A GENETIC INVESTIGATION OF ANTHROPOMETRIC TRAIT FEATURE SELECTION SCHEME (GIANT-FS) FOR IDENTIFICATION OF BIOMARKERS TO DIAGNOSE CANCER DISEASE

Thamizhselvi Elumalai

Research Scholar, Department of Computer Science and Engineering,
Puducherry Technological University, Puducherry, India
thamizhselviee@gmail.com

Geetha Vaithiyanathan

Associate Professor, Department of Information Technology,
Puducherry Technological University, Puducherry, India
vgeetha@pec.edu

Abstract – Due to its vigorous nature and level of reach, cancer disease still challenges the medical fraternity. Researches started to pay attention towards designing a system for diagnosing cancer disease at early stage of human life. In this direction, a helping hand data mining techniques play a pivot role in analyzing, identifying and selecting the features for diagnosing disease. In Recent years, Cancer biomarkers identification gained attraction with its significance in medical field. Hence, in this paper, a new feature selection scheme named Genetic Investigation of Anthropometric Trait- Feature Selection (GIANT-FS) for diagnosing cancer disease is proposed. Initially, the proposed scheme utilizes anthropometric traits for filtering the irrelevant features by which the optimal less number of features are selected and retained. Secondly, the selected features are then fed to Genetic Algorithm (GA) for identifying the optimal cancer biomarkers that helps in diagnosing the cancer disease. The proposed feature selection scheme is experimented with standard UCI data repository. The performance of the proposed scheme is evaluated with standard precision, recall, error estimation rate and also compared with recent existing systems.

Keywords: **Biomarkers; GIANT; Anthropometric; Feature Selection.**

1.Introduction

Datamining is a powerful tool to extract hidden predictive information's from large Database. There are three phases involved in datamining viz (i) Data preprocessing defines the step towards data preparation, (ii) Data modelling: - defines the step towards the predictive and descriptive algorithms and finally (iii) Data post processing: - defines the step towards visualization and evaluation of knowledge extraction. They were several datamining techniques have been developed and used in various fields that includes associative rule mining, classification, clustering, prediction, sequential patterns and feature selection, etc., This kind of technique have been applied to various applications such as market analysis, risk management, fraud detection, text mining, opinion mining, web analysis, financial field, natural disaster and medical health care system, etc., The nourishment of datamining techniques is clearly predictable in the medical care field. Since the patient information is very confidential and sensitive, datamining technique should strictly possess the below characteristics are handling patient data input, manage time and accuracy, avoid human errors and non-trivial knowledge extraction etc. In medical field, the researches mainly focused on diagnosis of disease for earlier detection and risk factor identification apart from prognosis, treatment, screening and monitoring of diseases. Diagnosis is often challenging, because signs and symptoms are nonspecific for several diseases. In this sense, we fixate on cancer diseases, in which abnormal gene were growth uncontrollably and destroy body tissue. The gene expression in cancer is used in molecular biology to query the expression of thousands of gene simultaneously that often helps in predicting the patients with reliable identification of cancer types. For this purpose, a particular type of gene with subset of organism's genomic that DNA will be expressed as mRNA. Therefore, the unique patterns of gene expression for a given cell is to be selected to predict the action of gene expression level. Datamining techniques were also useful in diagnosing distinct cancer types such as breast cancer, lung cancer, colon cancer, brain tumor, leukemia etc., with biological proof, it is obvious that, gene expression will be distinct for every individual with

DOI : 10.21817/indjcse/2023/v14i1/231401077

high level set of features. Hence, handling huge number of features for the purpose of diagnosis is known to be tedious. Therefore, a high dimensional feature selection datamining technique is important for characterizing gene expression that make it possible to identify the particular gene action, reduce the microarray data dimensionality, irrelevant gene filtration and overfitting etc. From the state of art, it is found that the feature selection techniques are widely divided into two types are filter method and wrapper method. The filter method is to evaluate the general characteristic of the training data to select a feature subset without relation to any learning algorithms. Whereas the wrapper method is requiring a predetermined induction algorithm which is to accesses the performance of the features selected. In late 1990's, feature selection in medical field were made with generic information. Even though biomarkers were discovered in the year of 1847 by Henry bence-john due to the advancement in technology, it has been reported that, biomarkers play a significant role in diagnosing cancer. The paradigm in which overproduced the tumor-specific protein that can be easily detected as a marker of cancer. The corresponding biomarkers for diagnosing the various cancer disease as listed in Table 1 is also incorporated with specific cutoff ranges of the biomarkers in the same table. The various cancer types biomarkers presented here is CEA (malignant effusion), CEA (peritoneal dissemination), Her-2/neu (stage IV breast cancer), Bladder Tumor antigen (urothelial carcinoma), Thyroglobulin (Thyroid metastasis), Alpha-fetaprotien (Heptocellular carcinoma), PSA (Prostate cancer), CA 125 (non-small cell lung), CA 19.9 (pancreatic cancer), CA 15.3 (breast cancer), IGF-II (Ovarian cancer), CD 98 (Lung cancer), Troponin I (myocardial infarction), B-Type natriuretic peptide (heart failure), etc.,. Disastrously, few markers immediately stand out as superior diagnostic tools and even fewer have been validated and approved. Still there is some obstacles to verify and validate the biomarkers in cancer is limited, and not yet validate due to the presence of huge dimensional genetic data's that leads to time complexity to diagnose the disease, hidden biomarkers, overfitting of genomes.

