https://doi.org/10.17048/AM.2020.122

Hanaa Salem

Communications & Computer Department, Faculty of Engineering, Delta University for Science&Technology, Egypt.

hana.salem@deltauniv.edu.eg

Gamal Attiya

Computer Science & Engineering Department, Faculty of Electronic Engineering, Menoufia University, Egypt.

gamal.attiya@yahoo.com

Nawal El-Fishawy

Computer Science & Engineering Department, Faculty of Electronic Engineering, Menoufia University, Egypt.

nelfishawy@hotmail.com

Multi-Agent based Intelligent Decision Support Systems for Cancer CLASSIFICATION

Abstract

There is evidence that early detection of cancer diseases can improve the treatment and increase the survival rate of patients. This paper presents an efficient CAD system for cancer diseases diagnosis by gene expression profiles of DNA microarray datasets. The proposed CAD system combines Intelligent Decision Support System (IDSS) and Multi-Agent (MA) system. The IDSS represents the backbone of the entire CAD system. It consists of two main phases; feature selection/reduction phase and a classification phase. In the feature selection/reduction phase, eight diverse methods are developed. While, in the classification phase, three evolutionary machine learning algorithms are employed. On the other hand, the MA system manages the entire operation of the CAD system. It first initializes several IDSSs (exactly 24 IDSSs) with the aid of mobile agents and then directs the generated IDSSs to run concurrently on the input dataset. Finally, a master agent selects the best classification, as the final report, based on the best classification accuracy returned from the 24 IDSSs.

The proposed CAD system is implemented in JAVA, and evaluated by using three microarray datasets including; Leukemia, Colon tumor, and Lung cancer. The system is able to classify different types of cancer diseases accurately in a very short time. This is because the MA system invokes 24 different IDSS to classify the diseases concurrently in parallel processing manner before taking the decision of the best classification result.
Keywords: Computer Aided Diagnosis (CAD), Intelligent Decision Support System (IDSS), Multi-Agent System, gene expression profile, machine learning, medical diagnosis.

1. Introduction

Medical diagnosis is the process of determining which disease or condition explains a person's symptoms and signs. The conventional method for diagnosing most of the existing diseases depends on human skills to recognize the occurrence of the convincing pattern. This age-old diagnosis method may subject to human mistake, imprecise diagnosis, time-consuming and labor intensive, and causes an unnecessary burden to radiologists. Moreover, by the right time of the diagnosis completed, it may already be at a critical stage.

Recently, Computer Aided Diagnosis (CAD) and machine learning systems have been developed and functional in order to support specialists in the determination of the diagnosis decision process [1]. Artificial Intelligence (AI) has become commonly used in health-related decision support systems (Raval et al., 2015, Bassen et al., 2013)

An artificial intelligent algorithm entitled DS-STM (Diagnosis Strategy of Serum Tumor Makers) is developed to provide decision support for physicians on the usage of different tumor markers and diagnosis of colorectal cancer (Shi et al., 2010). The DS-STM improves the diagnostic accuracy from 67,53% to 73,87% for the same validation dataset. A Clinical Decision Support System (CDSS) combining the results of classic and ancillary techniques is developed based on artificial neural networks, for intelligently diagnostic accuracy improvement (Bountris et al., 2014). The CDSS demonstrated high sensitivity (89,4%), high specificity (97,1%), high positive predictive value (89.4%), and high negative predictive value (97,1%), for detecting cervical intraepithelial neoplasia. An integrated expert system developed for diagnosis, prognosis, and prediction of breast cancer using soft computing techniques (Janghel et al., 2010; Kourou et al., 2015).

Other studies focused on building medical diagnosis systems based on agents (iantovics, 2008). A literature survey of agent-based applications in healthcare is discussed in (Isern and Moreno, 2008). The agent-based simulation in healthcare is continuously allowing the study of complex diseases such as diabetes or hepatitis C and even some kinds of cancer. Multi-agent systems have been applied in various disciplines of medical applications (Chakraborty and Gupta, 2014). A survey of multi-agent

based IDSS for medical classification problems is presented in (Salem et el., 2015). An agent-based IDSS for the home healthcare environment is developed in (Cervantes et al., 2007; Asadi et al., 2009; Vélez et al., 2009).

The proposed CAD system combines Intelligent Decision Support System (IDSS) and Multi-Agent (MA) system. The IDSS, the backbone of the entire CAD system, consists of two main phases; feature selection/reduction phase, and a classification phase. In the feature selection and reduction phase, eight diverse methods are developed. These methods include Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Correlated-based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR), Relief-F, Chi-Square, and Support Vector Machine with Recursive Feature Elimination (SVM-RFE). While, in the classification phase, three evolutionary machine learning algorithms are employed, J48, Naïve Bayes (NB), Genetic Algorithm (GA). On the other hand, the Multi-Agent (MA) system manages the entire operation of the CAS system. It first initializes several IDSSs (exactly 24 IDSSs) with the aid of mobile agents and then directs the generated IDSSs to run in concurrently. The MA system constructs an IDSS by selecting one of the eight feature selection methods and one of the three cancer classification algorithms. Finally, a master agent selects the best classification, as the final report, based on the best classification accuracy returned from the 24 IDSSs. Since the proposed CAD system allows 24 different IDSSs to be created and run concurrently for cancer diseases classification of an input dataset, this CAD system able to provide best classification accuracy in short time.

The proposed CAD system is implemented in JAVA and evaluated by using three microarray datasets including; Leukemia, Colon tumor, and Lung cancer. The main advantage of the proposed CAD system is that it classifies the cancer diseases accurately in a very short time. This is because cancer classification is done in parallel processing manner. Where, the MA system invokes 24 different IDSS to classify the diseases on the input dataset concurrently before taking a decision of the best classification result

The rest of this paper is organized as follows. Section 2 introduces the classification problem and its own challenges. Section 3 presents a literature survey of related work. Section 4 presents the proposed CAD system in details including the structure of both the IDSS and the MA system. Finally, the experimental results are discussed in Section 5 while the concluding remarks are lists in Section 6.

1.1 Problem Definition and Challenges

Medical diagnosis by most of the existing CAD systems depends on different sorts of information, such as medical lab tests (e.g. blood testing, Magnetic Resonance Angiography (MRA)), medical, or symptoms and different types of digital images. Nevertheless, different types of diagnosis tests have different risks. For examples, potential risks of X-rays are pregnant radiation risk and risk is increased for children. Potential dangers of blood testing are pain to acquire blood, a needle must be utilized,

and risk of Infection. Ultrasound images depends on quality and interpretation of the image (the skill of the individual doing the scan determine the quality and interpretation of the image) and there are other factors make a difference image quality, like the occurrence of air and calcified areas in the body and someone's body size. This paper tackles the problem of medical diagnosis and presents a Computer Aided Diagnosis (CAD) system for cancer diseases diagnosis by gene expression profiles of DNA microarray datasets. However, two challenges posed in microarray classification namely data dimensionality and classification accuracy.

