*International Journal of Scientific Research and Modern Education (IJSRME)*
*ISSN (Online): 2455 – 5630*
*(www.rdmodernresearch.com) Volume I, Issue I, 2016*

# DISEASE PREDICTION USING SPATIAL EM ALGORITHM BASED ON SEMI SUPERVISED APPROACH

**S. Kavitha* & K. Ramamoorthy****

* PG Scholar, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

** Assistant Professor, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

**Abstract:**

*Microarray technology is one of the important biotechnological means that allows recording the expression levels of thousands of genes simultaneously within a number of different samples. A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. Among the large amount of genes presented in gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. However, for most gene expression data, the number of training samples is still very small compared to the large number of genes involved in the experiments. When the number of genes is significantly greater than the number of samples, it is possible to find biologically relevant correlations of gene behavior with the sample categories or response variables. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories or response variables. So implement feature subset selection approach to reduce dimensionality, removing irrelevant data and increase diagnosis accuracy and presents learning method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data using Spatial EM algorithm. It can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement KNN (K- nearest neighbor classification) approach to diagnosis the diseases with severity levels. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. The experimental results prove that Spatial EM based classification approach provides improved accuracy rate in disease diagnosis.*

**Key Words:** Microarray, Gene Expression, Spatial EM, Scatter Matrix & Disease diagnosis

## 1. Introduction:

Classification and clustering are two major tasks in gene expression data analysis. Classification is concerned with assigning memberships to samples based on expression patterns, and clustering aims at finding new biological classes and refining existing ones. To cluster and/or recognize patterns in gene expression datasets, dimension problems are encountered. Typically, gene expression datasets consist of a large number of genes (attributes) but a small number of samples (tuples). Many data mining algorithms (e.g., classification association rule mining , pattern discovery, linguistic summaries and context-sensitive fuzzy clustering are developed and/or optimized to be scalable with respect to the number of tuples, so as not to handle a large number of attributes. To apply existing clustering algorithms to genes, various algorithms have been used. Well-known examples are: k-means algorithms, self-

organizing maps (SOM) and various hierarchical clustering algorithms. As for distance measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering genes. The genes regarded as similar by Euclidean distance may be very dissimilar in terms of their shapes or vice versa. It considers each gene as a random variable with n observations and measures the similarity between the two genes by calculating the linear relationship between the distributions of the two corresponding random variables. An empirical study has shown that Pearson's correlation coefficient is not robust to outliers and it may assign high similarity score to a pair of dissimilar genes. Recently, Spatial EM algorithms have been proposed to cluster both genes and samples simultaneously. Spatial EM algorithms aim at identifying subsets of genes and subsets of samples by performing simultaneous clustering of both rows and columns of a gene expression table instead of clustering columns and rows (genes and samples) separately. Specifically, these algorithms group a subset of genes and a subset of samples into a matrix such that the genes and samples exhibit similar behavior. Based on similar behavior, diseases can be classified using KNN classification techniques and evaluate the performance of the system.

**2. Literature Survey:**

Shaheena Bashir, "High breakdown Mixture Discriminant Analysis": The classification rules depend on the unknown parameters, which are to be estimated from the training data. In the presence of a number of outlying observations in the training data, the estimates of the unknown parameters can be unstable due to the undue influence of these atypical observations. High breakdown estimation is a procedure designed to remove this cause of concern, by producing estimators that are robust to serious distortion by outliers, eliminating the influence of such atypical observations. However, it is an important fact that in discriminate analysis, not only are the outliers a concern but also inliers. In the K-means clustering, the outliers for one group might be the inliers for others affecting the classification performance, while in case of mixtures of distributions, this situation may be even worse. The conventional maximum likelihood estimators are affected by the presence of outliers, and so break down. These non-robust estimators influence the discriminate function, leading to the poor classification. The mda approach resulted in the smallest errors of misclassification. It is because the mda approach with maximum likelihood estimators works well within the set of assumptions on which it is based. So, the standard mda approach based on the maximum likelihood method performed better, because the distributional assumption was satisfied in this case.