Marker	Disease	Cut Off	Sensitivity	Specificity
CEA	Malignant Pleural Effusion	NA1	57.5%	78.6%
CEA	Peritoneal Cancer Dissemination	0.5 ng/ml	75.8%	90.8%
Her-2/neu	Stage IV Breast Cancer	15 ng/mL	40%	98%
Bladder Tumor Antigen	Urothelial Cell Carcinoma	NA	52.8%	70%
Thyroglobulin	Thyroid Cancer Metastasis	2.3 ng/ml3	74.5%	95%
Alpha-fetoprotein	Hepatocellular Carcinoma	20 ng/ml	50%	70%
PSA	Prostate Cancer	4.0 ng/mL	46%	91%
CA 125	Non-Small Cell Lung Cancer	95 IU/mL	84%	80%
CA19.9	Pancreatic Cancer	NA	75%	80%
CA 15.3	Breast Cancer	40 U/ml	58.2%	96.0%
leptin, prolactin, osteopontin, and IGF-II	Ovarian Cancer	NA	95%	95%
CD98, fascin, sPIgR4, and 14-3-3 eta	Lung Cancer	NA	96%	77%
Troponin I	Myocardial Infarction	0.1 microg/L	93%	81%
B-type natriuretic peptide	Congestive Heart Failure	8 pg/mL	98%	92%

Table 1. Biomarkers Cutoff

2.Literature survey

Author	Technique	Theme	Advantages	Drawbacks	Metrics	Dataset
Mohan Allam et al[2022]	FS-BTCBO (Feature Selection-Binary Teaching Learning based optimization algorithm)	Compute the fitness of individuals for evaluating the malignant and benign tumors	Less number of features selected, High Accuracy rate	Sample erroring	Error rate, Accuracy	WDBC dataset
Negar Maleki et al [2021]	KNN method	KNN algorithm to find a correlation between the clinical information and data mining technique to support lung cancer prognosis	Minimize the overall miscalculation of the KNN, Avoid overfitting of data	More features needed	K parameters, Sensitivity, Specificity	Lung Cancer dataset
Punitha et al [2021]	IAIS-ABC-CDS	Integrated Artificial Immune System and Artificial BEE Colony to process effective feature selection	Enhancing Local search process efficiently, Less time	Absence of Global searching process.	F-score	Breast cancer
Md Akizar Rahman, Ravie [2020]	Two step Feature selection with 15 neuron Neural network	Classify the cancer disease with two step Feature selection method based on ANN	Reduce features, Accurate cancer detection with larger dataset		Error rate, TP, TN, FP, FN, ROC curve	WDBC dataset
Moloud abdar et al [2020]	Stacking and voting classifier	Nested ensemble with 2-meta classifier for detecting benign and malignant cancer	Improve classification algorithm by time	High dimensionality rate	Time complexity, Space complexity	WDBC dataset
Golnaz Sahebi et al[2020]	GeFe (Generalized wrapper feature selection) approach	A new operator for weighting features, improved the mutation and crossover operator and integrated nested error validation in GA process	Time complexity is low	Low clarity of imaging data causes high error rate	Accuracy, Sensitivity, Specificity, T-test	Lung cancer, Dermatology, Arrhythmia, WDBC, Hepatitis etc.,
Adeola Ogunleye et al[2020]	XGBoost model	Extreme gradient boosting is a base mode for chronic kidney disease diagnosis	Less no.of features , High performance rate	Time taken for classifying the classes.	Accuracy, Sensitivity, Specificity	CKD dataset
Beatriz Remeseiro et al [2019]	Feature selection in medical application	Review of Feature selection methods in diagnosing various disease	Find matrices include facial recognition, text classification	Low clarity of imaging leads to high error rate	TP, TN, FP, FN	Microarray data, Biomedical signal processing

Agnieszka woriak, Danuta [2018]	RCA(Reversed correlation) feature selection	Integrates a feature selection and clustering with statistical inference to improve medical diagnosis	Finding new dependencies between parameters, Relevant feature selection	Dynamic inference not focused	No. of features selected	CORONARY UCI data repository
Paul delmar et al [2017]	Identification of biomarkers	Statistical method and scoring algorithm helps to identify the biomarkers for diagnosis of diseases	Feasibility, Best feature selected	No optimal cutoff value for feature selection	Hazard ratio (HR)	Bevacizumab metastatic breast cancer
Xiaofeng zhu et al [2017]	Identify Single Nucleotide Polymorphisms (SNP)	Low rank graph regularized sparse regression mode to found association between SNP and brain imaging features.	Reduce noisy data, redundancy features	Focused on single brain imaging modality	Rank, Accuracy	Alzheimer’s disease
Wengang zhou et al [2014]	Biomarkers cancer Feature selection	Maximum relevance Binary particle swarm optimization (MRBPSO) method is to select best features and Class Dependent Multi category (CDMC) is to classify the class to identify biomarkers	Unique feature	Time complexity	Precision, recall, accuracy	Tumor datasets
Silu zhang et al [2017]	Selection method for breast cancer subtype	1-norm and 2-norm Support vector algorithm is to select the gene and subtype prediction	Discover new biomarkers found for therapy	Focused only on class features	Accuracy	PAM50
Zhi-cheng li et al [2017]	Identifying radiomics image signature in glioblastoma multiforme	Radiomics model	Identify MR images	Extraction of high throughput	Error rate	Brain imaging
Li-yeh chuang [2011]	Hybrid Feature selection method on microarray data	Correlation based feature selection(CFS), Taguchi genetic algorithm(TGA), K-nearest neighbor(KNN) combined together to extract gene features.	Eliminate irrelevant features, Fast coverage	TGA local search	Number of features selected, accuracy	Brain tumor, Leukemia, DLBLL

Vasileios ch.korfiatis et al [2013]	Wrapper feature selection to diagnosis of polythemia disease	LM-FM (Local maximization and floating maximization) method to find best possible subset features	Less time complexity	Not presented different diagnostic methods	Accuracy, Sensitivity, Specificity	Healthy primary polycythemia, Secondary polycythemia
Saima rathore et al [2017]	Minimum Redundancy maximum relevance(MRMR)feature selection	Classify colon biopsy images with MRMR to identify normal/ malignant tissues to predict cancer	Biopsy image color distribution is good, Reduction of features	Malignant images are different in cancer grades, Time complexity	Accuracy, Sensitivity, Specificity, F-score, MMC (Mathews coefficient), Kappa test	Colon biopsy images
Somsak Rakkeitwinaia et al [2015]	Principal feature analysis of minimum distribution overlap (PMDO)	PMDO is to Select the relevant features to avoid class overlap in each dimension to diagnose various cancer types	Large degree of overlap, Less time	Overfitting	Accuracy, Sensitivity, Specificity, T-test	Breast cancer, NSCL, CML(1), CML(2)
Nicole A. capela [2015]	Filter feature selection based human activity	Relief-F(RF), Correlation based feature selection (CFS), Fast correlation based filter (FCBF)	Select best Anthropometric traits	Not evaluate redundancy features, Not applicable for larger datasets.	RF level, CFS level, FCBF level, accuracy, error	Wearable smart phone data
Amit assa et al [2016]	Diagnosis of Coeliac Disease(CD) With anthropometric meaures	Multivariable Logistic regression	Less time	Not applicable for all age of human	P-Value(univarite analysis), Multi variable analysis(OR)	Coeliac Disease data
Mary Balliett and Jeanmarie Burke [2013]	Identification of low energy dietary	Convenience sampling technique	Time reduction on dietary plan	High Time complexity	Estimate body composition, Fasting hemoglobin level, Lipid profile	36 women, 13 men

