An Ensemble Deep Learning for Cancer Classification Using Unsupervised Deep Feature Extraction

Ms. R. Madhubala Shanmu[1]
Student of M.E
Department of CSE
Francis Xavier Engineering
College, Tirunelveli.

Mrs. P. Brundha[2]
Associate Professor
Department of CSE
Francis Xavier Engineering
College, Tirunelveli.

Dr.G.Aravind Swaminathan
Professor
Department of CSE
Francis Xavier Engineering
College, Tirunelveli.

Dr. R. Tino Merlin[4]
Associate Professor
Department of CSE
Francis Xavier Engineering
College, Tirunelveli.

Abstract - Antibody microarrays in a patent-pending series may be initiated and validated using microarray technology. A Microarray Data Analysis (MDA) is used to determine the patterns of hundreds of genes within a single experiment. There is a vast amount of gene expression data in the MD that may be used to diagnose malignancy. However, over-fitting and under-fitting issues arise due to the unbalanced class label instances present in microarray gene datasets and the initialization parameter value for the classifier. To get over this obstacle, this study proposes a stacking ensemble of Deep cluster-based DL systems for Cancer Classification. This system combines many learning models into a single, highly accurate prediction model. There are three distinct parts to the created model. To begin, we create a Modified Harmony Search Algorithm and a Modified Kernel-based Fuzzy C-Means (MHSAMKFC) to efficiently remove massive duplicate features. Second, to deal with uncertainties in the labeled training dataset and boost classifier performance, the MHSAMKFC with Convolutional Neural Network (CNN) classifier is suggested. Third, the ensemble technique, which employs several learning models to improve prediction accuracy, mitigates MHSAMKF0C over-fitting CNNs and under-fitting issues. En-MHSAMKFC-CNN describes the whole operation. In a conclusion, experiments are run on four Gene Expression Microarray (GEM) datasets to confirm that the En-MHSAMKFC-CNN enhances 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

Cancer is the second biggest cause of death worldwide, accounting for one out of every six deaths [1]. It is possible to reduce cancer-related mortality rates with early diagnosis and treatment. It is crucial to describe the unique characteristics of cancer valetudinarians, and patient-specific treatment plans are arranged due to the fact that indications differ from case to case. These characteristics can be most reliably extracted from the patient's genetic data. Thanks to significant developments in MD processing research during the last decade, it has become a useful tool for disease diagnosis [2]. Using microarrays based on genetic information, clinical pathology may identify, explain, and classify human illnesses like cancer. Cancer patients would benefit from earlier and more accurate diagnosis since it would lead to more effective therapy and more responsive malignancies.

Genealogical data generated by DNA microarrays is massive, and although some of it may be beneficial in the detection of cancer, the vast majority of it is both meaningless and noisy. Old, irrelevant, and distracting genomes lower the quality of data sets. Clinical framework development for the illness requires approaches to gene selection. especially when there aren't enough samples to go around [3]. Using a novel hybrid metaheuristic approach dubbed Training learning-based algorithm (TLBO) and Gravitational Search Algorithm (GSA), TLBOGSA [4] was developed for cancer classification (GSA). The search potential during the development stage is enhanced when gravitational search techniques are integrated with the instruction phase. However, this Feature Selection (FS) could not be effective in finding relevant genes because of the high complexity and low sample size of GEM data.

The MHSA [5] is an initiative with the goal of solving the dimensional curse problem by simplifying the process of locating important genes. However, whenever the Pitch Adjustment Rate (PAR) value is very near to zero, the algorithm's convergence speed may stall in the last rounds of the optimization process. The inability of traditional FCM to handle even slight discrepancies across clusters is addressed by the MKFC technique [6]. This method, however, is very vulnerable to noisy data, which often results in less useful genes. This research addresses these issues by integrating MHSA and MKFC for FS from array cancer datasets. Datasets are handled systematically using the MHSAMKFC technique.

It having a lot of data without a class label and being able to efficiently get rid of superfluous features.
Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and Artificial Neural Network (ANN) classifier are only few of the Machine Learning (ML) methods that have been used to categorize a sizable quantity of MD in the literature. Wrapper FS outperforms filter-based FS approaches in feature selection and classification processing. Two goals are met by a support vector machine (SVM) based classification and spider monkey optimization based FS [7]. Initially, we want to boost classifier accuracy while simultaneously decreasing the number of parameters. However, ML's cancer prediction algorithm is still difficult to use with little data and is very vulnerable to imprecision.

Thus, the DL Based cancer type classification [8] was developed to categorize the bigger GEM datasets. Genes that showed little variation across all of the data were eliminated using a Deep Neural Network (DNN) and several visualization techniques. To make use of the Convolutional layers, the high-dimensional information about the expressions was then combined into a 2-dimensional space. A three-layer convolutional neural network (CNN) with a learned neural structure was developed using the Guided Grad-[9] Cam's approach to choose landmark genes for categorization. As a consequence of using this strategy, accurate prognoses of cancer are more likely to be made. However, this causes a significant issue with computing time.

To boost the accuracy of the classification procedure, a DL-based Unsupervised CNN classifier is implemented. This classifier is trained to discover patterns in the data that allow for an accurate reconstruction of the training samples. For the purpose of minimizing the dimensionality reduction and uncertainties in the labeled training MD, this Unsupervised CNN is combined with the MHSAMKFC technique to generate a single optimum predictive model.

Over-fitting and under-fitting occur during cancer classification due to the MHSAMKFC-unbalanced CNN's class label occurrences in datasets and initialized parameter values for the classifier. As a means of addressing these concerns, MHSAMKFC-CNN has been augmented with a stacking ensemble, which combines different learning models to improve prediction accuracy. Using majority voting, the unlabeled data will be assigned to the class with the most votes among the predictions of the CNN classifiers in this Ensemble model. The suggested technique outperforms the gold-standard classifier on GEM Datasets for predicting cancer subtypes.

2. Literature survey

For the purpose of MD classification, an ensemble FS and enhanced discriminant principal components analysis feature extraction technique was designed [10]. In any case, the

The effectiveness of classifiers is heavily impacted by the thresholds chosen during pre-processing of datasets. For the purpose of organizing the microarray data, a centroid-based DNA choosing approach [11] was designed. However, when the number of data characteristics increased, the method's precision deteriorated.