Yixin Chen, "Depth-Based Novelty Detection and its Application to Taxonomic Research": The job of discovering and describing new species falls on taxonomists. The science of taxonomy has also been suffering from dwindling numbers of experts over the past few decades. Moreover, the pace of taxonomic research, as traditionally practiced, is very slow. In recognizing a species as new to science, taxonomists use a gestalt recognition system that integrates multiple characters of body shape, external body characteristics, and pigmentation patterns. They then make careful counts and measurements on large numbers of specimens from multiple populations across the geographic ranges of both the new and closely related species, and identify a set of external body characters that uniquely diagnoses the new species as distinct from all of its known relatives. The process is laborious and can take years or even decades to complete, depending on the geographic range of the species and believe that the pace of data gathering and analysis in taxonomy can be greatly increased through the integration of machine learning and data mining techniques into taxonomic research

and tackle one of the most important and challenging research objectives in taxonomy new species discovery and develop a novelty detection framework that avoids the above limitation of spatial depth. Specifically, introduce a new depth function, kernelized spatial depth (KSD), which defines the spatial depth in a feature space induced by a positive definite kernel. By choosing a proper kernel, e.g., Gaussian kernel, the contours of a kernelized spatial depth function conform to the structure of the data set. Consequently the kernelized spatial depth can provide a local perspectiveof the data set.

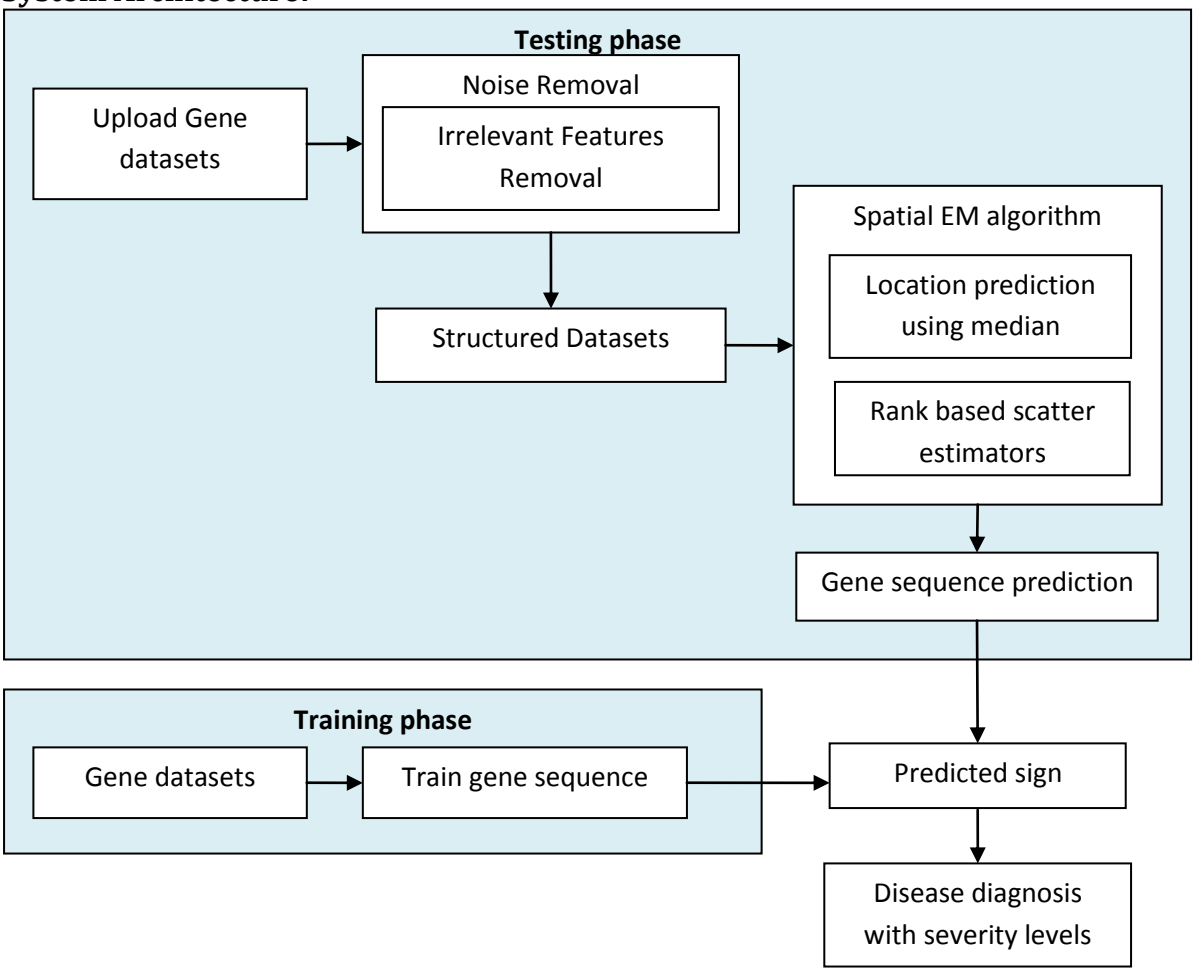
Yixin Chen, "Outlier Detection with the Kernelized Spatial Depth Function": Analyze a novel outlier detection framework based on the notion of statistical depths. Outlier detection methods that are based on statistical depths have been studied in statistics and computational geometry. These methods provide a center-outward ordering of observations. Outliers are expected to appear more likely in outer layers with small depth values than in inner layers with large depth values. Depth-based methods are completely data-driven and avoid strong distributional assumption. Moreover, they provide intuitive visualizationof the data set via depth contours for a low dimensional input space. However, mostof the current depth-based methods do not scale up with the dimensionality of the input space.For example, finding peeling and depth contours, in practice, require the computation of d dimensionalconvex hulls. Because each observation from a data set contributes equally to the valueof depth function, spatial depth takes a global view of the data set. Consequently the outliers' canbe called as "global" outliers. Nevertheless, many data sets from real-world applications exhibit more delicate structures that entail identification of outliers relative to their neighborhood, i.e., "local" outliers and develop an outlier detection framework that avoids the above limitation ofspatial depth.