3.Biological background of Anthropometric

The section discuss about the Anthropometry traits. Anthropometrics assessment methodology is one of the selection methods to measure the human position in different dimensionality with the help of various traits or features shown in Figure 1(a). The anthropometric indicator criteria used for the selecting features and they have been justified mainly on the basis of being correlated with other risk factors. During the measurement of anthropometric methods have inherent variation either due to biologic variation or due to error in measurements. There are some parament’s for giving quality assurance on anthropometric measurements like (i) Identification of certified lead anthropometries and trainer, (ii) Manual of standard operating procedure, (iii) Choice of subset equipment, (iv) Equipment calibration, (v) Standardization training and certification, (vi) Measurement resampling. These parameters are used to compute the quality of the anthropometric measurements. We conclude that there is a lack of consistency in the anthropometric field and therefore a more scientifically and theoretically study about anthropometry for the selection of features and use a cutoff points based on the high priority of the research fields.

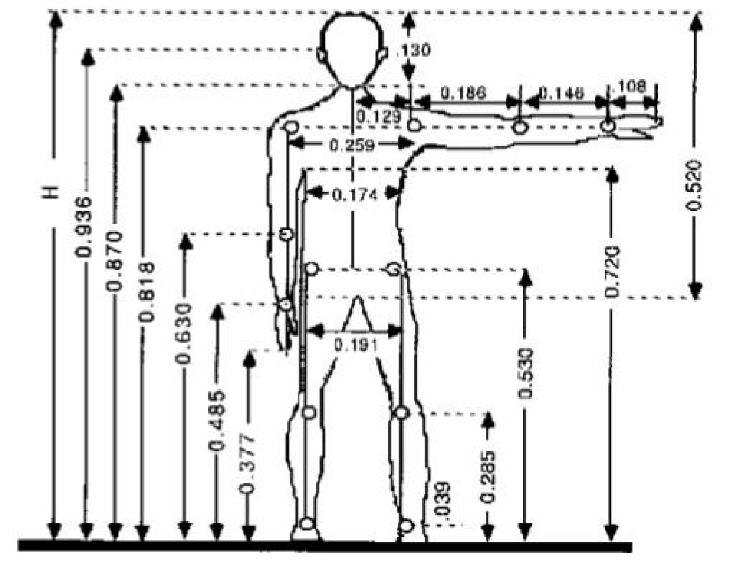

Fig 1(a). Anthropometric measures on human body

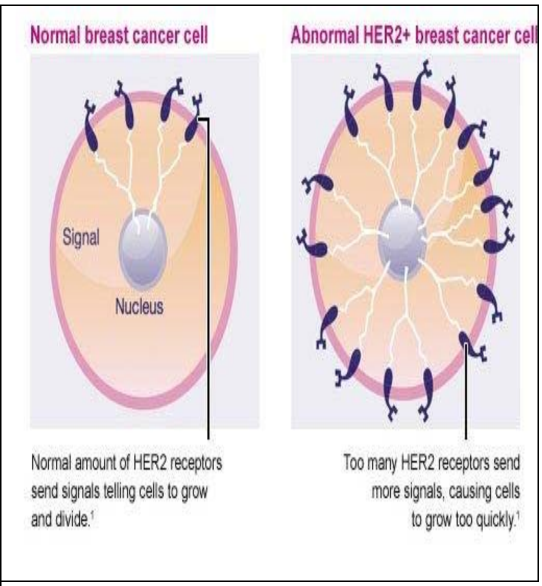
Normal breast cancer cell
Abnormal HER2+ breast cancer cell
Signal
Nucleus
Normal amount of HER2 receptors send signals telling cells to grow and divide.1
Too many HER2 receptors send more signals, causing cells to grow too quickly.1

Fig 1(b). Biomarkers measures

3.1. Biological Background of Genetic Investigation of ANthropometric Traits (GIANT)

This section confabulates about GIANT consortium made of many different groups of institutions, countries and the studies about large scale genetic datasets. This consortium is an international collaboration that seeks the experts to identify the loci which made a change in human body size and shapes, including height and measures of obesity. Anthropometry trait that has been studied by GIANT consortium. Genome – wide association study (GWAS) performed a study across a population’s genomes to find markers that track with particular traits and spots the genome-based heritability of height. The Genome wide association approaches have already proven their power to detect new gene variants in common complex disease condition shown in Figure 1(b).

Universally, the term GIANT defines the condition which caused by over-production of growth hormone in childhood reveals the human height in of ranges **7 feet** (**2.13 m**) to 9 feet (2.72 m) with the measures of anthropometry. In this direction, variation of human height named as Dwarfism (Short person), Normal, Giant (Tall person). In this sense, we proposed a work based on anthropometric with genetic process which leads to measure the cancer genome with genetic traits like protein rate, loci expression, biomarkers etc., to diagnosis the cancer disease. The main objectives of our proposed work is,

- To design an efficient feature selection scheme for a supervised selection procedure with a new similarity measure.
- To analysis the proposed method identifies less features (gene), to identify individual diseases.
- To develop an efficient algorithm to cope with continuous attributes rather than requiring the kind of attribute discretization in the preprocessing step.
- To investigate the performance of other kinds of heuristic function and updating strategy.
- To reduce the time consumption and error rate.
- To avoid the irrelevant features and reduce data dimensionality.

- To avoid overfitting of genetic data by filtration.

4. Proposed methodology

The proposed feature selection methods help to select the most distinguishing features (Gene) for the classification of different cancer by identifying biomarkers which is used to diagnosis the cancer. The Genetic Investigation of anthropometric trait Feature Selection (GAINT-FS) used to overcome the challenges of overfitting classes (gene) and identify the unique biomarker based on the gene action to diagnosis the cancer disease. GAINT-FS pre-selection a subset of candidate genes which dramatically decreases the dimensionality of biomarker featuring and improves the prediction performances, transparency, compactness and improve efficiency by decreasing human errors.

