Supplementary Material for

Automatic Identification of Novel Bacteria using Raman Spectroscopy and Gaussian Processes

Michael Kemmler[a], Erik Rodner[a],
Petra Rösch[b], Jürgen Popp[b,c], Joachim Denzler[a]

[a] *Computer Vision Research Group*
Department of Mathematics and Computer Science
Friedrich Schiller University of Jena, Germany
[b] *Institute of Physical Chemistry and Abbe Center of Photonics*
Friedrich Schiller University of Jena, Germany
[c] *Institute of Photonic Technology, Jena, Germany*

Abstract

The following material contains additional information to provide more insights of certain topics covered in the associated main manuscript. Experimental details such as parameter settings and the evaluation of the MLS approach, which are only of secondary interest for our paper, are incorporated here. Finally, recognition rates of all methods *with pre-processing* techniques prior to novelty detection are displayed.

S1. Implementation Details and Parameter Tuning Procedures

As in [1], the multi-class problem was tackled in one-vs-all fashion using a binary GP classifier with Laplace approximation and cumulative Gaussian likelihood. As covariance function, the extended isotropic exponential kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \theta_1^2 \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\theta_2^2)\right) \tag{1}$$

was used (in contrast to the kernel (5) for all other methods, which is based on a single hyperparameter). The hyparameters $[\theta_1, \theta_2]^T$ of the covariance function were estimated by maximizing marginal likelihood using the conjugate gradient optimizer `minimize` with 10 iterations for each binary one-vs-all problem. The additive noise component was set to a small value $\sigma_n^2 = 0.01$ to avoid numerical instabilities.

For SVDD, different values for outlier fraction parameter $\nu \in \{0, \ldots, 0.9\}$ were investigated ($\nu = 0$ meaning that the hard SVDD without slack variables is used). The Kernel KNN description method was used with exponential kernel κ, choosing among different sizes $K \in \{1, 5, 10, 25, 50\}$ of the nearest neighbor set of $\mathbf{x}_{NN}$. For the GMM, we followed the approach of Schmid *et al.* [2] using principle component analysis (PCA) as subspace reduction method and a full covariance matrix which is pooled over all strains. The number d of PCA components as well as the number k of normal distributions in the model were obtained by 10-fold cross-validation. Maximizing the average recognition rate on a 5×5-grid ($d \in \{10, 20, 30, 50, 80\}$

and $k \in \{5, 10, 20, 30, 50\}$), the optimum for our dataset was found to be $d = 30$ and $k = 30$. The MLS approach was re-implemented in Matlab and analyzed for a varying number $M \in \{10, 100, 500, 1000, 5000, 10000, 50000\}$ of simulated classes. Because of the randomized nature of the algorithm due to sampling, we always average over 100 runs to enable a robust performance analysis. As in GMMs, we projected the Raman spectra onto the first $d = 30$ PCA components. The latter step is also done for Parzen density estimation. Using a normal density with diagonal covariance as kernel, Silverman's rule of thumb[3] was used for estimating the bandwidth parameters for each dimension independently.

