INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at www.ijariit.com

An Optimized Classification of Human Cancer Disease for Gene Expression Data

K. Yamunadevi
yamunadevi.mail@gmail.com
Kaamadhenu Arts and Science College, Sathayamangalam, Tamil Nadu

R. Nagaraj
nagukasc@gmail.com
Kaamadhenu Arts and Science College, Sathaymangalam, Tamil Nadu

ABSTRACT

Classification of cancer is determined the appropriate treatment and helps to determine the prognosis. The Existing System was presented to classify the human cancer diseases predicted on the gene expression profiles. The existing approach is initially, the Information Gain (IG) is utilized for feature selection and Genetic Algorithm (GA) is used for feature reduction. Finally, the Genetic program is utilized for classifying the types of human cancer. However, the existing approach has high computation time and required a large amount of computational resource. To overcome this issue in this paper presented the Cuckoo Search (CS) optimization algorithm to optimize the threshold value of the features determined by the Information Gain (IG) and then Genetic programming (GP) is used for enhancing the performance of classifying human cancer. The proposed cuckoo search approach is nature inspired behavior and breeding process of cuckoo bird's optimization algorithm for generation of the global code book with one tuning parameter and it's applicable for both linear and nonlinear problems. The performance of the proposed approach is evaluated in terms of Classification Accuracy, Specificity, and Sensitivity.

Keywords: *Datamining, Cancer, Information Gain (IG), Optimization Algorithm.*

1. INTRODUCTION

Data mining [1] is used a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and patterns. These trends and patterns have formed the basis of predictive models that enable analysts to produce new observations from existing data. Gartner Inc.'s definition of data mining is the most comprehensive: the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, and by using pattern recognition technologies, as well as statistical and mathematical techniques.

Data mining [2] should be performed on very large or raw datasets using either supervised or unsupervised data mining algorithms. Note that data mining cannot occur without direct interaction with unitary data. The most successful data mining projects comply with the guidelines and steps in the Cross-Industry Standard Process for Data Mining (CRISP-DM). As the demand for data mining is increased and more algorithms are created, CRISP-DM ensures good practices that everyone can follow.

2. EXISTING METHODOLOGY

The Existing methodology [3] is presented to classify the human cancer diseases predicted on the gene expression profiles. The existing approach is initially, the Information Gain (IG) is utilized for feature selection and Genetic Algorithm is utilized for feature reduction. Finally, the Genetic program is utilized for classifying the types of human cancer. However, the existing approach has high computation time and required a large amount of computational resource. So in this paper presented the Cuckoo Search (CS) optimization algorithm is presented to optimize the threshold value of the features determined by the Information Gain (IG) and then Genetic programming (GP) is used for enhancing the performance of classifying human cancer.

The remainder of the section is described as following way. Section 3 provides the proposed methods and details of the proposed work. Section 4 describes the background study on methods related to various data mining methods to detect human cancer disease. Section 5 provides the experimental evaluation on public benchmarks and the corresponding critical discussion. Section 6 deals with the conclusion of the thesis. Section 7 deals with the scope for future work.

3. PROPOSED METHODOLOGY

In the proposed system, presented the Cuckoo Search (CS) optimization algorithm is presented to optimize the threshold value of the features determined by the Information Gain (IG) and then Genetic programming (GP) is used for enhancing the performance of classifying human cancer. The proposed cuckoo search approach is nature inspired behavior and breeding process of cuckoo bird's optimization algorithm for generation of the global code book with one tuning parameter and it's applicable for both linear and nonlinear problems.

3.1. Feature Selection by Information Gain (IG)

Feature selection is a preprocessing procedure expecting to select the most informative genes that can separate groups, i.e., cancer subtypes. The essential reason is to discover a reduced group band of features from a dataset to diminish the initial feature space dimensionality. Generally, cancers classification studies require the use of formal strategies of feature selection for just two explanations:

- To lower the computational requirements in experimental responsibilities, this helps the evaluation and exploration of data in these domains;
- To deliver controllable solutions in conditions of amounts and structure of variables (features), this helps medical interpretation.

Information Gain (IG) has been accounted to be the prevalent gene selection procedure. Univariate filter approaches have been broadly used in microarray data analysis. This pattern can be clarified by a number of reasons; the outcome gave by univariate gene rankings are natural and easy to understand. These simplified versions of output could satisfy the points and desires of biology, science and molecular-domain experts who interest for acceptance of results utilizing research laboratory procedures. In addition, filter approaches also offer less computational time to produce results which are an additional point to be favored by domain experts. In spite of the fact that we eliminate one high-ranked gene, it may not create any degradation of classification accuracy.

