Accurate Tumor Subtype Detection with Raman Spectroscopy via Variational Autoencoder and Machine Learning

Chang He,[†] Shuo Zhu,[†] Xiaorong Wu,[‡] Jiale Zhou,[‡] Yonghui Chen,[‡] Xiaohua Qian,[†] and Jian Ye,[*, †,§,⊥]

[†]State Key Laboratory of Oncogenes and Related Genes, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, P.R. China

[‡]Department of Urology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, P.R. China

[§]Shanghai Key Laboratory of Gynecologic Oncology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, P.R. China

[⊥]Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, P.R. China

Correspondence:

Jian Ye, School of Biomedical Engineering, Shanghai Jiao Tong University, No. 1954 Huashan Road, Shanghai, 200030, P. R. China, Tel: +86 18930726696, Email: yejian78@sjtu.edu.cn

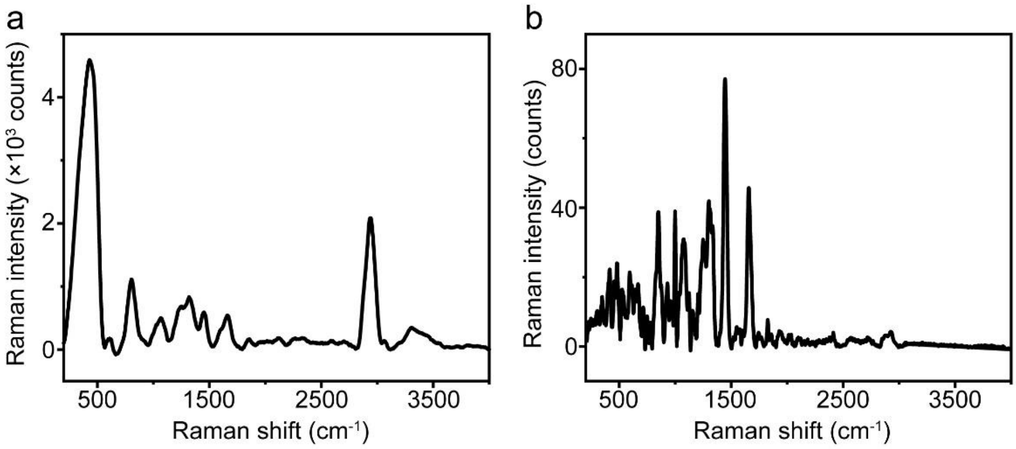

Figure S1. The example Raman spectra of (a) cell and (b) tissue samples in the range of 200-4000cm^{-1}.

