Electronic Supplementary Information

Feature Fusion of Raman Chemical Imaging and Digital Histopathology using Machine Learning for Prostate Cancer Detection

Trevor Doherty,[*a] Susan McKeever,[a] Nebras Al-Attar,[b,g] Tiarnán Murphy,[b] Claudia Aura,[c] Arman Rahman,[c] Amanda O'Neill,[d] Stephen P Finn,[e] Elaine Kay,[f] William M. Gallagher,[c] R. William G. Watson,[d] Aoife Gowen[‡b] & Patrick Jackman[‡a]

[*] *Corresponding author*

[a] *Technological University Dublin, City Campus, Grangegorman Lower, Dublin 7, Ireland.*

[b] *University College Dublin, School of Biosystems and Food Engineering, Belfield, Dublin 4, Ireland.*

[c] *University College Dublin, UCD School of Biomolecular and Biomedical Science, UCD Conway Institute, Belfield, Dublin 4, Ireland.*

[d] *University College Dublin, UCD School of Medicine, UCD Conway Institute, Belfield, Dublin 4, Ireland.*

[e] *Departments of Histopathology and Cancer Molecular Diagnostics, St. James Hospital and Trinity College Dublin.*

[f] *Department of Pathology, Beaumont Hospital, Beaumont Rd, Beaumont, Dublin 9, Ireland.*

[g] *Department of Laser and Optoelectronics Engineering, University of Technology, Baghdad, Iraq.*

[‡] These authors share last authorship.

Supplementary Figures

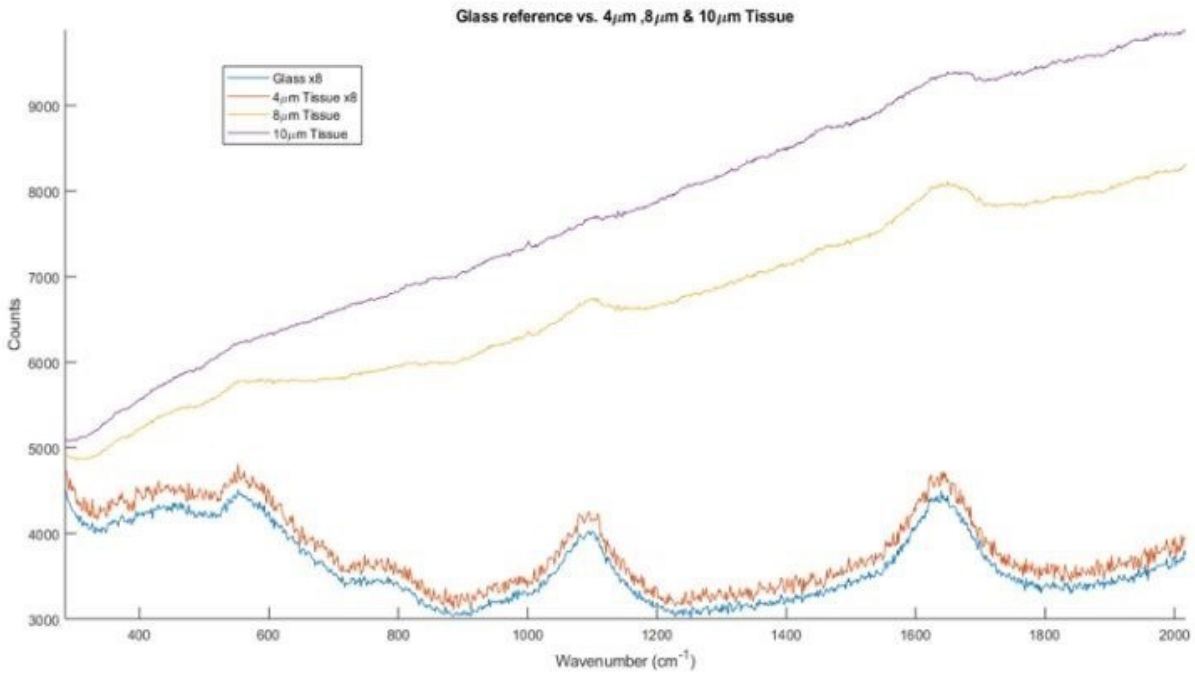

Fig. S1 Plot showing the effects of tissue thickness on background interference from the glass substrate. The spectra corresponding to the glass reference and 4µm tissue section have been multiplied by 8 in order to accommodate differences in scale

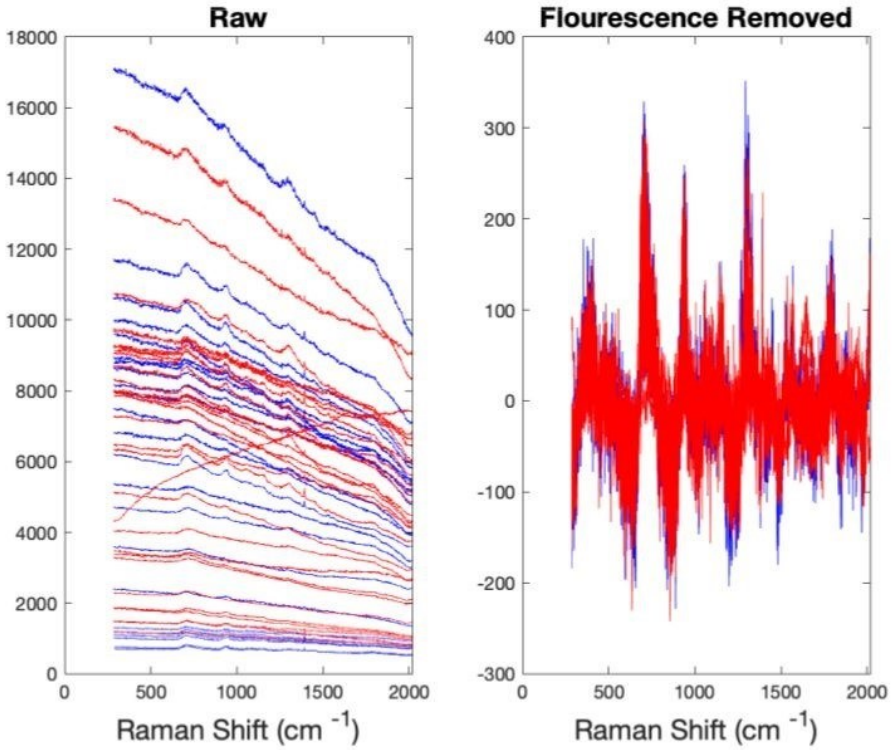

Fig. S2 Median Raman spectra of all samples in the reference set (see Table 4) for non-cancer/cancer

