# Characterization and Detection of Precancerous and Cancerous Cells Using Raman Spectroscopy and Machine Learning Algorithms

Uraib Sharaha[a,b], Daniel Hania[c], Dima Bykhovsky[d], Itshak Lapidot[#e,f], Mahmoud Huleihel[#a], Ahmad Salman[#g,]

[a]*Department of Microbiology, Immunology and Genetics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.*
[b]*Department of Biology, Science and Technology College, Hebron University, Hebron P760, Palestine.*
[c]*Department of Green Engineering, SCE - Shamoon College of Engineering, Beer-Sheva 84100, Israel.*
[d]*Electrical and Electronics Engineering Department, SCE-Sami Shamoon College of Engineering, Beer-Sheva 84100, Israel*
[e]*Department of Electrical and Electronics Engineering, Afeka Tel-Aviv Academic College of Engineering, Tel-Aviv 69107, Israel.*
[f]*LIA Avignon Université, 339 Chemin des Meinajaries, Avignon 84000, France*
[g] *Department of Physics, SCE-Sami Shamoon College of Engineering, Beer-Sheva 84100, Israel.*

*Corresponding authors :

Prof. Ahmad Salman
orcid.org/0000-0003-4953-8648
Tel: +972-8-6475794
e-mail: ahmad@sce.ac.il

Prof. Mahmoud Huleihel
Tel: +972-8-6479867
e-mail: mahmoudh@bgu.ac.il

# Contributed equally.

**Abstract**

In the current study, the characterization and detection of precancerous and cancerous cells were performed using Raman spectroscopy-based machine learning algorithms. Since all the Raman spectra have huge backgrounds, mainly due to fluorescence, pre-processing is very important. Figure S-1 includes the Raman spectrum of one of our measurements before and after baseline correction to exclude the background signature from the spectra before being analyzed by the machine learning algorithm. The most significant 60 spectral features distinguishing between Controls and Precancerous, Controls and Cancerous, and Precancerous and Cancerous, derived using ANOVA F-score, are detailed in Table S1a. We applied relative entropy as an alternative feature selection method and compared the results with the ANOVA F-score approach in Table S1 b.

The optimal feature subset for Control vs. Cancerous and Control vs. Precancerous was selected through manual evaluation of feature importance rankings, as shown in Figures S2a and S2 b, respectively.

It is important to relate all the Raman features of the spectrum to the biological molecule that composes the cell samples and relate their functional group vibration modes to the Raman shift spectrum. Table S1 details the Raman peaks in Raman shift spectra with their respective assignment from the literature.