(1) Initialize the process

Initialize the features with benchmark of various types of cancer. Each dataset corresponds to one discrete linguistic term of attribute, instance or classes. The proposed work begins with training set $\{X_i\}$ with n number of instances.

Where, $X_i = total\ number\ of\ input$
$\{X_i\}$= {1,2,3,…..,n}

(2) Partition Class dependent and Class independent

Construct the correlation matrix to split the classes as Class dependent (Positive class, i.e. Cancer malignant) and class independent (Negative class, i.e. Cancer begin) from the input set X_i. The matrix measures the linear dependence between the training set $\{X_i\}$, whereas the each input sets are correlated with one another and allow to identify the highest correlation termed as Class dependent (C_d) and the lowest correlation termed as Class independent (C_{id}). The below equation (1) represents overall input classes with high dimensionality in matrix formation. Equation (2) and (3) obtained from the equation (1) after the evaluation of the correlation matrix and split the classes into C_d *and* C_{id} that reduce the data dimensionality.

$$C[i\ ,id] = [d_{00}\ d_{01}\ d_{10}\ d_{11}\ ]\ [id_{00}\ id_{01}\ id_{10}\ id_{11}\ ] \quad (1)$$

$$C_d \quad = \{C_{d1}, C_{d2}, C_{d3},\dots, C_{dk}\} \quad (2)$$

$$C_{id} = \{C_{id1}, C_{id2}, C_{id3} \ldots, C_{idk}\ \} \quad (3)$$

$$t_{f(g)}\ \forall\ C_d, C_{id} \quad (4)$$

By determining the above equation (4), the classes are categorized as either positive or negative. Each class consist of set of feature $\{\ t_{f(g)} = 1,2,3, \dots, m\}$.

(3) Apply Anthropometric trait to select the features

From the set $t_{f(g)}$ which consist of m number of features with overfitting of unrelated features that leads to time consumption on disease diagnosis. In this sense, we proposed Anthropometric measures to filter out the optimal features for cancer diagnosis. Our proposed scheme present the anthropometric traits for the filtration of the best features by measuring the dimensionality of the features (gene) *shape, size* based on the action of gene by computing the Gene Growth ratio GG_r to the feature set $t_{f(g)}$ on both classesC_d, C_{id}. Here the term gene represents the features. The GG_r is obtained by performing the rate of Protein level (P_{level}) that might categorized into three levels. i.e. Low(P_l) , MediumP_m , High (P_h)by the function Transcription limitation($Tranc_{limit}$) and Translation limitation($Trans_{limit}$). Biologically represents, the Transcription limitation is the synthesis of protein from mRNA with DNA and the translation limitation is the synthesis of protein from mRNA. These two limits are the main pillar of the central dogma of molecular biology. The limitations have been measured with equation (8). Finally the highest GG_r are selected as a best optimal features to diagnose the cancer. Repeat the equation (5) to (10) until all the features visited in $t_{f(g)}$ and update the selected features in to the empty set $b(f_s)$.where $b(f_s)$ represents the null set that is to store the selected optimal feature list.

$$P_l = \sum_{l=0}^{d,id} Trans_{limit} + Tranc_{limit} \quad (5)$$

$$P_m = \sum_{m=0}^{d,id} Trans_{limit} + Tranc_{limit} \quad (6)$$

$$P_h = \sum_{h=0}^{d,id} Trans_{limit} + Tranc_{limit} \quad (7)$$

$$P_{level} = P_l \cap P_m \cap P_h \quad (8)$$

$$GG_r = P_{level}(t_{f(g)}) \quad (9)$$

$$GGr = \{ b(f_s) \sum_0^n t_{f(g)} \} \quad (10)$$

(4) Identification of Biomarkers with GA

Applied a Genetic algorithm to detecting the biomarkers to diagnose the cancer types with the cooperation of selected features $b(f_s)$. This section describes the genetic evolution process consisting of sequence of operators such as initial population, fitness function, parent selection, crossover and mutation .Hereby the selected features $b(f_s)$ with all the classes are taken in to the initial population. Input features subsist of number of genome that might be categorized as either positive biomarkers or negative biomarkers. The gene made of regular action termed as negative biomarkers that illustrates the stages of cancer as begin. Whereas the irregular action gene which made their action wrongly in human body is termed as positive biomarkers that illustrate the stage of cancer as malignant. Where, z represent number of selected features and m represents the number of parents.

$$bf_s = \{ b_{f1}, b_{f2}, b_{f3}, \ldots, b_{fz} \}$$
$$P_i = \{P_1, P_2, P_3, \ldots, P_m\} \quad (11)$$

