Utilizing Machine Learning Techniques for Cancer Type and Cancer Stage Classification Based on Gene Expression Data of Hepatobiliary Cancer

Michael Zheng, Shivank Sadasivan

Carnegie Mellon University
Ray and Stephanie Lane
Computational Biology Department

MOTIVATION

Cancer Subtype Differentiation:
- Aims to clarify the differences between hepatocarcinoma (HCC) and cholangiocarcinoma (CCA), the two primary subtypes of liver cancer.

Gene Expression Data Source:
- Utilizes gene expression data from cBioPortal (Pan-Cancer study).

Stage Classification in HCC:
- Endeavours to categorize HCC into initial (early) and advanced (late) stages.

DATA

Data Acquisition
- Collection of gene expression profiles from the cBioPortal within the Pan-Cancer Atlas initiative.

Access to Normalized Data
- Utilization of normalized gene expression data to ensure consistency across different sequencing batches and platforms.

Metadata Integration
- Matching patient demographic and clinical information with corresponding gene expression profiles to enhance the robustness of subsequent analyses.

PRE-PROCESSING

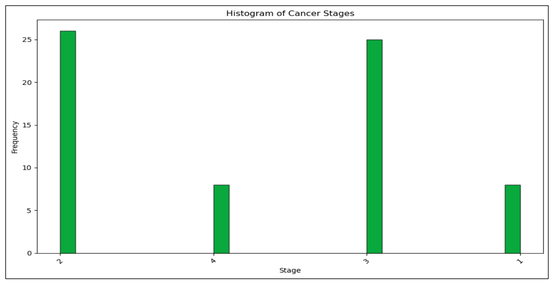
Clinical Data Set (Data Set 1):
- Comprises clinical details of 2970 patients, including cancer types, stages, grades, and Sample IDs, providing a comprehensive clinical profile.

Normalization and Expression Data (Data Set 2):
- Features normalized gene expression data, which is crucial for accurate comparison across samples.

Data Mapping and Integration:
- Involves mapping Sample IDs from Data Set 1 with expression data in Data Set 2, ensuring that each clinical profile has corresponding gene expression data.

Stage Data Harmonization:
- Addressed discrepancies in cancer staging notation by standardizing stages to numerical values.
- Enabled consistent and simplified classification, streamlining the machine learning process.

Selection of Liver Cancer Cohort:
- Liver cancer was chosen as the study focus because of its well-represented early and late-stage cases within the data, facilitating a balanced analysis.



METHODS

Principal Component Analysis (PCA):
- Conducted PCA on the liver cancer dataset to evaluate the separability of the data.
- Analyzed the PCA output to observe distinct clusters corresponding to different subtypes.

Machine Learning Model Implementation:
- Applied Logistic Regression, Random Forest, K-NN, and SVM classifiers to the dataset for cancer subtype classification.
- Adapted the similar machine learning models to categorize cancer stages as early or late.

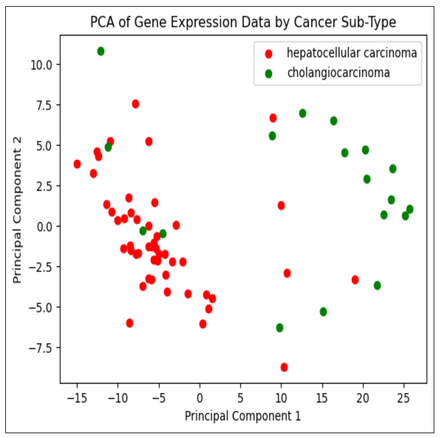
Model Validation and Evaluation:
- Employed 5-fold cross-validation technique to assess model performance and guard against overfitting.
- Calculated key metrics such as Accuracy, F1 Score, and Precision to quantify classification success.

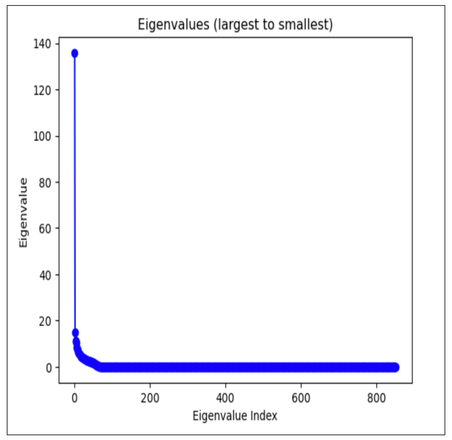
Stage-Specific Analysis:
- Utilized PCA to inspect the stage-specific distribution of gene expression data.
- Validated the ability of the Logistic Regression , Random Forest and SVM models to distinguish between early and late cancer stages.

