INNSPUB

Journal of Biodiversity and Environmental Sciences (JBES)
ISSN: 2220-6663 (Print) 2222-3045 (Online)
Vol. 21, No. 2, p. 92-99, 2022
http://www.innspub.net

RESEARCH PAPER OPEN ACCESS

Combining statistical and Fuzzy-rough classifiers for cancer Subtype prediction

Sukanta Majumder[1], Ansuman Kumar[*2], Anindya Halder[2]

[1]Department of Computer Science & Engineering, University of Kalyani, Kalyani, Nadia, West Bengal, India

[2]Department of Computer Application, North-Eastern Hill University, Tura Campus, Meghalaya, India

Article published on August 10, 2022

Key words: Cancer prediction, Statistical classifier, Fuzzy-rough classifier, Gene expression data.

Abstract

Cancer prediction from gene expression data is one of the challenging areas of research in the field of bioinformatics and machine learning. In gene expression data, labeled samples are very limited compared to unlabeled samples; and labeling of unlabeled data is expensive. Therefore, single classifier trained with limited training samples often fails to produce desired result. In this situation, combination of classifiers can be effective as its ensembles the results of individual classifiers which can improve the cancer prediction accuracy. In this article a novel method, combining statistical and fuzzy-rough classifiers (CSFRC) for cancer prediction is proposed which uses support vector machine, naive bayes as statistical classifiers and fuzzy-rough nearest neighbor classifier. The proposed method is able to deal the uncertainty, overlapping and indiscernibility usually present in cancer subtype classes of the gene expression data. The proposed method is validated on eight publicly available gene expression datasets. Experimental results suggest that the performance of the proposed method provides better results in comparison to other compared classifiers for cancer subtype prediction from gene expression data. The proposed method turns out to be very effective in cancer prediction from gene expression data particularly when the individual classifier result is not up to the mark with limited training samples.

***Corresponding Author:** Ansuman Kumar ✉ ansuman.kumar@gmail.com

Introduction

Cancer is one of the dangerous health problems across the world. There were approximately 18.1 million new cancer cases and 9.9 million cancer-related deaths worldwide reported in the year 2020 according to the Global cancer statistics. Therefore, the early detection and diagnosis of cancer are vital as it usually increases the chances of successful treatment. Conventional clinical methods for cancer sample classification rely on the clinical findings and the morphological appearance of the tumor. These methods are costly and time consuming. The recent development of microarray technology (Stekel, 2003) has allowed biologists to specify thousands of genes in a single experiment in order to produce comparatively low-cost diagnosis and prediction of cancer at early stage.

Several computational methods have been applied for analysis of microarray gene expression data using supervised (i.e., classification) (Dettling *et al.*, 2003), unsupervised (i.e., clustering) (Jiang *et al.,* 2004), semi-supervised clustering (Priscilla *et al.,* 2013), and semi-supervised classification (Halder *et al.,* 2014). Usually, the number of samples present in microarray gene expression data is very less compared to the number of genes (Du *et al.,* 2014); and the class subtypes present in dataset are often vague and overlapping in nature. Therefore, the single traditional classifiers often fail to achieve desired accuracy. In this situation, the combination of the classifiers (Kuncheva, 2004) is supposed to be valuable as it judiciously combines the predictions of the individual classifier to make the final decision which are expected to be better than any individual classifier.

Ensemble technique (i.e., combination of classifiers) is the learning model that achieves better results by combining the judgments of multiple base classifiers (Kuncheva, 2004). It uses many base classifiers, and combines their judgments in such a way that the combination result will improve the performance in comparison to any individual classifier (Kuncheva, 2004). The heterogeneity among the base classifiers and diversity in the training data set are the basic keys to success of ensemble technique.

