95
Transfer Learning : 3 layer NN	2708-27-1	0.789	0.718	0.832	0.885
Transfer Learning : 5 layer NN	2708-1000-100-10-1	0.895	0.829	0.937	0.952

Results trend - min metrics



Results trend - mean metrics



Why not CNN ?

- Usually used with intention of lower computation compared to fully connected network
- Used commonly in image recognition
  - Identify useful features relevant in multiple parts of the image
- Could not find such structure when features are genes
- To be further analyzed

Further Steps

- Current pipeline a very basic one - improvement required in all steps
- Better filtering and normalization
  - Use the supplementary data provided
  - Comparison among different filtering / normalization methods
- NN hyper parameter tuning
- Try out more complex NN models for prediction
  - Include dropout, regularization

Further Steps (cont)

- Better approaches for doing transfer learning
  - Currently only pre-training has been done, try out fine tuning i.e. training only the last few layers
  - To get equal num of features for NSCLC and GBM data try
    - PCA
    - UMap
    - Variable input autoencoder
- Try out simulations to get more data

THANK YOU