VOL. 17, NO. 11, JUNE 2022 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ELECTROPHORETIC REGISTERS: SIGNAL BRAIN CLASSIFICATION USING SUPPORT VECTOR MACHINE AND WAVELET IN MENINGITIS CONTAMINATED RATS

Luis Enrique Mendoza, Gonzalo Moreno and César Peña
University of Pamplona, Pamplona, Colombia
E-Mail: gmoren@hotmail.com

ABSTRACT

This paper proposes a new method for the classification of Capillary Electrophoretic Registers (CER) retrieved from the cerebrospinal fluid samples taken from meningitis contaminated rats. The proposed approach applies several signal processing tools such as wavelet analysis, dynamic programming, principal component analysis, and support vector machines (SVM) for data pre-processing, feature extraction. Furthermore, an algorithm is developed that detects zones in the Capillary Electrophoretic Registers (CER), where local energy variations between study groups are observed. This algorithm helps us to identify the effects that Klebsiella Pneumoniae bacteria produce in certain substances that are part of the cerebrospinal fluid samples. It is shown that Meningitis disease can be effectively detected with the proposed methods. Furthermore, we show that exploiting the information related to the local energy variation improves the classification correctness rate up to 97.3%. This classification performance is obtained using least square SVM as classification tools and the parameterized CER representation proposed in this paper.

Keywords: meningitis, capillary electrophoretic registers (CER), classification, support vector machine (SVM), signal processing, signal brain, rats.

Nomenclature

a	time location
b	dot product into w and support vectors obtained
b_1	scale or wideness associated with the function ψ
CER	Capillary Electrophoretic Registers
$C_{s,b}$	degree of similarity of the CER
d	integer value that sets the penalty for space insertion and the Score (A_i, B_i) function defines the value of the weight given to the alignment of the i element of the A sequence with j element of the B sequence
DPT	dynamic programming technique
$f(t)$	CER
$h(x)$	hyperplanes
KP	Klebsiella Pneumoniae
$k\langle x_i,x\rangle$	kernel function
	LS-SVM Least squares - Support vector machine
N	number of individuals used for training
	n and m A and B sequence lengths, respectively
PCA	principal component analysis
q	window scale factor
r	shift factor
ROI	region of interest
S	recursive array for dynamic programming
SVM	Support vector machine
t	time
U_h	eigenvalues of the covariance matrix
w	vector perpendicular to the classification hyperplane
WA	wavelet analysis
x_i	M-dimensional vector that contains the feature of the i-th individual
$x(t)$	original signal using wavelet analysis
y_i	class to which each individual belongs, {+1, -1}
Z_h	T-dimensional vector that represents the principal Components
α_i	i-th Lagrange coefficient
φ	represents the observation window
ψ	mother wavelet
ζ	variables vector, that is token about the 4th-levelapproximation coefficients (wavelet domain) after the sequence has been aligned

1. INTRODUCTION

Meningitis disease, also known as spinal meningitis, is an infection of the spinal cord and brain surrounding fluids that can be caused by a bacteria or viral infection [1, 2, 3]. This disease commonly occurs in children under five years old [4], making its fast diagnostic to prevent irreversible problems in the future a research challenge. Fever, headache, and loss of movement, among others, are the most common symptoms of this disease [5, 6]. In some cases, due to patient age, this disease does not appear or is not easily detected with a simple visual study performed by a specialist [7, 8]. Significant research efforts have been devoted to addressing the early detection of this disease using CER since it has been shown that the presence of the Klebsiella Pneumoniae (KP) bacteria produces variation in the concentration of certain substances (amino acids) [9]. Thus, by analysis, the changes in specific peaks of an electropherogram [10], the specialist can decide whether or not the meningitis virus has contaminated the sample in the study. These studies rely on a visual inspection and specialist experience, and, therefore, it is subject to human errors. It becomes necessary to develop automatic algorithms that can quickly detect the meningitis virus in their different phases, analyzing a CER. However, one of the challenges in processing CER is the variability

inherently observed in the data. More precisely, in a CER, each peak is associated with chemical substances, and its height represents the amount of concentration of the corresponding material. The time location of each peak in a CER is affected by the migration time shifts, dead zone, and variation of electropherogram duration [11, 12, 13], making it difficult to locate the same substance in multiple CER. In this work, an LS-SVM based classification approach is proposed that allows a quick classification of patients based on CER analysis. The proposed method shows that the meningitis virus can, indeed, cause changes in the amino acids presents in the tested sample. Furthermore, the proposed approach can also be used to detect zones in CER where exist variations of significant substances concentration between study groups.