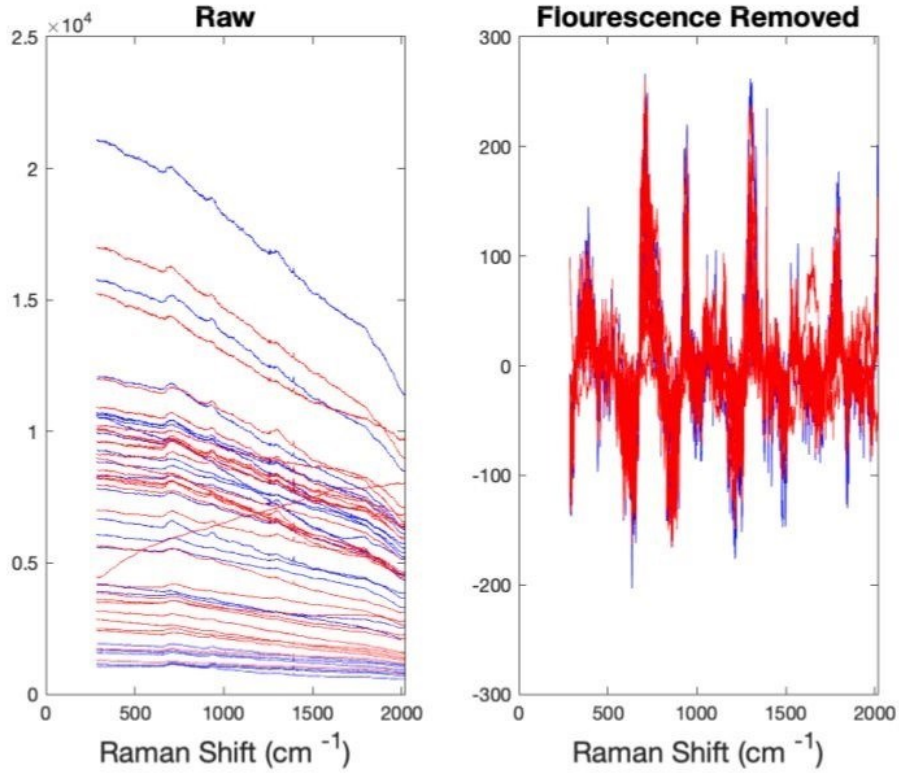

Fig. S3 Mean Raman spectra of all samples in the reference set (see Table 4) for non-cancer/cancer

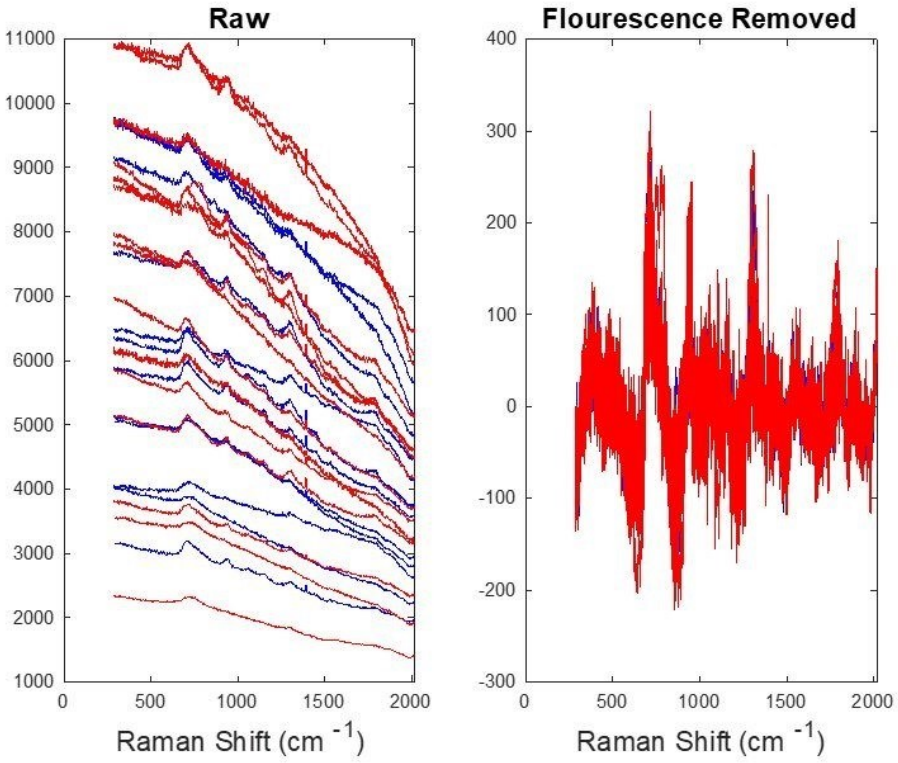

Fig. S4 Median Raman spectra of all samples in the reference set (see Table 4) for Gleason grade 3/grade 4

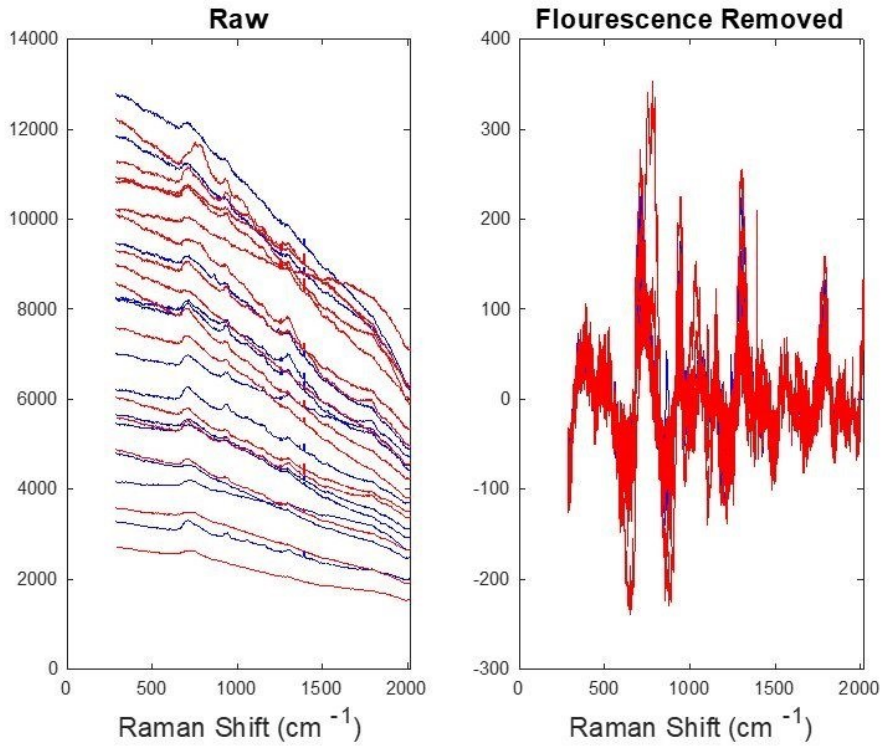

Fig. S5 Mean Raman spectra of all samples in the reference set (see Table 4) for Gleason grade 3/grade 4

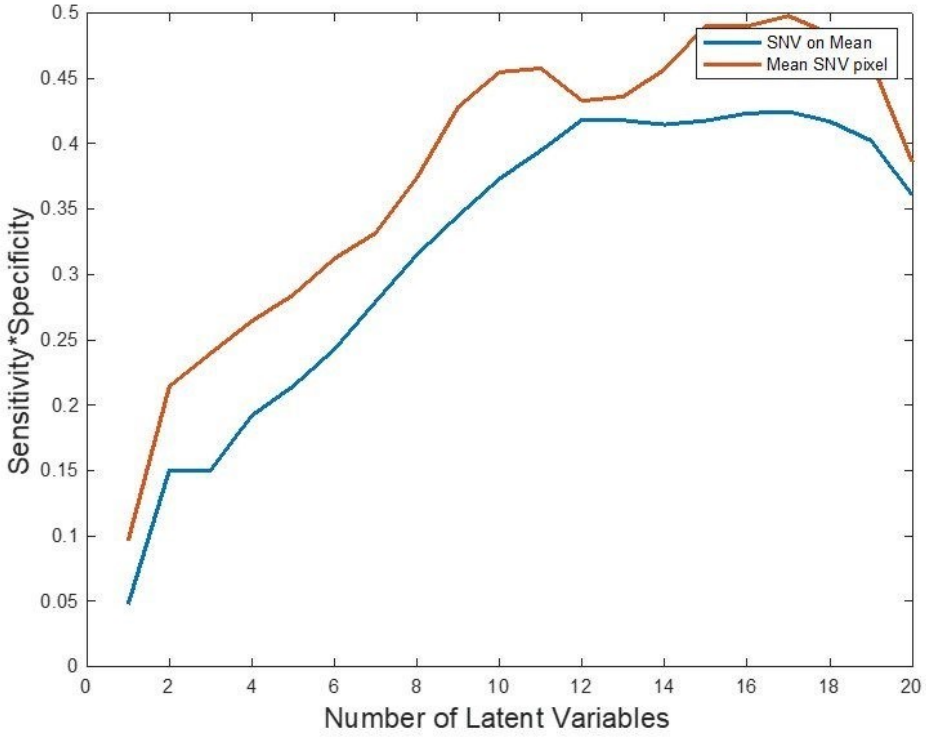

Fig. S6 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate preprocessing (SNV). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

