Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2022.00051

sigma
Journal of Engineering
and Natural Sciences
Mühendislik ve Fen
Bilimleri Dergisi

Research Article

Feature extraction for DNA capillary electrophesis signals based on discrete wavelet transform combined with multi-scale permutation entropy

Öyküm Esra YİĞİT[1,*], Ersoy ÖZ[1]

[1]Yildiz Technical University, Department of Statistics, Istanbul, Turkey

ARTICLE INFO

Article history
Received: 24 October 2020
Accepted: 09 April 2021

Keywords:
Discrete Wavelet Decomposition; DNA Capillary Electrophoresis Signal; Feature Extraction, Multiscale Permutation Entropy; Nucleic Acid Sequencing

ABSTRACT

DNA sequence classification is an important challenge in genomic studies due to non-linear and chaotic behavior of DNA oxidation signals of Adenine, Cytosine, Guanine, and Thymine bases. To achieve genotype identification of samples derived from biological sources accurately, Machine Learning (ML) methods have been commonly preferred instead of expert-based methods due to the ability in handling such these complex-structured biological sequences. Reducing the dimension without sacrificing important information that should not be omitted during the classification process is an important task in ML applications. This study presents a new feature extraction method to detect two sub-types of hepatitis nucleic acid trace files. The proposed method combines both discrete wavelet transform (DWT) and entropy. The DWT decomposes the bases signals up to three levels and thus all necessary information that is hidden in both spatial and frequency domains is aimed to captured. To achieve a good summarization of DNA trace files having different length, multi-scale permutation entropy (MPE) measures are then computed from approximate and detail coefficients of signals stored in the sub-bands. Different feature sets are extracted with the proposed method using real data covering 200 hepatitis DNA trace files and then fed to a simple memory-based learning classifier, k-NN. The classification performance of the proposed feature extraction method is compared against a method based on MPE features without wavelet decomposition. The results indicate, in classifying hepatitis DNA trace files, the average accuracy reaches up to nearly 99% with feature sets based on proposed method even at 30% training samples proportion.

Cite this article as: Yiğit ÖE, Öz E. Feature extraction for DNA capillary electrophesis signals based on discrete wavelet transform combined with multi-scale permutation entropy. Sigma J Eng Nat Sci 2022;40(3):474–489.

***Corresponding author.**
*E-mail address: oeyigit@yildiz.edu.tr

This paper was recommended for publication in revised form by Regional Editor Gülhayat Gölbaşı Şimşek

Published by Yıldız Technical University Press, İstanbul, Turkey

INTRODUCTION

In the last decades, deoxyribonucleic acid (DNA)-sequencing has become a growing research interest in the field of biological sciences. In 1977, two powerful approaches were introduced by Maxam and Gilbert [1] and Sanger, Nicklen and Coulson [2], which are based on chemical degradation and enzymatic synthesis, respectively [3]. Recently, new sequencing technologies have been developed for determining the complete or interested region of DNA sequence, and these technologies aim to serve users at affordable costs as well as produce results in a short time period. Capillary electrophoresis (CE) is a commonly-applied approach in studies of high-throughput DNA sequencing and separation [4]. Although the next generation sequencing (NGS) produce millions to billions more data than Sanger sequencing by CE within the same amount of time, the results of NGS are verified with the results obtained by CE. Besides, many institutions still prefer to use their legacy sequencer, CE, because this platform is fast, cost effective and preserve a familiar workflow. However, especially the abilities of being sensitive in detection and being cost-effective may turn into challenges in the case of huge number of subjects and hence, CE is preferred generally in small sized projects. The key principle of CE is the usage of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators which are labeled with different fluorescent dyes [5]. Base calling signals are recorded as fluorescence peaks and each peak show the nucleic acid sequence, Adenine (A), Cytosine (C), Guanine (G), and Thymine (T).

The accurate identification of virus subtypes is one of the main challenges for DNA sequencing centers. Since DNA data is indeed a 4-channel time series [6], it exhibits the complex-structured characteristics (e.g. being non-stationary and non-linear, being noisy data having outliers). For these reasons, it is often inadequate to visually distinguish the signals by a specialist and, moreover, it is prone to errors. Instead of such labor intensive methods, the original signal can be transformed to a different space where a set of features, called vector of features, is generated without losing the original information that are used in discrimination of subtypes. Changing the original signal into a useful vector of features is known as feature extraction and high classification performance is achieved with this pre-classification stage of virus subtypes recognition.

Some traditional smoothing and filtering methods used in denoising the sequential data for biological signal processing are based on Savitzky-Golay (SG), Fourier transform (FT) and Fast Fourier transform (FFT). These methods, which give efficient results with the stationary structure of the signal in the data, have a wide applicability especially in analytical chemistry. However, the fact that they are easily affected by the noise and thus the difficulty in processing non-stationary signals (having different shapes and widths) brought the need to use a transformation method that takes into account local information of the original signal in spatial domain as well as frequency. Wavelet transform (WT) which biological signals are represented in both spatial and frequency domains [7] has been effectively used in the feature extraction stage of classification [8–20]. More specifically, WT has been widely applied for denoising DNA CE signals that have overlapped peaks [21–26].

Although the obtained wavelet coefficients (WC) provide a well representation of the energy distribution in the spatial-frequency domain of the signal, a curse of dimensionality problem is caused when a high-dimensional feature space is used in the classification stage [27–29]. To carry out suitable reduction in dimension that produces classification with a high performance as well cost effective, some measures such as statistical-based [10,20,30,31], entropy-based [32–35] and both combined [36] are calculated over the set of WC and then used as a vector of features. Entropy-based features have a wide application in the studies of biological signal processing [37–43], in order to quantify the degree of disorder of an interested signal and can be effectively used to detect virus subtypes of DNA chromatograms (DNAC) [44].

In this study, a feature extraction method is proposed in the purpose of classifying the DNAC. The classification performance obtained from the proposed method is evaluated using Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV) trace files recorded from 200 hepatitis patients. The proposed method has two processing stages and these stages are executed sequentially. In the first stage, nucleic acid sequences, A, C, G and T, are decomposed into sub-bands using discrete WT (DWT). Thus, it is aimed to capture all necessary local information of DNAC which is hidden in the both spatial and frequency domains due to the artifacts. In the second stage, complexity of each of these sub-bands is quantified by multi-scale PE (MPE). Thus, it is intended to achieve a good summarization of trace files having different length and complex-structured characteristics. By executing two processing stages given above, different feature sets are generated. Kruskal-Wallis H tests are utilized to show the discrimination abilities of each generated sets. Thus, an additional computational load from redundant information is prevented. After, each feature set is used as an input to a simple classification algorithm, k-nearest neighbors (k-NN). The classification performance of the proposed feature extraction method is compared against a method based on MPE features without wavelet decomposition.

The remaining sections of this study are organized as follows. Following section includes materials and methods of the study. After, results are presented and a discussion and some concluding remarks are provided in last two sections.

MATERIAL AND METHODS

Dataset

DNA trace files are generally obtained with base-calling software. "Phred" which is the frequently-used by academic and commercial laboratories embedded in automated DNA analyzer, ABI-3730 (Applied Biosystems, Foster City, USA) is used in this study for the purpose of obtaining hepatitis DNA trace files. In total, 200 hepatitis DNA trace files are labelled as HBV (96 traces) and HCV (104 traces). The information of any trace file includes four nucleotides (bases), A, C, G and T, that has Gaussian shaped peaks. A single trace file has approximately 60,000 (4×15,000) data points and each trace has a different length from other traces. In order to obtain numerical representation of the bases of a trace, base signals are converted from SCF format to an array using "scfread" function in MATLAB 2019b software [45].

FEATURE EXTRACTION

To perform classification of HBV and HCV, extracting discriminative features from four-bases of hepatitis DNA trace files is an important stage. Following sub-sections provide a brief description of proposed feature extraction process which is based on discrete WT (DWT) and entropy measures.

Discrete wavelet transform

The WT decomposes a signal into a set of functions called wavelets. Wavelets obtained from a mother wavelet by dilating and shifting are small oscillatory waves and characterize the local information of signals in spatial-frequency domain. Continuous WT (CWT) and DWT are two types of wavelet analysis. Since the data generated by chromatography devices is discrete in nature, using DWT is recommended [46]. Wavelet is defined with the following equation:

$$\Psi(a,b,t) = \frac{1}{\left|\sqrt{a}\right|}\Psi\left(\frac{t-b}{a}\right) \tag{1}$$

where Ψ is the mother wavelet and a and b are the dilation (scale) and shift values, respectively. Using Eq. (1), wavelet coefficients are obtained. To achieve discrete transformation, the following discrete Ψ is used:

$$\Psi(m,n,t) = 2^{-\frac{m}{2}}\Psi(2^{-m}t-n) \tag{2}$$

where m and n demonstrate the scaling and location values. Wavelet coefficients are obtained an algorithm proposed by Mallat [47]. At first step, two digital filters, low-pass (LP) and high-pass (HP), are applied to the signals. Thus, DWT decomposes the original signal into two mutually orthogonal sets of wavelets, representing the low-frequency (high-scale) and high-frequency (low-scale) coefficients. The coefficients obtained from the LP filtering recorded as approximation coefficients, while HP filtering produces details coefficients. The useful information is included in the approximate part and the noise is included in the detail part [31]. The coefficients obtained from the first-level having the half frequency bandwidth of the original signal are down-sampled by a factor 2 [48]. The same procedure is repeated using the approximate coefficients of the first-level of decomposition in order to get two mutually orthogonal sets of wavelets for the second level of decomposition. At each step of the decomposition process, as the filtering and sub-sampling are applied, the frequency resolution is doubled, whereas the spatial resolution is halved.

