*Type of the Paper (Article)*

# Breast and Colon Cancer Classification from Gene Expression Profiles using Data Mining Techniques

**Mohammed Loey[1], Mohammed Wajeeh Jasim[2], Hazem M. EL-Bakry[3], Mohamed Hamed N. Taha[4], Nour Eldeen M. Khalifa[5]**

[1,2]Department of Computer Science, Faculty of Computers Artificial Intelligence, Benha University, Egypt.
[3]Department of Information Systems, Faculty of Computer & Information Sciences, Mansoura University, Egypt.
[4,5]Department of Information Technology, Faculty of Computers & Artificial Intelligence, Cairo University, Egypt.

**Abstract:** Early detection of cancer increases the probability of recovery. This paper presents an intelligent decision support system (IDSS) for the early diagnosis of cancer based on gene expression profiles collected using DNA microarrays. Such datasets pose a challenge because of the small number of samples (no more than a few hundred) relative to the large number of genes (on the order of thousands). Therefore, a method of reducing the number of features (genes) that are not relevant to the disease of interest is necessary to avoid overfitting. The proposed methodology uses the information gain (IG) to select the most important features from the input patterns. Then, the selected features (genes) are reduced by applying the grey wolf optimization (GWO) algorithm. Finally, the methodology employs a support vector machine (SVM) classifier for cancer type classification. The proposed methodology was applied to two datasets (Breast and Colon) and was evaluated based on its classification accuracy, which is the most important performance measure in disease diagnosis. The experimental results indicate that the proposed methodology is able to enhance the stability of the classification accuracy as well as the feature selection.

**Keywords:** machine learning, cancer diagnosis, grey wolf optimization algorithm, support vector machine, information gain, feature selection.

## 1. Introduction

Cancer is a common disease caused by certain abnormal changes to genes that are responsible for cell division and growth. These recognizable changes include mutations of the DNA that make up genes. Generally, cancer cells exhibit significantly more genetic changes than normal cells, although cancerous tumours show different specific combinations of genetic alterations in different people. However, a few of these recognizable changes may be the result of the cancer rather than its cause. As cancer grows, additional changes will occur [1]. Therefore, the early detection of cancer can improve the treatment possibilities and increase the survival rate of patients. Thus, developing appropriate methodologies that can effectively distinguish among tumour subtypes is vital. Early diagnosis of cancer is essential for sufficient and effective treatment because every cancer type requires specific treatment.

According to the World Cancer Organization, approximately 4,610 cases of central nervous system (CNS) tumours and various brain tumours were expected to be diagnosed in 2018 in children under the age of 20 in the United States. After leukaemia, brain cancer and other tumours of the CNS are the second most common type of cancer among children; the rate of such tumours has never

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



reached more than 26% among children under the age of one year [2, 3]. In 2019, 1,762,450 new cancer cases of brain and other nervous system tumours were reported in the United States, and the number of associated deaths was estimated to be 606,880. Thus, it is important to develop a methodology of detecting cancer in the early stages before the tumour worsens, thereby reducing the risk of death [4].

The conventional methods of diagnosing most existing diseases depend on human experience to recognize cases that correspond to confirmed data patterns. However, this age-old diagnosis methodology is subject to human error and imprecise diagnosis and is both time consuming and labour intensive, thus causing undue stress throughout the whole process. As an alternative, computer-aided diagnosis (CAD) systems based on machine learning have been continually improving and are employed to support specialists in the determination of diagnosis decisions [5, 6].

Most current CAD systems for medical diagnosis depend on diverse information, such as medical laboratory tests (e.g., blood tests and magnetic resonance imaging (MRI)), medical indicators (finger tremors and lung signs or symptoms), and various types of digital images (such as X-rays and ultrasound images). However, physical medical examinations pose a risk of transmission of infection through tools and other channels, such as scratching of the skin while taking a blood sample [7,8,9]. X-rays are harmful because of the exposure of body cells to radiation. The quality of ultrasound data depends on the accuracy and integrity of the image, which are affected by various factors, such as the presence of air between the surface of the skin and the tool and image blur. A system that depends on gene expression data collected using DNA microarrays can effectively solve these problems. Such a method can be used to diagnose cancer in the early stages, unlike other methods that use different kinds of image processing techniques. The challenges that arise in microarray classification are mainly centred on dimensionality and classification accuracy [8, 9].

Methodologies that depend on gene expression profiles have been able to detect cancer since their inception. In previous work, exhaustive efforts have been made to achieve the best results. Researchers have achieved excellent results in the classification of cancer based on gene expression profiles using various gene selection approaches and classifiers [10].

There is more than one approach to gene selection, including filter, wrapper, embedded and hybrid approaches, and every approach has its advantages and disadvantages. For example, the advantages of the filter approach are that it is very fast and computationally simple, whereas its main disadvantage is that each feature is measured separately, and thus, it does not consider the dependencies among features. The wrapper approach has the advantage of enabling an exhaustive search to generate optimal solutions, whereas its disadvantage is that it has a higher risk of overfitting than filter techniques do. The embedded approach has the same benefits as the wrapper approach while achieving better computational complexity; however, it is still prone to overfitting. Hybrid approaches can combine the advantages of various other approaches, but the time complexity may increase [11, 12].

This paper addresses the problem of medical diagnosis and presents an intelligent decision support system (IDSS) for cancer diagnosis based on gene expression profiles from DNA microarray datasets. DNA microarray technology has been efficiently applied to analyse gene expression in many experimental studies. Usually, the number of features (M) in a microarray dataset is very large (usually in the thousands), while the number of samples (N) is small (not exceeding hundreds) [13]. This paper proposes an IDSS for CNS cancer classification based on gene expression profiles. The

proposed system combines the information gain (IG), the grey wolf optimization (GWO) algorithm and the support vector machine (SVM) algorithm: the IG is used for selecting important genes (features) from the input matrix, GWO is used for feature reduction, and an SVM classifier is used for cancer diagnosis.

The remainder of this paper is organized as follows. Section 2 reviews some important previous works. Section 3 describes the proposed methodology; this section includes an overview of the IG filter approach for feature selection, the GWO algorithm for feature reduction and the SVM algorithm for classification. Section 4 describes the datasets used in this research. Section 5 reports and analyses the results, and Section 6 presents the conclusions and possibilities for future work.

**2. Related Works**

In all the previous studies listed below, gene expression profiles were used for the classification of cancer based on various methodologies. These methodologies were applied to datasets with a small number of samples, a large number of features, and the additional characteristics listed in table 1 below.

