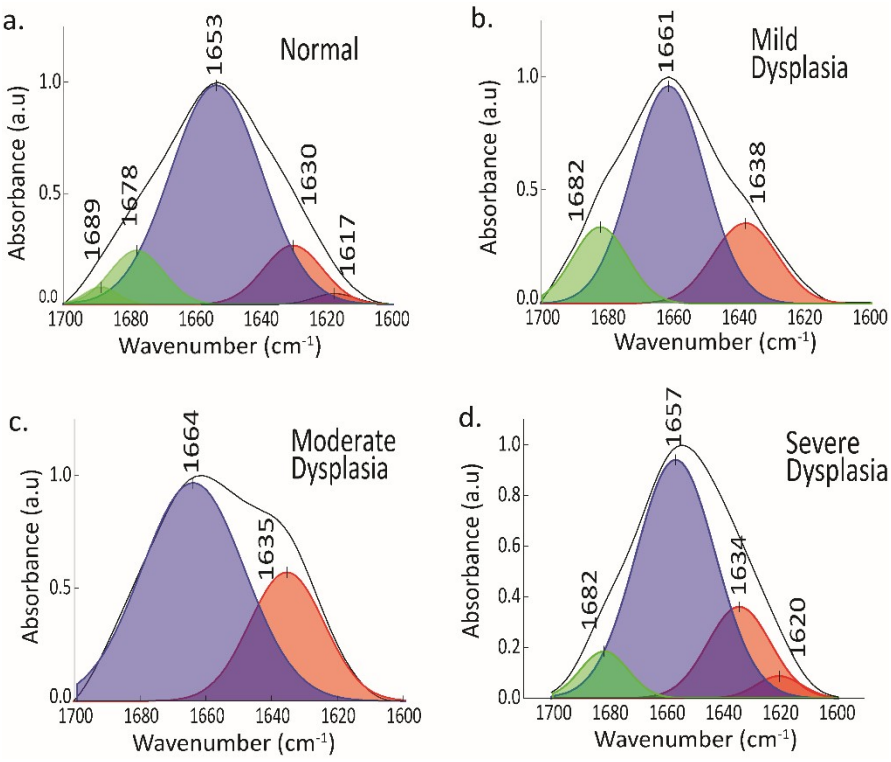
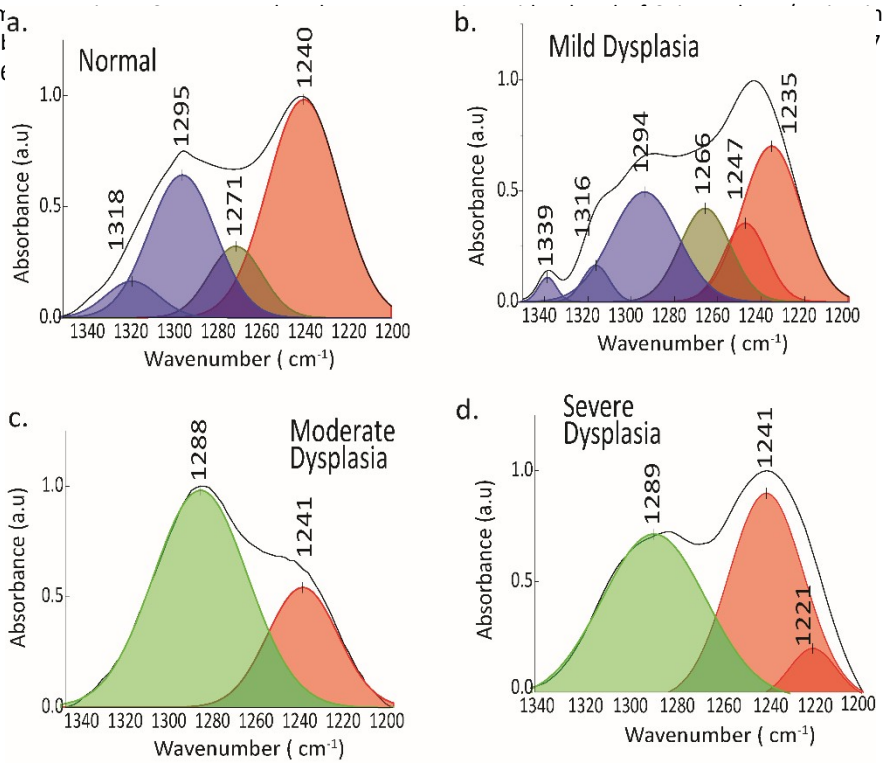
A comprehensive FTIR micro-spectroscopic analysis and classification of precancerous human oral tissue aided by Machine Learning