In order to choose the best characteristics for MD, a methodology [12] was created. This model ranked features for importance and utilized attributed grouping in the pipeline to get rid of noise. However, no attempts were made to rectify the dataset's imbalance if any were found. In order to accomplish local dimension reduction and classification of MD, a two-stage local dimensionality approach was proposed [13]. However, the precision of the two-stage local dimension method depends on the regularization value.

In [14] we described a Cooperative Co-evolution approach to FS (CCFS) that may be used in MD. Using the filter criteria in the objective function, a bidirectional gravitation search algorithm was used to explore the solution space according to the concept of coevolution theory. Unfortunately, this method required a lot of processing power to implement. In [15], a Bayesian Lasso quintile regression approach was introduced to characterize gene expression for GEM selection. Combining a skewed Laplace distribution for flaws with a graded hybrid of regular probability for regression coefficients, this technique is based on Bayesian MCMC assessment.

Through the use of dispersed parallel algorithms, a multiobjective instance selection model was developed [16] for MD. To better categorize the MD, this model chooses the most relevant features based on a variety of criteria, including feature number, classification error, and feature redundancy. However, it's possible that different objectives may collide with one another. The categorization of MD was given using a Partial Maximum Correlation Information (PMCI) approach [17]. To evaluate the importance of each feature, the perpendicular components were recovered from the attribute space. This approach, however, has a low F1 score. In order to extract and classify features from microarray gene expression cancer data, a discontinuous Bacterial Colony Optimization with a multi-size population (BCO-MDP) method was created [18]. On the other hand, without previous knowledge of datasets, it was difficult to discover an appropriate search space for high classification accuracy.

Attribute selection for high-dimensional data using Weight K-NN (WKNN) and GA was created [19]. The input degrees of the feature value was used in conjunction with GA to determine the optimal weighted sum for the involvement of the value in the element to the classification. The main drawback of this approach is the significant computational complexity it entails. Partition Relevant Analysis (PRA) and a reduction procedure were used to illustrate a balanced group hybrid technique [20]

In the second stage of PRA, we use methods of data dimensionality reduction to get rid of redundant and noisy indices. While promising, this approach requires more development before it can be used for complicated strategy functions. To address a highly nuanced problem in microarray data categorization, [21] researchers created a Grouping Genetic Algorithm with Extreme Learning Machine (GGA-ELM). In contrast, bigger datasets have little effect from this approach.

In [22], a stacking ensemble DL method based on a One-Dimensional Convolutional Neural Network (1D-CNN) approach was presented for cancer type prediction using TCGA data. The number of genes was cut down using the FS method of Least Absolute Shrinkage and Selection Operator (LASSO) regression. On the downside, this strategy had a heavy computational cost. To identify the best genes to prioritize when classifying microarray data, the Modified Gray Wolf Optimizer (MGWO) was used to create the resilient Minimum Redundancy Maximum Relevancy (rMRMR) filter approach [23]. Contrarily, the suggested mixture yields subpar categorization performance.

3. Proposed methodology

The genes themselves are the focus of gene expression studies. The process of gene selection involves locating the genes that are strongly linked to a certain group. One of the advantages of this method is that it may reduce the dimensionality of the dataset. Even more so, the application of classification renders many genes superfluous. Using gene selection reduces the possibility that irrelevant genes would be drowned out. As far as methods go in MDA, FS and clustering are at the top of the list. Therefore, the goal of this study is to propose an MHSAMKFC method for tackling the dimensionality issue on MD and picking relevant genes. The following is a condensed explanation of how this algorithm works.

MHSAMKFC

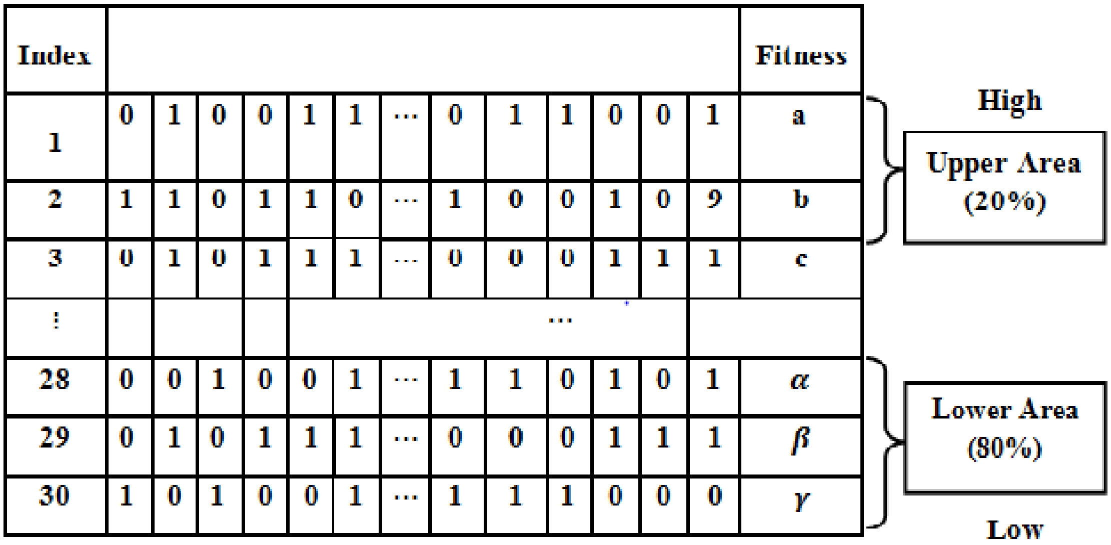
As briefly shown in [10], an MHSA is created for the FS procedure by adapting the preexisting HS.

Step 1 Constructing variables and Harmony Memory(HM)

To begin an HM project, it is necessary to set goals, choose a solid foundation from which to build, and develop harmony among all involved parties. It's not possible to use this method effectively without first grasping the significance of the parameters. Since HS is also an evolutionary algorithm, it may be likened to GAs. The genes in a chromosome of a GA are the most important and fundamental portions of the Hv. HM Size refers to the number of harmonies included inside a single HM (HMS). The initial harmony values, Hv, are chosen at random in the HS technique, and only a few of these values are used in the iterative process.

