A Hybrid Multiple Indefinite Kernel Learning Framework for Disease Classification from Gene Expression Data

Swetha S[*1], Dr. Srinivasan G N[2], Dr. Dayananda P[3]
Assistant Professor, Department of Information Science and Engineering, RV College of Engineering®, Bengaluru, Karnataka 560059, India[1*]
Professor (Retired), Department of Information Science and Engineering, RV College of Engineering®, Bengaluru, Karnataka 560059, India[2]
Department of Information Technology, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India[3]

***Abstract*—In recent years, Machine Learning (ML) techniques have been used by several researchers to classify diseases using gene expression data. Disease categorization using heterogeneous gene expression data is often used for defining critical problems such as cancer analysis. A variety of evaluated factors known as genes are used to characterize the gene expression data gathered from DNA microarrays. Accurate classification of genetic data is essential to provide accurate treatments to sick people. A large number of genes can be viewed simultaneously from the collected data. However, processing this data has some limitations due to noises, redundant data, frequent errors, increased complexity, smaller samples with high dimensionality, difficult interpretation, etc. A model must be able to distinguish the features in such heterogeneous data with high accuracy to make accurate predictions. So this paper presents an innovative model to overcome these issues. The proposed model includes an effective multiple indefinite kernel learning based model for analyze the gene expression microarray data, then an optimized kernel principal component analysis (OKPCA) to select best features and hybrid flow-directed arithmetic support vector machine (SVM)-based multiple infinite kernel learning (FDASVM-MIKL) model for classification. Flow direction and arithmetic optimization algorithms are combined with SVM to increase classification accuracy. The proposed technique has an accuracy of 99.95%, 99.63%, 99.60%, 99.51%, and 99.79% using the datasets including colon, Isolet, ALLAML, Lung_cancer, and Snp2 graph.**

Keywords*—*Gene expression; optimized kernel principle component analysis; multiple indefinite kernel learning; flow direction algorithm based support vector machine; arithmetic optimization algorithm

I. INTRODUCTION

The integration of data tends to be an emerging topic, whereas decision making based on metabolomics and genomic requires better prediction or diagnosis rather than the utilization of clinical data alone [1]. The prediction or classification of diseases dependent upon the medical data requires appropriate methodologies [2]. Machine learning has been widely playing a huge role, especially in biomedical researchers, over the past decades [3]-[4]. This process is partially because of greater advancements in data collection that have enabled the study of biomedical mechanisms of various diseases, particularly cancer [5]. When the gene expression data are adopted from microarrays comprising high density oligonucleotide arrays (HDOA) or complementary Deoxyribonucleic acid (CDNA), the classification methods are utilized for data examination and interpretation.

Disease classification using heterogeneous gene expression data is greatly utilized for determining fundamental issues like disease analysis and drug detection [6]-[7]. The gene expression data collected from DNA micro arrays are characterized through diverse evaluated variables known as genes [8]. An exact classification of the gene data is very important in order to be able to treat sick people appropriately [9]-[10]. The molecular founded connections over a scale can be analyzed through gene expression data, which can be examined by a significant tool called microarray [11]. A huge amount of genes from the collected data can be observed simultaneously. Still, there are certain drawbacks in processing these data as they comprise noises, redundant data, often prone to errors, increased complexity, smaller sample with large dimensionality, complex interpretation, etc.

Several researches were conducted previously using machine learning approaches to classify diseases using gene expression data [12]-[13]. The major reason behind searching for effective approaches is to predict the survival rates to grab better treatment. The feature selection approaches are highly efficient in eradicating the noisy features, redundant data and are significant in describing the biological features when minimizing the model complexity [14]. The chief focus of the feature selection approach is to reduce the data dimensionality, which improves the overall system performance [15]-[16]. The kernel is generally used to indicate a kernel trick, an approach of utilizing a linear classifier to solve non-linear issues [17]. The Kernel learning transforms linearly inseparable heterogeneous data over separable data. The kernel approaches are better, but parametric assumptions cannot be made and sensible over outliers.

Even though multi-task learning improves accuracy performance, a similarity measure between tasks is highly

required, which is not a priority for many diseases [18]-[19]. The utilization of transcriptomics subjects possesses diverse challenges in interpretation. To overcome this, multiple kernel learning (MKL) permits the combination of pathway data into prediction models that utilize transcriptomics, whereas the interpretation and accuracy can be enhanced. Several studies have been developed through the application of MKL over genomic data. Every MKL method can render an ordering for data type significance that delivers appropriate information [20]. The MKL aims to determine the kernel's best convex integrations to generate the best classifier. Diverse feature components of heterogeneous data with various kernel functions can be mapped to expose the data better in the new feature space.

A. Motivation

Predicting and classifying the disease from gene expression data is an extremely challenging task due to the intrinsic nature of the data. Heterogeneous data involves large differences between traits, making predictions difficult for any learning model. In addition, the size of the data is extremely large, leading to several complications. A prominent solution identified to this problem is using meta-heuristics that can optimally tune the parameters, resulting in higher accuracy. In order to achieve better performance of multiple kernel learning, machine learning models are generally adopted, among which the SVM model is highly preferred. Considering all the problems, this proposal focuses on developing a hybrid multi-core learning framework with the hybridization of an effective meta-heuristic with an SVM model to achieve a higher percentage of accuracy in disease classification. The main contributions of the proposed work are:

- To analyze the gene expression microarray data to attain higher accuracy in disease classification, an effective multiple indefinite kernel learning based model is proposed.
- A new approach of optimized kernel principal component analysis (OKPCA) is presented to decrease the dimensions of microarray data by eliminating the unimportant features from the feature space.
- In order to achieve higher accuracy in disease classification using gene expression data, several kernel functions such as radial basis function, sigmoid kernel, polynomial kernel and linear kernel are introduced and integrated into the hybrid flow-directed arithmetic SVM-based multiple infinite kernel learning (FDASVM-MIKL) classification frameworks.
- In order to validate its efficiency against the existing methods, extensive simulations of the proposed method are performed using different metrics.

The proposed research work is organized into various sections. The literature survey of disease classification from gene expression data directed by various researchers is described in Section II. The discussion of the proposed methodology for disease classification from gene expression data is described in Section III. Section IV discusses the simulations performed using simulation tools to analyze the outcomes of the proposed technique. Finally, the conclusion and the future scope of the proposed technique are provided in Section V with references.

II. Related Works

Most researchers have applied various methods to reduce dimensionality across heterogeneous gene expression data. Some of the prominent adopted models are examined as follows.

Liu et al. [21] presented a dimension reduction algorithm to enhance the classification performance and minimize the dimensionality. The dimensionality minimization was carried through a Weighted Kernel Principal Component Analysis (WKPCA) that builds the weights of the kernel function in accordance with the kernel matrix Eigenvalues. The feature dimensions were minimized through the multiple kernel functions combination. The t-class kernel functions were built to further enhance the efficacy of dimensional reduction. The classifiers like random forest, naive Bayes and Support Vector Machine were used to examine six real gene expression datasets. The major limitation faced in this approach was the non-flexibility of kernel function selection and the degraded embedding ability.

