

From Microorganisms to Prostate Cancer: Why SERS Spectra Need Machine Learning

Hao Zhu¹, Seda Kelestemur², Michael Grossberg³, Dorte M. Eisele²

¹M.S. Program in Data Science and Engineering, ²Department of Chemistry and Biochemistry,

³Department of Computer Science, The City College of New York, CUNY

Surface-enhanced Raman spectroscopy (SERS) records the vibrational fingerprints of many molecules near a plasmonic surface at once, which makes it a candidate for label-free cancer detection from blood. Whether the class can be read from a spectrum, however, depends on the task. We compared two SERS datasets of increasing difficulty. Spectra of three microorganisms (the yeast *Candida albicans* and the bacteria *Pseudomonas aeruginosa* and *Staphylococcus epidermidis*) were recorded at eight time points from 4 to 120 h, and their mean spectra differ visibly even at 4 h (Fig. 1a). A tabular foundation model, TabPFN [1] (version 3), given all time points of two experiments, classified single 4 h spectra of the third with a macro-averaged F1 score (macro-F1) of 0.945 ± 0.043 (mean $\pm$ standard deviation over three held-out experiments; chance 0.333). In a public serum SERS dataset of prostate cancer, benign prostatic hyperplasia and healthy controls [2], 59 patients after quality exclusions with 200 spectra each, the three class-mean spectra are nearly identical (Fig. 1b), and the same model reached a patient-level macro-F1 of only 0.52 (95% confidence interval [CI] 0.42-0.61). Under repeated cross-validation grouped by patient, a one-dimensional convolutional neural network (CNN) pretrained on 60,000 Raman spectra of 30 bacterial and yeast isolates [3] and fine-tuned on the serum spectra separated cancer from healthy with an area under the receiver operating characteristic curve (AUC) of 0.943 (CI 0.891-0.983), but cancer from benign, the clinically decisive comparison, only with 0.740 (0.629-0.841) (Fig. 1c). The same evaluation will be applied to spectra from the Eisele group's self-assembled nanotube substrates as they become available.

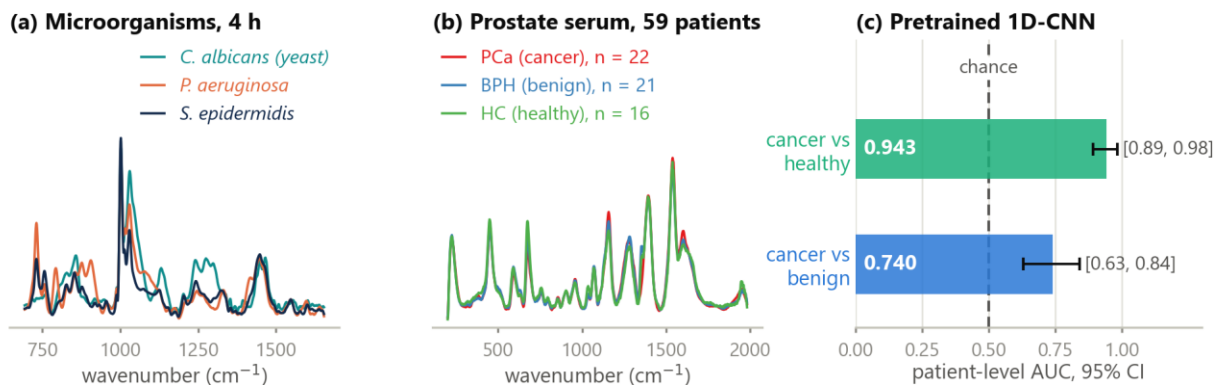


Figure 1. (a) Mean of min-max-scaled spectra of three microorganisms at 4 h, three experiments pooled. (b) Class-mean serum spectra after standard normal variate (SNV) scaling. (c) Patient-level AUC with 95% CI of the pretrained CNN; dashed line: chance.

References. [1] Hollmann, N. et al. *Nature* 637, 319 (2025). [2] Bi, X. et al. *Cell Rep. Med.* 5, 101579 (2024). [3] Ho, C.-S. et al. *Nat. Commun.* 10, 4927 (2019).