The impression behind IG is to choose features that reveal the most data about the classes. Superbly, such features are especially discriminative and happen in a single class. Information gain is a measure in view of entropy; it shows to what degree the whole entropy is decreased on the off chance that we know the estimation of a particular attribute. Accordingly, the IG value demonstrates the measure of information this attribute gives to the data set. IG value calculates for each feature, which decides whether this feature is to be chosen, or not. For checking the features so threshold value is utilized; if a feature has a greater IG value than the threshold, the feature is picked; else, it is not picked.

Let S be the set of n instances and C be the set of k classes. P(Ci, S) denotes the fraction of the instances in S that has class Ci. Then, the estimated information from this class membership is given by: A greater information gain will result in a greater probability of obtaining pure classes in a target class.

3.2. Feature Reduction by Cuckoo Search (CS) Optimization Algorithm

The proposed technique employs CS to reduce the features are determined by the IG.
The cuckoo search (CS) algorithm is nature inspired behavior and breeding process of cuckoo bird's optimization algorithm for generation of the global code book with one tuning parameter and it's applicable for both linear and nonlinear problems.

Cuckoo birds emits beautiful sounds, lay their eggs in the nests of host birds. If the host bird is recognized those eggs are not of the own it throws them away or abandons the nest and searches for a nest at any new location. Non-parasitic cuckoos, like most other non-passerines, lay white eggs, but many of the parasitic species lay colored eggs to match those of their passerine hosts. In some cases, female cuckoo can mimic the color and pattern of eggs of some selected host nets. The features are minimized the probability of eggs being thrown away from the nest and cause an increment in productivity of cuckoos further. Non-parasitic cuckoos leave the nest before they can fly, and some new world species have the shortest incubation periods among birds. The cuckoo breeding process is based on the current position of cuckoo and probability of better next position after a selected random walk with a number of chosen random step sizes.

This random walk plays a major role in the exploration, exploitation, intensification, and diversification of the breeding process. In general, this foraging of random walk and step size follows a probability distribution function. The probability distribution functions like Gaussian distribution, normal distribution, and Levy distribution. In cuckoo search, the random walk follows levy flight and step size follows levy distribution function. Levy flight is a random walk whose step follows the levy distribution function. In huge search space levy flight, random walk is better than Brownian walk because of its nonlinear sharp variation of parameters. The direction of the walk follows the uniform distribution function and steps of walk follow Mantegna's algorithm which gives both positive and negative numbers.
The Levy distribution function is

$$L(s,\gamma,\mu)=\begin{cases}\frac{\gamma}{2\pi}exp\left[-\frac{\gamma}{2(s-\mu)}\right]\frac{1}{(s-\mu)^{\frac{3}{2}}} & 0<\mu<s\infty \\ 0 & otherwise\end{cases} \quad (1)$$

Where the $\mu > 0$ is the minimum step and γ is the scale parameter. If the $s\to\infty$ then the Equation (1) becomes

$$L(s,\gamma,\mu)\approx\sqrt{\frac{\gamma}{2\pi}}\frac{1}{(s)^{\frac{3}{2}}} \quad (2)$$

In the cuckoo search, the generation of random walk steps is to depend on mostly Mantegna's algorithm. Depends on the Mantegna's algorithm the step size of random walk of the cuckoo is given by Equation (3)

$$\text{Step of random walk} = \frac{\mu}{(v)^{\frac{1}{\beta}}} \quad (3)$$

The μ and v are drawn from normal distribution or Gaussian distribution is given in equation (4) with β=2

$$L(s) = \frac{1}{\pi}\int_0^{\infty} \cos(\tau s)\, e^{-\tau\alpha^{\beta}} d\tau \quad (4)$$

From above equation

$$\mu \approx N(0, \sigma_{\mu}^2) \qquad v \approx N(0, \sigma_v^2) \quad (5)$$

Where the N () normal distribution function is given by Equation (6)

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad -\infty < x < \infty \quad (6)$$

Where

$$\sigma\mu = \left\{\frac{\mathbb{\Gamma}(1+\beta)sin\left(\frac{\pi\beta}{2}\right)}{\mathbb{\Gamma}\left[\frac{1+\beta}{2}\right]\beta 2^{\frac{(\beta-1)}{2}}}\right\}^{\frac{1}{\beta}} \text{ and } \sigma_v = 1 \quad (7)$$

Then the gamma equation (Γ) is given in equation (8)

$$\mathbb{\Gamma}(\beta) = \int_0^{\infty} e^{-t} t^{\beta-1} dt \quad (8)$$

The CS algorithm is optimized the threshold value of the features determined by the IG to enhance the performance of the classification of human cancer. The cuckoo search algorithm works with the following idealized rules. Each cuckoo lays one egg at a time and dumps it in a randomly chosen nest. The best nest with the high quality of eggs will carry over to the next generations. The number of available host nests is fixed and a host can discover an alien egg with the probability $P_a \in [0,1]$. In this case the host bird can either throw the egg away or abandon the nest so as to build a new nest in the new location.