Table 1. Characteristics of the datasets used in previous studies

| Dataset | Used in Ref | No. of samples | No. of features | Class 1 | Class 2 |
|---|---|---|---|---|---|
| Leukaemia72 | [11], [35], [33], [16], [10], [13] | 72 | 7129 | ALL (47) | AML (25) |
| Prostate | [11], [35], [33], [10], [14] | 136 | 12600 | Normal (59) | Tumour (77) |
| Lung Cancer-Ontario | [11] | 39 | 2880 | Non-relapse (15) | Relapse (24) |
| Lung Cancer-Michigar | [11], [35] | 96 | 7129 | Non-neoplastic (10) | Primary lung (86) |
| DLBCL Harvard | [11], [35] | 77 | 7129 | DLBCL (58) | FL (19) |
| Central Nervous System | [11], [35], [33], [16] | 60 | 7129 | Class 0 (39) | Class 1 (21) |
| Colon | [11], [35], [33], [16], [10], [14] | 62 | 2000 | Positive (22) | Negative (40) |
| Leukaemia | [14] | 72 | 3571 | ALL (47) | AML (25) |
| Prostate Outcome | [35] | 136 | 12600 | Normal (59) | Tumour (77) |
| Breast | [12] | 97 | 24,481 | Relapse (46) | Non-relapse (51) |
| Ovarian | [33], [35] | 253 | 15154 | Normal (91) | Cancer (162) |
| GCM | [33] | 190 | 16063 | - Multi-class- | |
| DLBCL Outcome | [35] | 58 | 7129 | Cured (32) | Fatal (26) |
| Leukaemia38 | [16] | 38 | 5000 | ALL (27) | AML (11) |
| Lymphoma | [33], [16], [10] | 96 | 4026 | - Multi-class- | |

Salem, Hanaa, et al [11] reported research on human cancer classification using gene expression profiles. The feature selection methodology used in this study exploited the IG for gene selection from the input microarray data. The methodology also exploited a genetic algorithm (GA) to reduce

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



the number of features selected based on the IG. The final task of cancer classification (or diagnosis) was accomplished by means of genetic programming (GP). The framework was verified by considering seven cancer gene expression datasets (Lung Cancer-Ontario, Leukaemia72, DLBCL Harvard, Prostate, Lung Cancer-Michigan, Colon, and Central Nervous System). The authors achieved classification accuracies of 85.48% (Colon), 86.67% (Central Nervous System), 97.06% (Leukaemia72), 74.4% (Lung Cancer-Ontario), 100% (Lung Cancer-Michigan), 94.8% (DLBCL Harvard) and 100% (Prostate).

As a hybrid gene selection technique, J. Bennet, C., et al [13] proposed an ensemble feature selection technique that is a mixture of the support vector machine-recursive feature elimination (SVM-RFE) approach and the Based Bayes error Filter (BBF) for attribute selection. These researchers employed SVM-RFE to sort the attributes and the BBF to remove redundant sorted attributes. The SVM algorithm was then used for classification. The best classification accuracy on the Leukaemia72 dataset reached 97.2%.

The authors of [35] presented an analysis of the behaviour of a GA with k-nearest-neighbours (KNN) and SVM classifiers on ten datasets. Using the GA, they reduced the number of features selected by three filters. In the final stage, the KNN and SVM algorithms were used for classification. The authors used 5-fold cross-validation, and on most datasets, the classification accuracy achieved with the SVM classifier was the same as that achieved with the KNN classifier; the results differed only for the Leukaemia72 dataset (Lung Cancer-Michigan: 100%, Ovarian: 100%, Central Nervous System: 81.25%, DLBCL Harvard: 100%, DLBCL Outcome: 77.27%, Prostate Outcome: 85.71%, Leukaemia72: 100% using SVM and 95.45% using KNN, Colon: 95%, Lung Harvard2: 100%, and Prostate: 92%).

In [33], an ensemble of five filters (IG, correlation-based feature selection (CFS), consistency-based, interaction, and ReliefF) and three classifiers (naïve Bayes, C4.5, and IB1) was proposed. The researchers used a simple voting scheme for classification. They applied their methodology to 10 microarray datasets with ten-fold cross-validation, and the best classification accuracies they obtained were 100% (Lung), 89.05% (Colon), 100% (Ovarian), 70% (Central Nervous System), 71.89% (Breast), 98.75% (Leukaemia72), 90.6% (Prostate), 68.42% (GCM), and 95.67% (Lymphoma).

In [16], the researcher applied a GA for gene selection in combination with four classifiers for cancer classification using a gene expression dataset. The classifiers used were naïve Bayes, SVM, oneR, and decision tree classifiers. The author analysed the results obtained by applying the methodology to six datasets, namely, Lymphoma, Lung, CNS, Colon, Leukaemia38, and Leukaemia72, on which the best classification accuracies were 97%, 99.4%, 82.3%, 88.8%, 100%, and 98.6%, respectively.

Salem, Hanaa, et al [12] presented research on the early classification of breast cancer based on gene expression profiles. Their system first extracts important genes from the input microarray data using the IG methodology and then exploits a GA to reduce the features selected in this way. The best results in this study were achieved with an IG threshold value of 0.7 for the breast cancer dataset; with this threshold, the features were initially reduced from 24,481 attributes to 45 attributes by the IG methodology and were further reduced to only 22 attributes by applying a GA with a population size of one hundred and twenty rounds of evaluation. The classification accuracy reached 100%.

Bouazza, Sara Haddou, et al [14] presented research on cancer classification using SVM and KNN classifiers. In this research, the effects achieved on three gene expression profile datasets (Prostate,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



Colon, and Leukaemia) were studied using multiple techniques for attribute selection (such as Fisher, ReliefF, SNR, and T-Statistics) with both KNN and SVM classifiers. The best results were obtained by combining the SNR attribute selection technique with the SVM classifier. The best classification accuracies achieved in this study with the SNR feature selector and the KNN classifier were 95% for the Colon and Prostate datasets and 100% for the Leukaemia dataset.

## 3. The Proposed Methodology

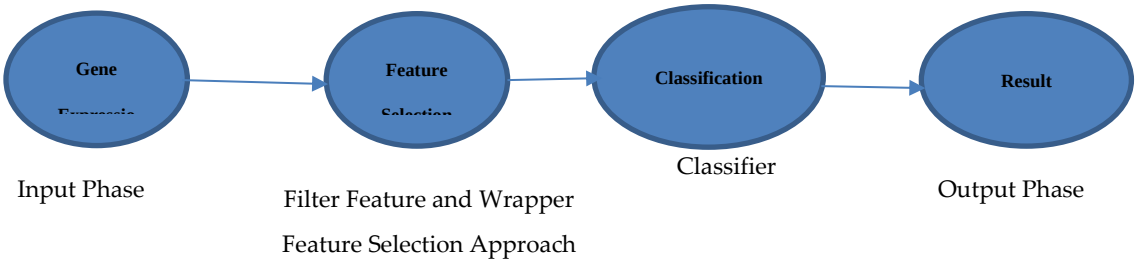


Figure 1. IG flowchart

The proposed methodology consists of three main stages. Once the data are input into the system, the IG filter first selects the most important features [15]. Then, the GWO algorithm reduces the number of selected features. The last stage of the system is to apply the SVM classifier to obtain specific cancer classification results. An overview of the methodology is shown in figure 1.

