TurBioHacks Hackathon

Ashna Narain
07/09/2025

Data and Methods

Dataset and Preprocessing

The dataset used in this study is the TCGA-LIHC set from UCSC Xena. For this analysis, I utilised the RNA-Seq data along with the clinical dataset with HBV status annotations.
The RNA-Seq dataset contains 424 samples, and is normalised via $log_2(count + 1)$. From this set, samples with HBV-positive and non-viral cases were selected and HCV samples were ignored.

NONVIRAL	219
HBV	147
HCV	11

To streamline the analysis further and reduce the possibility of noise in the data, the expression set was filtered to only retain the top 2000 most variable genes across samples. These gene expression values were then z-score normalised across samples for subsequent model building.

Model and Analysis

For a comprehensive evaluation of the available data, I built classifiers using four different methods (logistic regression, random forests, SVM (linear, RBF)) for analysing gene expression. For all models, data was split into a stratified 80/20 scheme for comparability across models.
For analysing these classifiers, the following parameters were evaluated for comparisons:

- ROC-AUC
- F1
- Accuracy

Along with these, plots for the ROC curves, the confusion matrices, and calibration curves are reported.

Metrics

Modality : Gene Expression

	LR	**RF**	**SVM (Linear)**	**SVM (RBF)**
ROC-AUC	0.698	0.685	0.707	0.718
F1	0.510	0.235	0.636	0.644
Accuracy	0.653	0.639	0.555	0.708
Split	Stratified 80/20	Stratified 80/20	Stratified 80/20	Stratified 80/20

Top 10 Features (Genes)

Logistic Regression

ENSG ID	Gene Name	Coefficient
ENSG00000204764	RANBP17	0.2726316521291810
ENSG00000174473	GALNTL6	0.229370275833444
ENSG00000143194	MAEL	0.2246317792569840
ENSG00000187957	DNER	0.223271047045332
ENSG00000120659	TNFSF11	0.2171910236729720
ENSG00000058404	CAMK2B	0.2136248184723280
ENSG00000284377	POLR2J3	0.2113861211355630
ENSG00000043355	ZIC2	0.2100730166436450
ENSG00000236816	ANKRD20A7P	0.2072911453327340
ENSG00000227097	RPS28P7	0.2055219566389790

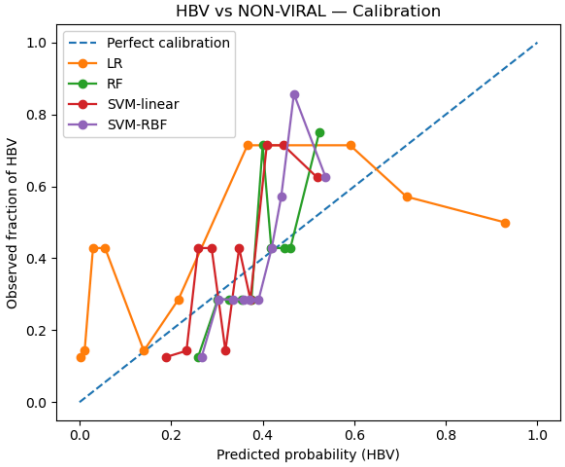
- ***DNER*** : DNER is often overexpressed in HCC cell lines and tissues and promotes HCC proliferation by regulating the activation of PI3K/AKT pathway. It has hence shown promise as a biomarker for HCC prognosis. [Liang et.al, 2018]
- ***TNFSF11/RANKL*** : RANKL has been shown to have a considerable role in the development of bone metastases in several malignant tumors, and RANKL -+ve patients have also been shown to have worse prognoses in HCC. [Sasaki et. al, 2007]
- ***ZIC2*** : ZIC2 has been shown to be upregulated in HBV-associated HCC tissues, and is also linked to poorer prognoses. [Sha et. al, 2021]

