International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I9P203

Volume 70 Issue 9, 20-33, September 2022
© 2022 Seventh Sense Research Group®

Original Article

An Unsupervised Deep Feature Selection and Ensemble Deep Learning Model for Cancer Classification

K. Prema[1], A. Kumar Kombaiya[2]

[1,2]*Department of Computer Science, Chikkanna Government Arts College, Tirupur, India.*

[1]*Corresponding Author :* premakar11phd@gmail.com

Received: 17 June 2022 Revised: 30 July 2022 Accepted: 10 August 2022 Published: 06 September 2022

Abstract *- Microarray technology is a principle to begin and verify the antibody microarrays in a registered series of patents. Within a particular trial, a Microarray Data Analysis (MDA) is utilized to identify the patterns of thousands of genes. The MD consists of a large volume of gene expression data for detecting cancer diseases. But, the imbalanced class label instances in microarray gene datasets and initialized parameter value for the classifier lead to over-fitting and under-fitting problems in cancer classification. Therefore, in this article, a stacking ensemble of Deep cluster-based Deep Learning (DL) systems for Cancer Classification is designed to overcome the abovementioned difficulty by using many learning models to build one ideal predictive model. The developed model is classified into three sections. First, a Modified Harmony Search Algorithm and Modified Kernel-based Fuzzy C-Means (MHSAMKFC) are developed to eliminate huge redundant features effectively. Second, the MHSAMKFC with Convolutional Neural Network (CNN) classifier is proposed to handle uncertainties in the labelled training dataset to improve the classifier performance. Third, the over-fitting and the under-fitting problem of MHSAMKFC-CNN is reduced by the ensemble method, which uses multiple learning models to provide better prediction accuracy. The whole process is termed to be En-MHSAMKFC-CNN. Finally, experimentation is carried out on four Gene Expression Microarray (GEM) datasets and verified that the En-MHSAMKFC-CNN improves the classification performance of SVM, KNN, RF and ANN classifiers.*

Keywords *- Microarray Data Analysis, Convolutional Neural Network, Fuzzy C-Means, Harmony Search Algorithm, Cancer Classification.*

1. Introduction

One in every six fatalities can be attributed to cancer, making it the second leading cause of mortality globally [1]. If cancer is detected and treated early, the death rates can be diminished. The distinguishing qualities of distinct cancer valetudinarians are vital to characterize, and patient-specific care is scheduled because the indicators vary from case to case. The genomic data from the patient is ideal for extracting these features. The tremendous advancement in MD processing research over the last decade has made it a powerful tool for illness diagnosis [2]. Clinical pathology recognizes, explains, and categorizes human diseases, including cancer, using microarrays based on genomic information. Cancer patients will be benefitted from more efficient treatment and more responsive cancers if discovered early and appropriately.

DNA microarrays produce a significant amount of genetic data that is possibly useful for cancer identification but is mostly useless and noisy. The presence of antiquated, unnecessary, and distracting genomes degrades data collections. Approaches to gene selection are critical for developing a clinical framework for the condition, particularly when samples are scarce [3]. TLBOGSA [4] was created for cancer categorization by utilizing a new hybrid metaheuristic method named Teaching learning-based algorithm (TLBO) and Gravitational Search Algorithm (GSA). When gravitational search mechanisms are combined with the teaching stage, the search potential during the growth period improves. However, due to the high dimensionality and small sample quantity of GEM data, this Feature Selection (FS) may not be efficient in identifying important genes.

The MHSA [5] is a project intended to overcome the dimensional curse issue by identifying relevant genes with fewer intricate issues. But, this method has a disadvantage in the last iterations when the Pitch Adjusting Rate (PAR) value is close to zero, which can cause the algorithm's convergence performance to stagnate. The MKFC method [6] solves the problem of classic FCM not being able to tolerate minor variations between clusters. However, this technique is very sensitive to noisy data that provide less informative genes. This paper combines both MHSA and MKFC for FS from microarray cancer datasets to solve these problems. The MHSAMKFC method deliberately handles the datasets having a large amount of data without class label and eliminate redundant feature effectively.

In literature, Machine Learning (ML) techniques like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and Artificial Neural Network (ANN) classifier were utilized to classify a large amount of MD. The wrapper FS method in processing features selection and classification performs better than filter-based FS methods. An SVM-based classification with spider monkey optimization based FS [7] classifies the cancer diseases with two objectives. The initial goal is to reduce the number of parameters while increasing classifier accuracy. But, the ML algorithm for cancer prediction is still challenging for small samples and is easily prone to susceptible error.

To solve this, the DL Based cancer type classification was introduced [8] to classify the larger GEM datasets. The Deep Neural Network (DNN) and visualization approaches such as layer-wise fetidness and Grad-Cam were used to screen out the genes with low deviation throughout all data. Then, the high-dimension expression information was integrated into a 2-dimensional space to suit the Convolutional layers. The idea of the Guided Grad-Cam [9] was used to develop a three-layer CNN with a trained neural structure for selecting notable genes for classification. This method provides better results for cancer prediction. But, it results in a high computational time problem.

DL-based Unsupervised CNN classifier is introduced to improve the classification process, which intends to learn data interpretations that can better reconstruct training samples. This Unsupervised CNN is integrated with the MHSAMKFC method to produce a single optimal predictive model to reduce the dimensionality reduction and uncertainties in the labelled training MD.

The MHSAMKFC-CNN has the imbalanced class label instances in datasets and initialized parameter values for the classifier, which leads to over-fitting and under-fitting problems while performing the cancer classification. To solve these issues, MHSAMKFC-CNN is enhanced with a stacking ensemble, which uses multiple learning models to provide better prediction accuracy. This Ensemble model is achieved by employing majority voting, where unlabeled data will assign a class with the highest number of votes among the CNN classifiers predictions. The developed method improves the classification performance compared to the standard classifier for the prediction of cancer type prediction on GEM Datasets.

2. Literature survey

A feature extraction strategy was developed using ensemble FS and improved discriminant independent component analysis for MD classification [10]. However, the datasets are pre-processed by setting thresholds which greatly influence the accuracy of classifiers. A centroid-based DNA choice technique was developed [11] for categorizing the microarray information. But, the accuracy of this method decreased with a large number of data features.

A framework was developed [12] to choose the top-ranked features for MD. This model used feature ranking techniques and attributed clustering in the pipeline to eliminate irrelevant features. However, if the dataset was imbalanced, efforts were not given to resolve this issue in the dataset. A two-stage local dimensionality technique was suggested [13] for local dimensionality reduction and classification of MD. However, the regularization parameter influences the accuracy of the two-stage local dimension approach.

A Cooperative Co-evolution method for FS (CCFS) was presented [14] on MD. A binary gravitational search algorithm was employed to search the solution space utilizing the principle of coevolution theory through filter criterion in the objective function. However, this technique had a high computational complexity. Bayesian Lasso quintile regression method was presented [15] to classify gene expression for GEM selection. This method combines Bayesian MCMC evaluation with a skewed Laplace distribution for defects and a graded hybrid of regular probabilities for regression coefficients.

A multiobjective attribute selection model was constructed [16] for MD through distributed parallel algorithms. This model selects the most significant features based on multiple objectives such as feature number, classification error and feature redundancy to classify the MD more effectively. However, there might be a possibility of conflicts between the various goals. A Partial Maximum Correlation Information (PMCI) method was presented [17] for the classification of MD. The orthogonal components were extracted from the attribute space to assess the significance of all attributes. However, this method has a poor F1 score. A discrete Bacterial Colony Optimization with a multi-size population (BCO-MDP) algorithm was developed [18] for feature selection and classification of microarray gene expression cancer data. However, finding a suitable search space for high classification accuracy was challenging without prior knowledge of datasets.

An attribute Selection scheme was developed [19] for large data dimension Data using Weighted K-NN (WKNN) and GA. GA was applied for computing the best weight vector for the contribution of the value in the component to the classification, which was equal to the input degree of the feature value. However, this method has high computational complexity. A weighted group hybrid method was demonstrated [20] using a Partition Relevance Analysis (PRA) and reduction process. It is accomplished by eliminating duplicate and noisy indexes using data dimensionality reduction techniques in the second phase of PRA. However, this method needs to work more on complex strategy functions. A Grouping Genetic Algorithm with Extreme Learning Machine (GGA-ELM) was developed [21] to resolve a maximally diverse issue in microarray data classification. However, this method has a low impact on larger datasets.