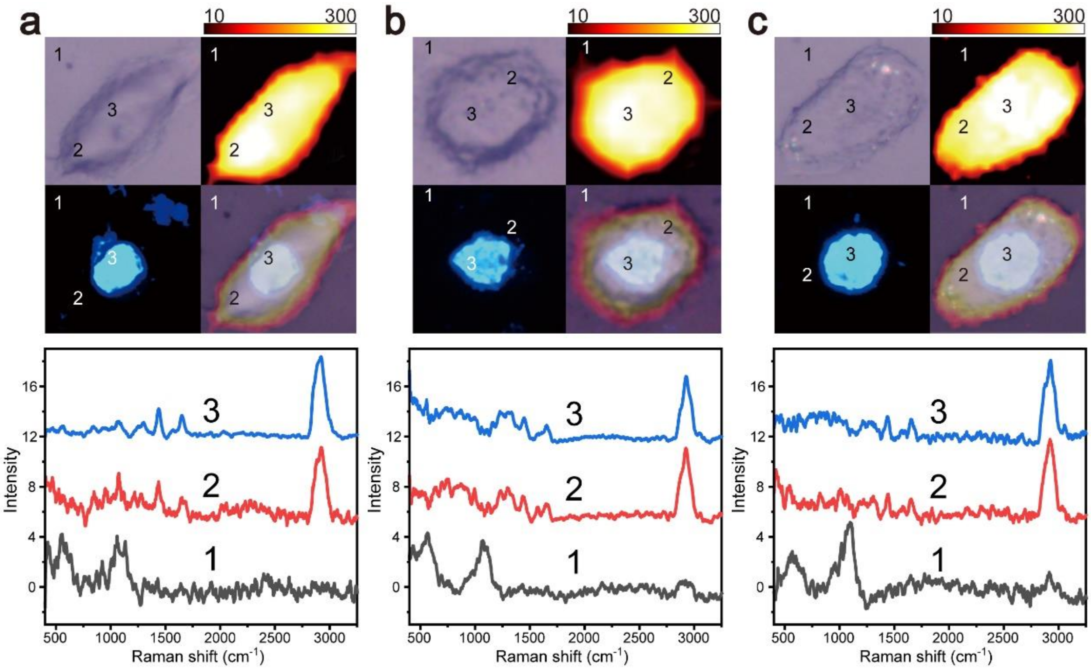

Figure S2. The Raman spectra of different positions of three NSCLC subtypes on quartz sheet (position 1: no cellular component, position 2: cytoplasm, position 3: nucleus). From left to right, the results of A549 (a), H460 (b), and H1299 (c) were shown.

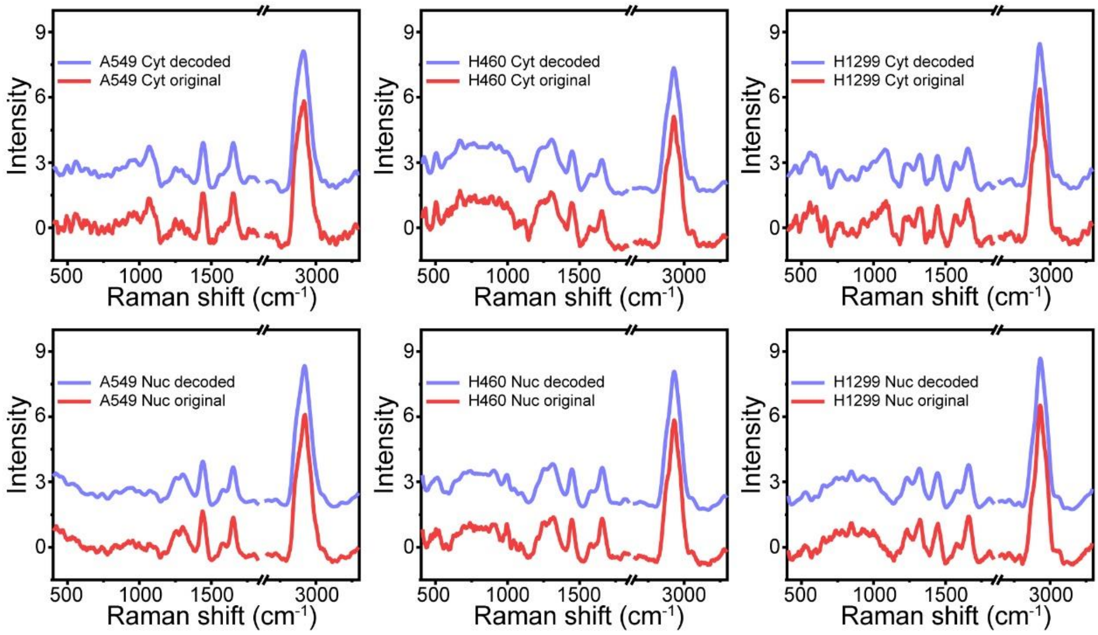

Figure S3. The stochastic Raman spectra (lower, red) and corresponding VAE-decoded spectra from different classes of NSCLC spectra (upper, blue).