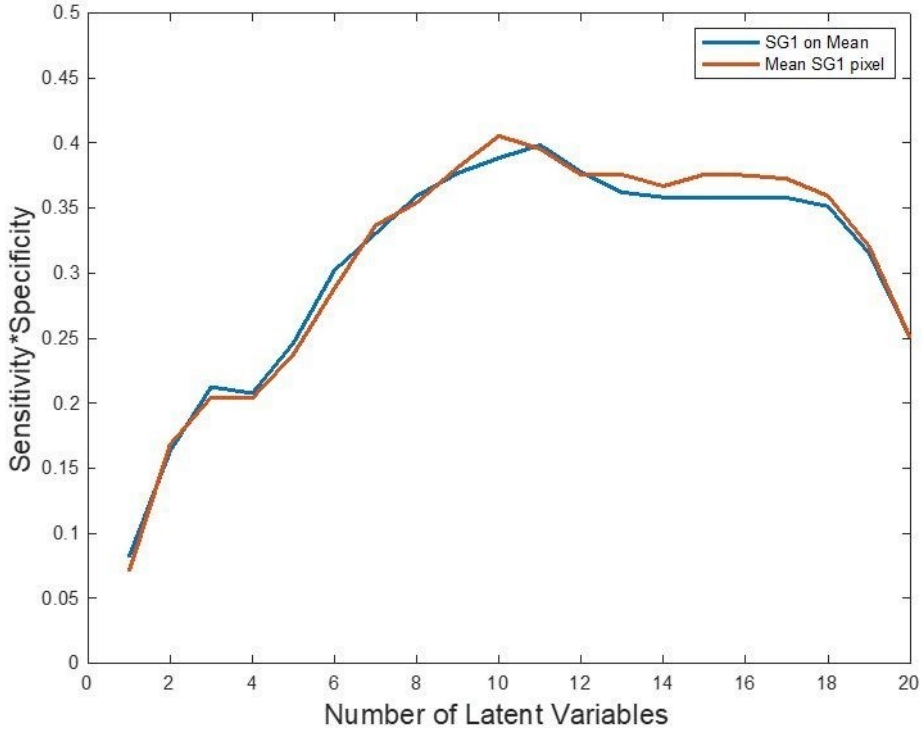

Fig. S7 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using 1st derivative Savitzky Golay pretreatment (SG1). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

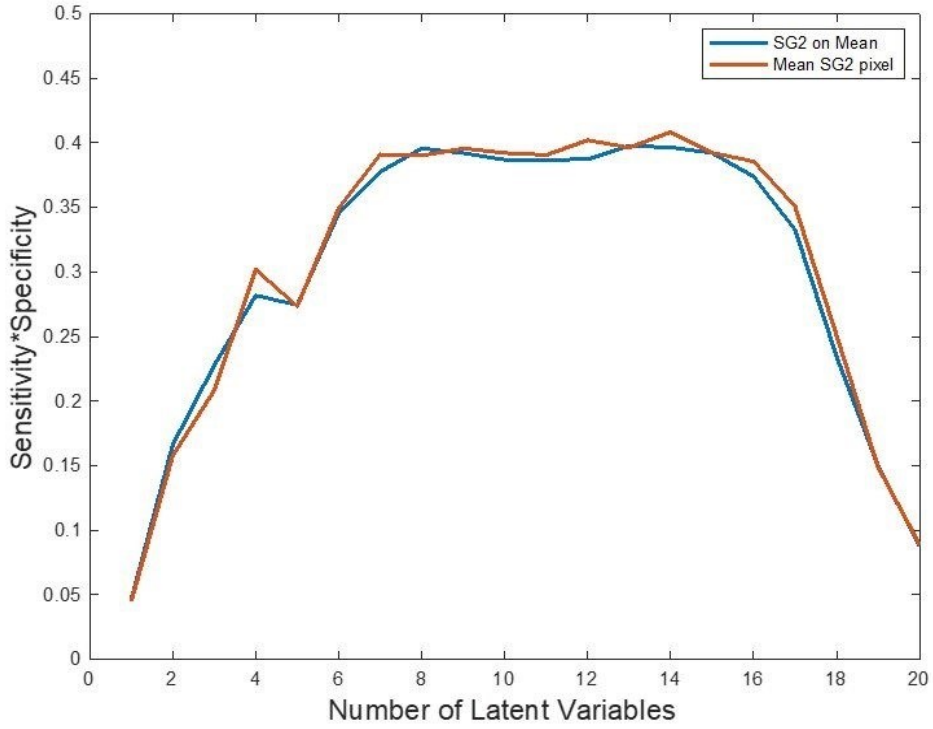

Fig. S8 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using 2nd derivative Savitzky Golay pretreatment (SG2). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

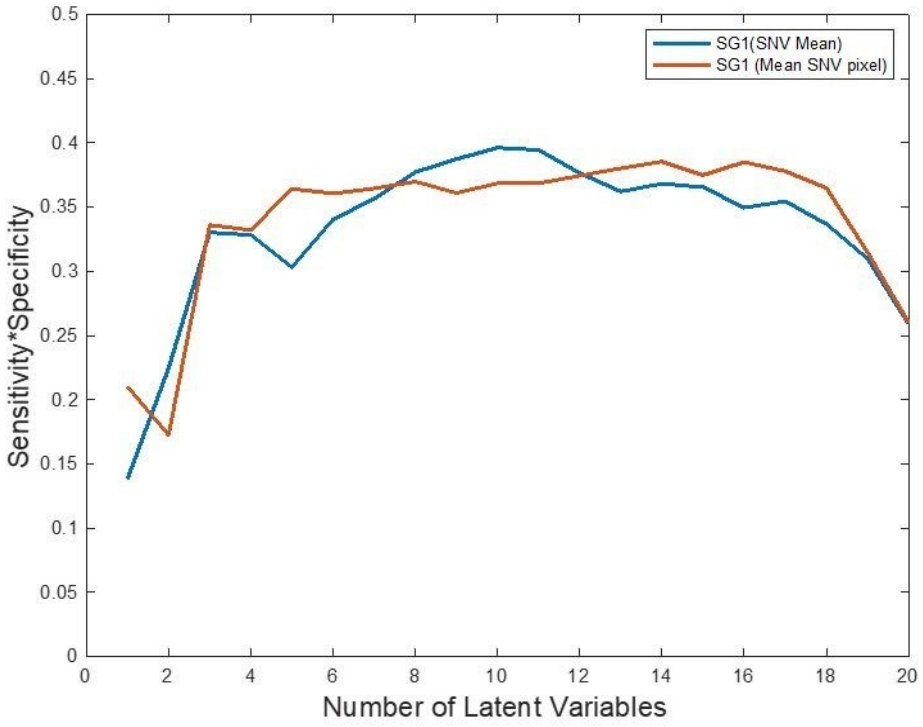

Fig. S9 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate preprocessing (SNV) followed by 1st derivative Savitzky Golay pretreatment (SG1). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

