How to Cite:
Ramya, P., & Reddy, T. B. (2022). An efficient filter based classification approach for microarray disease detection. *International Journal of Health Sciences*, *6*(S1), 7803–7819.
https://doi.org/10.53730/ijhs.v6nS1.6720

An efficient filter based classification approach for microarray disease detection

P. Ramya
Research scholar, Dept. of Computer Science & Technology, Sri Krishnadevaraya University Ananthapuram, India

Dr. T. Bhaskar Reddy
Professor, Dept. of Computer Science & Technology, Sri Krishnadevaraya University, Ananthapuram, India

***Abstract*---**As the size of the biomedical databases areincreasing day-by-day, finding anessential featureset for classification problem is complex due to large data size and sparsity problems. Microarray feature ranking and classification is one of the major challenges to scientific and medical researchers due to its high dimensional feature space and limited number of samples. Feature transformation, feature ranking and data classification are the essential components to improve the microarray cancer prediction on high dimensional datasets. In this work, a novel framework is designed and implemented to classify the high dimensional data with high true positive rate. In the proposed work, a hybrid feature transformation, hybrid feature selection and advance classification approach are implemented to improve the true positive rate and error rate of the disease prediction. A novel principal component ranking measure is integratedin order to find the subset of features for classification problem. Finally, a hybrid decision tree classifier is used to predict the classification accuracy on the selected features set. Experimental results proved that the present framework has better performance compared to the traditional models for variable microarray datasets.

***Keywords*---**Microarray data, feature selection, data transformation and classification.

Introduction

In general, microarray data contains a set of genes and their disease connections. The majority of traditional approaches to finding patterns on high-dimensional datasets are ineffective and computationally infeasible. As a result, processing all of the genes that aren't necessary throughout the classification process is tough.

In addition, the overall computational overhead rises dramatically. During the classification process, unwanted noise is produced. As a result, it is critical to identify a small number of genes that are typically involved in the classification process. A perfect blend of filter and wrapper systems is used in all classic gene selection techniques[1]. Filtering techniques are responsible for ranking each individual characteristic based on its usefulness. The association between each individual gene and its associated class label is taken into account during the ranking process. The above ranking process relies heavily on the univariate scoring system. Prior to the execution of categorization methods, the top rated genes are chosen. Wrapper techniques, on the other hand, necessitate the gene selection approach in order to integrate with a classifier. The primary goal of this method is to assess the categorization accuracy of each individual gene subset. According to the ranking of performance, the ideal subset of genes is discovered. Traditional filtering systems are incapable of measuring the link between distinct genes and are inefficient. During the biomedical diagnosis process, gene expression data is extremely important. According to current research, a small number of genes can result in high prediction accuracy during the cancer illness detection process[2]. A large proportion of genes are unrelated to the condition in question. As a result, throughout the medical data processing process, the gene selection technique is a very difficult work. Feature selection is thought to be the most powerful strategy for reducing the amount of data accessible.

The extreme classification model is a data classification extension of the classic neural network model. It divides the problem into a number of sub-problems and then combines them to obtain the best answer. The hidden layer's parameters comprise training data samples that are mapped to the output layer. The classic SLFN approach involves iterative parameter modifications, which might lead to problems. The proposed extreme classifier technique [3] solves these problems. As a result, the ELM technique delivers improved generalisation and learning speed. The majority of classical learning models for training SLFNs are slower than non-parametric alternatives. Because settings must be fine-tuned iteratively, this method works slowly. Furthermore, these models necessitate a large amount of computational memory, which increases the mapping process' overall calculation time. Extreme classifier [4] is an enhanced and somewhat modified version of the traditional SLFNs technique. This technique is used to improve the efficiency and performance of standard SLFNs. Furthermore, the majority of neural network-based learning schemes employ manual tweaking of control parameters (such as learning rate, learning epochs, and so on) as well as local minima. However, the extreme classifier is applied automatically, and no manual iterative adjustment is required. In ELM, the classification boundary isn't ideal, and it stays the same throughout the training phase. As a result, there's a potential that samples at the boundary will be misclassified. In comparison to other classic tuning-based systems, this strategy necessitates a large number of hidden neurons.