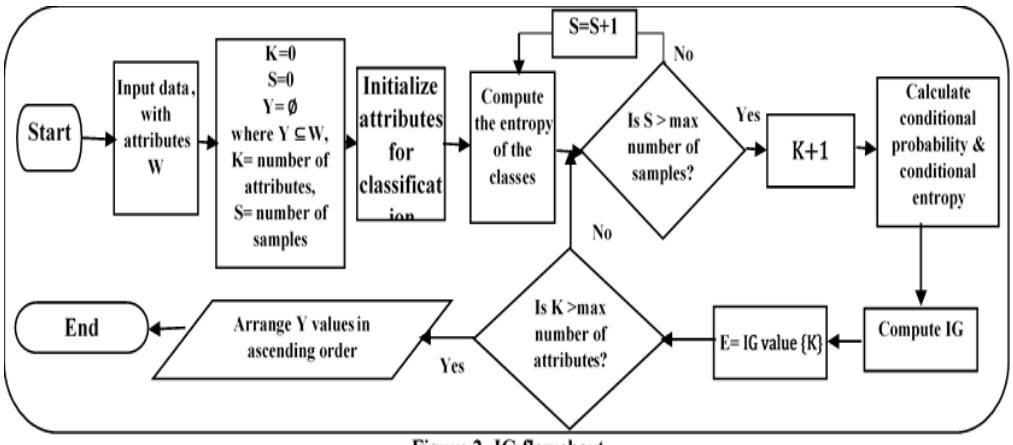


Figure 2. IG flowchart

### 3.1. Entropy and information gain (IG) for gene selection

Entropy is the basic concept used in information theory to compute the homogeneity of features; for example, when samples are fully homogeneous, they have an entropy equal to zero, whereas equally divided samples have an entropy value of one [17]. For a dataset with a high feature dimensionality and a small sample size, classification of the data is very difficult. Among the thousands of gene expression attributes that are usually investigated, only very few are relevant to a particular disease. Therefore, only the relevant features should be retained [16]. Proper investigation of the gene profiles will be helpful for selecting the genes that are most important for the classification process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



$E(Z) = - D^{+}\log_2(D^{+}) - D^{-}\log_2(D^{-})$ for a sample with negative and positive attributes.

The formula for entropy is as follows [18]:

$$\text{Entropy}\ (Z) = \sum_{i=1}^{V} -(D_k\ log_2 D_k) \qquad (1)$$

where the $D_i$ are the a priori probabilities of categorical variable $Z$ and $k$ is an index indicating a particular category in the classification system.

Consider the special case of two classification problems (where V is the number of classes).

Let j be a gene that may have n possible values ($j_1$, $j_2$, …, $j_n$). The entropy will be as follows:

$$\text{Entropy}\ (k/j) = -\sum_{j=1}^{n} p(j) \sum_{k=1}^{V} p\left(\frac{k}{j}\right)\ log_2\left(p\left(\frac{k}{j}\right)\right) \qquad (2)$$

where $p(\frac{k}{j})$ is the conditional probability of variable $K$ when attribute $J$ is constant, calculated over all attributes and classes. The calculation of IG mainly depends on the entropy [19]. Based on the distribution of the attributes in the dataset, the entropy is calculated for all attributes in the dataset. The data are then separated into groups of attributes. The entropy is calculated for each group separately, and the entropy values for all groups are combined to obtain the total entropy. The entropy based on individual groups of data is then subtracted from the entropy based on the entire data distribution [20]:

$$IG(J) = Entropy(S) - Entropy\ \left(\frac{k}{j}\right) \qquad (3)$$

When gene J and category K are not related, $IG(J) = Entropy(S) - Entropy\ \left(\frac{k}{j}\right) = zero$, whereas if they are related, then $Entropy(S) > Entropy\ \left(\frac{k}{j}\right)$, leading to IG(J) >0. There is a direct relationship between a larger difference between $J$ and $K$ and a stronger correlation between $J$ and $K$. A feature with a larger IG value is more important for classification. Therefore, genes with greater IG values are first chosen from among the original high-dimensional genes to be used as the basis for further gene selection [21].

The IG flowchart shown in figure 2 describes the steps of the IG algorithm. The input data set has a set of attributes W, and the required output is the selected subset Y of the original attributes W. First, the attributes to be considered for classification are initialized. Second, the entropy of all samples is computed for each class using equation (1). Then, the conditional probability for each value of a single attribute is calculated and is used to calculate the conditional entropy for every attribute via equation (2). The IG is computed using equation (3) for all attributes. The resulting IG values are arranged in ascending order, and all values that are above a certain threshold value are selected.

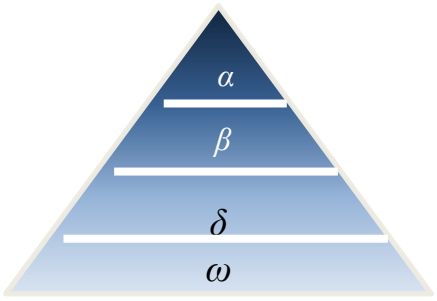


**Figure 3.** Hierarchy of grey wolves (dominance decreases from the top down) [22]

### 3.2. Grey wolf optimization (GWO) for feature reduction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



Wolves of the grey wolf (Canis lupus) species prefer to live in packs. On average, there are 5-12 members in one group. These animals live in groups governed by laws that maintain their hierarchical order, as shown in figure 3 [22]. At the top of the hierarchy is the leader, called the alpha, who may be male or female. The alpha is responsible for most of the pack's decisions, such as the places they hunt, the times at which they wake and sleep, and so on. All wolves in the pack follow the alpha [23]. The betas, who also may be male or female, constitute the second level in the hierarchy of grey wolves. They assist the alpha wolf in decision-making and coordinating other activities in the pack. A beta wolf is the most likely candidate to inherit the alpha's position when the alpha dies or becomes too old to lead. The third level in the hierarchy of grey wolves consists of wolves called subordinates (or deltas). Deltas follow the alpha and beta wolves, and the others follow the deltas. The deltas of the pack are divided into several categories, each of which is responsible for particular tasks.

