


RESEARCH

Open Access



# Automated detection of colon cancer using genomic signal processing

Safaa M. Naeem<sup>1</sup>, Mai S. Mabrouk<sup>2</sup>, Mohamed A. Eldosoky<sup>1</sup> and Ahmed Y. Sayed<sup>3\*</sup> 

## Abstract

**Background:** Disorders in deoxyribonucleic acid (DNA) mutations are the common cause of colon cancer. Detection of these mutations is the first step in colon cancer diagnosis. Differentiation among normal and cancerous colon gene sequences is a method used for mutation identification. Early detection of this type of disease can avoid complications that can lead to death. In this study, 55 healthy and 55 cancerous genes for colon cells obtained from the national center for biotechnology information GenBank are used. After applying the electron–ion interaction pseudopotential (EIP) numbering representation method for the sequences, single-level discrete wavelet transform (DWT) is applied using Haar wavelet. Then, some statistical features are obtained from the wavelet domain. These features are mean, variance, standard deviation, autocorrelation, entropy, skewness, and kurtosis. The resulting values are applied to the k-nearest neighbor (KNN) and support vector machine (SVM) algorithms to obtain *satisfactory* classification results.

**Results:** Four important parameters are calculated to evaluate the performance of the classifiers. Accuracy (ACC), F1 score, and Matthews correlation coefficient (MCC) are 95%, 94.74%, and 0.9045%, respectively, for SVM and 97.5%, 97.44%, and 0.9512%, respectively, for KNN.

**Conclusion:** This study has created a novel successful system for colorectal cancer classification and detection with the well-satisfied results. The K-nearest network results are the best with low error for the generated classification system, even though the results of the SVM network are acceptable.

**Keywords:** Colon cancer, Electron–ion interaction pseudopotential mapping method, Genomic signal processing, Discrete wavelet transform, Statistical features, Support vector machine, k-nearest neighbor

## Background

### Colorectal cancer

Colon cancer is the third most widespread cancer that affects both genders, after prostate cancer in men, breast cancer in women, and lung cancer in both sexes. Colon cancer starts in the large intestine (colon), the final part of the digestive tract [1, 2]. It usually affects older adults, although it can occur at any age. It typically starts as small, noncancerous clumps of cells called polyps, which form inside the colon. Over time some of these polyps

can convert into colon cancers [3]. Colon cancer is occasionally called colorectal cancer, a term, which merges colon cancer and rectal cancer that begins in the rectum [4].

### Recent research

In recent decades, researchers have used genomic signal processing (GSP) methods to solve a range of bioinformatics problems. This research falls into five broad parts. Firstly, the research could be performing cluster analysis of deoxyribonucleic acid (DNA) sequences [5], breast cancer diagnosis and detection using Wisconsin diagnostic breast cancer Database [6–9], cancer diagnosis and classification using DNA microarray technology [8, 10, 11] and classifying any gene sequence into diseased/

\*Correspondence: ahmed\_badr@m-eng.helwan.edu.eg

<sup>3</sup> Department of Engineering Mathematics and Physics, Faculty of Engineering El-Matara, Halwan University, Cairo, Egypt

Full list of author information is available at the end of the article

non-diseased state based on trinucleotide repeat disorders using DNA sequences [12].

Secondly, the mapping method applied to DNA sequences could be a Voss representation [5, 12, 13] or the EIIP method [13, 14].

Thirdly, the GSP algorithms used could be discrete Fourier transform (DFT) [5, 13], power spectral density (PSD) [5, 13], discrete wavelet transform (DWT) [12], moment invariants [14], statistical parameters [12], or fast Fourier transform [12].

Fourthly, the classifier used in the research could be the k-means algorithm [5, 7], linear discriminant analysis and support vector machine (SVM) [6, 7, 15], a Naive-Bayes classifier (NB) [7, 10], a deep convolutional neural network (CNN) [16, 17], multilayer perceptron (MLP) [7, 9], inception recurrent residual convolutional neural network [18, 19], probabilistic neural network [8], classification and regression tree (CART) [7], simple linear iterative clustering (SLIC), or optimal deep neural network (ODNN) [20].