A stacking ensemble DL technique for cancer type prediction utilizing TCGA data was described [22] based on a One-Dimensional CNN (1D-CNN) approach. Least Absolute Shrinkage and Selection Operator (LASSO) regression was employed as an FS approach to reduce the number of genes. However, this approach had a considerable computational overhead. The robust Minimum Redundancy Maximum Relevancy (rMRMR) filter strategy was developed [23] with Modified Gray Wolf Optimizer (MGWO) to determine the top-ranking genes in a microarray data classification. On the other hand, the proposed combination achieves poor classification accuracy.

3. Proposed methodology

The genes are the features of gene expression analysis. Gene selection is the procedure of identifying the genes most closely associated with a specific subclass. Lowering the dimensionality, as mentioned above, of the dataset is one of the benefits of this technique. Furthermore, when categorization is used, many genes become unnecessary. When gene selection is used, the risk of overwhelming the impact of relevant genes is lessened. In MDA, FS and clustering is by far the most popular technique. Hence, this research proposes an MHSAMKFC algorithm that seeks to overcome the dimensionality problem on MD and select meaningful genes. The complete working of this algorithm has been briefly explained below.

3.1. MHSAMKFC

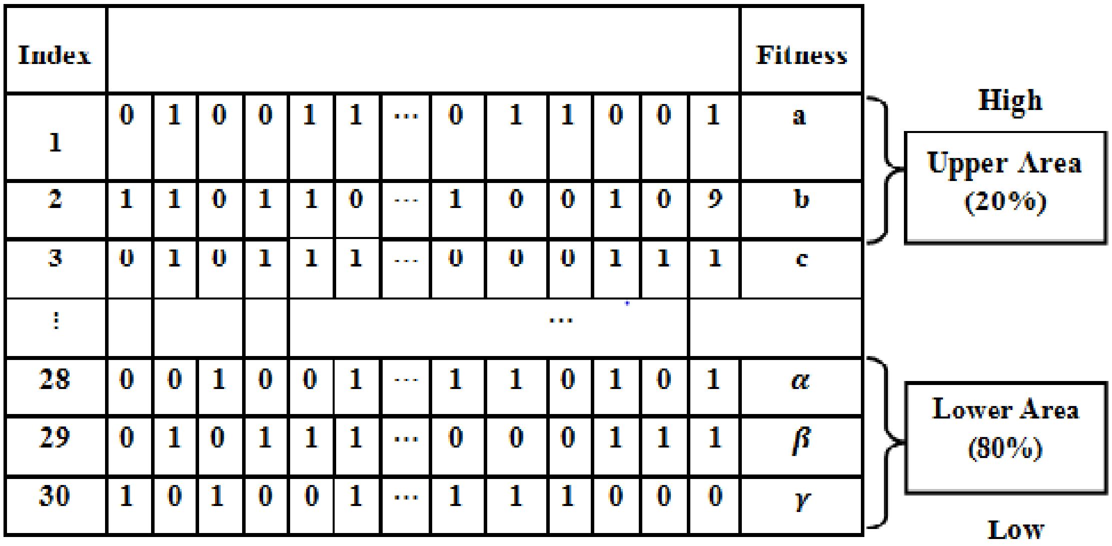
In this process, an MHSA is developed for the FS process by modifying the existing HS, which is briefly illustrated below in [10].

3.1.1. Step 1 Constructing variables and Harmony Memory (HM)

The first step is to define the scope of the HM project, choose a good starting point for your work, and establish parameters and harmony. The meaning of the parameters must be understood before using this procedure. HS can be compared to a Genetic Algorithm (GA) because it is an evolutionary algorithm. Genes are the essential portion of a Harmony Vector (Hv) and are the core parts of the chromosome in GAs. The amount of harmonies in a single HM is called the HM Size (HMS). Hv are random at the outset of the HS method's execution, and the iterative process relies on a small number of previously determined harmony values.

3.1.2. Step 2 Forming New Harmony by separating HM

Generating a New HM is similar to that of the present HS algorithm. Still, the observation will be carried out by dividing the HM into two parts, as illustrated in Figure 1. The topmost region comprises harmonies in the top 20% of fitness inside a single HM. Harmony Memory Considering Rate (HMCR) and PAR are not used for this area. As a result, the activation process does not add New Harmony. Rather than creating a diversity of combinations, when the combination is recombined within the harmony of the upper area, a combination of higher fit could be found, after which new harmonies are developed. The second area is the lower area in HM, in which the available harmonies form the latest harmonies by HMCR and PAR.



Index														Fitness
1	0	1	0	0	1	1	···	0	1	1	0	0	1	a
2	1	1	0	1	1	0	···	1	0	0	1	0	9	b
3	0	1	0	1	1	1	···	0	0	0	1	1	1	c
⋮					···									
28	0	0	1	0	0	1	···	1	1	0	1	0	1	α
29	0	1	0	1	1	1	···	0	0	0	1	1	1	β
30	1	0	1	0	0	1	···	1	1	1	0	0	0	γ

Fig. 1 Divided harmony memory

3.1.3. Step 3 Updating HM

Goodness-of-fit refers to the degree to which the classification model used in the piece works with the harmony selection used to make the classifications. Each harmony value determines the fit, sorted in the harmony sequence with the highest fitness. In Step 2, the longstanding binary harmonics with the smallest fit are corrected and eliminated to suit the scale of the originally provided HMS.

3.1.4. Step 4 Iterating previous Steps 2 and 3

Atpresent, there is no newly modified process. Steps 2 and 3 should be repeated as many times as the iteration req uires. The upper section discovers harmonies with a greater fitness level inside the combination with higher appropriateness as the total count of trials develops. The lower section preserves the benefits of the previous HS, namely, discovering combinations based on the diversifications. The greatest classification performance of two locations within one HM is saved in a text file

In MHSA, The Harmony Fitness (Hf) is evaluated using the Inter and Intra cluster distance, a cluster analysis used to discover overall distribution patterns and intriguing relationships among collected data features. The Intercluster distance I_d^r is the distance between two features belonging to two different clusters, whereas the Intra cluster's distance I_d^s is the distance between two features belonging to the same cluster, which is defined as follows

$$I_d^r = \sum_{i=1}^{N} \sum_{j=1}^{C} \sqrt{x_i - c_j} \qquad (1)$$

$$I_d^s = \sum_{i=1}^{N} \sum_{j=1}^{C} \sqrt{c_i - c_j} \qquad (2)$$

In Equations 1 and 2, N = Number of Clusters, C= Number of features under clustering, x_i denotes the feature under clustering, c_i represents the $i-th$ cluster, c_j = centroid of (same) cluster, i, j = Number of iterations. By using this equation, the Hf can be estimated to calculate the fitness value for the cluster features efficiently.

After the feature selection process, the collected features are further processed to the clustering method to eliminate the irrelevant features in the given datasets. The MKFC algorithm adds kernel information to the classic FC algorithm. It addresses the FC algorithm's inability to manage small changes within clusters, which has been created for effective data clustering. The kernel approach converts a high-dimensional feature space from a non-linear input data structure.

Kernel-based approaches entail conducting an arbitrary non-linear mapping from a d-size feature space R^d to a higher-size space (kernel space (K)). The kernel space may have an indefinite number of dimensions. Since the starting problem in the feature space may be non-linear and not exponentially distinct, increasing the number of dimensions is warranted.

MKFCM is divided into two sorts, the primary of which includes prototypes built in the attribute space. MKFCM-F will be the name of these clustering methods (with F standing for the feature space). The prototypes are preserved in the K in the second category, denoted as MKFCM-K, and so must be simulated in the feature space by generating an inverse mapping from kernel space to feature space. The MKFCM approach has the benefit that the hypotheses are stored in the feature space and are implicitly projected to the kernel space using the kernel operator.