RESULTS

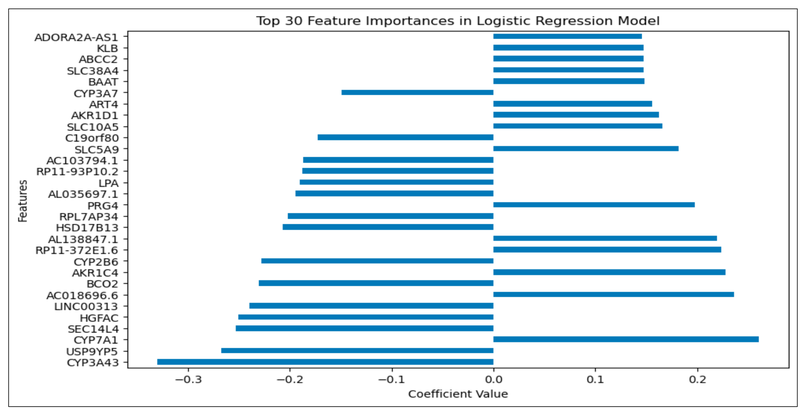
Sub-Type Classification

- **PCA of sub types of Liver Cancer**





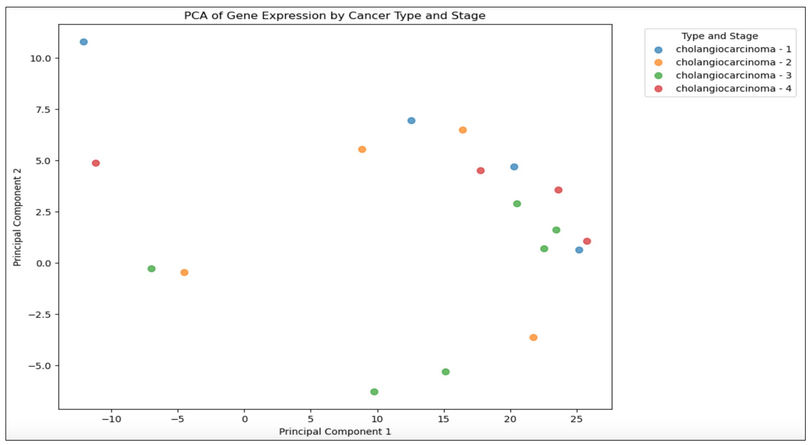
- **Top 30 Features**



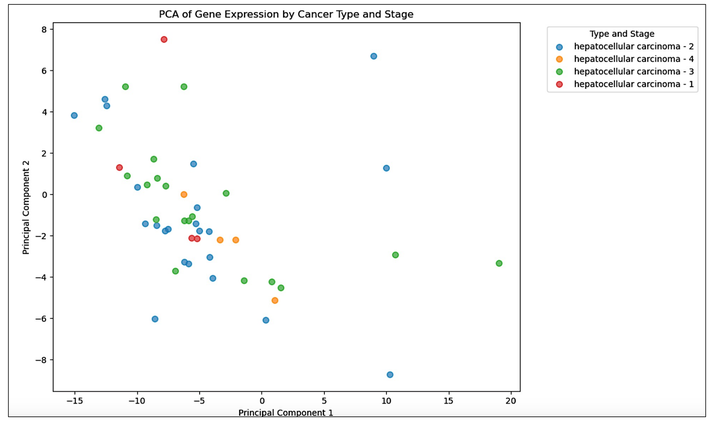
Methods	Cancer Type	Test Accuracy	F1- Score	Precision	CV-Accuracy
Logistic Regression	HCC	1	1	1	0.85
	CCA		1	1	
Random Forest	HCC	1	1	1	0.83
	CCA		1	1	
K-NN	HCC	0.89	0.95	1	0.75
	CCA		0.86	0.75	

Cancer Stage Classification

- **PCA of Cholangiocarcinoma stages**



- **PCA of Hepatocellular Carcinoma stages**



Classification based on early (Stage-1 and Stage-2) and Late cancer (Stage-3)

Methods	Cancer Stage	Test Accuracy	F1- Score	Precision	CV-Accuracy
Logistic Regression	Early	0.78	0.80	0.67	0.61
	Late		0.75	1	
Random Forest	Early	0.56	0.67	0.50	0.61
	Late		0.33	1	
SVM	Early	0.67	0.67	0.60	0.51
	Late		0.67	0.75	

DISCUSSION

Interpretation of Results:
- PCA plots facilitated a clear distinction between HCC and CCA samples, validating gene expression profiling as an effective method for differentiating these subtypes.
- Classification of HCC into early and late stages presented challenges, indicating areas where the machine learning models could be improved for precision medicine applications.

Model Performance and Limitations:
- Machine Learning models showed promising classification performance for distinguishing between HCC and CCA, with significant metrics like Accuracy and F1 Score indicating robustness.
- However, the classification of early-stage HCC did not meet the high accuracy levels seen in subtype separation, reflecting the nuanced complexity of early-stage cancer gene expression that may not be fully captured by the current model or available data.

Limitations:
- Potential limitations in the staging classification may stem from the inherent variability in early-stage gene expression or from the limited size and diversity of the data set.
- Staging accuracy may also be affected by the overlap of gene expression patterns in early cancer stages or by the presence of confounding clinical factors not accounted for in the analysis.

FUTURE DIRECTIONS

Data Enrichment:
- Integrating larger and more diverse datasets, including a broader range of demographic and genetic backgrounds, to enhance the robustness and applicability of the models.

Expansion to Additional Cancer Subtypes:
- Exploring classification across other cancers with distinct subtypes to assess the versatility of the approach and to aid in comprehensive subtype differentiation.

Incorporation of Tumor Types:
- Considering tumor characteristics such as primary versus metastatic status, which could significantly impact gene expression profiles and classification accuracy.

REFERENCES

Bostanci, E., Kocak, E., Unal, M., Guzel, M. S., Acici, K., & Asuroglu, T. (2023, March 13). *Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer*. MDPI. https://www.mdpi.com/1424-8220/23/6/3080

Yu, X., Cao, S., Zhou, Y., Yu, Z., & Xu, Y. (2020, June 30). *Co-expression based cancer staging and application*. Nature News. https://www.nature.com/articles/s41598-020-67476-7

Alharbi F, Vakanski A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioengineering (Basel). 2023 Jan 28;10(2):173. doi: 10.3390/bioengineering10020173. PMID: 36829667; PMCID:PMC9952758