Permutation and multi-scale permutation entropy

PE [49] quantify the complexity of time series having non-stationary, noisy and non-linear characteristics. For a given discrete time series $\{x(i)\}$ with length N, an embedding procedure is applied to the time series to generate the following m-dimensional vector:

$$X(i) = \{x(i), x(i+\tau), \ldots, x(i+(m-1)\tau)\} \tag{3}$$

where m and τ show the embedding dimension that is greater than or equal to 2 and embedding delay, respectively. The vector $X(i)$ is then arranged in an ascending order $\pi = r_0, r_1, \ldots, r_{m-1}$, where $x(i + r_0\tau) \leq x(i + r_1\tau) \leq \ldots \leq x(i + r_{m-1}\tau)$. For a given m, there are $m!$ possible order patterns π, referred as motifs. The relative frequency for each motif is expressed by $p(\pi)$ and the PE for a given embedding dimension is defined as:

$$\text{PE} = -\sum_{\{\pi\}} p(\pi)\log(p(\pi)) \tag{4}$$

In order to obtain scale-independent measure, PE is normalized by $PE/log(m!)$. The normalization ensures to get entropy values which are in the range between 0 and 1. The larger PE value is, the more complex the time series is.

Since PE is a single-scaled measure, it is not suitable for systems having structures on multiple spatial and temporal scales. Similar to the multi-scale entropy method introduced by [50], MPE [51] incorporates two different procedures. Firstly, for a given discrete time series $\{x(i)\}$ with length N, multiple coarse-grained time series $\{y^{(s)}\}$ are constructed. Each element of the coarse-grained series is obtained from the following formula

$$y_j^{(s)} = \frac{1}{s}\sum_{i=(j-1)s+1}^{js} x(i) \tag{5}$$

where $j = 1,2,\ldots, N/s$ and s is the scale parameter. Secondly, the complexity of each coarse-grained series are calculated with PE and plotted as a function of s.

CLASSIFICATION

K-NN is a simple memory-based classification algorithm which can be effectively used in both classification and estimation purposes. Since it serves as a robust classification technique for noisy time-series, the classification process of this study is carried out with k-NN. In this non-parametric method, subjects are classified based on the class of their known nearest neighbor. To determine the class, training set of the data and pre-defined neighborhood parameter (k) are required. The algorithm searches the space of training set for the k-nearest subjects based on a distance function or a similarity measure [34]. The performance of the k-NN clasifier depends on the type of the distance function and the value of k. Since Euclidean distance is the popularly used [46], the present study uses this metric. Also, the value of k is usually taken as small integer values with a positive sign and the value of k is taken from 1 to 5 in this study.

CLASSIFICATION PERFORMANCE EVALUATION

Different classification performance measures such as sensitivity (Se), specificity (Sp), accuracy (Acc) and Kappa statistic (κ) are used in evaluation. From the confusion matrix which is a useful tool for evaluating the classifier's performance, the number of correctly classified subjects (true positive-TP and true negative-TN) and their proportions (TP rate and TN rate) can be derived. TP rate and TN rate are generally known as Se and Sp, respectively. Acc gives an information about the proportion of overall correctly classified subjects (both TP and TN). κ is an agreement measure lies in the range between [−1,+1] for assessing the classification power of a classifier.

PROPOSED FEATURE EXTRACTION METHOD

The proposed method has mainly two processing stages that are executed sequentially, namely decomposition stage and feature extraction stage.

Decomposition Stage: For a trace file, firstly, signals in a single base (i.e., A, C, G or T) are pre-processed with DWT to decompose them into four sub-bands using third-level decomposition. For the first-level of decomposition, signals are passed through LP and HP filters and the approximate and detail coefficients are produced. Later, down-sampling is applied by a factor 2 according to Nyquist rule, and the approximate coefficients are fed into another LP and HP filters again. Thus, second-level coefficients are obtained. After down-sampling by a factor 2, the approximate coefficients are passed through the filters again and the third-level of coefficients are obtained.

Feature Extraction Stage: The coefficients of the sub-bands signals obtained from DWT are used to calculate entropy-based features. In the decomposition stage, it is created 4 sub-bands for a single base of a trace file. For each of these 4 sub-bands, entropy values are calculated in assessing the dynamics and complexity of signals located in these sub-bands. When the decomposition and feature extraction stages are repeated for all bases (A, C, G and T) and later for all trace files, a feature set which is then fed to any classifier can be created. Using various entropy estimators including embedding and spectral entropies, different feature sets can be generated.

Figure 1 shows an illustration of original (the first column) and decomposed signals (the second column) of a selected sample trace file from 200 hepatitis trace files. Signals coloured in green, blue, black and red demonstrate the signals of bases A, C, G and T, respectively. In the first panel of Figure 1, while the first column shows the original signals of base A, the second column gives the decomposed signals of base A into three levels using Daubechies 1 wavelet (dB1) [52]. Similarly, the second, the third and the fourth panels of Figure 1 illustrate the original and decomposed signals of base C, G and T, respectively, which belong to the same selected trace. As shown in the first columns of Figure 1, original signals in a trace span a wide range of intensities in spatial-frequency domain and contain overlapped peaks with different shapes (such as Gaussian). Smoothing or filtering is needed to such this data having non-stationary characteristics which may often represent the most important part of a signal sequence. To achieve successful separation between HBV and HCV trace files, the important information localized in both spatial and frequency domains can be derived using DWT (the decomposition stage). On the other hand, although the spatial-frequency characteristics of the signals are captured by DWT, the length of signals stored in four sub-bands of any trace file differs from the lengths of the signals in the sub-bands of other traces. The dimensionality problem occurs when a vector of features is desired to be formed used as an input for relevant classifier. To overcome this problem, each element of the feature vector should be defined in the same dimension. Entropy measures can be effectively applied to summarize the complex-structured signals with different lengths. The proposed method offers to use entropy measures for each-sub-bands of each bases and then to concatenate the calculated entropies in order to form a vector of features (feature extraction stage). Thus, a good summarization of trace files can be performed efficiently by executing two processing stages given above and a good discrimination between can be attained between the HBV and HCV DNAC.

EXPERIMENTAL SETUP

By executing two processing stages given above, different feature sets are generated. Each feature set is used as an input to k-NN classifier. The obtained classification performances using different feature sets are compared between