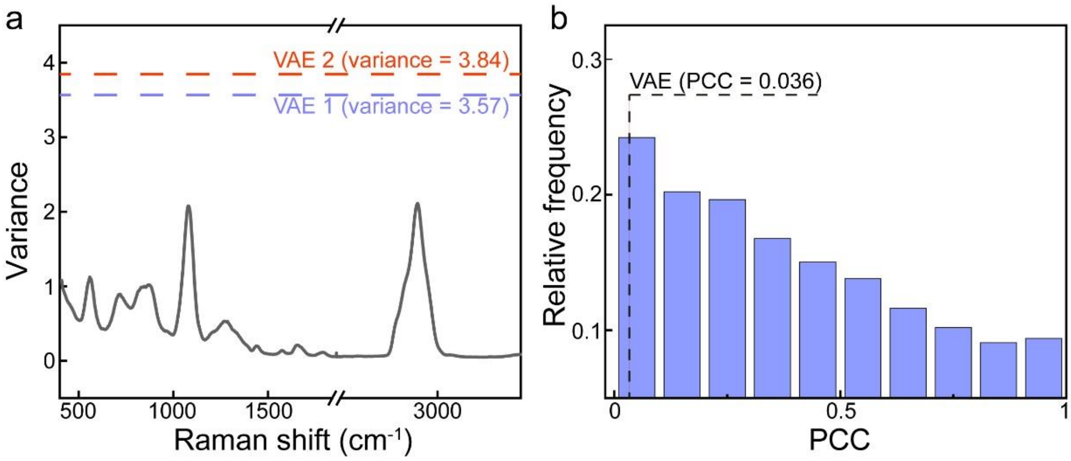

Figure S4. (a) The variance corresponding to each wavenumber in the training set of NSCLC cell samples. The dashed lines represent the variances of the two dimensions of data in the training set after VAE encoded respectively. (b) The histogram illustrates the relative frequencies of the PCC values for a pairwise comparison between all wave numbers of NSCLC cell samples. The dashed line represents the PCC values between the two dimensions of VAE encoded data.