The scout category is responsible for monitoring for threats to the pack [24]. The sentinel category is responsible for providing safety and protection for the pack. Elders are experienced wolves who are nominated to be future alphas or betas. Hunters assist the betas and alpha in hunting prey and providing food for the pack. Caretakers perform the caring tasks for wounded, ill, and weak wolves in the pack. All other wolves are omegas, who lie at the base of the hierarchy; they are the scapegoats. The omega wolves are subordinate to all other wolves in the hierarchy. They are the last to be allowed to eat. However, this does not mean that they are insignificant in the pack; without the omegas, the pack might collapse due to in-fighting. Furthermore, all wolves vent their violent tendencies by means of the omegas. This helps to maintain the hierarchical structure of the pack. The omegas also sometimes act as babysitters. Thus, all pack members participate in the leadership hierarchy. In the GWO algorithm, three primary hunting steps are performed for the purpose of optimization: seeking, encircling and attacking prey [25].

### 3.2.1. The mathematical model of GWO

In the mathematical formulation of GWO, the alpha ($\alpha$) represents the fittest solution, and the next best solutions are the beta and delta ($\beta$) and ($\delta$) solutions. Other solutions are regarded as omega ($\omega$) solutions. In the GWO algorithm, the leadership consists of alpha ($\alpha$), beta ($\beta$) and delta ($\delta$) wolves. The remaining omega ($\omega$) wolves are followers [30]. A mathematical representation of encircling behaviour is given by the following equations [27]:

$$\vec{E} = \left|\vec{F} \cdot \overrightarrow{Y_s}(i) - \vec{Y}(i)\right| \qquad (4)$$

$$\vec{Y}(i+1) = \overrightarrow{Y_s}(i) - \vec{B}.\vec{E} \qquad (5)$$

Here, $i$ represents the current iteration, $\vec{B}$ and $\vec{F}$ are coefficient vectors, $\overrightarrow{Y_s}$ is the prey's position vector and $\vec{Y}$ represents the grey wolf's position vector. The $\vec{B}$ and $\vec{F}$ vectors are calculated as follows:

$$\vec{B} = 2\vec{b}.\overrightarrow{m_1} - \vec{b} \qquad (6)$$

$$\vec{F} = 2.\overrightarrow{m_2} \qquad (7)$$

where the magnitude of $\vec{b}$ decreases linearly from 2 to 0 over multiple iterations and $\overrightarrow{m_1}$ and $\overrightarrow{m_2}$ are random vectors between [0,1].

The update parameter b controls the trade-off between exploitation and exploration. The parameter b is updated linearly from 2 to 0 as follows:

$$b = 2 - i\frac{2}{MxItr} \qquad (8)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



where MxIter is the overall number of iterations and $i$ is the number of the current iteration.

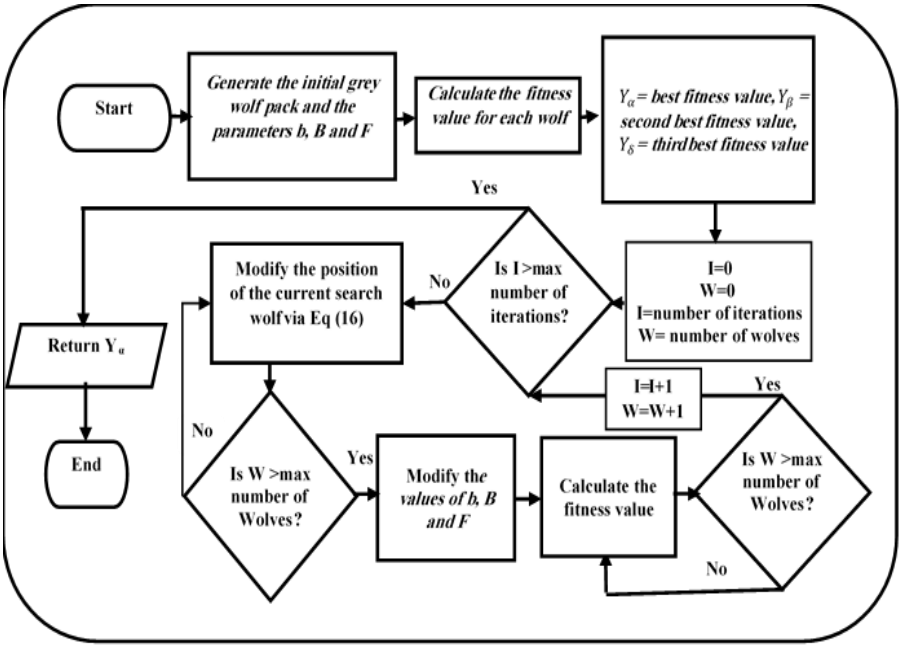


**Figure 4. GWO flowchart**

The grey wolves wish to identify the prey's location and encircle it. To this end, the alpha guides the pack in the hunt. However, the grey wolves have no idea of the area to be searched or the location of the prey. To represent this idea mathematically, we assume that the alpha represents the best solution available. The beta and delta wolves assist in inferring the location of the prey. For this purpose, we need to identify the three best values and update their positions to approach as close as possible to the optimal solution. The position update is performed in accordance with the following equations, and a flowchart of the GWO algorithm is shown in figure 4 [28].

$$\overrightarrow{E_\alpha} = \left|\overrightarrow{F_1} \cdot \overrightarrow{Y_\alpha} - \vec{Y}\right| \quad (9)$$

$$\overrightarrow{E_\beta} = \left|\overrightarrow{F_2} \cdot \overrightarrow{Y_\beta} - \vec{Y}\right| \quad (10)$$

$$\overrightarrow{E_\delta} = \left|\overrightarrow{F_3} \cdot \overrightarrow{Y_\delta} - \vec{Y}\right| \quad (11)$$

$$\overrightarrow{Y_1} = \overrightarrow{Y_\alpha} - \overrightarrow{B_1} \cdot (\overrightarrow{E_\alpha}) \quad (12)$$

$$\overrightarrow{Y_2} = \overrightarrow{Y_\beta} - \overrightarrow{B_2} \cdot (\overrightarrow{E_\beta}) \quad (13)$$

$$\overrightarrow{Y_3} = \overrightarrow{Y_\delta} - \overrightarrow{B_3} \cdot (\overrightarrow{E_\delta}) \quad (14)$$

$$\vec{Y}(i+1) = \frac{\overrightarrow{Y_1} + \overrightarrow{Y_2} + \overrightarrow{Y_3}}{3} \quad (15)$$

However, in the proposed system, the last equation (15) is modified as shown in equation (16) below:

$$\vec{Y}(i+1) = \mathrm{Max}\left(\overrightarrow{Y_1} + \overrightarrow{Y_2} + \overrightarrow{Y_3}\right) \quad (16)$$

### 3.3. Support vector machines (SVMs) for classification

The SVM technique is one of the most popular techniques in machine learning. It depends on points of similarity, similar to KNN. However, it does not require the calculation of the distances between a new unseen point and all other data points at hand; rather, only the vectors that will influence the decision-making process are considered. The SVM approach is based on the idea of maximizing the margins between different classes. The greater the certainty of a classifier is, the larger are the margins it provides [29]. SVM classification is based on two key ideas:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



- the notion of maximum margins and the concept of the kernel function.
- The area between a sample boundary and the nearest sample is called a margin. The support vectors represent these samples. In an SVM, the largest value is chosen to represent the margin.