1.1.1 Data Dimensionality

DNA microarray technology has been used in a wide variety of experimental researches for cancer disease prediction. However, because of the large number of features (in the request of thousands) and the little number of samples (basically not as much as a hundred) in this type of datasets, microarray data analysis faces the “large p- small n” paradigm also known as the execrate of dimensionality. Naturally, gene expression dataset keep a high dimension and a small sample size. This makes testing and training of general classification methods very hard. In general, only a relatively small number of gene expression data out of the total number of genes considered shows a significant correlation with a certain phenotype.

1.1.2 Classification Accuracy

Scientists in a wide spectrum of research topics investigate how to alleviate or to improve the life quality of people suffering from different terrible condition. An accurate prediction of different tumor types provides better treatment and toxicity minimization on patients. In spite of the extremely concentrated research strength, difficulties postured microarray classification are the accessibility of just a set number of samples in comparison and examination to the high dimensionality of the samples, and empirical variants in measured gene expression levels. The minor number of cancer samples typically available to train the model compared with the number of genes features present can reduce the performance of the classifier and intensification the risk of over-fitting. Cancer prediction based on gene expression data contains a great number of features, which needs a relatively large training set to learn a classifier with a small error rate.

2. Related Work

Recently, a wide range of publications focused on the classification of cancer diseases with the usage of the microarray datasets. In 2012, several techniques have been developed and tuned aiming to early detect breast cancer. An expert system that based on the artificial neural networks (ANN) are

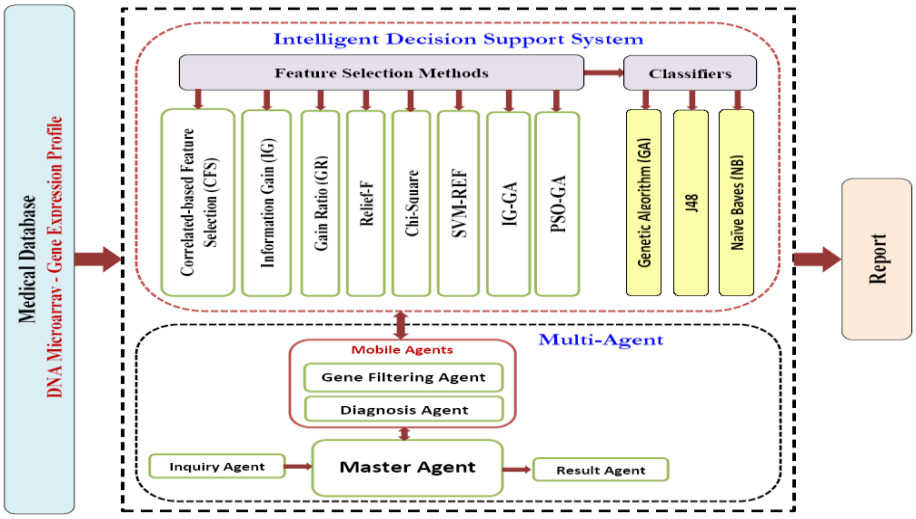
used for the automatic breast cancer diagnosis is becoming popular among researchers (Fatima and Amine, 2012). In the meantime, artificial intelligence is combined with fuzzy in a proposed hybrid model (Khashei et al., 2012). Naive Bayes classifier algorithm and J48 implementation of the C4.5 algorithm is applied to predict lung cancer survivability from an extensive data set with fifteen years of patient records (Dimitoglou et al., 2012). Several selection methods such as (plus a Random selection, Fast Correlation-Based Filter (FCBF), ReliefF and SVM-RFE) are used as a baseline technique and classifying techniques (Yu el al., 2012). Nearest Neighbor, Naive Bayes, and SVM with linear kernel are applies to 30 datasets involving different types of cancer (González et al., 2012). In 2013, a novel idea that based on using feature estimation method (Relief-F) to predicted feature selection method and merges it with the genetic algorithm that applied to discover a group of genes that can best differentiate between cancers subtypes or normal versus cancers instances. Many types of classification systems are applied to 5 cancers datasets demonstrates that there is no classification method can be used for all types (Hijazi and Chan, 2013). In 2014, a gene is chosen from each group depending on a specific ranking method forming the used set of genes. Information Gain is the most attribute evaluation methods that widely used. Where, the univariate filter provides features in ordered ranking form and then a threshold is required. At this stage SVM, Discontinuity Adaptive (DA), ANNs, GA, Naive Bayes (NB) and K-nearest neighbor (KNN) classifiers were used for classification (Wang, 2014). In (Lu et al., 2014), GA is used as a feature selection algorithm that using a specifically devised trace-based reparability criterion while the lung cancer microarray dataset is classified by using the back-propagation neural network (BPNN), SVM, and (KNN). In 2015, several gene selection strategies are developed depend on Bio-Inspired evolutionary approaches, such as, GA, PSO, or Ant Colony Optimization (ACO). These gene selection methods are equipped of scanning for ideal or near-ideal answers for complex and large spaces of conceivable solutions (Alshamlan et al., 2015; Kourou et al., 2015; Danjuma, 2015; Aziz et al., 2016). Then, the datasets are then classified using support vector machine (SVM) and Naïve Bayes (NB) classifiers. An IDSS for breast cancer diagnosis by using gene expression profiles is developed in (Hanaa et al., 2016; 2017). The methodology combines both Information Gain (IG) and Standard Genetic Algorithm (SGA). It first uses IG for feature selection, then uses Genetic Algorithm (GA) for feature reduction and finally uses Genetic Programming (GP) for cancer types' classification.

Although large number of research articles concerned with cancer classification by using microarray gene expression are presented, there is not known neither the best gene selection techniques nor the best classifiers to be used. The wide range of different gene selection algorithms and classifiers made it possible to combine any gene selection algorithm with any classifier to form a separate system.

3. Proposed CAD System

Figure 1 shows the proposed CAD system architecture. The system combines several intelligent decision support systems and multi-agent system. The IDSS performs the required cancer diseases classification while the multi-agent system allows 24 different IDSSs to run concurrently on the input dataset in order to accurately classify the diseases in very short time. The proposed system is called Multi Agent based parallel processing of Intelligent Decision Support systems (MA-IDSS).

Figure 1 Proposed CAD system architecture



3.1 Intelligent Decision Support Systems

From Figure 1, the proposed IDSS consists of two main phases; feature selection/reduction phase, and a classification phase. In the feature selection/reduction phase, eight diverse methods are developed. These methods include Correlated-based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR), Relief-F, Chi-Square, Support Vector Machine with Recursive Feature Elimination (SVM-RFE), Information Gain with Genetic Algorithm (IG-GA), and Particle Swarm Optimization with Genetic Algorithm (PSO-GA). While, in the classification phase, three evolutionary machine learning algorithms are employed, J48, Naïve Bayes (NB), Genetic Algorithm (GA). The IDSS first receives a gene expression profiles dataset and then performs the feature selection and classification process. The feature selection is performed by using one of the eight approaches while the cancer classification is performed by using one of the three algorithms. Since there are eight feature selection methods and three classification methods, then the proposed CAD system allows 24 different IDSSs to be created and used for cancer diseases classification. The idea is to support the proposed IDSS by different selectors and classifiers to classify the diseases accurately by selecting the appropriate selector/classifier.

3.1.1 Feature Selection Methods:

This section presents the proposed feature selection methods in some details.

