Classification of electrophoretic registers from meningitis contaminated rats

Luis E Mendoza[#1], Jose Luis Paredes[*2], Oscar E Gualdron[#3]

[#] Biomedical Engineering Research Group, University of Pamplona, ciudadela Universitaria, Pamplona –Colombia
[1]luís.mendoza@unipamplona.edu.co
[3]oegualdron@gmail.com
[*]Biomedical Engineering Research Group, University of the Andes, Merida Venezuela
[2]lparede@ula.ve

Abstract. This paper proposes a new method for classification of Capillary Electrophoretic Registers (CER) retrieved from cerebrospinal fluid sample taken from meningitis contaminated rats. The proposed approach applies several signal processing tools such as, wavelet analysis (WA), dynamic programming, principal component analysis (PCA) and support vector machines (SVM), for data pre-processing, feature extraction and CER classification. Furthermore, an algorithm is developed that detects zones in the CER where local energy variations between study groups (meningitis group and control group) are observed. This algorithm help us to identify the effects that Kliebsella Pneumonie (KP) bacteria produce in certain substances (aminoacids) that are part of the cerebrospinal fluid samples. It is shown that Meningitis disease can be effectively detected, analyzing the CER with the proposed methods. Futhermore, 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 (LS-SVM) as classification tools and the parameterized CER representation proposed in this paper.

Keywords: CER, Classification, LS-SVM, Signal processing.

I. INTRODUCTION

Meningitis disease also known as spinal meningitis is an infection of the spinal corn and brain surrounding fluids that can be caused by a bacteria or viral infection [1], [2]. This disease commonly occurs in children under 5 years old [3], making its fast diagnostic to prevent irreparable problems in the future a research challenge. Fever, headache and loss of movement, among others are the most common symptoms of this disease. 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 [4]. Great research efforts have been devoted to address the early detection of this disease using CER since it has been show that the presence of the *KP* bacteria produces variation in the concentration of certain substances (aminoacids) [5]. Thus, by analysis the changes in certain peaks of an electropherogram, the specialist can decide whether or not the sample in study has been contaminated by the meningitis virus. These studies rely on 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 analysing 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 a chemical substances and its height represents the amount of concentration of the corresponding substance. The time location of each peak in a CER is affected by the migration time shifts, dead zone and variation of electropherogram duration [6], [7],[8], making difficult to locate the same substance in multiple CER. In this work, a LS-SVM based classification approach is proposed that allows a quick classification of patients based on CER analysis. The proposed approach shows that the meningitis virus can, indeed, cause changes of the aminoacids presents in tested sample. Furthermore, the proposed approach can also be used to detect zones in CER where exist changes 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 appling a 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 [9] to align the CER. This alignment was used to overcome the problems caused by 'migration time variation' of each peak in CER. Third, PCA is used as a dimensional reduction technique to represent the CER with just a few parameters, where the number of 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 certain speaks of each CER in the meningitis group, are modified in amplitude, due to the presence of the bacteria *KP*. In order to identify these zones, an algorithm, termed multi-energy analysis, is implemented. The paper is organized as follows. Section 2 gives a brief overview about SVM. In Section 3, we explain the methodology used to select the feature vector. Section 4 presents the results and analysis. Finally, some concluding remarks are given in Section 5.

II. MATERIALS AND SUPPORT VECTOR MACHINE

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

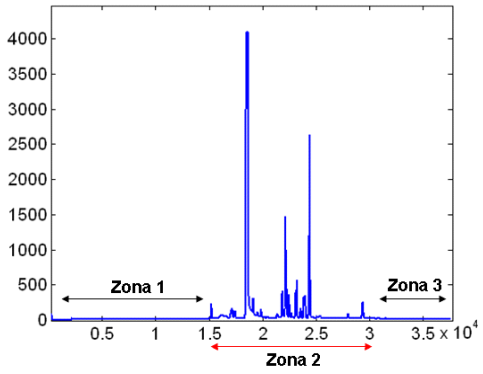

Figure 1 - Example of a CER. Source Author

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) [10], [11]. 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 is 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 [11]:

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

Where y_i belongs to {+1,-1} and indicates the class to which each individual belongs to for the application at hand, the control group is labelled as +1, whereas the -1 is the contamined group $\boldsymbol{x}_i$ is an M-dimensional vector that contains the feature of the *i-th* individual α_i is the *i-th* Lagrange coefficient, *N* is the number of individuals used for training, and *k(.,.)* denotes the kernel function. Finally, *b* is dot product into *w* and support vectors obtained, here *w* is a vector perpendicular to the classification hyperplane [11].

