International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

PMP-SVM: A Hybrid Approach for effective Cancer Diagnosis using Feature Selection and Optimization

Pinakshi Panda[1*], Sukant Kishoro Bisoy[2], Jyotsnarani Tripathy[3], Manmath Nath Das[4*]
[1]Department of Computer Science & Engineering
C. V. Raman Global University
Bhubaneswar,Odisha, India
e-mail: pinakshipanda@gmail.com
[2]Department of Computer Science & Engineering
C. V. Raman Global University
Bhubaneswar,Odisha, India
e-mail: sukantabisoyi@cgu-odisha.ac.in
[3]Department of CSE-AIML & IoT
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, Telangana, India
e-mail: jtjyotsna@gmail.com
[4]Department of AI&DS
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, Telangana, India
e-mail: manmathnath.das@gmail.com

[*]**Corresponding author's Email:** manmathnath.das@gmail.com, pinakshipanda@gmail.com

Abstract- Cancer disease is becoming a prominent factor in increasing the death ration over the world due to the late diagnosis. Machine Learning (ML) is playing a vital role in providing computer aided diagnosis models for early diagnosis of cancer. For the diagnosis process the microarray data has its own place. Microarray data contain the genetic information of a patient with a large number of dimensions such as genes with a small sample such as patient details. If the microarray is directly taken without reducing the dimension as the input to any ML model for classification, then Small Sample Size is the resulting issue. So, size of the microarray data needs to be reduces by using either of dimensionality reduction technique or the feature selection technique to increase the model's performance. In this work, proposed a hybrid model using Principal Component Analysis (PCA), Maximum Relevance Minimum Redundancy (MRMR), Particle Swarm Optimization (PSO) and Support Vector Machine (SVM) for cancer diagnosis. PCA and MRMR is used for feature selection and PSO is applied to get the optimized feature set. Finally, SVM is applied as the classification model. The proposed model is evaluated against multiple cancer microarray datasets to measure the performance in terms of accuracy, precision, recall, and F1 score. Result shows that proposed model performs better than existing state of art model.

Keywords- Microarray Data, Machine Learning, PCA, MRMR, PSO

I. INTRODUCTION

Around 18 million people were diagnosed with cancer in 2018, and of them, around 10 million perished because of a lack of a reliable diagnostic system, according to the world health organization and the national cancer institute. Therefore, the classification of cancer has developed into a significant research area. Many researchers have used machine learning to propose various models for classifying cancer. Today, a number of ai-based models are available via machine learning that can detect and classify various cancers. Several models [1,2] have been developed using biopsy information as an input. Biopsy and microarray datasets are two examples of the data sets that may be used to build the ml-based model. However, owing to the limited nature of the information included in the biopsy data, it is challenging to use the data for model construction. If the model needs more genetic information to produce an appropriate model, the developer now has access to it thanks to microarray data [3]. Better accuracy in the detection and categorization of cancer illness has elevated the creation and interpretation of microarray data in recent years [4]. Microarray technology, however, generates massive amounts of genes from a very small number of samples. Some of these genes are irrelevant to how the condition is classified. It is not simple to categorize microarray data because to its large feature count, noise, and computational complexity for its relatively small sample size. To overcome these constraints, several feature selection and classification methods have been developed. However, for all microarray datasets [5], neither feature selection nor classification methods do very well. Therefore, further research is needed into novel hybrid ways to reach an effective outcome. There are a variety of feature selection and classification

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

techniques developed to better classify microarray datasets. The breast, colon, ovarian, and leukemia datasets are used in the majority of the studies [6].

Microarray data and biopsy data are two examples of cancer diagnostic datasets. In the biopsy dataset, only the outcomes of laboratory tests on a certain number of individuals are included; no genetic information is provided. Microarray data is more suitable in constructing cancer disease diagnostic models because of the significance of genetic information in generating an accurate cancer diagnosis. Researchers are turning to microarray ideas to fill in the blanks in their biopsy data in order to diagnose cancer. Researchers benefit from microarray data since it is possible to track the activities of many genes with only one experiment [7]. Gene expression data might be obtained in large quantities from a single experiment utilizing microarrays. It's a fantastic chance for finding out whether a disease has a genetic component. Gene expression data, on the other hand, include high dimensionalities that are unimportant while looking for diseases. And the cancer microarray data set is noisy and not really informative. These features may interfere with correct labelling. It may be challenging to diagnose and classify conditions since there are so many genes but not enough patients (samples) to go around [8]. This is because there is duplication in the data that describes gene expression.

As the dimensionality of the microarray data is large it requires considerable time and effort to process it. Small sample size (SSS) is one of the greatest problems that might arise when the ML approach is applied directly to the model. When applied to raw data, feature selection approaches may overcome this issue. One of two techniques, feature selection and extraction [9], may be used to complete IT.