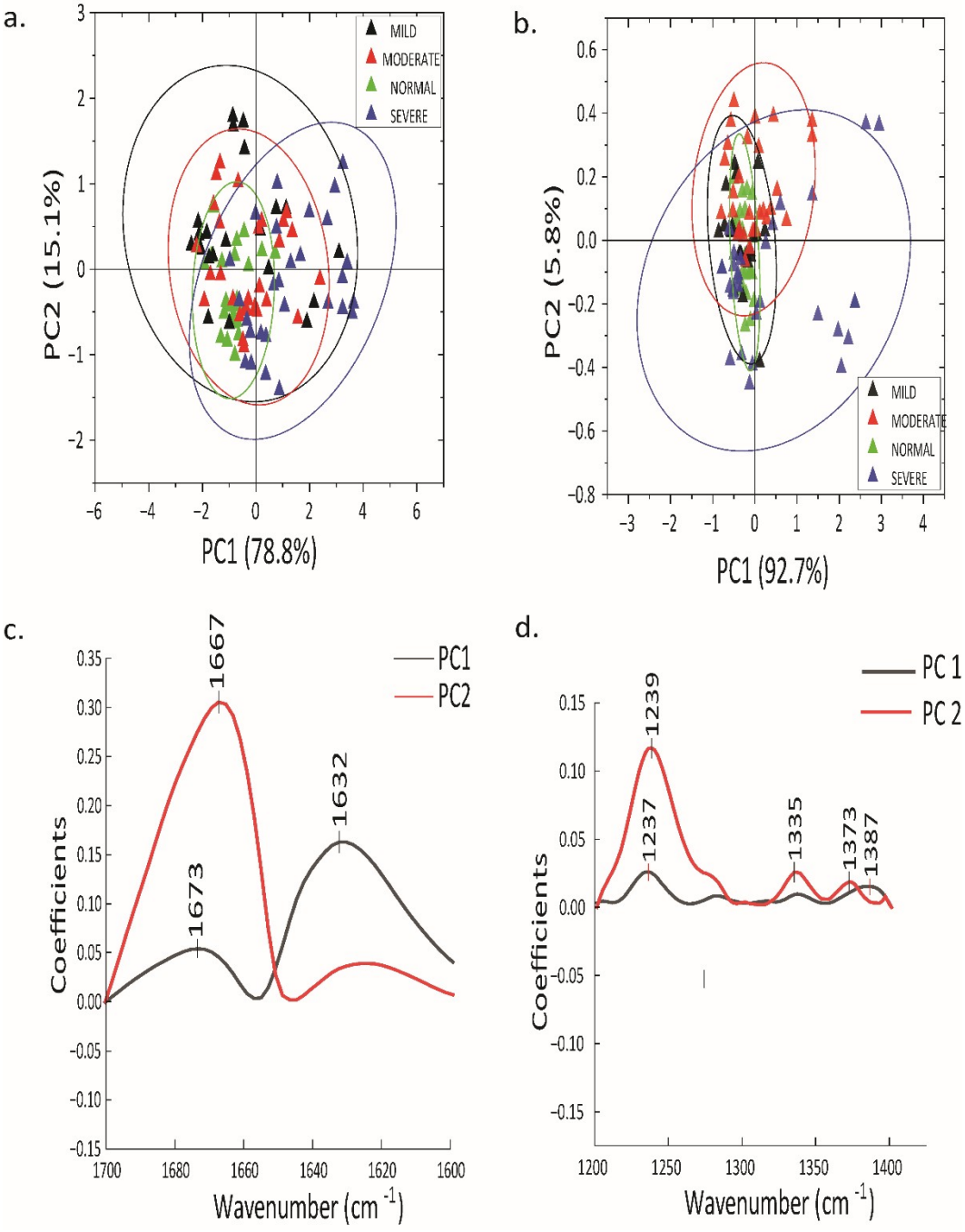
Supplementary Information



Supplen… a) Normal tissue, … cm−1), β-sheet(1…



Supplementary Figure S2: Deconvoluted FTIR spectra in amide III band of spinous layer/region in a) Normal tissue, b) Mild dysplasia, c) Moderate dysplasia and d) Severe dysplasia.1300 ± 5 (α-helix), 1288 ± 2 (β -turn), 1240 ± 2 (β - sheet) and random coil (1270 ± 5)



