Identification of Pathogenic Viruses Using Genomic Cepstral Coefficients with Radial Basis Function Neural Network

Emmanuel Adetiba, Oludayo O. Olugbara and Tunmike B. Taiwo

Abstract Human populations are constantly inundated with viruses, some of which are responsible for various deadly diseases. Molecular biology approaches have been employed extensively to identify pathogenic viruses despite the limitations of the approaches. Nevertheless, recent advances in the next generation sequencing technologies have led to a surge in viral genome sequence databases with potentials for Bioinformatics based virus identification. In this study, we have utilised the Gaussian radial basis function neural network to identify pathogenic viruses. To validate the neural network model, samples of sequences of four different pathogenic viruses were extracted from the ViPR corpus. Electron-ion interaction pseudopotential scheme was used to encode the extracted sample sequences while cepstral analysis technique was applied to the encoded sequences to obtain a new set of genomic features, here called Genomic Cepstral Coefficients (GCCs). Experiments were performed to determine the potency of the GCCs to discriminate between different pathogenic viruses. Results show that GCCs are highly discriminating and gave good results when applied to identify some selected pathogenic viruses.

Keywords Cepstral · Dengue · Ebolavirus · Electron · Enterovirus · Genomics · Hepatitis · Radial · Neural · Network

E. Adetiba (✉) · O.O. Olugbara · T.B. Taiwo
ICT and Society Research Group, Durban University of Technology, 1334, Durban 4000, South Africa
e-mail: emmanuela1@dut.ac.za

O.O. Olugbara
e-mail: oludayoo@dut.ac.za

T.B. Taiwo
e-mail: tunmike.bukola@yahoo.com

N. Pillay et al. (eds.), *Advances in Nature and Biologically Inspired Computing*, Advances in Intelligent Systems and Computing 419,
DOI 10.1007/978-3-319-27400-3_25

1 Introduction

Application of advanced technologies in molecular biology has greatly accelerated the ease with which known disease pathogens were identified within the last few years. Novel pathogens were discovered with ease using molecular approaches such as immune screening of cDNA libraries and polymerase chain reactions. Examples of pathogenic viruses that were discovered as a result of these efforts are the "hepatitis C" and "sin nombre" [1]. Despite this huge success, the effective identification of several viral pathogens has been elusive. On this account, the development of new techniques for identification of pathogens has become of interest. Modern DNA sequencing technologies hold promise because of the avalanche of genomic sequences of viral from laboratory and environmental surveillance studies that are made publicly available online. The availability of huge datasets has made automatic identification of species of the DNA sequences, an open challenge in Bioinformatics and Genomic Signal Processing (GSP) [2, 3].

In GSP, which is a somewhat new area in Bioinformatics, digital signal processing techniques are employed to analyse genomic data and the biological knowledge gained can be translated to a system based application [3]. Several studies reported in the literature have addressed the identification of species from nucleotide sequences using the digital signal processing techniques alongside the machine learning techniques. A classification model based on data mining and Artificial Neural Network (ANN) was developed for the identification of species from DNA sequences [3]. The authors mined nucleotide patterns from selected DNA sequences and used Multilinear Principal Component Analysis (MPCA) to reduce the dimensions of the mined patterns. The technique was validated on two different species and they reported good classification accuracies. In [4], a classification model was developed to classify eight different eukaryotes species. The authors utilized Frequency Chaos Game Representation (FCGR) to encode genomic sequences and they utilised a neural network technique to obtain the classification accuracy of 92.3 %. A model based on the Naïve Bayesian technique was used to classify archeal and bacterial genomes [5]. The authors used the dinucleotide composition of the genome sequences to report a classification accuracy of 85 %. In [6], the classification of proteins of three different species was reported. The authors used a Markov model to obtain classification accuracies 83.51, 82.12 and 66.63 % of the proteins of microbes, eukaryotes and archaea respectively.

However, the success of any genome identification (or classification) system is hugely dependent on critical factors such as the availability of valid datasets, feature extraction method that truly reflects the attributes of the genetic sequences, the classification algorithm and the objective evaluation of the identification system [7]. The immediate motivation for the study at hand, is the need to obtain a set of discriminating genomic features to improve the identification of species in genomic sequences [7]. The genomic sequences of four pathogenic viruses were extracted from the Virus Pathogen Database and Analysis Resource (ViPR) corpus. The extracted genomic sequences were numerically encoded using a low computational

Electron-Ion Interaction Potential (EIIP) scheme. Thereafter, a set of Genomic Ceptral Coefficients (GCCs) was computed from the encoded sequences and transmitted to the Gaussian Radial Basis Function (RBF) neural network to learn the genome sequences. Section 2 of this paper contains materials and methods, Sect. 3 contain the results and discussion and the paper concludes in Sect. 4.