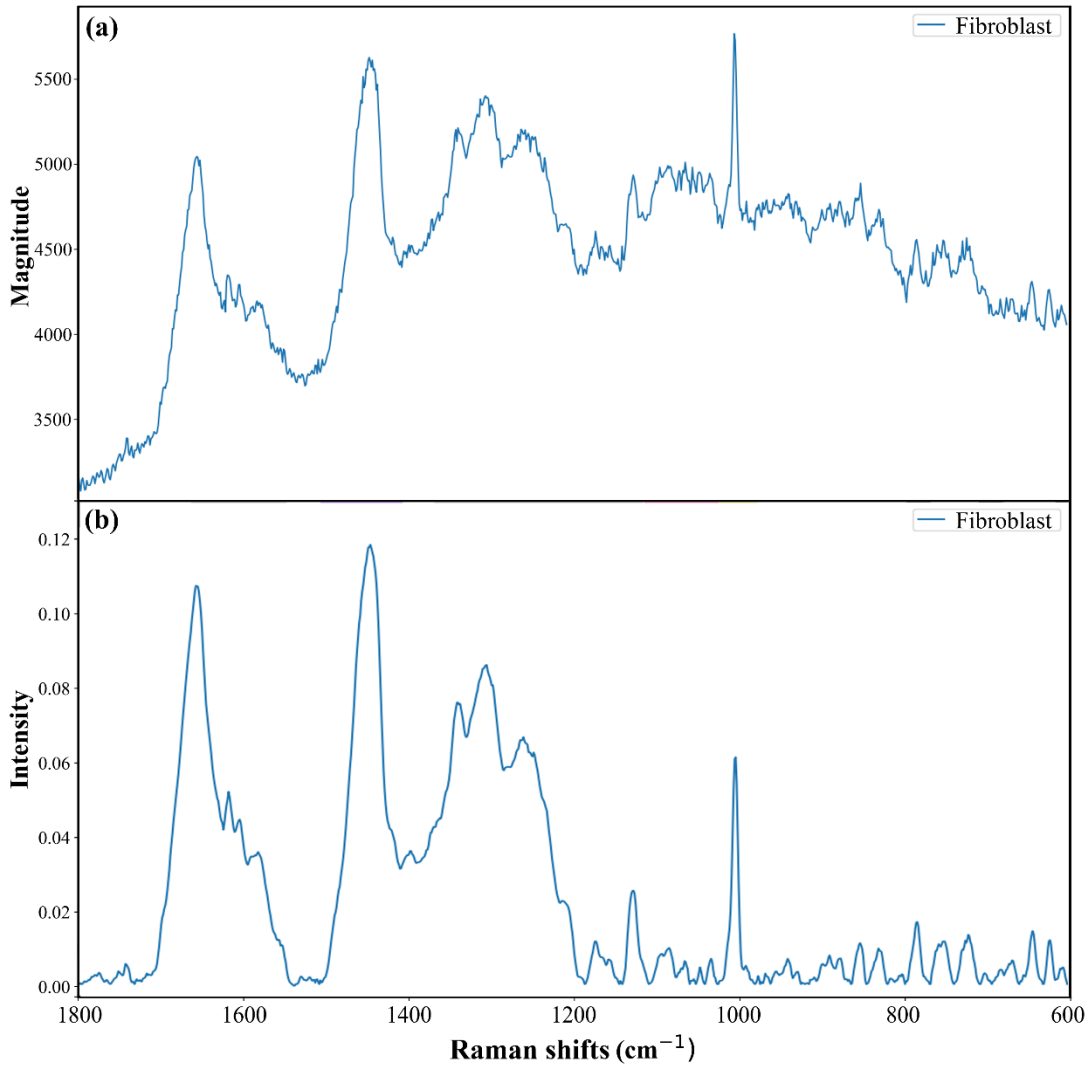


**Figure S1:** Typical Raman shift spectra of fibroblast cells: (a) before pre-processing in the 1800-400cm$^{-1}$ region, and (b) after pre-processing in the 1800-600cm$^{-1}$ region.

**Table S1a:** Top 60 features selected using the ANOVA F-score from the average Raman spectra (1800–600 $cm^{-1}$) across the three measurement sites for classification among the following pairs: Precancerous vs. Cancerous, Primary vs. Precancerous, and Primary vs. Cancerous.