A classifier can learn a set of rules, or the decision criterion, from a set of labelled data that has been annotated by an expert using machine learning. For compared to a system that relies solely on manual input, this approach provides for greater scaling and lower costs when classifying medical data. The majority of machine learning-based medical data classification research has focused on binary classifiers. That is, constructing a classifier using a set of positive and negative

examples in order to determine the membership of medical data in a class. Classification can take several forms, ranging from fully automated systems with minimal human participation [4] to semi-automated systems that combine human and machine intelligence.

The majority of classic ensemble classification models are run with a tiny feature space and a small amount of data. Traditional ensemble classifiers select a predetermined number of features for categorization as the size of the feature space grows. Learning classification models with all of the high-dimensional characteristics might cause major performance and scalability problems. Wrappers, filters, and embedded models are the three types of feature selection measures.

1. Feature selection in high-dimensional datasets is a problem.
2. The problem of disease prediction with a high rate of misclassification.
3. Using the parallel processing model to handle high-dimensional and huge datasets.

Related works

The ensemble learning group's Adaboost (Adaptive Boosting) is a meta learning-based method[5]. The AdaBoost's main goal is to improve the strong classifier by using a group of weak base classifiers. The AdaoBoost method is an iterative strategy in which a weak base classifier is chosen at each iteration to reduce the model's error rate. Machine learning models are severely hampered by high dimensionality. Most classification algorithms use feature selection metrics such as mutual information, correlation coefficient, rough-set, chi-square test, and others to pick a subset of features from a high-dimensional space in order to improve the precision of the classification process. They applied a PCA-based spectral filtering model to the original training data's high-dimensional characteristics. The principle component analysis and maximum likelihood estimation are the two reconstruction approaches employed. Randomization models employed a variety of distribution algorithms. Using the randomization operator and the randomization data, Bayesian analysis is employed in most of these approaches to estimate the original data distribution. On high-dimensional datasets, most standard techniques detect unsuitable and computationally infeasible patterns. As a result, processing all of the cancer patterns that aren't necessary throughout the classification phase is tough. As a result, the overall computational overhead rises dramatically. Feature selection is thought to be the most powerful strategy for reducing the amount of data accessible. Wrapper techniques and filter approaches are the two types of feature selection models used in the cancer classification process. To improve classification accuracy, the wrapper strategy evaluates search features or feature subsets. The filter technique assesses each feature independently of the classification algorithm, ranks the cancer features after review, and selects the best. Information, dependency, distance, and consistency are used in this assessment. Because of cross validation and repeated iteration to evaluate feature subsets, the wrapper model is slower than the filter model in general. Although subset selection is an NP-hard problem, the traditional wrapper model is more efficient because classification technique impacts total accuracy. However, due to the complicated interactions among features, detecting new illness patterns can become

challenging as the number of features in complex data grows. Feature ranking approaches calculate the measure for each feature and then rank them. These approaches choose the top'k' features based on their highest rank and discard those with lower feature ranks [6]. Information gain is one of the entropy value-based attribute selection measures.

Class-related and non-related attributes are found in good subsets of attributes. The approach is used to determine the degree of linkage between the characteristic vectors. The attributes are considered to have a strongly positive correlation when the coefficient of correlation between the two vectors is greater than "0." Similarly, if the correlation coefficient between these two characteristic vectors is smaller than "0," the functions are said to be negatively associated. If the correspondence coefficient of the two characteristic vectors is equal to "0," the features are said to be uncorrelated[4]. Chi-Square is a statistical analysis method. The independence is measured by the functional vector. With observed and anticipated values, the strength of the link between two random variables is examined. The pixels behind feature descriptors are used to optimise classification performance and provide a sparse representation. The descriptors should preferably be unaffected by operations such as scale, rotation, and lighting changes. Because of this invariance, descriptors can be matched across videos with different parameters. Each piece of medical data is scanned and translated into continuous data that is normalised. The large dimensionality and unbalanced nature of medical datasets are the key challenges. Traditional machine learning classifiers classify and predict disease using a subset of characteristics, with a high true negative rate and error rate. To compute the measure for each feature and rank them, attribute selection is utilised. These algorithms choose the top 'k' features based on their highest rank and discard those with lower feature ratings. One of the attribute selection measures based on entropy value is information gain. The mutual information of a target random variable, say P, and an independent random variable, say Q, is used in the information gain strategy. The fundamental drawback of this technique is that it favours features with big distinct values over those with smaller distinct values. The table below outlines the various attribute selection measures used in pattern creation decision trees. Table 0 compares various machine learning algorithms for detecting medical imbalances based on various parameters[7].An enhanced feature reduction based intrusion detection system was proposed by [8].[9]. The feature space is filtered in this model using feature ranking measures including information gain and correlation. The feature reduction strategy is applied when the feature ranking is completed successfully. The feature reduction strategy is accomplished by combining the ranks provided by the information gain and correlation procedure. The reduced features are used to train and test medical characteristics in a cancer dataset using feed forward neural medical. The pre-processing is done manually in this method, which is a major flaw in this model. For various types of concept drift, Liu et al developed a knowledge-maximized ensemble strategy [10]. They introduced an enhanced data stream classifier termed as knowledge maximised ensemble in this paper. As a result, limiting the amount of training data becomes difficult and complicated. Through the integration of several imbalance detection methodologies, this strategy can be influenced by many types of concept drift. Decision tree induction is a straightforward and effective classification method that generates a tree and a set