Various popular ensemble techniques are proposed in the literature, viz., Bagging, Boosting, AdaBoost, and Random Forest (Polikar, 2006). Ensemble techniques have the ability to handle small sample size and high dimensionality. Therefore, ensemble technique has been widely applied to microarray gene expression data. An outstanding review of ensemble techniques applied in bioinformatics may be found in (Yang *et al.,* 2010). Several pioneered work to classify cancer from the microarray gene expression data are proposed. Dettling and Buhlmann (Dettling *et al.,* 2003) proposed boosting for tumor classification with gene expression data. Osareh and Shadgar (Osareh *et al.,* 2013) provided an efficient ensemble learning method using RotBoost ensemble methodology. Valentini *et al.* (Valentini *et al.,* 2004) introduced bagged ensembles of support vector machines for cancer recognition.

However, those ensemble techniques are not able to deal with the uncertainty, ambiguity, over lapping ness and vagueness often present in the gene expression data. Therefore, in this article method combining statistical and fuzzy-rough classifiers (CSFRC) is proposed which utilizes the advantages of the statistical learning (using support vector machine and naive bayes) and rough fuzzy system (using fuzzy-rough nearest neighbor for uncertainty, ambiguity, vagueness and indiscernibility handling) in order to predict cancer subtypes from gene expression data (to improve the prediction accuracy of any individual classifier).

The remaining of the article is structured as follows. The preliminary study related to this article is briefly illustrated in Section 2. Section 3 presents a detailed description of the proposed CSFRC method. In Section 4, details of the experiments and analysis of the results are provided. Finally, conclusions are drawn in Section 5.

Materials and methods

Preliminary study

The proposed CSFRC method uses the concept of fuzzy set, rough set, Support vector machine (SVM) and Naive Bayes (NB). Thus, brief outline of those is provided below.

Fuzzy set theory

Fuzzy set theory developed by Zadeh (Zadeh, 1965) in the year 1965. Fuzzy set theory is an extension of crisp sets to handle vague and imprecise data. Fuzzy set *A* uses mapping from the universe X to the interval [0, 1]. The value *A(x)* for $x \in X$ is called the membership degree of *x* in *A*.

Rough set theory

Pawlak introduced rough set theory in early 1980s (Pawlak, 1982). Rough set theory can handle uncertainty, indiscernibility and incompleteness in the datasets. It begins with the idea of an approximation space, which is an ordered pair < *X, R* >, where *X* is the non-empty universe of discourse and *R* is an equivalence relation defined on *X*. *R* satisfies the reflexive, symmetric and transitive property. For each subset *A* of *X*, the lower approximation is defined as the union of all the equivalence classes which are fully included inside the class *A*, and the upper approximation is defined as the union of equivalence classes which have non-empty intersection with the class *A*.

Fuzzy-rough set theory

Fuzzy set theory can handle vague information, while rough set theory can handle incomplete information. These two theories are complementary to each other. Hybridization of these two concepts yields the idea of the fuzzy-rough set which is the pair of lower and upper approximations of a fuzzy set *A* in a universe *X* on which a fuzzy relation *R* is defined. The fuzzy-rough lower and upper approximations of *A* are defined respectively as follows (Radzikowska *et al.,* 2002):

$$(R \downarrow A)(x) = \inf_{y \in X} I(R(x,y), A(y)) \qquad (1)$$

$$(R \uparrow A)(x) = \sup_{y \in X} T(R(x,y), A(y)) \qquad (2)$$

Where, I is the Lukasiewicz implicator, T is the Lukasiewicz *t*-norms and $R(x,y)$ is the valued similarity of patterns x and y, $\inf$ is the *infimum* and $\sup$ represents the *supremum.*

Support Vector Machine

Support vector machine (SVM) (Vanitha *et al.,* 2015) is a supervised machine learning technique that can be used for classification as well as regression problems under statistical techniques. It handles non-linear decision boundaries of arbitrary complexity (Vanitha *et al.,* 2015). The decision boundary (a straight line in the case of a two-dimensional separation) is positioned to leave the largest possible margin on either side. Classification is done by the finding the hyper-plane that differentiates the two classes very well.