$$U_c = P_1 X P_2 \quad (12)$$

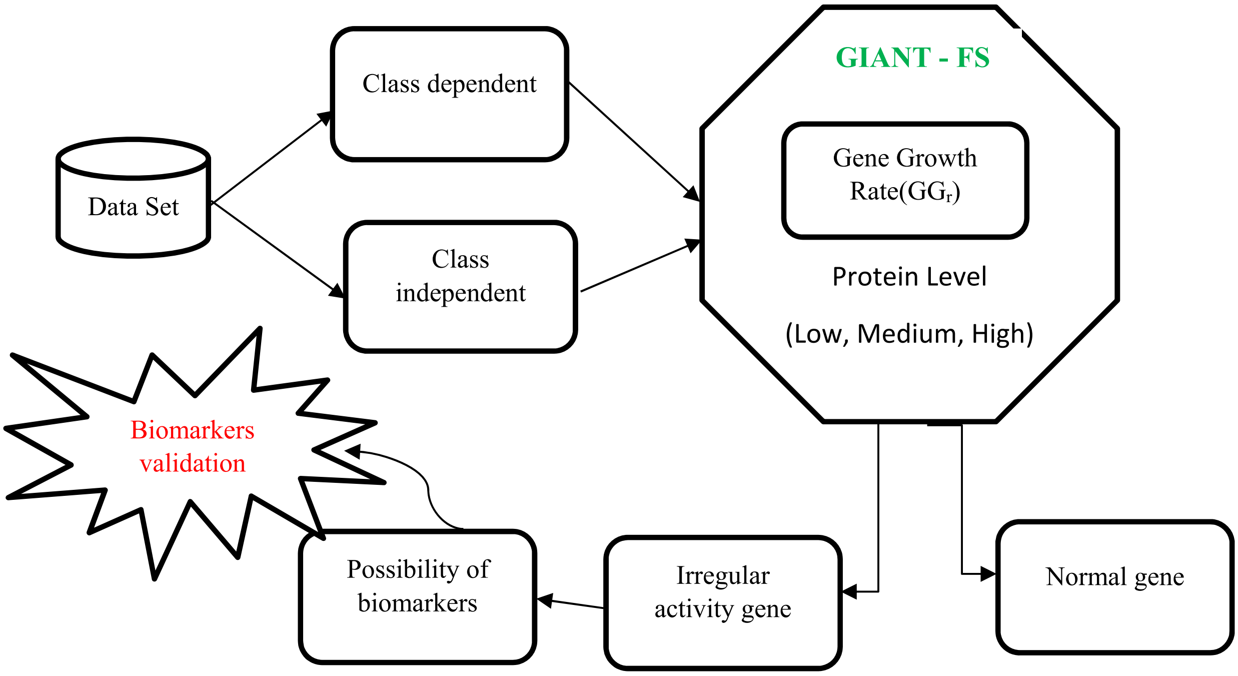

Fig 2. Working flow of GIANT-FS scheme

From the above set f_s , a random selection made to select the parents (P_i) from the equation (11) to generate the child (∂_i) with crossover operation. After the generation of new offspring ∂_i , the mutation process has starts to evaluate the biomarkers by constructing the person correlation matrix that formulate a gene pool network consist of collection of biomarkers as shown in table for the diagnosis of cancer types. The matrix shows the positive correlation (+1, Pos_{bio}) and negative correlation (-1 , Neg_{bio}) . From the matrix, a genetic pool network have been fabricate into -1, 0, +1. Finally we identify the biomarker and diagnose the cancer types that leads to improve the patient life time earlier. Above Fig 2. Represents the flow of our proposed GIANT-FS.

4.1 Proposed GIANT-FS Scheme

Step1: Inputs are initialized with x_i, where $i = 1,2,.,n$

$$X_i = 1,2, \ldots \ldots \ldots, n$$

Step2: Perform the correlation matrix to split the classes based on Class dependent (C_d) and Class independent (C_{id}).

Step3: Evaluate the features $t_{f(g)}$by computing the correlation matrix and classify the feature and its subset based on (C_d) and (C_{id}).

Where,

$$C_d = \{C_{d1}, C_{d2}, C_{d3}, \ldots ., C_{dk}\}$$

$$C_{id} = \{C_{id1}, C_{id2}, C_{id3} \ldots ., C_{idk}\}$$

$$t_{f(g)} \forall \; C_d, C_{id}$$

Step4: Apply anthropometric to select best features by following the steps below,

(i) Initialize the selection process with$t_{f(g)}$.

(ii) Compute the Gene Growth ratio GG_r by measuring the Rate of protein level P_{level} for each gene that falls under the category of low (P_l), medium (P_m), high (P_h) .

$$P_l = \sum_{l=0}^{d,id} Trans_{limit} + Tranc_{limit}$$

$$P_m = \sum_{m=0}^{d,id} Trans_{limit} + Tranc_{limit}$$

$$P_h = \sum_{h=0}^{d,id} Trans_{limit} + Tranc_{limit}$$

$$GG_r = P_{level}(t_{f(g)})$$

$$GGr = \{ b(f_s) \sum_0^n t_{f(g)}\}$$

(iii) Repeat step4 (eq 1) to step4 (eq 4) until all the features are visited.

Step5: Identify the biomarker with Genetic process.

(i) Initialize the set X_i with best filtered features $b(f_s)$.
where $(bf_s) = \{ b_{f1}, b_{f2}, b_{f3}, \ldots ., b_{fz}\}$

(ii) From the best fitted feature, select the parent feature (P_i) to generate the offspring's (∂_i) by using uniform crossover U_c operator. where,

$$P_i = \{P_1, P_2, P_3, \ldots ., P_m\}$$

$$U_c = P_1 X P_2$$

$$P_1 X P_2 = \partial_{1,2,\ldots i}$$

Step 6: Apply the mutation operator to evaluate best biomarker by following the steps given below:

(i) Evaluate pearson correlation matrix

(ii) Compute a strong gene pool network.

$$1\;0\;-1 \quad Gene\,Pool: \; -1\;1\;0 \quad -1\;-1\;1$$

(iii) Detect the biomarker which is more expressive in its action .i.e.Positive Biomarkers that cause hazardous stage of cancer.

$$Neg_{bio} = -1\,0\,-1\,0\,-1\,0\,0\,0\,-1 \qquad Pos_{bio} = 1\,0\,0\,0\,1\,0\,0\,0\,1$$

(iv) Repeat the step5 (i) to (ii) until found a hazardous biomarker.

Step6: Terminate the algorithm.

4. Experimentation Methodology and Result analysis

The experimental analysis has been performed for the proposed GIANT-FS and compared the result with other feature selection techniques for diagnose the cancer disease with the help of cancer Dataset obtained from UCI repository. The proposed scheme is implemented with Microsoft visual studio 2010 as front end and SQL Server management studio as back end. The experiments run without any termination condition for the taken datasets. Here, '10 fold cross validation' method has been used to classify the training set and test cases for all the techniques. Mat lab has been used to implement the existing feature selection techniques such as Class Dependent Multi category (CDMC), Local maximization and floating maximization (LM-FM).

4.1. Result analysis

These section illustrates the efficiency of the proposed GIANT-FS and the results has been compared with recent feature selection technique such as CDMC, LM-FM. Table 1 depicts the original number of features that are extracted form UCI data repository. Since the extracted features are in huge number, the proposed GIANT-FS selects the optimal number of features for different types of cancer disease. It is observed from Table 1 that, 11, 56 and 15 features are originally required for diagnosing breast cancer, lung cancer and leukemia respectively. But , the proposed scheme utilize only 8, 43, and 11 features for diagnosing the cancer types that leads to reduce the data dimensionality and overfitting of features. Likewise, the number of features selected by other selection techniques are also incorporated in the same table 1. From Table 2, it is evident that the proposed GIANT- FS taken in to consideration of less number of features without compromising the accuracy in diagnosing cancer disease. Figure 3 reveals the graphical representation of features selected for diagnosing the cancer types.