from a single dataset[11] to represent a model for many classes. The choice tree is a flowchart-like tree structure in which each inner knot represents a single attribute test, each branch represents a test result, and each leaf node represents a class. The root node is a tree's highest node. The decision tree classification technique comprises two phases: tree construction and tree pruning. CART was introduced by them (Classification and Regression Trees). To choose a characteristic from a list of feature space, different statistical measures are applied. CART uses both numerical and categorical attributes to form a decision tree and contains tools to deal with missing attributes[12]. It uses a large number of single fractional criteria, such as a gini index, gini ratio, and others, as well as one multi-variable (linear combination) in choosing the best partition and data is sorted at all nodes to get the best fractional point. The linear combination splitting criteria are utilised in regression analysis. The training data set is used to develop a classification model, whereas the test data record is used to validate the model. It's used to classify and predict fresh records that aren't yet trained or tested[13]. Controlled learning algorithms (such as clusters) are preferred to uncontrolled learning algorithms (such as clusters) because their prior knowledge of data log class labels simplifies the selection of features / attributes and, as a result, leads to the prediction / classification of accuracies. Some studies have been successful in using the Rough sets theory to the classification of various medical complications[14]. The error rates for rough sets were shown to be totally comparable to those of other computational techniques, and in many cases were much lower.

Proposed Model

Figure 1 depicts the overall design of the proposed paradigm. Each microarray gene illness dataset is first filtered to eliminate the gene's sparsity problem and missing values. The feature values are transformed using the gaussian transformation measure utilising a hybrid data transformation strategy. To strengthen the balance property of each feature and its class, each value is normalised. Each feature is converted using the gaussian transformation procedure in the first phase. The hybrid PCA technique is used to extract key characteristics in the second phase. Finally, for the prediction procedure, an efficient decision tree classifier is constructed to detect the essential cancer patterns.

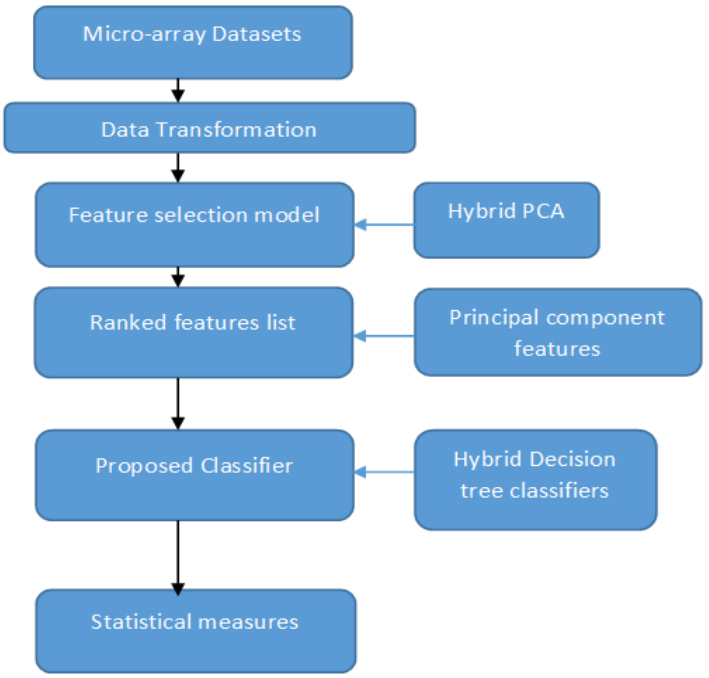

Figure 1: Proposed Gaussian filter based feature selection and classification model

Each microarray training dataset is pre-processed using the data transformation function to remove the variation among the data distribution. In the proposed work, a gaussian based data transformation function is used to normalize the input training data for clustering and wrapper feature ranking process in the mapper phase.