A. Motivation

Classification algorithms' ability to correctly classify instances depends on the dataset's specified characteristics. Some characteristics in a dataset may improve classification performance, while others may cause inaccurate classification due to irrelevant features. Thus, feature selection strategies reduce features to a manageable number. Reducing the amount of characteristics makes a good classification model cheaper to compute. Reducing features may lower diagnostic test costs in healthcare. First, a genuine medical dataset will likely include duplicates, missing values, noise, and biases owing to non-representative events. Data preprocessing is an important step in handing these outliers in the dataset. Second, these databases are vast and include a variety of information. One may need to choose the finest feature if it comes in many flavors. In case of Cancer diagnosis the dataset has its own foot print. To increase the efficacy while dealing with these kind data the feature selection and feature optimization process puts a vital impact on the diagnosis model's performance. .

B. Contribution

The contribution of this work is summarized below:

- Proposed hybrid model using PCA, MRMR, PSO and SVM.
- PCA and MRMR is used to select the relevant genes from the raw cancer data as the dimensionally is high. Then PSO is applied as the feature optimization algorithm.
- Proposed model is evaluated using different parameters with existing machine learning state-of-the-art algorithms.

C. Paper Organization

The rest of the sections arranged as follows. Section II and III hold the literature survey and background study respectively. Section IV focuses on the proposed methodology and the dataset description which are used during the research work. The empirical analysis is done in Section V. At the end overall conclusion of the work is presented in Section VI.

II. LITERATURE SURVEY

Using recursive feature elimination (RFE) and univariate ranking, the author of [10] developed a penalized logistic regression (PLR) model that may be used to predict outcomes (UR). According to the results of the trials, the suggested model surpasses others in terms of feature selection, test samples, and cross-validation on Microarray datasets with an accuracy level of 98.7 percent compared to the other models. A Bayesian classification and feature selection strategy based on logistic regression has been developed by the author in [11] for classification and feature selection. In past number of tests carried out on different Microarray datasets, including acute leukemia, tiny round blue cell tumors, and hereditary breast cancer, among others. It can be shown from the findings that the suggested model is successful in terms of identifying essential characteristics as well as accuracy in categorization classification. According to the results of the empirical investigation conducted in the relevant work, the greatest degree of accuracy reached was 97.7 %. To identify and predict Microarray data in a prostate cancer dataset, the author [12] used the PLR approach in conjunction with the top score pair (TSP) method. They then compared the results with Lasso, fisher discriminative analysis (FDA), and SVMs. According to the authors, the suggested technique beats previous approaches in terms of classification and prediction accuracy. The accuracy, AUC, and F1-Score of the suggested system have been used to evaluate its overall performance, as well as certain affecting elements. The accuracy level attained in the proposed study was 98.9 percent for Leukemia microarray data, which was the highest ever recorded.

Morais-Rodrigues et al [13] developed a strategy based on logistic regression for breast cancer categorization using a gene expression omnibus (GEO) data set, which they used in the GEO data series. In this suggested model, the authors examined all of the characteristics without decreasing any of them, and they said that it outperformed other models in terms of performance. The author of [13] has given a naive Bayes-based sequential feature extraction model for Microarray data categorization that is based on sequential feature extraction. Some studies were carried out on five microarray datasets, and it was stated that the suggested model had a much-improved performance when compared to the other models, with an accuracy of 99.1 percent. An LNB-MS model to discover biomarkers using a BPSO optimizer to classify cancer using microarray data was developed by Wu et al [15] and used in conjunction with a BPSO optimizer to classify cancer. It has been shown via tests that the suggested model provides the necessary assurances for gene selection.

According to Nagi and Bhattacharya [16], an ensemble approach referred to as SD-EnClass was developed for the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

categorization of Microarray cancer data. The improved classification accuracy obtained through the proposed approach is combined with stacking, bagging, and boosting methods in order to achieve even greater improvements in classification accuracy and overall performance. Based on KNN method and normal PSO technique a gene selection strategy is suggested [17] to differentiate small subset of beneficial genes from the rest of the population. The experiments were conducted using three Microarray datasets, namely, acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL) and, mixed linear leukemia (MLL), and small round blue cell tumor (SRBCT), and the results showed that the proposed model was effective in terms of computing time, the number of informative genes, classification accuracy on blind test samples, and other factors. Modified K-Nearest Neighbors (MKNN) is a classification approach introduced by the author in [18] that includes two Different situations, namely, SMKNN (Smallest) and LMKNN (Largest), in order to improve the performance of the KNN algorithm. The authors used six gene expression datasets, including colon tumors, lung cancer, leukemia cancer, ovarian cancer, lymphoma-DLBCL, and prostate cancer, among others, and compared the results with KNN, Weighted KNN, SVM, Fuzzy-SVM, and brain emotional learning (BEL) based on precision, recall, and classification accuracy parameters, and claimed that MKNN reduces the testing time when compared to other methods. The authors also claimed that MKNN to solve the challenges associated with unbalanced large-scale datasets, Mahfouz et al [19] developed an ensemble classifier constructed from KNN and the RF and GGA optimization techniques. This ensemble classifier was shown to be effective. In their paper, the authors stated that they had achieved improved accuracy when compared to base classifiers on a variety of datasets, including the CNS, Leukemia, Notterman, GDS3257, and Kentridge, among others. In [20], the author introduces a predictive model for the selection of genes or features based on two different techniques such as SVMs and fuzzy preference-based rough set (FPRS) for the selection of genes. Huo et al [21] have presented an SGL-SVM model that is based on the Sparse Group Lasso and the Support Vector Machine, respectively. The authors tested their approach using datasets derived from microarrays and next-generation sequencing (NGS). Based on the results of the studies, it has been shown that it is possible to achieve high classification accuracy on chosen highlighted genes with large dimension and small datasets for tumor.