III. METHODS

We use severals signal processing techniques, to find the most relevant parameterization (feature vector). Algorithm Wavelet analysis [12], [13], 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 [13]:

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

Where *f(t)* is the CER and $\psi(.)$ is the mother wavelet. a is the time location and b is the scale or wideness associated with the function $\psi(.)$. $C_{s,b}$ presents the degree of similarity of the CER with the signal $\psi(.)$. In this study, we use a symmetrical wavelet (*Symlet4*) as mother wavelet following the results reported in [11], [8] ,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). Threshold was chosen with the aim of avoiding lose any important 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 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. For solve the problem of migration time variation the dynamic programming technique, DPT [14] was choose. Mathematically the DPT or Nedleman and Wusch algorithm is represented mathematically as follows:

Initial conditions **Filling array**

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

$$S(i,j) = \max\begin{cases} S(i-1, j-1) + Score(A_i, B_i) \\ S(i, j-1) + d \\ S(i-1, j) + d \end{cases} \tag{3}$$

Where *S*, is the recursive array for dynamic programming, *d* is an 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, and finally *n* and *m* represent the A and B sequence lengths, respectively. Fig. 2, shows how the alignment is improved between two registers using the DPT.

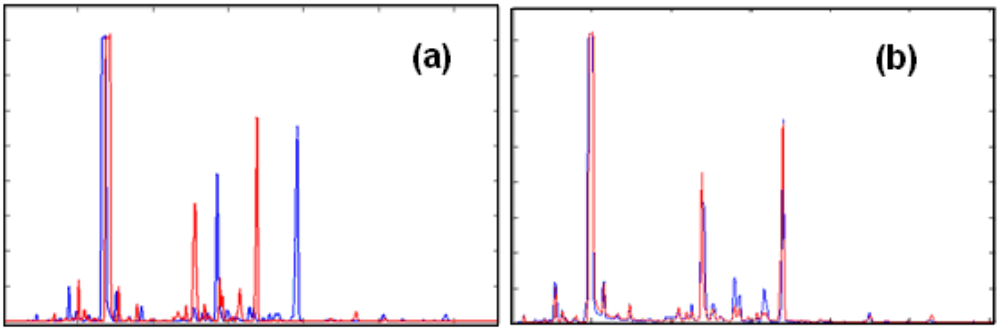

Figure 2 - CER, a) non-aligned register and b) aligned register with DPT. Source Author

Next, the parameterization stage, PCA technique [15] is used, making the process of training and classification faster and efficiency. PCA is mathematically defined as:

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

Where Z_h is a *T*-dimensional vector that represents the principal components, ζ is the variables vector, that is token about the 4th-level approximation coefficients (wavelet domain) after the sequence has been aligned, u_h are eigenvalues of the covariance matrix. *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 to detect the CER zone site where the greatest energy variability between the groups are 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 $\varphi(.)$ represents the observation window, *r* is the shift factor, and *q* is the window scale factor. $\varphi(t)$ is a function simple step with finite length τ , and likes *r* and *q*, it takes positive integer values.

IV. RESULTS

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

Fig. 4, 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; Fig. 3(a) and fig. 3(b) display the results with 25 and 32 principal components, respectively.

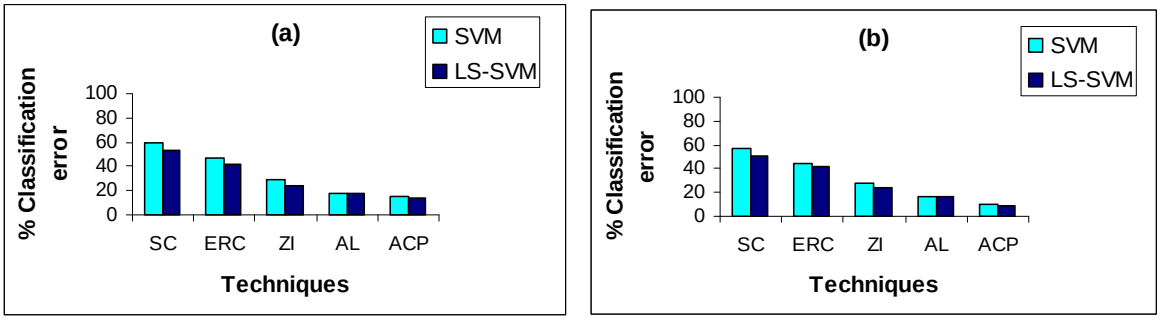

Figure 3 - Classification error, a) 25 y b) 32 principal components**.** Source Author

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, respect-tively where ROI stand for region of interest, SA stand for sequence alignment and DR stand for dimension reduction. Beside, observe that when we add more processing techniques the correct classification percentage increased, also observed that better percentage classification is find when have all join techniques. In this way, we present to techniques series for processing and classification the CER. The Fig. 4, shows *multienergy analysis* results.