Obviously, due to the fact that known kernel functions require only kernel functions to address problems in the kernel space, i.e., the inner development of the transform function. This variant of MKFCM is referred to as MKFCM-K when the concepts o_i are produced in the kernel space. The fundamental purpose of Equations 3,4 and 5 is to construct kernel space

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ij}^{m} \left\| \varphi\,(x_j) - o_i \right\|^2 \qquad (3)$$

$$u_{ij} = {}^{1}\Big/_{\sum_{h=1}^{c} (d\varphi_{ij}^2 / d\varphi_{ij}^2)^{1/(m-1)}} \qquad (4)$$

$$d\varphi_{ij}^2 = k\left(x_j x_j\right) - \frac{2\sum_{h=1}^{n} u_{ih}^{m}\, k(x_h x_j)}{\sum_{h=1}^{n} u_{ih}^{m}} + \frac{\sum_{h=1}^{n} \sum_{l=1}^{n} u_{ih}^{m}\, k(x_h x_l)}{\sum_{h=1}^{n} {u_{ih}^{m}}^{2}} \qquad (5)$$

Another type of MKFCM limitation is that the kernel space prototypes are basically mapped from the unique data space, otherwise the feature space. That is, the function is defined in Equation 6

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ij}^{m} \| \varphi\,(x_j) - \varphi(o_i) \|^2 \qquad (6)$$

This type of KFCM is mentioned as KFCM-F. Naturally, only $k(x, y) = exp\,(-\|x - y\|^2 / r^2)$ Gaussian kernel in Equation 7 is applied in KFCM, and since$k(x, x) = 1$ for Gaussian kernel

$$\begin{aligned} \left\| \varphi\,(x_j) - \varphi(o_i) \right\| &= <\ \varphi\,(x_j),\ \varphi\,(x_j) > + \\ &\quad <\ \varphi\,(o_i)\ \varphi\,(o_i) > -2\ \ \varphi\,(x_j)\varphi\,(o_i) \\ &= k\left(x_j, x_j\right) + k(o_i, o_i) - 2k(x_j, o_i\ ) \\ &= \ 2(1\text{-}(x_j, o_i\ )) \end{aligned} \qquad (7)$$

Here, $K\left(X_j, O_i\right)$can be considered as a robust distance measurement derived from the kernel space. For these KFCM-F applying Gaussian kernels, iteratively update the prototypes and memberships as Equation 8

$$\left\| \varphi\,(x_j) - \varphi(o_i) \right\| = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^{m} (-k(x_j, o_i\ )) \qquad (8)$$

```
**Algorithm 1 MHSAMKFC**
Input: Given Dataset D
Output: Best feature (Gene) cluster and $Hf$
\\HS algorithm: FS process
Apply the required variable BDR, HMCR, PAR and HMS
Assign $itr: = 0$ {iteration in progress}
Choose Harmony values (0 and 1)
BDR = HMS*0.2 // establish a top and bottom limit
For ($i = 1: i \leq HMS$), then
Develop primary harmony ($x_{new}$)
Perform Algorithm 2 to obtain cluster and $Hf$
End for
Repeat
For (J = 1: N) then            //HS in upper area
 $x_{new}$ = Arbitrarily chosen from $x_{(BDR+1)J}$to $x_{(HMS)j}$
end for
Create New Harmony ($x_{new}$)
Perform Algorithm 2 to obtain cluster and $Hf$
If ($Rand(0,1) < HMCR$) then //HS in lower area
 For($J = 1: N$) then
 $x_{new}$= Randomly select from$x_{(BDR+1)J}$to $x_{(HMS)j}$
 If ($Rand(0,1) < PAR$) then
  $x_{new}$ = $|x_{new}-1|$
 end if
 end for
 Generate new harmony ($x_{new}$)
Perform Algorithm 2 to obtain cluster and $Hf$
else
Develop a New Harmony randomly
Perform Algorithm 2 to obtain cluster and $Hf$
End if
if(fi$t(HM_{new(upper,lower)}) <$ fi$t( HM_{old})$)
Update HM
End if
Set$itr += 1$
Until ($itr < maxit$)
Determine the best harmony (Gene Feature and cluster)
```

```
**Algorithm 2.MKFC**
\\MKFCM: Clustering process
Fix $c, t_{max}, m > 1 \ and \ \varepsilon > 0$ for some positive constant;
Initialize the membership $u_{ik}^{0}$
$J_m = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{0} \|X_k - V_i\|^2$
For t =1, 2…, $t_{max}$, do:
        (a)Upgrade all prototypes $V_i^t$
        (b) Upgrade all memberships $U_{ik}^t$
Compute $E^{\mathrm{T}} = max_{i,k} | U_{ik}^t - U_{ik}^{t-1} |$, If$E^{\mathrm{T}} \leq \varepsilon$,
$U \in \{u_{ik} \in 0,1 \,|\, \sum_{i=1}^{c} u_{ik} = 1 \ \forall k \,; \ O < \sum_{k=1}^{N} u_{ik} < N, \forall i\}$
Stop: $else \ t = t + 1$ \\ number of clusters is obtained
```

The generated features are transferred to classifiers like SVM, KNN, RF, and ANN for effective cancer classifications to validate the efficacy of proposed FS and clustering approaches. But, these machine methods take a long time to classify the data features. So, the DL structure has been used for the classification process on GEM datasets.

3.2. MHSAMKFC-CNN

Once the data features are collected from the classifier, the Unsupervised CNN is used in this research work to reduce the time complexity and increase cluster performance by updating the cluster centres based on a reliable FS, which is a key component of this method to ensure its success. For the efficient performance of the clustering Algorithm, a CNN with the proposed MKFCMHS is briefly explained in Figure 2.

The CNN codes (the layer activations in a CNN before classification, including non-linearity) capture much information about the gene expression data. They have worked well as features for gene expression data used in many classification tasks. This work takes a step further in investigating the response of the individual layers to images of different classes.

Using CNN, layer activations are clustered using the MHSAMKFC technique. Cluster centroids are saved using this technique. If a particle's distance from the cluster's centroid is smaller than its distance from any other centroid, it is considered part of that cluster. Through experiments, MHSAMKFC-CNN determines the optimal centroids by switching between (1) assigning data points to categories based on the current centroids and (2) assigning data points to categories according to the actual centroids. (2) Choosing a centroid (the cluster's epicentre) based on the pre-existing grouping of data points. MHSAMKFC-CNN will be developed according to a dataset $D \in \mathbb{R}^{d \times k}$of k vectors (i.e., centroids), so thus that a data matrix $x_i \in \mathcal{R}^d i = 1, \dots, m$ can be projected to a code matrix s it that minimizes the error in reconstruction, which is defined as follows in Equations 9,10 and 11.

$$\underset{D \ \ s}{min} \sum_{i=1}^{N} \|D_i s_i - f(x_i, w)\|_2^2 \tag{9}$$

$$subject \ to \ \|s_i\| 0 \leq 1, \forall i \tag{10}$$

$$\|D_j\|_2 = 1, \forall i \tag{11}$$

where x_i denotes the source data and (x_i,) denotes the CNN function that calculates the gene expression data x_i With w weightiness and Dj is the jth column. The objective is to train a $D \in \mathbb{R}^{d \times k}$ and encoded vector of s_i , which will allow the initial CNN features to be reconstructed. Equation (9) yields a set of ideal cluster designations$\widehat{y_i}$, which are employed as substitute labels for CNN learning factors in ensembles. Each CNN's parameters were subsequently learnt by addressing Equation 12:

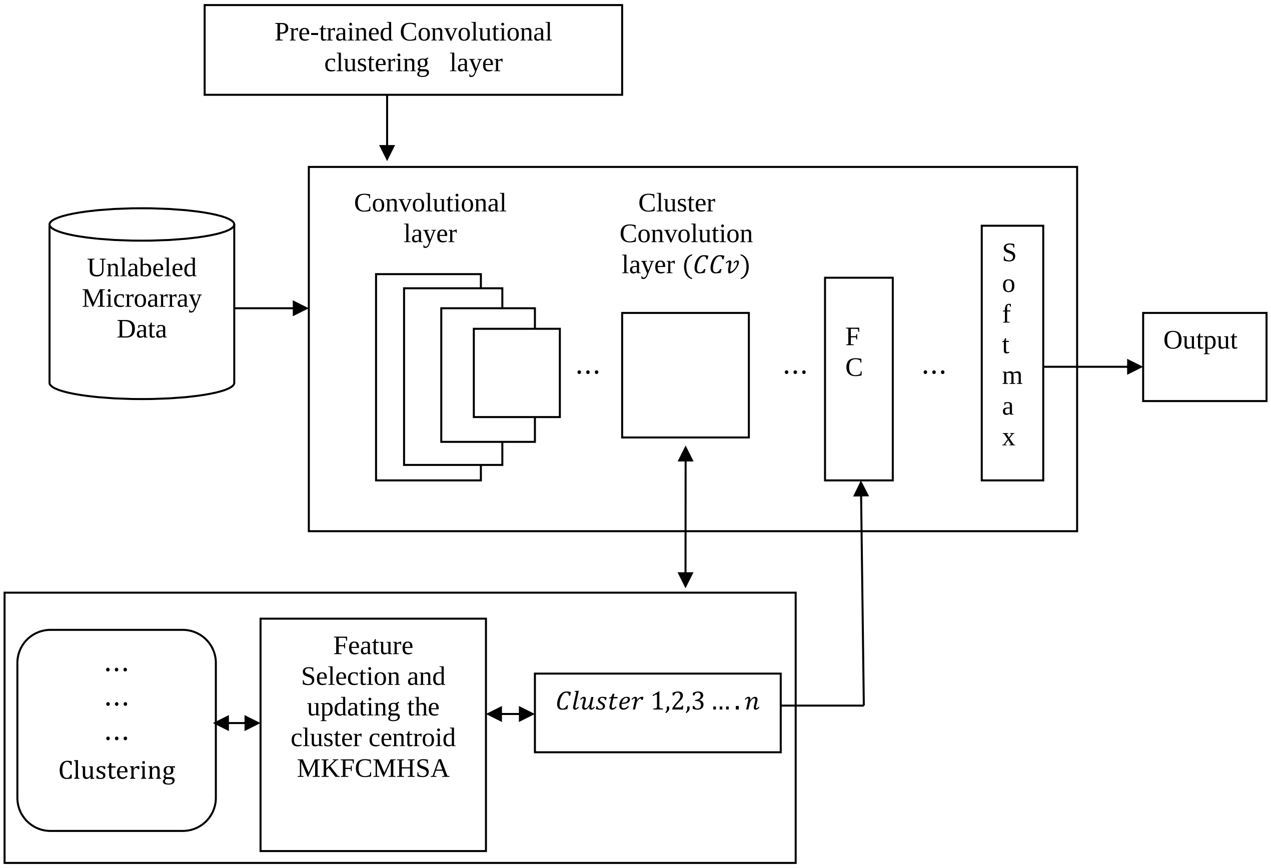

Pre-trained Convolutional clustering layer
Unlabeled Microarray Data
Convolutional layer
Cluster Convolution layer (CCv)
FC
Softmax
Output
…
Clustering
Feature Selection and updating the cluster centroid MKFCMHSA
$Cluster$ 1,2,3 ….n

Fig. 2 Structure for the CNN layer with MHSAMKFC System

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathrm{f}\,(x_i, \mathrm{w}), \hat{y_i}) \tag{12}$$

Stochastic Gradient Descent (SGD), as used in conventional CNN backpropagation, is used to reduce the cross entropy loss $\mathcal{L}$, in the equation mentioned above. Surrogate labels are generated from CNN features (see Equation 9) and used to fine-tune the CNNs' parameters during training with MHSAMKFC, a final iterative process (see Equation 12). Iterations of this process are performed till the clustering and failure have stabilized. Initially, a different procedure is devised to understand what kind of features are learnt in every layer.

1. Select n the number of clusters/classes.
2. Select, from the MD, subsets of equal size k, from each class. Thus in total, there are n_k features
3. Each information is passed through the pre-trained network, and their activations for all layers are recorded: In layer i there are n_k activations, which constitute the D_i, for the analysis at that layer described in the next step.
4. At layer i, the t D_i is clustered into n clusters using the MHSAMKFC-CNN algorithm, which is explained above
5. Analyze the clusters obtained at each layer concerning the original classes to which the corresponding features belong.

3.3. Ensemble of MHSAMKFC-CNN

The method of enhancing classifier efficiency by integrating the contributions of trained sub-models to tackle the identical categorization issue is known as the meta-learner, a prototype that learns to improve the base- learners' predictions and obtains the final result. As a result, the ensemble approach achieves better predictive performance on the MD for Cancer categorization than individual learners. An ensemble's generalizability decreases variance in predictions and assures that the most consistent and best possible projection is made. The Meta model develops to integrate the input predictions to generate a better final prediction than each of the base classifiers by taking the output of the sub-models MHSAMKFC- CNN with varied variables as input. The suggested stacking ensemble DL algorithm for the cancer prediction method on MD is shown in Fig. 3.

Algorithm 3. Stacking Ensemble Algorithm

Input: Data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Highest-level learning algorithms $L_1, \ldots \ldots, L_T$
Lowest-level learning algorithm L.

```
Process:
1. For $t = 1, \dots, T$ : %Train a highest-level learner by applying the
2. $h_t = L_T(D)$;% highest-level learning algorithm $L_T$
3. End
4. D* = ∅;        % Create a new database
5. For $i = 1, \dots, m$:
6.    For $t = 1, \dots, T$ :
7.       $z_{it} = h_t(x_i)$;
8.    end
9.    $D* = D* \cup ((z_{i1}, \dots\dots, z_{iT}), y_i)$;
10. end

11.   $h* = L(D^*)$;   % Apply the Lowest-level learning algorithm $L$ to the
                       % new data set D* to learn the second-level learner h*.

Output: $H(x) = h^*(h_1(x), \dots, h_T(x))$
```

4. Dataset Description

The effectiveness of the existing and proposed GEM dataset based on the cancer prediction method is implemented in MATLAB 2018a. It runs on a Microsoft Windows 7 with an Intel processor at 2.70 GHz and 4GB memory. Three GEM datasets, such as Leukemia, Lymphoma and prostate microarray, are collected for experimental purposes. These datasets are publicly available on the internet, listed in Table 1. From the collected data, 40% of data are used for training, and 60% are used for testing.

Table 1. Dataset Desecration

Data set	Instances	Features	Classes	Source
Leukemia	72	3572	2	https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia.html
Lymphoma	77	2647	2	https://ico2s.org/datasets/microarray.html
Prostate	102	2135	2	https://ico2s.org/datasets/microarray.html

5. Experimental results

The effectiveness of existing methods like GGA-ELM [21] and rMRMR-MGWO [23] and proposed methods MHSAMKFC, EN-MHSAMKFC using different algorithms like KNN, SVM, RF, ANN and CNN based on the abovementioned datasets. The performances are tested in terms of Accuracy, Precision, Specificity, Sensitivity and F1 score, which are briefly explained below.

5.1. Accuracy

The fraction of instances successfully categorized is described as accuracy. It's obtained by dividing the total proportion of accurately predicted sick (true positive) and normal (true negative) people by the overall number of classifications. It is calculated in Equation 13

$$Accuracy = \frac{TP+}{TP+TN+FP+FN} \tag{13}$$

Where TP denotes cancer patients correctly categorized as sick, FP denotes healthy people who are wrongly labelled as sick. TN denotes healthy individuals who are correctly identified as healthy. FN denotes sick people who are incorrectly classed as healthy. Table 2 depicts the comparison results of accuracy for proposed and existing techniques.

Table 2. Comparison of Accuracy

Datasets\ Classifiers	*Leukemia*	*Lymphoma*	*Prostate*
GGA-ELM	82.42	83.36	84.15
rMRMR-MGWO	84.13	85.24	85.67
MHSAMKFC - KNN	85.34	87.20	86.98
MHSAMKFC - SVM	87.42	89.24	89.67
MHSAMKFC - RF	90.76	91.87	91.34
MHSAMKFC ANN	92.24	93.74	95.14
MHSAMKFC-CNN	94.48	95.06	96.75
EN-MHSAMKFC-CNN	96.34	97.55	98.58

Fig. 4 displays the Accuracy of existing GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN and EN-MHSAMKFC-CNN techniques. In this analysis, EN-MHSAMKFC-CNN method is 16.88%, 14.51%, 12.88%, 10.20%, 6.148%, 4.444% and 1.968% for leukemia dataset; 17.02%,14.44%, 11.86%, 9.311%, 6.182%, 4.064%, 2.61% for Lymphoma dataset and 17.14%, 15.06%, 13.33%, 9.936 %, 7.926%, 3.615%,1.891% for Prostate dataset is higher than that of GGA-ELM, rMRMR-MGWO with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN methods respectively on given dataset. This analysis shows that the EN-MHSAMKFC-CNN can achieve better accuracy than other methods for microarray cancer classification.

5.2. Precision

The proportion of true positive incidents that are categorized as positive is known as precision.