Finally, the evaluation method used could be plotting [5], comparison [7], using improved binary particle swarm optimization (iBPSO) [10], calculation of the area under the curve of the receiver operating characteristic curve [17, 18], and calculating sensitivity, specificity, and accuracy [17, 20].

**Methods**

**Database sequence**

A vital source of genomic data is the search and retrieval system created in the NCBI GenBank [21] at the National Institutes of Health. In this research, 55 healthy genes and 55 cancerous genes of the colon are used. Each DNA sequence has a length of 400 nucleotides. The following is an example of the output of cancer data read by the *fastaread* function in *MATLAB R2017b*:

**Sequence = 'GCGATCGCCATGGCGGTGCAGCCGAAGGAGACGCTGCGAGTTGGAGAGCGCGGCCGAGGTGGGCTTCGTGCGCTTCTTCA.....etc.'**

*Description = 'AB489153.1 Synthetic construct DNA, clone: pF1KB3091, Homo sapiens MSH2 gene for mutS homolog 2, colon cancer, nonpolyposis type 1, without stop codon, in Flexi system'*

**Mapping method**

The electron-ion interaction pseudopotential (EIIP) numerical method is the most common representation rule used by many researchers [22–25]. The EIIP numerical values for A, G, C, and T in a DNA string are 0.1260, 0.0806, 0.1340, and 0.1335. These values represent the free electron energy distribution along the DNA sequence [26]. For example, if  $Y[n] = TATGGATCC$ , the corresponding EIIP numerical values,  $Y[e]$ , will be:

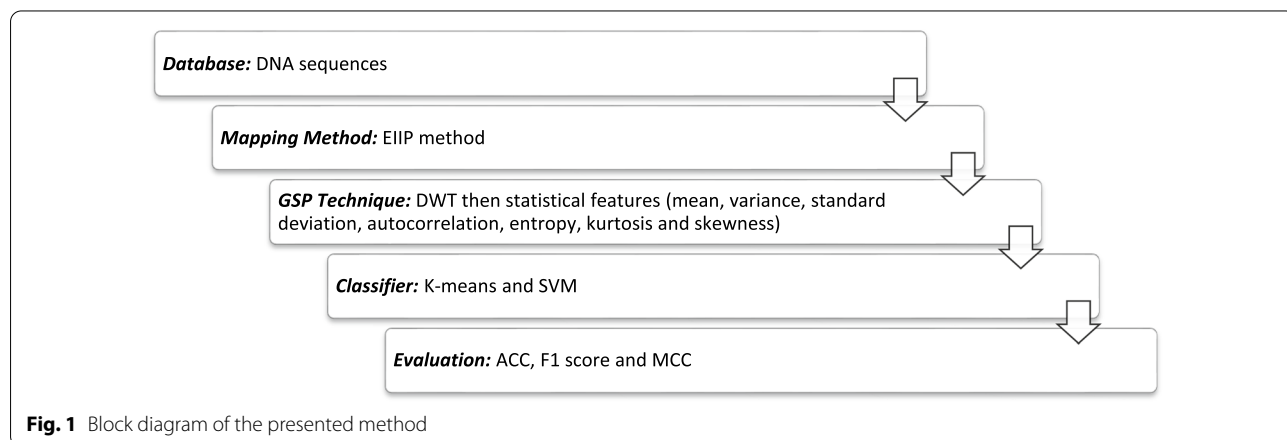
$$Y_e[n] = [0.1335 \ 0.1260 \ 0.1335 \ 0.0806 \ 0.0806 \ 0.1260 \ 0.1335 \ 0.1340 \ 0.1340] \tag{1}$$

In this study, a combination of those ideas was used, as well as other methods that are not listed above, for example, using DNA sequences as the classification database and using the k-nearest neighbor as the classifier. The following block diagram depicts the study steps, which are explained in detail later (Fig. 1).