Step 2 Forming New Harmony by separating HM

The process of creating a brand-new HM is quite similar to the current HS algorithm. Still, the HM will be split in half as shown in Figure 1 so that the observation may be made. The uppermost area is made up of the top 20% fittest harmonics within a particular HM. We don't utilize HMCR or PAR here since they need too much work. In this way, New Harmony is not included in the activation procedure. When the combination is recombined inside the harmony of the upper region, a combination of higher fit may be discovered, and then new harmonies would be produced. Second, in HM's lower register, you'll find HMCR and PAR's most recent harmonies.



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

Step 3 Updating HM

Goodness-of-fit measures how well the chosen classification model fits with the harmonic choice. The fitness is determined by the sequence of harmony values that maximizes fitness. Second, the scale of the given HMS is adjusted by correcting and removing the long-standing binary harmonics with the smallest fit.

Step 4 Iterating previous Steps 2 and 3

There is currently no improved procedure. Iterate through steps 2 and 3 as many times as necessary. As the total number of trials increases, the upper part finds harmony at a higher fitness level inside the combination with more appropriateness. The advantages of the original HS, such finding new combinations by considering variations, are maintained in the bottom part. Two sites inside a single HM have their highest classification performance recorded 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^r_d is the distance between two features belonging to two different clusters, whereas the Intra cluster's distance I^s_d is the distance between two features belonging to the same cluster, which is defined as follows

$$I^r_d = \sum_{i=1}^{N} \sum_{j=1}^{C} \sqrt{x_i - c_j} \quad (1)$$

$$I^s_d = \sum_{i=1}^{N} \sum_{j=1}^{C} \sqrt{c_i - c_j} \quad (2)$$

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.

In order to remove superfluous information from the provided datasets, the gathered features are subjected to a clustering algorithm after the feature selection procedure. To improve upon the traditional FC technique, the MKFC algorithm incorporates kernel information into the calculation. It was developed to remedy the inefficiency of the FC algorithm in handling incremental changes inside clusters. The kernel method takes a non-linear input data structure and transforms it into a high-dimensional feature space.

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 dimensionsis warranted.

Prototypes developed in the attribute space are the main kind of MKFCM. These clustering techniques will be referred to as MKFCM-F. (with F standing for the feature space). In the second class, designated MKFCM-K, the prototypes are kept in the K and must be mimicked in the feature space by means of a mapping in reverse from the kernel space to the feature space. The hypotheses in the MKFCM method are conveniently kept in the feature space and are then implicitly projected to the kernel space by use of the kernel operator.

As a matter of course, this is because the inner development of the transform function, i.e. the kernel space, may be tackled with only the help of additional kernel functions that are already known. When the concepts oi are generated in the kernel space, we refer to this version of MKFCM as MKFCM-K. Building kernel space is the primary goal of Equations 3, 4, and 5.

$$Q = \sum_{i=1}^{c} \sum^{N} u^m \| \varphi(\ ) - o^{\ 2} \quad (3)$$

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

$$d\varphi_{ij}^2 = k(x_j x_j) - \frac{2\sum_{h=1}^{n} u_{ih}^{m} (x_h x_j)}{\sum_{h=1}^{n} u_{ih}^{m}} + \frac{\sum_{h=1}^{n} \sum_{l=1}^{n} u_{ih}^{m} k_2(x_h x_l)}{\sum_{h=1}^{n} u_{ih}^{m}} \quad (5)$$