3.3 Classify the Cancer by Genetic Programming

In the proposed approach, cancer type classification is done by means of CS. In GP, a classification tree denotes classifier. It includes symbols from the function group (F) and the reduction group (R). The function group (F) embraces of arithmetic operators and the reduction group (R) embraces of a number of feature constants and variables characterized as takes after: (F) ={+,−, *,/}and(R) ={0. . . a number of genes, y1. . . yn}. The variables have denoted the significance of the expression degree of genes. The expression is assessed to compute the fitness of a candidate. From the microarray dataset, the features (y1. . . yn) are data. Each expression is evaluated. If evaluating an expression results in more than 0, it is classified as Class 1; otherwise, it is classified as Class 2. To compute the fitness value of the expression, the whole number of the accurate classification is counted. Enhanced classification accuracy is attained by the bigger fitness value.

3.4 Performance Evaluation

This section is presented the performance evaluation of the proposed IG/CS methodology. The proposed approach is evaluated in terms of classification accuracy, Specificity, and Sensitivity.

4. LITERATURE SURVEY

In this paper presented [4] Knowledge-based system for classification of Cancer for classification of breast cancer by using the clustering, noise removal, and classification techniques. In this scheme, the expectation maximization (EM) is used as the clustering method for clustering the data in the related groups. Then the classification and Regression trees are utilized to generate the fuzzy rules for classification of the breast cancer disease in the present knowledge-based system of fuzzy rule techniques. In order to overcome the multi collinearity issue include Principal component Analysts (PCA) in the present technique. The present technique is tested with Wisconsin Diagnostic Breast cancer and Mammographic mass datasets and the result shows that improve the prediction accuracy of breast cancer. However, the proposed approach does not perform well in large data sets.

In this paper presented [5] the Feature-Based cancer classification which is used as the single layer Artificial Neural network for classification of a cancer patient. In order to achieve the various feature reduction schemes the set of simple classifier utilized in this paper. Initially, the Principal component analysis (PCA), Factor Analysis (FA) and Discrete Fourier Transform (DFT) is utilized to reducing the dimension, after that these reduction dimensions are used to build the intelligent classifiers by using the various functional link Artificial Neural Network (FLANN). Since this approach has complexity issue.

In this paper presented [6] the Hybrid Approach for Cancer Classification to overcome the gene selection issue to the classification of cancer. Initially the MRMR is utilized in the pre-processing stage to select the top 100 genes, then the selected genes are fed in to the wrapper set up which contains the COA-HS algorithm and the SVM classification technique is used in present technique it is provided the high accuracy then finally classification performance of the selected genes are measured in terms of accuracy. However, this approach has convergence issue.

In this paper proposed [7] the Ensemble Gene selection method (EGS) to choose the multiple gene subsets for classification purpose. In this technique, the genes are chosen based on the conditional mutual information. The result shows that the present gene subset has the good discriminative capability for data classification. Additionally, the number of selected genes of the present techniques also finds out self adaptively. In order to increase the diversity of the present technique the initial points are allocated to various genes with highest information. If the multiple gene subsets have been obtained the present technique provides the train base classifiers and then the result is integrated by the majority voting strategy. The result shows that the present technique is outperform than the traditional techniques. Since the proposed approach had high computational cost.

In this paper presented [8] the application of pattern recognition and image processing techniques to examine the significance of the feature extraction from the fine needle aspiration biopsy images and the cause of the reducing large number features by utilizing the

various feature selection methods with a small loss of the classification accuracy. The present technique is used to reduce the size of the feature vector size then perform the classification with less information and deliver the satisfied results. The reduction of the feature vector leads to reduce the complexity of classification. The best classification feed forward to the neural network when the correlation measures are utilized. The present technique with correlation feature vector reach good performance and able classify the breast cancer data with high accuracy while the feature vector size reduced iteratively. Since the present approach does not perform well in high dimensional feature set.

