**Deep Neural Network-Driven Identification and Quantitative Validation of Key Driver Genes in Head and Neck Squamous Cell Carcinoma**

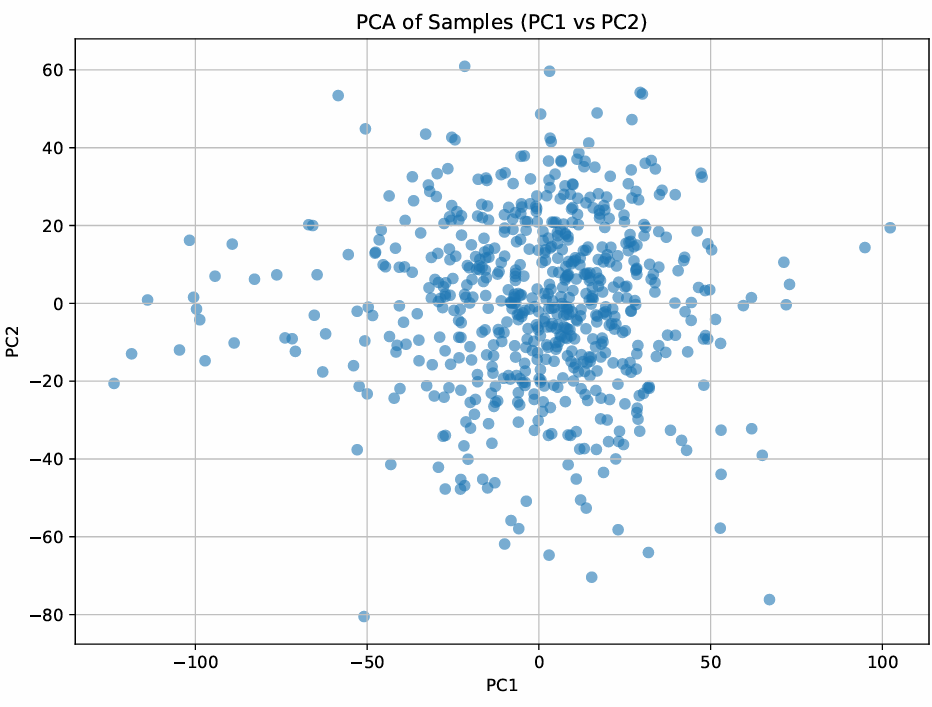
**Method**

1. Input Data
2. PCA
3. VAE Training using Two-fold
4. Runing Integrated Gradients for Each Fold
5. Ensemble Feature

In this step we will cluster latent features from each VAE model, which will help to keep the similar types of features together.

**RESULT**

Step 1

 A graph of a graph

AI-generated content may be incorrect.

**Step 1: PCA Dimensionality Reduction**

**1. Scree Plot (Explained Variance)**

“The scree plot of the top 50 principal components shows that each PC explains only a small fraction of the total variance—roughly 0.15–0.20% per component. There is no single dramatic ‘elbow,’ indicating that variance is distributed diffusely across many dimensions. By retaining the first 50 PCs (which together capture about 10% of the total variance), we reduce noise and dimensionality while preserving a representative sample of the overall structure in the data.”

**2. PC1 vs PC2 Scatter (Sample Distribution)**

“Plotting samples on the first two PCs reveals a fairly continuous cloud without tight clusters. This suggests that head-neck cancer samples do not naturally separate along just PC1 and PC2, but the spread indicates meaningful variation. Downstream models (VAEs) will therefore benefit from these 50 PCs to uncover more subtle, nonlinear structure that isn’t visible in the first two axes alone.”

**Step\_2\_VAE**

A graph with blue and orange bars

AI-generated content may be incorrect.

A graph with many dots

AI-generated content may be incorrect.

In this reconstruction‐error plot, we trained VAEs with latent dimensions of 5, 10, 25, 50, 75, and 100.

* Both **training** (blue) and **validation** (orange) error **steadily decrease** as we increase the latent size, from about **8.6** at 5 dims down to **6.8** (train) and **7.3** (val) at 100 dims.
* The **validation curve closely tracks** the training curve at every point, indicating **no sign of overfitting** even at higher dimensions.
* Notice that **most of the gain happens by 50 dimensions**—beyond that the error reduction **levels off**, so doubling to 100 dims only improves reconstruction marginally.

**Conclusion:** A latent space of **~50** dimensions captures almost all the signal (minimizes reconstruction error) without adding unnecessary complexity—making it our sweet spot for downstream analysis