In [22], the author proposes a novel model, referred to as the modified mutated firefly algorithm with SVM, which is intended to improve accuracy by finding feature subsets and hence improve overall performance. According to the authors, when compared to other methods, this suggested technique outperforms FA, MMFA, MMFA-DT, and MMFA-NB, among others, in terms of performance. Using a hybrid approach called binary biogeography optimization support vector machine recursive feature elimination (BBO-SVM-RFE), the author in [23] has offered a solution for the problems associated with feature selection that is more efficient (FS). Several tests are carried out on 18 benchmark datasets, and the authors claim that the suggested model beats others in terms of the number of features picked as well as accuracy, compared to other approaches. As a dimensionality reduction strategy, the author in [24] employed a Genetic Algorithm (GA) optimized minimal redundancy maximum relevance (MRMR) method as well as the C4.5 DT as a classification algorithm. The authors conducted their research using five microarray datasets, which included colon tumors, lung cancer, leukemia cancer, ovarian cancer, and breast cancer, among others. They concluded that dimensionality reduction techniques are critical in determining cancer classification accuracy and made the following claim:
Previously published research articles on breast cancer classification were evaluated by the author in [25]. These publications were based on the Wisconsin Breast Cancer Dataset (WBCD) and presented methodologies for breast cancer type classification. A combination of Random Forest (RF) and Extra Trees (ET) techniques built on Decision Tree (DT) strategies have been employed for the execution of the suggested methods, which take into account variables such as clump thickness, uniformity of cell size, mitoses and bland chromatin, among others. Following a comparison with other studies, the authors came to the conclusion that the presented methodologies may be included in the race to improve the accuracy of breast cancer categorization. Ram et al [26], implemented RF classifier on Microarray datasets, including Prostate, Leukemia, and Colon cancers among other things. A number of tests were carried out in R software, and the authors obtained accuracy and precision values of 87.39, 73.33, and 100 on colon cancer, prostate cancer, and leukemia cancer datasets based on certain defined key genes. These results were based on some chosen key genes. Abdulla and Khasawneh [27] have presented a new ensemble cost-sensitive feature selection technique, dubbed G-Forest, for the RF induction process, which they call the G-Forest algorithm. Using studies, it has been shown that G-Forest, on average, reduces expenditures by up to 56 percent while simultaneously increasing accuracy by up to 14 percent when compared to alternative ways. Wang et al [28] have presented an enhanced random forest-based rule extraction (IRFRE) technique, which they refer to as the IRFRE method. Based on the DT ensemble technique, this proposed method utilizes three breast cancer datasets: the Surveillance, Epidemiology, and End Results (SEER) dataset, the Wisconsin Original Breast Cancer (WOBC) dataset, and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to produce classification rules that are interpretable and accurate.

III. Background Study

In the current section different aspects of the ML including microarray data and different state-of-the-art algorithms are discussed.

A. Microarray Data

Microarray data deals with the genetic information of a patient. The main characteristics of the microarray dataset is the high dimensionality. With the advent of the microarray technology, the researchers are able to create a change in different disease diagnoses and prognoses in contrast to the non-genetic dataset or biopsy dataset. The microarray data contains various genetic information which can play a vital role in disease diagnosis. The main issue present with this dataset is the high dimensionality in terms of attributes. The below discussed things are the most renowned research challenges present behind the microarray data technology.

- Small Sample Size: In terms of patient information, microarray data include a large number of attributes as compared to the number of samples. As a result, in the case

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

of classification, the low sample size may be a concern. When the classification algorithms are applied, the result would be low accuracy. So, to increase the performance level in terms of accuracy level the dimension needs to be scaled down.