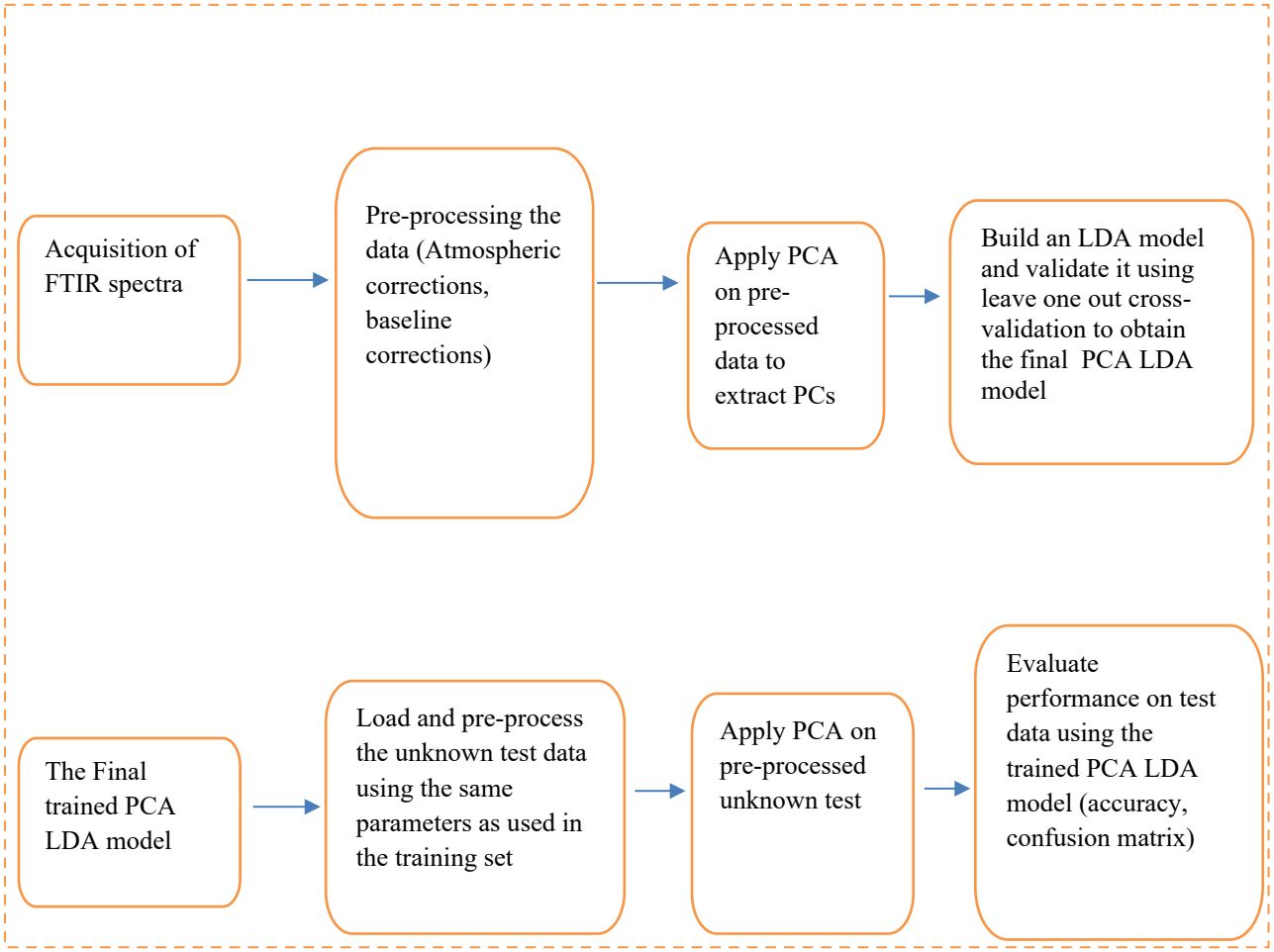
Supplementary Figure S3: a) PCA score plot in Amide I region, b) PCA score plot in amide III region, c) Loading plot in Amide I region and d) Loading Plot in amide III region.