Yiu-ming Cheung, "Maximum Weighted Likelihood via Rival Penalized EM for Density Mixture Clustering with Automatic Model Selection": Propose to learn the model parameters via maximizing a weighted likelihood, which is developed from the likelihood function of inputs with a designable weight. Under a specific weight design, then give out a maximum weighted likelihood (MWL) approach named Rival Penalized Expectation-Maximization (RPEM) algorithm, which makes the components in a density mixture compete with each other, and the rivals intrinsically penalized with a dynamic control during the learning. Not only are the associated parameters of the winner updated to adapt to an input, but also all rivals' parameters are penalized with the strength proportional to the corresponding posterior density probabilities. Compared to the EM, such a rival penalization mechanism enables the RPEM to fade out the redundant densities in the density mixture. In other words, the RPEM has the capability of automatically selecting an appropriate number of densities in density mixture clustering. The numerical simulations have demonstrated its outstanding performance on Gaussian mixtures and the color image segmentation problem. Moreover, show that a simplified version of RPEM actually generalizes the RPCCL algorithm so that it is applicable to ellipse-shaped clusters as well with any input proportion. Compared to the existing RPCL and its variants, this generalized RPCCL (G-RPCCL), as well as the RPCCL, circumvents the difficult pre-selection of the de-learning rate. Additionally, a special setting of G-RPCCL further degenerates to the RPCL and its Type A version, but meanwhile giving out a guidance to choose an appropriate delearning rate. Subsequently, propose a stochastic version of RPCL and its Type A variant, respectively, in which the difficult selection problem of the relearning rate is novelty circumvented. The experiments have shown the promising results of this stochastic implementation.

Kai Yu, "Robustness of the Affine Equivariant Scatter Estimator Based on the Spatial Rank Covariance Matrix": Use different approach to obtain equivariance property of spatial sign and rank covariance matrices under elliptical models without sacrifice of robustness. The basic idea is to take advantage of the fact that the spatial sign and rank functions preserve directional information but lose some measure on distance. Consequently, eigenvectors of the spatial sign and rank covariance matrices are able to capture principle components (orientation) of a data cloud (or underlying distribution), but Eigen values no longer reflect variation on those directions even for the rank covariance matrix in which some distance information is present in spatial rank function. The strategy is to replace each Eigen value with aunivariate scale estimator on the corresponding direction such that it depicts the proper variability. For consideration of robustness, the univariate scale functional must be robust, e.g. MAD (median of absolute deviation). And favor spatial rank covariancematrix over spatial sign covariance matrix because it is more efficient and there is no initial location estimator needed for computing rank vectors. Then call the resulting covariance matrix the modified spatial rank covariance matrix (MRCM).The main contributions of this paper are that study the robustness properties ofMRCM by the breakdown point and influence function. The finite sample breakdown points is obtained and show that the finite sample breakdown point can attain the upper boundby a proper choice of univariate scale estimator. The influence functions of eigenvalues andeigenvectors of the covariance matrix are derived and found to be bounded