For data with more than one dimension, an SVM classifier converts the data representation domain into a multi-dimensional domain and defines a hyperplane separating the data. The error (Err) and accuracy (Acc) of classification are used to evaluate the performance of an SVM classifier [14]:

$$Acc = (100 * (TruePo + TrueNe)) / (TrueNe + TruePo + FalseNe + FalsePo) \quad (17)$$

$$Err = (100 * (FalsePo + FalseNe)) / (TrueNe + TruePo + FalseNe + FalsePo) \quad (18)$$

where TruePo denotes the number of true positives, FalsePo denotes the number of false positives, TrueNe denotes the number of true negatives, and FalseNe denotes the number of false negatives.

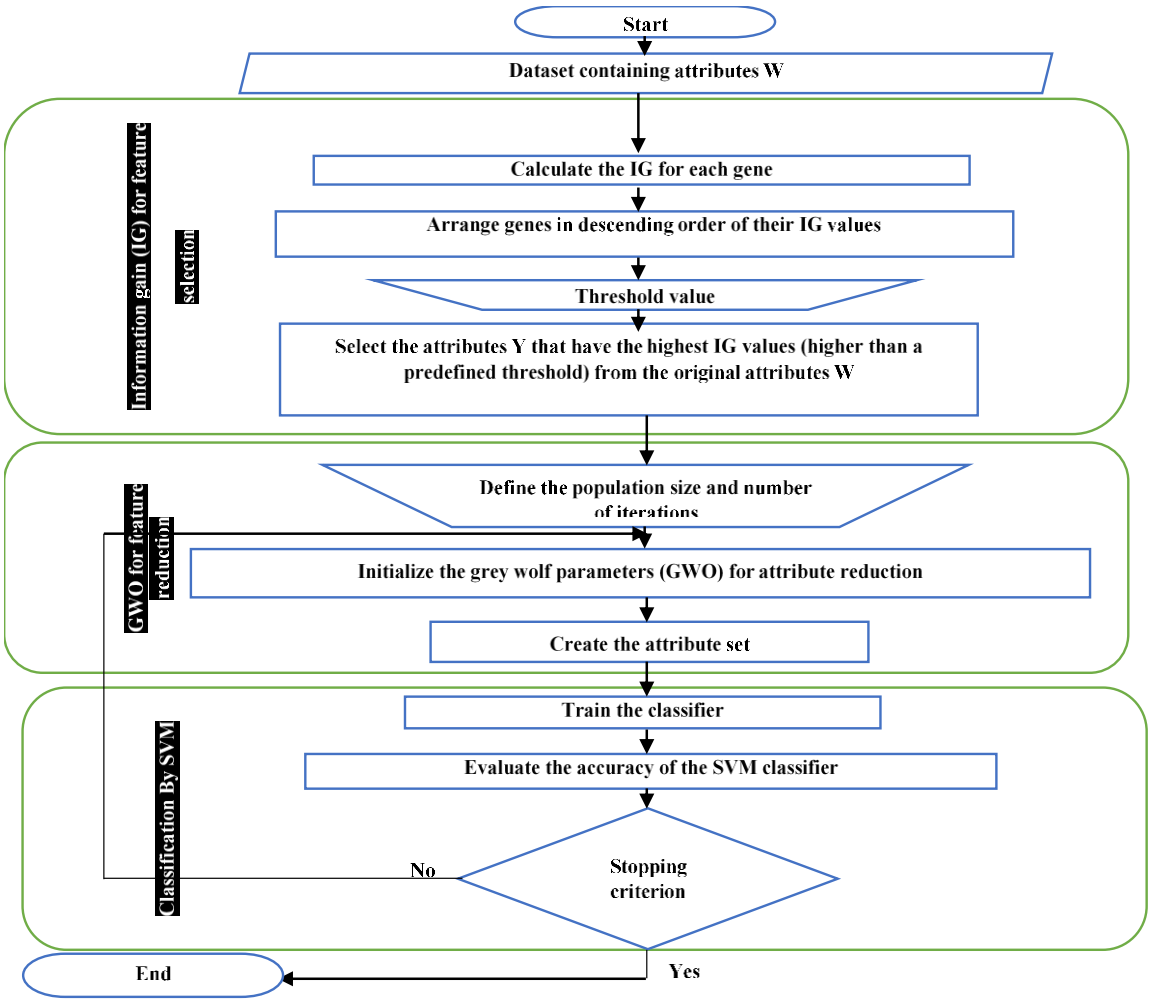


**Figure 5. Proposed system flowchart**

### 3.4. Proposed system workflow

Our system is based on an IDSS. The system works as shown in figure 5. First, the initial dataset, containing a set of attributes W, is entered into the system. i) Calculate the IG value for each gene and then arrange the genes in descending order of their IG values. ii) Select the attributes Y that have the highest IG values (higher than a predefined threshold) from among the attributes W. iii) Initialize the grey wolf parameters (GWO), such as the population size, $Y_i$, b, B, and E, and create the

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



attribute set. iv) Depending on the resulting wolves (selected feature subset), train an SVM classifier and evaluate its accuracy. v) Calculate the fitness value of each search wolf ($Y_\alpha$, $Y_\beta$, and $Y_\delta$) using the SVM accuracy function. vi) Update the positions of the current search agents using equation (16). vii) While the stopping condition (the maximum number of iterations) is not met, repeat steps (iv) and (v).

## 4. Performance Measures and Results

### 4.1 Microarray datasets

The proposed system was evaluated using skewed cancer gene expression datasets downloaded from the Kent Ridge Bio-medical Data Set website (http://datam.i2r.a-star.edu.sq/datasets/krbd/) [34]. The Kent Ridge Bio-medical Data Set Repository is an online repository of high-dimensional biomedical datasets including gene expression data, protein-profiling data and genomic sequence data that are related to classification. Each microarray dataset is in the form of a matrix that consists of M rows, corresponding to the samples, and N columns, corresponding to the genes. We used two sets of patient data to predict breast and colon cancer. Table 2 presents detailed information about the breast and colon cancer microarray datasets.