Naïve Bayes

Naïve Bayes algorithm (Chandra *et al.,* 2011) is also supervised learning algorithm. It is based on Bayes theorem and used for solving classification problems. Naïve Bayes classifier is one of the simple and most effective classification algorithms which helps in making the machine learning models that can make fast predictions (Chandra *et al.,* 2011).

Proposed method Combining Statistical and Fuzzy-Rough Classifiers

The proposed CSFRC method is combination of three diverse set of base classifiers, namely, Fuzzy-rough nearest neighbour (FRNN), Support vector machine (SVM) and Naive Bayes (NB). All three base classifiers are trained with the labeled training samples. Then after the test samples are classified by all the base classifiers to a certain class using the labeled training set. The ensemble decisions of the test samples are aggregated using majority voting technique applied on the predictions of different base classifiers. The block diagram of the proposed method is shown in Fig. 1 and details of the proposed method are described as follows.

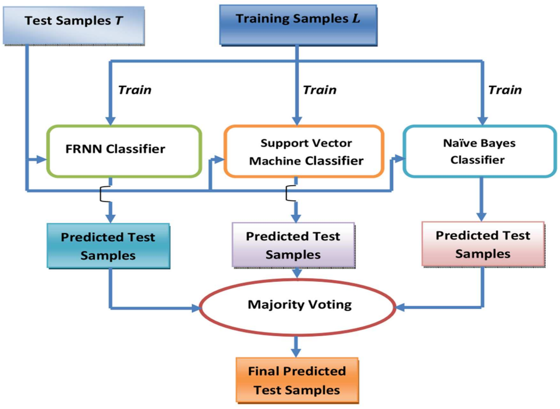

Fig. 1. Block diagram of the proposed method combining statistical and fuzzy-rough classifiers.

Fuzzy-Rough Nearest Neighbour Classifier

The fuzzy-rough nearest neighbour (FRNN) (Jensen *et al.,* 2008) classifier is used as one of the base classifiers in the proposed method. The detailed description of FRNN technique is given below:

Let, $L=\{< l_j \ , d_j > |\ j = 1\ to\ |L|\ and\ d_j \in C\}$

be the training set for individual FRNN classifier.

The details of the base classifier FRNN are described below.

1. Compute the k-nearest neighbour (kNN) labeled patterns closest to each of the test pattern (t) based on the Euclidean distance (compute the distance from the labeled pattern to test pattern).
2. The values of lower and upper approximations of test pattern (t) for belonging to each class C is calculated respectively as follows:

$$(R\downarrow C)(t) = \inf_{y\in kNN} I(R(t,y),C(y)) \quad (3)$$

$$(R\uparrow C)(t) = \sup_{y\in kNN} T(R(t,y),C(y)) \quad (4)$$

Where, I is the Lukasiewicz implicator, T is the Lukasiewicz t-norms and $R(t,y)$ is the valued similarity of test patterns X and labeled sample $y \in kNN$ is computed as:

$$R(t,y) = \frac{\sum_{y\in kNN}(\|t-y\|)^{\frac{2}{m-1}}}{(\|t-y\|)^{\frac{2}{m-1}}}; \quad (5)$$

where, $\|t-y\|$ is the distance of the test sample (t) from the labeled sample $y \in kNN$ (k-nearest neighbour labeled sample of test sample (t) and m ($1 < m < \infty$) is the fuzzifier. $C(y)$ is computed as:

$$C(y) = \begin{cases} 1, & if \quad y \in C; \\ 0, & Otherwise. \end{cases} \quad (6)$$

3. The test sample (t) is assigned to a particular class for which the average value of lower and upper approximations is highest. The assigned $ClassLabel$ (t) of test sample (t) is determined as follows:

$$ClassLabel(t) = \operatorname*{argmax}_j \left(\frac{(R\downarrow C_j)(t) + (R\uparrow C_j)(t)}{2} \right); \quad \forall t \quad (7)$$

Support Vector Machine Classifier

SVM classifier is also used as one of the base classifiers in the proposed CSFRC method. Thus, detailed description of SVM classifier is given below.