Rahimi *et al.* [22] presented a multi-task multiple KL approach with task clustering and developed a greatly time-effective solution. The proposed solution approach in this research was dependent upon the benders decomposition and clustering issue treatment through the determination of given tree structures in the graph. The method is called forest formulation and has been used to differentiate early and late stage cancers through the adoption of gene sets and genomic data. When the number of tasks and clusters gets maximized, the forest formulation approach is highly favorable due to computational performance. The time consumed in solving large scale instances was too high in the case of a multi-task multiple KL approach and clustering.

The microarray data was utilized for training deep learning approaches using extracted features. Almarzouki et al. [23] established an effective feature selection approach to maximize accuracy and minimize the classification time. The most significant genes were picked by eradicating the superfluous and duplicate information. Artificial Bee Colony (ABC) method using bone marrow pyruvate carboxylase gene expression data was employed in this research work. The features selected using the ABC algorithm were made as a wrapper based features selection system. The datasets of lung, kidney and brain cancer were utilized during testing and training. The characteristics of data were not effectively examined, and there was a greater possibility of losses.

The seven cancer datasets were collected initially from the Broad Institute GDAC Firehose comprising of isoform expression profile, survival information, gene expression profile and expression data of DNA methylation, respectively. Feng et al. [24] recommended kernel principal component analysis (KPCA) to extract the relevant features for every expression profile. The features are then fed over three similar kernel matrices through a Gaussian kernel function combined as a global kernel matrix. Finally, the features were applied over the spectral clustering algorithm to obtain clustering

outcomes. Due to the collection of abundant datasets, the dimensionality issue was not solved effectively.

To overcome the limitation of the increased computational effort of using a huge data set, Wani et al. [25] proposed an efficient method. The MKL founded gene regulatory network (GRN) inference method was presented in this research in which numerous heterogeneous datasets were combined together using the MKL paradigm. The GRN learning issue has been formulated as a supervised classification issue in which the genes are directed through a specified transcription factor differentiated from other non-regulated genes. In order to learn a huge scale GRN, a parallel execution construction was devised. Better accuracy rates and speedups can be obtained, but the data quality and redundancy issues were not solved effectively. Table I describes the major contribution with its corresponding merits and demerits of existing methods.

TABLE I. REVIEW OF EXISTING METHODS WITH THEIR MERITS AND DEMERITS

Author name and Reference	Technique	Contribution	Merits	Demerits
Liu *et al.* [21]	WKPCA	To develop a dimension reduction algorithm for enhancing the classification performance.	Minimization of dimensionality with less computational complexity.	Non-flexible and degraded embedding capability.
Rahimi *et al.* [22]	Forest formulation method	To distinguish late and early stage cancers using genomic data.	Better computational performance in aggregation.	Higher consumption of time.
Almarzouki *et al.* [23]	ABC algorithm	To develop an effective feature selection approach to eliminate the superfluous and duplicate information	The computational time can be minimized by using relevant features.	Ineffective characterization of gene data and high error possibilities.
Feng et al. [24]	KPCA	To extract the relevant features and obtain clustering results using a spectral clustering algorithm	An effectual global kernel matrix can be attained.	Greater dimensionality issue due to huge dataset.
Wani *et al.* [25]	MKL with GRN	To combine numerous datasets using the MKL paradigm and to revise a parallel execution framework.	The computational cost can be minimized.	Data quality and redundancy problems cannot be solved.

Based on the related work of the existing methods according to disease classification using gene expression data, various limitations negatively impact the performance of the method. Due to the disadvantages, such as the use of lesser datasets, non-flexible embedding capability, higher consumption of time, lower data quality, redundancy problem and lower classification accuracy, an effective proposed FDASVM-MIKL approach is developed. The additional limitations such as noises, redundant data, frequent errors, increasing complexity, smaller samples with high dimensionality, difficult interpretation, etc. This limitation can be reduced in appropriate feature extraction method, here in this research OKPCA proposed for analyzing expression data based on multiple indefinite kernel learning with and hybrid flow directed arithmetic SVM for gathering effective results to overcome the current limitations and increase classification accuracy.

III. PROPOSED METHODOLOGY

The heterogeneous nature of the data is a very difficult to process, predict and classify the disease utilizing gene expression data. The fact that the gene expression data are quite heterogeneous has been identified as one of the main challenges. In general, heterogeneous data have a wide range of feature variations, making it difficult for any learning model to make predictions. A model must be able to accurately distinguish between the characteristics of heterogeneous data in order to make reasonable predictions. Utilizing meta-heuristics to optimize the tuning of the parameters and increase accuracy has been proposed as one prominent solution to this problem. The support vector machine (SVM) model is highly preferred among the machine learning models when the task of multiple kernel learning is accomplished. This approach aims to consider all issues and create a hybrid multiple-kernel learning framework that combines an efficient meta-heuristic with an SVM model to increase the accuracy of disease classification. The Fig. 1 demonstrates the overall architecture for the proposed methodology.

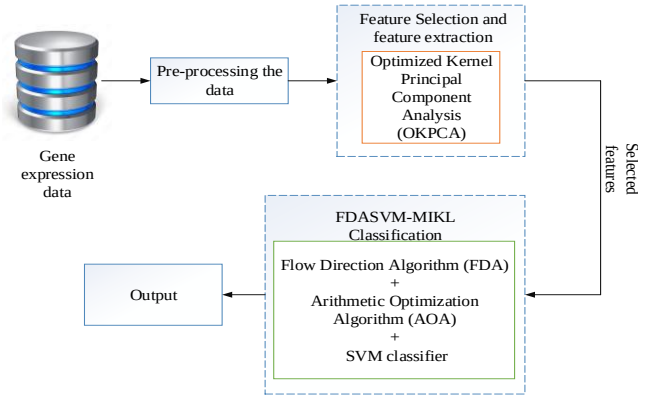

Fig. 1. Overall architecture of the proposed method.

The proposed framework comprises three major steps pre-processing, feature selection and classification. The dataset is initially brought through various processes to make it suitable for classification because it is diverse and has a large dimension. To improve the quality of the dataset, the outliers

are first eliminated. The missing values are then filled with mean values once the dataset has been examined for any missing values. After this step, the data set is then passed to the feature selection phase, in which the main features are extracted. The proposed study introduces the optimized kernel principal component analysis to identify the dataset's best features (OKPCA). The proposed hybrid flow directed arithmetic SVM based multiple indefinite kernel learning (FDASVM-MIKL) classification framework is then given the features chosen using the OKPCA technique. This proposed framework is associated with the SVM classifier, flow direction algorithm (FDA), and Arithmetic Optimization Algorithm (AOA). To increase overall performance, the SVM framework incorporates some kernels, containing the linear kernel, sigmoid kernel, polynomial kernel, and radial basis function.

A. Pre-processing

Pre-processing is a very essential step utilized to provide data cleansing that is useful for further analysis. Here, a standard pre-processing method is used. The dataset is then examined for non-value reduction processes, and the missing values are then filled with mean values. Outliers are removed to increase the quality of the data set. The dataset is then given to the feature selection stage, where the main features are extracted after this step.