**Genomic signal processing techniques**

1. Discrete Wavelet Transform

DWT transforms a signal into a group of basis functions called wavelets. DWT converts a discrete-time signal to its wavelet representation [27]. For DWT,



**Fig. 1** Block diagram of the presented method

there are various wavelets, which are widely divided into orthogonal and biorthogonal wavelets [28]. The orthogonal type was introduced by Hungarian mathematician Alfréd Haar [29]. The Haar DWT transform of a signal ( $S$ ) is generated by crossing it over a group of filters [30]. These are produced by passing a signal through a low-pass filter with impulse response ( $g$ ) resulting in a convolution, as follows:

$$F[m] = (S * g)[m] = \sum_{k=-\infty}^{\infty} S[m]g[m - k] \quad (2)$$

The signal is also passed over a high-pass filter ( $h$ ). The result gives two components, the first one, from the high-pass filter, is called the detail coefficients, and the other, from the low-pass filter, is called the approximation coefficients [31, 32]. In Fig. 2, the two filters are known as quadrature mirror filters, and they are linked to each other.

According to the rule of Nyquist, half of the signal frequencies are removed. As a result, the output of the low-pass filter in Fig. 2 is subsampled by two and processed by crossing it for another time over a new low-pass filter,  $g$ , and a new high-pass filter,  $h$ , with half cutoff frequency, as follows:

$$F_{low}[m] = \sum_{k=-\infty}^{\infty} S[m]g[2m - k] \quad (3)$$

$$F_{high}[m] = \sum_{k=-\infty}^{\infty} S[m]h[2m - k] \quad (4)$$

## 2. Statistical Features

After obtaining the DWT coefficients, some statistical features are extracted as follows:

### Mean

The arithmetical mean, called the average or the mathematical expectation, is the centric value of a group of numbers [33, 34]. To calculate the mean value  $\mu$  of a sequence  $S = [s_1, s_2, s_3, \dots, s_M]$  with length  $M$ , divide the sum of all sequence values by its length as in the following equation:

$$\mu = \frac{\sum_{i=1}^M s_i}{M} \quad (5)$$

### Variance

Variance ( $\sigma^2$ ) in probability theory and statistics is defined as the squared deviation expectation of a random variable from its mean. Informally, it quantifies how far a sequence of arbitrary numbers diverges from the mean value of the sequence [35]. It is determined by taking the differences between each number in the set and the mean. Then, the distinctions are squared, to make them positive. Finally, the sum of the squares is divided by the number of values in the set, as follows:

$$\sigma^2 = \frac{\sum_{i=1}^M (s_i - \mu)^2}{M} \quad (6)$$

where  $s_i$  is the  $i$ th data point,  $\mu$  is the mean of all data points, and  $M$  is the number of data points.

### Standard deviation

The standard deviation ( $\sigma$ ) is the square root of the variance ( $\sigma^2$ ) [35].

$$\sigma = \sqrt{\frac{\sum_{i=1}^M (s_i - \mu)^2}{M}} \quad (7)$$

where  $s_i$  is the  $i$ th data point,  $\mu$  is the mean of all data points, and  $M$  is the number of data points.

### Autocorrelation

Autocorrelation, also called serial correlation, is the association of a signal with a later copy of itself obtained via a delay function. It is a mathematical exemplification of the similarity between a given time series and a later version of itself over consecutive periods [36]. The method of calculation is the same as that used in the computation of the correlation between two different time series, excluding using the same time series twice: one time in its original form and another in a later form or in more time intervals [37]. The equation for the autocorrelation function is:

$$\rho_k = \frac{\sum_{t=k+1}^T (r_t - \mu_r)(r_{t-k} - \mu_r)}{\sum_{t=1}^T (r_t - \mu_r)^2} \quad (8)$$

where  $\rho_k$  are the autocorrelation coefficients,  $r_t$  is a data set sorted by ascending date,  $r_{t-k}$  is the same data set shifted by  $k$  units, and  $\mu_r$  is the average of the original data set.