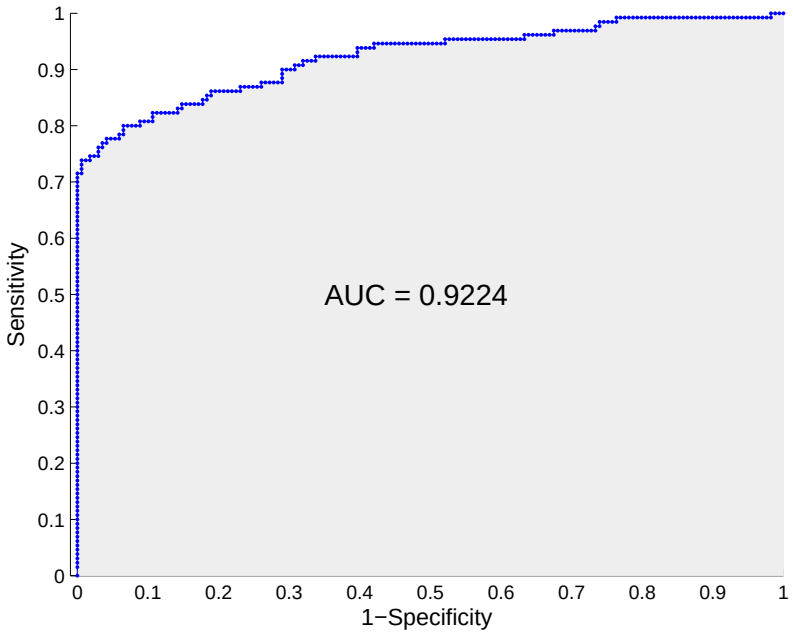

Figure S1: Example for a receiver operating characteristic (ROC) curve.

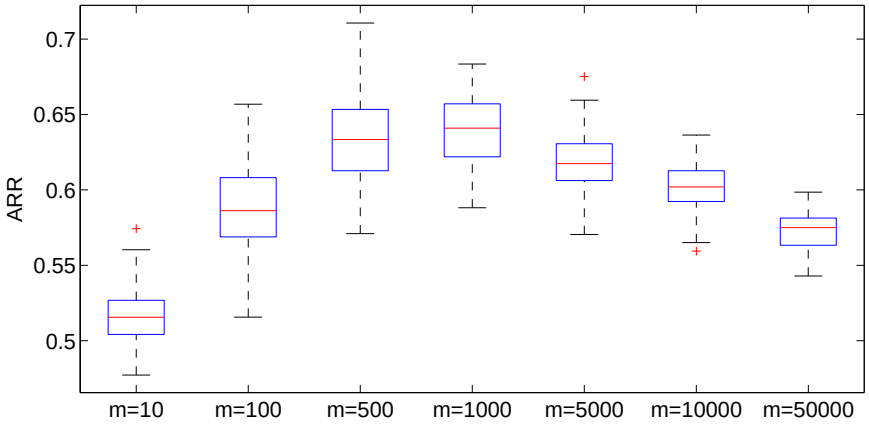

Figure S2: Novelty detection based on Maximum Likelihood with a varying number m of simulated outlier categories (MLS). The number of samples from the class prior is a critical parameter (optimum found at $m = 1000$), where under- or oversampling leads to inferior performance.

Table S1: False positives and false negatives of the best-performing novelty detection methods in the multi-class case. As 100 runs were taken for MLS due to its randomized nature, the two runs (MLS1 and MLS2) that lead to recognition rates equal to the median of the empirical ARR distribution are displayed.

			number of wrongly classified spectra					
	strain	#spectra	OVA	MLS1	MLS2	GMM	SVDD	GPR-V
known strains	*Bacillus sphaericus* DSM28	8	0	0	0	0	8	4
	Bacillus sphaericus DSM 396	7	0	1	1	0	7	5
	Bacillus subtilis DSM 347	8	0	2	3	1	2	4
	Escherichia coli DSM 1058	20	3	4	3	10	12	18
	Escherichia coli DSM 423	7	0	1	1	0	7	6
	Escherichia coli DSM 498	7	0	1	0	3	7	7
	Micrococcus luteus DSM 20030	6	0	2	3	6	1	6
	Micrococcus lylae DSM 20315	5	1	5	3	0	5	4
	Micrococcus lylae DSM 20318	5	2	5	4	3	0	5
	Staphylococcus cohnii DSM 20260	7	0	2	2	3	1	4
	Staphylococcus cohnii DSM 6669	8	3	3	5	4	0	7
	Staphylococcus cohnii DSM 6718	5	1	3	5	4	0	4
	Staphylococcus cohnii DSM 6719	5	1	2	2	3	0	3
	Staphylococcus epidermidis 195 Isolat	20	4	13	12	3	3	9
	Staphylococcus epidermidis ATTC 35984	7	3	7	7	0	0	5
	Staphylococcus warneri DSM 20036	5	0	0	0	0	0	3
novel strains	*Escherichia coli* DSM 426	24	22	13	12	16	6	1
	Escherichia coli DSM 5208	26	25	13	15	12	2	0
	Lactobacillus acidophilus DSM 9126	25	16	0	1	21	3	8
	Micrococcus luteus DSM 3906	45	8	20	23	10	33	0
	Staphylococcus hominis BCD 2684	21	5	5	2	16	16	4
	Streptococcus thermophilus DSM 20617	28	16	4	2	0	2	0

