Electronic Supplementary Material (ESI) for Sensors & Diagnostics.
This journal is © The Royal Society of Chemistry 2024

Ultra-low Dual Detection of Tetrahydrocannabinol and Cannabidiol in Saliva based on Electrochemical Sensing and Machine Learning: Overcoming Cross-Interferences and Saliva-to-Saliva Variations

Greter A. Ortega[1], Herlys Viltres[1], Hoda Mozaffari, Syed Rahin Ahmed, Seshasai Srinivasan,* Amin Reza Rajabzadeh*

School of Engineering Practice and Technology, McMaster University, 1280 Main Street West Hamilton, ON, L8S 4L8, Canada

Table S1. Summary of different literature studies on machine learning for electrochemical sensors.

Sensors	Analyte	Electrochemical Method	Sample	ML method	Ref
CuNPs/PEDOT-C4-COOH/3-electrodes-CWE	Maleic hydrazide	CV/oxidation	Spiked samples of onion, rice, potato, and cotton leaf	ANN / different traditional regression methods.	[1]
Disposable laser-induced porous graphene (LIPG)	Maleic hydrazide	DPV	PBS/ Potatoes and peanuts	Regression, Back-propagation ANN (BP-ANN), random forest (RF), and least squares support vector machine (LS-SVM)	[2]
Second-generation glucose-oxidase biosensor	Glucose	Amperometric	-	Regression, Partial Least Squares (PLS), Support Vector Machine for	[3]

(GOB).				Regression (SVMR-Lin) (SVMR-RBF) Artificial Neural Networks (ANN)	
nitrate reductase (NR)/3-electrodes-CWE	Nitrate	CV/reduction	Spiked lake water, vegetable juice, and fruit juice	Regression, Support Vector Machine (SVM)	[4]
CMC-MWCNTs/ MoS_2	Carbendazim	DPV/oxidation	Tea and rice	ANN model and traditional regression models	[5]
Pt/Ir working electrodes, Ag/ AgCl reference electrodes and platinum auxiliary electrodes with nano-platinum deposited.	Acetone	EIS	Water	Support vector machine (SVM) classification (0,1) (absence or presence of acetone) Training (Data set 80%, accuracy 98 %) Testing (Data set 20%, accuracy 97%)	[6]
Silver electrodes	Bacterias	EIS	Water	Classification, linear maximum likelihood estimation (MLE), linear discrimination analysis (LDA), and non-linear back propagation neural network (BPNN) methods. In the	[7]

				last case, 84 vectors (70% data) were used as training, while the remaining 36 vectors (30%) were divided equally between the testing and validation data sets. The accuracy was 100 % in all cases.	
laser-induced porous graphene WE	Salicylic acid	CV and LSV tests/oxidation	PBS/ lettuce and watermelon extracting solution.	Regression, ANN, and least squares support vector machine (LSSVM)	[8]
24 potentiometric sensors	Multiple analytes/bladder cancer	Potentiometric	urine	Logistic regression (LR), random forest (RF), extreme gradient boosting classifier (XGBC), support vector machine classifier (SVM), and voting classifier (VC). Classification yes/no bladder cancer. Training accuracy near 100 %, and the testing highest	[9]

				accuracy 80 %.	
Carbon, Prussian blue, Cobalt (II) phthalocyanine, Copper (II) oxide, Polypyrrole, and Palladium nanoparticles ink-modified carbon electrodes.	Heroin, morphine, codeine, paracetamol and caffeine	SWV	Water	PCA and Silhouette parameter calculation, K-nearest neighbor classifier. The accuracy presented is 100%.	[10]
Black phosphorene (BP) modified electrode	5-hydroxytryptamine	SWV	PBS	Regression, ANN algorithm	[11]
Reduced graphene and gold nanoparticles modifying glassy Carbon electrode (AuNPs/rGO/GCE)	Detection of Dopamine in the presence of epinephrine	DPV	Goat serum samples and artificial urine	POS-ANN model	[12]

Table S2. Summary of different papers about machine learning for the detection of THC.