B. Feature Selection and Extraction

The proposed study introduces the optimized kernel principal component analysis to identify the dataset's best features (OKPCA). This technique chooses the most important features from the dataset and ignores the rest, designed to decrease the dimensionality of the data.

1) Optimized kernel principal component analysis: Principal component analysis, which finds recurring patterns in the dataset with little information loss, is frequently used to reduce complex spectral datasets into understandable information. The strength and flexibility of principal component analysis are greatly enhanced by its clarity and conciseness. The important factor when using PCA is that it is a linear transformation and it can be written in the simple matrix form, which is given below:

$$B = HA \tag{1}$$

Where, B is a transformed data matrix, A is an original data matrix, and H is a transformation matrix. The corresponding eigenvectors $v_i, 1 \le i \le n$, and the necessarily non-negative eigenvalues (v_i) arranged in decreasing order. The transformation matrix is obtained by stacking the eigenvectors which is shown in equation (2).

$$H = \begin{bmatrix} \leftarrow \overline{v_1} \rightarrow \\ \vdots \\ \leftarrow \overline{v_n} \rightarrow \end{bmatrix} \tag{2}$$

The enhanced version of PCA, known as kernel principal component analysis (KPCA), can handle non-linear correlations between variables. It employs a non-linear function to translate the observed data into high dimensional space (kernel function). KPCA uses a non-linear mapping function $\omega(x)$ to translate an observed data matrix $datametics \in R^{M \times N}$ with N columns (variables) and M rows (observations). This can be calculated mathematically as shown in equation (3),

$$datametics \in R^{M \times N} \rightarrow \omega(x) \in R^{f} \tag{3}$$

Where f is the feature space. In kernel technique, kernel function and kernel matrix are known as $\mathrm{K}(x_i, x_j) = \omega(x_i)^T (x_j)$ *and* K, respectively. The appropriate kernel parameter is best discovered using this kernel-based strategy by generalizing the problem to an eigenvector one. The scatter error metric is used as the objective function of the problem.

The defined goal function computes the gradient and Hessian matrices, and the kernel parameter of the method is tweaked using the gradient values. Gradient (∇f) and Hessian $\left(\nabla^2 f\right)$ matrices are the popular optimization technique, as the search direction as it approaches the optimum that moves in the opposite direction of the positive gradient. Gradient and Hessian matrices for optimization are given in the equation below.

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix} \tag{4}$$

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \frac{\partial^2 f}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix} \tag{5}$$

This kernel-based approach best discovers the appropriate kernel parameter by generalizing the problem to an eigenvector. The scatter error metric is used as an objective function to solve the problem. The gradient and Hessian matrices are produced by the defined objective function, and

the gradient values are used to control the algorithm's kernel parameter. The hybrid flow directed arithmetic SVM based multiple indefinite kernel learning (FDASVM-MIKL) classification framework is then given to the features chosen using the OKPCA technique.

C. Classification using FDASVM-MIKL

The proposed hybrid Flow Directed Arithmetic SVM based multiple indefinite kernel learning (FDASVM-MIKL) frameworks are used for classification purposes. This framework combines an SVM classifier, a FDA and AOA which are described in below sections.

1) SVM based multiple indefinite kernel learning: Various kernels are incorporated into the SVM architecture to improve overall performance, containing the polynomial kernel, linear kernel, sigmoid kernel, and radial basis function. The input layer, hidden layer, SVM kernel layers, SVM output layer, and the voting layer form an original SVM based multiple indefinite kernel learning. An additive kernel model enhances the functionality of a standard kernel model. This model is obtained using the weighted linear sum of kernels.

a) Radial basis function (RBF): RBF kernels are the most usually utilized kinds of kernelization due to their similarity to the Gaussian distribution. The degree or similarity of proximity among two points X_1 *and* X_2 is determined by using the RBF kernel function. The mathematical representation of kernel is given in the following equation:

$$K(X_1, X_2) = \exp\left[-\frac{\|X_1 - X_2\|^2}{2\upsilon^2} \right] \quad (6)$$

Where υ is the variance or hyperparameter, and $\|X_1 - X_2\|$ is the Euclidean distance between two points X_1 *and* X_2.

b) Sigmoid kernel: The sigmoid kernel function is an activation function for artificial neurons and is similar to a two-layer perceptron neural network architecture. It is defined as equation (7):

$$K(X_1, X_2) = \tanh\left[\frac{X_1 . X_2 + coeff}{2\upsilon^2} \right] \quad (7)$$

The hyperbolic tangent, $\tanh$ is used to define this kernel. It can express intricate non-linear interactions when utilized with correctly calibrated parameters. However, this does not represent a true kernel since the sigmoid function might not be positive definite for some parameters.

c) Polynomial kernel: The kernel function is used with SVMs and other kernel models is termed as polynomial kernel, to represent the similarity of vectors (training samples) in a feature space via polynomials of the original variables machine learning can be used, allowing the learning of non-linear methods.

$$K(X_1, X_2) = \left[\frac{X_1 . X_2 + coef}{2\upsilon^2} \right]^p \quad (8)$$

Where *P* is the kernel parameter. It shows the similarity of vectors in the training dataset in a feature space over polynomials of the original variables that is utilized in the kernel.

d) Linear kernel: Linear kernels are used when the data can be linearly separated.

$$K(X_1, X_2) = X_1^T . X_2 \quad (9)$$

When the data is linearly separable or can be divided along a single line, a linear kernel is utilized in SVM. When a given data set contains many features, it is typically employed.

D. Hybrid Flow Directed Arithmetic SVM

The SVM classifier categorizes diseases in the data, and the hybrid FDA with the AOA method is used to optimize each kernel parameter. Performance evaluations are then conducted to determine the effectiveness of the proposed methods across a variety of datasets. Furthermore, analyses of gene expression data in various forms demonstrate the heterogeneity of the method. Support Vector Classification (C), known as Regularization Parameter, has a strictly positive value. This regularization parameter is optimized using a hybrid FDA-AOA.