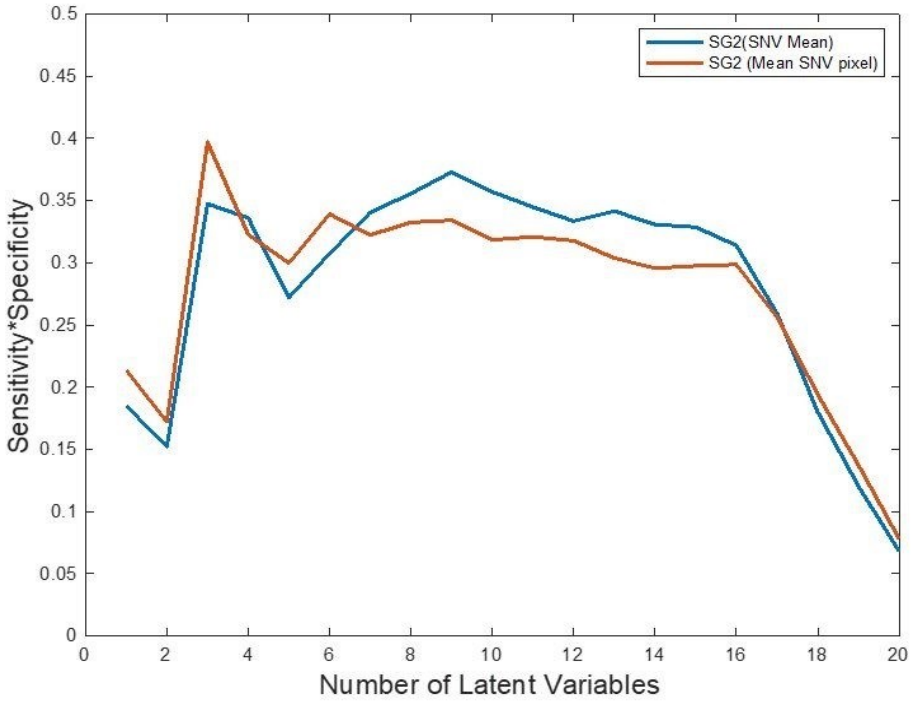

Fig. S10 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate preprocessing (SNV) followed by 2nd derivative Savitzky Golay pretreatment (SG2). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

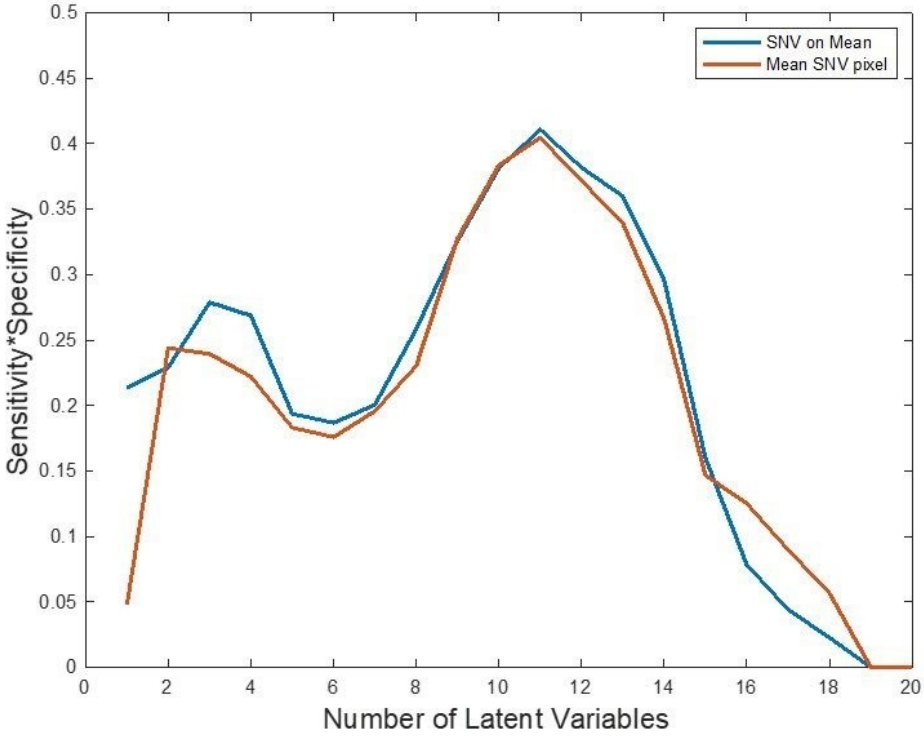

Fig. S11 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade 3/ grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate pre-processing (SNV). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

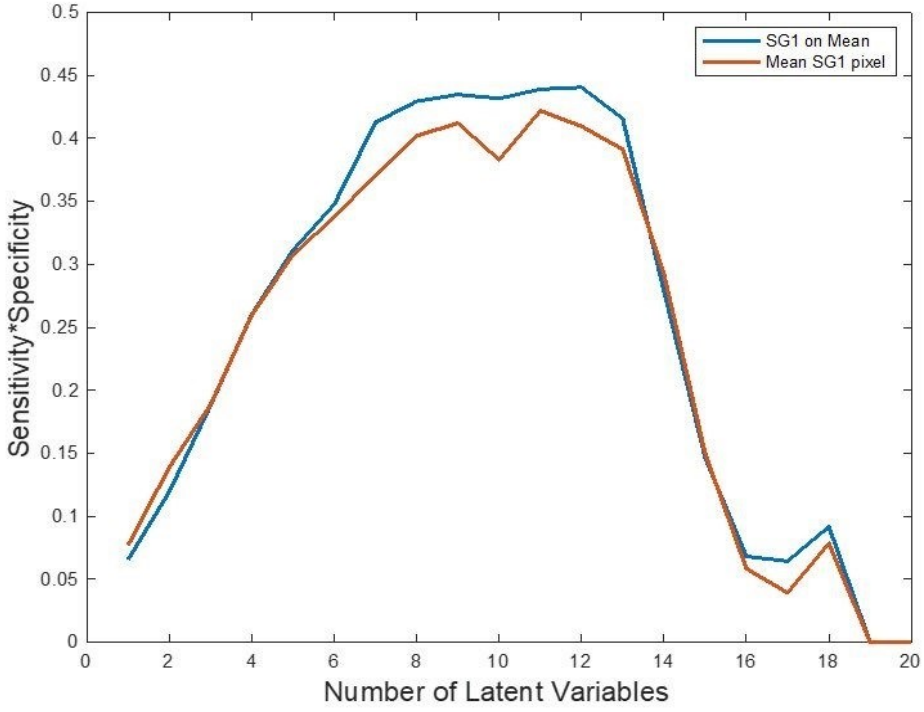

Fig. S12 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade 3/ grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using 1st derivative Savitzky Golay pretreatment (SG1). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