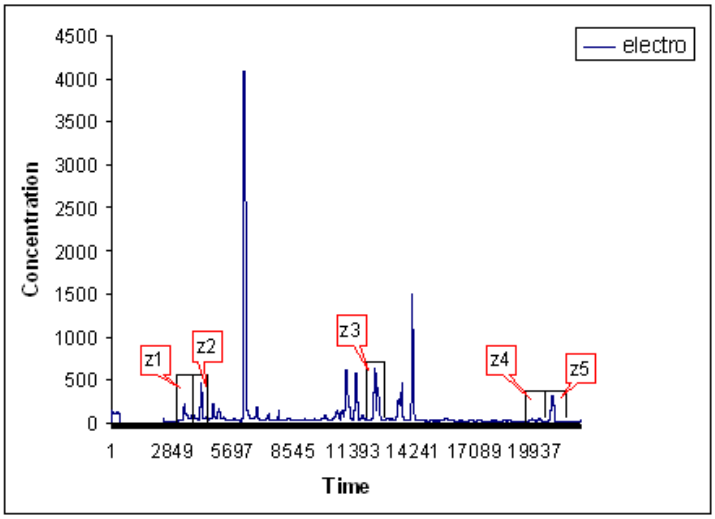

Figure 4 - Multi-energy analysis, detected *difference zones*, z_i**..** Source Authors

The algorithm identified five *difference zones* (z1, z2, z3, z4 and z5) where there are significant changes of 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 have a improve statistical difference (*t-value* decreases), which indicates that the areas in each group presents great difference of concentration. Note also, that if increased the rats number and the different zones number to found the value the *t-student* distribution decreased.

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, shows a summary of the percentage of classification error obtained using PCA, Multi-energy, 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.

TABLE 2. Average Classification Error Percentage (50 iterations)

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

Finally, 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 choose them as the best technique for these studies.

V. DISCUSSION AND CONCLUSIONS

The proposed algorithms showed to be capable of processing and classifying electrophoretic registers to detect presence of meningitis virus with a percentage of correct classification about 97.3%. 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. Multi-energy searching algorithm of *difference zones* also shows to be a good 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 time response.

REFERENCES

[1] C. Brandt, H. Simonsen, M. Liptrot, L. Segard, J. Lundgren, C. Ostergaard, N. Frimodt, and I. Rowland. "In vivo study experimental pneumococcal meningitis using magnetics resonance imaging," BMC medical imaging, pp. 1-11, 2008.
[2] C.T. Correa, "Pautas del manejo de la meningitis bacteriana en niños," (me falta acomodar info) vol 66, 2003.
[3] P. Maurer, E. Hoffman, H. Mast. "Bacterial meningitis after touch extraction," British Dental, pp. 69-71, Jan 2009.
[4] F. Schlenk, K. Frieler, A. Nagel, P. Vajkoczy, and A.S. Sarafzadeh. "Cerebral microdialysis for detection of bacterial meningitis in aneurismal subarachniod hemorrhage patients: A cohort study," Critical care, vol. 13, Jan 2008.
[5] Málly J, Baranyi M, Vizi ES. Change in the concentrations of amino acids in CSF and serum of patients with essential tremor. J neural transm. 1996.
[6] S. La, J. Cho, K. Han Kim, K. Rao. " Capillary electrophoretics profiling and pattern recognition analysis of urinary nucleosides from thyroid cancer patients," Analytical chemical, pp. 171-182, 2003.
[7] J. P. Schaeper, M.J. Sepaniak, "Electrophoresis," 2000, vol 21, pp. 1421-1429.
[8] Z. K. Shihabi, M. E. Hinsdale, "Electrophoresis," 1995, vol 16, pp. 2159-2163.
[9] G. Ceballos, J. Paredes and L. Hernández, "A novel approach for pattern recognition in capillary electroporesis data," Springerlink, pp. 150-153, 2008.
[10] C. Ton, C. Yang. "Feature selection for SVM: An application to hypertension diagnosis," Expert system with application, pp. 754-763, 2008.
[11] C. Lung, and C. Wang. "A GA-based feature selection and parameters optimization for support vector machine," Elsevier, pp. 231-240, 2006.
[12] X. Guo, L. Sun, G. Li, and S. Wang, "A hybrid wavelet analysis and support machine in forecasting development of manufacturing" Elservier, 2007.
[13] S. Mallat, "Wavelet bases: A wavelet tour of signal processing", book, pp. 220-320, 1999.
[14] A. Zifan, S. Saberi, M. Hassan and F. Towhidkhah, "Automated ECG segmentation using piecewise derivate dynamic time warping" International journal of biomedical sciences vol 3, 2007.
[15] C. Pérez, "Técnicas de análisis multivariantes de datos y aplicaciones con SPSS", edición 3, pp.672-700.