**3. Implementation:**

**System Architecture:**



**3.1 Disease Prediction Using Gene Clusters:**

**3.1.1 Datasets Acquisition:** In this module, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation.Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data or coarse data. After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records).

**3.1.2 Median Estimation:** To tackle the effect of outliers in cluster analysis to consider the Spatial EM clustering which replaces the squared Euclidean distances in the objective function of the k-means clustering with the absolute Euclidean distances. In spatial EM, can analyze coverage of the data before clustering begins. And propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it's the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested. The amalgamation of the single member cluster should be executed with the detachment of an object which is far from its cluster centroid when it is found to be beneficial. When no further amalgamations give an improvement, the transfer phase is reentered and continued until no more transfers or amalgamations can improve the clustering criterion value. In this module, can calculate the mean values for each gene features. These gene features listed as it is.

**3.1.3 Rank Based Scatter:** In this module, can create scatter matrix based on median values that are derived by clustering algorithm. Then construct scatter matrix and

reflecting as the within-cluster scatter, the between-cluster scatter and their summation the total scatter matrix. The determinant of a scatter matrix roughly measures the square of the scattering volume. And minimizing this measure is equivalent to both minimizing the intra-cluster scatter and maximizing the inter-cluster scatter. Based on scatter matrix, classification is performed in following modules.

**3.1.4 Disease Prediction:** Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings. KNN approach matches each neighborhood genes to predict the diseases. In this module, implement classifier design in semi supervised format. K nearest neighbor classifier allowed to access and provides predicted sign for corresponding diseases such as diabetic, leukemia and so on. Multi class framework can be implemented to predict the disease with severity levels. And diseases are predicted with various labeling.

**3.1.5 Evaluation Criteria:** In this module, the performance of the proposed semi-supervised algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separability index and classification accuracy of K-nearest neighbor rule. The proposed system provide improved accuracy rate in gene classification.

**3.2 System Implementation:**

**3.2.1 Spatiyal-EM Alogorithm:**

Spatial-EM modifies the component estimates on each M-step by spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability. Pseudo code of the algorithm is described as:

Initialization t = 0, $\mu_j, \sum_j = I, \tau_j = \frac{1}{K} for \forall_j$

Do until $\tau_j^t coverage for all j$

For j=1 to K

E-Step: Calculate $T_{ji}^t$

M-Step: Update $\tau_j^{t+1}$

Define$w_{ji}^t$, Find$\mu_j^{t+1}$, Find$(\sum_j^{t+1})^{-1}$ $and$ $(\sum_j^{t+1})^{-1/2}$

End

t=t+1

End

In spatial algorithm can first calculate the maximum coverage of data and then initialize all variables and perform Expectation and Maximization steps as in EM algorithm.The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The EM algorithm proceeds from the observation that the following is a way to solve these two sets of equations numerically.

### 3.2.2 KNN Algorithm:

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm has nothing to do with and is not to be confused with k-means, another popular machine learning technique. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. Often, the classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Neighbor or Neighborhood components analysis.

**Mathematical Model:**

KNN algorithm as derived as follows:

BEGIN
Input: D={(X1,C1),….,(Xn,Cn)}
X=(X1,…,Xn) new instance to be classified
For each labeled instance(Xi,Ci) calculate d(Xi,X)
Order d(Xi,X) from lowest to highest, (i-1,…,N)
Select the K nearest instances to X:$D_X^K$
Assign to X the most frequent class in $D_X^K$
End

## 4. Proposed System:

A gene-based clustering is used to group the gene patterns. Patterns are clustered based on genetic code transcriptions. The proposed methodology includes Spatial EM that can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement KNN (K- nearest neighbor classification) approach to diagnosis the diseases. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for recognizing biologically significant gene clusters with excellent predictive capability. Spatial-EM alters the component estimation on each M-stepby spatial median and rank covariance matrix to gainrobustness at the cost of increasing computational encumberand losing theoretical tractability. In spatial algorithm can first calculate the maximum coverage of data and then initialize all variables and perform Expectation and Maximization steps as in EM algorithm. The EM iteration swaps between to perform an expectation (E) step, which generates a function for the expect of the log-likelihood evaluated using the current estimation for the parameters, and maximization (M) step, which figures parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to decide the distribution of the latent variables in

the next E step. The EM algorithm proceeds from the observation that the following is a way to explain these two sets of equations numerically.