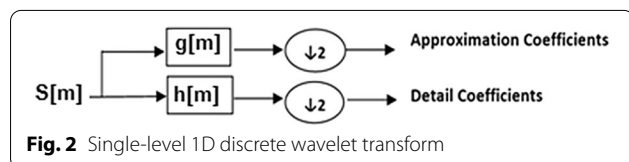


Fig. 2 Single-level 1D discrete wavelet transform

**Entropy**

Originally, Claude Shannon defined entropy as an aspect of his communication theory [38]. Shannon entropy provides vital information about repetitive sequences in whole chromosomes and is beneficial in finding evolutionary differences between organisms [39].

Shannon introduced the entropy,  $E$ , of a discrete random variable,  $Y$ , with possible values  $\{y_1, y_2, y_3, \dots, y_n\}$ , and probability mass function  $M(Y)$  as illustrated in [40, 41]:

$$E(Y) = - \sum_{i=1}^n M(y_i) \log_h M(y_i) \tag{9}$$

where  $h$  is the used logarithm base.

**Skewness and kurtosis**

In statistics, skewness is a measure of the asymmetry of the probability distribution of the variable around its mean. A symmetrical data set has a skewness of 0. It can be calculated as the averaged cubed deviation from the mean divided by the cubed standard deviation [42]. For defined data  $X_1, X_2, \dots, X_n$ , the equation for skewness, which represents the third moment, is as follows:

$$Skewness = \frac{\sum_{j=1}^n (X_j - \mu)^3 / n}{\sigma^3} \tag{10}$$

where  $\sigma$  is the standard deviation,  $\mu$  is the mean, and  $n$  is the data points' number.

It is used as a measure of the variable asymmetry and deviation from the normal distribution. It is called positively skewed distribution (right), where the most values are located on the left side of the mean, if the skewness value is greater than zero. It is called negatively skewed distribution (left), where the values are located on the right side of the mean, if the value is lower than zero. For the zero value (the mean value equals the median), the distribution is symmetrical about the mean value.

There is an incorrect concept that has appeared in different reports that kurtosis somehow measures the peakedness (flatness, pointiness, or modality) of a distribution, despite statisticians' efforts to set the record straight. In statistics, kurtosis is the measurement of the probability distribution tailedness of a variable [43]. The kurtosis value is related to the distribution tail-heaviness, not its peak. For defined data  $X_1, X_2, \dots, X_n$ , the equation for kurtosis, which represents the fourth moment, is as follows:

$$Kurtosis = \frac{\sum_{j=1}^n (X_j - \mu)^4 / n}{\sigma^4} \tag{11}$$

where  $\sigma$  is the standard deviation,  $\mu$  is the mean, and  $n$  is the number of data points.

The result is usually compared to the kurtosis of the normal distribution (Mesokurtic distribution), which equals three. A distribution is called a Leptokurtic distribution if the kurtosis value is more than three. In this case, it has more intensive tails than the Mesokurtic distribution. A distribution is known as a Platykurtic distribution if the kurtosis value is less than three. It has fewer tails than the normal distribution.

**Classifier**

In this research, two kinds of classifiers were used, and then their results were compared. They were K-nearest neighbors (KNN) and support vector machine (SVM).

1. K-nearest neighbors

The KNN algorithm is an unsupervised machine learning algorithm, and it is one of the most widely used classification methods. KNN is a case-based algorithm, so it does not require a learning step. It handles the training samples using a distance function and a separation function. It is based on the categories of the closest neighbors [7, 8]. When a new item is rated, it must be compared to others using a similarity scale, then KNNs are taken into regard, and the distance between the new item and the neighbor is used as the weight [44]. Various methods are used to calculate this distance. The most common technique is the Euclidean distance between the two vectors  $y_{ir}$  and  $y_{jr}$  which can be measured as stated in [45]:

$$d(y_i, y_j) = \sqrt{\sum_{r=1}^n (y_{ir} - y_{jr})^2} \tag{12}$$

The performance of the method depends on the K value selected and the distance cutoff used. The K value represents the number of neighbors chosen to specify the new element class.