| | Primary-Cancerous | | Primary-Precancerous | | Precancerous-Cancer | |
|---|---|---|---|---|---|---|
| | Wavenumber | Score | Wavenumber | Score | Wavenumber | Score |
| 1 | 1317 | 216 | 1315 | 123 | 1323 | 33 |
| 2 | 1316 | 209 | 1313 | 120 | 1322 | 32 |
| 3 | 1318 | 206 | 1316 | 120 | 1324 | 32 |
| 4 | 1315 | 195 | 1317 | 108 | 1703 | 30 |
| 5 | 1321 | 195 | 1246 | 108 | 1702 | 30 |
| 6 | 1331 | 192 | 1248 | 107 | 1701 | 29 |
| 7 | 1322 | 191 | 1249 | 107 | 1700 | 29 |
| 8 | 1326 | 191 | 1312 | 107 | 1747 | 29 |
| 9 | 1329 | 191 | 1250 | 102 | 1326 | 29 |
| 10 | 1333 | 190 | 1425 | 102 | 1321 | 29 |
| 11 | 1334 | 190 | 1098 | 98 | 1698 | 29 |
| 12 | 1320 | 189 | 1245 | 104 | 1697 | 28 |
| 13 | 1332 | 189 | 1097 | 96 | 1261 | 29 |
| 14 | 1324 | 188 | 1318 | 92 | 1263 | 29 |
| 15 | 1323 | 186 | 1624 | 104 | 1748 | 29 |
| 16 | 1328 | 186 | 1625 | 104 | 1746 | 28 |
| 17 | 1336 | 185 | 1094 | 97 | 1264 | 28 |
| 18 | 1327 | 181 | 1099 | 95 | 1705 | 27 |
| 19 | 1337 | 175 | 1093 | 97 | 1690 | 27 |
| 20 | 1313 | 172 | 1102 | 94 | 1327 | 27 |
| 21 | 1624 | 168 | 1627 | 101 | 1260 | 28 |
| 22 | 1625 | 168 | 1100 | 94 | 1320 | 27 |
| 23 | 1572 | 165 | 1623 | 99 | 1513 | 28 |
| 24 | 1627 | 162 | 1095 | 94 | 1215 | 27 |
| 25 | 1338 | 161 | 1628 | 97 | 1695 | 27 |
| 26 | 1571 | 159 | 1244 | 91 | 1693 | 27 |
| 27 | 1628 | 159 | 1426 | 92 | 1265 | 27 |
| 28 | 1573 | 157 | 1251 | 92 | 1696 | 27 |
| 29 | 1234 | 147 | 1424 | 93 | 1689 | 26 |
| 30 | 1233 | 147 | 1092 | 92 | 1691 | 27 |
| 31 | 1623 | 158 | 1243 | 87 | 1692 | 27 |
| 32 | 1629 | 157 | 781 | 86 | 1214 | 26 |
| 33 | 1575 | 152 | 1311 | 92 | 1269 | 26 |
| 34 | 1232 | 143 | 1629 | 93 | 1259 | 26 |
| 35 | 1217 | 145 | 1622 | 92 | 1217 | 26 |
| 36 | 1630 | 153 | 1575 | 88 | 1687 | 26 |
| 37 | 1230 | 143 | 1573 | 87 | 1272 | 26 |
| 38 | 1312 | 147 | 1572 | 87 | 1267 | 26 |
| 39 | 1339 | 142 | 1241 | 86 | 1318 | 26 |
| 40 | 1632 | 149 | 1103 | 87 | 1744 | 25 |
| 41 | 1570 | 146 | 1090 | 86 | 1266 | 26 |
| 42 | 1235 | 141 | 780 | 85 | 1270 | 26 |
| 43 | 1633 | 148 | 1576 | 85 | 1706 | 25 |
| 44 | 1218 | 142 | 782 | 84 | 1328 | 25 |
| 45 | 780 | 145 | 1620 | 87 | 1686 | 26 |
| 46 | 1229 | 142 | 1253 | 85 | 1271 | 26 |
| 47 | 1622 | 145 | 1446 | 84 | 1274 | 26 |
| 48 | 1634 | 146 | 1630 | 89 | 1637 | 24 |
| 49 | 1219 | 142 | 1484 | 84 | 1258 | 25 |
| 50 | 778 | 142 | 1320 | 85 | 1685 | 25 |
| 51 | 1220 | 142 | 1485 | 84 | 718 | 25 |
| 52 | 1638 | 145 | 1571 | 85 | 1638 | 25 |
| 53 | 1635 | 146 | 1445 | 85 | 1218 | 25 |
| 54 | 1576 | 140 | 1483 | 84 | 1635 | 24 |
| 55 | 1237 | 138 | 1619 | 86 | 1749 | 24 |
| 56 | 1637 | 147 | 1310 | 84 | 1254 | 25 |
| 57 | 1639 | 143 | 1240 | 84 | 1684 | 25 |
| 58 | 1215 | 140 | 1427 | 85 | 1256 | 25 |
| 59 | 1099 | 140 | 1336 | 83 | 1707 | 25 |
| 60 | 1100 | 139 | 1412 | 83 | 1329 | 26 |

**Table S1b:** Comparison between the selected features using ANOVA F-score and relative entropy methods similar to Table S1a.