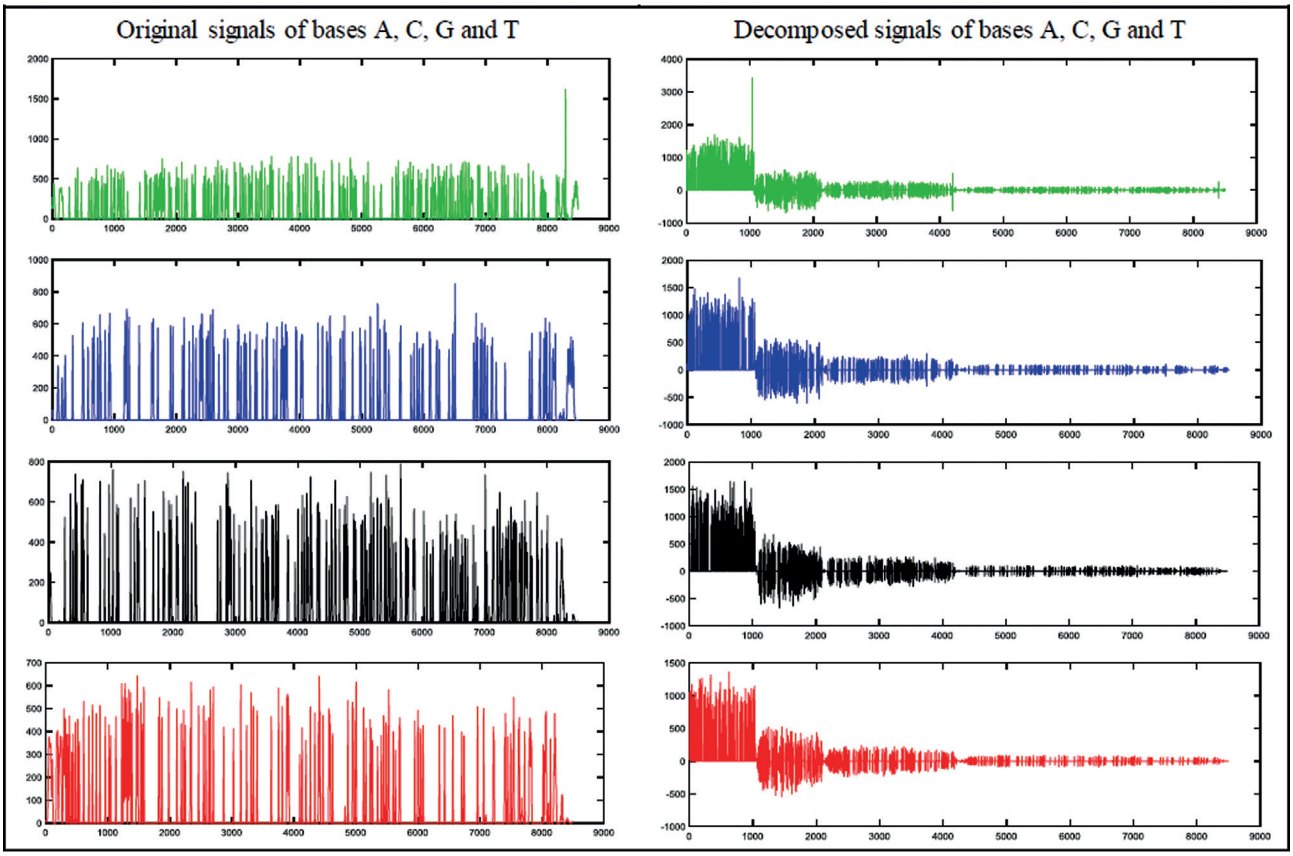

Figure 1. Original and decomposed signals of a trace file.

Table 1. Description of feature sets extracted in Step 1

Base	**Feature Set**					
	PE$_{(A,C,G,T)}$		**MPE2**$_{(A,C,G,T)}$		**MPE3**$_{(A,C,G,T)}$	
	Feature Name	**Description**	**Feature Name**	**Description**	**Feature Name**	**Description**
A	$PE_{(A)}$	PE of A	$MPE2_{(A)}$	MPE with s=2 of A	$MPE3_{(A)}$	MPE with s=3 of A
C	$PE_{(C)}$	PE of C	$MPE2_{(C)}$	MPE with s=2 of C	$MPE3_{(C)}$	MPE with s=3 of C
G	$PE_{(G)}$	PE of G	$MPE2_{(G)}$	MPE with s=2 of G	$MPE3_{(G)}$	MPE with s=3 of G
T	$PE_{(T)}$	PE of T	$MPE2_{(T)}$	MPE with s=2 of T	$MPE3_{(T)}$	MPE with s=3 of T
	all_$MPE_{(A,C,G,T)}$					
	$PE_{(A)}$, $PE_{(C)}$, $PE_{(G)}$, $PE_{(T)}$, $MPE2_{(A)}$, $MPE2_{(C)}$, $MPE2_{(G)}$, $MPE2_{(T)}$, $MPE3_{(A)}$, $MPE3_{(C)}$, $MPE3_{(G)}$, $MPE3_{(T)}$					

the proposed feature extraction method (given in Step 2) and a method based on entropy features without wavelet decomposition (given in Step 1). The steps of experimental setup are given below:

Step 1: In this step, four different feature sets based on MPE are considered. To create the first feature set, MPE values are calculated for a single base of a trace file. In order to compute entropy values, MPE parameters, m and τ, are set at 3 and 1, respectively, as suggested in [49], [53]. Besides, the scale parameter s is taken as 1 (leading to single scale MPE, namely PE). The same procedure is applied to all bases of a trace and then to all trace files. Hence, the first feature set which is then fed to k-NN classifier is created. This set consists of 4 vector of features ($PE_{(A)}$, $PE_{(C)}$, $PE_{(G)}$, $PE_{(T)}$) and it is demonstrated as $PE_{(A,C,G,T)}$. This step is repeated to form second (namely $MPE2_{(A,C,G,T)}$) and third (namely $MPE3_{(A,C,G,T)}$) feature sets which are based on MPE with scale parameter $s = 2$ and $s = 3$, respectively. As in $PE_{(A,C,G,T)}$, the second and the third feature sets consist of 4 vectors of features. The last feature set is formed by combining all features vectors from the first, the second and the third feature sets. Thus, the fourth feature set consists of 12 vector of features and it is demonstrated by all_MPE$_{(A,C,G,T)}$. Table 1 shows the description of extracted feature sets in this step.

Step 2: As in Step 1, four sets of features are used in this step. After obtaining the coefficients produced for any of 4 sub-bands (decomposition stage) of a base of a trace file, entropy-based features are calculated. To form the first feature set, MPE parameters,m, τ and s are set at 3, 1 and 1, respectively (leading to single scale MPE, namely PE) and PE values are computed for all sub-bands of a base of a trace. The same procedure is applied to all bases of a trace, and then to all trace files. Hence the first feature set is created. This set consists of 16 vector of features, of which 4 are PE values of base A (SB$_1$_PE$_{(A)}$, SB$_2$_PE$_{(A)}$, SB$_3$_PE$_{(A)}$, SB$_4$_PE$_{(A)}$), 4 are PE values of base C (SB$_1$-PE$_{(C)}$, SB$_2$_PE$_{(C)}$, SB$_3$_PE$_{(C)}$, SB$_4$_PE$_{(C)}$), 4 are PE values of base G (SB$_1$_PE$_{(G)}$, SB$_2$_PE$_{(G)}$, SB$_3$_PE$_{(G)}$, SB$_4$_PE$_{(G)}$) and 4 are PE values of base T (SB$_1$_PE$_{(T)}$, SB$_2$_PE$_{(T)}$, SB$_3$_PE$_{(T)}$, SB$_4$_PE$_{(T)}$). In addition, Kruskal-Wallis H test is utilized to show the discrimination abilities of each sub-band of a base. Thus, an additional computational load from a redundant information is prevented. This step is repeated to form second and third feature sets with the same m and τ parameters, but with different scale parameters such as $s = 2$ and $s = 3$, respectively. The last feature set is formed by combining all features vectors from the first, the second and the third feature sets. Thus, the fourth feature set consists of 48 vector of features and it is demonstrated by all_wMPE$_{(A,C,G,T)}$. Appendix A shows the description of extracted the first, the second and the third feature sets which are demonstrated by wPE$_{(A,C,G,T)}$, wMPE2$_{(A,C,G,T)}$ and wMPE3$_{(A,C,G,T)}$, respectively. Besides, the Kruskal-Wallis H test results applied on all extracted feature sets and the resultant p-values are shown in Appendix A. As it is shown, the discrimination ability among all vector of features is statistically significant (p-values < 0.001).

Block diagram of the experimental setup is given in Figure 2. Two different feature extraction methods such as entropy features with and without wavelet decomposition are utilized. The results of classifications using all extracted feature sets based on these two methods are obtained with different training sample proportions (from 5% to 30%, increased by %5) and with different k values (from 1 to 5). To avoid random selection effect, each classification process for a feature set is repeated 100 times and the average values of performance measures (μ_{ACC}, μ_{κ}, μ_{Se} and μ_{Sp}) are considered to make performance evaluation. Also, it is aimed to show that each classification process performed with different feature set does not suffer from an over-fitting problem. For this aim, it is expected to obtain non-complementary values of μ_{Se} and μ_{Sp} [54] in each classification process. Pairwise comparisons of the performance measures between {PE$_{(A,C,G,T)}$—wPE$_{(A,C,G,T)}$}, {MPE2$_{(A,C,G,T)}$—wMPE2$_{(A,C,G,T)}$}, {MPE3$_{(A,C,G,T)}$—wMPE3$_{(A,C,G,T)}$} and {all_MPE$_{(A,C,G,T)}$ —all_wMPE$_{(A,C,G,T)}$} are made with different sample proportions and k values.

RESULTS

In this study, different feature sets are extracted for 200 hepatitis DNA trace files according to the proposed feature extraction method which is based on entropy features with wavelet decomposition. In assessing the summarization ability of the proposed method, different feature sets with the same entropy parameters are also extracted using entropy features without wavelet decomposition. All feature sets are then fed to k-NN classifier with different k and training sample proportions. Pairwise comparisons of the classification performance measures between extracted feature sets based on entropy features with and without wavelet decomposition are given in following tables (from Table 2 to Table 7 for training sample proportions from 5% to 30%).