Technique	Sample	ML method	LOD	Accuracies	Ref
Color-based lateral flow+immunoassay	16 drugs and ethanol THC/saliva	Multilayer perceptron artificial neural network (MLP-ANN)	THC-50 ng/mL	Overall were Training 100 %, Validation	[13]

		classification “very positive”, “positive”, “doubtful”, “negative”, “very negative” or “undetermined” (VP, P, D, N, VN, U)		92 %, and Testing 89.7 %	
FTIR	THC/ cannabis inflorescence	The Savitzky-Golay 2nd derivative (polynomial order: 2, window size: 3) and standard normal variate (SNV) were the preprocessing. Two types of regressors were trained using the preprocessed datasets: Genetic Algorithm and Ensemble regression models.	-	**N/A**	[14]
EIS/anti-THC *(incubation 15 min)*	THC-BSA/saliva THC/saliva	Binary classification of THC+/− Two Logistic Regression models – one without and one with K-folds cross-validation and two Support Vector machines (SVM) – one with a linear kernel and one with a radial bias kernel.	100 pg/mL	**N/A**	[15]

s-SWCNTs chemiresistor	THC/ breath	Random forest (RF), k-nearest neighbor (kNN), and support vector machine classifier (SVC) were used to classify the recovery traces as containing THC or not.	0.163 ng	**N/A**	[16]

Table S3. Examples of filters, swabs, and collectors used to collect and filtrate the saliva samples.

Filters-diameter-pore size	Swabs/collectors
PTFE-25 mm-0.2 µm	PureSal/Filtration(Swab + squeeze)
PES-25 mm-0.2 µm	NeoSal (Swab + buffer) 1:4
PVDF-25 mm-0.2 µm	SalivaBio swab (Swab + squeeze)
Nylon-25 mm-0.2 µm	SalivaBio swab + Pure Sal filter
Nylon-25 mm-0.45 µm	POREX OFCD-100 (No filter)
Nylon-13 mm 0.45 um*	POREX OFCD-201-SRF (with filter)
wwPTFE NanoSEP-0.2 µm*	POREX OFCD-100 +glass wool
wwPTFE NanoSEP-0.45 µm*	POREX OFCD-100 swab +glass wool
wwPTFE-13mm-0.45 µm*	N/A
wwPTFE-13mm-0.2 µm*	N/A
wwPTFE-25mm-0.2 um*	N/A
Glass wool (Pyrex 3950)	*Pall company

PTFE-Polytetrafluoroethylene, PES-Polyethersulfone, PVDF- Polyvinylidene, wwPTFE-water wettable polytetrafluoroethylene

Table S4. Interference experiments detail.

Experiments	THC based- Sensor (m-Z-THC)	Experiments	CBD based-Sensor (m-Z-CBD)
	Total electrodes P-Z/m-Z (1 Saliva)		Total electrodes P-Z/m-Z (1 Saliva)
[THC]= 0 ng/mL [CBD]= 0 ng/mL	8 (1/7m-Z-THC)	[CBD]= 0 ng/mL [THC]= 0 ng/mL	8 (1/7m-Z-CBD)
[THC]= 0 ng/mL [CBD]= 10 ng/mL	4 (1/3m-Z-CBD)	[CBD]= 0 ng/mL [THC]= 10 ng/mL	4 (1/3 m-Z-CBD)
[THC]= 0 ng/mL [CBD]= 50 ng/mL	4	[CBD]= 0 ng/mL [THC]= 50 ng/mL	4
[THC]= 2 ng/mL [CBD]= 0 ng/mL	8	[CBD]= 2 ng/mL [THC]= 0 ng/mL	8
[THC]= 2 ng/mL [CBD]= 10 ng/mL	4	[CBD]= 2 ng/mL [THC]= 10 ng/mL	4
[THC]= 2 ng/mL [CBD]= 50 ng/mL	4	[CBD]= 2 ng/mL [THC]= 50 ng/mL	4
[THC]= 5 ng/mL [CBD]= 0 ng/mL	8	[CBD]= 5 ng/mL [THC]= 0 ng/mL	8
[THC]= 5 ng/mL [CBD]= 10 ng/mL	4	[CBD]= 5 ng/mL [THC]= 10 ng/mL	4
[THC]= 5 ng/mL [CBD]= 50 ng/mL	4	[CBD]= 5 ng/mL [THC]= 50 ng/mL	4
Total of electrodes (1 Saliva)	**48**	**N/A**	48
Total of electrodes (6 Salivas)	**576**		

S2.2 Machine Learning algorithms.

