Supplemental Information

An outlier removal method based on PCA-DBSCAN for blood-SERS data analysis

Miaomiao Liu, [a] Tingyin Wang, *[a] Qiyi Zhang, [a] Changbin Pan, [a] Shuhang Liu, [a] Yuanmei Chen, *[b] Duo Lin, [a] and Shangyuan Feng *[a]

[a] Key Laboratory of OptoElectronic Science and Technology for Medicine of Ministry of Education, Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou, 350117, China. Email: tywang@fjnu.edu.cn; syfeng@fjnu.edu.cn

[b] Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou, Fujian, 350001, China. Email: chenym091@hotmail.com

Table S1 Results summary for KNN classifier

Classifier: K-Nearest Neighbors (KNN)					
Eps	MinPts	Accuracy	Precision	Recall	F1-score
		0.8907	0.9073	0.8871	0.8953
0.0004	4	**0.9765**	**0.9814**	**0.9741**	**0.9774**
	6	0.9571	0.9648	0.9314	0.9451
	8	0.9463	0.9675	0.9406	0.9514
0.0006	4	0.9602	0.9675	0.9457	0.9548
	6	0.9540	0.9639	0.9465	0.9540
	8	0.9598	0.9665	0.9444	0.9536
0.0008	4	0.9162	0.9314	0.9014	0.9131
	6	0.9101	0.9272	0.8911	0.9045
	8	0.9213	0.9368	0.9049	0.9173
0.001	4	0.9050	0.9245	0.8843	0.8987
	6	0.9050	0.9245	0.8843	0.8987
	8	0.9274	0.9407	0.9150	0.9254
0.0012	4	0.8889	0.9114	0.8711	0.8855
	6	0.8889	0.9114	0.8711	0.8855
	8	0.8889	0.9114	0.8711	0.8855
0.0014	4	0.9006	0.9199	0.8884	0.9005
	6	0.8889	0.9114	0.8711	0.8855
	8	0.8889	0.9114	0.8711	0.8855
0.0016	4	0.9006	0.9215	0.8850	0.8985
	6	0.9006	0.9215	0.8850	0.8985
	8	0.9056	0.9235	0.8928	0.9047

Table S2 Results summary for SVM classifier

Classifier: Support Vector Machine (SVM)					
Eps	MinPts	Accuracy	Precision	Recall	F1_score
		0.8234	0.8918	0.8450	0.8604
0.0004	4	**0.8824**	0.9247	0.8551	**0.8782**
	6	0.8405	0.9133	0.7605	0.7929
	8	0.8725	**0.9319**	0.8216	0.8588
0.0006	4	0.8807	0.9193	**0.8557**	0.8760
	6	0.8793	0.9224	0.8451	0.8684
	8	0.8678	0.9134	0.8345	0.8579
0.0008	4	0.8324	0.8798	0.7977	0.8195
	6	0.8315	0.8792	0.7943	0.8168
	8	0.8652	0.9009	0.8469	0.8645
0.001	4	0.8268	0.8738	0.7891	0.8105
	6	0.8268	0.8738	0.7891	0.8105
	8	0.8380	0.8859	0.8062	0.8283
0.0012	4	0.8444	0.8742	0.8110	0.8285
	6	0.8444	0.8742	0.8110	0.8285
	8	0.8444	0.8742	0.8110	0.8285
0.0014	4	0.8453	0.8782	0.8185	0.8359
	6	0.8444	0.8742	0.8110	0.8285
	8	0.8444	0.8742	0.8110	0.8285
0.0016	4	0.8453	0.8777	0.8151	0.8329
	6	0.8453	0.8691	0.8156	0.8315
	8	0.8556	0.8840	0.8276	0.8447

Table S3 Results summary for Random Forest classifier

Classifier: Random Forest					
Eps	MinPts	Accuracy	Precision	Recall	F1_score
		0.9118	0.9147	0.8906	0.8997
0.0004	4	0.9118	0.9147	0.8906	0.8997
	6	0.8528	0.9148	0.7657	0.7936
	8	0.8792	0.8678	0.8376	0.8502
0.0006	4	0.8977	0.9240	0.8891	0.9004
	6	**0.9195**	**0.9470**	**0.8925**	**0.9113**
	8	0.8793	0.8861	0.8779	0.8781
0.0008	4	0.7486	0.7419	0.7239	0.7304
	6	0.8933	0.9163	0.8525	0.8705
	8	0.8539	0.8364	0.8327	0.8338
0.001	4	0.8603	0.8993	0.8154	0.8355
	6	0.8715	0.9074	0.8465	0.8655
	8	0.8045	0.8805	0.7439	0.7623
0.0012	4	0.8889	0.9255	0.8634	0.8827
	6	0.8444	0.8776	0.8225	0.8403
	8	0.8889	0.9164	0.8607	0.8782
0.0014	4	0.8840	0.8890	0.8605	0.8706
	6	0.8278	0.8356	0.8117	0.8195
	8	0.8389	0.8535	0.8037	0.8183
0.0016	4	0.8674	0.8798	0.8417	0.8541
	6	0.8785	0.8963	0.8694	0.8795
	8	0.8778	0.8866	0.8561	0.8672