Table 2 shows all pairwise comparison results of k-NN classifications based on two feature extraction methods at 5% training sample proportion. Compared to the method based on entropy features without wavelet decomposition, higher μ_{ACC} and μ_{κ} are obtained in all pairwise comparisons

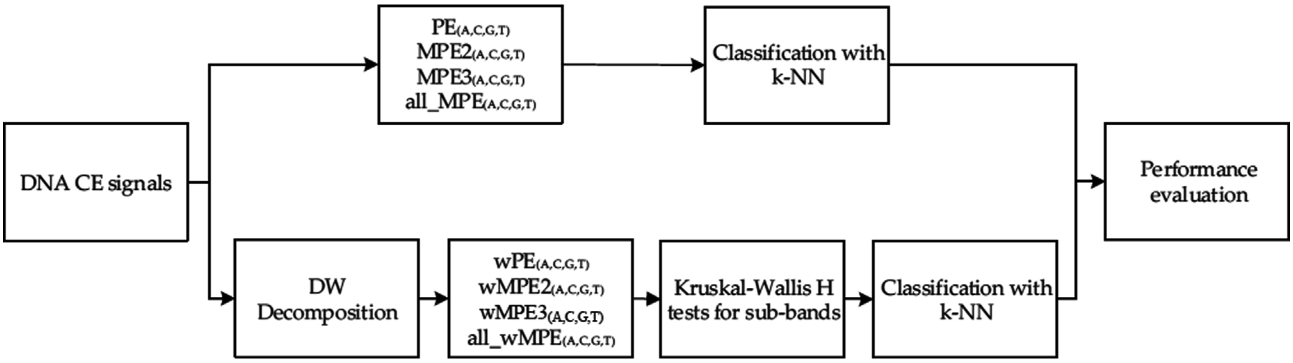

Figure 2. Block diagram of the experimental setup.

Table 2. Pairwise comparisons of feature sets at 5% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1	PE$_{(A,C,G,T)}$	0.907	0.816	0.898	0.918	wPE$_{(A,C,G,T)}$	0.920	0.842	0.903	0.940
2		0.848	0.700	0.793	0.910		0.875	0.753	0.841	0.914
3		0.835	0.675	0.777	0.901		0.857	0.717	0.817	0.902
4		0.782	0.569	0.717	0.856		0.819	0.641	0.785	0.858
5		0.724	0.457	0.646	0.815		0.747	0.501	0.719	0.785
1	MPE2$_{(A,C,G,T)}$	0.903	0.808	0.868	0.942	wMPE2$_{(A,C,G,T)}$	0.906	0.814	0.875	0.942
2		0.843	0.691	0.788	0.906		0.883	0.768	0.855	0.914
3		0.842	0.687	0.790	0.899		0.866	0.735	0.860	0.876
4		0.774	0.552	0.729	0.826		0.822	0.647	0.804	0.845
5		0.726	0.459	0.659	0.805		0.805	0.613	0.783	0.833
1	MPE3$_{(A,C,G,T)}$	0.892	0.787	0.857	0.931	wMPE3$_{(A,C,G,T)}$	0.917	0.836	0.907	0.929
2		0.852	0.708	0.794	0.917		0.888	0.777	0.877	0.900
3		0.834	0.674	0.773	0.903		0.886	0.774	0.871	0.904
4		0.795	0.594	0.749	0.848		0.864	0.732	0.853	0.879
5		0.736	0.481	0.642	0.844		0.840	0.682	0.847	0.836
1	all_MPE$_{(A,C,G,T)}$	0.896	0.794	0.872	0.923	all_wMPE$_{(A,C,G,T)}$	0.910	0.823	0.890	0.934
2		0.855	0.713	0.818	0.897		0.880	0.762	0.872	0.891
3		0.830	0.666	0.751	0.918		0.875	0.753	0.868	0.885
4		0.797	0.603	0.713	0.892		0.808	0.620	0.768	0.855
5		0.741	0.487	0.698	0.793		0.765	0.534	0.766	0.771

Table 3. Pairwise comparisons of feature sets at 10% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1	PE$_{(A,C,G,T)}$	0.936	0.872	0.933	0.939	wPE$_{(A,C,G,T)}$	0.954	0.909	0.948	0.961
2		0.915	0.832	0.901	0.932		0.934	0.869	0.932	0.937
3		0.907	0.817	0.891	0.927		0.931	0.862	0.940	0.921
4		0.899	0.799	0.886	0.913		0.917	0.835	0.920	0.914
5		0.864	0.732	0.826	0.908		0.906	0.815	0.902	0.912
1	MPE2$_{(A,C,G,T)}$	0.939	0.879	0.929	0.951	wMPE2$_{(A,C,G,T)}$	0.948	0.896	0.934	0.964
2		0.921	0.843	0.910	0.934		0.935	0.870	0.928	0.942
3		0.906	0.815	0.892	0.924		0.929	0.858	0.928	0.930
4		0.896	0.794	0.883	0.913		0.923	0.848	0.934	0.913
5		0.892	0.787	0.881	0.907		0.916	0.833	0.924	0.909
1	MPE3$_{(A,C,G,T)}$	0.932	0.865	0.923	0.943	wMPE3$_{(A,C,G,T)}$	0.932	0.865	0.938	0.927
2		0.906	0.814	0.881	0.935		0.922	0.845	0.916	0.930
3		0.908	0.819	0.898	0.923		0.931	0.862	0.943	0.918
4		0.887	0.776	0.861	0.917		0.929	0.858	0.943	0.914
5		0.885	0.774	0.864	0.910		0.926	0.852	0.942	0.909
1	all_MPE$_{(A,C,G,T)}$	0.937	0.875	0.920	0.956	all_wMPE$_{(A,C,G,T)}$	0.958	0.916	0.953	0.963
2		0.922	0.844	0.909	0.937		0.935	0.871	0.929	0.943
3		0.921	0.843	0.923	0.920		0.936	0.872	0.947	0.924
4		0.891	0.784	0.862	0.923		0.922	0.845	0.930	0.915
5		0.876	0.757	0.846	0.913		0.919	0.840	0.930	0.909

with proposed feature extraction method. μ_{ACC} difference of each pairwise comparison range between 0.3% and 10.4%. More specificly, the highest μ_{ACC} difference (10.4%) is observed between the classifications using feature sets of MPE3$_{(A,C,G,T)}$ (73.6%) and wMPE3$_{(A,C,G,T)}$ (84.0%), at k = 5. The best classification performance is μ_{ACC} = 92.0% and it is obtained with wPE$_{(A,C,G,T)}$ feature set at k = 1. Also, it is observed that as k increases, the classification performance of k-NN classifier using both entropy features with and without wavelet decomposition decreases (based on μ_{ACC} and μ_{κ}). In addition, according to the μ_{Se} and μ_{Sp}, over-fitting does not occured in any of classification process performed with different feature sets.

The pairwise comparison results of k-NNclassification performances using feature sets based on entropy features with and without wavelet decomposition are given in Table 3 (at 10% training sample proportion). As it is seen, entropy features with wavelet decomposition generate higher μ_{ACC} and μ_{κ} in nearly all comparisons (except comparison between MPE3$_{(A,C,G,T)}$ and wMPE3$_{(A,C,G,T)}$, at k=1. Here, both feature sets generate the equal μ_{ACC} and μ_{κ}) as compared with entropy features without wavelet decomposition. When the μ_{ACC} of each pairwise comparison is considered, the highest μ_{ACC} difference is obtained as 4.3% between the feature sets of all_MPE$_{(A,C,G,T)}$ and all_wMPE$_{(A,C,G,T)}$, at k=5. Also, the classification with wPE$_{(A,C,G,T)}$ produce μ_{ACC} as 95.4% (at k=1), which is very close to the highest μ_{ACC}. Also, as k increases, the classification performance of k-NN classifier using both entropy features with and without wavelet decomposition decreases in general. Besides, computed μ_{Se} and μ_{Sp} suggest that over-fitting problem does not happen in any classification process.

Table 4 shows the pairwise comparisons of the classification results at 15% training sample proportion. According to the μ_{ACC} and μ_{κ}, entropy features with wavelet decomposition produce better classification performances for all pairwise comparisons than entropy features without wavelet decomposition. The μ_{ACC} difference of each pairwise comparison range from 0.1% to 1.8%. The highest μ_{ACC} is obtained in the classification using feature set of all_wMPE$_{(A,C,G,T)}$ (96.7%). Also, the classifications with wPE$_{(A,C,G,T)}$ (96.4%, at k=1) and wMPE2$_{(A,C,G,T)}$ (96.2%, at k=1) produce similar performances. In addition, overfitting problem does not occurred in any of classification process according to the μ_{Se} and μ_{Sp}.