Gaussian based Data Pre-processing

Input : Training microarray dataset D, F(D): Feature space of D.
Output: Gaussian Filtering or Transformed data KD.
Procedure:

1. Read input data D.
2. For each feature F[i]in feature space F(D)
3. Do
4. Apply Gaussian transformation on each feature as

5. $\text{GeneKernelTransform}(F[i]) = \phi = (\sum (F[i]) / \max\{F[i]\}).\frac{1}{\sqrt{2\Pi}} e^{-(F[i]-\mu(F[i]))/\sigma(F[i])}$

6. If(ϕ>0)

7. Then

8. Normalize F[i] using Min-max normalization $[0, \phi]$

9. Else

10. Normalize F[i] using Min-max normalization with [R1,R2]

$$x' = \frac{x - \min\_(x)}{\max\_(x) - \min\_(x)} * (R2 - R1) + R1$$

11. End if

12. Done

Here, gaussian normalization and min-max range normalization are used to improve the data distribution with uncertain values. R1 and R2 represents the minimum and maximum values of each feature. This approach is used to remove the sparsity problem and data normalization problem in high dimensional datasets.

Algorithm: HybridPCA (HPCA)

In the proposed feature ranking model, traditional PCA algorithm is enhanced to find the most essential features in the given feature space. PCA is used to find the most significant features using the covariance between the features and its Eigen vectors. If the covariance between the features represents positive, then it indicates the strong relationship among the features. Similarly, if the covariance between the features represents negative, then it indicates the weak relationship among the features.

Input: Normalized training data.
Output: Ranked features.
Step 1: Input Normalized data ND.
Step 2: Mean of the normalized data is computed using eq.(1)

$$\mu_D = \sum \text{ND}[i] / N \quad \text{-----(1)}$$

Step 3: Let F={f[0],f[1]….f[m]} be the feature space with m features.
Find the candidate features pairs

CF={(f[0],f[1]),(f[0],f[2]),(f[0],f[3])……(f[m],f[0])….}.

For m feature space we will get $\frac{m!}{(m-2)!2!}$ candidate sets.

For each pair of candidate features CF

Do

Compute covariance between features as

$$Cov(CF\{x,y\}) = \frac{\sum_{i=1}^{n}(CF[x_i]-\mu_{CF[x_i]})(CF[y_i]-\mu_{CF[y_i]})}{(n-1)} \quad \text{---(2)}$$

Done

Step 4: Compute the eigen vector and values using the eq.(3) and eq.(4)

$$EigenValues[] = Det(\lambda I - COV(CF)) = 0 \quad \text{---- (3)}$$

Here I is the identity matrix of same dimension as COV(CF). The corresponding Eigen vector is given as

$$(\lambda I - COV(CF))v = 0 \quad \text{------(4)}$$

Here the optimal eigen sum is computed as

$$OptimalEigenSum = \frac{\sum EV[i])}{MaxProb\{F(EV[i]), C[m])\}}; \quad \text{----(5)}$$

for m classes

Step 5: Selecting the highest ranked principal components using the maximum conditional probability computation measure. Sort the eigen values according to descending order. The highest eigen value is the principal component of the dataset and it is more significant for classification problem.

Proposed Classification Algorithm

Step 1: Partition the given training dataset into 'm' classes.
Step 2: For each partition
Step 3: do
Step 4: Apply the decision tree classification feature selection measures one each partition.

Proposed hybrid measure is used to minimize the runtime (ms) and to improve the true positivity and false negative rate on large datasets. Since, network training data have nominal attributes and numerical attributes, proposed measure is used to find the nominal association between the numerical and nominal attributes using the following measures for node selection.