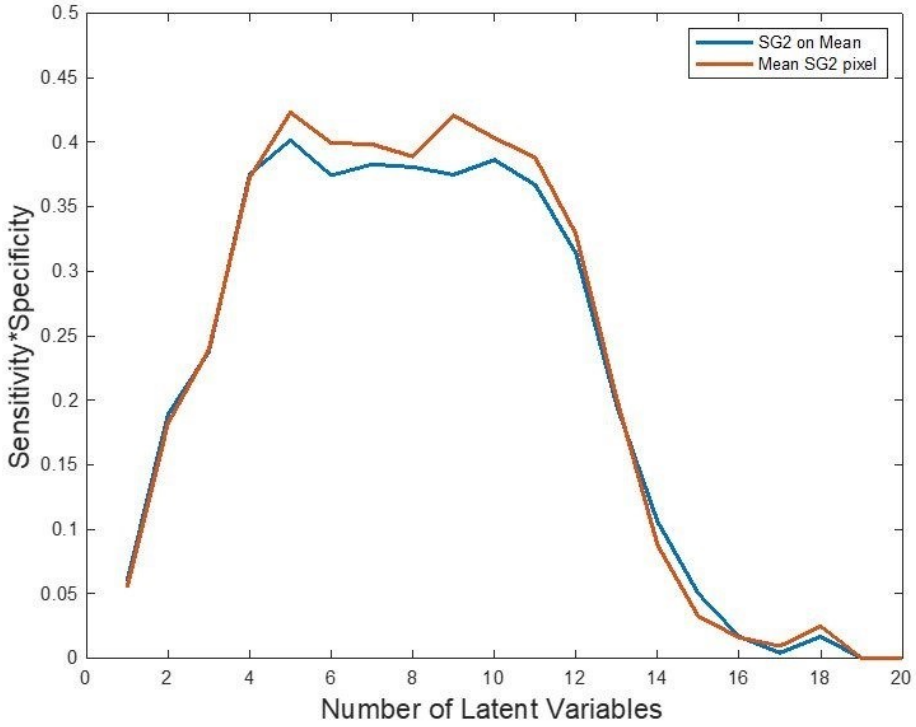

Fig. S13 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade 3/ grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using 2nd derivative Savitzky Golay pretreatment (SG2). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

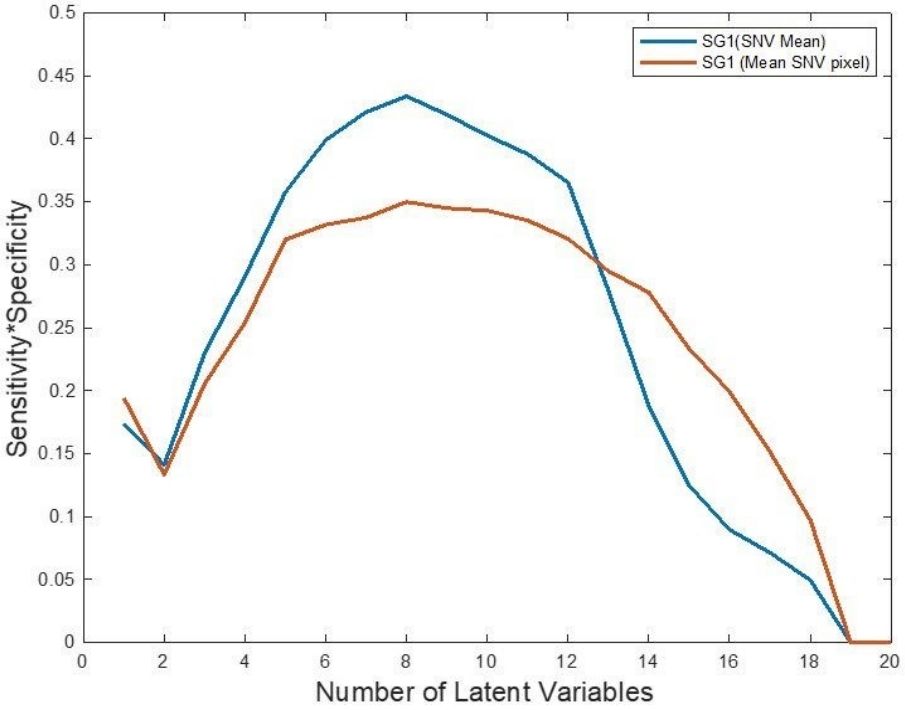

Fig. S14 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade 3/ grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate pre-processing (SNV) followed by 1st derivative Savitzky Golay pretreatment (SG1). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

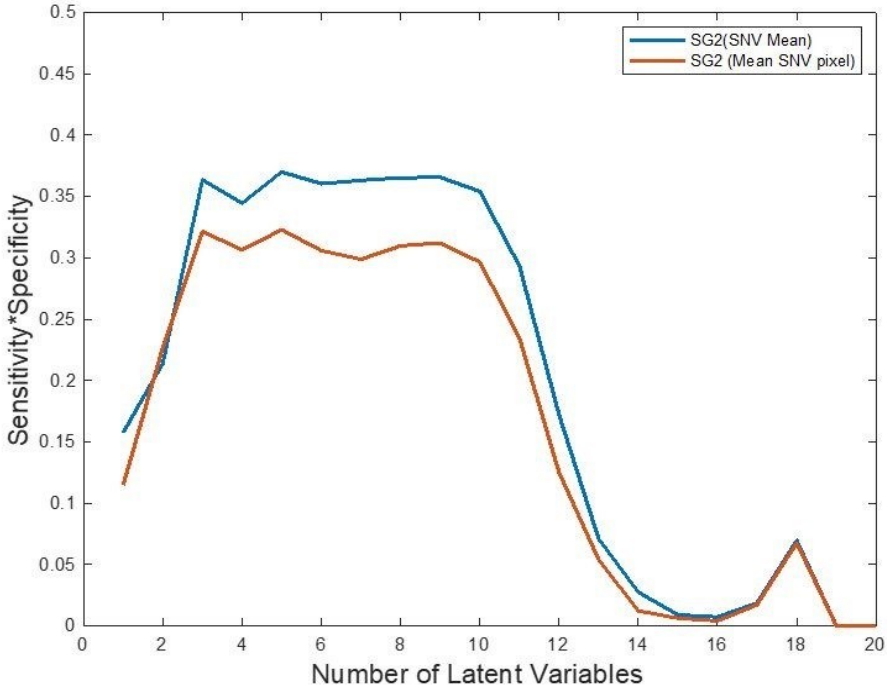

Fig. S15 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade 3/ grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using standard normal variate pre-processing (SNV) followed by 2nd derivative Savitzky Golay pretreatment (SG2). Curves compare the application of pretreatment to individual pixel spectra vs. the mean spectra over all pixels