The following sigmoid function (Vanitha *et al.,* 2015) with two parameters A and B is used to calculate the confidence score to classify the test sample (t).

$$SVM(t, C_i) = \frac{1}{1+e^{A.f(t)+B}}\ ;\ \forall t \qquad (8)$$

The modality of sigmoid function is controlled by parameter A and B, and $f(t)$ is the standard output value of test sample (t) in SVM (Vanitha *et al.,* 2015). Thus, the class of test samples can be determined by the Equation 8.

Naïve Bayes Classifier

Naïve bayes classifier is also involved in the proposed CSFRC method as a base classifier. Thus, detailed description of NB classifier is given below.

It is a simple probabilistic based method, which can predict the class membership probabilities (Chandra *et al.,* 2011). The classifier will predict that the test sample (t) belongs to the class with the highest posterior probability, conditioned on t. That is, the NB classifier predicts that the sample t belongs to the class C_i, if and only if $P(C_i|t) > P(C_j|t)$ for $1 \leq j \leq$ m, j ≠ i. The class C_i for which $P(C_i|t)$ is maximized is called the Maximum Posteriori Hypothesis (Chandra *et al.,* 2011). Bayes theorem is given in Equation 9.

$$P(C_i|t) = \frac{P(t|C_i)P(C_i)}{P(t)}; \forall t \ (9)$$

Majority voting for final class assignment

The majority voting method (Marak *et al.,* 2021) is applied to assign the final class for each test sample, once all three base classifiers make a prediction (for each test sample). Here, each classifier cast a vote (or predict) in the form of class label for each test sample and the final decision / prediction is made for a test sample which gets the maximum vote (as class label).

Results and discussion

In this section, we provide the details of microarray gene expression cancer datasets used for the experiments followed by the compared method and performance evaluation measures. Experimental results and analysis of the results are summarized at the end.

Description of Datasets

We have used eight real life microarray gene expression cancer datasets namely, Colon Cancer, Brain tumor, SRBCT, Lymphoma, Prostate Cancer, Ovarian Cancer, Leukemia, Lung Cancer datasets in this article. These datasets are publicly available at www.stat.ethz.ch/dettling/bagboost.html (Dettling, 2004) and technology agency for science and research, kent ridge bio-medical dataset repository (http://datam.i2r.astar.edu.sg/datasets/krbd/ index. html). The dataset is a collection of the samples and each sample is described by gene expression values and their class label information. Detail descriptions of the used datasets are given below.

Colon Cancer dataset contains 40 samples of cancerous patients and 22 samples of normal patients. Each sample comprises of 2000 gene expression values.

Brain Tumor dataset contains 42 samples distributed in 5 classes of brain tumor viz., medulloblastomas, malignant gliomas, atypical teratoid/rhabdoid tumors, primitive neuroectodermal tumors, human cerabella. Numbers of samples for these classes are 10, 10, 10, 8 and 4 respectively. There are 5597 genes in each sample.

Small round blue cell tumors (SRBCT) dataset consists of 63 samples. Among them, 12 samples are of neuroblastoma (NB), 20 samples are of rhabdomyosarcoma (RS), 8 samples are of Burkitt's lymphoma (BL) and 23 samples are of Ewing's sarcoma (ES). Each sample comprises of 2308 genes expression values.

Lymphoma dataset contains 62 samples and each sample is having 4026 genes. There are 3 classes of lymphoma viz., diffuse large B-cell lymphoma, follicular lymphoma and chronic lymphocytic leukemia.

Prostate cancer dataset contains 102 samples in which 52 observations are from prostate cancer tissues and 50 are from normal patients. The expression profile contains 6033 genes.

Ovarian cancer dataset consists of 203 samples in which 91 samples are normal and 162 samples are cancerous. There are 15154 genes in each sample.