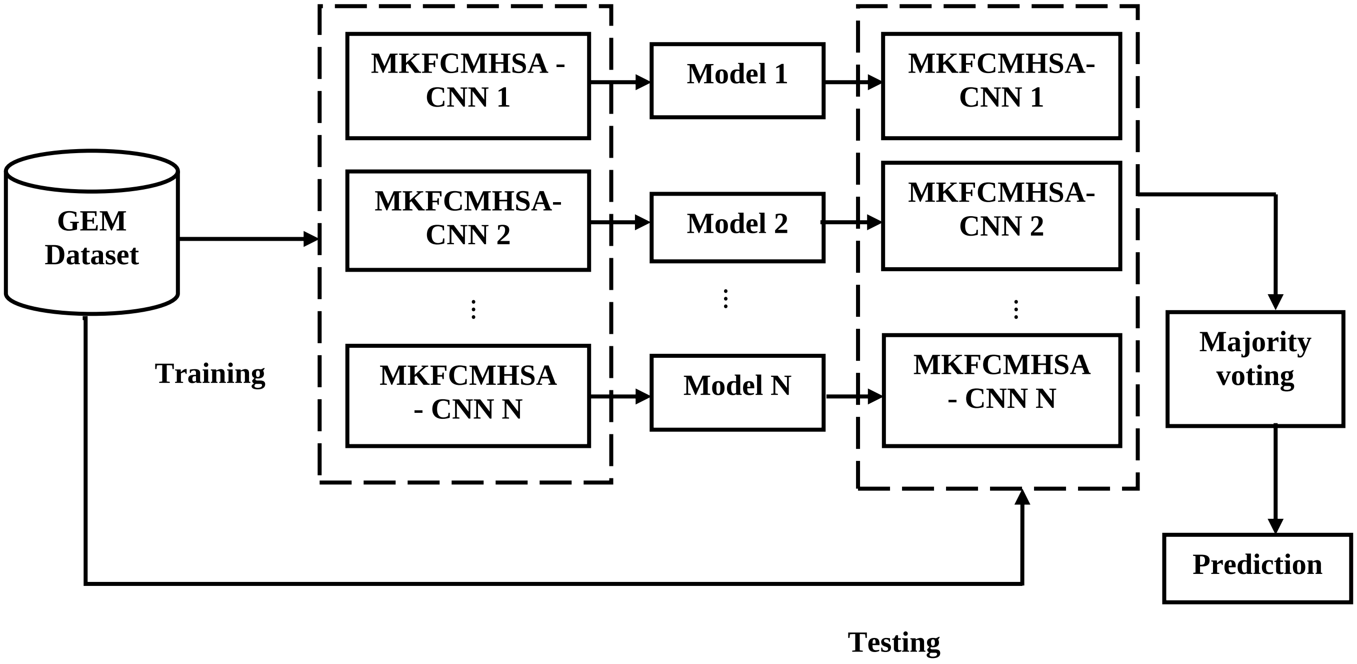


Fig. 3 Stacking ensemble with MHSAMKFC-CNN

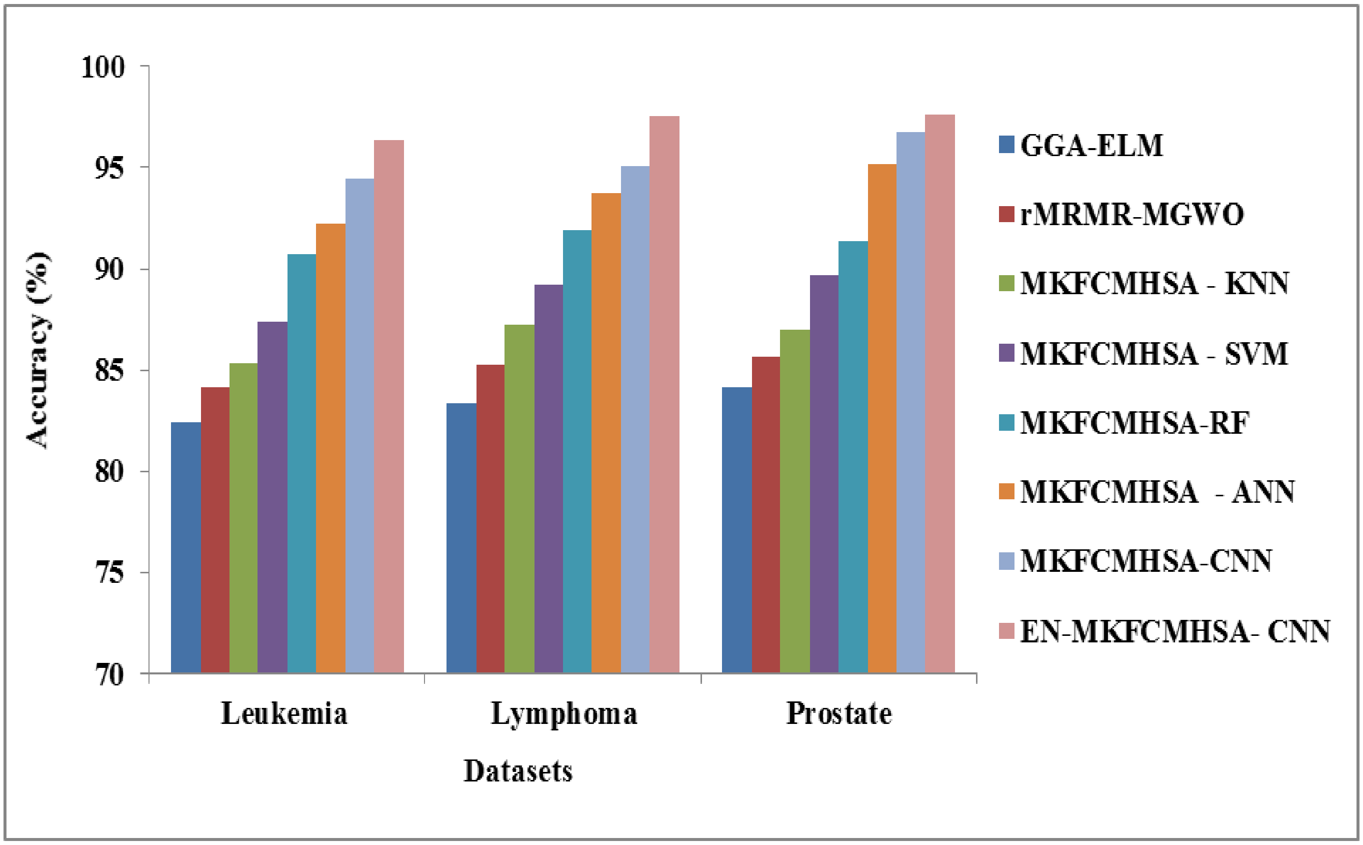


Fig. 4 Comparison of Accuracy

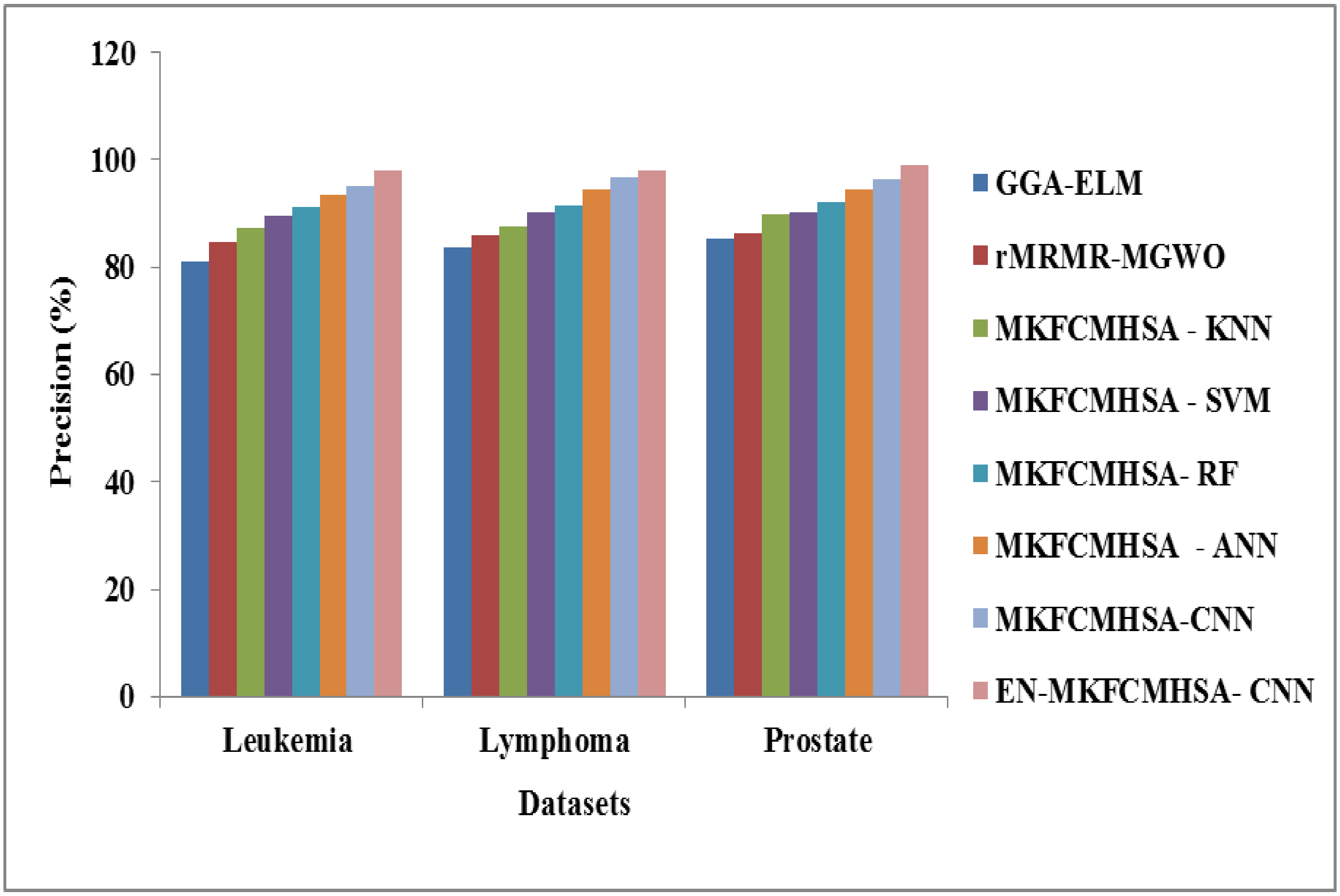


Fig. 5 Comparison of Precision

It is calculated in Equation 14,

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

Table 3 shows the comparison results of precision for proposed and existing methods. Figure 5 displays the precision of existing GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN and EN-MHSAMKFC- CNN techniques.

In this analysis, EN-MHSAMKFC-CNN method is 20.84%, 15.87%, 12.20%, 9.512%, 7.661%, 5.088%, and 3.17% for leukemia dataset; 16.94%, 14.02%, 11.92%, 8.502%, 6.937%, 3.543%, and 1.34% for Lymphoma dataset and 14.58%, 10.16%, 9.563%, 7.307%,4.663% and 2.52% Prostate dataset is higher than that of GGA-ELM, rMRMR-MGWO with proposed MHSAMKFC – KNN, SVM, RF, ANN and CNN methods respectively on given dataset. This analysis shows that the EN-MHSAMKFC-CNN can achieve better precision than other methods for microarray cancer classification.

Table 3. Comparison of Precision

Datasets\ Classifiers	***Leukemia***	***Lymphoma***	***Prostate***
GGA-ELM	81.17	83.70	85.31
rMRMR-MGWO	84.65	85.84	86.38
MHSAMKFC - KNN	87.42	87.45	89.85
MHSAMKFC - SVM	89.57	90.21	90.34
MHSAMKFC - RF	91.11	91.53	92.24
MHSAMKFC ANN	93.34	94.53	94.57
MHSAMKFC-CNN	95.07	96.58	96.54
EN-MHSAMKFC-CNN	98.09	97.88	98.98

5.3. Specificity

Specificity quantifies the rate at which original negatives are accurately identified as such. The formula is as follows in Equation 15:

$$Specificity = \frac{TN}{FP+TN} \qquad (15)$$

Table 4 shows the comparison results of Specificity for proposed and existing methods

Figure 6 displays the Specificity of existing GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN and EN-MHSAMKFC-CNN techniques. In this analysis, EN-MHSAMKFC-CNN method is 16.58%, 15.80%, 13.07%, 10.49%, 8.211%, 4.702%, and 2.168% for leukemia dataset; : 17.91%, 14.42%, 11.26%, 7.983%, 7.155%, 4.632%, and 2.655% for Lymphoma dataset and : 17.57%, 14.43%, 10.00%, 8.326%, 6.219%, 3.821%, 1.228% for Prostate dataset is higher than that of GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN and CNN methods respectively on given dataset. This analysis shows that the EN-MHSAMKFC-CNN can achieve better Specificity than other methods for microarray cancer classification.

Table 4. Comparison of Specificity

Datasets\ Classifiers	*Leukemia*	*Lymphoma*	*Prostate*
GGA-ELM	83.65	82.93	84.10
rMRMR-MGWO	84.21	85.46	86.41
MHSAMKFC - KNN	86.24	87.89	89.89
MHSAMKFC - SVM	88.26	90.56	91.28
MHSAMKFC - RF	90.12	91.26	93.09
MHSAMKFC ANN	93.14	93.46	95.24
MHSAMKFC-CNN	95.45	95.26	97.68
EN-MHSAMKFC-CNN	97.52	97.79	97.68

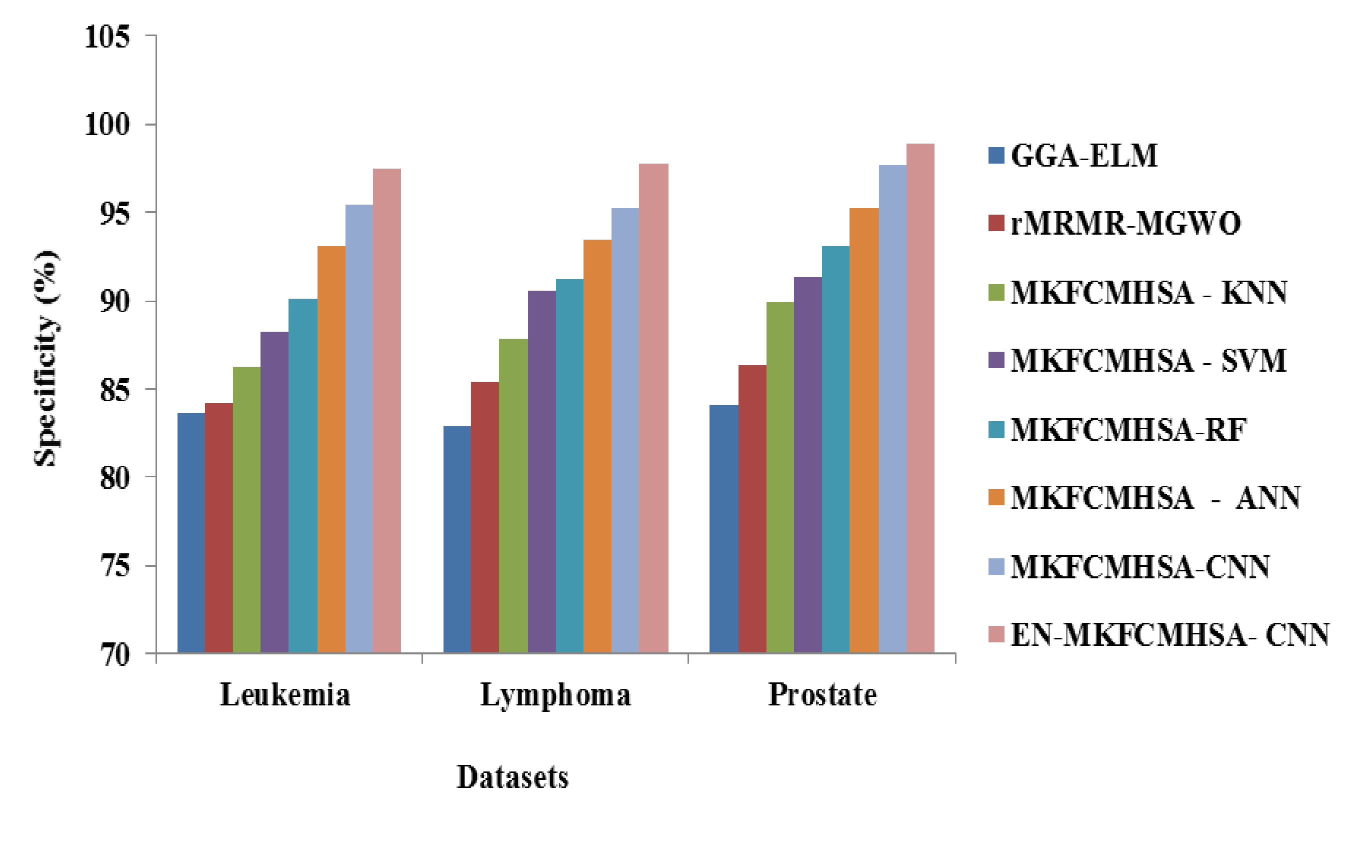


Fig. 6 Comparison of Specificity

5.4. Sensitivity

The definition of sensitivity is the proportion of correctly identified positives (e.g., the percentage of sick people who are correctly identified as having the condition). The formula is as follows in Equation 16:

$$Sensitivity = \frac{TP}{TP+FN} \qquad (16)$$

Table 5 shows the comparison results of sensitivity for proposed and existing methods.