In this paper proposed [9] the novel approach for classification for predicting cancer through gene expression profiles which are build with supervised learning hidden markov models (HMMs). The gene expression of each tumour is designed by HMM and the prominent discriminant genes are chosen by the present techniques are depends on the modification of the analytic hierarchy process (AHP). The modified AHP is allowed quantitative factors which are used to rank the outcomes of the individual gene selection methods such as t-test, entropy, receiver operating characteristic curve, Wilcoxon test, and signal to noise ratio. The result shows that the HMM is the powerful tool for cancer classification better than the classical classification techniques. The combinations of AHP-HMM is provided the better stability and robustness to the selection of gene and improve early detection, ease of use to the treatment of cancers in effective and efficient manner. However, the proposed approach has a high mortality rate.

In this paper presented [10] the random projection (RP) technique for reducing the high dimensional features into low dimensional space with the short duration to predict the classification of cancer disease. In order to improve the accuracy of the random projection technique, it's combining with other techniques such as Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Feature Selection (FS). The different combinations of the methods are tested with the microarray dataset. The result shows that the feature selection with random projection improves the classification accuracy better than the PCA and LDA. Since the proposed approach is required more observation.

In this paper presented [11] the feature subset based ensemble method to classify the multiple cancers by utilizing the miRNA expression data in order to generate the multiple subsets the feature relevance and redundancy considers. The present techniques are utilized the C4.5 decision tree algorithm and SVM algorithm for classification. The present techniques are tested with the sequence based miRNA expression datasets and validated with the 10 fold and leave one out cross validations. The result shows that the present technique reaches higher prediction accuracy than the traditional ensemble technique. However, the proposed approach is not provided probabilistic framework.

In this paper proposed [12] a simple yet very effective method for cancer classification using the very few gene expression. The aim of the present technique is the finding the smallest gene subsets for accurate cancer classification from micro array data by using supervised machine learning algorithms (SVM). The present techniques are involved in two-phase such as chosen some important genes by using the 2-way Analysis of Variance (ANOVA) ranking scheme, then test with the SVM classifier it provides the good accuracy. Since the proposed approach has a high mortality rate.

In this paper presented [13] the analysis of feed forward neural network and the island differential evolution propagation algorithm is utilized to train this network. The aim of the present technique is built the effective tool for constructing the neural models which help to the proper classification of different classes of breast cancer. The present techniques are proposed two different migration topologies such as random topology and torus topology. The performances are tested with Wisconsin Breast Cancer Diagnosis problem and the result shows that the random topology is provided good classification accuracy compare to torus topology. Since the proposed approach is required more observation.

In this paper proposed [14] the novel Gauss Newton Representation based Algorithm (GNRBA) for classification of breast cancer. It uses the sparse representation with feature selection and evaluates the sparsity in a computationally efficient way. Then the present technique is proposed new gauss Newton based classifier to find optimal weights for training samples for classification. The present techniques are tested with Wisconsin breast cancer database and Wisconsin Diagnosis breast cancer database from the UCI machine learning repository. The result shows that the present technique provides better accuracy, sensitivity, specificity, confusion matrices compare to traditional approaches. This approach is very sensitive to outliers.

In this paper presented [15] the analysis of categorizing and automated classification of breast cancer by using the multi scale basic image features (BIF) and Local Binary Patterns (LBP) combined with the random decision trees classifier is used for the classification of breast cancer. The present techniques are demonstrated the text-based classification of Hematoxylin and Eosin (H&E) images from IBC. The result shows that the multi-scale approach provides the good accuracy. Still this is required more training.

In this paper presented [16] the Machine Learning (ML) approaches for predicting cancer. The various predictive models are discussed based on ML techniques as well as various input features and data samples. The ML is the branch of artificial intelligence which is used to relate the problem of learning from the data samples in the concept of inference. Each learning process contains two phases. (i) Estimation of unknown dependencies in a system from the given dataset. (ii) Then the usage of the estimated dependencies to prediction the new outputs of the system. In this work, the two main methods used such as supervised learning and unsupervised learning. However, this approach has complexity is an early prediction.

In this paper presented [17] the various data mining techniques in the several types of lung cancer datasets to enhance the lung cancer diagnosis. In this technique, the most effective model to predict patients with lung cancer appears to be the Naive bayes which are used to follow the IF-THEN rules, decision trees, and neural networks. The decision tree result is easier to read and interpret. The present techniques for predicting lung cancer can be further enhanced and expanded. Since this approach has complexity issue.

In this paper presented [18] the several data mining techniques to diagnosis and prediction of breast cancer. The prediction of the outcome of the disease is the one of the complex tasks to enhance the data mining applications. The usage of the computers with automated tools, the large volumes of the medical data are gathered and available within the medical research groups. The data mining techniques are a popular research tool for a medical researcher to predictions of the exploit patterns and related with a large number of variables which is used to improve the prediction of disease using the historical datasets. The several data mining techniques such as Decision trees, Digital Mammography classification using association rule mining and ANN, Association rule-based classifier, neural network based classifier system, Naive bayes classifier, support vector machine, logistic regression and Bayesian network. The result shows that the Bayesian network is performing well to predict out Breast Cancer and diagnosis. However, the Bayesian networks require large amount of probability data.