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

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

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

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

Here, $(X_j,\ O_i)$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

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

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 limitFor ($i = 1: i \leq HMS$), then
Develop primary harmony ($x_{new}$)
Perform Algorithm 2 to obtain cluster and $Hf$
End forRepeat
For (J = 1: N) then          //HS in upper area
 $x_{new}$ = Arbitrarily chosen from $_{(BDR+1)}$to $x_{(HMS)j}$
end for
Create New Harmony ($x_{new}$)
Perform Algorithm 2 to obtain cluster and $Hf$
If ((0,1) $< HMCR$) then //HS in lower areaFor($J = 1: N$)
 then
 $x_{new}$= Randomly select from$_{(BDR+1)J}$to $x_{(HMS)j}$
 If ((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(fit($HM_{new(upper,lower)}$) < fit( $HM_{old}$))
Update HMEnd 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 $u^0_{ik}$
$J_m = \sum_{i=1} \sum_{k=1}^{N} u^0_{ik} \|X_k - V_i\|^2$
For t =1, 2…, $t_{max}$, do:
        (a) Upgrade all prototypes $V_i^t$
        (b) Upgrade all memberships $U_{ik}^t$
Compute $E^T = max_{i,} | U^t_{ik} - U^{t-1}_{ik} |$, If$E^T \leq \varepsilon$,
$U \in \{u_{ik} \in 0,1 \mid \sum_{i=1}^{c} u_{ik} = 1 \; \forall k ; \; O < \sum_{k=1}^{N} u_{ik} < N, \forall i\}$
Stop: $else \; t = t + 1$ \\ number of clusters is obtained
```

To verify the effectiveness of suggested FS and clustering methods, the obtained features are passed to classifiers including SVM, KNN, RF, and ANN for accurate cancer classifications. The downside is that these machine-based techniques are time-consuming. difficulty in identifying patterns in the data and determining how to categorize it. As a result, classification of GEM datasets has been performed using a DL architecture.

MHSAMKFC-CNN

Key to the success of this approach is the employment of an Unsupervised Convolutional Neural Network (CNN) to update the cluster centers based on a dependable FS after the data features have been acquired from the classifier. Figure 2 provides a quick summary of how a CNN equipped with the suggested MKFCMHS contributes to the effective performance of the clustering Algorithm.

Several pieces of information regarding the gene expression data are encoded in the CNN codes (the layer activations in a CNN prior to classification, which may include non-linearity). They have proven useful as characteristics for numerous classification applications using gene expression data. In this effort, we go farther in exploring how the distinct layers react to various types of pictures.

CNN employs the MHSAMKFC method to cluster layer activations. This method is useful for preserving cluster centers. A particle is local to a cluster if its average distance from the cluster's centroid is less than its average distance from any other centroid. By alternating between (1) allocating data points to categories based on the current centroids and (2) assigning data points to categories based on the actual centroids, MHSAMKFC-CNN is able to experimentally identify the best centroids. (2) Selecting an epicentre (centroid) for the cluster from the preexisting grouping of data points. In order to create MHSAMKFC-CNN, a dataset will be used as a guide. $D \in \mathbb{R}^{d\times k}$of k vectors (i.e., centroids), so thus that a data matrix $x_i \in R^d i = 1, \ldots, m$ 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.

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

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

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

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

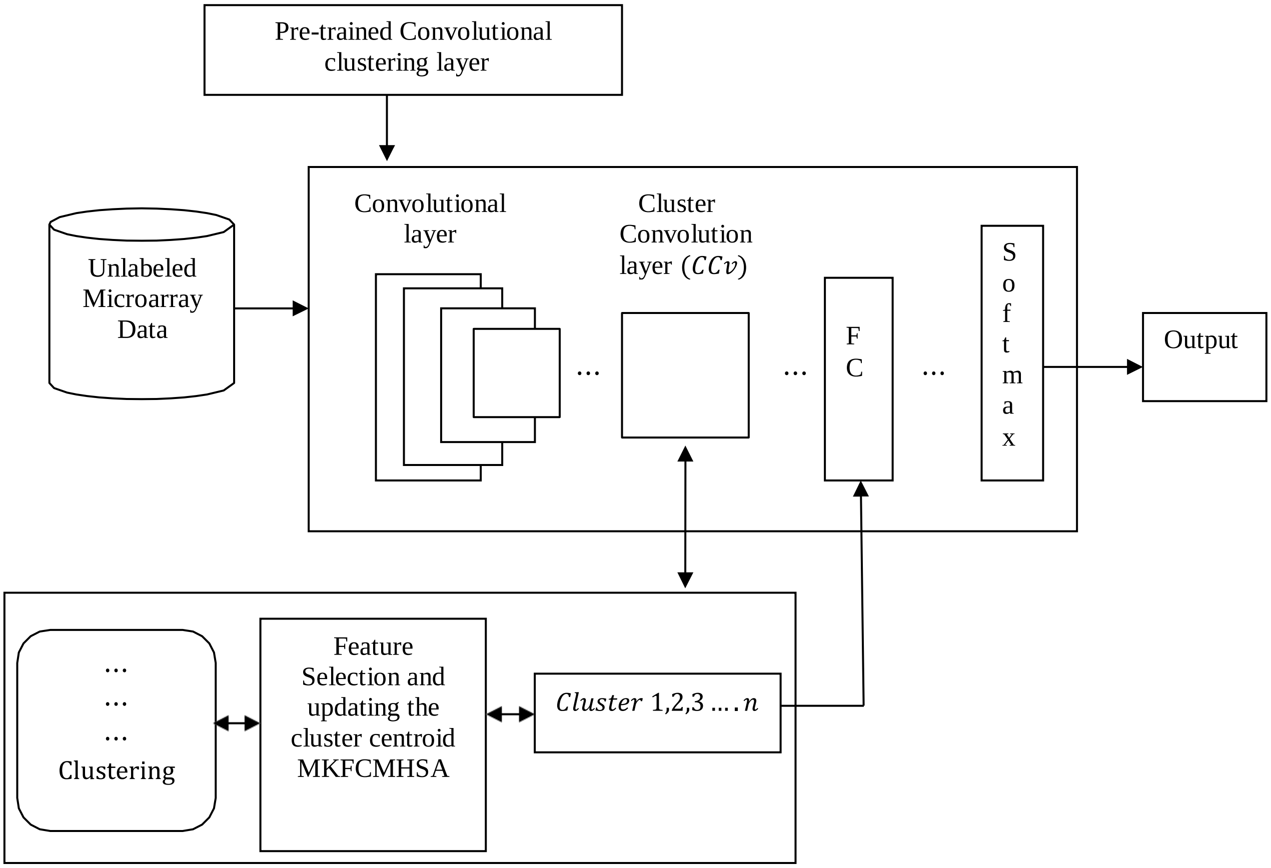

Pre-trained Convolutional clustering layer
Unlabeled Microarray Data
Convolutional layer
Cluster Convolution layer (CCv)
…
…
F C
…
S o f t m a x
Output
…
…
…
Clustering
Feature Selection and updating the cluster centroid MKFCMHSA
$Cluster$ 1,2,3 … . n

Fig. 2 Structure for the CNN layer with MHSAMKFC System

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} (\mathrm{f}\,(x_i\,, \mathrm{w}), \hat{y}_i) \qquad (12)$$

The cross entropy loss L is minimized using Stochastic Gradient Descent (SGD), a technique also used in standard CNN backpropagation. During the last round of training using MHSAMKFC, surrogate labels are created from CNN features (see Equation 9) and utilized to fine-tune the parameters of the CNNs (see Equation 12). This procedure is repeated until the clustering and failure have reached a steady state. To begin, a unique method is developed to determine what traits are taught at each successive layer.

1. First, decide on n as the total number of groups.
2. Second, from the MM, choose k-sized subsets representing each class. Therefore, there are nk characteristics in all.
3. Third, the data is fed into the pre-trained network, and the activations of all of its layers are recorded. The Di for the analysis at layer I will be nk activations, which will be detailed in the next section.
4. According to the MHSAMKFC-CNN algorithm, the t Di is divided into n clusters at layer .
5. Analyze the clusters obtained at each layer concerning the original classes to which the corresponding features belong.

Ensemble of MHSAMKFC-CNN

The meta-learner is a prototype that learns to enhance the predictions of the base-learners and produces the end result; it is used to increase classifier efficiency by merging the efforts of sub-models trained to address the same classification problem. As a consequence, the ensemble method outperforms individual learners in terms of prediction performance on the MD for Cancer classification. The capacity to generalize the results of an ensemble improves prediction accuracy and guarantees a stable, high-quality forecast. By using the output of the MHSAMKFC- CNN sub-models with different variables as input, the Meta model learns to combine the predictions and provide a better final prediction than each of the basic classifiers. Figure 3 depicts the recommended stacking ensemble DL algorithm for the cancer prediction technique on MD..

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 $t = 1, . . . , T$ : %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 database5. For $i = 1, . . . , m$:
6.    For $t = 1, . . . , T$ :
7.       $z_{it} = h(x_i)$;
8.    end
9.    $D * = D * \cup ((z_{i1}, ….., z_{iT}), y_i)$;
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: $(x) = h^*(h_1(x), . . . , h_T(x))$
```