The pairwise comparsions of the feature sets' classifications results at 20% training sample proportion are provided in Table 5. μ_{ACC} and μ_{κ} of all pairwise comparisons favor the classifications that used the feature sets based on entropy features with wavelet decomposition. The highest μ_{ACC}

Table 4. Pairwise comparisons of feature sets at 15% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1		0.957	0.914	0.963	0.950		0.964	0.928	0.953	0.976
2		0.937	0.875	0.934	0.942		0.945	0.891	0.944	0.948
3	PE$_{(A,C,G,T)}$	0.944	0.889	0.961	0.928	wPE$_{(A,C,G,T)}$	0.945	0.890	0.950	0.940
4		0.924	0.849	0.926	0.923		0.937	0.875	0.949	0.925
5		0.923	0.847	0.932	0.915		0.935	0.872	0.954	0.916
1		0.958	0.916	0.958	0.958		0.962	0.925	0.957	0.969
2		0.944	0.888	0.940	0.948		0.950	0.900	0.951	0.949
3	MPE2$_{(A,C,G,T)}$	0.943	0.885	0.954	0.930	wMPE2$_{(A,C,G,T)}$	0.949	0.898	0.956	0.942
4		0.934	0.868	0.936	0.932		0.941	0.882	0.948	0.935
5		0.923	0.846	0.932	0.914		0.940	0.881	0.961	0.920
1		0.939	0.879	0.930	0.951		0.940	0.879	0.945	0.934
2		0.931	0.863	0.922	0.941		0.932	0.865	0.938	0.927
3	MPE3$_{(A,C,G,T)}$	0.931	0.863	0.937	0.927	wMPE3$_{(A,C,G,T)}$	0.945	0.890	0.970	0.919
4		0.924	0.848	0.924	0.924		0.942	0.885	0.968	0.914
5		0.925	0.851	0.935	0.916		0.939	0.878	0.966	0.910
1		0.962	0.925	0.964	0.961		0.967	0.935	0.964	0.971
2		0.948	0.896	0.945	0.952		0.953	0.907	0.945	0.962
3	all_MPE$_{(A,C,G,T)}$	0.946	0.892	0.958	0.933	all_wMPE$_{(A,C,G,T)}$	0.952	0.905	0.964	0.941
4		0.934	0.870	0.946	0.923		0.945	0.890	0.963	0.926
5		0.929	0.860	0.942	0.917		0.940	0.881	0.964	0.915

Table 5. Pairwise comparisons of feature sets at 20% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1	$PE_{(A,C,G,T)}$	0.964	0.928	0.967	0.960	$wPE_{(A,C,G,T)}$	0.971	0.942	0.960	0.983
2		0.951	0.903	0.953	0.951		0.960	0.920	0.956	0.965
3		0.951	0.902	0.969	0.933		0.958	0.917	0.969	0.947
4		0.950	0.901	0.969	0.931		0.951	0.902	0.962	0.940
5		0.945	0.891	0.971	0.919		0.950	0.900	0.973	0.925
1	$MPE2_{(A,C,G,T)}$	0.964	0.927	0.964	0.963	$wMPE2_{(A,C,G,T)}$	0.967	0.934	0.959	0.976
2		0.957	0.913	0.958	0.955		0.958	0.915	0.958	0.958
3		0.955	0.910	0.968	0.942		0.959	0.919	0.976	0.942
4		0.945	0.891	0.963	0.927		0.954	0.909	0.972	0.936
5		0.940	0.881	0.956	0.924		0.953	0.907	0.980	0.925
1	$MPE3_{(A,C,G,T)}$	0.941	0.882	0.922	0.961	$wMPE3_{(A,C,G,T)}$	0.942	0.883	0.947	0.936
2		0.934	0.869	0.920	0.950		0.936	0.871	0.940	0.931
3		0.948	0.897	0.961	0.935		0.948	0.897	0.974	0.920
4		0.940	0.881	0.945	0.936		0.945	0.890	0.977	0.912
5		0.941	0.883	0.961	0.920		0.950	0.899	0.984	0.913
1	$all\_MPE_{(A,C,G,T)}$	0.970	0.941	0.974	0.967	$all\_wMPE_{(A,C,G,T)}$	0.973	0.947	0.966	0.982
2		0.959	0.918	0.955	0.964		0.965	0.930	0.966	0.965
3		0.957	0.915	0.957	0.958		0.969	0.938	0.981	0.956
4		0.944	0.889	0.953	0.935		0.953	0.905	0.969	0.936
5		0.942	0.884	0.960	0.923		0.954	0.909	0.986	0.921

Table 6. Pairwise comparisons of feature sets at 25% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1	$PE_{(A,C,G,T)}$	0.967	0.935	0.974	0.960	$wPE_{(A,C,G,T)}$	0.980	0.960	0.972	0.988
2		0.955	0.911	0.963	0.947		0.967	0.934	0.958	0.977
3		0.961	0.923	0.982	0.940		0.965	0.931	0.971	0.960
4		0.953	0.907	0.972	0.935		0.957	0.915	0.973	0.942
5		0.952	0.904	0.977	0.926		0.956	0.912	0.969	0.942
1	$MPE2_{(A,C,G,T)}$	0.972	0.944	0.973	0.971	$wMPE2_{(A,C,G,T)}$	0.972	0.944	0.964	0.981
2		0.961	0.922	0.962	0.961		0.964	0.929	0.961	0.969
3		0.960	0.921	0.976	0.945		0.965	0.929	0.973	0.956
4		0.950	0.901	0.967	0.934		0.959	0.918	0.970	0.947
5		0.948	0.897	0.972	0.923		0.959	0.917	0.983	0.932
1	$MPE3_{(A,C,G,T)}$	0.945	0.891	0.933	0.959	$wMPE3_{(A,C,G,T)}$	0.945	0.891	0.949	0.943
2		0.937	0.875	0.919	0.958		0.937	0.875	0.935	0.939
3		0.949	0.899	0.957	0.942		0.950	0.899	0.973	0.924
4		0.946	0.893	0.953	0.940		0.947	0.893	0.975	0.917
5		0.946	0.892	0.964	0.927		0.951	0.902	0.983	0.917
1	$all\_MPE_{(A,C,G,T)}$	0.972	0.944	0.971	0.974	$all\_wMPE_{(A,C,G,T)}$	0.981	0.962	0.977	0.985
2		0.964	0.929	0.962	0.968		0.971	0.942	0.971	0.972
3		0.963	0.926	0.969	0.957		0.971	0.942	0.974	0.968
4		0.956	0.913	0.968	0.944		0.962	0.925	0.967	0.957
5		0.953	0.907	0.978	0.927		0.958	0.915	0.983	0.930

Table 7. Pairwise comparisons of feature sets at 30% training sample proportion

k	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}	Feature Set	μ_{ACC}	μ_{κ}	μ_{Se}	μ_{Sp}
1	$PE_{(A,C,G,T)}$	0.971	0.943	0.976	0.967	$wPE_{(A,C,G,T)}$	0.982	0.964	0.970	0.995
2		0.961	0.923	0.966	0.957		0.973	0.946	0.966	0.981
3		0.969	0.938	0.989	0.947		0.972	0.945	0.969	0.976
4		0.961	0.921	0.981	0.939		0.964	0.927	0.968	0.959
5		0.959	0.919	0.992	0.924		0.963	0.926	0.969	0.956
1	$MPE2_{(A,C,G,T)}$	0.973	0.946	0.972	0.974	$wMPE2_{(A,C,G,T)}$	0.973	0.946	0.963	0.983
2		0.965	0.930	0.964	0.967		0.967	0.934	0.956	0.980
3		0.964	0.928	0.965	0.963		0.968	0.937	0.975	0.962
4		0.956	0.913	0.969	0.943		0.966	0.933	0.979	0.953
5		0.955	0.911	0.979	0.931		0.968	0.937	0.982	0.955
1	$MPE3_{(A,C,G,T)}$	0.948	0.896	0.938	0.959	$wMPE3_{(A,C,G,T)}$	0.948	0.896	0.950	0.945
2		0.941	0.883	0.935	0.949		0.942	0.885	0.939	0.946
3		0.951	0.903	0.952	0.952		0.953	0.906	0.973	0.931
4		0.947	0.895	0.958	0.938		0.949	0.897	0.976	0.920
5		0.947	0.894	0.966	0.926		0.951	0.903	0.982	0.919
1	$all\_MPE_{(A,C,G,T)}$	0.976	0.951	0.981	0.971	$all\_wMPE_{(A,C,G,T)}$	0.985	0.970	0.975	0.995
2		0.971	0.942	0.976	0.966		0.977	0.954	0.970	0.985
3		0.970	0.940	0.986	0.953		0.979	0.958	0.982	0.977
4		0.965	0.931	0.973	0.957		0.971	0.942	0.980	0.961
5		0.960	0.920	0.985	0.934		0.963	0.927	0.984	0.943

difference is found as 1.3% for the pairwise comparisons between the feature sets of $MPE2_{(A,C,G,T)}$ and $wMPE2_{(A,C,G,T)}$, at k=5. The best classification performance in terms of μ_{ACC} is produced with $all\_wMPE_{(A,C,G,T)}$ feature set, at k=1 (97.3%). Also, classification with feature set $wPE_{(A,C,G,T)}$ produce the similar performance at k = 1 (97.1%). The over-fitting does not observed according to the μ_{Se} and μ_{Sp}.

