Machine Learning Biomarker Discovery and AI-Based Protein Design for Low-Cost Oral Cancer Screening

Ravidath, Shyam (School: Sunset High School)

Oral squamous cell carcinoma (OSCC) accounts for over 90% of oral cancers, with 5-year survival rates below 60% due to late-stage diagnosis. Current diagnostic approaches rely on invasive surgical biopsy costing $845 per procedure. This project developed an integrated machine learning and AI protein design pipeline for point-of-care OSCC diagnostics, with a secondary therapeutic arm. Proteomics data from CPTAC HNSCC (182 samples, 9,382 proteins) trained five classification models: Logistic Regression, Random Forest, Gradient Boosting, SVM, and ensemble. External validation on PXD025701 (28 samples, 45% protein overlap) achieved AUC = 0.73, rising to 0.79 when evaluated by paired concordance reflecting the dataset's matched tumor/adjacent-normal design. Consensus analysis across four methods identified hemopexin (HPX) as the primary biomarker ($F=380.32$, $p<0.001$). RFdiffusion designed 32 therapeutic binders (86 residues) targeting HPX's heme-binding pocket (best pLDDT 0.809, i_pTM 0.233). RFAntibody generated 24 diagnostic nanobodies (best pLDDT 0.920, i_pAE 7.39 Å). A lateral flow assay was validated at OHSU using biotin-streptavidin-HRP chemistry, demonstrating dose-dependent detection. Clinical modeling projects 89% cost reduction ($7.5M savings per 10,000 screened) versus universal biopsy, achieving 97.4% sensitivity while avoiding 91% unnecessary biopsies. This pipeline integrates computational biomarker discovery, AI-designed proteins, and experimental validation to enable $9 point-of-care screening for high-risk populations.

Awards Won:
Fourth Award of $600