4. Dataset Description

Using MATLAB 2018a, we test the efficacy of both the current and the planned GEM datasets for cancer prediction. It has 4GB of RAM, an Intel CPU running at 2.70 GHz, and Microsoft Windows 7. To conduct experiments, we gather three GEM datasets, including data on leukemia, lymphoma, and prostate microarray. Table 1 contains links to publicly accessible online datasets. For every 100 pieces of data that are gathered, only 40 are utilized 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

Methods like KNN, SVM, RF, ANN, and CNN are used to analyze the efficacy of existing methods like GGA-ELM [21] and rMRMR-MGWO [23] as well as proposed methods like MHSAMKFC and EN-MHSAMKFC. Below is a quick explanation of the five metrics used to evaluate a performance: accuracy, precision, specificity, sensitivity, and F1 score.

Accuracy

Accuracy is defined as the proportion of examples that fit into the intended categories. It is calculated by dividing the number of correct classifications (both positive and negative) by the total number of persons classified. Equation 13 is used to determined.

$$Accuracy = \frac{TP+}{TP+TN+FP+FN} \qquad (13)$$

When it comes to cancer, TP stands for those who have been accurately diagnosed as sick, whereas FP represents those who have been incorrectly diagnosed as ill. People who have been determined to be healthy are denoted with the letter TN. FN stands for false negatives, or those who are really ill but are misdiagnosed. Accuracy results for proposed and current approaches are compared in Table 2.

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

– Accuracy of the current GGA-ELM and rMRMR-MGWO is shown in Fig. 4, along with the suggested MHSAMKFC -KNN, SVM, RF, ANN, CNN, and EN-MHSAMKFC-CNN methods. For the leukemia dataset, the EN-MHSAMKFC-CNN method outperformed GGA-ELM and rMRMR-MGWO by 16.88%; for the lymphoma dataset, it outperformed them by 17.02%; for the prostate dataset, it outperformed them by 17.14%; and for the melanoma dataset, it outperformed them by 7.926%. Evidence from this study demonstrates that the EN- MHSAMKFC-CNN outperforms competing approaches for microarray cancer classification.