All pairwise comparison results of classifications at 25% training sample proportion are given in Table 6. When compared with the entropy features without wavelet decomposition, higher μ_{ACC} and μ_{κ} are obtained with the proposed method in most of pairwise comparisons. μ_{ACC} and μ_{κ} take the same value between classifications using feature sets of $MPE2_{(A,C,G,T)}$ and $wMPE2_{(A,C,G,T)}$ at k=1 and $MPE3_{(A,C,G,T)}$ and $wMPE3_{(A,C,G,T)}$ at k=2. The highest μ_{ACC} (98.1%) is observed in the classification using the feature set $all\_wMPE_{(A,C,G,T)}$ at k=1. Also, with feature set $wPE_{(A,C,G,T)}$, the similar classification accuracy is achieved (98.0%) at the same k. When μ_{ACC} differences of pairwise comparisons are considered, the highest difference (1.3%) is observed between the feature sets $PE_{(A,C,G,T)}$ and $wPE_{(A,C,G,T)}$. In addition, according to the μ_{Se} and μ_{Sp}, over-fitting problem does not appear in any classification process performed with different feature sets.

The last table includes pairwise comparisons of k-NN classifications using the entropy features with and without wavelet decomposition at 30% training sample proportion (Table 7). In most of all pairwise comparisons, it is observed that feature sets extracted with the proposed feature extraction method produce better classification performances than the feature sets based on entropy without wavelet decomposition in terms of μ_{ACC} and μ_{κ}. The same μ_{ACC} and μ_{κ} are obtained between the pairwise comparisons of feature sets $MPE2_{(A,C,G,T)}$ and $wMPE2_{(A,C,G,T)}$ and $MPE3_{(A,C,G,T)}$ and $wMPE3_{(A,C,G,T)}$, at k=1. The highest μ_{ACC} is found as 98.5% is observed in the classification with $all\_wMPE_{(A,C,G,T)}$ at k=1. Also, with feature set $wPE_{(A,C,G,T)}$, the similar classification peformance is produced with the same k (98.2%). Besides, the highest μ_{ACC} difference is found as 1.3% between the pairwise comparison of classifications using the feature sets of $MPE2_{(A,C,G,T)}$ and $wMPE2_{(A,C,G,T)}$ at k=5. As in other training sample proportions, over-fitting problem does not evident in any classification process according to the μ_{Se} and μ_{Sp} results.

DISCUSSION

Machine learning (ML) can be seen as a very useful tool in the interpretation of genomic data [55] and has been extensively used for the purpose of genomic sequence classification with the rapid development of information technologies in recent years [55–59]. Especially, for DNA sequencing data, a very common and important challenge is discriminating the genes belonging different classes since distinguishing the signals by visually is almost impossible. ML methods present promising performances in various tasks such as recognition and categorization to overcome this challenge as shown in many important studies [6, 60–68].

One of the important pre-processing stage of the ML is reducing the dimension [69] of the high-dimensional raw data without sacrificing the useful information. Various feature extraction methods have been proposed in an attempt to find most compact and informative feature sets [70] to achieve high classification ability with small error rate. For the purpose of obtaining correct reflection of signals, some smoothing and filtering techniques such as SG and FT can be used. However, these traditional approaches are not efficient in CE signal denoising because CE signals have different shapes and widths in both spatial and frequency domains [4, 71]. As stated in [26], WT is an appropriate denoising tool for the DNA CE signals because the important information that is hidden in both spatial and frequency domains can be captured by WT. Some pioneer studies have been conducted to show the applicability of this transformation technique in the DNA CE signals [21–26]. However, the length of signals located in each sub-bands still differs among bases of DNA trace files after decomposition. The differences in lengths lead to a dimensionality problem in forming a vector of features for the purpose of using it as an input to the relevant classifier. Thus, an additional transformation stage is needed for each element of the feature vector to have same dimension.

Entropy is a very powerful and well-known statistical measure in quantifying the complexity of biological signals and has been widely used, especially in the studies of EEG signals [37–43], as a feature extraction method to obtain high classification accuracy. However, limited studies deal with the complexity of signals stored in each nucleic base (A, C, G and T). In the previous study [44], it was shown that the entropy-based features calculated for each bases of DNA trace files produce remarkable results in discriminating hepatitis DNA trace files as HBV and HCV. Support vector machines (SVM) classifier was used with different kernel functions and parameter optimization was performed for SVM parameters.

In this study, on the other hand, a feature extraction method which combines both DWT and entropy is proposed. Proposed method is defined with two stages that are executed sequentially. While first stage involves the DW decomposition of bases signals into sub-bands, the second stage involves forming the vector of features based on calculating the entropy values for these sub-bands. The real data set covering 200 patients' DNA trace files are used. Among 200 patients having hepatitis, 96 of them was labelled as HBV and 104 of them was labelled as HCV. Different feature sets are generated using the proposed method with different MPE parameters. Then, generated feature sets are fed to a memory-based classifier, k-NN, due to the simplicity [72]. Different values of k from 1 to 5 are considered. In order to assess the performance of the proposed feature extraction method, it is compared to the classification results of generated feature sets based on MPE without wavelet decomposition. Different training sample sizes ranging from 5% to 30 in 5% increments are handled. Each classification process with different feature sets and training sample proportions is repeated 100 times to avoid bias caused by random selection. The average of performance measures obtained from the 100 runs are computed (μ_{ACC}, μ_{κ}, μ_{Se} and μ_{Sp}). Pairwise comparisons of the average performance measures between feature sets generated with two different feature extraction methods, (i.e., entropy features with and without wavelet decomposition) are made. The results demonstrate that features extracted with proposed method produce higher μ_{ACC} and μ_{κ} at all training sample proportions (i.e., from 5% to 30%) compared to the entropy features without wavelet decomposition even using a simple memory-based learning classifier (k-NN) that does not require any parameter optimization. The highest μ_{ACC} are obtained almost in all pairwise comparisons for the feature set of all_wMPE$_{(A,C,G,T)}$ at k=1. Also, wPE$_{(A,C,G,T)}$ feature set produce similar μ_{ACC} values with the same k. With proposed feature extraction method, the μ_{ACC} range from 92.0% to 98.5%. As it is expected, when the training proportion increases, the average classification accuracy increases. This shows, even at 30% training samples proportion, the classification performance reaches up to nearly 99% by entropy features with wavelet decomposition. When the classification results of feature sets are considered individually, the highest μ_{ACC} is obtained at k=1 among all values of k (for all training samples proportions).

The proposed feature extraction method combined the DWT and entropy has several advantages against the entropy-based features without wavelet decomposition. The proposed method has the following advantages:

- In all pairwise classification comparisons of feature sets, entropy based features with wavelet decomposition produce better performance measures compared to the feature extraction method which is based on entropy but without wavelet decomposition. It is shown that wavelet decomposition is an important stage in the feature extraction process in order to localize the nitrogen-containing bases (i.e., A, C, G and T) in DNA. After decomposition of signals in all bases, the complexities of bases' sub-bands are measured by MPE and, in conclusion, very satisfactory classification performances are obtained.
- As the training sample proportion decreases, μ_{ACC} range between pairwise comparison of classifications based on entropy features with and without wavelet decomposition is widening. Considering that the training samples of real DNA sequence data are naturally low, this result is very important.

CONCLUSION

In this study, a new feature extraction method is proposed for the classification of DNA trace files based on entropy features with wavelet decomposition. In addition to enabling the detection of hidden information stored in each bases of trace files, this method also allows to assess the chaotic nature of this information at the local level. To achieve better identification of the genotypes of viruses such as hepatitis, this method can be effectively used in the studies classification of DNAC.

Although a simple memory-based learning classifier is used in this study, satisfactory classification performances are obtained with various feature sets generated by the proposed method. Different classifiers which require the parameter optimization can also be considered for future studies if the purpose is improving the classification performance.

REFERENCES

[1] Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci 1977;74:560–564.

