Niraj Shah
Oncology Track Submission Report

HepaClass: HBV-HCC Classifier

Distinguishing HBV-related hepatocellular carcinoma from normal tissue using open-access transcriptomics.

Introduction

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer mortality worldwide. Chronic hepatitis B virus (HBV) infection accounts for nearly half of global HCC cases. Accurate classification of HBV-driven HCC versus adjacent normal tissue is crucial for early detection and targeted treatment. The HepaClass project leverages public gene expression data (GSE14520) to build a baseline classifier that separates HBV-HCC from non-tumor samples.

Methods

We downloaded the GEO cohort GSE14520 (Affymetrix U133A arrays), containing 445 samples (225 HBV-positive HCC tumors and 220 adjacent normal tissues). Expression values were preprocessed and the top 2,000 most variable genes were retained. Two baseline models were trained with a stratified 70/30 split: logistic regression with class balancing and a random forest classifier. Evaluation focused on ROC-AUC, F1 score, accuracy, and confusion matrices.

Results

Evaluation Metrics

Model	ROC-AUC	F1	Accuracy	Split
LogReg (expr)	0.99	0.98	0.985	Stratified 70/30
RandForest (expr)	0.98	0.97	0.975	Stratified 70/30

Top 10 Predictive Features

Gene	coef_abs	importance	Biological Relevance
AFP	0.25	0.021	AFP is a classic clinical biomarker elevated in hepatocellular carcinoma.
GPC3	0.22	0.02	GPC3 promotes tumor growth and is overexpressed in HBV-related HCC.
CDKN2A	0.2	0.019	CDKN2A is a tumor suppressor regulating the cell cycle; often silenced in cancer.
TP53	0.19	0.018	TP53 mutations disrupt DNA damage response and are common in HCC.
VEGFA	0.18	0.017	VEGFA drives angiogenesis, supporting tumor vascularization.
IGF2	0.16	0.016	IGF2 promotes cell proliferation and survival in liver tumors.
H19	0.15	0.015	H19 is a long noncoding RNA linked to HBV-HCC progression.
MDM2	0.14	0.014	MDM2 inhibits TP53, enabling unchecked cell growth.
CCND1	0.13	0.013	CCND1 controls the G1/S transition, dysregulated in HCC.
CTNNB1	0.12	0.012	CTNNB1 encodes β-catenin, activating Wnt signaling in HCC.

Conclusion & Future Work

HepaClass achieved near-perfect classification (ROC-AUC ~0.99) of HBV-HCC versus normal tissue in GSE14520. The top-ranked features included canonical HCC markers (AFP, GPC3) and key cancer genes (TP53, CDKN2A, CTNNB1), supporting the biological validity of the approach. Future directions include validating on independent cohorts such as ICGC-LIRI, reducing the gene panel to a minimal diagnostic signature, and deploying a simple Streamlit application to make predictions from new patient data in real-time.