Precision

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

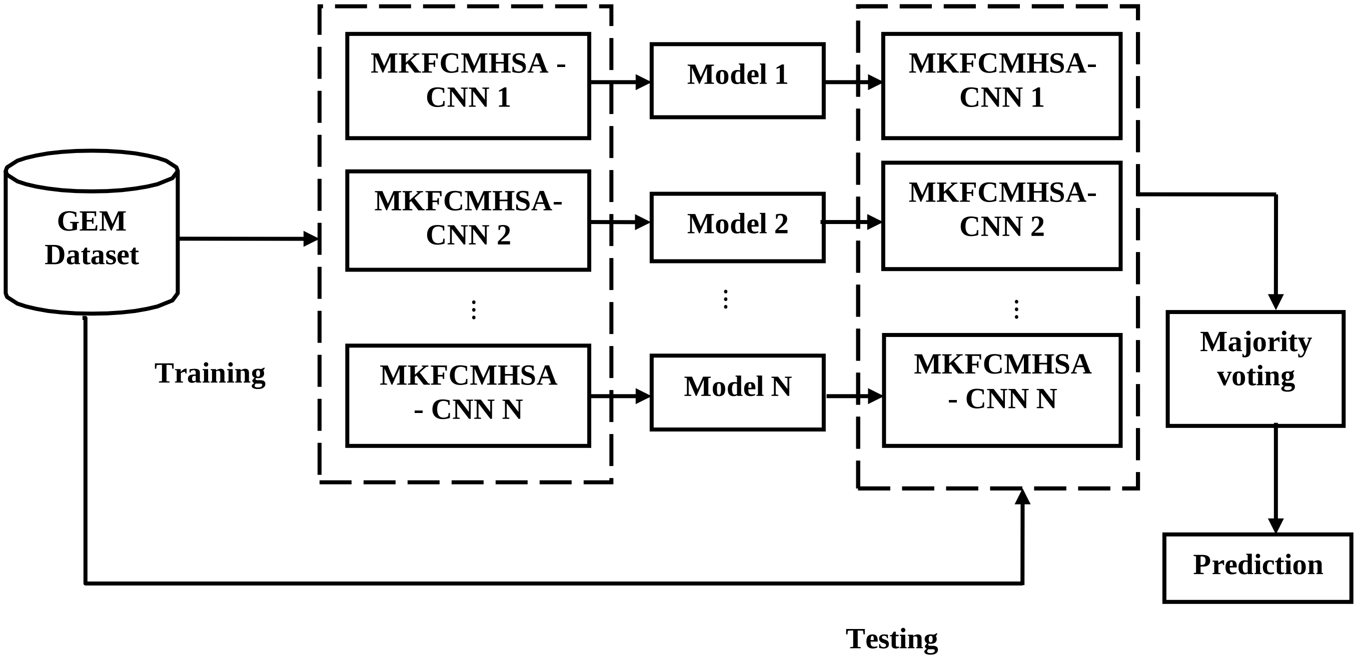


Fig. 3 Stacking ensemble with MHSAMKFC-CNN

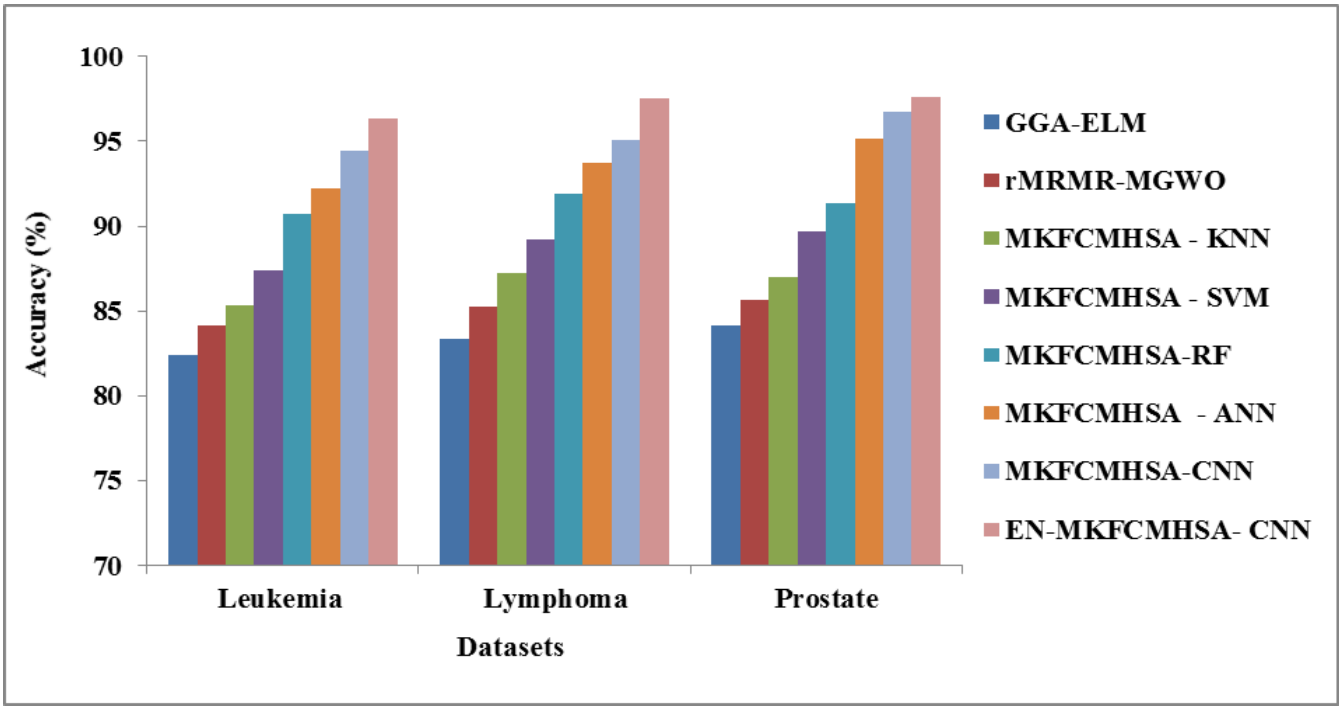


Fig. 4 Comparison of Accuracy

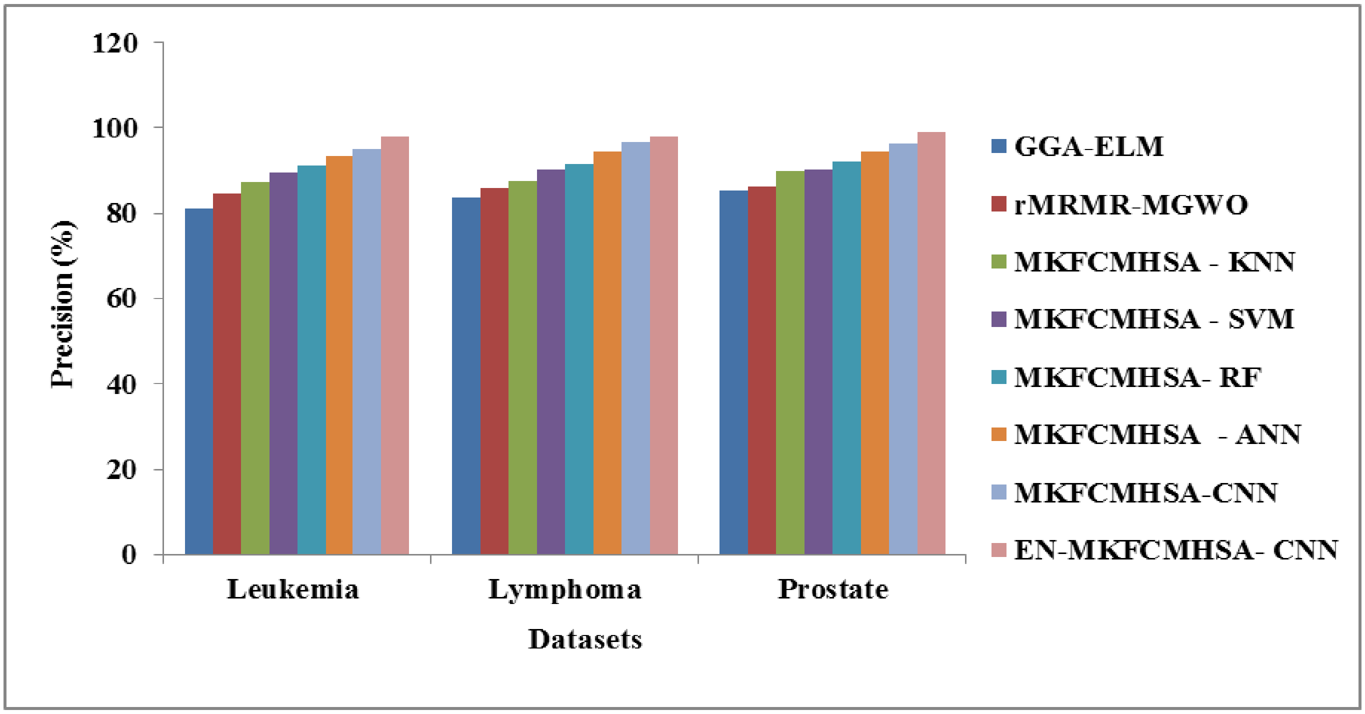


Fig. 5 Comparison of Precision

It is calculated in Equation 14,

$$Precision = \frac{TP}{TP+FP} \qquad (14)$$

The accuracy comparisons of the proposed and current approaches are shown in Table 3. Compare the accuracy of the currently used GGA-ELM and rMRMR-MGWO to that of the proposed MHSAMKFC- KNN, SVM, RF, ANN, CNN, and EN-MHSAMKFC- CNN in Figure 5.