1. **Correlation-Based Feature Selection (CFS)**

The degree of redundancy between the features to get subsets of the feature is known as CFS (Koprinska et al., 2015). The evaluator intends to discover the subsets of features that are separately exceedingly corresponded with the class yet have low inter-correlation.

$$r_{sc} = \frac{n\,\overline{r_{si}}}{\sqrt{n+n(n-1)\overline{r_{ii}}}} \qquad (1)$$

Where, rsc is the correlation among feature subsets and the class, n is the features number, rsi is the average of the correlations among the features and the class, and rii is the average inter-correlation between features.

2. **Information Gain (IG)**

IG is a feature ranking technique. Y is a discrete random variable with probability function P, its entropy is characterized by:

$$H(Y) = -\sum_i P(y_i)\,\log_2(P(y_i)) \qquad (2)$$

For the classification framework, class S is variable, so the entropy of the classification framework can be characterized as:

$$H(S) = -\sum_{i=1}^{l} P(s_i)\,\log_2(P(s_i)) \qquad (3)$$

For a gene Y, it may have n possible values (y1, y2, …… , yn). The corresponding conditional entropy is

$$H\left(S/y\right) = -\sum_{j=1}^{n} P(y_j)\,\sum_{i=1}^{l} P(s_i/y_j)\,\log_2(P(s_i/y_j)) \qquad (4)$$

$$IG(Y) = H(S) - H(S/Y) \qquad (5)$$

If gene X and category S are not relevant IG(Y) = H(S) – H(S/Y) = zero. Well, if relevant, H(S) > H(S/Y), i.e., IG(Y) = H(S) – H(S/Y) > 0. The larger the difference is, the stronger the correlation between Y and S. Therefore, when choosing genes, usually choose genes with great information gain to signify the original high-dimensional gene first, and use them as an origin for further gene selection (Xu and Jiang , 2015).

3. **Gain Ratio (GR)**

Gain Ratio incorporates “split information” of features into an Information Gain statistic. The “split information” of a gene is obtained by measuring how broadly and uniformly it splits the data. We should think about again a microarray dataset has a set of classes denoted as si, (i = 1, ..., m), and every feature f has a set of possible values denoted as K (Mwadulo , 2016).

The gain ratio of a feature f is given as:

$$\text{GainRatio}(f) = \frac{\text{Information Gain}(f)}{\text{Split}(g)} \quad (6)$$

In which:

$$\text{Split}(f) = -\sum_{k \in K} \sum_{i=1}^{m} \frac{|S_k|}{|S|} \log \frac{|S_k|}{|S|} \quad (7)$$

Where, S_k is the subset of S of which feature f has value k.

4. Relief-F

The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. In particular, it tries to locate a good estimate of the following probability to appoint as the weight for every feature f (Dittman et al., 2012).

$$w_f = P(\text{a different value of } f/\text{different class}) - P(\text{a different value of } f/\text{same class}) \quad (8)$$

5. CHI (X^2 Statistics)

Chi-Squared is the basic statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is really independent of the class value (Canedo et al., 2016). The X2 statistic measure how far away from the real value is the expected value:

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,i}} \quad (9)$$

In this equation, n is the number of various values of the feature in equation, c is the number of classes in Oi,j is the number of instances with value i which are in class j, and Ei,j is the predictable number of instances with value i and class j, taking into account (p·q)/n. The bigger this chi-squared statistic, the more unlikely it is that the distribution of values and classes are independent; that is, they are related, and the feature in question is relevant to the class.

6. Support Vector Machine with Recursive Feature Elimination (SVM-RFE)

The SVM-RFE is an application of RFE using the weight magnitude as a ranking criterion (Zhang and Huang, 2015).

Algorithm SVM-RFE:
Inputs: Features for training Y0 = [y1, y2,... yk,... yL]T
Classes c = [c1, c2,... ck,... cl]T
Initialize:
Subgroup of remaining features subf = [1, 2,... n]
Ranke Feature Matrix ra = [ ]
Repeating till subf = [ ]
Restrict training Subgroup of remaining features to significant feature index Y= Y0(:, subf)

Classifier training γ= train-SVM(Y, c)

Calculate the weight vector of (subf) $w_v = \sum_k \gamma_k rc_k y_k$

Calculate criteria of ranking criteria rci = $(w_{vi})2$, for all i

Search for the feature has smallest (rc) f = argmin(rc)

Bring up to date ranked feature matrix ra = [subf(f), ra]

Remove the feature has lowest ranked criterion value subf = subf (1:g − 1, g + 1:length(subf))

Output: Matrix of feature ranked values ra.

7. IĢ-ĢA

The Information Gain/Genetic Algorithm (IĢ-ĢA) is a hybrid feature selection and reduction method of microarray dataset. This method utilized a two-phase method to execute feature selection then reduction. In the first phase, IĢ value was computed for every gene. In the second phase, all the selected features must comply with a threshold. Therefore, feature selection was once again performed, this time profiting by ĢA to diminish the features.

8. Particle Swarm Optimization/Genetic Algorithm (PSO/GA)

The Particle Swarm Optimization/Genetic Algorithm (PSO-ĢA) is a hybrid feature selection and reduction strategy of microarray dataset (Hanaa et al., 2016). The algorithm works as follows:

1) Randomly create initial population.
2) Introductory population of M*N with the binary system. Where, M is the number of particles in a swarm, and N stands for the length of an individual.
3) Calculate fitness values of the individuals by the fitness function.
4) Update positions and velocity of the particles.
5) If the current iteration does not satisfy termination condition, go to step 3. Otherwise, the current population is the final solution.
6) Perform feature reduction by using GA.

3.1.2 Classification Methods

This section presents the proposed classification methods in some details.

1. Ĵ48 Algorithm (Decision tree)

The Ĵ48 classifier is a simple C4.5 decision tree for classification. A binary tree was constructed. In the topmost 10 procedures in data mining, J48 has been recorded (Wu et al., 2008). In the classification issue, the decision tree methodology is the most valuable. Via this procedure, a tree is developed to show the classification procedure (Bhargava et al., 2013).

J48 Algorithm:

Input:

Ţ // Dataset for Train

Output:

D // Decision tree

DTCONSTRUCT (*Ţ)

{ D=ϕ; D= Build root node and marker with splitting feature;

D= Insert arc to root node for each split establish and marker;

For every arc ensures

Ţ = Database produced by applying splitting establish to Ţ;

If stopping point extended for this pathway,

D'= construct leaf node and marker with the proper class;

Else

D'= DTCONSTRUCT(D); D= Insert D' just before arc;}

2. Naïve Ɓayes classifier

The Naïve Ɓayes is a simple probabilistic classifier that determines a set of probabilities by excluding the frequency and blends of values in a certain dataset. The algorithm utilizes Ɓayes hypothesis and adopts all features to be independent given the value of the class variable (Patil and Sherekar, 2013). This conditional independence assumption rarely holds true in real world applications, hence, the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems (Dimitoglou et al., 2012). The probability that a document d with vector $y = < y_1,...,y_n>$ belongs to hypothesis h is:

$$P((h_1|y_i) = \frac{P(y_i|h_1).\ P(h_1)}{P(y_i|h_1).\ P(h_1)+P(y_i|h_2).\ P(h_2)} \quad (10)$$