1) Flow direction algorithm: FDA is evaluated using the direct runoff flow in basins, which is the main focus. The FDA calculates flow velocity based on each individual slope, which falls steeply toward its near neighbors. The FDA introduces new tools for performing optimization. The neighborhood radius decreases from high to low values by defining a washbasin filling technique that helps in escaping local solutions. The FDA algorithm applies the relationship below to determine the initial position of flows:

$$flow[X(i)] = mc + \Re(vc - mc) \quad (10)$$

Where, $flow[X(i)]$ denotes the location of the flow i^{th}, mc *and* vc denote the lower and upper limits of the decision variables, and $\Re$ denotes a uniformly distributed random number between zero and one. Additionally, it is assumed that each flow is surrounded by one or more neighborhoods, whose positions are determined by the relationship shown in below equation:

$$neighbor[X(j)] = flow[X(i)] + \Re(n).\kappa \quad (11)$$

The normal distribution with a mean of 0 and standard deviation of 1 where, $\Re(n)$ is a random variable. $neighbor[X(j)]$ represents the neighbor at the j^{th} position. The large numbers for this parameter shows the searching in a large range, while small numbers (κ) limit searching to a small range.

$$\kappa = (\Re * X\Re - \Re * (flow[X(i)])) * \|best(X) - flow[X(i)]\| * G \quad (12)$$

Where, G is a non-linear weight, $\Re$ is a random number between $0 and \infty$, $X\Re$ is a random location and $\Re$ is a random number with uniform distribution. This relationship's first term demonstrates that $flow[X(i)]$shifts to a random position $X\Re$. The Euclidian distance between $best(X) and\ flow[X(i)]$is lowered to zeros for the second term when iteration is increased, closing the gap between the two. Thus, the local search is not working. In the third term, the (G) is determined as follows:

$$G = \left(\left(1 - \frac{itr}{itr_{\max}}\right)^{2*\Re(n)}\right) * \left(\overline{\Re} * \frac{itr}{itr_{\max}}\right) * \overline{\Re} \tag{13}$$

Where, the random vector with uniform distribution is represented as $\overline{\Re}$. The following relationship determines the new place of the flow.

$$newflow[X(i)] = flow[X(i)] + \upsilon * \frac{flow[X(i)] - neighbor[X(j)]}{\|flow[X(i)] - neighbor[X(j)]\|} \tag{14}$$

$$neighbor[X(j)] = flow[X(i)] + \Re(n) * \kappa \tag{15}$$

$$\upsilon = \Re(n) * M_0 \tag{16}$$

$$M_0(i, j, z) = \frac{fitnessflow[X(i)] - neighbor[X(j)]}{low[X(i,z)] - neighbor[X(j,z)]} \tag{17}$$

Where, $fitnessflow[X(i)]\ and\ neighbor[X(j)]$ are the substitute for the objective value of the $neighbor(j)$and the $flow(i)$. The (z) component reflects the size of the issue. $newflow[X(i)]$ displays the updated position $flow(i)$. Fig. 2 shows the flowchart based on the FDA.

This FDA method starts with the initial population of the search space or drainage basin. The flows then shift to a low height position by achieving best outcome or output point with lowest height.

2) Arithmetic Optimization Algorithm (AOA): AOA is a meta-heuristic algorithm that uses the distribution behavior using four major arithmetic operators in mathematics, such as multiplication, subtraction, division and addition. To carry out the optimization processes in various search areas, AOA is arithmatically modeled and placed into action. This meta-heuristic technique uses population data to solve optimization problems without finding their derivatives. Initialization, exploration, and exploitation are the three important phases of the optimization process.

a) Initialization phase: The best optimized solution is regarded as the best candidate solution in each iteration of the AOA optimization process. The optimization procedure in AOA starts with a collection of candidate solutions (S) , as shown in the matrix, which is generated randomly.

$$S = \begin{bmatrix} X_{1,1} & \cdots & \cdots & X_{1,k} & X_{1,m-1} & X_{1,m} \\ X_{2,1} & \cdots & \cdots & X_{2,k} & \cdots & X_{2,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{M-1,1} & \cdots & \cdots & X_{M-1,k} & \cdots & X_{M-1,m} \\ X_{M,1} & \cdots & \cdots & X_{M,k} & X_{M,m-1} & X_{M,m} \end{bmatrix} \tag{18}$$

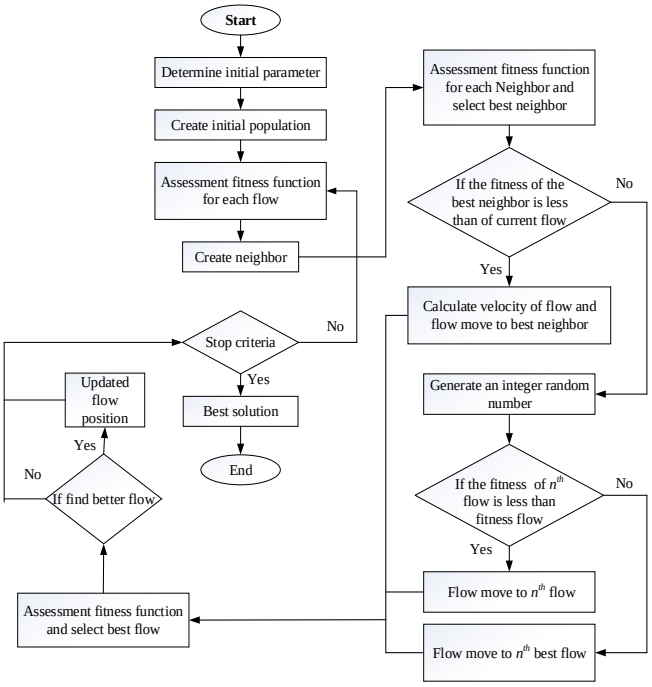

Fig. 2. Flowchart based on flow direction algorithm.

The search phase before it starts to function (i.e., exploitation or exploration), AOA should be select. Math Optimizer Accelerated (MOA) function is the coefficient derived from equation (18) and provided in the subsequent search phases.

$$MOA\_current_{itr} = \mu + current_{itr} \times \left(\frac{\eta - \mu}{\eta_{itr}} \right) \quad (19)$$

Where, the function value at the i^{th} iteration is represented as $MOA\_current_{itr}$. The $current_{itr}$ stands for the current iteration (η_{itr}) between 1 and the maximum number of iterations. The minimum and maximum values for accelerated function are indicated by the μ and η, respectively.

b) Exploration phase: This section introduces the exploring behavior of AOA. To determine the better outcome, two major search processes (division search approach and multiplication search approach) are used by AOA's exploration operators to randomly explore the search area at different positions. This is the most basic rule that can approximate the actions of arithmetic operators. For the condition $a > MOA$ (a is a random number), the exploration search is accomplished by the MOA function. The exploratory phase is evaluated using the position updating equation given below:

$$X_{i,k}(current_{itr}+1) = \begin{cases} \frac{Best(X_{i,k})}{(MOP \cdot \varepsilon)} \times \left((UB_{value(k)} - LB_{value(k)}) \cdot \gamma + LB_{value(k)} \right), & b < 0.5 \\ (Best(X_{i,k}) \cdot MOP) \times \left((UB_{value(k)} - LB_{value(k)}) \cdot \gamma + LB_{value(k)} \right), & otherwise \end{cases} \quad (20)$$

Where, $Best(X_{i,k})$ indicates the k^{th} result in the next iteration, $X_{i,k}(current_{itr})$ represents the k^{th} location of the i^{th} solution at the existing (current) iteration, and $X_i(current_{itr}+1)$ denotes the i^{th} solution in the following iteration. (ε) is a tiny integer number, $LB_{value(k)}$ and $UB_{value(k)}$ stand for the lower bound and upper bound values of the k^{th} position. The control parameter (γ), which is fixed equal to 0.5, is used to alter the search process.

$$MOP(current_{itr}) = 1 - \frac{current_{itr}^{1/\beta}}{\eta_{itr}^{1/\beta}} \quad (21)$$