Fig. 7 displays the Sensitivity of existing GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN and EN-MHSAMKFC-CNN techniques. In this analysis, EN-MHSAMKFC-CNN method is 16.58%, 15.80%, 13.07%, 10.49%, 8.211%, 4.702%, and

Table 5. Comparison of Sensitivity

Datasets\ Classifiers	***Leukemia***	***Lymphoma***	***Prostate***
GGA-ELM	84.78	83.17	82.99
rMRMR-MGWO	85.67	85.23	85.19
MHSAMKFC - KNN	87.32	88.45	88.49
MHSAMKFC - SVM	89.39	89.27	90.61
MHSAMKFC - RF	92.51	93.24	92.94
MHSAMKFC ANN	95.12	96.02	93.26
MHSAMKFC-CNN	97.89	98.94	96.19
EN-MHSAMKFC-CNN	84.78	83.17	82.99

2.168% for leukemia dataset; 17.91%, 14.42%, 11.26%, 7.983%, 7.155%, 4.632%, and 2.655% for Lymphoma dataset and 17.57%, 14.43%, 10.00%, 8.326%, 6.219%, 3.821%, 1.228% for Prostate dataset is higher than that of GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN and CNN methods respectively on given dataset. This analysis shows that the EN-MHSAMKFC-CNN can achieve better Specificity than other methods for microarray cancer classification.

5.5. F1-Score

The harmonic mean of precision and recall is the F1 score. It is calculated in Equation 17

$$F1 - score = \frac{2\times precision\times recall}{precision+recall} \quad (17)$$

Table 6 shows the comparison results of the F1-score for proposed and existing methods.

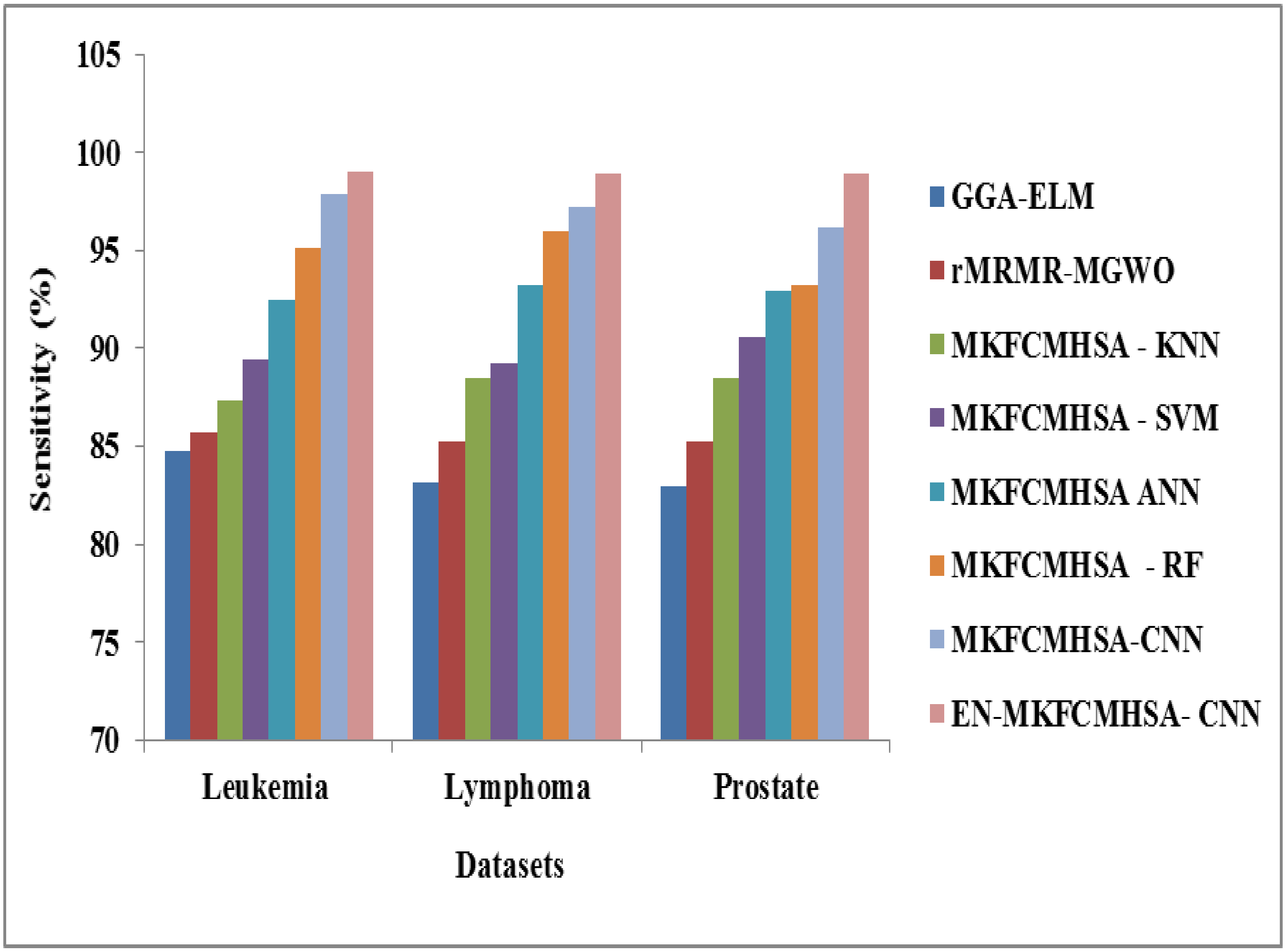


Fig. 7 Comparison of Sensitivity

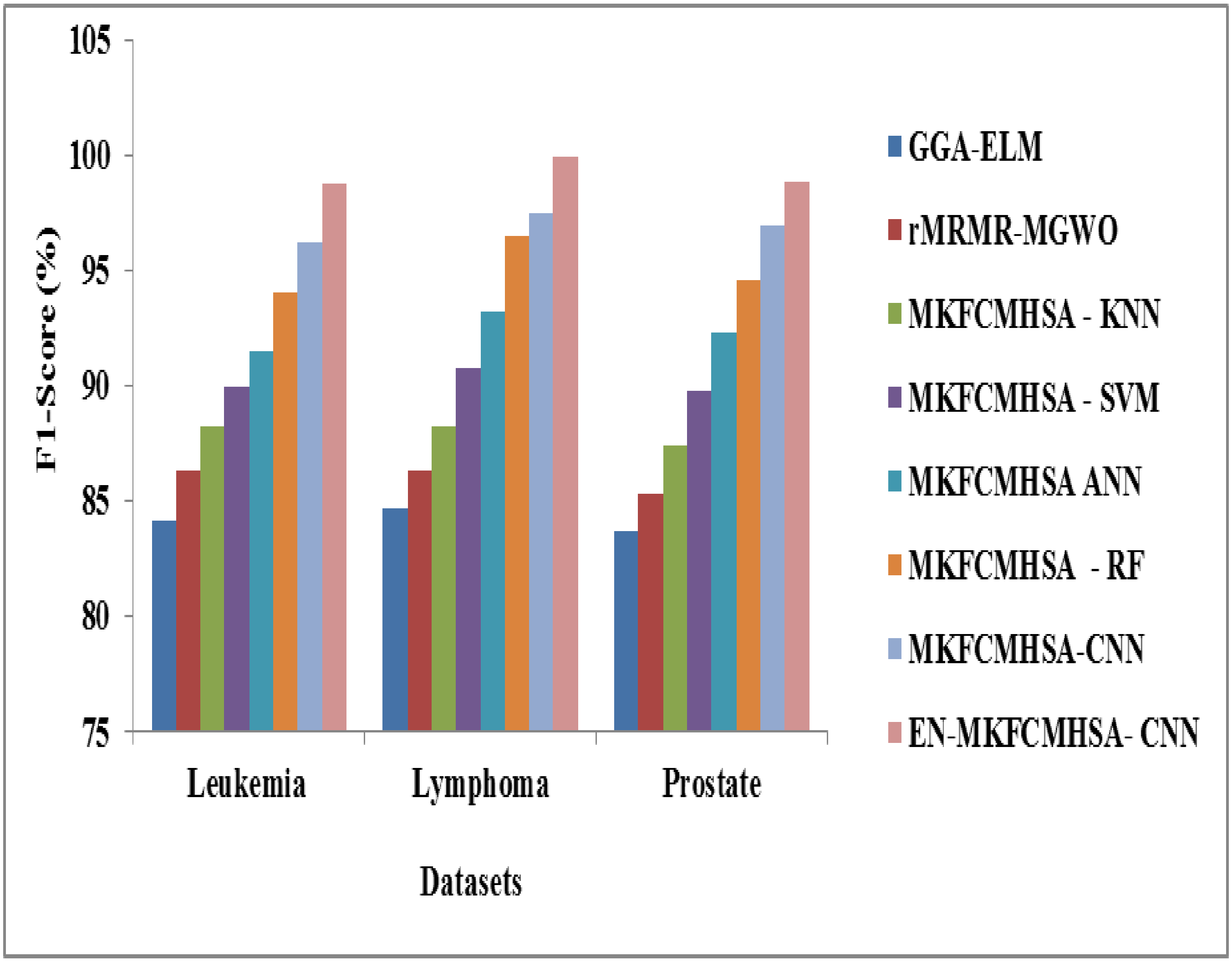


Fig. 8 Comparison of F1-Score

Table 6. Comparison of F1-Score

Datasets\ Classifiers	*Leukemia*	*Lymphoma*	*Prostate*
GGA-ELM	84.10	84.68	83.65
rMRMR-MGWO	86.32	86.33	85.29
MHSAMKFC - KNN	88.24	88.24	87.37
MHSAMKFC - SVM	89.92	90.73	89.78
MHSAMKFC - RF	91.52	93.18	92.31
MHSAMKFC ANN	94.03	96.45	94.57
MHSAMKFC-CNN	96.24	97.46	96.94
EN-MHSAMKFC-CNN	98.75	99.89	98.82