Table 2. Descriptions of the datasets

| Dataset | Classes | Genes | Samples | Class Distribution |
|---|---|---|---|---|
| **Breast [34]** | relapse, non-relapse | 24481 | training samples: 78, test samples: 19 | training samples: 34 relapse & 44 non-relapse, test samples: 12 relapse & 7 non-relapse |
| **Colon [34]** | normal ("positive"), tumour ("negative") | 2000 | 62 | positive: 22 negative: 40 |

The breast cancer dataset contains 24,481 genes arranged in a matrix. The rows of the matrix represent the genes (features), while the columns represent the samples/instances (patients). This microarray dataset is divided into two matrices: a training matrix and a test matrix. Table 2 summarizes the details of the breast cancer microarray dataset. The training dataset contains prognosis results for 78 patients, 34 of whom are relapse cases and 44 of whom are non-relapse cases. The 34 relapse patients are those for whom distant metastases were observed within 5 years, while the remaining 44 non-relapse instances represent patients who remained cured of the disease for at least 5 years after preliminary diagnosis. The test matrix contains 12 relapse instances and 7 non-relapse instances. For each test, two important criteria were utilized for observational assessment of the performance: the number of selected genes (features) and the classification accuracy [34].

Colon cancer, also called colorectal cancer, is a type of cancer caused by uncontrolled cell growth in the colon, rectum, or vermiform appendix. The two classes in the colon cancer dataset are somewhat different from those in the previous one. In the breast cancer dataset, all samples were collected from cancer patients, and the objective of the proposed classification system is to determine to which type of cancer a new sample belongs. By contrast, the colon cancer dataset contains data on 62 colon

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



adenocarcinoma specimens taken from patients, 40 of which were real tumours and the other 22 of which were not tumours. Therefore, the objective of the classification system is to determine whether a new sample is a tumour. The gene expression data matrix contains the expression results for the 2000 genes with the highest minimal intensities across the 62 tissue samples. Accordingly, the entire gene expression data matrix has dimensions of 2000*62. The training data matrix here has dimensions of 2000*32, and the test data matrix has dimensions of 2000*30. Note that the genes are organized in the matrix in order of descending minimal intensity. This means that the expression values are not normalized with respect to the mean intensity in each experiment [34, 35].

**4.2. Accuracy analysis**

The number of true positives (TruePo) is the number of positive cases that are correctly detected. The number of true negatives (TrueNe) is the number of negative cases that are correctly detected. The number of false positives (FalsePo) is the number of negative cases that are diagnosed as positive. The number of false negatives (FalseNe) is the number of positive cases that are diagnosed as negative. The accuracy represents how close the predictions come to the actual values. A high accuracy and high precision indicate that the test procedure functions well with a meaningful hypothesis. The general equation for accuracy is as shown in equation (17) [27].

**4.3. Results analysis**

In this section, the proposed system is benchmarked on a gene expression profile dataset for breast and colon cancer that has been utilized by other researchers [33, 11, 16, 35, 34]. Three methodologies for classifying microarray datasets are considered. In the first, the SVM classifier is first applied without any feature selection, and then the wrapper feature selection approach based on the GWO algorithm is applied in combination with the same classifier on the same dataset; the results obtained in this way are presented in table 3. In the second methodology, the filter feature selection approach based on the IG algorithm is applied in combination with the SVM classifier; the results are shown in table 4. The third methodology involves a hybrid feature selection approach using both the IG algorithm and GWO in combination with SVM classification; this methodology achieves the best classification accuracy, as also shown in table 4.

C#.net 2018 was used to implement the proposed system. The Weka tool suite version 3.8 was employed in C#.net to apply the IG filtering approach to each dataset for attribute selection. Then, the number of selected attributes was reduced by GWO, programmed in C#.net. Finally, the SVM classifier was called from Weka into C#.net to determine the final classification accuracy. The proposed system uses 5-fold cross-validation [32].

Table 3 shows the results and parameter values for the first tested methodology. The breast cancer dataset contains 24482 genes; when classification was performed on this dataset using the SVM classifier alone, the classification accuracy did not exceed 65%. When the data were first subjected to GWO with 35 wolves and 75 iterations, the classification accuracy increased to 71.795%, and the number of considered genes was reduced to 16055; when the same original data were subjected to both IG filtering and GWO before SVM classification, the classification accuracy reached 88.46%, and the number of genes was reduced to 455 , as shown in table 4. Table 3 also shows the results and parameter settings for the colon cancer dataset. This dataset contains 2000 genes, and when classification was performed on this dataset using the SVM classifier alone, the classification accuracy did not exceed 63%. When the data were first subjected to GWO with 120 wolves and 160

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



iterations, the classification accuracy increased to 85.484%, and the number of considered genes was reduced to 999; when the same original data were subjected to both IG filtering and GWO before SVM classification, the classification accuracy reached 90.32%, and the number of genes was reduced to 70, as shown in table 4.

Table 3. Classification accuracy achieved with the SVM classifier alone and with SVM in combination with GWO using 5-fold cross-validation

| **Dataset** | **SVM** | | **SVM + GWO** | | | |
|---|---|---|---|---|---|---|
| | No. of genes | Acc in % | No. of wolves | No. of iterations | No. of genes | Acc in % |
| Breast | 24482 | 65 | 25 | 50 | 16285 | 70.512 |
| | | | 35 | 75 | **16055** | **71.795** |
| | | | 50 | 100 | 16122 | 70.512 |
| | | | 100 | 20 | 16104 | 70.512 |
| | | | 100 | 25 | 12259 | 70.512 |
| Colon | 2000 | 63 | 20 | 30 | 1311 | 83.87 |
| | | | 30 | 30 | 1014 | 83.87 |
| | | | 40 | 30 | 1012 | 83.87 |
| | | | 50 | 30 | 963 | 83.87 |
| | | | 200 | 50 | 1017 | 83.87 |
| | | | 75 | 100 | 1335 | 85.484 |
| | | | 100 | 120 | 1013 | 83.87 |
| | | | 120 | 160 | **999** | **85.484** |