2 Materials and Methods

In this section, we describe the viral dataset, the EIIP nucleotide mapping scheme, the cepstral analysis technique on which the GCCs is based, the RBF neural network classifier and the experiments performed. All the computational techniques described were implemented in MATLAB R2012a.

2.1 Dataset

We extracted genome sequences of featured viruses from the Virus Pathogen Database and Analysis Resource (ViPR) corpus for our experimentations. The ViPR corpus provides free access to records of gene sequences and proteins of various viral pathogens so as to facilitate research and development of diagnostics, vaccines and therapeutics. The extracted viruses are Ebolavirus, Dengue virus, Hepatitis C and Enterovirus D68. These viruses are currently classified as featured viruses on the ViPR databases because they are responsible for diseases that are presently attracting serious attention from health sectors, scientists and governments worldwide [22]. Complete genome sequences of seven strains of the Bundibuggo specie of Ebolavirus were extracted while ten complete genomes of each of the remaining three viruses were extracted from the ViPR for our experiments. We extracted a small dataset (37 instances of virus sequences) in order to examine the efficacy of the GCCs. The sequence length of each of these viruses is shown in Table 1. As illustrated in the Table, the length of each of the genomic sequences varies from one virus to the other and from one strain of the virus to another strain of the same virus. For instance, the sequence length of the Ebolavirus varies from 18939 to 18941 while that of Enterovirus D68 varies from 420 to 809.

Table 1 The range of the genome sequence length of the selected viruses

Virus	Range of the length of genome sequence
Ebolavirus	18939–18941
Enterovirus D68	420–809
Dengue virus	10176–15287
Hepatitis C virus	9220–9587

2.2 Electron-Ion Interaction Potential (EIIP)

Nucleotide sequence is a string of four distinct characters representing four nucleotides, which are A (Adenine), C (Cytosine), G (Guanine) and T (Thymine) [2, 8]. To apply DSP techniques to process these characters, it is necessary to first convert them into numeric sequences. This problem was solved by Voss [9] using four binary indicator sequences, which are $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$. These indicator sequences take the value of one or zero depending on whether or not the corresponding character exists at n location. These four binary indicator sequences are said to contain some redundancy and they can be transformed into three non-redundant sequences as reported in [10, 11]. Moreover, the authors in [12] proposed a nucleotide encoding scheme by replacing the four indicator sequences with just one sequence and they named the scheme "EIIP indicator sequence". Doing so, they calculated the energy of delocalized electrons in amino acid and nucleotides as the Electron-Ion Interaction Pseudopotential (EIIP). Substituting the EIIP values for A, C, G and T in a nucleotide string $x[n]$, we obtain the "EIIP indicator sequence" that represents the distribution of the energies of free electrons along the sequence [13]. The EIIP values for the four nucleotides are shown in Table 2, where A = 0.1260, G = 0.0806, C = 0.1340 and T = 0.1335. Using the EIIP scheme, the computational overhead of binary indicator sequence is significantly reduced by 75 % [12, 13].

The corresponding Discrete Fourier Transform (DFT) of EIIP encoded sequence $u_i[n]$ where i = A, G, C or T is:

$$U_i[k] = \sum_{n=0}^{N-1} u_i[n] e^{\frac{-j2\pi kn}{N}}, \quad k = 0, 1, 2, \ldots N-1 \tag{1}$$

and the power spectrum is defined as:

$$P[k] = |U_A[k]|^2 + |U_G[k]|^2 + |U_C[k]|^2 + |U_T[k]|^2 \tag{2}$$

2.3 Cepstral Analysis Technique

A cepstrum is defined as the inverse DFT of the logarithmic magnitude of the DFT of a signal as illustrated with a block diagram in Fig. 1. In other words, a cepstrum