Where, $P(h_1|y_i)$ is posterior probability, $P(h_1)$ is the prior probability associated with hypothesis h_1. From different hypotheses,

$$P(y_i) = \sum_{j=1}^{n} P(y_i|h_j)\, P(h_j) \quad (11)$$

Thus,

$$P(h_1|y_i) = \frac{P(y_i|h_1)\, P(h_1)}{P(y_i)} \quad (12)$$

Algorithm NB:

Input:

Y= {Y_1,….,Y_n} // Features values. Ω = T_1 ×…× T_n // The set of all feature sets

C, c∈ {0,…,n −1} // Number of classes . A hypothesis h: α →{0,…,n−1} // assigns a class to any given set of variables is defined as a classifier. f_c(y), c= 0,…,n−1

h(y)= $\text{argmax}_{c \in \{0,\ldots,n-1\}} f_c(y)$ //The classifier selects the class with the maximum discriminant function on a given set of variables

The Bayes classifier h^*(y) uses the posterior probabilities given a set of variables as the discriminant function. Hence, the Bayes' discriminant function can be written as f^*(y)= P(Y=y |C=c) (C=c) //P(Y=y |C=c) P(C=c) the class-conditional probability distribution (CPD) (Rish, 2001).

h^*(y)=argmax_c P(Y=y |C=c) P(C=c) (13)

Output:

We can get the naïve Bayes classifier. By applying the supposition of features are autonomous given the class, $f_c^{NB}(y) = \prod_{j=1}^{n} P(Y_j = y_j | C = c)\, P(C = c)$ (14)

3. Genetic Algorithm (GA) Classifier

In the proposed system, a branch of GA called Genetic Programming (GP) is used as a classifier. The fundamental distinction between GA and GP is individual's structure. GA individuals have string organized while GP's individuals are trees. The GP based classifier is represented by a classification tree. GP methods have been used widely in optimizing classification problems due to their flexibility and adaptability. GP is an evolutionary based on optimization technique. In our algorithm, each individual in GP is an ensemble of decision trees. Each terminal is a single decision tree, while the function is one of the arithmetic operators: F = {+, -, *, /} and T comprises of 10 constants then the value of the expression level of genes is represented by variables, T = {0… 9, y_1... y_n} [28].

The microarray dataset contains information for the variables (y1… yn). For an applicant to assess the fitness, its expression is assessed. On the off chance that the end result of assessing feature is greater than 0, it is categorized as Class 1. Else, it is categorized as Class 2. The feature is assessed with data in the training series. The number of aggregates the right classification is considered as the fitness value of the feature.

3.2 Multi-Agent System

The Multi-Agent (MA) system manages the entire operation of the proposed CAD system. It directs 24 different IDSSs to run concurrently on the input dataset in order to accurately classify the

diseases in very short time. As shown in Figure 1, the MA system consists of Gene Filtering Agent, Diagnosis Agent, Master Agent, Inquiry Agent, and Result Agent. The different agents have specific properties of singular intelligent agents and cooperate with each other to reach the main goal. In the following, agent properties, functions and responsibilities are describes is some details. In addition, the cooperation between agents and the sequence diagram of messages are describes is some details.

3.2.1 Agents Properties and Functions

1. **Gene Filtering Agent**

Gene Filtering Agent gets the objective of forming sets of genes from that gene actually convoluted in the analysis of diseases to be diagnosed.

2. **Diagnosis Agent**

Diagnosis agent creates categories using the DNA microarray samples predicated on the quantifiable data of gene expression profiles.

3. **Master Agent**

Master Agent plays maestro role in the proposed System.

4. **Inquiry Agent**

Inquiry agent is able to browse single instance or more to be classified by the master agent of the proposed system.

5. **Result Agent**

Result agent interrelates with the end user to get his requirements and provides him the results.

3.2.2 Agents Responsibilities

Table 1 summarizes the responsibilities given to each agent of the proposed CAD system.

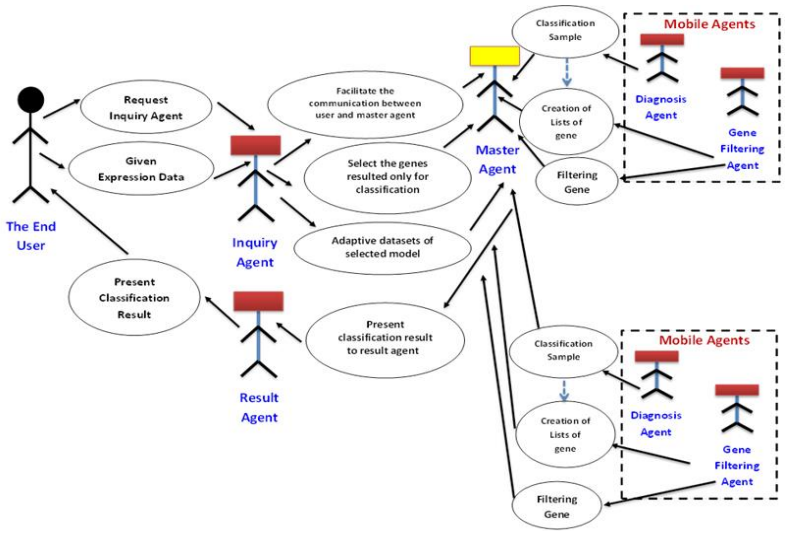
Table 1: Agent Responsibilities

Agent	Responsibilities
Gene Filtering Agent	Get lists of genes - Filter of genes - Request information of genes
Diagnosis Agent	Find patterns of samples - Classifying of samples - Evaluate filter genes
Master Agent	Coordinate other agents -Select the best model of classification - Communication between agents - Interaction with the inquiry agent and result agent
Inquiry Agent	Communication between user and master agent - Select the genes resulted only in classification - Interaction with the master agent
Result Agent	Interaction with the user - Communication with the master agent. - Show the results

3.2.3 Cooperation between Agents

Figure 2 shows the use case of the MA system while the sequence diagram of different messages between all the system agents is shown in Figure 3.

Figure 2: Use Case Diagram of the MA system

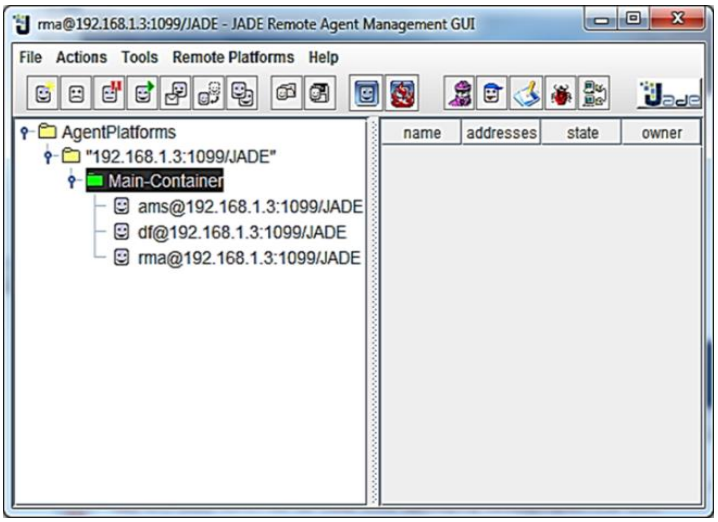


3.3 Proposed CAD System Implementation

In this research work, the proposed CAD system is implemented by using the Java Agent Developer Framework (JADE). JADE is a software framework fully implemented in the Java language. It simplifies the implementation of multi-agent systems through a middle-ware that complies with the Foundation for Intelligent Physical Agents (FIPA) specifications and through a set of graphical tools that support the debugging and deployment phases. A JADE-based system can be distributed across machines (which not even need to share the same OS) and the configuration can be controlled via a

remote GUI, as shown in Figure 3. The configuration can be even changed at run-time by moving agents from one machine to another when required.