2. Support vector machines

In learning systems, SVMs or networks are a supervised-learning method related to learning techniques that analyze data for detection and classification studies [6]. An SVM creates a hyperplane as a resolution surface to classify input data into a high-dimensional feature space. The hyperplane can differentiate between the different class patterns and increase the class margin. Patterns represent a set of points grouped to be separated by distinct lines for various categories. The points are assigned and classified according to which aspect of the

line they belong to [7, 46]. This process leads to a linear classification generated by the SVM, while the use of a kernel produces a nonlinear classification [15].

Each algorithm was used separately for classification and provided parameters for comparison. In this study, 35 normal colorectal genes and 35 cancerous genes were used as training data, and the testing data included 20 normal colorectal genes and 20 cancerous genes.

*MATLAB R2017b* was used to perform the analysis. The *fitcknn* function was used for creating a KNN with a default number of neighbors  $k=1$ , while the *fitcsvm* function was used for generating the SVM.

**Results**

Three important parameters were calculated for the performance evaluation of proposed method. They are Matthews correlation coefficient (MCC), F1 score, and ACC. They can be estimated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{13}$$

(ACC: 0 is the worst value; 100 is the best)

$$F1_{score} = \frac{2TP}{2TP + FP + FN} \times 100\% \tag{14}$$

(F1 Score: 0 is the worst value; 100 is the best)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

(MCC: - 1 is the worst value; + 1 is the best) where four confusion-matrix parts (FP, FN, TP, and TN) stand for false positive, false negative, true positive, and true negative values, respectively.

The MCC is more informative than the ACC or the F1 score in evaluating the performance of a binary classifier, because it takes into the balance rates of the FP, FN, TP, and TN [47]. For example, in a set of 200 elements, 180 are positive, and only 20 are negative. After applying the classifier, the following results are obtained:

$$FP = 20, TP = 180; FN = 0, TN = 0.$$

The previous inputs give F1 score = 94.73% and accuracy = 90%. Although these results look impressive, MCC would be indefinite, as the FN and TN would be zeroes, thus the denominator of Eq. 15 would be zero. The F1 score relies on which class is positive, which MCC does not [49].

The extracted features were used as an input to a KNN and an SVM network separately, and the results of each were compared. In this study, training data of 35 normal colorectal genes and 35 cancerous genes were used, and

**Table 1** Results of the two classifiers

	K-nearest network	SVM network
TP	19	18
FP	1	2
TN	20	20
FN	0	0

**Table 2** Comparison of calculated parameters for the two classifiers

	K-nearest network	SVM network
ACC %	97.5	95
F1%	97.44	94.74
MCC	0.9514	0.9045

testing data of 20 normal colorectal genes and 20 cancerous genes were used.

Table 1 shows the TP, FP, TN, and FN values obtained from the two classifiers.

From the previous values, Table 2 can be created using Eqs. 13–15.

**Discussions**

The KNN algorithm identified 19 cancer genes and 20 normal genes out of a total of 20 each (TP = 19, FP = 1, TN = 20, and FN = 0), while the SVM network recognized 18 cancer genes and 20 normal genes (TP = 18, FP = 2, TN = 20, and FN = 0) (Table 1).

The results of both methods were satisfactory. KNN gives 97.5% accuracy, 97.44% F1 score and 0.9512 MCC, while SVM network gives 95% accuracy, 94.74% F1 score and 0.9045 MCC (Table 2).

In comparison, achieving a higher ACC, higher F1 score, and higher MCC is evidence that the classification process is more successful, and the classifier is more effective. These results indicate that the classifier can recognize the required target with minimum errors. From the research results, the KNN classifier could achieve the research purpose of differentiating between normal and cancerous colorectal genes using GSP methods.

The results indicate the success of using GSP methods for cancer recognition and diagnosis. Table 3 provides a comparison of the results obtained in the current work to those of other studies according to the database used, method, classifier, and output.