Table 2 Electron-Ion Interaction Pseudopotential for Nucleotides

Nucleotide	EIIP value
A	0.1260
G	0.0806
C	0.1340
T	0.1335

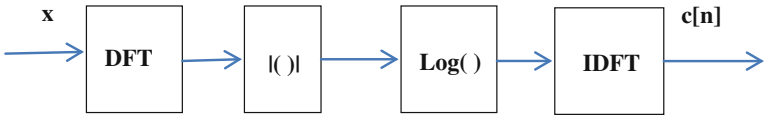

Fig. 1 Block diagram of the computational components of a signal cepstrum

can be considered as a spectrum of logarithmic spectrum that comprises of logarithmic amplitude scale, but linear frequency scale [14]. Cepstrum analysis is used as a tool for detection of periodicity in a spectrum because the harmonic structure of the spectrum is emphasized by the logarithmic amplitude scale. Areas where cepstral coefficients are applied include radar analysis, speech processing, marine exploration, diastolic heart sound analysis and electroencephalogram pattern classification [14–16]. This current study utilizes cepstral analysis to obtain Genomic Cepstral Coefficients (GCCs) for identification of pathogenic viruses from their genome sequences. This effort strongly aligns with the practice in the field of Genomic Signal Processing (GSP) in which DSP techniques are applied to solve biological problems based on nucleotide sequences [9–12].

In this study, the real and complex cepstral [15] are considered to obtain real and complex genomic cepstral features.

The real cepstrum of a signal $x[n]$ is calculated by computing the natural logarithm of the magnitude of its Fourier transform and taking the inverse Fourier transform of the result given as:

$$c_x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left|X(e^{jw})\right| e^{jwn} dw \tag{3}$$

The complex cepstrum of the signal is also computed by calculating the complex natural logarithm of the Fourier transform and taking the inverse Fourier transform of the result using:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\log\left|X(e^{jw})\right| e^{jw} + j \arg(X(e^{jw}))] e^{jwn} dw \tag{4}$$

The phase of the signal is represented as *arg*() in Eq. (4). Both the real and the complex cepstral analysis produce cepstral coefficients and truncating the coefficients at different linear frequency scale allows the preservation of different amount of spectral details. It has been reported in the literature that the first 12 to 15 coefficients of a cepstrum are a compact representation of the spectral envelope [16].

Radial Basis Function Neural Network

An ANN is a parallel biologically inspired computational system that mimics the configuration of the human nervous system. The human brain, which is the central

organ in the nervous system is made up of 10^{14}–10^{15} interconnections of neurons and learning involves adjustments to the synaptic connections between these neurons [17]. Artificial neuron was motivated principally from the structure and functions of the human brain and in order to learn problem that cannot be handled by one neuron, an aggregate of several neurons called ANN are engaged. The two main types of ANN architectures are the feed-forward and recurrent networks. In feed-forward architecture, signal flows from input to output stringently in a forward path while there are feedback paths in recurrent networks.

Radial Basis Function (RBF) and Multilayer Perceptron (MLP) neural network are feed-forward networks that have been used extensively in the Bioinformatics and artificial intelligence research communities for pattern classification [18–20]. However, we selected the RBF neural network in this study as a biologically inspired computational platform to identify the GCCs extracted from selected pathogenic viruses. This is because the RBF neural network is reported to be very robust to input noise, always guarantee high accuracy and have lower design rigour than MLP neural network [21]. The input layer of the RBF neural network receives input signal $\boldsymbol{x}$ and transmits it to the hidden layers where the signal is processed and further transmitted in a forward direction to the output layer to generate the output signal $\boldsymbol{y}$. To process information received in the hidden layer, an RBF network utilizes several kernel functions such as Gaussian, cubic, thin plate spline, Cauchy and inverse multiquadric. However, one of the most commonly used kernel is the Gaussian function, which is used in this study. This kernel function is represented as:

$$f(x) = \exp(-\frac{x^2}{\sigma^2}) \tag{6}$$

Where σ is the width or scaling parameter that characterises the input space under the influence of the basis function. The output $\boldsymbol{y}$, from the RBF neural network is defined as:

$$y_k = \sum_{j=1}^{N} \sigma_{j,k} f(\left\| c_j - x \right\|_2) \tag{7}$$

Where c_j is the centroid of the jth basis function, N is the number of neurons in the hidden layer and $||x||$ is the radial distance of its argument, which is usually taken as Euclidean distance.

The RBF neural network used in this study was configured to appropriately identify virus sequences in our experimental dataset of GCCs. The number of neurons in the input layer vary from 12 to 15 because this range of elements was tested for the GCCs. The number of neurons in the hidden layer by default is equal to the number of instances in the training dataset which is equal to 37 in this study. The dataset is partitioned to 70 % training, 15 % validation and 15 % testing. Meanwhile, the output layer contains 4 neurons because there are four virus classes in our experimentation dataset with each virus representing a class. The other

configurations of the RBF neural network in this study are MSE goal of 0 and spread of 0.1. To perform an evaluation of the results computed by the RBF neural network, we utilized four different widely used metrics, which are the accuracy, Mean Square Error (MSE), sensitivity and specificity [8, 23].