| | Primary-Cancerous | | Primary-Precancerous | | Precancerous-Cancer | |
|---|---|---|---|---|---|---|
| | Wavenumber | | Wavenumber | | Wavenumber | |
| | ANOVA F-score | relative entropy | ANOVA F-score | relative entropy | ANOVA F-score | relative entropy |
| 1 | 1317 | 1317 | 1315 | 1315 | 1323 | 1323 |
| 2 | 1316 | 1316 | 1313 | 1316 | 1322 | 1322 |
| 3 | 1318 | 1318 | 1316 | 1313 | 1324 | 1324 |
| 4 | 1315 | 1321 | 1317 | 1317 | 1703 | 1703 |
| 5 | 1321 | 1315 | 1246 | 1246 | 1702 | 1261 |
| 6 | 1331 | 1331 | 1248 | 1248 | 1701 | 1263 |
| 7 | 1322 | 1322 | 1249 | 1249 | 1700 | 1326 |
| 8 | 1326 | 1326 | 1312 | 1312 | 1747 | 1702 |
| 9 | 1329 | 1329 | 1250 | 1624 | 1326 | 1701 |
| 10 | 1333 | 1333 | 1425 | 1245 | 1321 | 1700 |
| 11 | 1334 | 1320 | 1098 | 1625 | 1698 | 1748 |
| 12 | 1320 | 1334 | 1245 | 1250 | 1697 | 1747 |
| 13 | 1332 | 1332 | 1097 | 1425 | 1261 | 1698 |
| 14 | 1324 | 1324 | 1318 | 1093 | 1263 | 1697 |
| 15 | 1323 | 1323 | 1624 | 1098 | 1748 | 1321 |
| 16 | 1328 | 1328 | 1625 | 1094 | 1746 | 1264 |
| 17 | 1336 | 1336 | 1094 | 1097 | 1264 | 1746 |
| 18 | 1327 | 1327 | 1099 | 1627 | 1705 | 1705 |
| 19 | 1337 | 1337 | 1093 | 1095 | 1690 | 1690 |
| 20 | 1313 | 1313 | 1102 | 1623 | 1327 | 1327 |
| 21 | 1624 | 1572 | 1627 | 1099 | 1260 | 1260 |
| 22 | 1625 | 1625 | 1100 | 1102 | 1320 | 1696 |
| 23 | 1572 | 1624 | 1623 | 1628 | 1513 | 1265 |
| 24 | 1627 | 1338 | 1095 | 1318 | 1215 | 1513 |
| 25 | 1338 | 1571 | 1628 | 1100 | 1695 | 1320 |
| 26 | 1571 | 1627 | 1244 | 1424 | 1693 | 1693 |
| 27 | 1628 | 1234 | 1426 | 1251 | 1265 | 1691 |
| 28 | 1573 | 1233 | 1251 | 1244 | 1696 | 1692 |
| 29 | 1234 | 1628 | 1424 | 1426 | 1689 | 1215 |
| 30 | 1233 | 1573 | 1092 | 1092 | 1691 | 1695 |
| 31 | 1623 | 1623 | 1243 | 1243 | 1692 | 1689 |
| 32 | 1629 | 1629 | 781 | 781 | 1214 | 1687 |
| 33 | 1575 | 1575 | 1311 | 1311 | 1269 | 1269 |
| 34 | 1232 | 1232 | 1629 | 1629 | 1259 | 1214 |
| 35 | 1217 | 1630 | 1622 | 1622 | 1217 | 1744 |
| 36 | 1630 | 1217 | 1575 | 1575 | 1687 | 1259 |
| 37 | 1230 | 1312 | 1573 | 1573 | 1272 | 1272 |
| 38 | 1312 | 1230 | 1572 | 1103 | 1267 | 1217 |
| 39 | 1339 | 1570 | 1241 | 1241 | 1318 | 1267 |
| 40 | 1632 | 1632 | 1103 | 1572 | 1744 | 1270 |
| 41 | 1570 | 1235 | 1090 | 1090 | 1266 | 1706 |
| 42 | 1235 | 780 | 780 | 780 | 1270 | 1266 |
| 43 | 1633 | 1633 | 1576 | 1576 | 1706 | 1686 |
| 44 | 1218 | 1339 | 782 | 1446 | 1328 | 1318 |
| 45 | 780 | 1634 | 1620 | 1630 | 1686 | 1328 |
| 46 | 1229 | 1229 | 1253 | 1620 | 1271 | 1274 |
| 47 | 1622 | 1622 | 1446 | 1320 | 1274 | 1258 |
| 48 | 1634 | 1218 | 1630 | 782 | 1637 | 1685 |
| 49 | 1219 | 1219 | 1484 | 1484 | 1258 | 1637 |
| 50 | 778 | 1635 | 1320 | 1253 | 1685 | 1271 |
| 51 | 1220 | 1638 | 1485 | 1485 | 718 | 718 |
| 52 | 1638 | 1220 | 1571 | 1445 | 1638 | 1638 |
| 53 | 1635 | 778 | 1445 | 1310 | 1218 | 1635 |
| 54 | 1576 | 1637 | 1483 | 1571 | 1635 | 1218 |
| 55 | 1237 | 1237 | 1619 | 1427 | 1749 | 1749 |
| 56 | 1637 | 1576 | 1310 | 1412 | 1254 | 1256 |
| 57 | 1639 | 1639 | 1240 | 1240 | 1684 | 716 |
| 58 | 1215 | 1215 | 1427 | 1483 | 1256 | 1254 |
| 59 | 1099 | 1100 | 1336 | 1619 | 1707 | 1329 |
| 60 | 1100 | 1099 | 1412 | 1422 | 1329 | 1684 |