- Class Variability: This is one more well-known issue present behind the microarray data concept where some of the classes will act as the major class and others will be the minor one. The major class will contain a huge number of the attributes and samples as compared to the others which can result in a class imbalance which is also responsible for degrading the performance in terms of accuracy. The class imbalance can be categorized into two different kinds such as the oversampling and under-sampling and data pre-processing is the only solution for both the issues
- Data Outlier: All of the attributes present in the microarray dataset are not used some of them can degrade the performance and those are called the outliers. These outliers can be removed by introducing different feature selection and extraction techniques.
- Data Redundancy: In microarray data, the dimension is very high with the presence of redundant attributes. These attributes could affect the algorithmic performance in case of the classification and prediction. Hence before going to the classification or prediction problem the redundant attributes can be deleted and it can be done in the data pre-processing step.

B. Machine Learning Algorithms

1) Neural Network:

In this case, the organic neural network of the human brain serves as the inspiration for a kind of information processing system. ANN provides a mathematical model of the neuron found in the human brain. A node with an activation function may stand in for a single neuron. Values are applied as input to the node, and those values are multiplied by the corresponding weights before being added to the total. The total is sent to the output only if it is more than a predetermined limit. Neural networks are large-scale networks of neurons used for data analysis and pattern recognition. Classification and cluster analysis are two of its primary applications. In this setup, the algorithm may self-improve by having its weights and biases adjusted automatically.

2) K Nearest Neighbors (KNN):

It is a kind of machine learning algorithm that records every single example that is available and then classifies new cases based on how closely they are comparable to the stored cases. This approach is simple to implement since it does not need the building of the model or the fine-tuning of its parameters.

3) Decision Tree (DT):

The decision tree algorithm is a kind of supervised machine learning method used mostly for classifying characteristics and their related values into distinct buckets. The central node of a decision tree represents an attribute test, the leaf nodes represent the class labels, and the branch represents conjunctions of attributes that correspond to those class labels, much like a flowchart. The criterion for sorting things is the whole route from the trunk to the branches.

IV. POPOSED WORK

This section deals with the dataset description and proposed hybrid method in details.

A. Description of Microarray Dataset

Four different types of Cancers such as Colon Cancer, Breast Cancer, Prostate Cancer., and Lung Cancer are used for this work. The information of different cancer type used is shown in Table 1.

TABLE I. MICROARRAY DATASET USED

Cancer Name	Sample	Attributes	Class
Lung	62	2178	2
Colon	63	2002	2
Prostate	103	341	2
Breast	32	570	2

B. Methodology

In this world PCA and MRMR have been used as the feature selection approach. These strategies for dimensionality reduction are carried out in stages. At first, the PCA is applied on the rough set, which was regarded the input. Then the featured set undergoes one more feature selection algorithm known as MRMR or Maximum Relevance Minimum Redundancy. The objective to apply two feature selection techniques to find best features which can be considered for the classification for providing high performance. PSO algorithm is implemented as the optimization algorithm. Finally Support Vector Machine (SVM) is applied with the classification as an objective.

1) Support Vector Machine (SVM):

The supervised Machine Learning technique known as a Support Vector Machine (SVM) [5] Classification and regression are helped by this method. A maximum-margin ultimate hyperplane may be defined via calculation. The existence of these support vectors characterizes this hyper plane. You can achieve a better hyperplane with less training samples. A reduced number of training samples is employed to attain better classification accuracy.

2) Principal Component Analysis (PCA):

The key concept present behind the PCA is to reduce the high dimension of the microarray data. This can be achieved by forming the featured spaces known as principal components (PCs) which are not correlated to each other. The working procedure of the PCA is described as follows [3].

- Calculate the mean of each dimension of the dataset.
- Calculate the covariance matrix.
- Calculate the eigenvector and its corresponding eigenvalue.
- Choose K highest eigenvalue to form the PC.

Lowering the PC population will help in increasing the accuracy. A high PC will result in low efficiency as compared to a low PC.

3) Maximum Relevance Minimum Redundancy (MRMR):

It is a feature selection technique which favors characteristics having a high correlation with the class (output) and a low correlation with one another. The Pearson correlation coefficient may be used to calculate the correlation between

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

features and the F-statistic can be used to calculate the correlation with the class (relevance) for continuous features (redundancy). Then, using a greedy search to maximize the objective function, which is a function of relevance and redundancy, features are chosen one by one. The MID (Mutual Information Difference criteria) and MIQ (Mutual Information Quotient criterion) are two forms of objective functions that reflect the difference or quotient of relevance and redundancy, respectively. The MRMR feature selection strategy for temporal data necessitates various pre-processing approaches that flatten temporal data into a single matrix ahead of time. This might lead to the loss of potentially crucial information within temporal data.