The results for the EN-MHSAMKFC-CNN approach are as follows: 20%, 15.75%, 12.20%, 9.512%, 7.661%, 5.088%, and 3.17% for the leukemia dataset; 16.9%, 14.02%, 11.92%, 8.502%, 6.937%, 3.543%, and 1.34% for the lymphoma dataset; 14.58, 10.16, 9.563, 7.307, 4.66, and 2.52% for the lymphoma dataset

On the prostate dataset, the suggested MHSAMKFC - KNN, SVM, RF, ANN, and CNN techniques all perform better than GGA-ELM, rMRMR-MGWO. The results of this study demonstrate that the EN-MHSAMKFC-CNN outperforms competing approaches for microarray cancer classification in terms of accuracy.

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

Specificity

The rate of correctly identifying original negatives is the measure of specificity. Equation 15 shows the equation that must be used:

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

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

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 methodis 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

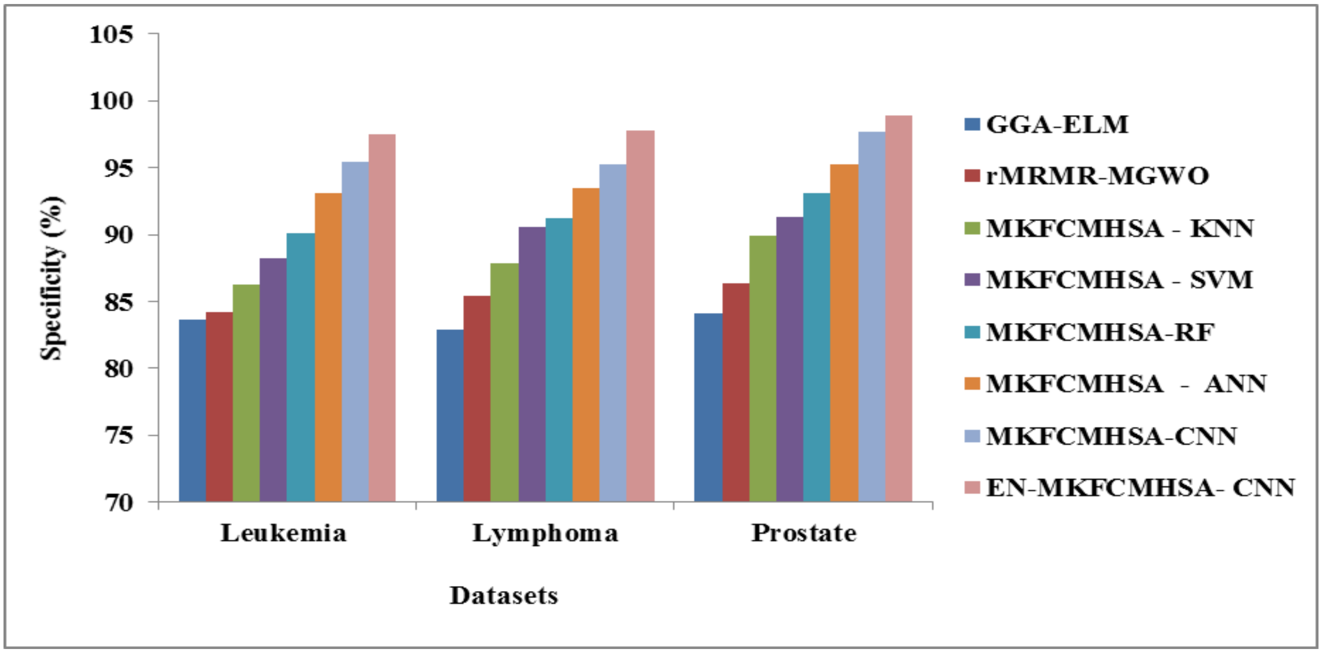


Fig. 6 Comparison of Specificity

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

given dataset. This analysis shows that the EN- MHSAMKFC-CNN can achieve better Specificity than othermethods for microarray cancer classification.

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 formulais as follows in Equation 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 methodis 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 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

F1-Score

The harmonic mean of precision and recall is the F1score. 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-scorefor proposed and existing methods.