Leukemia dataset is having 72 samples distributed in two classes namely, lymphoblastic leukemia and myeloid leukemia. Each sample is described by 3571 genes.

Lung Cancer dataset consists 203 samples in which 139 samples of lung adenocarcinomas, 20 samples of pulmonary carcinoids, 21 samples of squamous cell lung carcinomas, 6 samples of small-cell lung carcinomas and 17 normal lung samples. Each sample contains expression values of 12600 genes. The summary of the datasets used for the experiments is provided in Table 1.

Table 1. Summary of eight microarray gene expression datasets used for the experiments.

Datasets	Samples	Genes	Classes
Colon Cancer	62	2000	2
Brain Tumor	42	5597	5
SRBCT	63	2308	4
Lymphoma	62	4026	3
Prostate cancer	102	6033	2
Ovarian cancer	253	15154	2
Leukemia	72	3571	2
Lung Cancer	203	12600	5

Comparison with others methods

The performance of the proposed EnFRNN method is compared with three methods namely, Fuzzy *k*-Nearest Neighbour (FKNN) (Keller *et al.,* 1985), Fuzzy-Rough Nearest Neighbour (FRNN) (Jensen *et al.,* 2008) and Ensemble based Fuzzy-Rough Nearest Neighbour (EnFRNN) (Kumar *et al.,* 2020).

Fuzzy k- Nearest Neighbour Classifier

Fuzzy *k*-Nearest Neighbour (FKNN) (Keller *et al.,* 1985) is an extension of the *k*-Nearest Neighbour (KNN) classifier.

In *KNN* algorithm, equal weightage is given to all the *k*-nearest neighbours to calculate the predicted class of a test data. FKNN algorithm assigns fuzzy membership of a test pattern in each class. That class is taken to be the predicted class (of that test pattern) for which the fuzzy-membership is maximum.

Fuzzy-Rough Nearest Neighbour Classifier

Fuzzy-Rough Nearest Neighbour (FRNN) classifier (Jensen *et al.,* 2008) is the combination of fuzzy and rough sets theories. It uses the concept of upper and lower approximations to assign the class label information to the test pattern. The values of lower and upper approximations of a decision class are computed based on the *k*-nearest neighbours of a test pattern.

Ensemble based Fuzzy-Rough Nearest Neighbour Classifier

Ensemble based Fuzzy-Rough Nearest Neighbour (EnFRNN) method (Kumar *et al.,* 2020) combines the predictions of all three FRNN base classifier with the help of majority voting that improves the classification accuracy. Here fuzzy set deals the vagueness, ambiguity and rough set deals the uncertainty, incompleteness and indiscernibility present in the gene expression data.

Table 2. Summary of the average experimental results (in terms of accuracy, precision, recall, macro F_1, micro F_1 and kappa) of 10 simulations achieved by different methods viz., FKNN, FRNN, EnFRNN and the proposed method CSFRC performed on eight microarray gene expression datasets.