The main goal is to use mutual information (MI) to determine the largest dependence between a collection of characteristics X and class c. Once the marginal probabilities p(m) and p(n) and the joint probability p(m, n) between these two characteristics are known, the MI between them may be calculated successfully. MI can be calculated as in Equations 1 and 2.

$$MI(m,n) = \sum_{n\in N}\sum_{m\in M} p(m,n)\log\left(\frac{p(m,n)}{p(m)p(n)}\right) \dots\dots\dots\dots\dots\dots (1)$$

$$R_d = \frac{1}{|X|^2}\sum_{m_i m_j} MI(m_{;}; m_j) \dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Where R_d = Redundancy Measure, X is the total number of features, MI is the mutual information between i and j feature. Minimum R_d must be chosen. The MI value between the feature and the target action is used to calculate the relevance measure. If the MI value is low, it suggests that the characteristic and the target action have a poor relationship. Relevance can be measured as follows in equation 3:

$$R_l = \frac{1}{|X|}\sum_{m_i m_j} MI(m_{;}; m_j) \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Where R_l is the Relevance measure, X is the total number of features, and MI is the mutual information between i and j feature. Maximum R_l must be chosen. The steps of the mRMR features are given as follows:

Step 1: For features X {X1,X2,X3,…..Xn}

Step 2: Calculate $R_l = MI(Xi, C)$ where C is the class.

Step 3: Initiate $R_d = 0$

For every feature X

$R_d += MI\ (X_{;}; X_j)$

Step 4: MRMR [Xi]= $R_l - R_d$

Step 5: Feature set = sort (MRMR [Xi])

4) Particle Swarm Optimization (PSO):

The evolutionary approach known as particle swarm optimization (PSO) takes cues from occurrences like bird swarming to get a desirable outcome. The initial population in PSO is a chaotic swarm. Each swarm travels in the selected D-dimensional search space. Even when in motion, each particle stays precisely where it should be. With a maximum speed of Spmax, the current location of each individual vector is denoted by the vector M = M1, M2, M3,.....,Mn. Within the defined search area, each swarm is permitted to move at a speed between [Spmin, Spmax]. The velocity of a particle in the swarm is a crucial factor in this optimization process. The velocity (v) of a particle i at (t+1)th iteration can be defined as follows.

$$v_i^{t+1} = v_i^t + \alpha_1\omega_1(L_i^t - x_i^t) + \alpha_2\omega_2(G^* - x_i^t) \dots\dots\dots\dots (4)$$

α_1 and α_2 are the acceleration coefficient of the particle. ω_1 and ω_2 are uniformly doistributed random numbers. L_i^t is the local best of the particle i at t[th] iteration. G^* is the global best of the swarm. The position of a particle (i) at tth iteration can be defied as

$$x_i^{t+1} = x_i^t + v_i^{t+1} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

Algorithm 1: Pseudocode for the proposed work.

REQUIRE: Datasets D ← {D1, D2,D3,D4}, Feature Set (F) ← {f_1, f_2,….., f_n}
OUTPUT: Performance measures ← { Ac,Pr,Sn,Sp,Er,Fv}
D_i∈D, apply data scaling
for D_i∈D, apply PCA
  for D_i∈D, apply MRMR
    $D_{featured} \leftarrow f_j$
  end for
end for
Apply PSO () to $D_{featured}$
  Initialize particle population ← $\{X_1, X_2, \dots X_i\}$, v_i, L_i^t, G^*
  Define objective function F_{min}()
  Find G^* in k
  while (t < max_iteration**)**
    for j ← 1 to i
      Calculate v_i^{t+1} using equation 4
      Calculate x_i^{t+1} using equation 5
      Update G^*
    end for
  end while

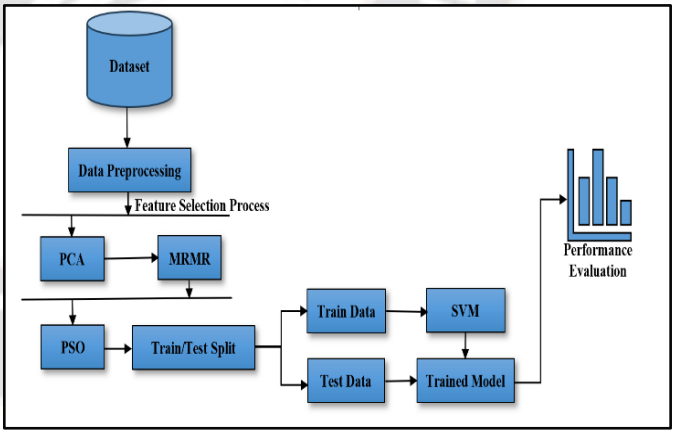

Figure 1. Proposed Model

V. Result and Discussion

Proposed work is implemented using a system having i3 processor with 8GB RAM, 1TB HDD, 256GB SSD. To the raw dataset the feature reduction technique PCA is applied to reduce the number of feature selection algorithm has been applied. After applying PCA the MRMR has again been applied as the feature selection method to obtain the featured dataset. To the featured dataset the PSO algorithm has been applied as the optimizer for obtaining the optimized dataset upon which

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

the SVM algorithm is applied as the classifier. Table 2 shows the exact number of features that has been selected at each stage. Table 3 shows empirical analysis of the proposed model. Classification Accuracy (CA), Sensitivity, Specificity, F1 score, Mathews Correlation Coefficient (MCC), and Precision have been considered as the evaluation criteria which can be defined as equation 6-11. Figure 2, and 3 show the MCC and Precision comparison of the proposed system in contrast to some state-of-the-art ML algorithms.