Hybrid Entropy Measure for Hoeffding Tree Construction

Let NDbe the normalized selected feature, cbrt is the cube root, chiVal is the chisquare value. The hybrid entropy of the Hoeffding tree is given by

$$n = \sum \mathrm{ND[i].log(ND[i])}$$

$$\mathrm{Ent}(D_p) = \frac{(n + \log(\sum \mathrm{ND\,[i]})}{\sqrt{(\sum(\mathrm{ND[i]} * (\mathrm{ND[i]} - \mu_{ND})^2}}$$

$$\mathrm{HCondEntropy}(D_p) = \frac{-\mathrm{Math.\,cbrt(entropyConditional(ND[i])} * \mathrm{total)} * n}{(\mathrm{CramersV(ND)} + \mathrm{chiVal(ND)})}$$

$$\mathrm{CramersV}(D_P) = \mathrm{Math.\,sqrt(chiVal(ND)}/\left(\sum \mathrm{ND[i].} * \min\{\mathrm{nrow, ncolumns}\}\right)$$

$$\mathrm{chiVal(ND)} = \text{yates chisquare value for ND.}$$

$$\mathrm{entropyConditionalent(ND)} = -(\sum \mathrm{ND[i].log(ND[i])}/(\log m * \sum \mathrm{ND[i]})$$

where m represents m classes.

Proposed Random forest attribute selection measure (RFASM)

$$\text{Modified Gain} = e^{-n/(\log 2 * \sum \mathrm{ND[i]})} + \mathrm{Gain(D)}$$

$$\mathrm{HRTASM} = \frac{-n * \sqrt[3]{\mathrm{Ent(ND)}}}{(\mathrm{ND[i].\,chiVal(ND)})^3}$$

Step 5: Construct the decisiontree using the enhanced attribute selection measure to improve the true positive rate of the classification.

Feature selection based Ensemble Classifier for gene pattern discovery

Input: Feature Maximum features Max; Ranked Features RF.

Output: Gene patterns

Procedure:

1. For each partition in input data M[i]
2. Do
3. Let the set of gene feature ranking measure are represented as GF.
4. GF[]={"HPCA","PCA","PSO","ICA"};
5. Let the set of base classifiers are denotes as C[]={"SVM","GFSRandomForest"," GFSHoeffdingTree"};
6. Apply feature selection method GF[] on the partition using feature selection measures.
7. Sort features using the gene ranking values.
8. Select Max ranked features RF[] from the sorted list for ensemble gene pattern discover

9. Apply the classification models C[] on the partition using
10. ClassPredictions CP[]={};
11. For each classifier C[i] do
12. Do
13. If(CP[0]=="SVM")
14. CP[0]=Classify(C[0],RF[]);
15. Else if(CP[1]==" GFSRandomForest")
16. CP[1]=Classify(C[1],RF[]) using random forest attribute selection measure for decision tree construction.
17. Else if(CP[2]==" GFSHoeffdingTree")
18. CP[2]=Classify(C[2],RF[]) using hoeffding attribute selection measure for decision tree construction.
19. To select the optimal gene for disease prediction the majority voting of the CP[0],CP[1] and CP[2] are considered to improve the highest true positive rate and accuracy.
20. End for
21. End for

Experimental Results

To evaluate the performance of the proposed model to the existing models, different microarray datasets were selected from the biomedical repository. Different dataset used for experimental evaluation are summarized in Table 1. In the experimental results, 10% of the training data are used as testing data for performance evaluation. Proposed feature selection-based ensemble methods increase the performance of true positive rate and accuracy on entire high dimensional datasets. Proposed model uses the entire training data set for construction of decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models. From the experimental results, it is clear that proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

From the experimental results, it is clear that proposed ensemble classification improves the overall true positive and false negative rate. Also, the main advantage of using proposed model is to reduce the error rate on high dimensional features.

Table 1
Datasets and Its Characteristics

Micro array Datasets	Gene sets	Data-Type
Prostate	2136	Continuous/Numeric
Lymphoma	5000	Continuous/Numeric
DLBCL-Stanford	4000	Continuous/Numeric
Breast cancer	24481	Continuous/Numeric
Leukemia	7129	Continuous/Numeric

Proposed model increase the performance of accuracy and error rate on entire high dimensional microarray datasets. Proposed classification model uses high dimensional data set to generate decision patterns; therefore, the prediction accuracy of each cross validation tends to be more accurate than the traditional ensemble classification models.