S2. Experiments using Pre-processing Techniques

The following section includes recognition rates when pre-processing techniques (reduction to fingerprint region, background subtraction) are employed prior to classification.

Table S2: Novelty detection results solely based on the fingerprint region 540–1800 cm^{-1}.

(a) multi-class case

method	specificity	sensitivity	ARR
GMM	77 (45.6%)	100 (76.9%)	61.2%
Parzen	169 (100.0%)	0 (0.0%)	50.0%
SVDD ($\nu = 0$)	89 (52.7%)	78 (60.0%)	56.3%
Kernel KNN ($K = 50$)	2 (1.2%)	130 (100.0%)	50.6%
GPR-M	21 (12.4%)	124 (95.4%)	53.9%
GPR-V	155 (91.7%)	45 (34.6%)	63.2%
MLS ($m = 1000$)	119.75 (70.9%)	76.45 (58.8%)	**64.8**%

(b) one-class case

method	specificity	sensitivity	ARR
GMM	23 (13.6%)	127 (97.7%)	55.7%
Parzen	169 (100.0%)	0 (0.0%)	50.0%
Kernel KNN ($K = 10$)	31 (18.3%)	119 (91.5%)	54.9%
SVDD ($\nu = 0$)	29 (17.2%)	123 (94.6%)	55.9%
GPR-M	38 (22.5%)	119 (91.5%)	**57.0**%
GPR-V	27 (16.0%)	126 (96.9%)	56.4%

Table S3: Novelty detection results using a background removal algorithm for eliminating fluorescence artefacts.

(a) multi-class case

method	specificity	sensitivity	ARR
GMM	89 (52.7%)	128 (98.5%)	**75.6**%
Parzen	169 (100.0%)	0 (0.0%)	50.0%
Kernel KNN ($K = 50$)	3 (1.8%)	130 (100.0%)	50.9%
SVDD ($\nu = 0.1$)	119 (70.4%)	73 (56.2%)	63.3%
GPR-M	87 (51.5%)	104 (80.0%)	65.7%
GPR-V	169 (100.0%)	2 (1.5%)	50.8%
MLS ($m = 1000$)	119.15 (70.5%)	74.68 (57.4%)	64.0%
OVA	54 (32.0%)	101 (77.7%)	54.3%

(b) one-class case

method	specificity	sensitivity	ARR
GMM	59 (34.9%)	130 (100.0%)	67.5%
Parzen	169 (100.0%)	0 (0.0%)	50.0%
Kernel KNN ($K = 10$)	42 (24.9%)	126 (96.9%)	60.9%
SVDD ($\nu = 0.1$)	28 (16.6%)	129 (99.2%)	57.9%
GPR-M	73 (43.2%)	108 (83.1%)	63.1%
GPR-V	130 (76.9%)	99 (76.2%)	**76.5**%

References

[1] M. Kemmler, J. Denzler, P. Rösch, J. Popp, in *Proc. DAGM*, **2010**, pp. 81–90.
[2] U. Schmid, P. Rösch, M. Krause, M. Harz, J. Popp, K. Baumann, *Chemometr. Intell. Lab.* **2009**; *96*, 159 .
[3] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, **1986**.