TABLE II. DIMENSIONALITY AFTER FEATURE REDUCTION AND OPTIMIZATION

Dataset	After PCA	After MRMR	After PSO
Lung	1317	876	234
Colon	1090	671	187
Prostate	208	116	51
Breast	319	179	91

$$CA = \frac{Tr_{Pos} + Tr_{Neg}}{Tr_{Pos} + Tr_{Neg} + Fl_{Pos} + Fl_{Neg}} \quad \text{.....(6)}$$

$$Sensitivity = \frac{Tr_{Pos}}{Tr_{Pos} + Fl_{Neg}} \quad \text{.....(7)}$$

$$Specificity = \frac{Tr_{Neg}}{Tr_{Neg} + Fl_{Pos}} \quad \text{.....(8)}$$

$$F1 = 2\,x\,\frac{\left(\frac{Tr_{Pos}}{Tr_{Pos} + Fl_{Pos}} * \frac{Tr_{Neg}}{Tr_{Neg} + Fl_{Pos}}\right)}{\frac{Tr_{Pos}}{Tr_{Pos} + Fl_{Pos}} + \frac{Tr_{Neg}}{Tr_{Neg} + Fl_{Pos}}} \quad \text{..... (9)}$$

$$MCC = \frac{(Tr_{Pos} + Tr_{Neg})(Fl_{Pos} + Fl_{Neg})}{\sqrt{(Tr_{Pos} + Fl_{Pos})(Tr_{Pos} + Fl_{Neg})(Tr_{Neg} + Fl_{Pos})(Tr_{Neg} + Fl_{Neg})}} \quad \text{..... (10)}$$

$$Precision = \frac{Tr_{Pos}}{Tr_{Pos} + Fl_{Pos}} \quad \text{..... (11)}$$

Where Tr_{Pos} , Tr_{Neg}, Fl_{Pos}, *and* Fl_{Neg} is true positive, true negative, false positive and false negative, respectively. Performacne of different models are presented below in Table 3.

TABLE III. PERFORMANCE ANALYSIS OF THE PROPOSED WORK

Dataset	Model	CA (%)	F1 (%)	Specificity (%)	Sensitivity (%)
Lung Cancer	KNN	0.925	0.921	0.923	0.925
	Naïve Bayes	0.861	0.862	0.866	0.861
	Neural Network	0.878	0.879	0.878	0.884
	Decision Tree	0.845	0.842	0.844	0.845
	Logistic Regression	0.878	0.878	0.878	0.878
	Proposed	**0.988**	**0.987**	**0.988**	**0.988**
Colon Cancer	KNN	0.85	0.851	0.853	0.85
	Naïve Base	0.861	0.862	0.866	0.861
	Neural Network	0.878	0.879	0.878	0.884
	Decision Tree	0.845	0.842	0.844	0.845
	Logistic Regression	0.883	0.883	0.883	0.884
	Proposed	**0.967**	**0.966**	**0.961**	**0.967**
Pro-state Cancer	KNN	0.861	0.862	0.862	0.861
	Naïve Base	0.856	0.867	0.862	0.866
	Neural Network	0.878	0.879	0.878	0.886
	Decision Tree	0.839	0.835	0.84	0.839
	Logistic Regression	0.918	0.91	0.901	0.901
	Proposed	**0.972**	**0.971**	**0.973**	**0.972**
Breast Cancer	KNN	0.85	0.851	0.851	0.85
	Naïve Base	0.85	0.852	0.858	0.85
	Neural Network	0.878	0.879	0.878	0.887
	Decision Tree	0.834	0.833	0.832	0.834
	Logistic Regression	0.908	0.91	0.908	0.908
	Proposed	**0.95**	**0.949**	**0.949**	**0.95**