The proposed approach consists of several stages. First, a low-resolution representation of the CER is obtained by applying a wavelet analysis (WA) on the raw data. Furthermore, in the wavelet domain, a hard-thresholding operator is applied to the detail wavelet coefficients to reduce the high-frequency noise and identify the not-active zone of each CER. Second, we extend the work recently reported in Ceballos, Paredes, and Hernandez [14] to align the CER. This alignment was used to overcome the problems caused by the 'migration time variation' of each peak in CER. Third, principal component analysis (PCA) is used as a dimensional reduction technique to represent the CER with just a few parameters, where the number of the parameter used is tailored to minimize the classification error. Finally, a stage for extraction of relevant zones or zones of o interest CER was implemented. In this last stage, we show how particular speaks of each CER in the meningitis group are modified in amplitude due to the presence of the bacteria KP. To identify these zones, an algorithm, termed multi-energy analysis, is implemented. The paper is organized as follows. Section II gives a brief overview of SVM. In Section III, we explain the methodology used to select the feature vector. Section IV presents the results and analysis. Finally, some concluding remarks are given in Section V.

2. MATERIALS AND SUPPORT VECTOR MACHINE

CER used in this work was obtained from a database gated at the Laboratory of Behavioral Physiology at the Department of Physiology in the Andes University (Venezuela) consisting of 57 registers, one for each rat in the study. Additionally, 9 out of the 57 registers were used as ¨blind registers¨. These registers have the characteristic were not known to that group belong. In general, in all the CER, exist two types the zones (active and not-active zone): how is observed in Figure-1. Here, the *x-label* and y-label represent the duration and amplitude or concentration, respectively, of the CER.

To develop the classification process, we use the broadly known classification/regression tool named SVM, in particular the most recent version, the least-squares based SVM (LS-SVM) [15-19]. This tool is chosen because it copes with the problem of CER; the like is known as low reproducibility. The problem of low reproducibility in CER has produced for instrument system proper to high sensibility a different variable like, for example, temperature. LS-SVM and SVM obtain a classification throughout hyperplanes, which linearly divides two classes. This is denoted by *h(x)*; the SVM and LS-SVM techniques can be written as follows, Eq. (1):

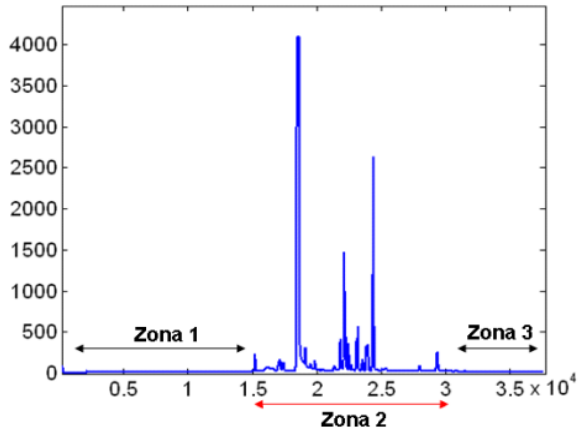

Figure-1. Example of a CER. Source Author

$$h\ (x) = \sum_{i=1}^{N} \alpha_i\ y_i\ k\left\langle x_i, x\right\rangle + b \tag{1}$$

Where y_i indicates the class to which each individual belongs, the control group is labeled as +1, whereas the -1 is the contaminated group [16].

3. METHODS

We use several signal processing techniques to find the most relevant parameterization (feature vector). Algorithm Wavelet analysis [20, 21], is initially used to find the region of interest (ROI) (eliminating zones that have not relevance in every CER), mitigate noise and compress the signal. Mathematically, wavelet transform is defined as [20]:

$$C_{s,b} = \frac{1}{b_1} \int_{-\infty}^{\infty} f(t) * \psi \left(\frac{t-a}{b_1} \right) dt \tag{2}$$

In this study, we use asymmetrical wavelet (Symlet4) as mother wavelet following the results reported in [14, 20] and 7th level decomposition for the location of the region of interest (ROI) on each CER. This region is also obtained with the help of a threshold (1% of the maximum coefficient at level 7 decomposition). A threshold was chosen the avoid losing any critical peak or substance in the registers. The region of interest is considered starting at the instant when a first detailed coefficient exceeds the threshold value and ends when the last coefficient of this level is lower than the threshold. Denoising and signal compression (resolution decrease) are also done using a symlet4 mother wavelet but with 4th-level decomposition. We noticed that this is the most

appropriate level for denoising since higher level decomposition, the CER loses relevant characteristics. To solve the problem of migration time variation, the dynamic programming technique (DPT) in Zifan, Saberi, Moradi, and Towhidkhah [22] was chosen. Mathematically the DPT or Nedleman and Wusch algorithm is represented as follows:

Initial conditions

$$S(0,0)=0$$
$$S(i,0)=id \quad 1\le i\le n$$
$$S(0,j)=jd \quad 1\le j\le m \tag{3}$$

Filling array

$$S(i,j)=\max\begin{cases}S(i-1,j-1)+Score\ \ (A_i,B_i)\\ S(i,j-1)+d\\ S(i-1,j)+d\end{cases}$$

Figure-2, shows how the alignment is improved between two registers using the DPT

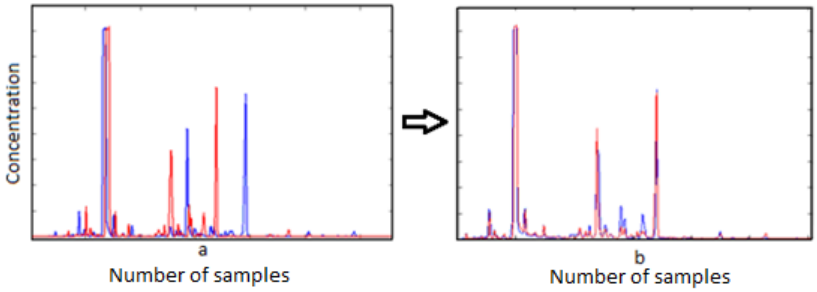

Figure-2. CER, a) non-aligned register and b) aligned register with DPT. Source Author

Next, the parameterization stage, PCA technique in López [23] is used, making the process of training and classification faster and efficient. PCA is mathematically defined as:

$$Z_h=\zeta u_h \tag{4}$$

Dimension reduction is achieved by selecting up to 32 principal components, which represent around 95% of the total energy.

Next, Multi-energy analysis is proposed as a technique to recognize areas in the CER where there are substances concentrations changes between registers coming from contaminated and control groups. The algorithm it is capable of detecting the CER zone site where the most significant energy variability between the groups is located. Mathematically the energy function is defined as follows:

$$e(r,q)=\int_{-\infty}^{\infty}\left[x(t)*\varphi\left(\frac{t-r}{q}\right)\right]^2 dt \tag{5}$$

Where, x(t) represents the parameterized signal using wavelet analysis, the φ represents the observation window, r is the shift factor, and q is the window scale factor.

4. RESULT

In this work to training and validation sets are randomly selected from the database. The final classification error percentage was calculated by averaging the rates of correct classification obtained in each iteration. The average was on fifty repetitions.

Figure-3 shows the evolution of the algorithm's classification error as the described pre-conditioning and parameterization techniques are incorporated in the definition of the feature vector. Note that the classification performance tends to improve as a new technique is added. Different numbers of principal components were used; Figure-4(a) and Figure-4(b) display the results with 25 and 32 principal components, respectively.

Where SC, ERC, ZI, AL, and ACP denote the classification error using: raw data (original electrophoretic registers), original+denoising registers, original + denoising + ROI registers, original+denoising + ROI + SA registers, and original + denoising + ROI + SA + DR registers, where ROI stands for a region of interest, SA stand for sequence alignment and DR stand for dimension reduction. Besides, observe that when we add more processing techniques, the correct classification percentage increased, also noted that better percentage classification is founded when we have all join methods. In this way, we present techniques series for processing and classification of the CER. Figure 3 shows multi-energy analysis results.