5. RESULT AND DISCUSSION

The performance of proposed approach IG/CS is evaluated in terms of Accuracy, Precision, Recall, F-Measure, and Specificity. The experimental result shows that the proposed approach IG/CS is achieved a better result than existing IG/GA approach.

5.1 Data Set Descriptions

5.1.1 Leukemia Dataset

Leukemias are primary disorders of bone marrow. They are malignant neoplasms ofhematopoietic stem cells. The total number of genes to be tested is 7129, and a number of samples to be tested is 72, which are all acute leukemia patients, either acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML).

5.1.2 DLBCL Dataset

Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas (FL) are two B-cell lineage malignancies that have very different clinical presentations, natural histories and response to therapy. However, FLs frequently evolve over time and acquire the morphologic and clinical features of DLBCLs and some subsets of DLBCLs have chromosomal translocations characteristic of FLs. The gene-expression based classification model was built to distinguish between these two lymphomas.

5.1.3 Lungcancer Dataset

It is a disease in which certain lung cells don't function right, divide very fast, and produce too much tissue forming a lung tumor. There are 181 tissue samples among which 31 samples belong to MPM and 150 belong to ADCA. Each sample is described by 12533 genes. “Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma”

5.2 Accuracy

The accuracy is the defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy can be calculated from the formula given as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

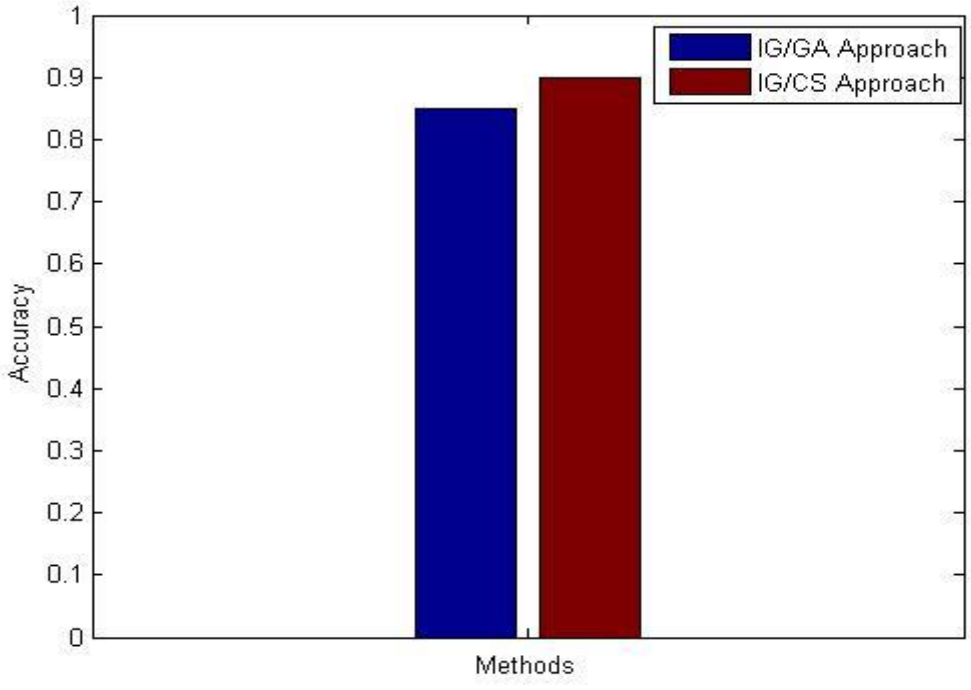

Figure 4.1 Comparison of Accuracy

Figure 4.1 shows that the comparison of Accuracy values of Proposed Information Gain and Cuckoo Search optimization Approach (IG/CS) and Existing Information Gain and Genetic Algorithm (IG/GA). The result shows that the Accuracy value of the proposed IG/CS system is high compared to existing IG/GA approach. For example, an accuracy value of the proposed IG/CS is 0.90 which is higher than the accuracy value of IG/GA approach.