Experiments

In this study, two experiments were performed to determine the potency of complex and real GCCs to discriminate between different pathogenic viruses. The purpose of the first experiment is to use the complex GCCs to identify pathogenic viruses, while the purpose of the second experiment is to use real GCCs to identify pathogenic viruses. In the experiments, the number of GCCs used as features varied from 12 to 15 in order to determine their effects on the identification accuracy.

In the first experiment, the nucleotide sequences of all the virus sequences were first encoded using the EIIP scheme. Thereafter, we obtained the vectors of complex GCCs, which were transmitted to train the configured RBF neural network to learn virus identification task. In the second experiment, we used the same nucleotide encoding scheme (EIIP) to encode the genome sequences of pathogenic viruses. The real GCCs were thereafter computed to train the RBF neural network to learn the task of virus identification.

3 Experimental Results and Discussion

The spectral plots of the complex GCCs for all the viruses obtained from the first experiment are shown in Fig. 2. These plots clearly show that the complex cepstrum of each virus is unique. For instance, the spectral shape of Dengue virus has a downward peak at the end of the spectrum. The Ebolavirus spectral shape shows both upward and downward peaks at the end of the spectrum. The Enterovirus D68 has a dense spectrum with peaks at the beginning and terminates with a downward peak. The Hepatitis C virus has intermittent dense spectral details across the length of the spectrum that terminates in an upward spike.

In this first experiment, although the number of GCCs varied from 12, 13, 14 to 15, when training the RBF neural network to learn GCCs, we obtained the same results for all the performance metrics. The identification results yield an accuracy of 83.3 %, MSE of 0.0497, sensitivity of 0.9063 and specificity of 0.9520 as shown in Table 3. These results imply that retaining any number of elements of the complex GCCs from 12 to 15 does not affect the performance of the pathogenic virus identification system.

Figure 3 shows the plots of the real GCCs for all the virus sequences in the second experiment. These plots show that the real GCCs of each virus is unique. It can be observed from the plots that Dengue virus has many tiny spikes across the length of the spectrum, Ebola virus has a flat spectrum with a few tiny spikes, Enterovirus has a zig-zag spectrum while Hepatitis C virus has a dense spectrum with a spectral envelope of low amplitude.

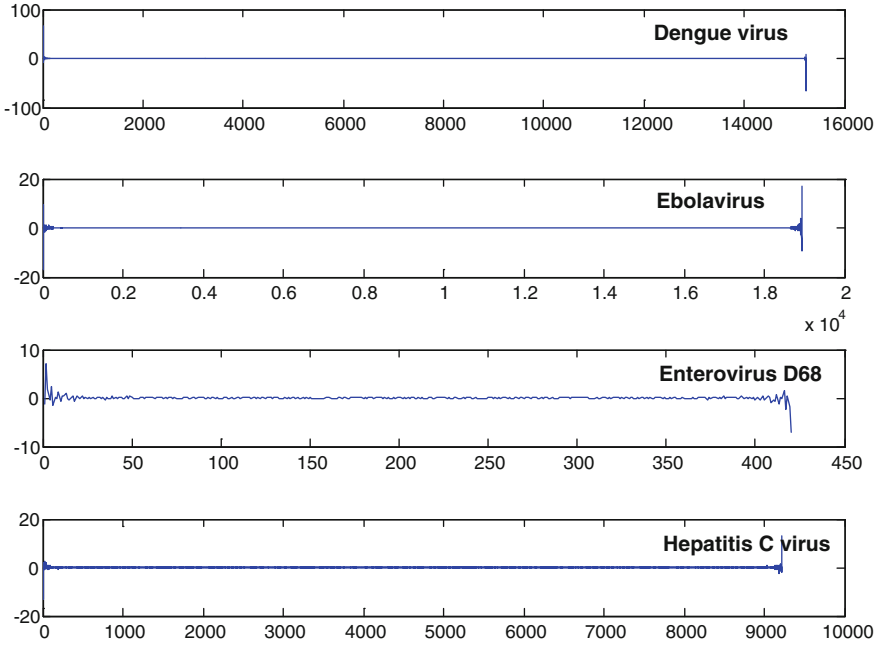

Fig. 2 The complex Genomic Cepstral Coefficients of the four viruses

Table 3 Results of the first experiment

No of elements in GCCs	Accuracy	MSE	Sensitivity	Specificity
12	0.8330	0.0497	0.9063	0.9520
13	0.8330	0.0497	0.9063	0.9520
14	0.8330	0.0497	0.9063	0.9520
15	0.8330	0.0497	0.9063	0.9520