Random Forest (RF) was used to classify the concentration of THC present in saliva for the purposes of this study. Random Forest is an ensemble machine learning method that combines a group of different Decision Trees, where each tree trains a subset of the training set with randomly selected predictors among all features. The training dataset will be divided repeatedly into

subspaces based on an attribute that offers maximum information gain. This splitting process is robust to outliers and multicollinearity. As a result, Decision Tree and Random Forest algorithms perform well when features have different scales and do not require feature scaling. A Decision Tree method is prone to overfitting since its structure can mimic the data closely. This problem is mainly resolved by introducing the concept of randomness in Random Forest methods. In this study, the optimal number of random trees was 200.

Artificial Neural Network (ANN) is another powerful ML technique used in this paper for classification. A neuron (perceptron) is the essential component of a Dense Neural Network structure, with weighted inputs and a bias. A neuron introduces non-linearity to the system through a proper activation function. The model often initiates by choosing minimal random weights and biases. Later, it adjusts these values based on a gradient descent algorithm to minimize a loss function. Vanishing and exploding gradients are the main problems of Neural Networks. This issue can be addressed by an appropriate activation function, batch normalization, and implementing gradient clipping. Overfitting is a common issue for Neural Networks, especially with a limited amount of data. Strategies like early stopping, regularization techniques, and dropout can be possible solutions. Finding a suitable architecture for an ANN is a trial and error process and depends on datasets. The dense network structure used in this study consists of three hidden layers with 32,64 and 128 nodes, respectively. Rectified Linear Unit function used for hidden layers' activation function and softmax for the output layer. Additionally, regularization techniques, as well as dropout, are implemented.

This study used support Vector Machine (SVM) as another alternative for classification and regression. SVM technique finds a hyperplane to separate different classes by maximizing an acceptable margin between the hyperplane and the nearest points of a class. Support vectors are the outliers and the closest data points in each category to the hyperplane. These vectors play a critical role in the positioning of the hyperplane. In many datasets, finding the hyperplane in low-dimension space to separate the classes is impossible. In other words, the hyperplane exists in higher-dimension, and datasets must be transformed into high-dimension space. As a result, the kernel trick is often used to map the datasets to the new dimension based on only the similarity and distances between two points in the original dimension. The concept of SVM techniques for regression is the same, finding a hyperplane that fits the maximum number of points. Contrary to regular regression models, the objective is not to minimize the sum of squared errors, but instead

to find a maximum acceptable margin of error to fit the training set along the hyperplane. The distances between data points play a crucial role in Machine learning algorithms like SVM; hence feature scaling, and dimensionality reduction are highly recommended for SVM methods. The SVM approaches can be costly in memory requirements and time computational power. Moreover, SVM techniques are susceptible to noises that can lead to overfitting. This study used a radial basis function kernel for classification and a moderate regularization parameter for regression.

This study used a Logistic Regression classifier for binary classification of interaction between THC and CBD. Logistic Regression predicts the probability of a point belonging to a binary class. It uses a sigmoid function and a threshold criterion to calculate the probability. It is an easy model to understand and implement; nonetheless, it is prone to overfitting when the number of predictors is higher than the number of instances.

Finally, this work used Principal Component Analysis (PCA) for dimensionality reduction and different preprocessing techniques, including Standard Scaler and non-linear Power Transformer for feature scaling. Small to medium size datasets with a large number of features are at high risk of overfitting. Dimensionality reduction techniques intend to preserve datasets' information in lower space and reduce the complexity of the model and possible multicollinearity in the system. Consequently, the computational time, memory requirement, and noise and redundancy decrease while accuracy improves. PCA techniques find a subspace (hyperplane) to transfer data while maintaining the original variances. Among feature scaling methods, a standard scaler algorithm modifies the mean and variance of each feature. On the other hand, A Power Transformer is a non-linear transformer that changes the correlation and distances between data points.