5.3 Precision

Precision value is evaluated according to the feature classification at true positive prediction; false positive.It is expressed as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

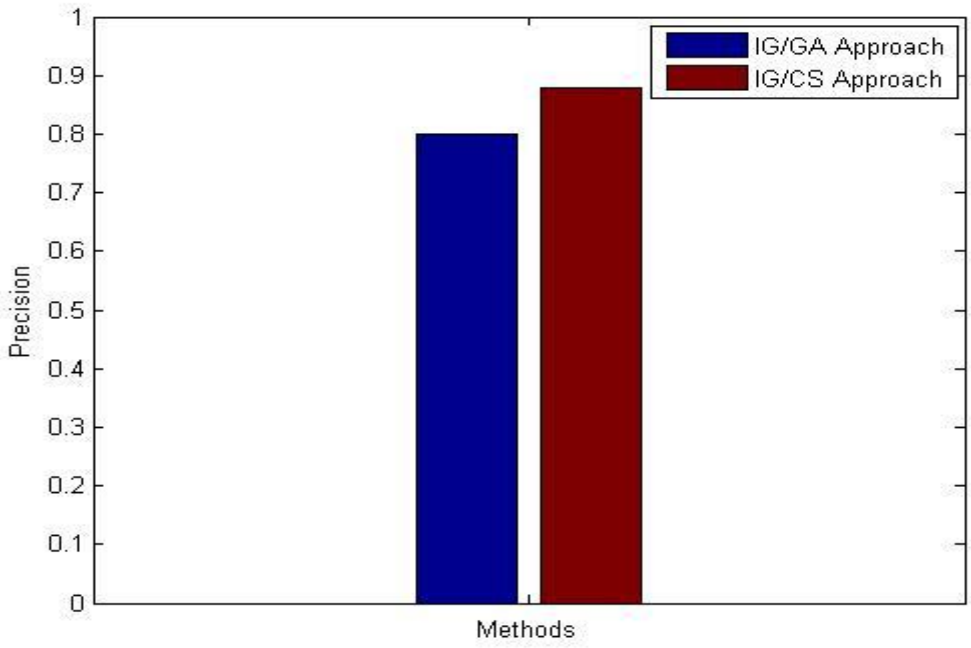

Figure 4.2 Comparison of Precision

Figure 4.2 shows that the comparison of Precision values of Proposed Information Gain and Cuckoo Search optimization Approach (IG/CS) and Existing Information Gain and Genetic Algorithm (IG/GA). The result shows that the Precision value of the proposed IG/CS system is high compared to existing IG/GA approach. For example, the Precision value of the proposed IG/CS is 0.88 which is higher than the Precision value of IG/GA approach.

5.4 Recall

Recall value is evaluated according to the feature classification at true positive prediction, false negative. It is given as,

$$\text{Recall} = \frac{\text{Truepositive}}{(\text{Truepositive} + \text{Falsenegative})}$$

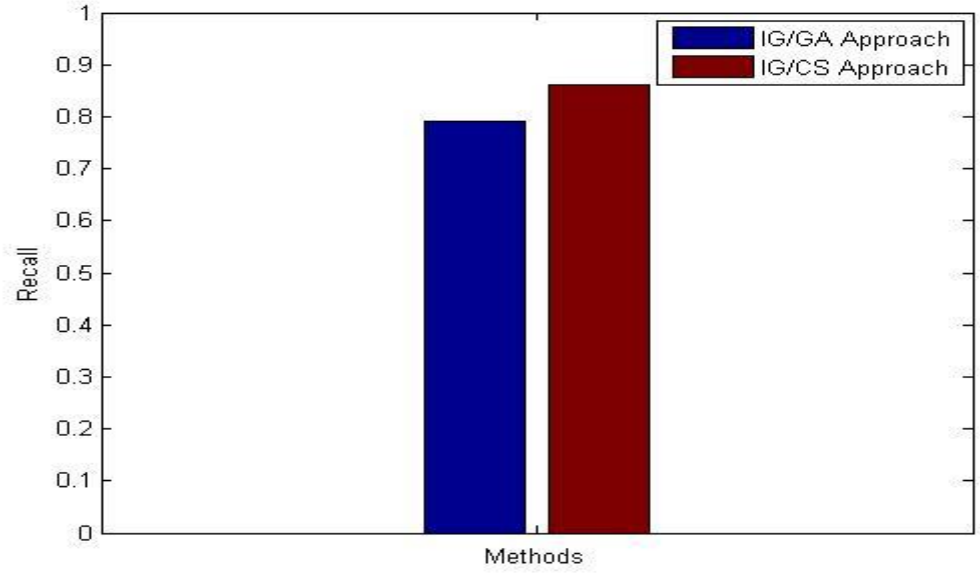

Figure 4.3 Comparison of Recall

Figure 4.3 shows that the comparison of Recall values of Proposed Information Gain and Cuckoo Search optimization Approach (IG/CS) and Existing Information Gain and Genetic Algorithm (IG/GA). The result shows that the Recall value of the proposed IG/CS system is high compared to existing IG/GA approach. For example, Recall value of the proposed IG/CS is 0.86 which is higher than the Recall value of IG/GA approach.