**Gene Classification:**

Microarray classification approaches based on machine learning algorithms applied to DNA microarray data have been shown to have statistical and medical relevance for a variety of diseases. One particular machine learning algorithm, KNN, has exposed promise in a variety of biological classification tasks, including gene expression microarrays. KNNs are powerful classification systems based on regularization techniques with excellent performance in many practical classification problems. The Support Vector Machine is rooted in statistical learning theory. It is different from the other classification method in the sense that KNN tries to maximize the separation between samples of two classes. Normally, only a subset of the data samples determines the decision hyper plane. Suppose the n data samples belong to two classes $\{(x_1, y_1), \ldots, (x_n, y_n)\}, x_i \epsilon \Re^m \, and \, y_t$ =1 or -1. A smaller value of the first term corresponds to better generalization, while the fewer positive values of the slack variables in the second term correspond to fewer misclassifications on the training samples. When the later is equal to zero, the training samples are linearly separable and there is no misclassification.

**Advantages:**

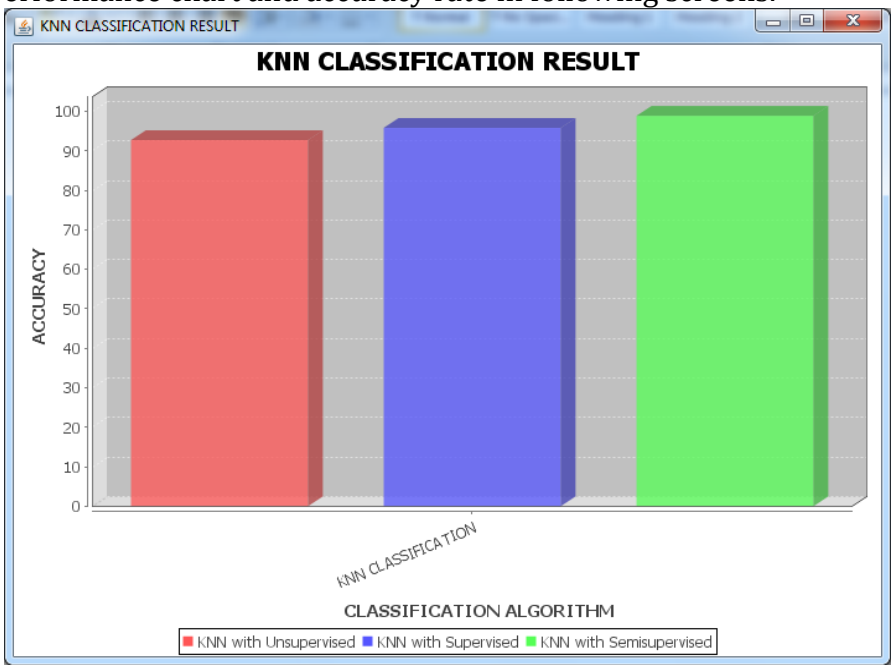
- ✓ Outliers are predicted efficiently in gene expression data.
- ✓ Automatic clustering process is done.
- ✓ Efficiently diagnosis the diseases using classification performance.
- ✓ Partial and full data can be handled properly.
- ✓ Noises eliminated and predict the diseases accurately
- ✓ Severity levels are calculated.