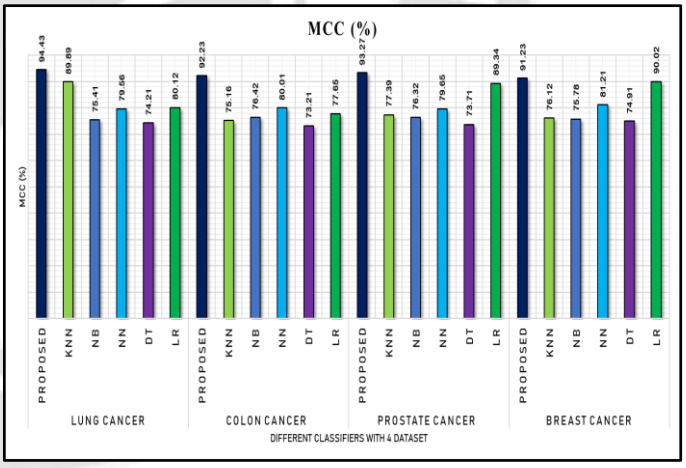

Figure 2. Mathew's Correlation Coefficient (MCC) of diffent algorithms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

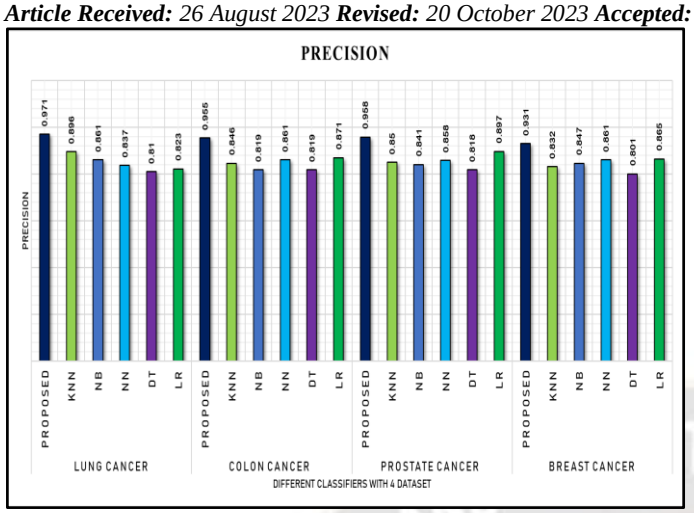

Figure 3. Precision value of diffent algorithms

The empirical analysis shows that the propose model gets highest accuracy level 98.8% with F1-Score, Specificity and Sensitivity values 98.7%, 98.8%, and 98.8% , respectively for Lung Cancer.

VI. CONCLUSION

For effective cancer disease classification, the microarray data plays a vital role. However, dealing with the microarray data creates a lot of issues including the small sample size issue. To avoid the issues the feature selection and feature optimization techniques came into front as an emerging solution. The current research work four different types of cancer datasets are considered such as Lung Cancer, Colon Cancer, Prostate Cancer, and Breast Cancer. To the dataset the PCA and MRMR algorithms are used for the feature selection algorithm. Then to reduced feature set PSO algorithm is used to determine the optimized feature set from the dataset. Finally, the SVM algorithm is used for classification purpose. The performance of the proposed work is compared to different machine learning state-of-the-art algorithms. The empirical analysis of this research work shows that the proposed model outperforms other algorithm with an accuracy of 98.8% in case of the Lung Cancer.

REFERENCE

[1] K. Kanti Ghosh et al., "Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data," Expert Systems with Applications, vol. 169, p. 114485, May 2021, doi: https://doi.org/10.1016/j.eswa.2020.114485.

[2] M. Takamiya, K. Saigusa, and K. Dewa, "DNA microarray analysis of hypothermia-exposed murine lungs for identification of forensic biomarkers," Legal Medicine, vol. 48, p. 101789, Feb. 2021, doi: https://doi.org/10.1016/j.legalmed.2020.101789.

[3] A. Mirzal, "Statistical Analysis of Microarray Data Clustering using NMF, Spectral Clustering, Kmeans, and GMM," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–1, 2020, doi: https://doi.org/10.1109/tcbb.2020.3025486.

[4] K. Kourou, G. Rigas, C. Papaloukas, M. Mitsis, and D. I. Fotiadis, "Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks," Computers in Biology and Medicine, vol. 116, p. 103577, Jan. 2020, doi: https://doi.org/10.1016/j.compbiomed.2019.103577.

[5] Y. Peng, "A novel ensemble machine learning for robust microarray data classification," Computers in Biology and Medicine, vol. 36, no. 6, pp. 553–573, Jun. 2006, doi: https://doi.org/10.1016/j.compbiomed.2005.04.001.

[6] Y. Chen and Y. Zhao, "A novel ensemble of classifiers for microarray data classification," Applied Soft Computing, vol. 8, no. 4, pp. 1664–1669, Sep. 2008, doi: https://doi.org/10.1016/j.asoc.2008.01.006.

[7] J.-H. Hong and S.-B. Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming," Artificial Intelligence in Medicine, vol. 36, no. 1, pp. 43–58, Jan. 2006, doi: https://doi.org/10.1016/j.artmed.2005.06.002.

[8] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, vol. 16, no. 10, pp. 906–914, Oct. 2000, doi: https://doi.org/10.1093/bioinformatics/16.10.906.