Fig. 8 displays the F1-Score of existing GGA-ELM, rMRMR-MGWO, with proposed MHSAMKFC – KNN, SVM, RF, ANN, CNN and EN-MHSAMKFC-CNN techniques. In this analysis, EN-MHSAMKFC-CNN method is 17.41%,14.39%, 11.91%, 9.819%, 7.899%, 5.019%, and 2.608% for leukemia dataset; 17.96%, 15.70%, 13.20%,10.09%, 7.201%, 3.566%, and 2.493% for Lymphoma dataset; 18.13%, 15.86%,13.10%, 10.06%, 7.052%, 4.494% and 1.939% for Prostate dataset is higher than that of GGA-ELM, rMRMR-MGWO , with proposed MHSAMKFC – KNN, SVM, RF, ANN and CNN methods respectively on given dataset. This analysis shows that the EN-MHSAMKFC-CNN can achieve a better F1-Score than other methods for microarray cancer classification.

6. Conclusion

This research proposes methods for developing an efficient microarray cancer detection system with high classification accuracy results. Initially, MHSAMKFC was developed to handle datasets having large amounts of data without class labels and eliminate redundant features effectively. Then, the MHSAMKFC-CNN method was introduced to eliminate the classification susceptible to errors problem in machine learning methods and reduce the time of CNN classification. Finally, a stacked ensemble model is proposed that uses multiple learning models to produce one optimal predictive model to handle the over-fitting and under-fitting problems of the classifier. To conclude, the experimental results prove that the proposed EN-MHSAMKFC-CNN method has better classification results than other existing methods for cancer prediction.

References

[1] K. D. Miller, A. Goding Sauer, A. P. Ortiz, S. A. Fedewa, P. S. Pinheiro, G. Tortolero-Luna, And R.L. Siegel, "Cancer Statistics for Hispanics/Latinos," *Ca: aCancer Journal for Clinicians*, vol.68, no.6, pp.425-445, 2018.

[2] J. D. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, And N. Papadopoulos, "Detection And Localization of Surgically Resectable Cancers with A Multi-Analyte Blood Test," Science, vol.359, no.6378, pp.926-930, 2018.

[3] S . Farjana Farvin, And S . Krishna Mohan., "A Comparative Study on Lung Cancer Detection Using Deep Learning Algorithms," *SSRG International Journal of Computer Science And Engineering*, vol.9, no.5, pp.1-4, 2022, *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V9I5P101.

[4] A.K Shukla, P. Singh, And M. Vardhan, "Gene Selection for Cancer Types Classification Using Novel Hybrid Metaheuristics Approach," *Swarm And Evolutionary Computation*, vol. 54, pp.100661, 2020.

[5] J. H. Bae, M. Kim, J. S. Lim, And Z. W. Geem, "Feature Selection for Colon Cancer Detection Using K-Means Clustering And Modified Harmony Search Algorithm," Mathematics, vol.9, no.5, pp.570, 2021.

[6] C. Y. Yu, Y. Li, A. L. Liu, And J. H. Liu, "A Novel Modified Kernel Fuzzy C-Means Clustering Algorithm on Image Segmentation," *In 2011 14th Ieee International Conference on Computational Science And Engineering IEEE*, pp. 621-626, 2011.

[7] R. R. Rani And D. Ramyachitra, "Microarray Cancer Gene Feature Selection Using Spider Monkey Optimization Algorithm And Cancer Classification Using Svm," *Procedia Computer Science*, vol.143, pp.108-116, 2018.

[8] B. Lyu And A. Haque, "Deep Learning Based Tumor Type Classification Using Gene Expression Data," *In Proceedings of the 2018 Acm International Conference on Bioinformatics, Computational Biology And Health Informatics*, pp.89-96, 2018.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh And D. Batra, "Grad-Cam: Visual Explanations From Deep Networks Via Gradient-Based Localization," *In Proceedings of the IEEE International Conference on Computer Vision*, pp.618-626, 2017.

[10] M. Mollaee, And M. H. Moattar, "A Novel Feature Extraction Approach Based on Ensemble Feature Selection and Modified Discriminant Independent Component Analysis for Microarray Data Classification," *Biocybernetics And Biomedical Engineering.* Vol.36, no.3, pp.521-529, 2016.

[11] S. Guo, D. Guo, L. Chen And Q. Jiang, "A Centroid-Based Gene Selection Method for Microarray Data Classification," *Journal of Theoretical Biology*, vol.400, pp.32-41, 2016.

[12] B. Sahu, S. Dehuri And A. K Jagadev, "Feature Selection Model Based on Clustering And Ranking In Pipeline for Microarray Data," *Informatics In Medicine Unlocked*, vol.9, pp.107-122, 2017.

[13] S. Guo, D. Guo, L. Chen And Q. Jiang, "A L1-Regularized Feature Selection Method for Local Dimension Reduction on Microarray Data," *Computational Biology And Chemistry*, vol.67 , pp.92-101, 2017.

[14] M. K. Ebrahimpour, H. Nezamabadi-Pour And M. Eftekhari, "Ccfs: A Cooperating Coevolution Technique for Large Scale Feature Selection on Microarray Datasets," *Computational Biology And Chemistry*, vol.73, pp.171-178, 2018.

[15] Z. Y. Algamal, R. Alhamzawi And H. T. M. Ali, "Gene Selection for Microarray Gene Expression Classification Using Bayesian Lasso Quantile Regression," *Computers In Biology And Medicine*, vol. 97, pp. 145-152, 2018.

[16] B. Cao, J. Zhao, P. Yang, P. Yang, X. Liu, J. Qi And K. Muhammad, "Multiobjective Feature Selection for Microarray Data Via Distributed Parallel Algorithms," *Future Generation Computer Systems*, vol.100, pp.952-981, 2019.

[17] M. Yuan, Z. Yang And G. Ji, "Partial Maximum Correlation Information: A New Feature Selection Method for Microarray Data Classification," *Neurocomputing*, vol.323 , pp.231-243, 2019.

[18] H. Wang, L. Tan, And B. Niu, "Feature Selection for Classification of Microarray Gene Expression Cancers Using Bacterial Colony Optimization with Multi-Dimensional Population," *Swarm And Evolutionary Computation*, vol.48, pp.172-181, 2019.

[19] S. Li, K. Zhang, Q. Chen, S. Wang, And S. Zhang, "Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors And Genetic Algorithm," *IEEE Access*, vol.8, pp.139512-139528, 2020.

[20] N. Ilc, "Weighted Cluster Ensemble Based on Partition Relevance Analysis with Reduction Step," *IEEE Access*, vol.8 , pp.113720-113736, 2020.

[21] P. García-Díaz, I. Sánchez-Berriel, J.A. Martínez-Rojas, And A. M. Diez-Pascual, "Unsupervised Feature Selection Algorithm for Multiclass Cancer Classification," *Genomics*, vol.112, no.2, pp.1916-1925, 2020.

[22] M. Mohammed, H. Mwambi, I. B. Mboya, M. K Elbashir, And B. A. Omolo, "Stacking Ensemble Deep Learning Approach To Cancer Type Classification Based on Tcga Data," *Scientific Reports*, vol.11, no.1, pp.1-22, 2021.

[23] O. A. Alomari, S. N. Makhadmeh, M. A Al-Betar, Z.A.A Alyasseri, I. A. Doush, A. K. Abasi, 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.