[2] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 1977;74:5463–5467. [CrossRef]

[3] Swerdlow H, Zhang JZ, Chen DY, Harke HR, Grey R, Wu S, et al. Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence. Anal Chem 1991;63:2835–2841. [CrossRef]

[4] Wang Y, Gao Q. Spatially adaptive stationary wavelet thresholding for the denoising of DNA capillary electrophoresis signal. J Anal Chem 2008;63:768–774. [CrossRef]

[5] Karger BL, Guttman A. DNA sequencing by CE. Electrophoresis 2009;30:S196–S202. [CrossRef]

[6] Öz E, Kaya H. Support vector machines for quality control of DNA sequencing. J Ineq Appl 2013;85:1–9. [CrossRef]

[7] Combes JM, Grossman A. Ychamitchian, P. Wavelets. The Time-Frequency Methods and Phase Space. Berlin, Heidelberg: Springer Verlag; 1989. [CrossRef]

[8] Adeli H, Zhou Z. Dadmehr, N. Analysis of EEG records in an epileptic patient using wavelet transform. J Neurosci Meth 2003;123:69–87. [CrossRef]

[9] Amin HU, Malik AS, Ahmad RF, Badruddin, N.; Kamel N, Hussain M, Chooi WT. Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. Australas Phys Eng Sci Med 2015;38:139–149. [CrossRef]

[10] Cvetkovic D, Übeyli ED, Cosic I. Wavelet transform feature extraction from human PPG, ECG, and EEG signal responses to ELF PEMF exposures: A pilot study. Digit Signal Process 2008;18:861–874. [CrossRef]

[11] Gandhi T, Panigrahi BK, Anand S. A comparative study of wavelet families for EEG signal classification. Neurocomputing 2011;74:3051–3057. [CrossRef]

[12] Guo L, Rivero D, Dorado J, Rabunal JR, Pazos A. Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. J Neurosci Meth 2010;191:101–109. [CrossRef]

[13] Hazarika N, Chen JZ, Tsoi AC, Sergejew A. Classification of EEG signals using the wavelet transform. Signal Process 1997;59:61–72. [CrossRef]

[14] Jahankhani P, Kodogiannis V, Revett K. EEG signal classification using wavelet feature extraction and neural networks. In: IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06); 2006. [CrossRef]

[15] Kousarrizi MRN, Ghanbari AA, Teshnehlab M, Shorehdeli MA, Gharaviri A. Feature extraction and classification of EEG signals using wavelet transform, SVM and artificial neural networks for brain computer interfaces. In: International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing; 2009. [CrossRef]

[16] Mahmoodabadi SZ, Ahmadian A, Abolhasani MD. ECG feature extraction using Daubechies wavelets. In: Proceedings of the fifth IASTED International conference on Visualization, Imaging and Image Processing; 2005.

[17] Murugappan M, Rizon M, Nagarajan R, Yaacob S, Zunaidi I, Hazry D. EEG feature extraction for classifying emotions using FCM and FKM. Int J Comput Commun 2007;1:21–25.

[18] Prochazka A, Kukal J, Vysata O. Wavelet transform use for feature extraction and EEG signal segments classification. In: 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008); 2008. [CrossRef]

[19] Sadati N, Mohseni HR, Maghsoudi A. Epileptic seizure detection using neural fuzzy networks. In: IEEE International Conference on Fuzzy Systems, 2006. [CrossRef]

[20] Subasi A. EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl 2007;32:1084–1093. [CrossRef]

[21] Cao W, Chen X, Yang X, Wang E. Discrete wavelets transform for signal denoising in capillary electrophoresis with electrochemiluminescence detection. Electrophoresis 2003;24:3124–3130. [CrossRef]

[22] Ceballos GA, Paredes JL, Hernández LF. Pattern recognition in capillary electrophoresis data using dynamic programming in the wavelet domain. Electrophoresis 2008;29:2828–2840. [CrossRef]

[23] Gao Q, Lu Y, Sun D, Zhang D. A multiscale products technique for denoising of DNA capillary electrophoresis signals. Meas Sci Technol 2013;24:065004. [CrossRef]

[24] Olazábal V, Prasad L, Stark P, Olivares JA. Application of wavelet transforms and an approximate deconvolution method for the resolution of noisy overlapped peaks in DNA capillary electrophoresis. Analyst 2004;129:73–81. [CrossRef]

[25] Perrin C, Walczak B, Massart DL. The use of wavelets for signal denoising in capillary electrophoresis. Anal Chem 2001;73:4903–4917. [CrossRef]

[26] Wang Y, Gao Q. Spatially adaptive stationary wavelet thresholding for the denoising of DNA capillary electrophoresis signal. J Anal Chem 2008;63:768–774. [CrossRef]

[27] Duda RO, Hart PE, Stork DG. Pattern Classification; USA: John Wiley & Sons; 2012.

[28] Li H, Jiang T, Zhang K. Efficient and robust feature extraction by maximum margin criterion. In: Thrun S, Saul LK, Schölkopf B, editors. Advances in Neural Information Processing Systems, London: MIT Press; 2004.

[29] Sakarya U. Dimension reduction using global and local pattern information-based maximum margin criterion. Signal Image Video Process 2016;10:903–909. [CrossRef]

[30] Kandaswamy A, Kumar CS, Ramanathan RP, Jayaraman S, Malmurugan N. Neural classification of lung sounds using wavelet coefficients. Comput Biol Med 2004;34:523–537. [CrossRef]

[31] Mahaphonchaikul K, Sueaseenak D, Pintavirooj C, Sangworasil M, Tungjitkusolmun S. EMG signal feature extraction based on wavelet transform. In: The 2010 ECTI International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON2010); 2010. [CrossRef]

[32] Ji N, Ma L, Dong H, Zhang X. EEG signals feature extraction based on DWT and EMD combined with approximate entropy. Brain Sci 2019;9:201. [CrossRef]

[33] Kumar Y, Dewal ML, Anand RS. Relative wavelet energy and wavelet entropy based epileptic brain signals classification. Biomed Eng Lett 2012;2:147–157. [CrossRef]

[34] Sharma R, Pachori RB, Acharya UR. An integrated index for the identification of focal electroencephalogram signals using discrete wavelet transform and entropy measures. Entropy 2015;17:5218–5240. [CrossRef]

[35] You Y, Chen W, Li M, Zhang T, Jiang Y, Zheng X. Automatic focal and non-focal EEG detection using entropy-based features from flexible analytic wavelet transform. Biomed Signal Process Cont 2020;57:101761. [CrossRef]

[36] Panda R, Khobragade PS, Jambhule PD, Jengthe SN, Pal PR, Gandhi TK. Classification of EEG signal using wavelet transform and support vector machine for epileptic seizure diction. In: International Conference on Systems in Medicine and Biology; 2010. [CrossRef]

[37] Acharya UR, Molinari F, Sree SV, Chattopadhyay S, Ng KH, Suri JS. Automated diagnosis of epileptic EEG using entropies. Biomed Signal Process Cont 2012;7:401–408. [CrossRef]

[38] Acharya UR, Sree SV, Ang PCA, Yanti R, Suri JS. Application of non-linear and wavelet based features for the automated identification of epileptic EEG signals. Int J Neural Syst 2012;22:1250002. [CrossRef]

[39] Arunkumar N, Ramkumar K, Venkatraman V, Abdulhay E, Fernandes SL, Kadry S, et al. Classification of focal and non focal EEG using entropies. Pattern Recognit Lett 2017;94:112–117. [CrossRef]

[40] Bhattacharyya A, Pachori RB, Upadhyay A, Acharya UR. Tunable-Q wavelet transform based multiscale entropy measure for automated classification of epileptic EEG signals. Appl Sci 2017;7:385. [CrossRef]

[41] Michielli N, Acharya UR, Molinari F. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Comput Biol Med 2019;106:71–81. [CrossRef]

[42] Sharma R, Pachori RB, Acharya UR. Application of entropy measures on intrinsic mode functions for the automated identification of focal electroencephalogram signals. Entropy 2015;17:669–691. [CrossRef]

[43] Yuan Q, Zhou W, Li S, Cai D. Epileptic EEG classification based on extreme learning machine and nonlinear features. Epilepsy Res 2011;96:29–38. [CrossRef]

[44] Öz E, Aşkın ÖE. Classification of hepatitis viruses from sequencing chromatograms using multiscale permutation entropy and support vector machines. Entropy 2019;21:1149. [CrossRef]

[45] MATLAB (2017), Version 9.2.0 (R2017a), Natick, MA, The MathWorks Inc.

[46] Kumar K. Standardising the chromatographic denoising procedure. Anal Meth 2018;10:4189–4200.