Datasets	Methods	Accuracy (%)	Overall Precision	Overall Recall	Macro F_1	Micro F_1	Kappa
Colon Cancer	FKNN	80.69±8.28	0.8467	0.8237	0.8029	0.8350	0.6255
	FRNN	90.86±4.74	0.9078	0.9128	0.9006	0.9098	0.8040
	EnFRNN	96.85±2.48	0.9667	0.9661	0.9642	0.9662	0.9288
	CSFRC	97.17±1.41	0.9668	0.9769	0.9668	0.9688	0.9300
Brain Tumor	FKNN	67.81±7.66	0.6692	0.7901	0.6433	0.7224	0.5812
	FRNN	82.77±8.24	0.8227	0.8648	0.7914	0.8423	0.7772
	EnFRNN	87.04±3.59	0.8691	0.8652	0.8323	0.8667	0.8296
	CSFRC	89.21 ±4.97	0.8673	0.8902	0.8327	0.8780	0.8479
SRBCT	FKNN	71.45±4.37	0.7918	0.7727	0.7140	0.7818	0.6239
	FRNN	83.09±5.56	0.8586	0.8197	0.8129	0.8386	0.7678
	EnFRNN	89.15±5.16	0.9155	0.8552	0.8648	0.8841	0.8479
	CSFRC	90.91±5.96	0.9274	0.8864	0.8870	0.9064	0.8743
Lymphoma	FKNN	96.25±1.01	0.9786	0.9218	0.9474	0.9493	0.9202
	FRNN	97.33±1.26	0.9875	0.9431	0.9630	0.9647	0.9431
	EnFRNN	97.40±0.96	0.9886	0.9323	0.9566	0.9596	0.9368
	CSFRC	96.43±1.56	0.9833	0.9259	0.9498	0.9538	0.9239
Prostate Cancer	FKNN	67.55±10.89	0.6736	0.7444	0.6425	0.7047	0.3471
	FRNN	86.12±7.96	0.8613	0.8738	0.8594	0.8675	0.7224
	EnFRNN	90.64±3.84	0.9069	0.9150	0.9058	0.9110	0.8130
	CSFRC	91.73±4.47	0.9200	0.9173	0.9170	0.9187	0.8345
Ovarian Cancer	FKNN	87.07±7.50	0.8563	0.8704	0.8525	0.8626	0.7100
	FRNN	90.76±7.04	0.9149	0.9145	0.9027	0.9144	0.8101
	EnFRNN	95.26±2.52	0.9555	0.9486	0.9489	0.9519	0.8983
	CSFRC	95.98±1.79	0.9688	0.9495	0.9573	0.9590	0.9147
Leukemia	FKNN	75.59±5.77	0.7879	0.7668	0.7482	0.7772	0.5162
	FRNN	81.76±11.95	0.8408	0.8356	0.8106	0.8381	0.6425
	EnFRNN	88.28±7.04	0.8933	0.8739	0.8741	0.8835	0.7533
	CSFRC	88.38±5.56	0.8818	0.8735	0.8727	0.8775	0.7477
Lung Cancer	FKNN	61.81±8.43	0.7895	0.6061	0.6070	0.6852	0.4414
	FRNN	68.94±7.46	0.8300	0.6364	0.6613	0.7195	0.5147
	EnFRNN	73.66±6.02	0.8646	0.6240	0.6572	0.7244	0.5556
	CSFRC	72.68±7.32	0.8773	0.6749	0.7125	0.7629	0.5571

Performance Evaluation Measures

In this article, we have used six different kinds of validity measures namely, (i) percentage accuracy, (ii) precision, (iii) recall, (iv) macro averaged F_1 measure, (v) micro averaged F_1 measure (Kumar *et al.,* 2019) and (vi) kappa (Cohen, 1960) to assess the performance of the methods.

Experimental Results and Analysis

The average experimental results of 10 simulation runs (on random selection of labelled/training patterns) in terms of percentage accuracy, precision, recall, macro F_1, micro F_1 and kappa obtained by all the methods (viz., FKNN, FRNN, EnFRNN and the proposed CRFRC) performed on eight microarray gene expression datasets are reported in Table 2.

Best results are shown in bold font in the Table 2. The standard deviations of accuracies of 10 simulations are also shown using ± sign corresponding to each percentage accuracy in Table 2. It is seen from the Table 2, that the proposed CSFRC method performed better in terms all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) over other methods namely, FKNN, FRNN and EnFRNN for six datasets. EnFRNN method performed better compared to the proposed method only for lymphoma dataset.

Conclusions

This article presents combining statistical and fuzzy-rough classifiers for cancer subtype prediction from microarray gene expression datasets. Cancer subtype classes are usually overlapping and indiscernible in nature which can be handled by the fuzzy-rough set theory. Therefore, in this article the proposed method CRFRC utilizes the advantages of the statistical learning (using support vector machine and naive bayes) and rough fuzzy system (using fuzzy-rough nearest neighbour for uncertainty, ambiguity, vagueness and indiscernibility handling) to predict cancer subtypes classes from gene expression data to further improve the prediction accuracy of any individual classifier.