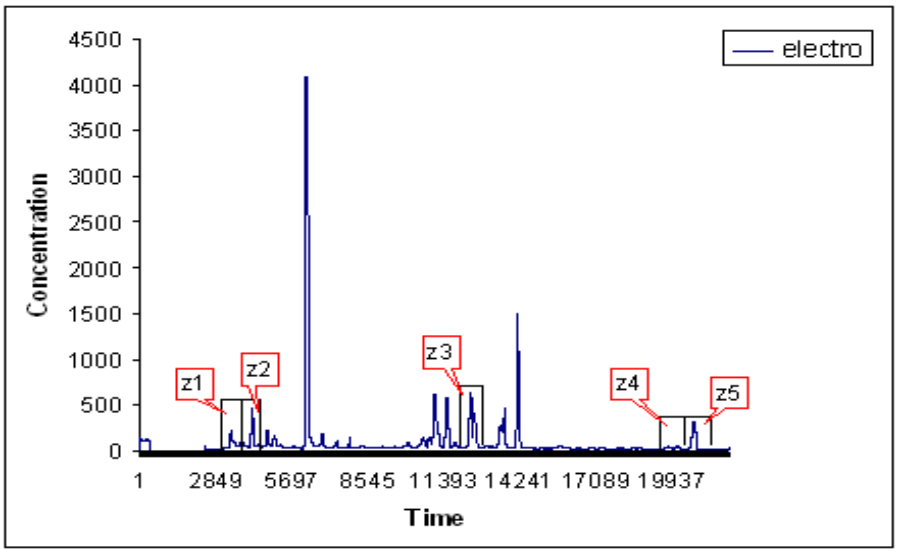

Figure-3. Multi-energy analysis, detected difference zones, zi. Source Author

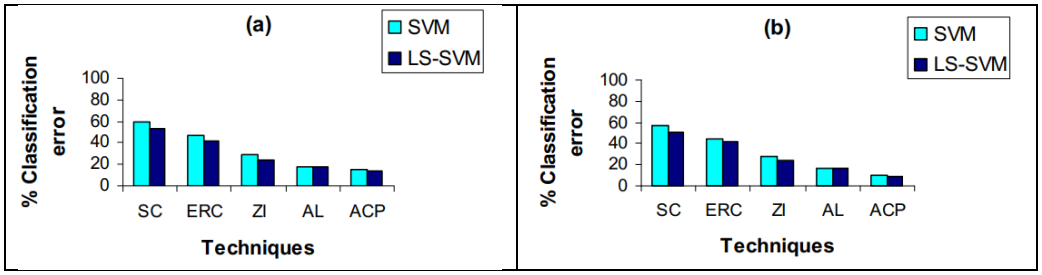

Figure-4. Classification error, a) 25 y b) 32 principal components. Source Author

The algorithm identified five different zones (z1, z2, z3, z4, and z5), where there are significant changes in energy between registers coming from control and meningitis-contaminated rats. Additionally, a t-student test was carried out to validate the significance of these changes. Table-1 shows the results. Note that if the number of 'difference zones' is increased, the study has an improved statistical difference (t-value decreases), which indicates that the areas in each group present a significant difference of concentration. Note also that if increased the rat's number and the different zones number to found, the value of the t-student distribution decreased.

Table-2 shows a summary of the percentage of classification error obtained using PCA, Multi-energy, the region of interest and PCA+IA as training and classification patterns. Note that the best result was found when PCA components and IA and LS-SVM classifier were used for feature extraction and classification, respectively.

Finally, it shows the execution times in seconds for the data choose. PCA+IA with SVM time 0.512 and PCA+IA with LS-SVM time 0.129, which with the above results, leads to accepting them as the best technique for these studies

Table-1. T-student distribution test.

Difference zones	5 Rats	10 Rats	15 Rats
2	0,000500000	0,000450000	0,000008950
3	0,000085000	0,000056320	0,000000045
5	0,000003650	0,000000789	0,000000003

Table-2. Average classification error percentage (50 iterations).

	Training Patterns			
Classification Algorithm	**PCA (32)**	**Multi-energy**	**Interest areas, IA**	**PCA + IA**
SVM	10,4%	12,9%	11,5%	6,3%
LS-SVM	2,7%	10,8%	5,5%	2,7%

5. CONCLUSIONS

The proposed algorithms showed to be capable of processing and classifying electrophoretic registers to detect the presence of meningitis virus with a percentage of correct classification of about 97.3%. The PCA technique allows the training and classification process to be faster and more efficient. The algorithm it is capable of detecting the CER zone site where the most significant energy variability between the groups are located.The proposed approach can be used as an assistant tool for quick and early diagnosis of this disease presence; This new tool can also serve as a specialist support to diagnose and quickly develop a preventive measure to reduce the pathological consequences that meningitis or another disease that cause alterations in the CER involves. The multi-energy searching algorithm of different zones also shows to be an excellent tool to find when specific substances in the register have more variations of concentration if the virus is present. So, specialists can focus their attention only on these areas, improving their timely response.

REFERENCES

[1] Brandt C. T., Simonsen H., Liptrot M., Søgaard L. V., Lundgren J. D., Østergaard C. & Rowland I. J. 2008. In vivo study of experimental pneumococcal meningitis using magnetic resonance imaging. BMC medical imaging. 8(1): 1.

[2] Correa C., Troncone A., Rodríguez L., Carreño M., Bedoya C. & Narváez R. 2003. Management guidelines for bacterial meningitis in children: Overview. AVPP, 66(Supl 3): 2-7.