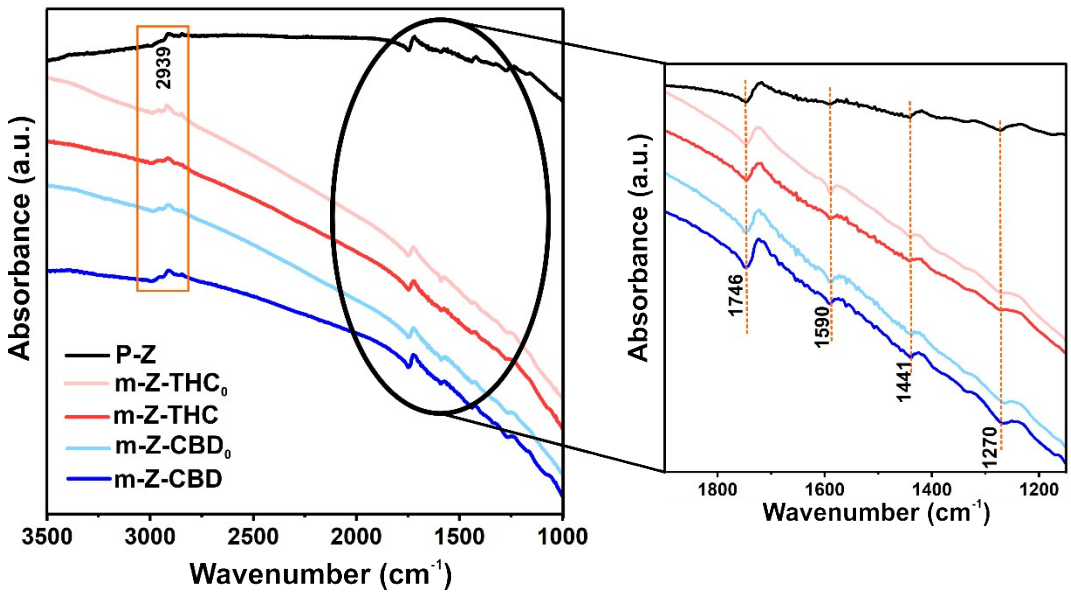

Figure S1. FTIR spectra of pristine and modified electrodes.