The effectiveness of the proposed method is tested using eight real life microarray gene expression cancer datasets in terms of different validity measures viz., accuracy, precision, recall, F_1-measures and kappa. It is observed from the experimental results that the proposed CRFRC method performed better in terms all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) for almost seven datasets out of eight datasets. In future, robustness of the proposed CRFRC method may further be tested on other kind of gene expression datasets such as microRNA.

References

Chandra B, Gupta M. 2011. Robust approach for estimating probabilities in naïve Bayesian classifier for gene expression data. Expert Systems with Applications **38(3),** 1293-1298.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20(1),** 37-46.

Dettling M, Buhlmann P. 2003. Boosting for tumor classification with gene expression data. Bioinformatics **19(9),** 1061-1069.

Dettling M. 2004. Bagboosting for tumor classification with gene expression data. Bioinformatics **20(18),** 583-593.

Du D, Li K, Li X, Fei M. 2014. A novel forward gene selection algorithm for microarray data. Neurocomputing **133,** 446-458.

Halder A, Misra S. 2014. Semi-supervised fuzzy k-NN for cancer classification from microarray gene expression data. In Proceedings of the 1st International Conference on Automation, Control, Energy and Systems (ACES 2014) (IEEE Computer Society Press) 1-5.

Jensen R, Cornelis C. 2008. A new approach to fuzzy-rough nearest neighbour classification. In: Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing **310-319,** 2008.

Jiang D, Tang C, Zhang A. 2004. Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering **16(11),** 1370-1386.

Keller JM, Gray MR, Givens JA. 1985. A fuzzy K - nearest neighbor algorithm. IEEE Transactions on Systems, Man and Cybernetics **15(4),** 580-585.

Kumar A, Halder A. 2019. Active learning using fuzzy-rough nearest neighbour classifier for cancer prediction from microarray gene expression data. International Journal of Pattern Recognition and Artificial Intelligence **34(1),** p. 2057001.

Kumar A, Halder A. 2020. Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. Engineering Applications of Artificial Intelligence **91,** p. 103591.

Kuncheva LI. 2004. Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, 2nd ed.

Marak DCB, Halder A, Kumar A. 2021. Semi-supervised Ensemble Learning for Efficient Cancer Sample Classification from miRNA gene expression data. New Generation Computing (Springer) 1-27.

Osareh A, Shadgar B. 2013. An efficient ensemble learning method for gene microarray classification. BioMed Research International **2013(1),** 1-10.

Pawlak Z. 1982. Rough sets. International Journal of Computer and Information Science **11(5),** 341-356.

Polikar R. 2006. Ensemble based systems in decision making", IEEE Circuits and Systems Magazine **6(3),** 21-45.

Priscilla R, Swamynathan S. 2013. A semi-supervised hierarchical approach: two-dimensional clustering of microarray gene expression data. Frontiers of Computer Science **7(2),** 204-213.

Radzikowska AM, Kerre EE. 2002. A comparative study of fuzzy rough sets. Fuzzy Sets and Systems **126,** 137-156.

Stekel D. 2003. Microarray Bioinformatics. 1st ed., Cambridge University Press, Cambridge, UK.

Valentini G, Muselli M, Ruffino F. 2004. Cancer recognition with bagged ensembles of support vector machines. Neurocomputing **56,** 461-466.

Vanitha CDA, Devaraj D, Venkatesulu M. 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science **47,** 13-21.

Yang P, Yang YH, Zhou BB, Zomaya AY. 2010. A review of ensemble methods in bioinformatics. Machine Learning **5(4),** 296-308.

Zadeh L. 1965. Fuzzy sets. Information and Control **8(3),** 338-353.