AERO- HNSCC: An Autoencoder-Based Risk Stratification Model for Head and Neck Squamous Cell Carcinoma Individual Report (2024)

Jucheng Hu

# I. INTRODUCTION

This study introduces a novel model, AutoEncoder Risk Stratification for Oncology in Head and Neck Squamous Cell Carcinoma (AERO-HNSCC), which transcends the traditional biomarker identification method by integrating autoencoders (AEs) and multi-omics data for prognostic analysis of Head and Neck Squamous Cell Carcinoma (HNSCC). AERO-HNSCC uniquely utilizes a deep AE for pre-processing multi-dimensional data, including protein expression, RNA-Seq, and clinical information, enhancing the predictive accuracy of patient overall survival (OS). Developed and validated on The Cancer Genome Atlas (TCGA) HNSCC dataset, AERO-HNSCC demonstrates a significant success in risk stratification with an average precision of 73% and a statistically significant p-value of 0.0021 in the log-rank test for survival analysis. These quantitative evaluations demonstrate that the AERO-HNSCC encoded data establishes a robust association with patient OS and can accurately predict OS. Compared with raw multi-omics data for risk stratification and signature-identification-centred (SIC) methods, this new method shows comparable or superior performance. In conclusion, this research contributes a novel computational framework for HNSCC prognosis, its success underlines the potential for AE-based methods to transcend traditional biomarker discovery and to offer broader, more universal solutions in oncological research.

# II. Personal Contribution

## General Contributions

The general contribution to this research project was multifaceted and instrumental in advancing the study towards its objectives. Efforts were primarily focused on four key areas: literature review and analysis, data access and preprocessing, development of a universal testing framework, and comprehensive report drafting.

Literature Review and Analysis: A thorough examination of related works was conducted, meticulously reviewing studies that utilized Lasso and Lasso-Cox SIC methods, as well as various AE methodologies including deep AE, sparse AE, denoising AE, and Variational AE. This investigation allowed for the distillation of critical steps of SIC methods from the literature. The process of normalizing, transforming, and analyzing gene expression data was meticulously identified and summarized, culminating in the identification of specific gene signatures relevant to the research.

Data Accessing, Exploration, and Cleaning: Through the use of R programming, scripts were developed to access the TCGA dataset. Developed python functions focusing on extracting, cleaning, and preprocessing vital data such as protein expression and RNA-Seq information. This task involved overcoming challenges such as handling missing gene data from the signature and resolving ambiguities related to repeated gene names by integrating gene IDs with gene names. Ming Wang has similar contribution on the extract and cleaning of RNA data at this stage, who developed similar programs and merged data to tsv files.

Universal Test Kit Development: The development of a universal testing kit was spearheaded, designed to evaluate the dimensionality reduction performance of AEs. This comprehensive toolkit, equipped with functionalities for result visualization, PCA, K-mean clustering, SVM, and AE architecture design and training, was pivotal in the development and validation of the AERO-HNSCC model. This tool not only facilitated the research on protein side but was also utilized for RNA side by Xi Chen, demonstrating its versatility and applicability.

Paper Write-up: The drafting of key sections of the research paper, including the abstract, introduction, a detailed review of related works encompassing both SIC and AE-based methods, and the entirety of the segments detailing the AERO-HNSCC model's application on protein expression, was undertaken. This encompassed the experimental design, results analysis, discussion of limitations, and conclusions, as well as suggestions for future research directions. The independent effort in articulating the findings and methodologies contributed significantly to the coherence and depth of the final manuscript.

## B. Contributions on The Protein Side

# III. Project Assessment

IV. Team Assessment

## A. Overall Assessment

## B. Individuals Assessment