**Table S1:** The assignments of major Raman peaks in Raman shift spectra recorded from biological samples with their respective assignment

| **Wavenumber ($cm^{-1}$)** | **Assignment** |
|---|---|
| 1720–1745 | C=O stretching vibrations of lipids (triglycerides and cholesterol esters) |
| 1710–1716 | C=O antisymmetric stretching: RNA and purine base |
| 1705–1690 | C=O antisymmetric stretching vibrations: RNA, DNA |
| 1654 | Amide I: C=O (80%) and C–N (10%) stretching, N–H(%10) bending vibrations: proteins α-helix |
| 1630–1640 | Amide I: C=O (80%) and C–N (10%) stretching, N–H (10%) bending vibrations: proteins *β*-structure |
| 1610, 1578 | C4-C5 and C=N stretching in imidazole ring of DNA, RNA |
| 1515 | Aromatic tyrosine ring |
| 1540–1550 | Amide II: N–H (60%) bending and C–N (40%) stretching vibrations: proteins *α*-helix |
| 1530 | Amide II: N–H (60%) bending and C–N (40%) stretching vibrations : proteins β-structure |
| 1467 | CH2 bending vibrations: lipids and proteins |
| 1455 | CH3 bending and CH2 scissoring vibrations: lipids and proteins |
| 1370–1400 | COO– symmetric stretching and CH3 bending vibrations: lipids, proteins |
| 1330–1200 | Amide III: proteins |
| 1230–1244 | PO2 − antisymmetric stretching vibrations: RNA, DNA and phospholipids |
| 1060, 1050 | C–O stretching vibrations: deoxyribose/ribose DNA, RNA |
| 1003 | Phenylalanine (ring-breathing) |
| 925–929 | Sugar vibrations in the backbone of DNA-Z form |
| 967 | C–C and C–N stretch $PO3^{2-}$ stretching (DNA) |
| 957 | CH3 deformation (lipid, protein) |
| 936 | C–C residue α-helix |
| 921 | C–C stretch proline |
| 898 | C–C stretch residue |
| 870 | C-DNA |
| 853 | Ring breathing Tyr-C–C stretch proline |
| 828, 833 | Out of plane breathing Tyr; PO2 − asymmetric stretching, DNA (B form) |
| 807 | A-DNA |
| 786 | DNA-RNA (PO2 −) symmetric stretching |
| 746 | Thymine |
| 727 | Adenine |