Table 2
Comparative analysis of present approach to the traditional approaches by using accuracy on different Microarray dataset

Model	DLBCL	Prostate	Lymphoma	BreastCancer
PCA+Ensemble	87.45	91.43	90.54	85.35
ACO+Ensemble	86.35	89.13	91.33	82.54
Fuzzy PCA+Ensemble	83.24	87.34	89.23	84.67
KM+SNR+Ensemble	94.65	91.43	92.53	86.75
KM+t-test+Ensemble	95.74	94.64	95.09	87.14
KM+SAM+Ensemble	91.53	89.13	87.45	91.43
HPCA+SVM+HRFDT+HHDT	96.85	95.81	96.98	93.64
HSNR+SVM+HRFDT+HHDT	97.98	95.14	96.13	93.18
HT-test+SVM+HRFDT+HHDT	97.15	94.92	95.93	94.82
Gaussian based Deep neural network	98.93	96.35	97.14	94.92
Proposed Model	99.23	98.45	98.71	97.93

Table 2, describes the performance of the proposed model on all cancer datasets. Here, all the cancer datasets are evaluated using the proposed model to find the average true positive rate and precision rate on the high dimensional datasets. From the table, it is visualized that the present approach has better true positive rate and precision over the existing models.

Table 3
Gene-Disease features extraction using the proposed PCA model

ABC	ACO	Chi-square	Mutual information	Information Gain	Genetic Algorithm	PCA	Proposed PCA
58	59	56	78	63	56	51	41
53	60	54	78	60	52	56	37
55	62	54	74	59	62	69	43
56	63	52	82	57	53	60	46
58	54	55	63	58	60	72	48
56	50	59	85	64	54	52	44
65	61	65	75	58	58	69	42
53	62	65	77	55	62	75	50
64	57	63	79	50	62	56	41

65	64	52	85	59	57	69	36
60	62	71	69	62	56	50	36
56	55	61	70	57	61	70	38
59	61	70	67	65	58	74	42
56	53	67	69	62	60	58	45
54	58	59	69	51	55	60	39
51	54	72	74	60	57	72	45
53	59	64	61	57	53	60	46
60	61	59	85	50	58	65	44
56	59	62	76	62	61	69	38
58	52	62	71	51	60	54	38

Table 3, illustrates the performance of gene-disease feature extraction using the proposed PCA approach on large datasets. From the table1, it is clearly shown that the present feature extraction procedure has high filtering rate as compared to the existing approaches.

Table 4
Performance analysis of computational runtime (ms) with different traditional feature selection models

Features Size	ABC	ACO	Chisquare	Mutual information	Information Gain	Genetic Algorithm	PCA	Proposed_PCA
GeneDisease-100	5417	7117	6297	7230	7024	6638	6481	4747
GeneDisease-200	5365	6071	6529	6957	5917	6253	6516	3965
GeneDisease-300	6200	5753	6609	5959	6882	7099	6469	3474
GeneDisease-400	6122	6801	5514	6055	5729	7430	5584	3495
GeneDisease-500	6727	5488	5676	7488	6685	6103	6753	4809
GeneDisease-600	5774	5592	6089	6823	6273	5996	6597	4292
GeneDisease-700	7205	6524	6045	6060	7449	7448	6319	3888
GeneDisease-800	6976	5859	7340	6036	6544	6855	5864	4723
GeneDisease-900	7080	6097	5710	6692	5661	6866	7365	4488
GeneDisease-1000	5854	6731	7118	7087	5916	5541	5552	4368
GeneDisease-1100	6431	7548	7224	5597	7359	7022	6253	4470
GeneDisease-1200	6595	5404	5767	5989	7086	5950	7552	3995

GeneDisease-1300	6077	6919	7018	5357	6242	5645	6886	4335
GeneDisease-1400	7350	5536	6953	5711	6927	7133	7061	4439
GeneDisease-1500	7536	7139	5695	5414	6211	6909	6438	4303
GeneDisease-1600	7557	7237	7425	6421	6729	6922	5686	3618
GeneDisease-1700	7032	7402	5520	6334	5614	5942	5840	4524
GeneDisease-1800	7556	5435	7183	5417	5700	7236	5630	4734
GeneDisease-1900	5669	6159	5589	5708	7281	5900	6588	4568
GeneDisease-2000	6318	7119	6253	5603	6852	5925	7090	4178

Table 4, describes the performance of computational runtime(ms) of gene-disease feature extraction using the proposed PCA approach on large datasets. From the table3, it is clearly shown that the present feature extraction procedure has low computation runtime as compared to the existing approaches.