From Table 3, the best accuracy obtained from the related studies was 96.7% [7], and this study reached 97.5% accuracy.

**Table 3** Comparison between different studies of classifications

Study	Data set	Method	Classifier	Output
Mendizabal-Ruiz et al. [5]	DNA sequence	Voss representation DFT PSD	K-means algorithm	Clustering
Ali et al. [7]	WBCD		MLP KNN CART NB SVM	ACC = 96.7% ACC = 96.27% ACC = 91% ACC = 93.62% ACC = 96.42%
Indu et al. [10]	DNA Microarray	Integrating correlation-based feature selection model	iBPSO	ACC about 92%
Fang et al. [15]	Ultrasound (US) images	Region of interest extraction based on SLIC	SLIC	ACC up to 92.05%
Zhou et al. [17]	Magnetic resonance imaging (MRI) images	Localizing the lesions at dynamic contrast-enhanced MRI data in a weakly supervised manner	CNN	ACC = 95%
Lakshmanaprabu et al. [20]	Computed tomography (CT) images	The deep features extracted from a CT lung images	ODNN	ACC = 94.56%
Presented Study	DNA sequences:	DWT Statistical features	KNN SVM	ACC = 97.5% ACC = 95%

## Conclusions

Many researchers worldwide have studied cancer, hoping to detect this disease at an early stage so that they could reduce its risk, which often leads to death. The basic concept of the presented study is that cancer is considered to be a genetic disease. The EIIP method was used to convert the DNA sequences from strings into number values so that GSP could be applied in the feature extraction step, and suitable classifiers were selected. Single-level DWT was applied using Haar wavelets. Then, the statistical features mean, variance, standard deviation, autocorrelation, entropy, skewness, and kurtosis were obtained from the wavelet domain. Finally, the resulting values were input into KNN and SVM networks. The KNN results were the best, with low error for the classification system, although the results of the SVM were acceptable. An automated system was therefore generated for the classification and detection of colorectal cancer with good results, avoiding the disadvantages of traditional methods. These traditional detection methods include collecting blood, urine, or stool sample from the patient and testing it in the laboratory. That takes a long time, requires experienced examiners, and the probability of error is relatively high. In future work, other GSP features can be used, and different classifiers can be chosen to improve the results.

## Abbreviations

ACC: Accuracy; CART: Classification and regression tree; CT: Computed tomography; DFT: Discrete Fourier transform; DNA: Deoxyribose nucleic acid; DWT:

Discrete wavelet transform; EIIP: Electron-ion interaction pseudopotential; FN: False negative; FP: False positive; GSP: Genomic signal processing; iBPSO: Improved binary particle swarm optimization; KNN: K-nearest neighbor; MCC: Matthews correlation coefficient; MLP: Multilayer perceptron; MRI: Magnetic resonance imaging; NB: Nave Bayes; ODNN: Optimal deep neural network; PSD: Power spectral density; SLIC: Simple linear iterative clustering; SVM: Support vector machine; TN: True negative; TP: True positive; US: Ultrasound.

## Acknowledgements

Not applicable.

## Authors' contributions

SMN, MSM, MAD, and AYS contributed to conception and supervision of study. SMN and MSM contributed to research techniques. SMN and MSM contributed to analysis and interpretation of the data. SMN contributed to writing of the paper. MAD and AYS contributed to critical review. SMN, MSM, MAD, and AYS contributed to clinical assessment. All authors read and approved the final manuscript.

## Funding

Not applicable.

## Availability of data and materials

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Biomedical Engineering Department, Faculty of Engineering, Helwan University, Cairo, Egypt. <sup>2</sup>Biomedical Engineering Department, Faculty of Engineering, Misr University for Science and Technology (MUST University), Cairo, Egypt. <sup>3</sup>Department of Engineering Mathematics and Physics, Faculty of Engineering El-Matariya, Helwan University, Cairo, Egypt.