5.5 F-Measure

F-measure is calculated from the precision and recall value. It is calculated as:

$$\text{f} - \text{measure} = 2 \times \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)$$

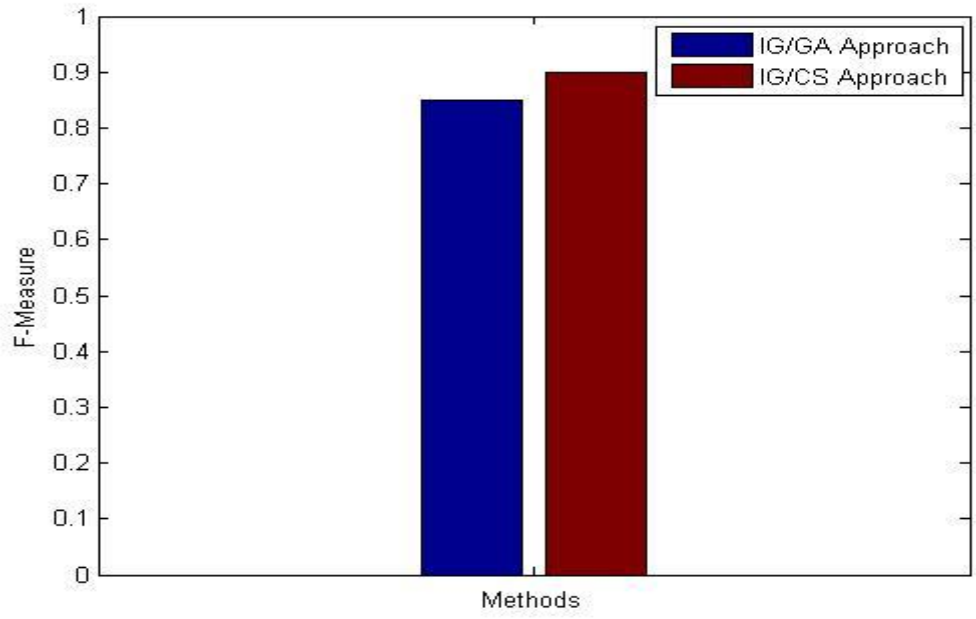

Figure 4.4 Comparison of F-Measure

Figure 6.4 shows that the comparison of F-Measures values of Proposed Information Gain and Cuckoo Search optimization Approach (IG/CS) and Existing Information Gain and Genetic Algorithm (IG/GA). The result shows that the F-Measure value of the proposed IG/CS system is high compared to existing IG/GA approach. For example, F-Measure value of the proposed IG/CS is 0.90 which is higher than the F-Measure value of IG/GA approach.

5.6 Specificity

Specificity relates to the test's ability to correctly detect patients without a condition. Consider the example of a medical test for diagnosing a disease. Specificity of a test is the proportion of healthy patients known not to have the disease, who will test negative for it. Mathematically, this can also be written as:

$$\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives+Number of false positives}}$$