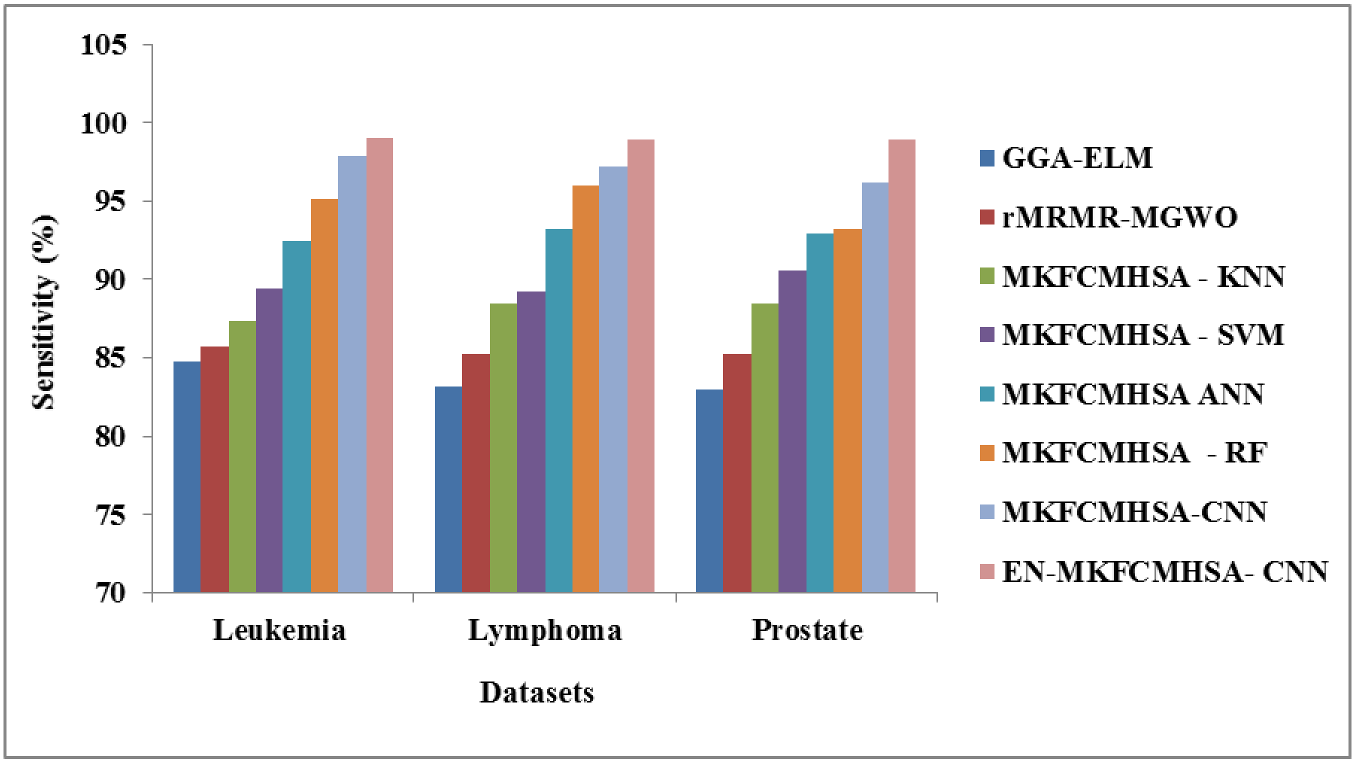


Fig. 7 Comparison of Sensitivity

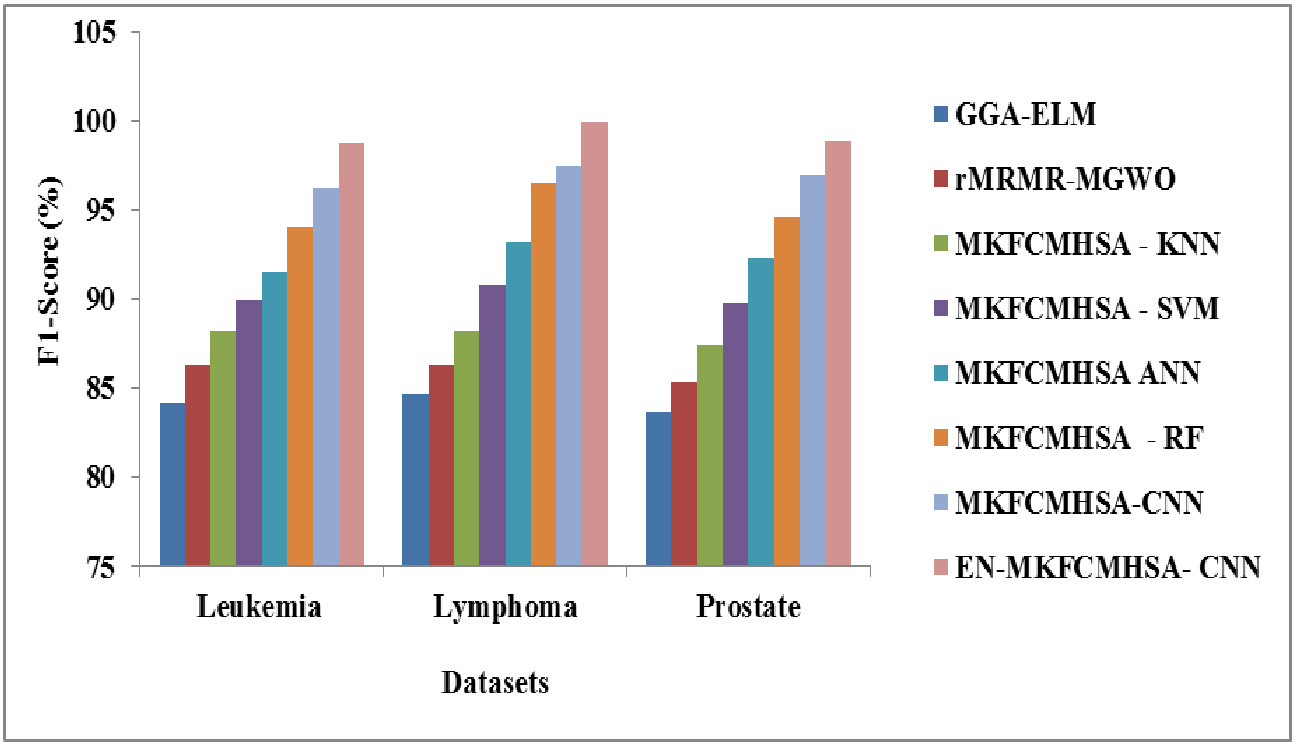


Fig. 8 Comparison of F1-Score

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 methodis 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

A microarray cancer detection system with excellent classification accuracy is proposed as a consequence of this study's findings. To efficiently remove duplicate features from big datasets lacking class labels, MHSAMKFC was first designed. To solve the time-consuming issue of CNN classification and the susceptibility of machine learning to mistakes, the MHSAMKFC-CNN approach was developed. Finally, a stacked ensemble model is developed to deal with the classifier's over-fitting and under-fitting issues by combining many learning models into a single optimum prediction model. As a consequence, the experimental findings demonstrate that the proposed EN- MHSAMKFC-CNN technique achieves superior classification results compared to the state-of-the-art methods currently used 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 SurgicallyResectable 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 MetaheuristicsApproach,” *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 AndModified 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 MicroarrayData,” *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 FeatureSelection 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 BayesianLasso 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 DataClassification,” *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 ColonyOptimization 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-NearestNeighbors 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 forMulticlass 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.