Received: 19 March 2021 Accepted: 30 July 2021

Published online: 12 October 2021

**References**

1. Thanikachalam K, Khan G (2019) Colorectal cancer and nutrition. *Nutrients* 11(1):164. <https://doi.org/10.3390/nu11010164>
2. Vuik F, Nieuwenburg S, Bardou M et al (2019) Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. *Gut* 68:1820–1826
3. Mármol I, Sánchez-de-Diego C, Pradilla DA, Cerrada E, Rodriguez MJ (2017) Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int J Mol Sci* 18(1):197. <https://doi.org/10.3390/ijms18010197>
4. Kuipers EJ, Grady WM, Lieberman D et al (2015) Colorectal cancer. *Nature reviews. Disease Primers* 1:15065. <https://doi.org/10.1038/nrdp.2015.65>
5. Mendizabal-Ruiz et al (2018) Genomic signal processing for DNA sequence clustering. *PeerJ* 6:e4264. <https://doi.org/10.7717/peerj.4264>
6. David A Omondigbe et al (2019) Machine learning classification techniques for breast cancer diagnosis. 2019. *IOP Conference Series: Materials Science and Engineering* 495:012033
7. Ali Al BA (2019) Comparative analysis of nonlinear machine learning algorithms for breast cancer detection. *Int J Mach Learn Comput* 9(3)
8. Fogliatto FS, Anzanello MJ, Soares F, Brust-Renck PG (2019) Decision support for breast cancer detection: classification improvement through feature selection. *Cancer Control* 26(1):1073274819876598
9. Alickovic E, Subasi A (2020) Normalized Neural Networks for Breast Cancer Classification. In: Badnjevic A, Škrbić R, Gurbeta Pokvić L (eds) *CMBEBIH 2019. CMBEBIH 2019. IFMBE proceedings*, vol 73. Springer, Cham
10. Indu J, Vinod KJ, Renu J (2018) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Appl Soft Comput* 62:203–215
11. Serhat K, Kemal A, Mete Celik (2020) Diagnosis and classification of cancer using hybrid model based on relief and convolutional neural network. *Medical Hypotheses*. 137:10957
12. Shen T, Nagai Y, Udayakumar M, Narasimhan K, Shriram RK, Arvind MN, Elamaran V (2019) Automated genomic signal processing for diseased gene identification. *J Med Imaging Health Inform* 9(6):1254–1261
13. Naeem SM, Mabrouk MS, Eldosoky MA (2017) Detecting genetic variants of breast cancer using different power spectrum methods. In: 2017 13th international computer engineering conference (ICENCO), Cairo, pp 147–153
14. Sayed AY, Naeem SM, Mabrouk MS, Eldosoky MA (2020) New method for cancer classification using moment invariants and artificial neural network. In: 2020 9th international conference on mathematics and information sciences (ICMIS), 6–8 Feb 2020, Aswan, Egypt
15. Fang Z, Zhang W, Ma H (2020). Breast Cancer Classification with Ultrasound Images based on SLIC. *Proceedings of 9th international conference frontier computing (FC)*, pp 235–248
16. Coudray N, Moreira AL, Sakellaropoulos T, Fenyo D, Razavian N, Tsigiros A (2018) Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *BioRxiv*, pp. 197574. <https://doi.org/10.1101/197574>
17. Zhou J, Luo LY, Dou Q et al (2019) Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. *J Magn Reson Imaging* 50(4):1144–1151. <https://doi.org/10.1002/jmri.26721>
18. Alom MZ, Yakopcic C, Nasrin MS, Taha TM, Asari VK (2019) Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *J Digit Imaging* 32(4):605–617
19. Mesut T, Burhan E, Zafer C (2020) Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with uotencoders. *Medical Hypotheses*. February. Volume 135:109503
20. Lakshmanaprabu SK, Mohanty SN, Shankar K, Arunkumar N, Ramirez G (2019) Optimal deep learning model for classification of lung cancer on CT images. *Futur Gener Comput Syst* 92:374–382
21. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44(D1):D7–D19. <https://doi.org/10.1093/nar/gkv1290>