Where, $current_{itr}$ stands for the current iteration, η_{itr} stands for the maximum number of iterations, and MOP is a coefficient. The $MOP(current_{itr})$ function value at the i^{th} iteration is signifies the parameter, $MOP(current_{itr})$. The term (β) specifies the exploitation accuracy over the iterations, and it is a key parameter.

c) Exploitation phase: The exploitation approach of AOA is described in this section. According to the arithmetic operators, the mathematical representation utilizing any subtraction or addition produced very high dense outcomes related to the exploitation search process. As a result, the exploitation search finds the almost ideal answer that can be determined after numerous attempts (iterations). The exploitation operators (Addition and Subtraction) of AOA investigate the search area systematically in some dense regions and take a method to determine the better result. According to two major search approaches (i.e., Addition search strategy and Subtraction search strategy), modeled below:

$$X_{i,k}(current_{itr}+1) = \begin{cases} (Best(X_{i,k}) - (MOP)) \times \left((UB_{value(k)} - LB_{value(k)}) \cdot \gamma + LB_{value(k)} \right), & c < 0.5 \\ (Best(X_{i,k}) + (MOP)) \times \left((UB_{value(k)} - LB_{value(k)}) \cdot \gamma + LB_{value(k)} \right), & otherwise \end{cases} \quad (22)$$

A deep search is used to fully use the search space. The other operator addition will not be considered till the first operator subtraction in this phase (first rule in equation (22)), which is conditioned by $c < 0.5$. If not, the subtraction will be replaced with the second operator addition to complete the current task. The partitions from the previous phase methods are analogous to those in this phase.

Pseudo-code for AOA:

```
Initialize the AOA parameters, where β, γ.
Initialize the positions of the solution, (i=1,……N)
while (current_itr < η_itr) do
    Compute the fitness function for the solution given.
    Find the best solutions.
    Update the MOA value from equation (19).
    Update the MOP value using (21).
    for (i=1 to solutions) do
      for (k=1 to positions) do
        Create random values between [0,1] (a, b and c)
        if a >MOA then
            Exploration phase
            if b > 0.5 then
                Use division math operator ("÷").
                Update the i^th position of the solution using equation (20)
            else
                Use multiplication math operator ("×").
                Update the i^th position of the solution using equation (20)
            end if
        else
            Exploitation phase
            if c > 0.5 then
                Use Subtraction math operator ("−").
                Update the i^th position of the solution using equation (22)
            else
                Use Addition math operator ("×").
```

27. Update the i^{th} position of the solution using equation (22)
28. **end if**
29. **end if**
30. **end for**
31. **end for**
32. $current_{itr} = current_{itr} + 1$
33. **end while**
34. Return the best solution (X).

In AOA, generating a randomized set of populations is the first step in the optimization process. Every solution improves its position in relation to the best solution found. The parameter *MOA* is similarly increased from 0.2 to 0.9 to emphasize exploitation and exploration. When, $a > MOA$ the candidate solutions effort to diverge from the near-optimal result, and when $b < MOA$ they attempt to meet to the near-optimal result.

IV. Results and Discussion

This section describes the experimental outcomes of the proposed FDASVM-MIKL classification. The proposed work performance is evaluated by using a simulation tool in PYTHON. Some of the simulation parameters are given in the Table II.

TABLE II. Simulation Parameters

Parameters	Values
Regularization Parameter	1.0
Kernel functions	Linear kernel, Sigmoid kernel, Polynomial kernel and Radial Basis Function
Iteration	1000
Intercept scaling	1.0
Fit intercept	True
Random state	None

Several current approaches are examined to assess the proposed categorization performance. The next subsections provide descriptions of the dataset, representations of various performance metrics, analyses, and comparisons.

A. Dataset Description

The data utilized for assessing the performance of gene expression classification through the FDASVM-MIKL based approach is gathered from the datasets Colon, Prostate_GE, Isolet, Lung_ cancer, ALLAML, snp2graph and the download link of each dataset is given below:

1) Colon dataset (http://biogps.org/dataset/tag/colon/): It is a well-known dataset for expression data analysis (cancer). Seven criteria and 90 samples are included. Dataset based on the human species.

2) Prostate_GE https://wiki.cancerimagingarchive.net/display/Public/QIN+PROSTATE): This dataset includes multi-parametric data gathered for staging or detecting prostate cancer.

3) Isolet (https://archive.ics.uci.edu/ml/datasets/isolet): Real characteristics (genetic variation) with several variables are present. There are 697 attributes and a total of 7797 instances.

4) Lung_ cancer (https://archive.ics.uci.edu/ml/datasets/lung+cancer): Integer characteristics make up this multivariate dataset. It primarily has 56 features and 32 occurrences.

5) ALLAML (https://www.kaggle.com/datasets/nikhilsharma00/leukemia-dataset): This dataset, which can be used for training, mainly consists of 7129 probes, 38 samples of bone marrow, and it is associated with Leukemia.

6) snp2graph(https://www.kaggle.com/code/ilfiore/ncbi-check-reference-genome-version/data?select=snp2graph_full.csv): SNPs are identified and selected from gene variation associated with different types of human cancers.

B. Performance Metrics

Various performance metrics, including accuracy, F measure, sensitivity, specificity, and recall statistics, are taken into consideration while evaluating the effectiveness of the proposed gene expression data classification. The mathematical expressions used to describe various metrics are shown in the following representation.

1) Accuracy: The entire count of accurate predictions over the entire total amount of predictions is called accuracy. The accuracy can be mathematically expressed as,

$$Accu\,racy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

Where TP means true positive, FP describes false positive, TN represents true negative, and FN indicates false negative.

2) Precision: The percentage of reliable predictions from the predictor model that correctly anticipated positive cases from all positive predictions is called precision.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{24}$$

3) Recall: The ratio of appropriate items chosen to the total number of appropriate objects is known as recall.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{25}$$

4) F1-score: The harmonic means of the Positive predictive value (PPV) and the Recall or True positive rate (TPR) is determined as the F measure. It can be mathematically signified as,

$$F_{measure} = 2\,\frac{PPV \times TPR}{PPV + TPR} \tag{26}$$

5) Specificity: The number of negative outcomes over the whole number of accurately negative samples. The specificity rate can be mathematically denoted as,

$$Specificity = \frac{TN}{TN + FP} \quad (27)$$

C. Performance Analysis

The analysis contains the main performance metrics, including accuracy, F-measure, recall, sensitivity and specificity, which are considered when comparing the performance of proposed and current techniques. The explanation of the performance includes a description and a graphical representation. The dataset such as Colon, Isolet, ALLAML, Lung CANCER, Prostate and Snp2 are used to classify gene expression into six categories. The performance comparison of accuracy is given in Table III.

TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED METHOD

Performance Metrics	ALLML dataset	Colon dataset	Isolet dataset	Lung Cancer dataset	Prostate dataset	Snp2 dataset
Accuracy	0.9960	0.9995	0.9963	0.9951	0.9971	0.9979
Specificity	0.9927	0.9954	0.9949	0.9876	0.9889	0.9900
Precision	0.9896	0.9937	0.9911	0.9983	0.9930	0.9921
F1-Score	0.9748	0.9830	0.9830	0.9821	0.9938	0.9794
Recall	0.9817	0.9994	0.9988	0.9873	0.9876	0.9829

The comparison of Accuracy, Specificity Precision, F1-score and Recall with its values is represented in Table III. The accuracy of the proposed approach using the ALLML dataset is 99.60%, the Colon dataset is 99.95 %, the Isolet dataset is 99.63%, the Lung Cancer dataset is 99.51%, the Prostate dataset is 99.71%, and the Snp2 dataset is 99.79% obtained. The performance values of Specificity Precision, F1 score and Recall are also given in the table. The performance examination of the proposed Accuracy, Sensitivity, Precision, F1-score, Specificity and Recall is given in Fig. 3.

Accuracy, Specificity, Precision, Recall and F1-score performance is greater than the existing method. Table IV shows the accuracy performance comparison of existing and proposed approaches using the Colon dataset.

The accuracy of the proposed FDASVM-MIKL method is 99.95%. The existing method includes DNN, Improved DNN, CNN, and RNN with accuracy performance of 91.4%, 91.4%, 82.8% and 84%, respectively. Related to the existing methods proposed FDASVM-MIKL approaches has a better accuracy outcome. The performance comparison of the proposed and existing methods using the Colon dataset is represented graphically in Fig. 4.

Fig. 5 presents the comparative graphical representation of the proposed method using current approaches. Table V presents an accuracy comparison of proposed and current approaches using the Isolet dataset.

The proposed FDASVM-MIKL approach has an accuracy value of 99.63%. The accuracy performance of the existing methods, including SVM with multiplicative kernel combination (GKML), ElasticNet-SVM, Multiple indefinite kernel learning based FS (MIK-FS), and SVM with l1 norm regularizer (11-SVM), is 96.01%, 81.589%, 88.03%, and 94.86%, respectively. The proposed FDASVM-MIKL approach has improved accuracy performance compared to the current methods. The performance study of the proposed and current approaches utilizing the Colon dataset is visually depicted in Fig. 5.

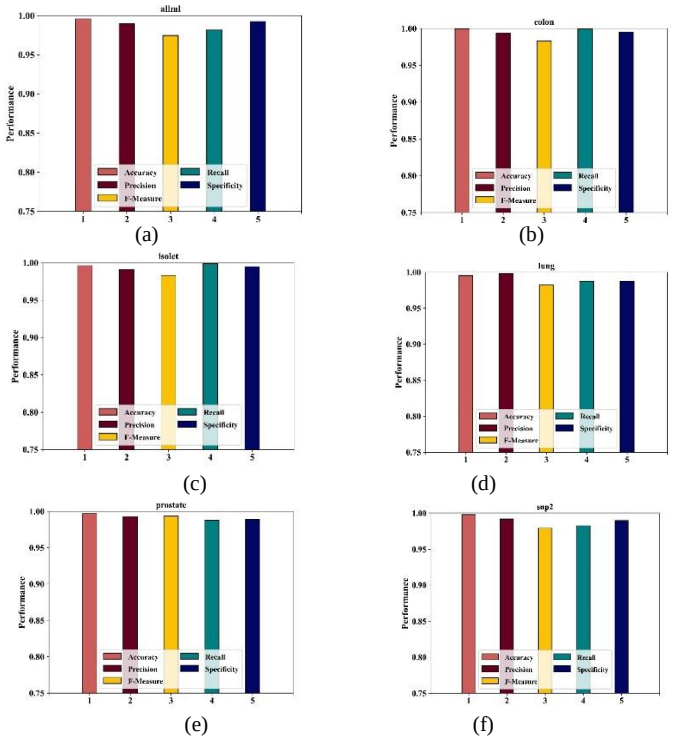

Fig. 3. Performance analysis of the proposed method using a different dataset.

TABLE IV. ACCURACY COMPARISON OF PROPOSED AND CURRENT METHODS USING THE COLON DATASET

Colon dataset	
Methods	**Accuracy**
DNN	0.914
Improved DNN	0.914
CNN	0.828
RNN	0.84
Proposed FDASVM-MIKL method	0.9995

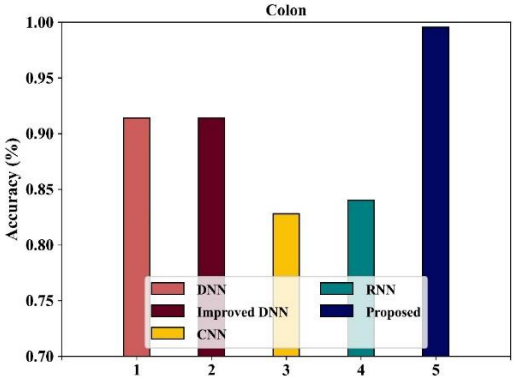

Fig. 4. Accuracy performance analysis of proposed and current methods using the Colon dataset.

TABLE V. ACCURACY COMPARISON OF PROPOSED AND CURRENT APPROACHES USING THE ISOLET DATASET

Isolet dataset	
Methods	**Accuracy**
GKML	0.9601
ElasticNet-SVM	0.81589
MIK-FS	0.8803
11-SVM	0.9486
Proposed FDASVM-MIKL method	0.9963

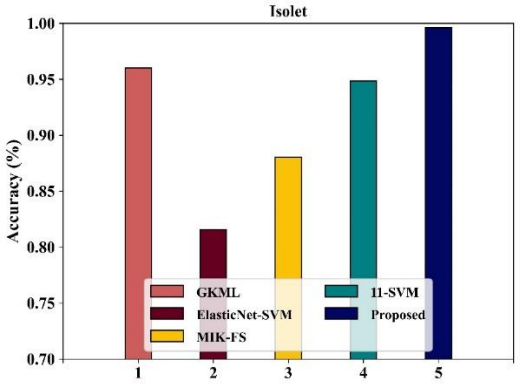

Fig. 5. Accuracy performance analysis of proposed and current approaches using the Isolet dataset.

The graphical comparison of the proposed method with the known methods is shown in Fig. 5. Table VI represents the accuracy comparison of the proposed and existing methods using the Prostate dataset.

TABLE VI. ACCURACY COMPARISON OF PROPOSED AND EXISTING METHODS USING THE PROSTATE DATASET

Prostate dataset	
Methods	**Accuracy**
DNN	0.892
Improved DNN	0.932
CNN	0.892
RNN	0.924
Proposed FDASVM-MIKL method	0.9971

The accuracy of the proposed FDASVM-MIKL approach is 99.71%. The accuracy performance of the existing methods, including DNN, Improved DNN, CNN, and RNN, is 89.2%, 93.2%, 89.2%, and 82.4%, respectively. The proposed FDASVM-MIKL approach has improved accuracy performance compared to the current methods. The performance study of the proposed and current approaches utilizing the Prostate dataset is visually depicted in Fig. 6.