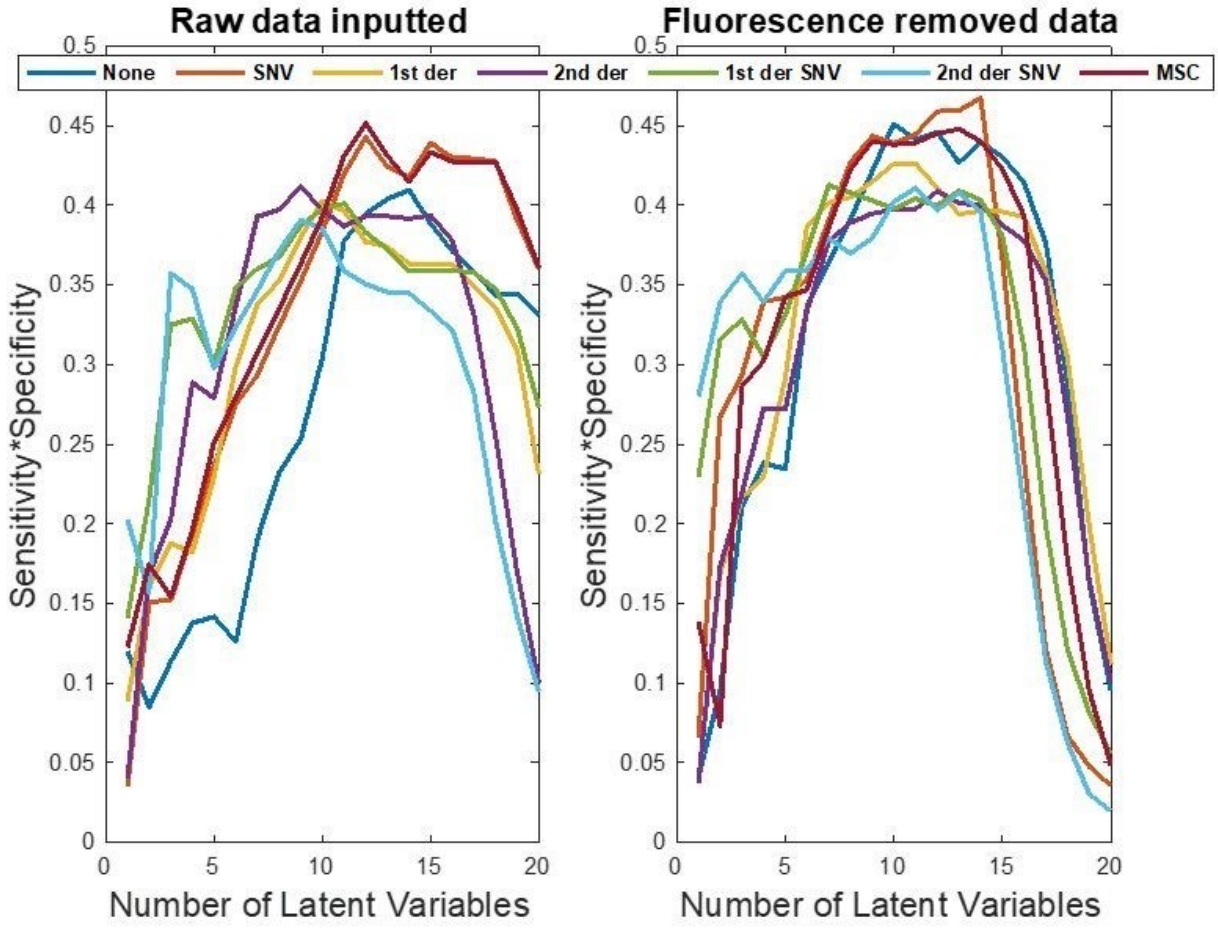

Fig. S16 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using untreated Raman spectra and six spectral pretreatments: standard normal variate pre-processing (SNV), 1st derivative Savitzky Golay pretreatment ('1st der', window size = 15 points, polynomial order = 3), 2nd derivative Savitzky Golay pretreatment ('2nd der', window size = 15 points, polynomial order = 3), combinations of SNV followed by 1st or 2nd derivative pretreatment and multiplicative scatter correction (MSC). The optimal PLS-DA model parameters were selected as fluorescence removal followed by standard normal variate preprocessing and a PLS-DA model with 12 latent variables

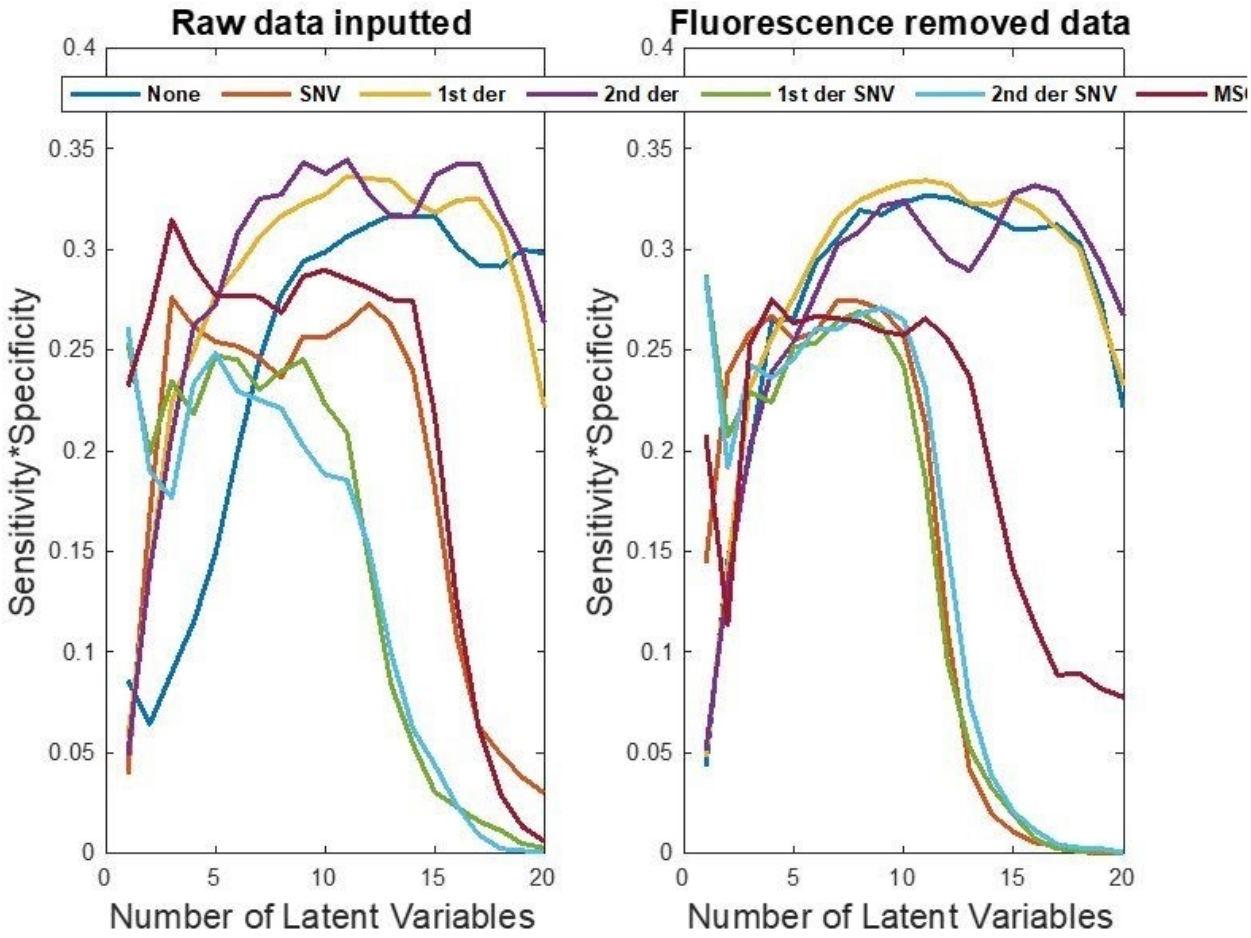

Fig. S17 Product of sensitivity and specificity for PLS-DA models constructed on median Raman spectra for non-cancer/cancer classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using untreated Raman spectra and six spectral pretreatments: standard normal variate pre-processing (SNV), 1st derivative Savitzky Golay pretreatment ('1st der', window size = 15 points, polynomial order = 3), 2nd derivative Savitzky Golay pretreatment ('2nd der', window size = 15 points, polynomial order = 3), combinations of SNV followed by 1st or 2nd derivative pretreatment and multiplicative scatter correction (MSC). The optimal PLS-DA model parameters were selected as raw spectra followed by 2nd derivative Savitzky-Golay preprocessing and a PLS-DA model with 9 latent variables.