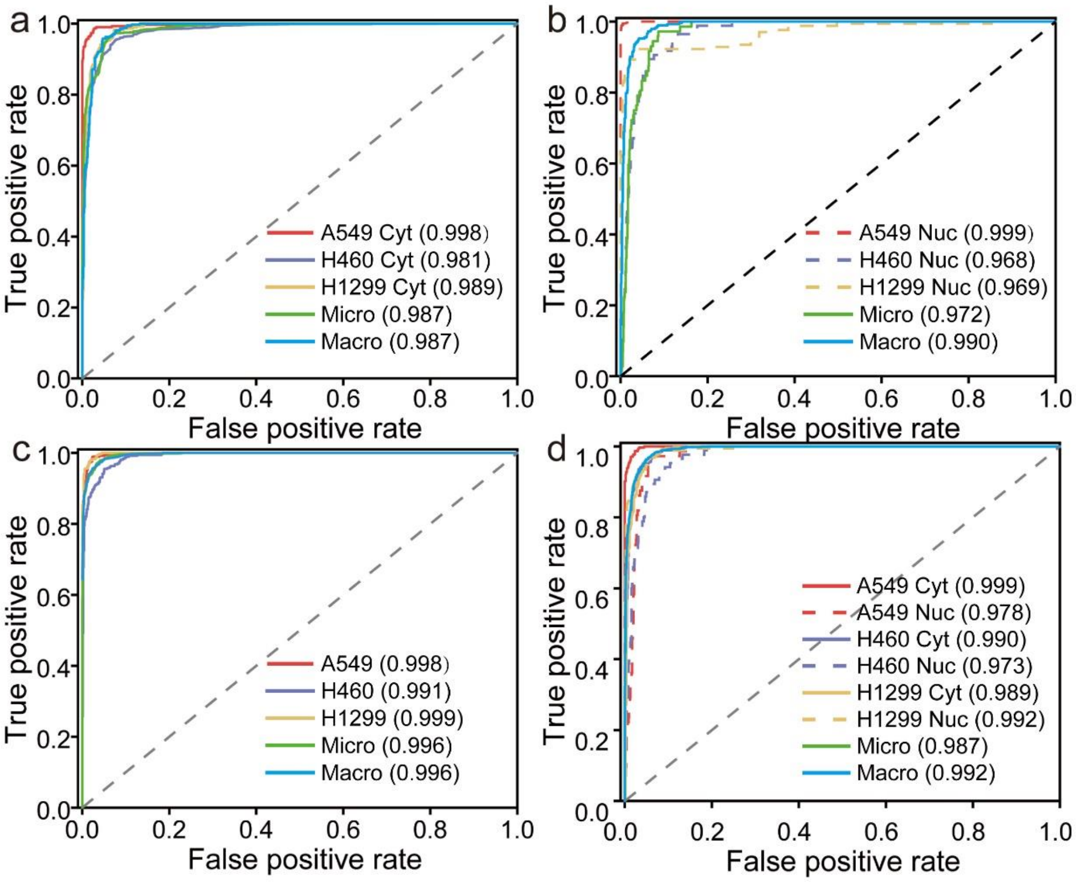

Figure S5. ROC curves and AUC values of four Gaussian Naïve Bayes (Gaussian NB) models based on different types of data sets: (a) cytoplasmic model, (b) nuclear model, (c) cell model, (d) six-classes model. The ROC results of different subtypes of NSCLC cells were shown as the curves (A549: red, H460: violet, H1299: brown, Micro-average: green, Macro-average: blue) and AUC values were shown in the legends. The gray dashed lines represent the ROC curves for the completely random guesses.

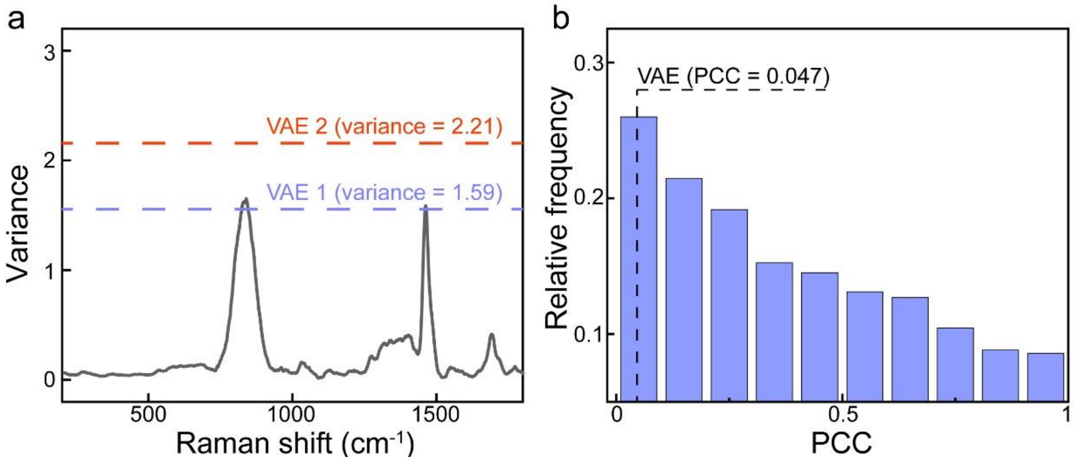

Figure S6. (a) The variance corresponding to each wavenumber in the training set of kidney tissue samples. The dashed lines represent the variances of the two dimensions of data in the training set after VAE encoded respectively. (b) The histogram illustrates the relative frequencies of the PCC values for a pairwise comparison between all wave numbers of kidney tissue samples. The dashed line represents the PCC values between the two dimensions of VAE encoded data.

Table S1. The Gaussian NB result of the distinction between three subtypes of NSCLC.

model	level	LOPOCV (%)	accuracy (%)
nuclear model	spectrum	96.5	96.6
	cell	/	100
cytoplasmic model	spectrum	95.7	95.4
	cell	/	100
cell model	spectrum	97.2	97.4
	cell	/	100
six-classes model	spectrum	89.6	89.6
	cell	/	100

Table S2. The classification result of different dimension reduction methods on six classes of NSCLC cells spectra.

	Gaussian NB		LDA		SVM(Linear)	
	LOPOCV (%)	accuracy (%)	LOPOCV (%)	accuracy (%)	LOPOCV (%)	accuracy (%)
PCA	78.6	76.3	80.9	77.5	79.8	77.9
t-SNE	85.8	84.7	84.5	84.3	86.9	86.1
UMAP	87.2	87.1	88.3	88.2	87.7	86.8
VAE	89.6	89.6	87.9	87.7	88.3	88.6

Table S3. The classification results of nine models on kidney cancer spectra.

model	LOPOCV (%)	accuracy (%)	AUC
RF	78.7	77.5	0.805
SVM(RBF)	78.4	77.0	0.803
MLP	77.2	76.3	0.795
SVM(Linear)	77.3	76.1	0.793
LR	76.8	75.6	0.773
KNN	75.4	72.5	0.761
Ada	74.3	70.0	0.721
Gaussian NB	71.6	69.0	0.694
LDA	69.7	67.8	0.694