Figure 3: Agent Platform

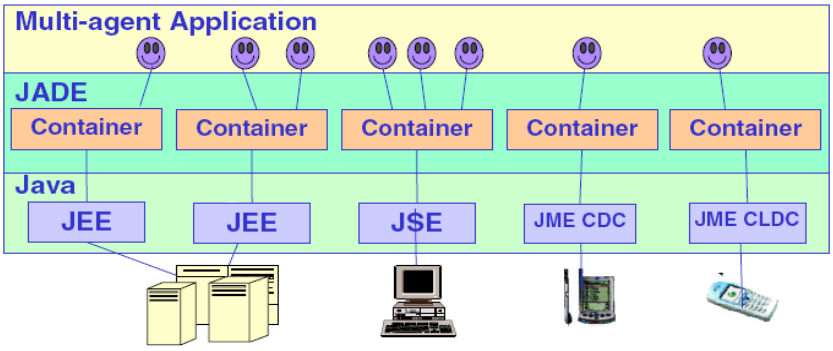


Besides the agent abstraction, JADE provides a simple yet powerful task execution and composition model, peer-to-peer agent communication based on the asynchronous message-passing paradigm, which facilitates the development of a distributed system. The intelligence, initiative, information, resources and control of agents can be fully distributed to mobile terminals as well as on computers in the fixed network (Bellifemine et al., 2008).

A JADE-based application is composed of a collection of active components called Agents. Each agent has a unique name. Each agent is a peer since he can communicate in a bi-directional way with all other agents. Each agent lives in a container (that provides its runtime) and can migrate within the platform. The Agent Communication Language (ACL) is the main output of FIPA. Common patterns of agent conversations have been formalized into interaction protocols that provide agents with a library of patterns to achieve common tasks (Kravari and Bassiliades, 2015).

JADE includes both the libraries (i.e. the Java classes) required to develop application agents, and the run-time environment that provides the basic services and that must be active on the device before agents can be executed. Each instance of the JADE run-time is called container (since it ‘‘contains’’ agents). Figure 4 draws the architecture of a JADE agent system deployed on a set of heterogeneous computing nodes.

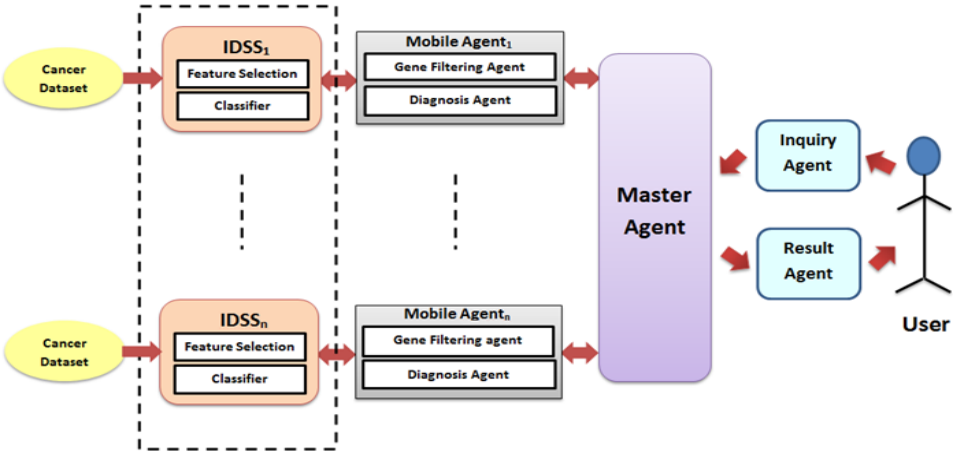
Figure 4: Architecture of JADE Agent System



3.4 Proposed CAD System Operation

As mentioned previously, the proposed CAD system consists of two main parts IDSS and multi-agent system. Figure 5 shows a typical block diagram for the proposed CAD system operation. From Figure 2, the master agent manages whole process of the CAD system. It first accepts an inquiry from the inquiry agent to diagnose a given dataset. Then, the master agent directs mobile agents to initialize several IDSSs (exactly 24 IDSSs) and directs the different IDSSs to run concurrently. Where, each mobile agent builds an IDSS by selecting one the eight feature selection methods and one of the three classification algorithms that described previously in Figure 1. After the IDSS performs the cancer classification of the input dataset, the mobile agent returns the result to the master agent. Finally, the master agent selects the best result from the returned results and sends the classification report to the user through the result agent, based on the best classification accuracy returned from the 24 IDSSs. From Figure 5, the entire process of cancer diseases diagnosis is automated through the distribution of IDSSs to be run concurrently. In the proposed system, gene identification and tumor classification is distributed in several operational agents with different ability and many tasks could be solved in parallel with the knowledge of other experts. Hence, cancer classification is done in distributed computing manner by the operational IDSSs.

Figure 5: The Proposed CAD System Operation



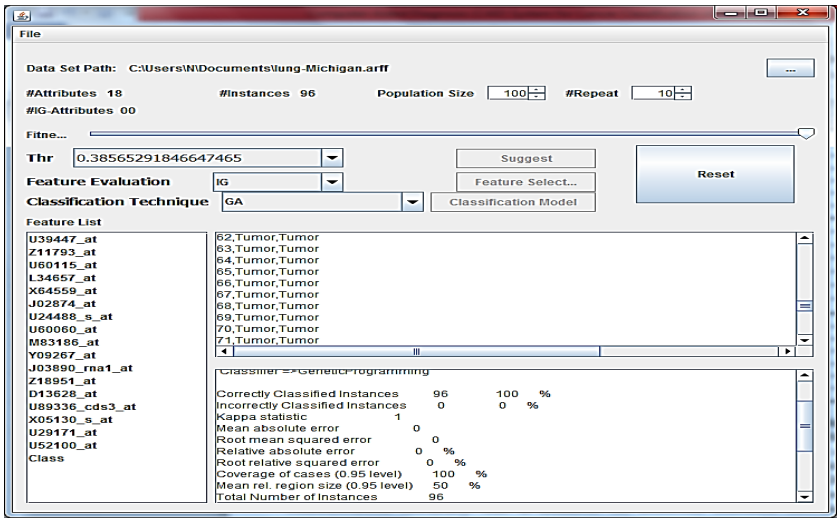
3.4.1 System GUI

The developed CAD system is a fully automated system. However, in the proposed CAD system, every IDSS may be invoked manually or automatically to be run over JADE runtime within a main-container. Figure 6 shows the GUI of the proposed system defined by JADE specification. As shown in Figure 6, in addition to loading data to be analyzed, the developed system composed of three distinct stages; feature selection, classification stage and evaluation stage.