Recall

Table 5
Performance analysis of recall using different traditional classification frameworks

Features_Size	SVM	Random_Forest	CNN	RNN	HNN	Proposed Model
GCDisease-100	0.83	0.8	0.82	0.86	0.83	0.98
GCDisease-200	0.87	0.87	0.85	0.92	0.8	0.96
GCDisease-300	0.84	0.84	0.89	0.89	0.82	0.98
GCDisease-400	0.89	0.76	0.9	0.91	0.86	0.96
GCDisease-500	0.78	0.79	0.76	0.81	0.84	0.96
GCDisease-600	0.83	0.74	0.71	0.82	0.83	0.97
GCDisease-700	0.85	0.78	0.79	0.83	0.82	0.97
GCDisease-800	0.74	0.74	0.82	0.89	0.92	0.98
GCDisease-900	0.86	0.8	0.81	0.81	0.8	0.96
GCDisease-1000	0.84	0.75	0.82	0.9	0.92	0.97
GCDisease-	0.75	0.8	0.78	0.88	0.92	0.98

1100						
GCDisease-1200	0.74	0.77	0.83	0.88	0.84	0.97
GCDisease-1300	0.83	0.86	0.79	0.82	0.85	0.98
GCDisease-1400	0.78	0.77	0.82	0.84	0.9	0.97
GCDisease-1500	0.89	0.86	0.73	0.8	0.84	0.96
GCDisease-1600	0.86	0.78	0.7	0.91	0.81	0.97
GCDisease-1700	0.71	0.77	0.86	0.92	0.86	0.97
GCDisease-1800	0.82	0.74	0.84	0.9	0.81	0.96
GCDisease-1900	0.71	0.83	0.83	0.83	0.81	0.98
GCDisease-2000	0.87	0.86	0.71	0.91	0.85	0.97

Table 4, describes the performance of recall ofgene-disease classification using the proposed classification framework on large datasets. From the table4, it is clearly shown that the present framework has high computational recall as compared to the existing frameworks.

Accuracy

Table 5
Performance analysis of accuracy using different traditional deep learning frameworks

Features_Size	SVM	Random_Forest	CNN	RNN	HNN	Proposed Model
GCDisease-100	0.71	0.88	0.79	0.86	0.89	0.97
GCDisease-200	0.75	0.88	0.85	0.8	0.94	0.97
GCDisease-300	0.86	0.86	0.83	0.91	0.82	0.98
GCDisease-400	0.88	0.87	0.91	0.9	0.94	0.97
GCDisease-500	0.78	0.75	0.73	0.83	0.93	0.97
GCDisease-600	0.81	0.73	0.85	0.93	0.87	0.98
GCDisease-700	0.76	0.88	0.77	0.85	0.9	0.98
GCDisease-800	0.72	0.77	0.8	0.84	0.86	0.98
GCDisease-	0.76	0.81	0.89	0.85	0.86	0.97

900						
GCDisease-1000	0.74	0.86	0.83	0.82	0.82	0.97
GCDisease-1100	0.78	0.87	0.8	0.82	0.88	0.98
GCDisease-1200	0.81	0.84	0.71	0.82	0.87	0.97
GCDisease-1300	0.85	0.81	0.9	0.89	0.81	0.97
GCDisease-1400	0.86	0.81	0.73	0.85	0.89	0.98
GCDisease-1500	0.7	0.86	0.79	0.82	0.85	0.97
GCDisease-1600	0.87	0.8	0.76	0.85	0.91	0.97
GCDisease-1700	0.76	0.72	0.74	0.91	0.93	0.98
GCDisease-1800	0.75	0.79	0.74	0.81	0.84	0.97
GCDisease-1900	0.82	0.75	0.75	0.88	0.84	0.97
GCDisease-2000	0.8	0.87	0.83	0.88	0.87	0.98

Table 5, describes the performance of accuracy of gene-disease classification using the proposed classification frameworks on large datasets. From the table5, it is clearly shown that the present framework has high computational accuracy as compared to the existing frameworks.