[47] Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 1989;11:674–693. [CrossRef]

[48] Ocak H. Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. Expert Syst Appl 2009;36:2027–2036. [CrossRef]

[49] Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 2002;88:1–4. [CrossRef]

[50] Costa M, Goldberger AL, Peng CK. Multiscale entropy analysis of complex physiologic time series. Phys Rev Lett 2002;89:1–4. [CrossRef]

[51] Aziz W, Arif M. Multiscale permutation entropy of physiological time series. In: Proceedings of the 9th

International Multitopic Conference (INMIC '05); 2005. [CrossRef]

[52] Daubechies I. Ten Lectures on Wavelets; Philadelphia: Siam; 1992. [CrossRef]

[53] Nalband S, Prince AA, Agrawal A. Entropy-based feature extraction and classification of vibroarthographic signal using complete ensemble empirical mode decomposition with adaptive noise. IET Sci Meas Tech 2018;12:350–359. [CrossRef]

[54] Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer Inform 2014;13:(Suppl 1):145–158. [CrossRef]

[55] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–332. [CrossRef]

[56] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Chen MS, Yu PS, Liu B, editors. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Berlin: Springer; 2002. [CrossRef]

[57] Randhawa GS, Hill KA, Kari L. ML-DSP: Machine learning with digital signal processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. BMC Genom 2019;20:267. [CrossRef]

[58] Remita MA, Halioui A, Diouara AAM, Daigle B, Kiani G, Diallo AB. A machine learning approach for viral genome classification. BMC Bioinform 2017;18:208. [CrossRef]

[59] Xing Z, Pei J, Keogh E. A brief survey on sequence classification. SIGKDD Explor 2010;12:40–48. [CrossRef]

[60] Avogadri R, Valentini G. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artif Intell Med 2009;45:173–183. [CrossRef]

[61] Dixit P, Prajapati GI. Machine learning in bioinformatics: A novel approach for dna sequencing. In: Fifth International Conference on Advanced Computing & Communication Technologies; 2015. [CrossRef]

[62] Khashei M, Hamadani AZ, Bijari M. A fuzzy intelligent approach to the classification problem in gene expression data analysis. Knowl Based Syst 2012;27:465–474. [CrossRef]

[63] Molla M, Waddell M, Page D, Shavlik J. Using machine learning to design and interpret gene-expression microarrays. AI Mag 2004;25:23.

[64] Öz E, Kurt S, Asyalı MH, Kaya H, Yücel Y. Feature based quality assessment of DNA sequencing chromatograms. Appl Soft Comput 2016;41:420–427. [CrossRef]

[65] Ranawana R, Palade V. A neural network based multi-classifier system for gene identification in DNA sequences. Neural Comput Appl 2005;14:122–131. [CrossRef]

[66] Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics 2018;34:1666–1671. [CrossRef]

[67] You W, Wang K, Li H, Jia Y, Wu X, Du Y. Classification of DNA sequences basing on the dinucleotide compositions. In: Second International Symposium on Computational Intelligence and Design; 2009. [CrossRef]

[68] Zhou Q, Jiang Q, Wei D. A new method for classification in DNA sequence. In: 6th International Conference on Computer Science Education (ICCSE); 2011. [CrossRef]

[69] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: Science and Information Conference; 2014. [CrossRef]

[70] Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications. Netherlands: Springer Verlag; 2008.

[71] Liu BF, Sera Y, Matsubara N, Otsuka K, Terabe S. Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis. Electrophoresis 2003;24:3260–3265. [CrossRef]

[72] Zhang D, He J, Zhao Y, Luo Z, Du M. Global plus local: a complete framework for feature extraction and recognition. Pattern Recognit 2014;47:1433–1442. [CrossRef]

Appendix A. Description of feature sets extracted in Step 2.

		HCV	**HBV**
Feature Set	**Description**	**Chi-Square Test Statistics**[a] **(p-value)**	**Chi-Square Test Statistics**[a] **(p-value)**
wPE$_{(A,C,G,T)}$			
SB$_1$_PE$_{(A)}$	PE for sub-band 1 of the base A	98.791	271.238
SB$_2$_PE$_{(A)}$	PE for sub-band 2 of the base A	(<0.000)	(<0.000)
SB$_3$_PE$_{(A)}$	PE for sub-band 3 of the base A		
SB$_4$_PE$_{(A)}$	PE for sub-band 4 of the base A		
SB$_1$_PE$_{(C)}$	PE for sub-band 1 of the base C	138.899	243.668
SB$_2$_PE$_{(C)}$	PE for sub-band 2 of the base C	(<0.000)	(<0.000)
SB$_3$_PE$_{(C)}$	PE for sub-band 3 of the base C		
SB$_4$_PE$_{(C)}$	PE for sub-band 4 of the base C		
SB$_1$_PE$_{(G)}$	PE for sub-band 1 of the base G	140.171	275.139
SB$_2$_PE$_{(G)}$	PE for sub-band 2 of the base G	(<0.000)	(<0.000)
SB$_3$_PE$_{(G)}$	PE for sub-band 3 of the base G		
SB$_4$_PE$_{(G)}$	PE for sub-band 4 of the base G		
SB$_1$_PE$_{(T)}$	PE for sub-band 1 of the base T	105.363	242.486
SB$_2$_PE$_{(T)}$	PE for sub-band 2 of the base T	(<0.000)	(<0.000)
SB$_3$_PE$_{(T)}$	PE for sub-band 3 of the base T		
SB$_4$_PE$_{(T)}$	PE for sub-band 4 of the base T		
wMPE2$_{(A,C,G,T)}$			
SB$_1$_MPE2$_{(A)}$	Multi-scale PE (s=2) for sub-band 1 of the base A	25.877	254.360
SB$_2$_ MPE2$_{(A)}$	Multi-scale PE (s=2) for sub-band 2 of the base A	(<0.000)	(<0.000)
SB$_3$_ MPE2$_{(A)}$	Multi-scale PE (s=2) for sub-band 3 of the base A		
SB$_4$_ MPE2$_{(A)}$	Multi-scale PE (s=2) for sub-band 4 of the base A		
SB$_1$_ MPE2$_{(C)}$	Multi-scale PE (s=2) for sub-band 1 of the base C	54.433	237.080
SB$_2$_ MPE2$_{(C)}$	Multi-scale PE (s=2) for sub-band 2 of the base C	(<0.000)	(<0.000)
SB$_3$_ MPE2$_{(C)}$	Multi-scale PE (s=2) for sub-band 3 of the base C		
SB$_4$_ MPE2$_{(C)}$	Multi-scale PE (s=2) for sub-band 4 of the base C		
SB$_1$_ MPE2$_{(G)}$	Multi-scale PE (s=2) for sub-band 1 of the base G	62.145	275.357
SB$_2$_ MPE2$_{(G)}$	Multi-scale PE (s=2) for sub-band 2 of the base G	(<0.000)	(<0.000)
SB$_3$_ MPE2$_{(G)}$	Multi-scale PE (s=2) for sub-band 3 of the base G		
SB$_4$_ MPE2$_{(G)}$	Multi-scale PE (s=2) for sub-band 4 of the base G		
SB$_1$_ MPE2$_{(T)}$	Multi-scale PE (s=2) for sub-band 1 of the base T	27.688	236.742
SB$_2$_ MPE2$_{(T)}$	Multi-scale PE (s=2) for sub-band 2 of the base T	(<0.000)	(<0.000)
SB$_3$_ MPE2$_{(T)}$	Multi-scale PE (s=2) for sub-band 3 of the base T		
SB$_4$_ MPE2$_{(T)}$	Multi-scale PE (s=2) for sub-band 4 of the base T		
wMPE3$_{(A,C,G,T)}$			
SB$_1$_MPE3$_{(A)}$	Multi-scale PE (s=3) for sub-band 1 of the base A	71.140	322.014
SB$_2$_ MPE3$_{(A)}$	Multi-scale PE (s=3) for sub-band 2 of the base A	(<0.000)	(<0.000)
SB$_3$_ MPE3$_{(A)}$	Multi-scale PE (s=3) for sub-band 3 of the base A		
SB$_4$_ MPE3$_{(A)}$	Multi-scale PE (s=3) for sub-band 4 of the base A		

		HCV	**HBV**
Feature Set	**Description**	**Chi-Square Test Statistics**[a] **(p-value)**	**Chi-Square Test Statistics**[a] **(p-value)**
SB_{1}– $MPE3_{(C)}$	Multi-scale PE (s=3) for sub-band 1 of the base C	115.019	221.649
SB_{2}– $MPE3_{(C)}$	Multi-scale PE (s=3) for sub-band 2 of the base C	(<0.000)	(<0.000)
SB_{3}– $MPE3_{(C)}$	Multi-scale PE (s=3) for sub-band 3 of the base C		
SB_{4}– $MPE3_{(C)}$	Multi-scale PE (s=3) for sub-band 4 of the base C		
SB_{1}– $MPE3_{(G)}$	Multi-scale PE (s=3) for sub-band 1 of the base G	76.747	276.995
SB_{2}– $MPE3_{(G)}$	Multi-scale PE (s=3) for sub-band 2 of the base G	(<0.000)	(<0.000)
SB_{3}– $MPE3_{(G)}$	Multi-scale PE (s=3) for sub-band 3 of the base G		
SB_{4}– $MPE3_{(G)}$	Multi-scale PE (s=3) for sub-band 4 of the base G		
SB_{1}– $MPE3_{(T)}$	Multi-scale PE (s=3) for sub-band 1 of the base T	81.993	229.165
SB_{2}– $MPE3_{(T)}$	Multi-scale PE (s=3) for sub-band 2 of the base T	(<0.000)	(<0.000)
SB_{3}– $MPE3_{(T)}$	Multi-scale PE (s=3) for sub-band 3 of the base T		
SB_{4}– $MPE3_{(T)}$	Multi-scale PE (s=3) for sub-band 4 of the base T		

a. Kruskal Wallis Test