[9] A. Sharma and K. K. Paliwal, "Cancer classification by gradient LDA technique using microarray gene expression data," Data & Knowledge Engineering, vol. 66, no. 2, pp. 338–347, Aug. 2008, doi: https://doi.org/10.1016/j.datak.2008.04.004.

[10] J. Zhu, "Classification of gene microarrays by penalized logistic regression," Biostatistics, vol. 5, no. 3, pp. 427–443, Jul. 2004, doi: https://doi.org/10.1093/biostatistics/kxg046.

[11] X. Zhou, K.-Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," Journal of Biomedical Informatics, vol. 37, no. 4, pp. 249–259, Aug. 2004, doi: https://doi.org/10.1016/j.jbi.2004.07.009.

[12] H. Zhao, S. Qi and Q. Dong, "Predicting prostate cancer progression with penalized logistic regression model based on co-expressed genes," 2012 5th International Conference on BioMedical Engineering and Informatics, Chongqing, China, 2012, pp. 976-980, doi: 10.1109/BMEI.2012.6512948.

[13] . Morais-Rodrigues et al., "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression," Gene, vol. 726, p. 144168, Feb. 2020, doi: https://doi.org/10.1016/j.gene.2019.144168.

[14] L. Fan, K.-L. Poh, and P. Zhou, "A sequential feature extraction approach for naïve bayes classification of microarray data," Expert Systems with Applications, vol. 36, no. 6, pp. 9919–9923, Aug. 2009, doi: https://doi.org/10.1016/j.eswa.2009.01.075.

[15] M. -Y. Wu, D. -Q. Dai, Y. Shi, H. Yan and X. -F. Zhang, "Biomarker Identification and Cancer Classification Based on Microarray Data Using Laplace Naive Bayes Model with Mean Shrinkage," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 6, pp. 1649-1662, Nov.-Dec. 2012, doi: 10.1109/TCBB.2012.105.

[16] S. Nagi and D. Kr. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 2, no. 3, pp. 159–173, Jun. 2013, doi: https://doi.org/10.1007/s13721-013-0034-x.

[17] S. Kar, K. Das Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," Expert Systems with Applications, vol. 42, no. 1, pp. 612–627, Jan. 2015, doi: https://doi.org/10.1016/j.eswa.2014.08.014.

[18] S. M. Ayyad, A. I. Saleh, and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," Biosystems, vol. 176, pp. 41–51, Feb. 2019, doi: https://doi.org/10.1016/j.biosystems.2018.12.009.

[19] M. A. Mahfouz, A. Shoukry, and M. A. Ismail, "EKNN: Ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis," Artificial Intelligence in Medicine, vol. 111, p. 101985, Jan. 2021, doi: https://doi.org/10.1016/j.artmed.2020.101985.

[20] U. Maulik and D. Chakraborty, "Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification," in IEEE Transactions on NanoBioscience, vol. 13, no. 2, pp. 152-160, June 2014, doi: 10.1109/TNB.2014.2312132.

[21] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso," Journal of Theoretical Biology, vol. 486, p. 110098, Feb. 2020, doi: https://doi.org/10.1016/j.jtbi.2019.110098.

[22] N. Almugren and H. M. Alshamlan, "New Bio-Marker Gene Discovery Algorithms for Cancer Gene Expression Profile," in IEEE Access, vol. 7, pp. 136907-136913, 2019, doi: 10.1109/ACCESS.2019.2942413.

[23] D. Albashish, A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, "Binary biogeography-based optimization based SVM-RFE for feature selection," Applied Soft Computing, vol. 101, p. 107026, Mar. 2021, doi: https://doi.org/10.1016/j.asoc.2020.107026.

[24] Irne Mabarti, "Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4.5 Decision Tree," Journal of Data Science and Its Applications, vol. 3, no. 1, pp. 38–47, May 2020, doi: https://doi.org/10.34818/jdsa.2020.3.37.

[25] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," Computers in Biology and Medicine, vol. 128, p. 104089, Jan. 2021, doi: https://doi.org/10.1016/j.compbiomed.2020.104089.

[26] M. Ram, A. Najafi, and M. T. Shakeri, "Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest," Iranian Journal of Pathology, vol. 12, no. 4, pp. 339–347, Dec. 2017, doi: https://doi.org/10.30699/ijp.2017.27990.

[27] M. Abdulla and M. T. Khasawneh, "G-Forest: An ensemble method for cost-sensitive feature selection in gene expression microarrays," Artificial Intelligence in Medicine, vol. 108, p. 101941, Aug. 2020, doi: https://doi.org/10.1016/j.artmed.2020.101941.

[28] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," Applied Soft Computing, vol. 86, p. 105941, Jan. 2020, doi: https://doi.org/10.1016/j.asoc.2019.105941.