Table 4. Classification accuracy achieved with SVM classification in combination with IG feature selection with a threshold value of zero using 5-fold cross-validation

| **Dataset** | | **IG + SVM** | | **IG + GWO + SVM** | | | |
|---|---|---|---|---|---|---|---|
| No. of genes before IG selection | | No. of genes | Acc in % | No. of iterations | No. of wolves | No. of genes | Acc in % |
| Breast | 24482 | 715 | 82 | 50 | 20 | 470 | 88.46 |
| | | | | 70 | 50 | 478 | 88.46 |
| | | | | 120 | 100 | **455** | **88.46** |
| | | | | 20 | 12 | 504 | 87.17 |
| | | | | 50 | 15 | 457 | 88.46 |
| Colon | 2000 | 135 | 87.096 | 25 | 13 | 81 | 90.32 |
| | | | | 30 | 15 | 80 | 90.32 |
| | | | | 50 | 20 | 66 | 90.32 |
| | | | | 75 | 50 | **70** | **90.32** |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



Table 5. Classification accuracy achieved with SVM classification in combination with IG feature selection with multiple IG thresholds using 5-fold cross-validation

| Dataset | IG + SVM with multiple IG threshold values | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | threshold = 0.17 | | threshold = 0.2 | | threshold = 0.198 | | threshold = 0.29 | |
| | No. of genes | Acc | No. of genes | Acc | No. of genes | Acc | No. of genes | Acc |
| Breast | 350 | 80.77 | 398 | 84.61 | 441 | 83.3 | 28 | 78.2 |
| Colon | 135 | 87.09 | 108 | 85.48 | 117 | 87.9 | 31 | 82.25 |

Tables 5 and 6 show the parameter settings and the results obtained with the proposed system with and without GWO, respectively, using 5-fold cross-validation. The parameter settings are the same as those in table 4 except that the IG threshold value is varied. As seen from tables 4 and 5, when the threshold value was changed to 0.17 or to 0.2 or more, the classification accuracy achieved was lower than the best result achieved with a threshold value of zero. However, although better accuracy results were achieved with no threshold IG value, using a threshold made it possible to reduce the number of features from 7129 to 32, thereby decreasing the time and memory consumption needed for the classification process.

Table 6. Classification accuracy achieved with IG + GWO + SVM with multiple IG thresholds using 5-fold cross-validation

| Dataset | IG + GWO + SVM | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. of wolves | No. of iterations | Acc with multiple IG threshold values and different numbers of genes | | | | | | | |
| | | | threshold = 0.17 | | threshold = 0.2 | | threshold = 0.198 | | threshold = 0.29 | |
| | | | No. of genes | Acc | No. of genes | Acc | No. of genes | Acc | No. of genes | Acc |
| Breast | 20 | 50 | 338 | 88.46 | 282 | 91.026 | 290 | 89.74 | 20 | 83.3 |
| | 50 | 70 | 349 | 89.74 | 260 | 91.026 | 249 | 91.026 | 17 | 83.3 |
| | 100 | 120 | 337 | 89.74 | 272 | 92.307 | **250** | **94.87** | 16 | 84.61 |
| | 120 | 150 | 351 | 91.025 | 245 | 92.307 | 270 | 93.59 | 18 | 84.61 |
| Colon | 13 | 25 | 74 | 88.7 | 50 | 88.7 | 74 | 90.322 | 17 | 94.322 |
| | 15 | 30 | 75 | 90.322 | 77 | 90.322 | 78 | 90.322 | 23 | 95.935 |
| | 20 | 50 | 64 | 90.322 | 70 | 90.322 | 82 | 90.322 | 17 | 94.322 |
| | 50 | 75 | 85 | 90.322 | 62 | 90.322 | 56 | 90.322 | **16** | **95.935** |

Table 7. Best results with multiple IG threshold values

| Dataset | Threshold | No. Genes | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| Breast | 0.198 | 250 | 94.87 | 0.95 | 0.90 | 0.92 |
| Colon | 0.29 | 16 | 95.935 | 0.952 | 0.909 | 0.93 |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 February 2020 doi:10.20944/preprints202002.0324.v1

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



Table 8. Accuracy comparison of several different classifiers

| Reference | Dataset | Classifier | Accuracy in % |
|---|---|---|---|
| [33] | Colon, Breast | C4.5 | 76.19, 61 |
| | | naïve Bayes | 52.14, 51.89 |
| | | IB1 | 73.38, 60.22 |
| This work | Colon, Breast | SVM | 63, 65 |

Table 7 summarizes the best results obtained when applying the proposed methodology to the two datasets (Breast and Colon). In table 8, we review the classification accuracies of several different classifiers for comparison with the SVM classifier.

Tables 9 and 10 show the differences between the classification accuracies of different methodologies based on hybrid feature selection approaches (filter and wrapper approaches) when applied to CNS, colon and breast cancer data. As shown, the best results achieved with the proposed methodology are 94.87% (Breast) and 95.935% (Colon). A comparison of the experimental results reveals that the proposed system offers improved sample classification accuracy. These experimental results show that the proposed strategy is able to improve the stability of the feature selection results as well as the sample classification accuracy.

Table 9. Classification accuracy of the proposed methodology vs. other methodologies on Breast.

| Reference | Methodology | | | Accuracy in % |
|---|---|---|---|---|
| [34] | ReliefF + 3-NN | | | 70.96 |
| [12] | IG + GA | | | 100% |
| [35] | Optimized Fuzzy Rule Generation (OFRG) algorithm | | | 94 |
| [36] | filtering and normalization + PSO + SVM | | | 94 |
| | filtering and normalization + GA + SVM | | | |
| This work | IG | GWO | SVM | 94.87 |

Table 10. Classification accuracy of the proposed methodology vs. other methodologies on Colon.

| Reference | Methodology | | | Accuracy in % |
|---|---|---|---|---|
| [35] | T-Statistics, SNR, F-Test | GA | SVM | 85 |
| | T-Statistics, SNR, F-Test | GA | KNN | 85 |
| [34] | Random + SVM | | | 88.41 |
| [14] | Fisher, T-Statistics, SNR and ReliefF + KNN and SVM | | | 95% |
| [26] | IG + GA + PG | | | 85.48 |
| This work | IG | GWO | SVM | 95.935 |

**5. Conclusions**

In this research, an enhanced IDSS is proposed based on IG feature selection, the GWO algorithm and SVM classification. The proposed system employs the IG method for initial feature selection, while GWO is used to reduce the number of selected features to enable more accurate sample

Peer-reviewed version available at Symmetry 2020, 12, 408; doi:10.3390/sym12030408



classification by the SVM. Two microarray datasets are used as benchmarks to evaluate the proposed methodology. The experimental results indicate that the proposed methodology is able to enhance the stability of the classification accuracy as well as the feature selection. The best results are obtained when combining the IG approach with both the GWO and SVM algorithms; the classification accuracy reaches 94.87% for breast cancer data and 95.935% for colon cancer data. In future work, additional classifiers should be added to the system. In addition, there is a possibility of testing the system on other benchmarks, especially binary-class datasets and test the reliability of diagnosis after repeated sampling of tissue from the same patient.