Table S4 Results summary for Adaboost classifier

Classifier: Adaboost					
Eps	MinPts	Accuracy	Precision	Recall	F1_score
		0.7869	0.8371	0.7551	0.7767
0.0004	4	0.8294	0.8972	0.8263	0.8383
	6	0.9632	0.9591	0.9489	0.9534
	8	0.7584	0.8768	0.7550	0.7569
0.0006	4	0.9205	0.9152	0.9277	0.9196
	6	0.9598	0.9552	0.9542	0.9546
	8	**0.9655**	**0.9688**	**0.9542**	**0.9607**
0.0008	4	0.9162	0.9263	0.9122	0.9167
	6	0.8989	0.9065	0.8890	0.8948
	8	0.8708	0.8951	0.8722	0.8806
0.001	4	0.9218	0.9274	0.9289	0.9239
	6	0.9106	0.9156	0.9118	0.9102
	8	0.9274	0.9382	0.9293	0.9304
0.0012	4	0.9278	0.9310	0.9289	0.9292
	6	0.9278	0.9310	0.9289	0.9292
	8	0.9278	0.9310	0.9289	0.9292
0.0014	4	0.7459	0.8209	0.7605	0.7594
	6	0.9333	0.9396	0.9334	0.9357
	8	0.9278	0.9350	0.9250	0.9291
0.0016	4	0.8785	0.9117	0.8915	0.8892
	6	0.8508	0.8716	0.8210	0.8366
	8	0.9278	0.9328	0.9267	0.9293

Table S5 Results summary for GNB classifier

Classifier: Gaussian Naive Bayes					
Eps	MinPts	Accuracy	Precision	Recall	F1_score
		85.79	86.10	85.86	85.81
0.0004	4	**0.9176**	0.9125	**0.9183**	0.9128
	6	0.8896	0.9083	0.8525	0.8722
	8	0.8792	0.8949	0.8437	0.8645
0.0006	4	0.9148	**0.9173**	0.9169	**0.9151**
	6	0.8851	0.8826	0.8758	0.8780
	8	0.8851	0.8756	0.8758	0.8746
0.0008	4	0.8994	0.9047	0.9043	0.9021
	6	0.8989	0.9039	0.9038	0.9015
	8	0.8933	0.8960	0.8993	0.8952
0.001	4	0.8771	0.8854	0.8701	0.8752
	6	0.8771	0.8854	0.8701	0.8752
	8	0.8883	0.8961	0.8872	0.8891
0.0012	4	0.8944	0.8968	0.8931	0.8931
	6	0.8944	0.8968	0.8931	0.8931
	8	0.8944	0.8968	0.8931	0.8931
0.0014	4	0.8895	0.8982	0.8957	0.8934
	6	0.8944	0.8968	0.8931	0.8931
	8	0.8944	0.8968	0.8931	0.8931
0.0016	4	0.8840	0.8873	0.8912	0.8856
	6	0.8950	0.8951	0.9011	0.8956
	8	0.9056	0.9093	0.9098	0.9077

DBSCAN operates by defining a neighbourhood around each data point and identifying core points, which have enough neighbouring points within a specified distance (Eps). It then expands the clusters by connecting core points to their density-reachable neighbors.[36] The key parameters of DBSCAN are the distance threshold (Eps) and the minimum number of points required to form a core point (MinPts). These parameters influence the cluster formation and can be adjusted based on the characteristics of the dataset. Assuming the spectral data sample set is defined as $D = (s_1, s_2, ..., s_m)$, the density concepts involved in DBSCAN are defined as follows:

(1) Eps-Neighbourhood: For a sample $s_j \in D$, its ε-neighbourhood contains a subset of samples from the dataset D that are at a distance less than or equal to Eps from s_j, denoted as $N\varepsilon(s_j) = \{s_i \in D | \text{distance}(s_i, s_j) \leq \varepsilon\}$. The number of samples in this subset is denoted as $|N\varepsilon(s_j)|$.

(2) Core Object: For a sample $s_j \in D$, if its ε-neighbourhood corresponding to $N\varepsilon(s_j)$ contains at least MinPts samples, i.e., $|N\varepsilon(s_j)| \geq \text{MinPts}$, then s_j is considered a core object.

(3) Density direct: If a sample s_i is located in the ε-neighbourhood of sample s_j, and s_j is a core object, then s_i is said to be directly reached by the density of s_j.

(4) Density reachable: For samples s_i and s_j, if there exists a sample sequence $p_1, p_2, ..., p_T$ satisfying that $p_1 = s_i$, $p_T = s_j$, and p_{T+1} is directly reached by p_T in terms of density, then s_j is considered reachable by the density of s_i.

(5) Density connectivity: For samples s_i and s_j, if there exists a core object sample s_k such that both s_i and s_j are reachable by the density of s_k, then s_i and s_j are considered density-connected.