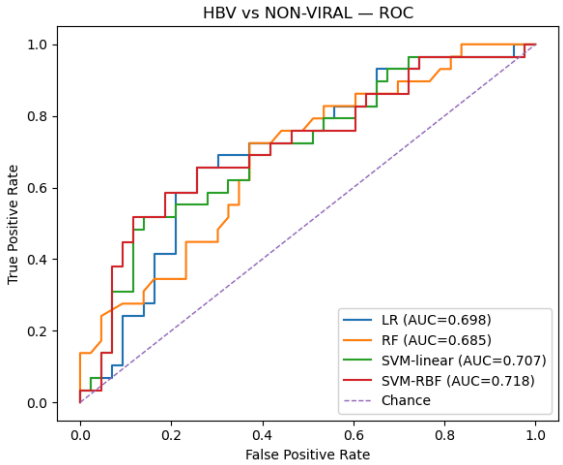
Random Forests

ENSG ID	Gene Name	Importance
ENSG00000144366	GULP1	0.0040431043991236
ENSG00000226251	LINC02608	0.0036336568139726
ENSG00000162391	FAM151A	0.0031108545904213
ENSG00000109181	UGT2B10	0.0030493092643789
ENSG00000282301	CYP3A7-CYP3A51P	0.0028040391324968
ENSG00000043355	ZIC2	0.0027339057282056
ENSG00000158874	APOA2	0.0026410313151625
ENSG00000106128	GHRHR	0.002577011715775
ENSG00000163071	SPATA18	0.0023812274829569
ENSG00000066230	SLC9A3	0.0023181031689141

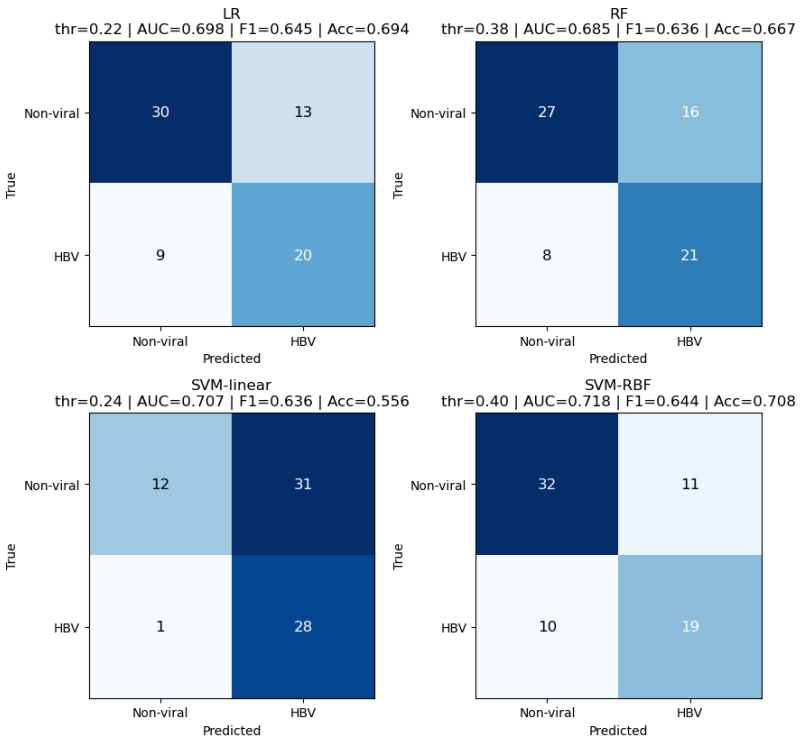
- ***GULP1*** **:** GULP1 has been shown to be a diagnostic biomarker for HCC: GULP1 is often overexpressed in patients with HCC. [Kim et.al, 2025]
- ***APOA2*** : Overexpression of small hepatitis B virus surface antigen (SHBs) has been shown to decrease APOA2 levels in SHBs-expressing hepatoma cells, and conversely suppression of SHBs results in increased APOA2 expression. [Wu et. al, 2024]
- ***SPATA18*** : SPATA18 has been shown to be targeted by HBV-DNA in intrahepatic cholangiocarcinoma (ICC). [Li et. al, 2021]

Plots





Overall, the calibration for Logistic Regression (Brier Score : 0.2576) is better than Random Forests (Brier Score : 0.2208).
In terms of the ROC, SVM-RBF exhibits the best AUC value. All four models perform better than chance.



Note : AI tools have been utilised only for generation of parts of the codes utilised. However, the ideas and approaches presented here are all original and independently formulated by me.