## 6. References

[1] D. Walker, A. Bendel, C. Stiller, D. Indelicato, S. Smith, M. Murray and A. Bleyer, "Central nervous system tumors," in Cancer in adolescents and young adults, Springer, 2017, pp. 335-381.

[2] D. A. Walker, A. Bendel, C. Stiller, P. Byrne and M. Soka, "Central Nervous System Tumors," in Cancer in Adolescents and Young Adults, W. A. Bleyer and R. D. Barr, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 151-183.

[3] "cancer.net," American Society of Clinical Oncology (ASCO), [Online]. Available: https://www.cancer.net/cancer-types/central-nervous-system-childhood/view-all. [Accessed 8 2018].

[4] Siegel, R.L., Miller, K.D. and Jemal, A. , "Cancer statistics, 2019," CA A Cancer J Clin , vol. 69 , pp. 7-34 , 2019.

[5] R. R. Janghel, A. Shukla, R. Tiwari and R. Kala, "Intelligent decision support system for breast cancer," in International Conference in Swarm Intelligence, 2010.

[6] A. B. Al-Badareen, M. H. Selamat, M. Samat, Y. Nazira and O. Akkanat, "A Review on Clinical Decision Support Systems in Healthcare," Journal of Convergence Information Technology, vol. 9, p. 125, 2014.

[7] W. A. Berg, L. Gutierrez, M. S. NessAiver, W. B. Carter, M. Bhargavan, R. S. Lewis and O. B. Ioffe, "Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer," Radiology, vol. 233, pp. 830-849, 2004.

[8] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," Computerized medical imaging and graphics, vol. 31, pp. 198-211, 2007.

[9] M. E. Ahsen, T. P. Boren, N. K. Singh, B. Misganaw, D. G. Mutch, K. N. Moore, F. J. Backes, C. K. McCourt, J. S. Lea, D. S. Miller and others, "Sparse feature selection for classification and prediction of metastasis in endometrial cancer," BMC genomics, vol. 18, p. 233, 2017.

[10] V. Elyasigomari, D. A. Lee, H. R. C. Screen and M. H. Shaheed, "Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification," Journal of biomedical informatics, vol. 67, pp. 11-20, 2017.

[11] H. Salem, G. Attiya and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," Applied Soft Computing, vol. 50, pp. 124-134, 2017.

[12] H. Salem, G. Attiya and N. El-Fishawy, "Early diagnosis of breast cancer by gene expression profiles," Pattern Analysis and Applications, vol. 20, pp. 567-578, 2017.

[13] J. Bennet, C. Ganaprakasam and N. Kumar, "A Hybrid Approach for Gene Selection and Classification using Support Vector Machine.," International Arab Journal of Information Technology (IAJIT), vol. 12, 2015.

[14] S. H. Bouazza, N. Hamdi, A. Zeroual and K. Auhmani, "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers," in Intelligent Systems and Computer Vision (ISCV), 2015, 2015.

[15] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," Advances in bioinformatics, vol. 2015, 2015.

[16] J.-Y. Yeh, "Applying data mining techniques for cancer classification on gene expression data," Cybernetics and Systems: An International Journal, vol. 39, pp. 583-602, 2008.

[17] J. C. Baez, T. Fritz and T. Leinster, "A characterization of entropy in terms of information loss," Entropy, vol. 13, pp. 1945-1957, 2011.

[18] L. Chen, K. Wu and Y. Li, "A load balancing algorithm based on maximum entropy methods in homogeneous clusters," Entropy, vol. 16, pp. 5677-5697, 2014.

[19] M. W. Mwadulo, "A Review on Feature Selection Methods For Classification Tasks," International Journal of Computer Applications Technology and Research, vol. 5, pp. 395-402, 2016.

[20] O. Okun, Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations, Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2011.

[21] M. Bramer, Principles of data mining, vol. 180, Springer, 2007.

[22] S. Mirjalili, S. M. Mirjalili and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46-61, 2014.

[23] L. D. Mech, "Alpha status, dominance, and division of labor in wolf packs," Canadian Journal of Zoology, vol. 77, pp. 1196-1203, 1999.

[24] D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision support system for medical diagnosis using data mining," International Journal of Computer Science Issues, vol. 8, pp. 147-153, 2011.

[25] C. Muro, R. Escobedo, L. Spector and R. P. Coppinger, "Wolf-pack (Canis lupus) hunting strategies emerge from simple rules in computational simulations," Behavioural processes, vol. 88, pp. 192-197, 2011.

[26] E. Emary, H. M. Zawbaa, C. Grosan and A. E. Hassenian, "Feature Subset Selection Approach by Gray-Wolf Optimization," in Afro-European Conference for Industrial Advancement, Cham, 2015.

[27] X. Song, L. Tang, S. Zhao, X. Zhang, L. Li, J. Huang and W. Cai, "Grey Wolf Optimizer for parameter estimation in surface waves," Soil Dynamics and Earthquake Engineering, vol. 75, pp. 147-157, 2015.

[28] E. Emary, H. M. Zawbaa and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," Neurocomputing, vol. 172, pp. 371-381, 2016.

[29] S. Marsland, Machine learning: an algorithmic perspective, Chapman and Hall/CRC, 2015.

[30] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau and others, "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, p. 436, 2002.

[31] S.-B. Cho and H.-H. Won, "Machine learning in DNA microarray analysis for cancer classification," in Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19, 2003.

[32] A. Isaksson, M. Wallman, H. G{\"o}ransson and M. G. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," Pattern Recognition Letters, vol. 29, pp. 1960-1965, 2008.

[33] Bolón-Canedo, V. Sánchez-Maroño, N. Alonso-Betanzos, A., "An ensemble of filters and classifiers for microarray data classification," Pattern Recognition, vol. 45, pp. 531-539, 2012.

[34] Alonso-González, Carlos J. Moro-Sancho, Q. Isaac Simon-Hurtado, Arancha Varela-Arrabal, Ricardo, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods," Expert Systems with Applications, vol. 39, pp. 7270-7280, 2012.

[35] C. Gunavathi and K. Premalatha, "Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification," Int J Comput Electr Autom Control Inf Eng, vol. 8, pp. 1490-7, 2014.

[36] Paul, Amit Sil, Jaya Mukhopadhyay, Chitrangada Das, “Gene selection for designing optimal fuzzy rule base classifier by estimating missing value”, Applied Soft Computing, vol. 55, pp. 276-288, 2017.

[37] Niloofar Yousefi Moteghaed, Keivan Maghooli, and Masoud Garshasbi, “Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine”, J Med Signals Sens, vol. 8, pp. 1-11, 2018.