[3] Kalita J., Singh R. K., Misra U. K. & Kumar S. 2018. Evaluation of cerebral arterial and venous system in tuberculous meningitis. Journal of Neuroradiology, 45(2): 130-135. http://doi.org/10.1016/j.neurad.2017.09.005

[4] Maurer P., Hoffman E., Mast H. 2009. Bacterial meningitis after touch extraction. British Dental. pp. 69-71.

[5] Lin Y. H., Chang Y. W., Yang S. H., Chang H. H., Lu M. Y. & Fan P. C. 2015. Gliomatosis cerebri with spinal metastasis presenting with chronic meningitis in two boys. Journal of the Formosan Medical Association, 114(9): 886-890. http://doi.org/10.1016/j.jfma.2012.10.024

[6] Misra U. K., Kumar M. & Kalita J. 2018. Seizures in tuberculous meningitis. Epilepsy Research, 148(October): 90-95. http://doi.org/10.1016/j.eplepsyres.2018.10.005

[7] Schlenk F., Frieler K., Nagel A., Vajkoczy P. & Sarrafzadeh A. S. 2009. Cerebral microdialysis for detection of bacterial meningitis in aneurysmal subarachnoid hemorrhage patients: a cohort study. Critical Care. 13(1): R2.

[8] Lucas M. J., Brouwer M. C. & van de Beek D. 2016. Neurological sequelae of bacterial meningitis. Journal of Infection, 73(1): 18-27. http://doi.org/10.1016/j.jinf.2016.04.009

[9] Mally J., Baranyi M. & Vizi E. S. 1996. Change in the concentrations of amino acids in CSF and serum of patients with essential tremor. Journal of neural transmission. 103(5): 555-560.

[10] Giridharan V. V., Simões L. R., Dagostin V. S., Generoso J. S., Rezin G. T., Florentino D., Barichello T. 2017. Temporal changes of oxidative stress markers in Escherichia coli K1-induced experimental meningitis in a neonatal rat model. Neuroscience Letters, 653, 288-295. http://doi.org/10.1016/j.neulet.2017.06.002

[11] La S., Cho J., Kim J. H. & Kim K. R. 2003. Capillary electrophoretic profiling and pattern recognition analysis of urinary nucleosides from thyroid cancer patients. Analytica chimica acta. 486(2): 171-182.

[12] Shihabi Z. K. & Hinsdale M. E. 1995. Some variables affecting reproducibility in capillary electrophoresis. Electrophoresis. 16(1): 2159-2163.

[13] Schaeper J. P. & Sepaniak M. J. 2000. Parameters affecting reproducibility in capillary electrophoresis. Electrophoresis: An International Journal. 21(7): 1421-1429.

[14] Ceballos G., Paredes J. L. & Hernandez L. F. 2007. A novel approach for pattern recognition in capillary electrophoresis data. In IV Latin American Congress on Biomedical Engineering, Bioengineering Solutions for Latin America Health (pp. 150-153). Springer, Berlin, Heidelberg.

[15] Su C. T. & Yang C. H. 2008. Feature selection for the SVM: An application to hypertension diagnosis. Expert Systems with Applications. 34(1): 754-763.

[16] Huang C. L. & Wang C. J. 2006. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with applications. 31(2): 231-240.

[17] Mujahid Khan A., Fayyaz M. & Gilani S. A. M. 2017. Thermal and Visible Image Fusion: a Machine Learning Approach. International Journal on Media Technology. 1(2).

[18] Marrugo N., Amaya D. & Ramos O. 2018. Comparison of Multi-Class Methods of Features Extraction and Classification to Recognize EEGs Related with the Imagination of Two Vowels. International Journal on Communications Antenna and Propagation (IRECAP). v 8, n 5. https://doi.org/10.15866/irecap.v8i5.12709

[19] Moloi K., Jordaan J. & Hamam Y. 2020. The Development of a High Impedance Fault Diagnostic Scheme on Power Distribution Network. International Review of Electrical Engineering (IREE). v 15, n1. https://doi.org/10.15866/iree.v15i1.17074

[20] Guo X., Sun L., Li G. & Wang S. 2008. A hybrid wavelet analysis and support vector machines in forecasting development of manufacturing. Expert Systems with Applications. 35(1-2): 415-422. v 1, n 2.

[21] Mallat S. 1999. A wavelet tour of signal processing. Elsevier. pp. 220-320.

[22] Zifan A., Saberi S., Moradi M. H. & Towhidkhah F. 2007. Automated ECG Segmentation Using Piecewise Derivative Dynamic Time Warping.

[23] López C. P. 2004. Multivariate data analysis techniques. Pearson Educación. pp. 672-700.