- Feature Selection: A method analysis the major core attributes relative to the dataset and decides dependency score.
- Classification Model: Focuses on building an artificial model to classify the cancer diseases using a classification mechanism.
- Evaluation Model: Evaluates the selected IDSS by using standard performance evaluation measures.

Through the GUI shown in Figure 7, the system may be used individually to analyze a given dataset manually by first selecting an appropriate feature selection method and an appropriate classifier. Then, the system shows the final report after completing the diagnosis process.

Figure 6: GUI of the Proposed System



4. Performance Evaluation

4.1 Dataset

In this study, three gene expression datasets, downloaded from the Kent Ridge Biomedical Dataset website (http://datam.i2r.a-star.edu.sg/datasets/krbd/), are used. Table 1 presents detailed information about these datasets. The datasets categorized into Leukemia, Colon tumor, and, Lung cancer-Michigan.

Table 1: Three Gene Expression Datasets of Human Cancer

Datasets	Diagnostic Task	Number of Samples	Number of Genes	Class Distribution
Leukemia	Acute Myelogenous Leukemia (AML), Acute Lymphoblastic Leukemia (ALL)	72	7129	AML: 25 ALL: 47
Colon tumor	Normal ("Positive"), Tumor ("Negative")	62	2000	Positive: 22 Negative: 40
Lung cancer-Michigan	Non-neoplastic lung samples, Primary lung adenocarcinomas samples	96	7129	Non-neoplastic: 10 Primary lung: 86

4.2 Performance Metrics

Various performance metrics are used. True Positive (TP): the number of positive cases correctly detected. True Negative (TN): the number of negative cases correctly detected. False Positive (FP): the number of negative cases diagnosed as positive. False Negative (FN): the number of positive cases diagnosed as negative. These performance metrics are first computed and then used to compute Classification Accuracy (CA), sensitivity, and specificity of the IDSS according to the following equations.

Sensitivity,True positive rate (TPR),and Recall = ((TP))/(((TP) + (FN)) (15)

Specificity = ((TN))/(((TN)+(FP))) (16)

Classification Accuracy= (TP+TN)/(TP+FP+TN+FN) (17)

Precision = ((TP))/(((TP)+ (FP))) (18)

F=2 . (Precision . Recall)/(Precision+Recall) (19)

5. Experimental Results

To evaluate which of the proposed IDSS models performs better than the others; each IDSS is applied on the three cancer datasets. In each IDSS, one feature selection method and one classifier are used.

Although the results are different between datasets, one of the proposed IDSS provides an optimal feature-classifier combination. Nevertheless, not only one system is the best system for all datasets. IG/GA/GA is the best for leukemia dataset. It carried out an accuracy of 100% with 5 selected features only, as Shown in Figure 7. From Figure 8, GR/GA is the best for colon dataset, where it has a classification accuracy of 90.32 % with a number of feature selections equals to 12, then GR/J48 has the same classification accuracy but a number of selected features is 39. For Lung cancer-Michigan dataset, IG/GA with ṆB classifier reached to the accuracy of classification 100% and a number of selected features is 17, as shown in Figure 9.

Figure 7: Performance Measures of Different Systems for Leukemia Dataset

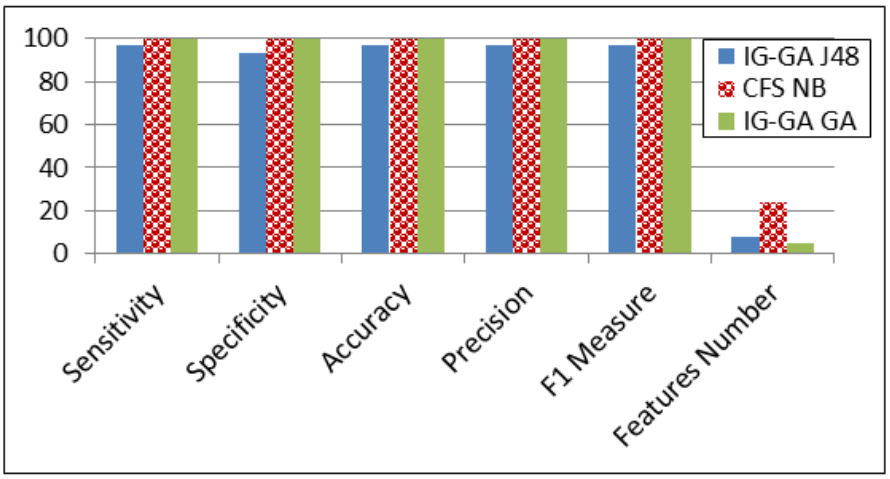


Figure 8: Performance Measures of Different Systems for Colon Dataset

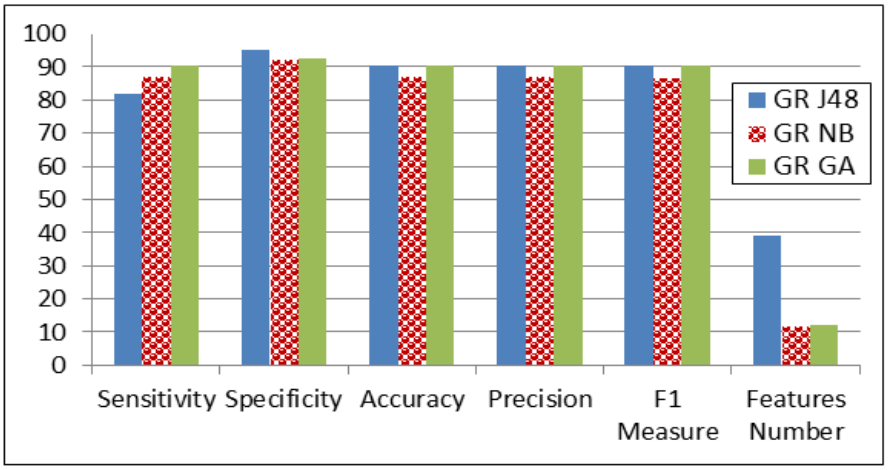


Figure 9: Performance Measures of Different Systems for Lung Dataset

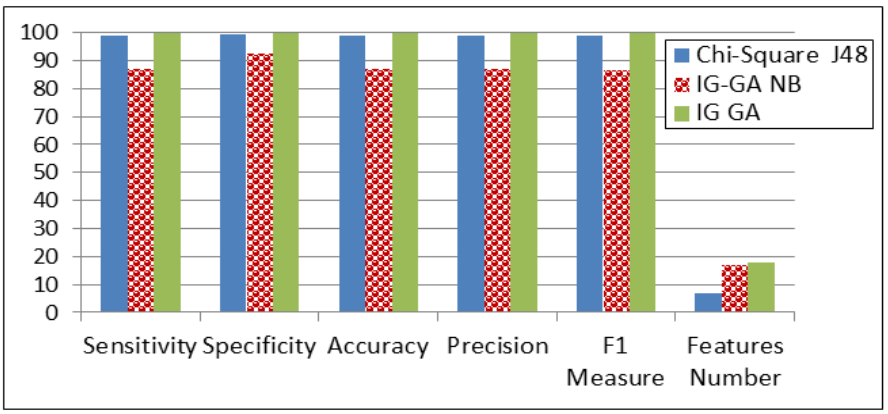


Table 2 presents a comparative analysis between the proposed system and K. J. Danjuma system [25]. By comparing the experimental results, the proposed system improves the classification accuracy, where the accuracy rates close to 100%. The experimental results demonstrate that the proposed procedure can improve the constancy of the feature selection as well as the sample accuracy of classification. The highest AUC values for each dataset with the number of selected genes that gives these highest values are shown in Table 3.