The results obtained in this second experiment after training the RBF neural network to learn the real GCCs are tabulated in Table 4. These results show that the same performance values were obtained for real GCCs with 12 and 13 elements, whereas the results improved for real GCCs with 14 elements. There is also an improvement in the results with respect to sensitivity and specificity when the numbers of elements in the real GCCs were increased from 14 to 15. The best performance accuracy of 97.3 %, MSE of 0.0309, sensitivity of 0.9919 and specificity of 0.9919 were obtained with 15 element real GCCs.

Comparatively, Tables 3 and 4 show the results of real GCCs to be better than the results obtained with complex GCCs, which led to the realization of our experimental objective. The result of this study is significant because each pathogenic virus with sequence length ranges of 18939–18941, 420–809, 10176–15287,

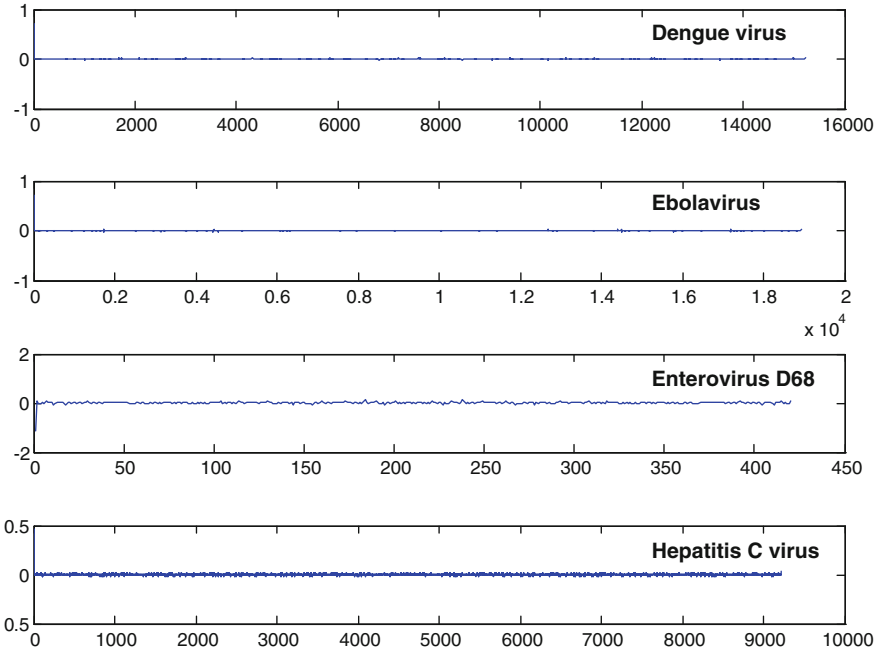

Fig. 3 The real Genomic Cepstral Coefficients of the four viruses

Table 4 Results of the second experiment

No of elements in GCCs	Accuracy	MSE	Sensitivity	Specificity
12	0.9460	0.0541	0.9393	0.9824
13	0.9460	0.0541	0.9393	0.9824
14	0.9730	0.0309	0.9773	0.9773
15	0.9730	0.0309	0.9919	0.9919

and 9220–9587 can be represented using 15 element real GCCs with good discriminating power. Despite an extensive literature search for a similar study that used the same dataset and neural network for pathogenic virus identification, the closest work we found was Karthika et al. [4]. The authors used FCGR to encode the genomic sequences of eight eukaryote species and neural network as the classifier to obtain an accuracy of 92.3 %. This previous study in [4] corroborates our position of the need for discriminating features from genomic sequences for species identification. Hence, the high level of accuracy, sensitivity, specificity and low MSE achieved in this study is a promising endeavour. This result forms the basis for proposing a genomic computational model that incorporates the EIIP scheme, 15 element real GCCs and RBF neural network for accurate identification of pathogenic viruses from genome sequences.

4 Conclusion

In this study, we have been able to apply cepstral analysis technique to obtain genomic cepstral coefficients from genome sequences of pathogenic viruses. These features were incorporated into the proposed computational model for accurate identification of pathogenic viruses from genome sequences. The efficacy of the model has been validated using appropriate data and evaluation metrics. The implementation of the model holds a lot of promises for genomic based diagnosis of microbial diseases in humans, DNA Barcoding, bio-diversity study and wildlife forensic. In the future, we hope to improve the robustness of the proposed model by incorporating more pathogenic viruses and validating the model on a large data set. In addition, we hope to experiment with other classification algorithms that can enhance the performance of the model when more pathogenic virus sequences are elicited.