Conclusion

To find the key feature sets from a vast number of feature space, an ensemble classification technique with weighted function is applied. The proposed model efficiently classifies huge data with high dimensionality because the weights in the deep neural network are optimised using the weighted function and the logistic function. The majority of traditional feature transformation methods, such as log transformation and min-max normalisation, are unaffected by data distribution and outliers. The heuristic activation function and ensemble classification metrics are used to improve traditional PSO and ABC based ensemble learning. To effectively address these issues, it is necessary to create an algorithm that can discover trustworthy disease candidates based on existing gene-disease connections that have been validated through biological experiments. To increase the true positive rate and error rate of disease prediction, the suggested study uses a hybrid feature transformation, hybrid feature selection, and advance classification approach. In order to discover the subset of features for the classification issue, a novel principal component ranking metric is integrated. Finally, on the selected features set, a hybrid decision tree classifier is utilised to predict classification accuracy. The current

framework outperforms existing models for changeable microarray datasets, according to the findings of experiments.

References

1. M. Ghosh S. Begum R. Sarkar D. Chakraborty U. Maulik "Recursive memetic algorithm for gene selection in microarray data" Expert Systems with Applications vol. 116 pp. 172-185 2019.
2. Z. Rustam I. Primasari D. Widya "Classification of cancer data based on support vectors machines with feature selection using genetic algorithm and laplacian score" AIP Conference Proceedings vol. 2023 no. 1 pp. 020234 2018.
3. V.B. Canedo N.S. Marono "A Review of Microarray Datasets and Applied Feature Selection Methods" Information Sciences pp. 111-135 2014.
4. Q. Su "A Cancer Gene Selection Algorithm Based on the K-S Test and CFS" Biomed Research International pp. 1-6 2017.
5. M. Morovvat A. Osareh "An Ensemble of Filters and Wrappers for Microarray Data Classification" Machine Learning and Applications: An International Journal (MLAIJ) vol. 3 no. 2 June 2016.
6. N Matamala MT Vargas R González-Cámpora R Miñambres et al. "Tumor microRNA expression profiling identifies circulating microRNAs for early breast cancer detection" Clin Chem vol. 61 no. 8 pp. 1098-106 Aug 2015.
7. K. Yan L. Ma Y. Dai W. Shen Z. Ji D. Xie "Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis" International Journal of Refrigeration vol. 86 pp. 401-409 2018.
8. H. Lu J. Chen K. Yan Q. Jin Y. Xue Z. Gao "A hybrid feature selection algorithm for gene expression data classification" Neurocomputing vol. 256 pp. 56-62 2017.
9. K. Yan Z. Ji H. Lu J. Huang W. Shen Y. Xue "Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units" IEEE Transactions on Systems Man and Cybernetics: Systems 2017.
10. Y. Liu H. Lu K. Yan H. Xia C. An "Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification" Computational intelligence and neuroscience 2016.
11. C. Braicu D. Gulei B. De Melo Maia I. Berindan-Neagoe G. A. Calin "Mirna expression assays" in Genomic Applications in Pathology Springer pp. 65-92 2019.
12. T. Setoyama H. Ling S. Natsugoe G. A. Calin "Non-coding rnas for medical practice in oncology" The Keio journal of medicine vol. 60 no. 4 pp. 106-113 2011.
13. M. Ghosh S. Begum R. Sarkar D. Chakraborty U. Maulik "Recursive memetic algorithm for gene selection in microarray data" Expert Systems with Applications vol. 116 pp. 172-185 2019.
14. J. Krawczuk T. Łukaszuk "The feature selection bias problem in relation to high-dimensional gene data" Artif. Intell. Med. vol. 66 pp. 63-71 2016.
15. H. Öztoprak M. Toycan Y.K. Alp et al. "Machine-based classification of ADHD and non-ADHD participants using time/frequency features of event-related neuroelectric activity" Clin. Neurophysiol. vol. 128 no. 12 pp. 2400-2410 2017.

16. P. Viday Sagar, Nageswara Rao Moparthi, Ch. Mukesh “Smart Meter Analytics for Optimizing the Utilization of Electricity using Arima, Navie & Holt Winter” International Journal of Innovative Technology and Exploring Engineering Vol 8, PP 585-590 (2019)