**5. Experimental Results:**

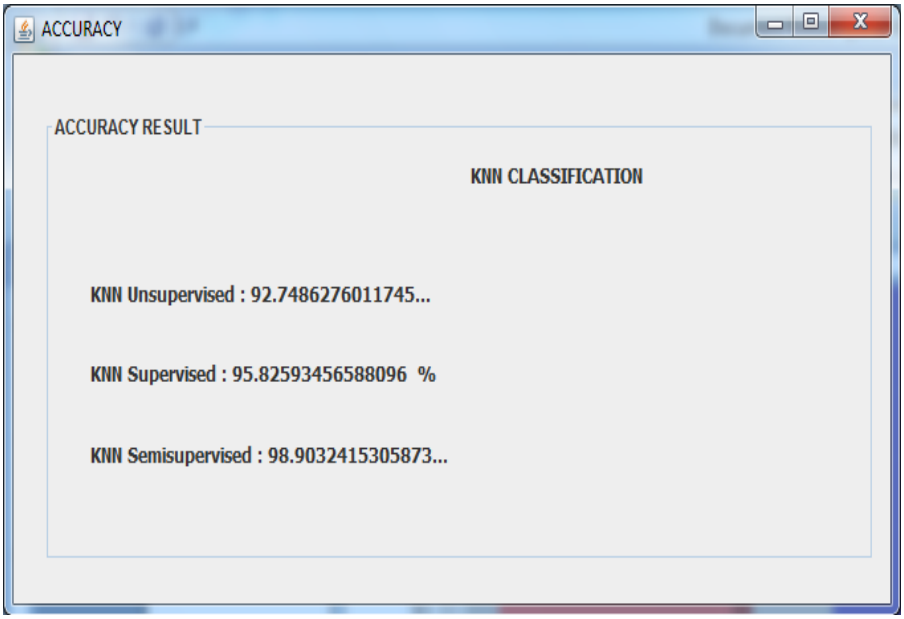
Experimental results can evaluate the performance of the system using Accuracy rate. The accuracy rate is calculated using true positive, false positive, true negative and false negative metrics. So the accuracy rate is defined as:

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

And show performance chart and accuracy rate in following screens.



**5.1 KNN Classification Result:**




**5.2 Accuracy Result:**

KNN based unsupervised, KNN based supervised and KNN based semi supervised approaches are evaluated and accuracy rate is listed above. Based on evaluation, KNN based semi supervised approach provide 97-98% accuracy in gene clustering and disease prediction.

**6. Conclusion and Future Work:**

**6.1 Conclusion:**

Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Gene expression data generated by microarray experiments offer tremendous potential for advances in molecular biology and functional genomics. This paper reviewed both classical and recently developed clustering algorithms, which have been applied to gene expression data, with promising results. The proposed semi-supervised spatial EM clustering algorithm is based on measuring mean values and scatter matrix using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised EM gene selection algorithm with accuracy rate. An important finding is that the proposed semi-supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. And predict the diseases with severity levels in improved accuracy rate.

**6.2 Future Work:**

We can extend the work to implement various classification algorithms to improve the accuracy rate at the time of disease prediction.

**7. References:**

1. S. Bashir and E. M. Carter, “High breakdown mixture discriminant analysis,” J. Multivariate Anal., vol. 93, no. 1, pp. 102–111, 2005.
2. C. Biernacki, G. Celeux, and G. Govaert, “An improvement of the NEC criterion for assessing the number of clusters in a mixture model,” Pattern Recognit.Lett., vol. 20, pp. 267–272, 1999.

3. B. Brown, "Statistical uses of the spatial median," J. Roy. Stat. Soc., B, vol. 45, pp. 25–30, 1983.
4. M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression fidata by using support vector machines," Proc. Nat. Acad. Sci., vol. 97, no. 1, pp. 262–267, 2000.
5. N. A. Campbell, "Mixture models and atypical values," Math. Geol., vol. 16, pp. 465–477, 1984.
6. G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," Classification J., vol. 13, pp. 195–212, 1996.
7. Y. Chen, Bart H. Jr, X. Dang, and H. Peng, "Depth-based novelty detection and its application to taxonomic research," in Proc. 7th IEEE Int. Conf. Data Mining, Omaha, Nebraska, 2007, pp. 113–122.
8. Y. Chen, X. Dang, H. Peng, and H. Bart Jr., "Outlier detection with the kernelized spatial depth function," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 288–305, Feb. 2009.
9. Y. Chueng, "Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 750–761, Jun. 2005.
10. X. Dang and R. Serfling, "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties," J. Stat. Inference Planning, vol. 140, pp. 198–213, 2010.
11. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., B, vol. 39, pp. 1–38, 1977.