Table 2: No. of Features in different cancer types

TECHNIQUES	BREAST CANCER (11)	LUNG CANCER (56)	LUEKEMIA (15)
GIANT-FS	8	24	11
CDMC	9	36	14
LM-FM	9	47	13

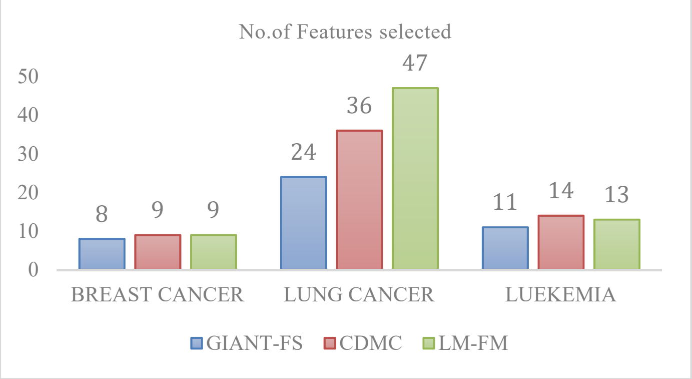

Fig 3. Best Features

In order to check the credibility of the proposed GIANT-FS scheme, the standard measures via sensitivity, specificity and accuracy are used. The results of the performance measures for the proposed scheme is presented in table 3. The Proposed GIANT scheme could achieve an optimal value of 0.95, 0.82, 0.91 for sensitivity, specificity, accuracy in diagnosis of breast cancer. Similarly, the performance result of proposed scheme for lung cancer and leukemia also evaluated and associate in the same table 2. Figure 4 illustrate the graphical representation of precision and recall values.

TECHNIQUE	DATASETS	Sensitivity	Specificity	Accuracy
GIANT	BREAST CANCER	0.95	0.82	0.91
	LUNG CANCER	0.95	0.66	0.84
	LEUKEMIA	0.91	0.73	0.84
CD	BREAST CANCER	0.79	0.64	0.75
	LUNG CANCER	0.91	0.40	0.60
	LEUKEMIA	0.88	0.69	0.80
LM-FM	BREAST CANCER	0.91	0.88	0.90
	LUNG CANCER	0.90	0.63	0.81
	LEUKEMIA	0.91	0.70	0.82

Table 3: Precision for different types of cancer

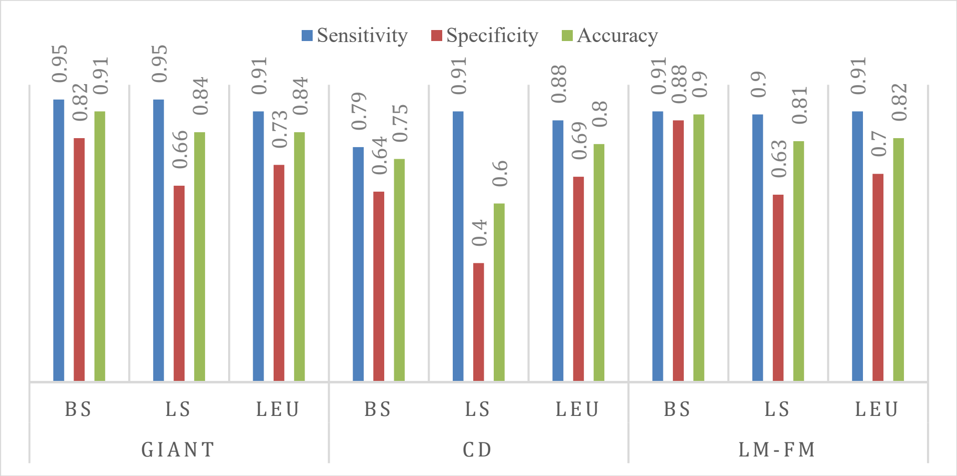

Fig 4. Precision and Recall

The Performance of the proposed scheme is illustrated with standard false positive, false negative and false discovery rate to measure the error estimation rate. The proposed scheme could achieve a false positive rate of 0.17, false negative rate of 0.04 and false discovery rate of 0.10 for diagnosing the breast cancer. Following the same way, Lung cancer could achieve 0.33, 0.05 and 0.17 respectively for leukemia cancer. Likewise, the error estimation rate is computed for existing scheme and presented in same table. From table 3, it is also proved that, there are few cases in which the LMFM scheme outperformance the proposed scheme. In spite of the shortfall, the overall performance in error estimation, the proposed scheme reveals a good performance compared with other existing schemes. Figure 5 graphically represents the error estimation rate of the proposed and existing schemes while diagnosing the disease.

TECHNIQUES	DATASETS	FALSE POSITIVE	FALSE NEGATIVE	FALSE DISCOVERY RATE
GIANT	BREAST CANCER	0.17	0.04	0.10
	LUNG CANCER	0.33	0.05	0.17
	LEUKEMIA	0.26	0.08	0.16
CD	BREAST CANCER	0.35	0.20	0.16
	LUNG CANCER	0.6	0.08	0.52
	LEUKEMIA	0.3	0.11	0.18
LM-FM	BREAST CANCER	0.11	0.08	0.06
	LUNG CANCER	0.3	0.09	0.17
	LEUKEMIA	0.26	0.08	0.18