The graphical comparison of the proposed method with the known approaches is shown in Fig. 6. Using the ALLAML dataset, Table VII compares the accuracy of the proposed and current approaches.

The accuracy of the proposed FDASVM-MIKL technique is 99.60 %. The current methods include rMRMR-nMGWO, Random Forest, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Nets, and Decision Tree accuracy performances are 98.3%, 48.6%, 87.5%, 98.7%, and 83.3%, respectively. The result shows that the accuracy performance of the proposed FDASVM-MIKL methodology is greater when compared with existing approaches. Fig. 7 illustrates the performance analysis of proposed and current approaches using the ALLAML dataset.

The graphical comparison of the proposed technique with the known methods is shown in Fig. 7. Using the Lung cancer dataset, Table VIII compares the accuracy of the proposed and current approaches.

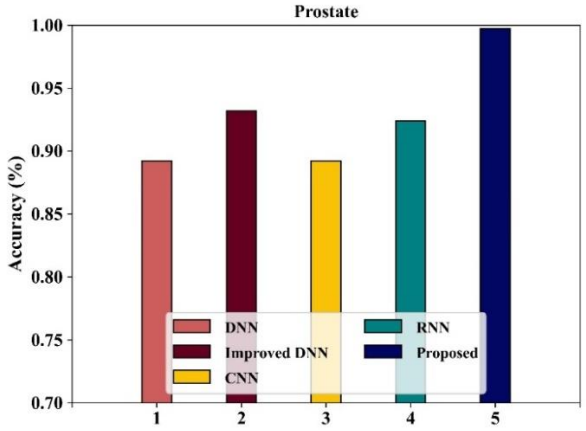

Fig. 6. Accuracy performance analysis of proposed and current approaches using the Prostate dataset.

TABLE VII. ACCURACY COMPARISON OF PROPOSED AND CURRENT APPROACHES USING THE ALLAML DATASET

ALLAML dataset	
Methods	**Accuracy**
rMRMR-\nMGWO	0.983
LASSO	0.486
Random Forest	0.875
Elastic Nets	0.987
Decision Tree	0.833
Proposed FDASVM-MIKL method	0.9960

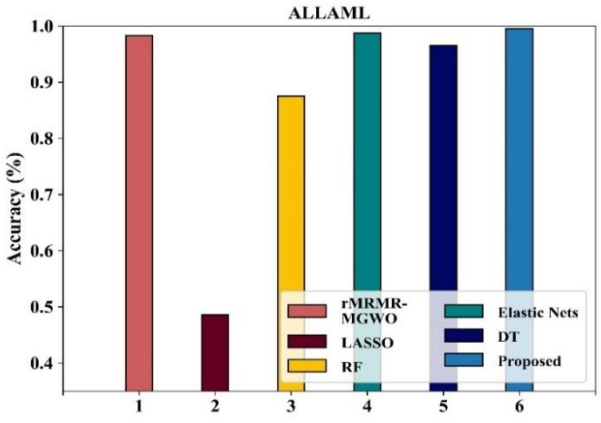

Fig. 7. Accuracy performance analysis of proposed and current approaches using the ALLAML dataset.

TABLE VIII. ACCURACY COMPARISON OF PROPOSED AND CURRENT APPROACHES USING THE LUNG CANCER DATASET

Lung_cancer dataset	
Methods	**Accuracy**
rMRMR-\nMGWO	0.975
LASSO	0.793
Random Forest	0.916
Elastic Nets	0.975
Decision Tree	0.876
Proposed FDASVM-MIKL method	0.9951

The accuracy of the proposed FDASVM-MIKL approach is 99.51%. The accuracy performance of the existing methods, including robust Minimum Redundancy Maximum Relevancy- Gray wolf optimizer algorithm (rMRMR-\nMGWO), Random Forest, Elastic Nets, Least Absolute Shrinkage and Selection Operator (LASSO) and Decision Tree, is 97.5%, 79.3%, 91.6%, 97.5% and 87.6%, respectively. The proposed FDASVM-MIKL approach has improved accuracy performance compared to the current methods. The performance study of proposed and current approaches utilizing the Lung cancer dataset is visually depicted in Fig. 8.

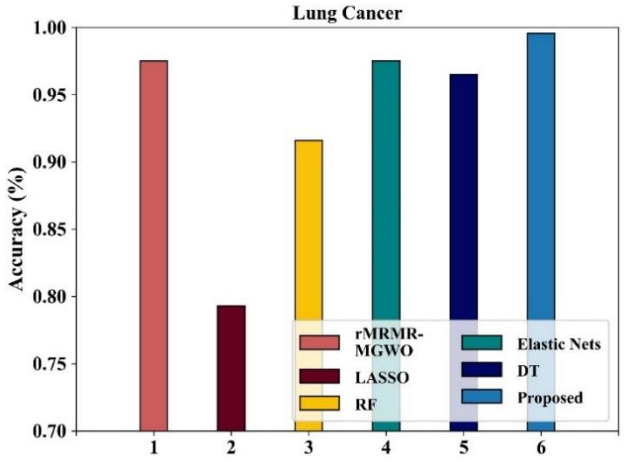

Fig. 8. Accuracy performance analysis of proposed and current approaches using the Lung cancer dataset.

The comparison employs several epoch counts. Fig. 9 compares the Loss epochs of Train, Test and Validation of the proposed system.

TABLE IX. ERROR RATE OF THE PROPOSED METHOD USING DATASETS

Dataset	Error rate
ALLML	0.0039
Colon	0.0004
Isolet	0.0036
Lung_Cancer	0.0048
Prostate	0.0028
Snp2_graph	0.0020

The training, testing and validation losses of the proposed approach on the provided dataset are plotted as a function of epoch number in Fig. 9. In the comparison, various epoch counts are employed. The Error rate of the proposed method for each dataset is illustrated in Table IX.

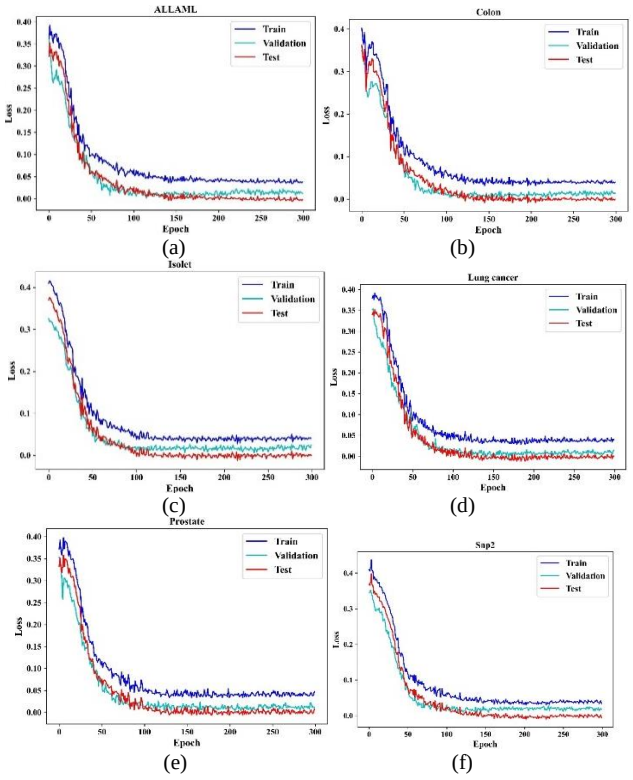

Fig. 9. Training, testing and validation epochs vs. loss.