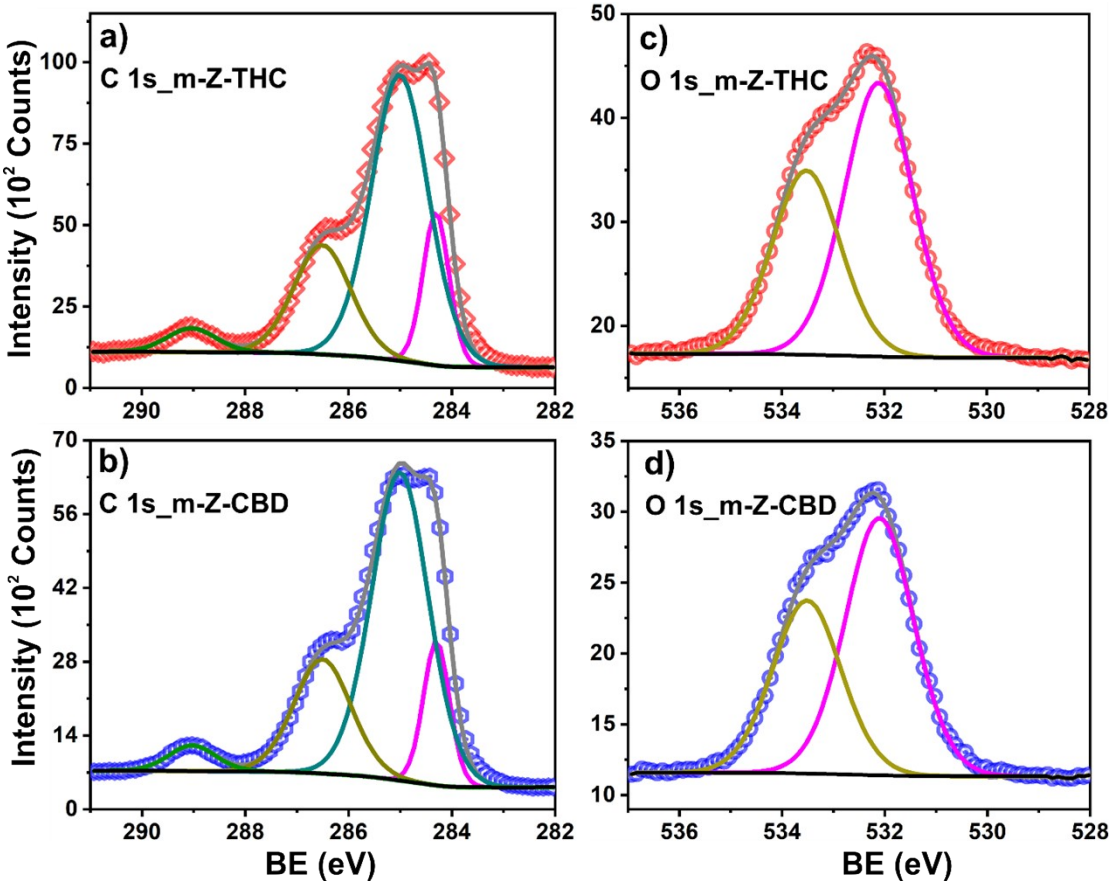

Figure S2. a), b) C1s and **c), d)** O1s high-resolution spectra before and after Zensor working electrode modification.

Table S5. XPS survey data (atomic percentage) for the most concentrated elements present in the materials.

Samples	Elements (At. %)						
	C 1s	**O 1s**	**N 1s**	**Cl 2p**	**S 2p**	**P 2p**	**Si 2p**
P-Z	82.2	8.7	0.4	7.6	0.3	0.1	0.8
m-Z THC_0	82.7	9.3	0.3	7.1	0.4	-	0.1
m-Z THC	84.1	9.1	0.3	6.0	0.3	0.1	0.2
m-Z CBD_0	82.4	9.9	0.2	6.8	0.4	0.1	0.2
m-Z CBD	83.9	9.4	0.4	5.8	0.2	0.1	0.1

Table S6. The peak-fitting results of high-resolution C 1s signal of materials.

Samples	Assignment	E_B (eV)	FWHM (eV)	At. %
Pristine	**C1s** $_{C=C\ aromatic}$	284.4	0.6	16.2
	C1s $_{C-C,\ C-H}$	285.0	1.3	57.6
	C1s $_{COH,\ C-O-C,\ C-Cl}$	286.5	1.3	22.0
	C1s $_{O-C=O}$	289.0	1.3	4.2
m-Z THC_0	**C1s** $_{C=C\ aromatic}$	284.3	0.6	14.4
	C1s $_{C-C,\ C-H}$	285.0	1.3	58.9
	C1s $_{COH,\ C-O-C,\ C-Cl}$	286.5	1.3	22.5
	C1s $_{O-C=O}$	289.0	1.1	4.3
m-Z THC	**C1s** $_{C=C\ aromatic}$	284.4	0.7	17.0
	C1s $_{C-C,\ C-H}$	285.0	1.3	58.5
	C1s $_{COH,\ C-O-C,\ C-Cl}$	286.5	1.3	21.3
	C1s $_{O-C=O}$	289.0	1.1	3.2
m-Z CBD_0	**C1s** $_{C=C\ aromatic}$	284.3	0.6	12.9
	C1s $_{C-C,\ C-H}$	285.0	1.3	60.5
	C1s $_{COH,\ C-O-C,\ C-Cl}$	286.5	1.3	22.4
	C1s $_{O-C=O}$	289.0	1.1	4.2
m-Z CBD	**C1s** $_{C=C\ aromatic}$	284.4	0.6	14.7
	C1s $_{C-C,\ C-H}$	285.0	1.3	61
	C1s $_{COH,\ C-O-C,\ C-Cl}$	286.5	1.3	20.6
	C1s $_{O-C=O}$	289.1	1.1	3.7

Table S7. The peak-fitting results of high-resolution O 1s signal of materials.