Table 4: Error Rate

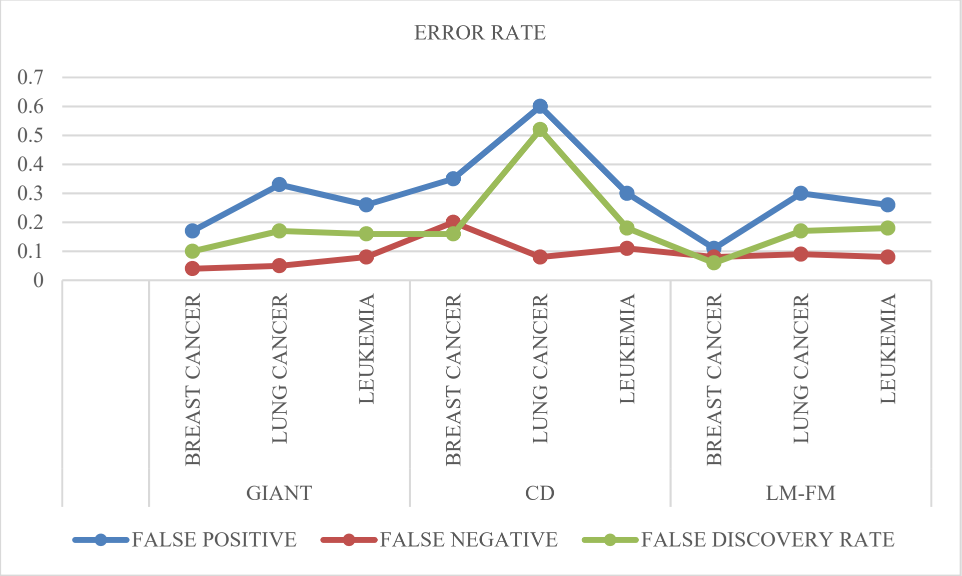

Fig 5. Error reduction

5. Discussion

From the above statistics and discussion, it is clearly observed that the new feature selection scheme named Genetic Investigation of Anthropometric Trait- Feature Selection (GIANT-FS) for diagnosing cancer disease is perform better than the other classification techniques such as CDMC, LM-FM. The performance measures on error estimation shows that LM-FM and CDMC techniques are slightly increases compared to proposed GIANT; Although the proposed GIANT-FS improves its predictive accuracy by selecting less number of features; And also calculating the Precision and Recall show that the GIANT-FS method used to identification of biomarker gene to predict the cancer disease. This indicates that the GIANT-FS method is effective in selecting the features and identify biomarkers with subjective knowledge.

6. Conclusion

In this paper, a feature selection method for identification of biomarker gene for cancer is proposed. The importance of cancer diagnosis in the real world scenario is clearly visible and understood. A detailed survey on feature selection technique for cancer diagnosis is also carried out. From the survey, a motivation for designing a new feature selection technique GIANT-FS for classifying the classes and to identify biomarker's for diagnose cancer. The results show that our proposed GIANT-FS achieved higher classification accuracies. The proposed method identified less biomarker gene which helps to identify the cancer patient quickly. A lot of these marker genes are confirmed to be real biomarkers by literature. GIANT-FS improve the computation efficiency and are robust against overfitting. These methods can be applied to any feature selection tasks in other research fields.

Funding

No funding is provided for the preparation of manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Adeola Ogunleye et al, 2020, "XG boost model for chronic kidney disease diagnosis", IEEE/ ACM transaction, vol 17, Issue 6, PP: 2131-2140
[2] Agnieszka woriak, Danuta Zakrzewska, 2018, "Integrating correlation based feature selection and clustering for improved cardiovascular disease diagnosis", Overcoming Big data barriers in Machine learning technique for the real life applications , Vol 2018, Article 2520706.
[3] Amit assa et al,2016,"Anthropometric measures and prevalence trends in adolescents with coeliac disease: a population based study",BMT, pp.139-144.
[4] Ana pc candido et al,2011,"Anthropometric measurements and obesity diagnosis in schoolchildren",ISSN, ACTA PAEDIATRICA,pp.1651-2227
[5] Anagnostou, T, 2003, Artificial neural networks for decision-making in urologic oncology. European Urology, 43(6), pp: 596–603.
[6] Banzhaf W, 1998,Genetic programming—an introduction: on the automatic evolution of computer programs and its applications. San Francisco; Morgan Kaufmann.
[7] Beatriz Remeseiro, Veronica Bolon, 2019, "A review of feature selection in medical applications", Computer in Biology and medicine, Vol 112, 103375.

[8] Biao jie, 2017, “Temporally Constraint Group Sparse Learning for longitudinal data analysis in Alzheimer’s disease”, IEEE transactions on biomedical engineering.

[9] Celia C. Bojarczuka, 2007, “A constrained-syntax genetic programming system for discovering classification rules”: application to medical data sets. Artificial Intelligence in Medicine, 30 , pp:27-48.

[10] Chapman. P,2000, CRISP-DM 1.0 step-by-step data mining guide. In The CRISP-DM consortium.

[11] D.Deopa et al,2013”Anthropometric measurements of external ear of medical students in uttarakhand region ”,Elsevier on anatomical society,vol.62,pp.79-83.

[12] Dorata Lorkiewicz-Muszynska et al ,2015, “Accurancy of the Anthropometric measurements of skeletonized skulls with corresponding measurements of their 3d reconstruction obtained by CT Scanning”,Schweizerbart ,vol.7313,pp.293-301

[13] Duen-Yian Yeh ,2011, “ A predictive model for cerebrovascular disease using data mining”. Expert Systems with Applications, 38(7),pp: 8970–8977.

[14] Engelbercht,2002, Computational Intelligence. John Wiley&Sons,West Succex, England, pp.345-367

[15] G. Lashkia,2003, “ An inductive learning method for medical diagnosis”, Pattern Recognition Letter . 24(1-3) pp: 273–282.

[16] Golnaz Sahebi et al,2020, “Ge Fe: A Generalized wrapper feature selection approach for optimizing classification performance”, Computer in Biology and Medicine- Elsevier, Vol 125

[17] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[18] Huang, M.-J ,2007, “ Integrating data mining with case based reasoning for chronic diseases prognosis and diagnosis”. Expert Systems with Applications, 32(3),pp: 856–867.

[19] Huang, Y,2007 Feature selection and classification model construction on type 2 diabetic patients’ data. Artificial Intelligence in Medicine, 41(3) ,pp: 251–262.

[20] Humar, K.2008, “,Design of a hybrid system for the diabetes and heart diseases”. Expert Systems with Applications, 35(1-2), pp:82–89 .

[21] Ismail M. Anwar, Khalid M. Salama, and Ashraf M. Abdelbar, 2015, “ Instance Selection with Ant Colony Optimization”, Procedia Computer Science, pp.248–256

[22] Ismail M. Anwar, Khalid M. Salama, Ashraf M. Abdelbar,2015, ADR-Miner: An Ant-Based Data Reduction Algorithm for Classification, pp.515-521

[23] Jairus Hihn, Tim Menzies,2015 Data Mining Methods and Cost Estimation Models, IEEE/ACM International Conference on Automated Software Engineering Workshop, pp.5-9

[24] Koyuncugil, A,2008, Donor research and matching system based on data mining in organ transplantation. Journal of Medical Systems, 34,(3),pp: 251–259.