References

1. Wang, D., Urisman, A., Liu, Y.T., Springer, M., Ksiazek, T.G., Erdman, D.D., DeRisi, J.L.: Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. **1**(2), 257–260 (2003)
2. Mabrouk, M.S.: A study of the potential of EIIP mapping method in exon prediction using the frequency domain techniques. Am. J. Biomed. Eng. **2**(2), 17–22 (2012)
3. Sathish Kumar, S., Duraipandian, N.: An effective identification of species from DNA sequence: a classification technique by integrating DM and ANN. Int. J. Adv. Comput. Sci. Appl. **3**(8), 104–114 (2012)
4. Karthika, V., Nair, V.V., Gopinath, D.P.: Classification of organisms using frequency-chaos game representation of genomic sequences and ANN. In: Proceedings of the 10th National Conference on Technological Trends (NCTT09), pp. 243–247, 6–7 Nov 2009
5. Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I., Coster, J.: Capturing Whole —Genome characteristics in short sequences using a naive Bayesian classifier. Genome Res. **11**, 1404–1409 (2001)
6. Zanoguera, F., De Francesco, M.: Protein classification into domains of life using Markov chain models. In: Proceeding of the Computational Systems Bioinformatics Conference, pp. 517–519 (2004)
7. Song, C., Shi, F.: Prediction of protein subcellular localization based on Hilbert-Huang transform. Wuhan Univ. J. Nat. Sci. **17**(1), 48–54 (2012)
8. Adetiba, E., Olugbara, O.O.: Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. Sci. World J. **2015**, id. 786013, 1–17 (2015)
9. Voss, R.F.: Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. **68**(25), 3805 (1992)
10. Anastassiou, D.: Frequency-domain analysis of biomelecular sequences. Bioinformatics **16** (12), 1073–1081 (2000)
11. Anastassiou, D.: Genomic signal processing. IEEE Signal Process. Mag. **18**(4), 8–20 (2001)
12. Nair, A.S., Sreenadhan, S.P.: A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation **1**(6), 197–202 (2006)

13. Pirogova, E., Simon, G.P., Cosic, I.: Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model. IEEE Trans. Nanobiosci. **2**, 63–69 (2003)
14. Akay, M.: Biomedical Signal Processing, pp. 113–135. Academic Press (2012)
15. Oppenheim, A.V., Schafer, R.W.: Digital Signal Processing. Englewood Cliffs, Prentice-Hall (1975)
16. Thakur, S., Adetiba, E., Olugbara, O.O., Millham, R.: Experimentation using short-term spectral features for secure mobile internet voting authentication. Math. Prob. Eng. **2015**, id. 564904, 1– 21 (2015)
17. Adetiba, E., Ekeh, J.C., Matthews, V.O., Daramola, S.A., Eleanya, M.E.U.: Estimating an optimal backpropagation algorithm for training an ANN with the EGFR Exon 19 nucleotide sequence: an electronic diagnostic basis for non-small cell lung cancer (NSCLC). J. Emerg. Trends Eng. Appl. Sci. **2**(1), 74–78 (2011)
18. Kurban, T., Beşdok, E.: A comparison of RBF neural network training algorithms for inertial sensor based terrain classification. Sensors **9**(8), 6312–6329 (2009)
19. Oyang, Y.J., Hwang, S.C., Ou, Y.Y., Chen, C.Y., Chen, Z.W.: Data classification with radial basis function networks based on a novel kernel density estimation algorithm. IEEE Trans. Neural Netw. **16**, 225–236 (2005)
20. Lee, C.C., Chung, P.C., Tsai, J.R., Chang, C.I.: Robust radial basis function neural networks. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **29**(6), 674–685 (1999)
21. Derks, E.P.P.A., Pastor, M.S., Buydens, L.M.C.: Robustness analysis of radial base function and multi-layered feed-forward neural network models. Chemometr. Intell. Lab. Syst. **28**(1), 49–60 (1995)
22. Pickett, B.E., Greer, D.S., Zhang, Y.: Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses **4**, 3209–3226 (2012)
23. Zou, K.H., O'Malley, A.J., Mauri, L.: Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation **115**(5), 654–657 (2007)