V. CONCLUSION AND FUTURE SCOPE

This paper proposes a unique method for the analysis of expression data based on multiple indefinite kernel learning, OKPCA and hybrid FDA-AOA is also employed. The PYTHON platform is used to carry out the proposed strategy. The evaluation results are also taken into account for various types of data sources. The accuracy of the classifications is used to determine the performance. The accuracy of the proposed technique using the colon dataset is 99.95%, the accuracy of the proposed technique using the Isolet dataset is 99.63 %, for ALLAML is 99.60%, using the Lung cancer dataset is 99.51%, for the prostate dataset is 99.71% and the proposed method accuracy using the Snp2 dataset is 99.79%. It is clear from the results that the Isolet database performs better. The comparative result of the proposed strategy demonstrated that it is more accurate than other existing methods. In the future, this study will be extended to include improved techniques and advanced classification approaches.

REFERENCES

[1] M. Ota, Y. Nagafuchi, H. Hatano, K. Ishigaki, C. Terao, Y. Takeshima, H. Yanaoka et al, "Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases," Cell, vol. 184, no. 11, pp. 3006-3021, 2021.

[2] O.A. Alomari, S.N. Makhadmeh, M.A. Al-Betar, Z.A. Alkareem Alyasseri, I.A. Doush, A.K. Abasi, M.A. Awadallah, and R.A. Zitar, "Gene selection for microarray data classification based on Gray Wolf

Optimizer enhanced with TRIZ-inspired operators," Knowledge-Based Systems, vol. 223, pp. 107034, 2021.

[3] E. Schaafsma, C.M. Fugle, X. Wang, and C. Cheng, "Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy," British journal of cancer, vol. 125, no. 3, pp. 422-432, 2021.

[4] J.P. Dumanski, J. Halvardson, H. Davies, E. Rychlicka-Buniowska, J. Mattisson, B.T. Moghadam, N. Nagy et al, "Immune cells lacking Y chromosome show dysregulation of autosomal gene expression," Cellular and Molecular Life Sciences, vol. 78, no. 8, pp. 4019-4033, 2021.

[5] A. Rahimi, and M. Gönen, "A multi-task multiple kernel learning formulation for discriminating early-and late-stage cancers," Bioinformatics, vol. 36, no. 12, pp. 3766-3772, 2020.

[6] F. Bao, Y. Deng, M. Du, Z. Ren, S. Wan, K.Y. Liang, S. Liu et al, "Explaining the genetic causality for complex phenotype via deep association kernel learning," Patterns, vol. 1, no. 6, pp. 100057, 2020.

[7] Z. Cai, R.C. Poulos, J. Liu, and Q. Zhong, "Machine learning for multi-omics data integration in cancer," Iscience, pp. 103798, 2022.

[8] M. Palazzo, P. Yankilevich, and P. Beauseroy, "Latent regularization for feature selection using kernel methods in tumor classification," arXiv preprint arXiv: 2004.04866, 2020.

[9] A. Cabassi, S. Richardson, and P.D. Kirk, "Kernel learning approaches for summarising and combining posterior similarity matrices," arXiv preprint arXiv: 2009.12852, 2020.

[10] R. Qi, J. Wu, F. Guo, L. Xu, and Q. Zou, "A spectral clustering with self-weighted multiple kernel learning method for single-cell RNA-seq data," Briefings in Bioinformatics, vol. 22, no. 4, pp. bbaa216, 2020.

[11] L. Guo, X. Zhang, Z. Liu, X. Xue, Q. Wang, and S. Zheng, "Robust subspace clustering based on automatic weighted multiple kernel learning," Information Sciences, vol. 573, pp. 453-474, 2021.

[12] P.J. Baddoo, B. Herrmann, B.J. McKeon, and S.L. Brunton, "Kernel learning for robust dynamic mode decomposition: linear and non-linear disambiguation optimization," Proceedings of the Royal Society, vol. 478, no. 2260, pp. 20210830, 2022.

[13] W. Duan, "Sparse Bayesian kernel learning for high-dimensional regression and classification," PhD diss., 2022.

[14] S. Reddy, A. Kumar, K.Z. Ghafoor, V.P. Bhardwaj, and S. Manoharan, "CoySvM-(GeD): Coyote Optimization-Based Support Vector Machine Classifier for Cancer Classification Using Gene Expression Data," Journal of Sensors, 2022.

[15] A. Cabassi, and P.D. Kirk, "Multiple kernel learning for integrative consensus clustering of omic datasets," Bioinformatics, vol. 36, no. 18, pp. 4789-4796, 2020.

[16] S. Li, L. Jiang, J. Tang, N. Gao, and F. Guo, "Kernel fusion method for detecting cancer subtypes via selecting relevant expression data," Frontiers in Genetics, vol. 11, pp. 979, 2020.

[17] M.O. Adebiyi, M.O. Arowolo, and O. Olugbara, "A genetic algorithm for prediction of RNA-seq malaria vector gene expression data classification using SVM kernels," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 2, pp. 1071-1079, 2021.

[18] E. Kim, and Y. Chung, "Comparison and optimization of deep learning-based radiosensitivity prediction models using gene-expression profiling in National Cancer Institute-60 cancer cell lines," Nuclear Engineering and Technology, 2022.

[19] S.R. Price, D.T. Anderson, T.C. Havens, and S.R. Price, "Kernel Matrix-Based Heuristic Multiple Kernel Learning," Mathematics, vol. 10, no. 12, pp. 2026, 2022.

[20] B. Ulmer, M. Odenthal, R. Buettner, W. Roth, and M. Kloth, "Diffusion kernel-based predictive modeling of KRAS dependency in KRAS wild type cancer cell lines," NPJ systems biology and applications, vol. 8, no. 1, pp. 1-11, 2022.

[21] W.B. Liu, S.N. Liang, and X.W. Qin, "A novel dimension reduction algorithm based on weighted kernel principal analysis for gene expression data," PloS one, vol. 16, no. 10, pp. e0258326, 2022.

[22] A. Rahimi, and M. Gönen, "Efficient multi-task multiple kernel learning with application to cancer research," IEEE Transactions on Cybernetics, 2021.

[23] H.Z. Almarzouki, "Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile," Journal of Healthcare Engineering, 2022.

[24] J. Feng, L. Jiang, S. Li, J. Tang, and L. Wen, "Multi-omics data fusion via a joint kernel learning model for cancer subtype discovery and essential gene identification," Frontiers in genetics, vol. 12, pp. 647141, 2021.

[25] N. Wani, and K. Raza, "MKL-GRNI: A parallel multiple kernel learning approach for supervised inference of large-scale gene regulatory networks," PeerJ Computer Science, vol. 7, pp. e363, 2021.