S-

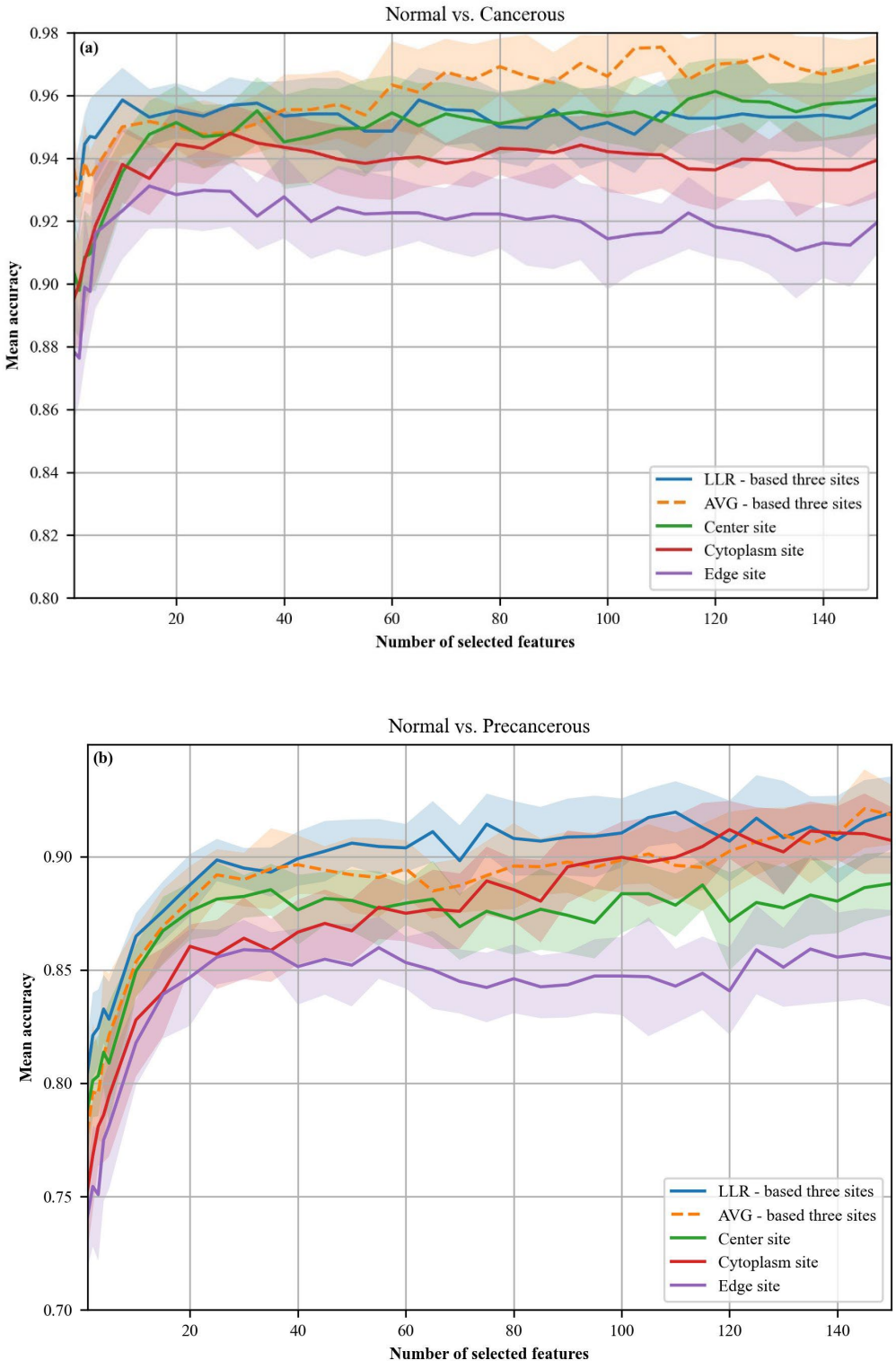


**Figure S2:** LR binary classification results comparing two approaches: (i) LLR-based decision logic applied across three measurement sites (center, cytoplasm, and membrane) and (ii) classification performed separately for each site without decision logic. (a) Normal vs. Cancerous, and (b) Normal vs. Precancerous.