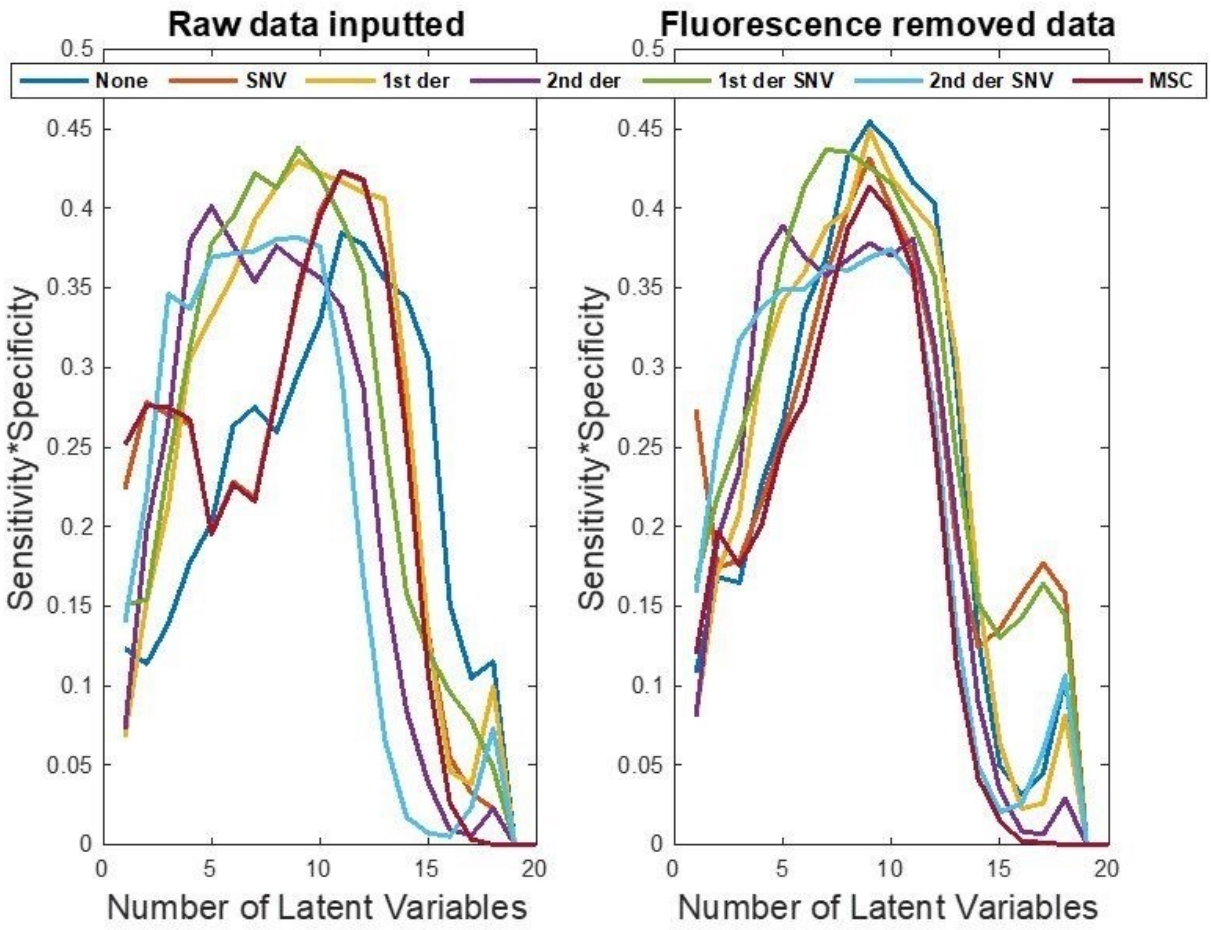

Fig. S18 Product of sensitivity and specificity for PLS-DA models constructed on mean Raman spectra for Gleason grade3/grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using untreated Raman spectra and six spectral pretreatments: standard normal variate pre-processing (SNV), 1st derivative Savitzky Golay pretreatment ('1st der', window size = 15 points, polynomial order = 3), 2nd derivative Savitzky Golay pretreatment ('2nd der', window size = 15 points, polynomial order = 3), combinations of SNV followed by 1st or 2nd derivative pretreatment and multiplicative scatter correction (MSC). The optimal PLS-DA model parameters were selected as fluorescence removal followed by either no pretreatment or 1st derivative Savitzky-Golay preprocessing and a PLS-DA model with 9 latent variables