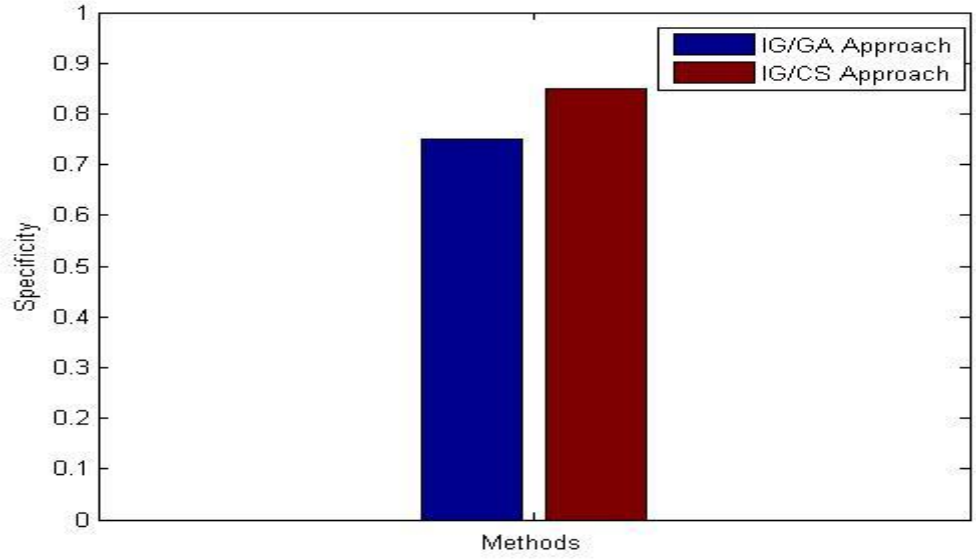

Figure 4.5 Comparison of Specificity

Figure 4.5 shows that the comparison of Specificity values of Proposed Information Gain and Cuckoo Search optimization Approach (IG/CS) and Existing Information Gain and Genetic Algorithm (IG/GA). The result shows that the Specificity value of the proposed IG/CS system is high compared to existing IG/GA approach. For example, Specificity value of the proposed IG/CS is 0.85 which is higher than the Precision value of IG/GA approach.

Table 4.1 Comparisons of Accuracy, Precision, and Recall, F-Measure and Specificity,

Metrics	IG/GA Approach	IG/CS Approach
Accuracy	0.85	0.90
Precision	0.80	0.88
Recall	0.79	0.86
F-Measure	0.85	0.90
Specificity	0.75	0.85

6. CONCLUSION

Classification of cancer predicated on gene expression data is an encouraging research area in the field of data mining. The suggested algorithm tended to the issue of early diagnosis cancer any particular one of the world's most genuine health issues. In this paper, a new methodology is presented to classify human cancers diseases predicated on the gene expression profiles. In this proposed scheme, initially the IG is utilized for feature selection and CS is used for feature reduction and finally, the GP is used to classifying the cancer disease. The CS approach is used to optimize the threshold value of the features determined by the IG. The experimental result shows that the proposed IG/CS approach is outperformed than existing approach IG/GP approach in terms of Classification Accuracy, Specificity, and Sensitivity.

7. SCOPE FOR FUTURE WORK

In the future work is integrated various kinds of genomic data (interaction between gene expression profile and protein-protein dataset) to develop and improved the classification accuracy and reliability when contrasted with utilizing gene expression alone. Also, it can be extended to perform classification on high dimensional data also.

8. REFERENCE

[1] Bharati, M., & Ramageri, M. (2010). Data mining techniques and applications.
[2] Antonie, M. L., Zaiane, O. R., & Coman, A. (2001, August). Application of data mining techniques for medical image classification. In Proceedings of the Second International Conference on Multimedia Data Mining (pp. 94-101). Springer-Verlag.
[3] Salem, H., Attiya, G., & El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, 124-134.
[4] Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133-144.
[5] Mahapatra, R., Majhi, B., & Rout, M. (2012). Reduced feature based efficient cancer classification using single layer neural network. Procedia Technology, 6, 180-187.
[6] Elyasigomari, V., Lee, D. A., Screen, H. R., & Shaheed, M. H. (2017). 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, 67, 11-20.
[7] Liu, H., Liu, L., & Zhang, H. (2010). Ensemble gene selection for cancer classification. Pattern Recognition, 43(8), 2763-2772.
[8] Jeleń, Ł., Krzyżak, A., Fevens, T., & Jeleń, M. (2016). Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies. Computers in biology and medicine, 79, 80-91.
[9] Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hidden Markov models for cancer classification using gene expression profiles. Information Sciences, 316, 293-307.
[10] Xie, H., Li, J., Zhang, Q., & Wang, Y. (2016). Comparison among dimensionality reduction techniques based on Random Projection for cancer classification. Computational biology and chemistry, 65, 165-172.
[11] Piao, Y., Piao, M., & Ryu, K. H. (2017). Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. Computers in biology and medicine, 80, 39-44.
[12] Bharathi, A., & Natarajan, A. M. (2010). Cancer Classification of Bioinformatics data using ANOVA. International journal of computer theory and engineering, 2(3), 369.
[13] Thein, H. T. T., & Tun, K. M. M. (2015). An approach for breast cancer diagnosis classification using a neural network. Advanced Computing, 6(1), 1.
[14] Dora, L., Agrawal, S., Panda, R., & Abraham, A. (2017). Optimal breast cancer classification using Gauss–Newton representation based algorithm. EXPERT SYSTEMS WITH APPLICATIONS, 85(1), 134-145.
[15] Reis, S., Gazinska, P., Hipwell, J., Mertzanidou, T., Naidoo, K., Williams, N., & Hawkes, D. J. (2017). Automated Classification of Breast Cancer Stroma Maturity from Histological Images. IEEE Transactions on Biomedical Engineering.
[16] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.
[17] Krishnaiah, V., Narsimha, D. G., & Chandra, D. N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies, 4(1), 39-45.
[18] Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:1205.1923.

