**Translating Multi-Modal Machine-Learning Insights into Clinically Actionable Subtype-Specific Biomarkers with an Emphasis on Biomarker Interactions**

Ying Hu, Chunhua Yan, Qingrong Chen, and Daoud Meerzaman

Center for Biomedical Informatics & Information Technology, CBIIT, NCI, NIH

**Background**  
Accurate classification of tumor subtypes is pivotal for precision oncology, yet existing approaches rarely integrate **multiple** heterogeneous data types or explicitly model gene–gene interactions. Recent advances in machine learning (ML) now enable joint analysis of multi-omics profiles and features to uncover robust, biologically interpretable biomarkers.

**Objective**  
This project will develop an integrated ML framework that:

1. **Predicts tumor subtypes** from combined RNA-seq (transcriptomic) and whole slide images (WSIs) data.
2. **Identifies key genes and gene–gene interactions** that drive distinctions between subtypes.
3. **Elucidates functional roles** of these genes and interactions through gene-set enrichment and network-diffusion analyses.

**Methods**

*Data Harmonization* RNA-seq counts will be normalized and batch-corrected, while WSIs will be pre-processed and summarized into quantitative radiomic features.

*Predictive Modeling* Six complementary classifiers—glmnet, k-nearest neighbors, naïve Bayes, random forest, linear SVM, and XGBoost—will be trained with cross-validation, and their ensemble performance will be evaluated for subtype prediction.

*Feature Selection & Interaction Mining* Feature selection coupled with stability selection will identify candidate subtype-associated genes. The **vivid** R package will detect synergistic gene pairs whose interactions significantly improve classification.

*Functional Interpretation* Candidate genes and gene pairs will undergo gene-set enrichment analysis (GSEA) using Gene Ontology, KEGG, Hallmark, and Reactome sets. Protein–protein-interaction (PPI) network diffusion on Reactome and NeST networks will then reveal higher-order functional modules.

*Network Construction* Significant genes and interactions will be integrated into a directed, interpretable gene network that highlights putative regulatory cascades distinguishing tumor subtypes.

**Expected Outcomes**

* A rigorously benchmarked multimodal ML pipeline for accurate tumor-subtype prediction.
* A ranked list of subtype-defining genes and gene–gene interactions with robust statistical support.
* An interactive gene-network visualization to guide experimental validation and therapeutic-target discovery.