Table 2: 10-fold Cross-Validation Performance Evaluation for Lung Cancer Dataset

Performance Metrics	K. J. Danjuma [25]			Proposed IDSS		
	MLP	J48	NB	J48	NB	GA
Correctly Classified Instances	82.3	81.1	74.4	98.9	100	100
Mean absolute error	21.6	22.7	29.0	1.04	0	0
Root mean squared error	38.6	39.7	44.9	10.21	0	0
Relative absolute error	44.5	46.7	59.7	5.37	0	0
Root relative squared error	78.4	80.6	91.2	33.39	0	0
True Positive (TP) Rate	82.4	81.9	74.5	99	100	100
False Positive (FP) Rate	20.7	20.2	33.2	0.1	0	0
Precision	82.5	81.8	76.8	99.1	100	100
Recall	82.4	81.9	74.5	99	100	100
F-Measure	82.1	81.8	72.7	99	100	100
ROC Area (AUC)	84.7	82.2	79.2	99.4	100	100

Table 3: Highest AUC Values for different classifiers with best values of selected features for different datasets

Microarray Datasets	R. Aziz [26]		Proposed IDSS		
	SVM	NB	J48	NB	GA
Leukemia	0.9468 (35)	0.9536 (30)	0.939 (8)	1 (24)	0.944 (7)
Colon Tumor	0.9126 (30)	0.8566 (25)	0.844 (39)	0.887 (12)	0.894 (12)
Lung Cancer-Michigan	-	-	0.994 (7)	0.998 (14)	1 (18)

6. Conclusion

In this paper, an accurate, fast CAD system is developed for cancer diseases classification by gene expression profiles of DNA microarray dataset. The proposed CAD system is constructed based on Intelligent Decision Support System (IDSS) and Multi-Agent (MA) system. The IDSS combines eight feature selection methods and three evolutionary machine-learning classification methods. While, the MA system is implemented to system manages the entire operation of the CAD system. The MA system invokes 24 different IDSS with the aid of mobile agents and then directs the generated IDSSs to run concurrently to classify the disease on the input dataset before taking a decision. The proposed system is implemented in JAVA, evaluated using three gene expression profile datasets of cancer diseases (Leukemia, colon and Lung cancer-Michigan) and compared with most recent systems. The main benefit of the proposed CAD system is that the system classified the cancer diseases accu-

rately in a very short time. This is because cancer classification is done in parallel processing manner by 24 different IDSSs before taking a decision of the best classification result. In addition, the system is able to maximize the cancer classification accuracy and minimize the number of selected genes against other approaches. In addition, the proposed methodology may be applied on different datasets.

References

D. Raval, D. Bhatt, M. K. Kumhar, and V. Parikh, D. Vyas, "Medical Diagnosis System using Machine Learning", International Journal of Computer Science & Communication, Vol. 7, pp. 177-182, 2015.

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 (JCIT), Vol. 9, No. 2, pp. 125-135, March 2014.

D. Bassen, S. Nayak, X. Chong Li and M. Sam, "Clinical Decision Support System (CDSS) for the Classification of Atypical Cells in Pleural Effusions", Procedia Computer Science, Elsevier, Vol. 20, No. 2, pp. 379–384, 2013.

https://doi.org/10.1016/j.procs.2013.09.290

J. Shi, Q. Su, C. Zhang, G. Huang, and Y. Zhu, "An intelligent decision support algorithm for diagnosis of colorectal cancer through serum tumor markers", Computer methods and Programs in Biomedicine, Elsevier, Vol. (100), pp. 97-107, 2010.

https://doi.org/10.1016/j.cmpb.2010.03.001

P. Bountris, M. Haritou, A. Pouliakis, N. Margari, M. Kyrgiou, A. Spathis, A. Pappas, I. Panayiotides, E. A. Paraskevaidis, P. Karakitsos, and D. D. Koutsouris, "An Intelligent Clinical Decision Support System for Patient-Specific Predictions to Improve Cervical Intraepithelial Neoplasia Detection", BioMed Research International, Hindawi, Vol. 2014, pp. 1-20, 2014.

https://doi.org/10.1155/2014/341483

R. R. Janghel, A. Shukla, R. Tiwari, and R. Kala, "Intelligent Decision Support System for Breast Cancer", Proceedings of the International Conference on Swarm Intelligence, Springer Lecture Notes in Computer Science, pp. 351-358, 2010.

https://doi.org/10.1007/978-3-642-13498-2_46

K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine Learning Applications in Cancer Prognosis and Prediction”, Computational and Structural Biotechnology Journal, Vol. 13, pp. 8-17, 2015.

https://doi.org/10.1016/j.csbj.2014.11.005

B. L. Iantovics, “Agent-Based Medical Diagnosis Systems”, Computing and Informatics, Vol. 27, No. 4 2008.

D. Isern and A. Moreno, “A Systematic Literature Review of Agents Applied in Healthcare”, Journal of Medical Systems, Springer, pp. 40-43, 2015.

https://doi.org/10.1007/s10916-015-0376-2

S. Chakraborty, and S. Gupta, “Medical Application using Multi Agent System - A Literature Survey”, Sougata Chakraborty et al Int. Journal of Engineering Research and Applications, Vol. 4, pp. 528-546, 2014.

H. Salem, G. Attiya, and N. El-Fishawy, “A Survey of Multi-Agent based Intelligent Decision Support System for Medical Classification Problems”, International Journal of Computer Applications, Vol. 123, pp. 20-25, 2015.

https://doi.org/10.5120/ijca2015905529

L. Cervantes, Y. S. Lee, H. Yang, S. h. Ko, and J. Lee, “Agent-Based Intelligent Decision Support for the Home Healthcare Environment”, Chapter in Advances in Hybrid Information Technology, Vol. 4413 of the series Lecture Notes in Computer Science, Springer, pp. 414-424, 2007.

https://doi.org/10.1007/978-3-540-77368-9_41

R. Asadi, N. Mustapha, and N. Sulaiman, “A Framework for Intelligent Multi Agent System Based Neural Network Classification Model”, International Journal of Computer Science and Information Security (IJCSIS),Vol. 5, No. 1, pp. 168-174, 2009.