22. Trad CH, Fang Q, Cosic I (2003) Protein sequence comparison based on the wavelet transform approach. *Protein Eng* 15(3):193–203
23. Ghosh A, Barman S (2013) Prediction of prostate cancer cells based on principal component analysis technique. *Procedia Technology-Int Conference Computational Intelligence: Modeling Techniques and Applications (CIMTA)*, pp 37–44
24. Wassfy HM, Abd Elnaby MM, Salem ML, Mabrouk MS, Zidan AA (2016) Eukaryotic gene prediction using advanced DNA numerical representation schemes. In: *Proceedings of fifth international conference advances in applied science and environmental engineering (ASEE)*, Kuala Lumpur, Malaysia
25. Nair SA, Sreenadhan SP (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformatics* 1(6):197–202
26. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11(7):674–693
27. Prakash SN, Khan AM (2020) MRI image compression using multiple wavelets at different levels of discrete wavelets transform. *J Phys Conf Ser* 1427:012002
28. Haar A (1910) Zur Theorie der orthogonalen Funktionensysteme. *Math Ann* 69(3):331–371
29. Zhang D (2019) Wavelet transform. In: *Fundamentals of image data mining. Texts in computer science*. Springer, Cham
30. Ghorpade A, Katkar P, Transform I (2014) Image compression using Haar transform and modified fast Haar wavelet transform. *Int J Sci Technol Res* 3:3–6
31. Chun-Lin (2010). Tutorial of the Wavelet Transform. Taipei, Taiwan
32. Mean, Median and Mode, <http://www.mathcentre.ac.uk>, math center. Accessed January 02, 2021
33. Nicholas N, Watier CL, Sylvain C (2011) What does the mean mean? *J Stat Educ* 19(2)
34. Keijo R (2011). *Statistics 1*. (Translation by Jukka-Pekka Humaloja and Robert Piché)
35. Thomas BF, Stanley RJ, Carter HR (1984) *Advanced econometric methods*. Springer, New York, pp 205–236
36. Autocorrelation (2006). *Encyclopedia of Measurement and Statistics*. SAGE Publications. 30 Aug. 2009. [http://www.sage-e-reference.com/statistics/Article\\_n37.html](http://www.sage-e-reference.com/statistics/Article_n37.html)
37. Shannon, Claude EA (1948) Mathematical theory of communication. *Bell Syst Tech J* 27(3):379–423
38. Thanos D, Li W, Provata A (2018) Entropic fluctuations in DNA sequences. *Physica A* 493:444–457. <https://doi.org/10.1016/j.physa.2017.11.119>
39. Tenreiro MJ (2012) Shannon entropy analysis of the genome code. *Math Prob Eng* 1–2. <https://doi.org/10.1155/2012/132625>
40. Das J, Barman S (2017) DSP based entropy estimation for identification and classification of homo sapiens cancer genes. *Microsyst Technol* 23(9):4145–4154
41. Chattopadhyaya A, Chattopadhyay S, Bera JN, Sengupta S (2016). Wavelet decomposition based skewness and kurtosis analysis for assessment of stator current harmonics in a PWM-fed induction motor drive during single phasing condition. *AMSE J Ser Adv B* 59(1):1–14
42. Westfall PH (2014) Kurtosis as peakedness. 1905–2014. *R.I.P. Am Stat* 68(3):191–195. <https://doi.org/10.1080/00031305.2014.917055>
43. Hadi AH, Ahmed KA, Sara AW (2018) Frequency hopping spread spectrum recognition based on discrete Fourier transform and skewness and kurtosis. *Int J Appl Eng Res* 13(9) 7081–7085
44. Negnevitsky M (2005) *Artificial intelligence: a guide to intelligent systems*. Pearson ch. 6, pp 175–179
45. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2001) Support vector clustering. *J Mach Learn Res* 2:125–137
46. Ten CD (2017) Quick tips for machine learning in computational biology. *BioData Min* 10(1):1–5

47. Chicco D, Jurman G (2020) The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom* 21(1):6

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---