Samples	Assignment	E_B (eV)	FWHM (eV)	At. %
Pristine	**O1s** $_{C=O}$	532.5	1.5	68.3
	O1s $_{O^*\text{-}(C=O)\text{-}C,\ C\text{-}O\ aromatic}$	533.5	1.6	31.7
m-Z THC_0	**O1s** $_{C=O}$	532.1	1.6	59.8
	O1s $_{O^*\text{-}(C=O)\text{-}C,\ C\text{-}O\ aromatic}$	533.5	1.6	40.2
m-Z THC	**O1s** $_{C=O}$	532.3	1.8	63.1
	O1s $_{O^*\text{-}(C=O)\text{-}C,\ C\text{-}Oaromatic}$	533.5	1.8	36.9
m-Z CBD_0	**O1s** $_{C=O}$	532.1	1.6	59.8
	O1s $_{O^*\text{-}(C=O)\text{-}C,\ C\text{-}Oaromatic}$	533.5	1.6	40.2
m-Z CBD	**O1s** $_{C=O}$	532.2	1.7	54.4
	O1s $_{O^*\text{-}(C=O)\text{-}C,\ C\text{-}Oaromatic}$	533.4	1.7	45.6

References

1 Y. Sheng, W. Qian, J. Huang, B. Wu, J. Yang, T. Xue, Y. Ge and Y. Wen, *Microchimica Acta*, 2019, **186**, 543.

2 L. Xu, R. Wu, X. Zhu, X. Wang, X. Geng, Y. Xiong, T. Chen, Y. Wen and S. Ai, , DOI:10.1039/d1ay01261d.

3 F. F. Gonzalez-Navarro, M. Stilianova-Stoytcheva, L. Renteria-Gutierrez, L. A. Belanche-Muñoz, B. L. Flores-Rios and J. E. Ibarra-Esquer, *Sensors (Switzerland)*, 2016, **16**, 1–13.

4 J. Massah and K. Asefpour Vakilian, *Biosystems Engineering*, 2019, **177**, 49–58.

5 X. Zhu, P. Liu, Y. Ge, R. Wu, T. Xue, Y. Sheng, S. Ai, K. Tang and Y. Wen, *Journal of Electroanalytical Chemistry*, 2020, **862**, 113940.

6 Y. Rong, A. V Padron, K. J. Hagerty, N. Nelson, S. Chi, N. O. Keyhani, J. Katz, S. P. A. Datta, C. Gomes and E. S. Mclamore, *Analyst*, , DOI:10.1039/c8an00065d.

7 S. Ali, A. Hassan, G. Hassan, C. H. Eun, J. Bae, C. H. Lee and I. J. Kim, *Scientific Reports*, 2018, **8**, 1–11.

8 M. Li, P. Zhou, X. Wang, Y. Wen, L. Xu, J. Hu, Z. Huang and M. Li, *Computers and Electronics in Agriculture*, 2021, **191**, 106502.

9 R. Belugina, E. Karpushchenko, A. Sleptsov, V. Protoshchak, A. Legin and D. Kirsanov, *Talanta*, 2021, **234**, 122696.

10 D. Ortiz-Aguayo, K. De Wael and M. Del Valle, , DOI:10.1016/j.jelechem.2021.115770.

11 Y. Zhu, T. Xue, Y. Sheng, J. Xu, X. Zhu, W. Li, X. Lu, L. Rao and Y. Wen, *Microchemical Journal*, 2021, **170**, 106697.

12 Z. Rao, B. Guo, J. Zu, W. Zheng, Y. Xu and Y. Yang, *IEEE Sens J*, 2024, **24**, 7463–7472.

13 A. Carrio, C. Sampedro, J. L. Sanchez-Lopez, M. Pimienta and P. Campoy, *Sensors (Switzerland)*, 2015, **15**, 29569–29593.

14 R. Deidda, F. Coppey, D. Damergi, C. Schelling, L. Coïc, J. L. Veuthey, P. Y. Sacré, C. De Bleye, P. Hubert, P. Esseiva and É. Ziemons, *Journal of Pharmaceutical and Biomedical Analysis*, , DOI:10.1016/j.jpba.2021.114150.

15 H. Stevenson, A. Bacon, K. M. Joseph, W. R. W. Gwandaru, A. Bhide, D. Sankhala, V. N. Dhamu and S. Prasad, *Scientific Reports*, 2019, **9**, 1–11.

16 S. I. Hwang, N. G. Franconi, M. A. Rothfuss, K. N. Bocan, L. Bian, D. L. White, S. C. Burkert, R. W. Euler, B. J. Sopher, M. L. Vinay, E. Sejdic and A. Star, 2021, **16**, 11.