H. G. Vélez, M. Mier, M. J. Sapé, T. N. Arvanitis, J. M. G. Gómez, M. Robles, P. H. Lewis · S. Dasmahapatra, D. Dupplaw, A. Peet, C. Arús, B. Celda, S. V. Huffel, and M. L. Ariet, “HealthAgents: Distributed Multi-Agent Brain Tumor Diagnosis and Prognosis”, Applied Intelligence, Springer, Vol. 30, pp. 191-202, 2009.

https://doi.org/10.1007/s10489-007-0085-8

B. Fatima and C. M. Amine, “A Neuro-Fuzzy Inference Model for Breast Cancer Recognition”, International Journal of Computer Science & Information Technology (IJCSIT), Vol. 4, No 5, Pages163-173, October 2012.

https://doi.org/10.5121/ijcsit.2012.4513

M. Khashei, A. Z. Hamadani and M. Bijari, “A Fuzzy Intelligent Approach to the Classification Problem in Gene Expression Data Analysis”, Knowledge-Based Systems, Elsevier, Vol. 27, Pages 465–474, 2012.

https://doi.org/10.1016/j.knosys.2011.10.012

G. Dimitoglou, J. A. Adams, and C. M. Jim,” Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability”, Journal of Computing, Vol. 4, Pages 1-9, 2012.

H. Yu, J. Ni, Y. Dan, and S. Xu,” Mining and Integrating Reliable Decision Rules for Imbalanced Cancer Gene Expression Data Sets”, Tsinghua Science and Technology, Vol. 17, Pages 666-673, December 2012.

https://doi.org/10.1109/TST.2012.6374368

C. J. A. González, Q. I. M. Sancho, A. S. Hurtado, R. V. Arrabal,” Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods”, Expert Systems with Applications, Vol. 39, Pages 7270–7280, 2012.

https://doi.org/10.1016/j.eswa.2012.01.096

H. Hijazi and C. Chan,” A Classification Framework Applied to Cancer Gene Expression Profiles”, Journal of Healthcare Engineering, Vol. 4, Pages 1-30, 2013.

https://doi.org/10.1260/2040-2295.4.2.255

X. Wang, ”Identification of Marker Genes for Cancer Based on Microarrays Using a Computational Biology Approach”, Current Bioinformatics, National Institutes of Health, Vol. 9, pages 1-10, 2014.

https://doi.org/10.2174/1574893608999140109115649

C. Lu, Z. Zhu, and X. Gu,”An Intelligent System for Lung Cancer Diagnosis Using a New Genetic Algorithm Based Feature Selection Method”, Journal of Medical Systems, Vol. 38, pages pages 1-9, 2014.

https://doi.org/10.1007/s10916-014-0097-y

H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, ”The Performance of Bio-Inspired Evolutionary Gene Selection Methods for Cancer Classification Using Microarray Dataset”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 4, No. 3, Pages 166-170, May 2015.

https://doi.org/10.7763/IJBBB.2014.V4.332

K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, ”Machine learning applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal, Vol. 13, Pages 8-17, 2015.

https://doi.org/10.1016/j.csbj.2014.11.005

K. J. Danjuma, “Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients”, International Journal of Computer Science Issues (IJCSI), Volume 12, Issue 2, March 2015.

[26] R. Aziz, C. K. Verma, and N. Srivastava, “A Fuzzy based Feature Selection from Independent Component Subspace for Machine Learning Classification of Microarray Data” Genomics Data, Elsevier, Vol. 8, pp. 4-15, 2016.

https://doi.org/10.1016/j.gdata.2016.02.012

Hanaa Salem, Gamal Attiya, and Nawal El-Fishawy, “Early Diagnosis of Breast Cancer by Gene Expression Profiles,” Pattern Analysis and Applications, pp.1-12, 2016. Doi:10.1007/s10044-016-0574-7 Online: 1 September 2016.

https://doi.org/10.1007/s10044-016-0574-7

Hanaa Salem, Gamal Attiya, and Nawal El-Fishawy, “Classification of Human Cancer Diseases by Gene Expression Profiles,” Applied Soft Computing Journal, Vol. 50, No. 1, pp.124-134, 2017.

https://doi.org/10.1016/j.asoc.2016.11.026

I. Koprinska, M. Rana , V. G. Agelidis, “Correlation and Instance-Based Feature Selection for Electricity Load Forecasting”, Knowledge-Based Systems, Elsevier, Vol. 82, pp. 29-40 , 2015.

https://doi.org/10.1016/j.knosys.2015.02.017

J. Xu and H. Jiang, “An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier”, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), IEEE, pp. 273-276, 2015.

M. W. Mwadulo, “A Review on Feature Selection Methods for Classification Tasks”, International Journal of Computer Applications Technology and Research, Vol. 5, Issue 6, pp. 395 - 402, 2016.

https://doi.org/10.7753/IJCATR0506.1013

D. Dittman, T. Khoshgoftaar, R. Wald and A. Napolitano, “Similarity analysis of feature ranking techniques on imbalanced DNA microarray datasets”, International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1-5, 2012.

https://doi.org/10.1109/BIBM.2012.6392708

V. B. Canedo, N. S. Maroño, and A. A. Betanzos, “Feature selection for high-dimensional data”, Progress in Artificial Intelligence, Vol. 5, Issue 2, pp. 65–75, 2016.

https://doi.org/10.1007/s13748-015-0080-y

L. Zhang and X. Huang, “Multiple SVM-RFE for multi-class gene selection on DNA Microarray data”, International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1-6, 2015.

https://doi.org/10.1109/IJCNN.2015.7280417

Hanaa Salem, Gamal Attiya, and Nawal El-Fishawy, “An Optimization Approaches Using PSO/GA/GA and IG/GA/GA Hybrid Algorithms based on Gene Cancer Classification”, Proceedings of the 1st International Conference on Advanced Technology and Applied Sciences (ICaTAS2016), Malaysia, 6-7 September 2016.

X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.F. McLachlan,A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, “Top 10 algorithms in data mining”, Knowledge Information System, Vol. 14, pp. 1–37, 2008.

https://doi.org/10.1007/s10115-007-0114-2

N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria, “Decision Tree Analysis on J48 Algorithm for Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 6, pp. 114-119, June 2013.

T. R. Patil and S. S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification”, International Journal of Computer Science And Applications, Vol. 6, No.2, pp. 256-262, Apr 2013.

G. Dimitoglou, James, A. Adams and C. M. Jim, “Comparison of the C4.5 and a Naïve Bayes Classifier for the Prediction of Lung Cancer Survivability”, Journal of Computing, Vol. 4, Issue 8, pp. 3-12, 2012.

I. Rish, “An empirical study of the naive Bayes classifier”, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Vol. 22, pp. 41-46, 2001.

F. Bellifemine, G. Caire, A. Poggi, and G. Rimassa, “JADE: A Software Framework for Geveloping Multi-Agent Applications”, Information and Software Technology, Elsevier, Vol. 50, pp. 10-21, 2008.

https://doi.org/10.1016/j.infsof.2007.10.008

K. Kravari and N. Bassiliades, “A Survey of Agent Platforms”, Journal of Artificial Societies and Social Simulation, Vol. 18, 2015.

https://doi.org/10.18564/jasss.2661

http://datam.i2r.a-star.edu.sg/datasets/krbd/.