[25] Li-yeh chung et al, 2011“A Hybrid selection method for DNA microarray data”, Computer in Biology medicine,vol.41(4), pp.228-37

[26] Louise G H goh et al,2015 “Anthropometric measurements of general and central obesity and the perdiction of cardiovarular disease risk in woman : a cross-sectional study ”,BMJ

[27] M. Bishop, 2006, Pattern Recogition and Machine Learning, Springer, New York, pp.656-672

[28] Mary Balliett DC and Jeanmarie.R.Burke ,2013, “changes in Anthropometric measurements ,body composition ,blood pressure ,lipid profile ,and testosterone in patient participating in a low-energy dietary intervention”, Elsevier on found of chiropractic medicine,vol.12,pp.3-14

[29] Md Akizar Rahman, Ravie, 2020, “An Enhancement in cancer classification accuracy using a two step feature selection method on Artificial neural network with 15 neurons”, Research center for cyber security,12(2),271.

[30] Michie, D,1994, Machine learning, neural and statistical classification. Ellis Horwood.

[31] Milos Radovic et al, 2017, “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data”, BMC Bioinformatics, vol.9

[32] Mohan Allam, 2022, “Optimal feature selection using binary teaching learning based optimization algorithm”, Journal of King saud university- computer and information science, Vol 34, Issue 2

[33] Moloud abdar et al, 2020, “A new nested ensemble technique for automated diagnosis of breast cancer”, Pattern Recognition- Elsevier, Vol 132, PP:123-131

[34] Muhammad Asif, Jamil Ahmed, 2015, “Analysis of Effectiveness of Apriori and Frequent Pattern Tree Algorithm in Software Engineering Data Mining”, International Conference on Intelligent Systems, Modelling and Simulation,pp.28-33

[35] Myoung-Jong Kim ,2003, The discovery of experts’ decision rules from qualitative bankruptcy data using genetic algorithms . Expert Systems with Applications, 25,,pp: 637–646.

[36] Negar Maleki et al, 2021, “A K-NN method for lung cancer prognosis with the use of a Genetic algorithm for feature selection”, Expert systems with applications – Elsevier, Vol 164

[37] Nicole A.capela et al,2015, “Feature selection for wearable smartphone-Based human activity recognition with able bodied, elderly and stroke patient”, PLOS.0124414

[38] Olfa Ben Ahmed,2016, Recognition of Alzheimer’s disease and Mild cognitive Impairment with multimodel image-derived biomarkers and Multiple Kernel Learning, Elsevie.

[39] Paul Delmar, 2017, “Innovative methods for the identification of predictive biomarker signature in oncology: Application to bevacizumab”, Contemporary clinical trials communications.

[40] Peter Clark , 1989,The CN2 induction algorithm . Machine Learning , 3(4), pp:261-283.

[41] Punitha et al, 2021, “An Automated Breast cancer diagnosis using feature selection and parameter optimization in ANN”, Computer and Electrical engineering -Elsevier, Vol 90,106958

[42] Shao, 2016, “An Organelle correlation-Guided feature selection approach for classify multi-label subcellular Bio –images”, ieee transactions on computational biology and bioinformatics.

[43] Silu zhang, 2017, “Novel gene selection method for breast cancer intrinsic subtypes from two large cohort study”, IEEE International Conference on Bioinformatics and Biomedicine.

[44] Somsak Rakkeitwinaia, Chidchanok Lursinsapa, Chatchawit Aporntewana, Apiwat Mutirangurab,2015, “ New feature selection for gene expression classification based on degree of class overlap in principal dimensions”, Computers in Biology and Medicine, pp.292–298

[45] Taiping zhang,2016, “Learning Proximity Relations for Feature Selection”, IEEE transactions on knowledge and data engineering.

[46] Tamer Ucar ,2011, “ Predicting existence of mycobacterium tuberculosis on patients using data mining approaches” , Procedia Computer Science,Vol 3, pp: 1404–1411.

[47] Tan, P.-N, 2006 Introduction to data mining. Boston: Pearson Education Inc.

[48] Teng-eow tan and yao-chung chung,2012”Association of Anthropometric measurements with components of metabolic syndrome and carotid Intima-media thickness in young healthy Taiwanese”, Elsevier on medical ultrasound ,vol.20,pp.210-214

[49] Turker Tekin Erguzela, Cumhur Tasb, Merve Cebic,2015, “ A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders”, Computers in Biology and Medicine, pp.127–137,

[50] Vasileios ch.Korfiatis et al,2013, “A Classification system based on a new wrapper feature selection algorithm for the diagnosis of primary and secondary Polycythemia”, Elsevier, vol.43(12), pp.2118-2126.
[51] Wengang Zho,2014, “A novel class dependent feature selection method for cancer biomarker discovery”, Computers in Biology and Medicine.
[52] Xiaofeng Zhu, 2017, “ Low rank Graph Regularised Structured Sparse Regression for Identifying Genetic Biomarkers”, IEEE transactions on big data.
[53] Yoo. I, 2012, “ Data mining in healthcare and biomedicine: A survey of the literature”. Journal of Medical Systems, 36(4) pp: 2431–2448,
[54] Zhi-cheng li,2017, “Identifying a radiomics imaging signature for prediction of overall survival in glioblastoma multiforme”, Biomedical Engineering International Conference.

Authors Profile

Thamizhselvi Elumalai received her M.Tech. in Network and Internet Engineering at Pondicherry University and her B.Tech. in Computer Science and Engineering in RGCET at Pondicherry University of Puducherry, India. She is currently pursuing her Ph.D. degree in the department of CSE at Puducherry Technological University under the academic supervision of Dr. Geetha Vaithiyanathan, an associate professor from same university. She is interested in conducting research in data mining and medical data mining.

Geetha Vaithiyanathan is currently an associate professor in the Department of Information Technology at Puducherry Technological University of Puducherry, India. She received her Ph.D. (CSE) and M.Tech. (CSE) from Pondicherry University and her B.Tech. (CSE) from Pondicherry Engineering College. Her research interests include distributed computing, cloud computing, security solutions, and data mining. She has authored more than 55 papers in journals, conferences, and books.