Parameter	Mild dysplasia	Moderate dysplasia	Severe dysplasia
Age (Years)	55±10 21 Male 3 Female	52±5 31 Male 3 Female	53±5 35 Male 2 Female
Habit Smoking Smoking with smokeless tobacco Smokeless tobacco	 17 Male 4 Male 3 Female	 23 Male 8 Male 3 Female	 30 Male 5 Male 2 Female

Predictive Model					
		Normal	**Mild**	**Moderate**	**Severe**
LDA	**Normal**	24 Samples (100%)	0 (0%)	0 (0%)	0 (0%)
	Mild	1 samples (4.55%)	19 samples (86.36%)	1 samples(4.55%)	1 samples (4.55%)
	Moderate	0 (0%)	1 samples (3.4%)	28 samples (93.10%)	0(0%)
	Severe	0 (0%)	0 (0%)	4 samples (13.3%)	26 samples (86.6%)
Cross-validation	**Normal**	24 samples (100%)	0 (4.1%)	0 (0%)	0 (0%)
	Mild	2 samples (9.09%)	17 samples (77.2%)	3 samples (13.6%)	0 samples (4.5%)
	Moderate	0 (0%)	1 samples (3.4%)	28 samples (96.5%)	0 (0%)
	Severe	0 (0%)	1 samples (6.2%)	1 samples (6.2%)	14 samples (87.5%)

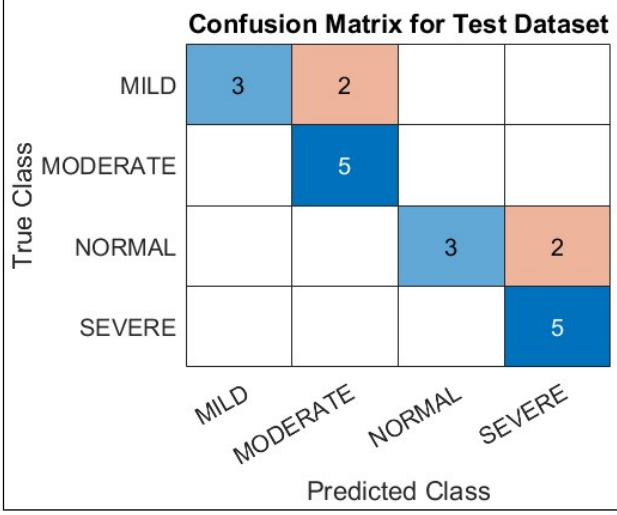
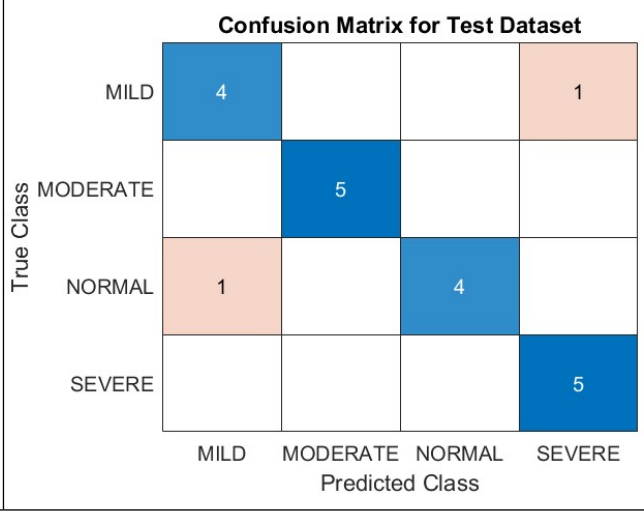
Supplementary Table S1: Clinicopathological characteristics of the study groups.

Supplementary Table S2: PCA-LDA classification in 1700 – 1600 cm^{-1} region



Supplementary Figure S4*:* Training model of the external data set

Initially, a trained PCA LDA model was built and validated using leave-one-out-cross-validation. Then, an unknown dataset was acquired from FTIR analysis of 20 samples which was not involved in the model development (Normal = 5, Mild dysplasia = 5, Moderate dysplasia = 5, and Severe dysplasia =5), and the same pre-processing steps were applied to it as used for the training data. Next, the trained PCA-LDA model was used to classify the test data. In the amide I region the accuracy achieved was 80%, whereas in the amide III region, the accuracy achieved is around 90%.

Confusion matrix in amide I region	Confusion matrix in amide III region
	

Supplementary Figure S5: Confusion Matrix based on the above training model for both Amide I and Amide III regions.