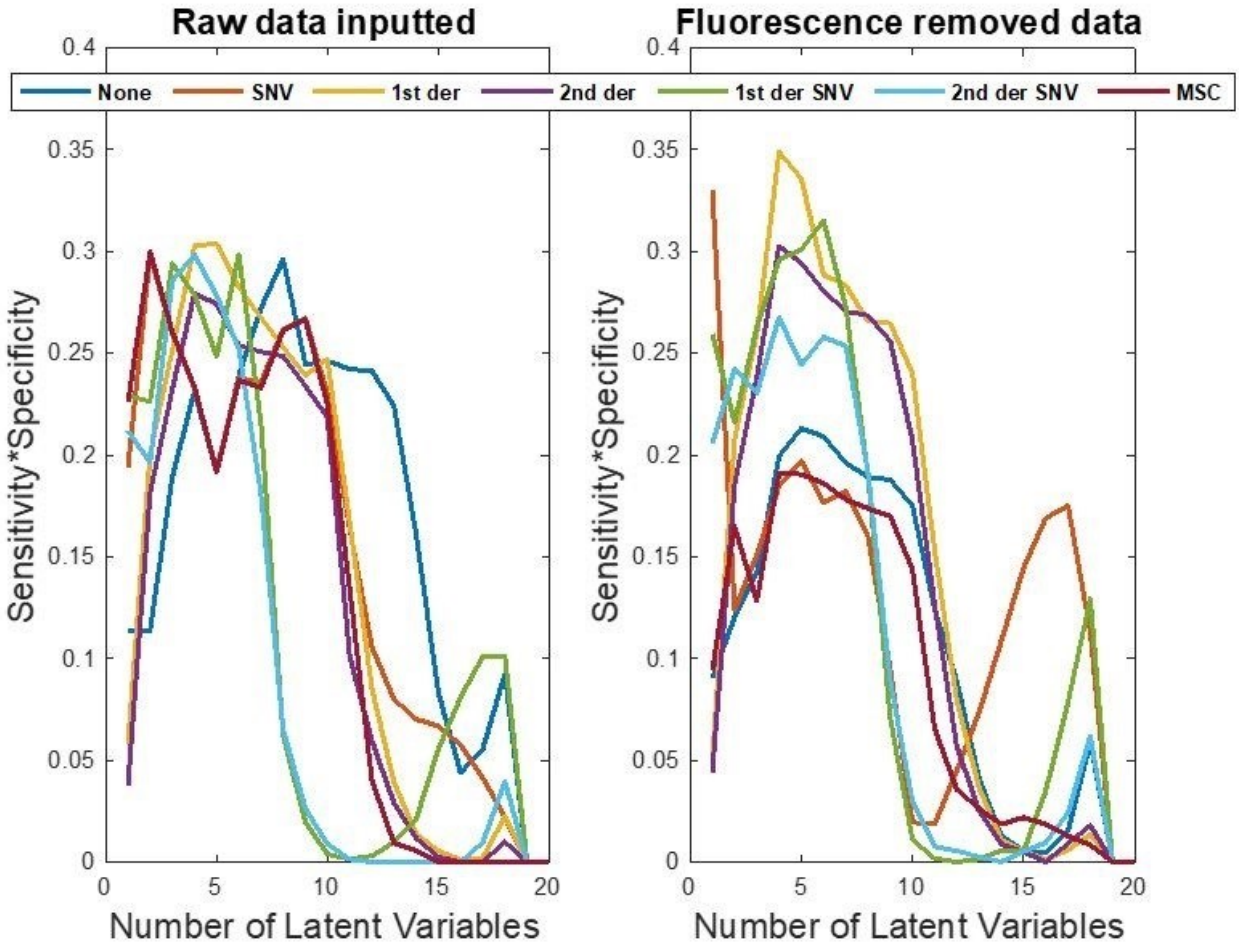

Fig. S19 Product of sensitivity and specificity for PLS-DA models constructed on median Raman spectra for Gleason grade3/grade 4 classification. Values shown are the mean calculated over 200 random splits of the reference image set. Models were constructed using untreated Raman spectra and six spectral pretreatments: standard normal variate pre-processing (SNV), 1st derivative Savitzky Golay pretreatment ('1st der', window size = 15 points, polynomial order = 3), 2nd derivative Savitzky Golay pretreatment ('2nd der', window size = 15 points, polynomial order = 3), combinations of SNV followed by 1st or 2nd derivative pretreatment and multiplicative scatter correction (MSC). The optimal PLS-DA model parameters were selected as fluorescence removal followed by 1st derivative Savitzky-Golay preprocessing and a PLS-DA model with 4 latent variables

Supplementary Tables

Table S1 Results of Tissue Thickness Experiment

Tissue Thickness	Summary of Results
4 μm	The tissue was too thin to be easily seen under the microscope without staining. No useable spectra could be acquired.
8 μm	The tissue was easily found under the microscope and useable spectra were acquired, however, there was noticeable interference from the glass substrate.
10 μm	The tissue was easy to find with the microscope and interference from the substrate was reduced in comparison to the 8 μm section.
16 μm	The tissue began peeling from the slide and did not remain useable.

Table S2 Confusion matrix for 12 latent variable PLS-DA model constructed on mean Raman spectra (fluorescence removal followed by standard normal variate preprocessing) for Non-Cancer (NC) /Cancer (C) classification using 5-fold cross validation

	Predicted NC	**Predicted C**
Actual NC	36	23
Actual C	15	53

Table S3 Confusion matrix for 9 latent variable PLS-DA model constructed on median Raman spectra (raw spectra followed by 2nd derivative Savitzky-Golay preprocessing) for Non-Cancer (NC) /Cancer (C) classification using 5-fold cross validation

	Predicted NC	**Predicted C**
Actual NC	32	27
Actual C	26	42

Table S4 Confusion matrix for 9 latent variable PLS-DA models constructed on mean Raman spectra (pretreated by fluorescence removal followed by 1st derivative Savitzky-Golay preprocessing) for Gleason Grade 3/ Gleason Grade 4 (G3/G4) classification using 5-fold cross validation

	Predicted G3	**Predicted G4**
Actual G3	22	17
Actual G4	21	9

Table S5 Confusion matrix for 9 latent variable PLS-DA models constructed on mean Raman spectra (fluorescence removal with no pretreatment) for Gleason Grade 3/ Gleason Grade 4 (G3/G4) classification using 5-fold cross validation

	Predicted G3	**Predicted G4**
Actual G3	20	19
Actual G4	18	12

Table S6 Confusion matrix for 4 latent variable PLS-DA models constructed on median Raman spectra (fluorescence removal followed by 1st derivative Savitzky-Golay preprocessing) for Gleason Grade 3/ Gleason Grade 4 (G3/G4) classification using 5-fold cross validation

	Predicted G3	**Predicted G4**
Actual G3	19	20
Actual G4	20	10

Table S7 Details of non-cancer/cancer (NC/C) and Gleason grade 3/grade 4 (G3/G4) models including diagnosis, modalities and number of reference sets. 5-fold cross-validation was used, which incorporated samples from disjoint groups of patients.

Model No.	**Diagnosis**	**Imaging Modality**	**No. of Reference Sets**
1	NC/C	DP	1
2		RCI	1
3		DP+RCI	1
4	G3/G4	DP	1
5		RCI	1
6		DP+RCI	1
7		DP	10
8		RCI	10
9		DP+RCI	10

Table S8 Parameters pertaining to optimal SIFT/BoVW/SVM classifiers that used samples from a single randomly selected set of patients as a reference set. Diagnostic tasks are non-cancer/cancer (NC/C) and Gleason grade 3/grade 4 (G3/G4) with digital pathology (DP), Raman Chemical Imaging (RCI) and multimodal (DP+RCI) imaging modalities

Diagnosis	**Imaging Modality**	**Dictionary Size 1**	**Dictionary Size 2**	**C**	γ	**Kernel**
NC/C	DP	300		2^{11}	2^{0}	rbf
	RCI		50	2^{8}	2	rbf
	DP+RCI	300	10	2^{12}	2^{-1}	rbf
G3/G4	DP	200		2^{7}	2^{3}	rbf
	RCI		5	2^{13}	2^{3}	rbf
	DP+RCI	300	5	2^{8}	2^{2}	rbf

Table S9 Parameters pertaining to optimal SIFT/BoVW/SVM classifiers. Each imaging modality was investigated using results from 10 different models, each using different sets of samples for reference images. The diagnostic task is Gleason grade 3/grade 4 (G3/G4) differentiation with digital pathology (DP), Raman Chemical Imaging (RCI) and multimodal (DP+RCI) imaging modalities

Imaging Modality	Reference Set No.	Dictionary Size 1	Dictionary Size 2	C	γ	Kernel
DP	1	100		2^{12}	2^{0}	rbf
	2	500		2^{8}	2^{1}	rbf
	3	500		2^{9}	2^{2}	rbf
	4	1000		2^{7}	2^{3}	rbf
	5	50		2^{14}	2^{-1}	rbf
	6	500		2^{9}	2^{1}	rbf
	7	100		2^{12}	2^{0}	rbf
	8	200		2^{7}	2^{3}	rbf
	9	1000		2^{11}		linear
	10	500		2^{9}	2^{1}	rbf
RCI	1		10	2^{4}	2^{3}	rbf
	2		25	2^{9}	2^{-1}	rbf
	3		5	2^{2}	2^{2}	rbf
	4		10	2^{15}	2^{-1}	rbf
	5		5	2^{5}	2^{-1}	rbf
	6		100	2^{8}	2^{2}	rbf
	7		25	2^{2}	2^{2}	rbf
	8		5	2^{13}	2^{3}	rbf
	9		25	2^{6}	2^{1}	rbf
	10		25	2^{6}	2^{1}	rbf
DP+RCI	1	50	5	2^{14}	2^{-1}	rbf
	2	200	50	2^{8}	2^{1}	rbf
	3	100	5	2^{8}	2^{1}	rbf
	4	1000	10	2^{12}	2^{-2}	rbf
	5	50	5	2^{12}	2^{-2}	rbf
	6	200	5	2^{11}	2^{-1}	rbf
	7	50	50	2^{8}	2^{1}	rbf
	8	300	5	2^{8}	2^{2}	rbf
	9	1000	5	2^{9}	2^{2}	rbf
	10	75	5	2^{14}	2^{-3}	rbf