APPLICATION OF DIMENSIONALITY REDUCTION ON CLASSIFICATION OF COLON CANCER USING ICA AND K-NN ALGORITHM

Rasheed G. Jimoh [1], Ridwan M. Yusuf [1], Yusuf O. Olatunde [1], Saheed Yakub Kayode [2]

[1] **Computer Science Department, University of Ilorin, Ilorin, Kwara State, Nigeria**
[2] **Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria**

Corresponding author: Yusuf O. Olatunde, hizyyusuf@gmail.com

ABSTRACT: Several sectors including engineering, health, academics and so on deals with very large number of information and few specimens. This highlights the need of a technique to improve data accuracy in order to enable professionals such as biologists, clinicians and so on to comprehend the structure of a complex microarray dataset and the gene expression in cells when reduced. This study employs Independent Component Analysis for feature extraction before using *k*-Nearest Neighbor algorithm to classify colon cancer dataset which contains DNA microarray gene expression data with 2000 features and 62 samples. The experiment was performed using MATLAB 2015a. The result shows that the dimensionality reduction applied improve the classification performance in terms of accuracy, sensitivity, specificity and precision by 11.3%, 25.2%, 36.3% and 12.8% respectively.

***KEYWORDS*:** Microarray, k-NN, Dimensionality Reduction, ICA, Colon Cancer.

1. INTRODUCTION

Dimension reduction is an effective and essential tool used to analyze microarray datasets ([APM11]). A lot of algorithms and feature extraction techniques have been put forward in literature the reduction of dimensionality ([RT15]). Principal Component Analysis (PCA) is one of the most widely used and common dimensionality reduction techniques, it is seen as an unsupervised technique and relatively effective tool, but it's not considered as efficient for dataset that are complex and of high dimension ([APM11]). Therefore, there is need to address the inability of PCA to precisely retrieve the genuine latent features of complex datasets ([APM11]). Data in a very high dimensional space often exists in a lower dimensional space and unsupervised feature extraction technique such as PCA may not be totally efficient.

Independent Component Analysis (ICA) is a feature extraction method that utilizes the presence of free factors in multivariate information and breaks down an input dataset into statistically independent components ([X+05]). Independent Component Analysis can lessen the impacts of noise and is ideal for separating diverse signals. Recently, ICA have been used in extracting expression modes of genes in microarray analysis, Moreover, a recent report showed that ICA could enhance the biological legitimacy of the features measured with other methods ([LSC11]).

In this study, Independent Component Analysis (ICA) is use for feature extraction. It is a supervised feature extraction dimensionality reduction method. It choice result from the efficiency rate offered by supervised techniques compare to unsupervised techniques, because the connections among the latent features are much more effective ([LSC11]).

1.1 Independent Component Analysis (ICA)

Independent component analysis (ICA) is named as a computational strategy or system utilized for separating grouped flag into its diminished subcomponents. The "mixed drink party issue" can be portrayed as basic routine with regards to ICA, in which the central discourse signals are split from a specimen information including people speaking together inside a room. This situation is generally translated by considering the non-presence of time delays or echoes. A fundamental factor to be considered is that if N sources are available, at that point in any event N estimations (for example, Amplifiers) are required to mine the primal signs. ICA calculation has the ability to utilize higher request insights which may contain generous corresponding information. This is basically esteemed when the dispatch of information differentiates strikingly from its transcendent parameters, which for no less than two reasons is commonly the circumstance when managed microarray articulation information. Most ICA calculations require "brightened" information, by methods for a personality covariance network. This is more than a factual essential, wherein the calculation goes for crumbling the information past its initial two minutes which is a definitive objective of PCA and its related strategies ([C+14]).

1.2 *K*-Nearest Neighbor

Nearest neighbor search is among the predominant learning and classification methods as presented by ([FH51]), it has been turned out to be a straightforward yet capable acknowledgment calculation. In ([CH67]), it was demonstrated that the choice decide performs well considering that no unequivocal information of the information is accessible. *k*-NN govern is utilized to indicate the basic speculation of this technique, this is a circumstance whereby another example is characterized into the class with the individuals that are most present inside the K Nearest Neighbors, can be utilized to obtain great assessments of the Bayes blunder and its likelihood of mistake asymptotically approaches the Bayes blunder ([DH73]). There are three (3) confinements for the conventional *k*-NN content arrangement ([ST10]):

1. High calculation complexity: keeping in mind the end goal to distinguish the k closest neighbor tests, all similitudes between the preparation tests must be figured. The KNN classifier is not any more ideal when the quantity of preparing test is less. In any case, should the preparation set contain an enormous number of tests, at that point the KNN classifier needs more opportunity to ascertain the similitudes. There are three (3) approaches to take care of this issue which are diminishing the measurements of the element space, utilizing littler datasets and utilizing enhanced calculation.
2. Dependency on the training set: The classifier can be produced just with the preparation tests and it doesn't make utilization of any extra information. This influences the calculation to rely upon the preparation to set too much and recalculation is required paying little mind to whether a little change on preparing set exists
3. No weight difference between samples: All the preparation tests are dealt with similarly; no distinction exists between tests of information with modest number and tests of information with enormous number. So, it doesn't coordinate the genuine wonder where the examples have uneven appropriation ordinarily.

2. RELATED WORK

In 2014, Trstenjak, Mikac and Donko ([TMD14]) introduced the system of *k*-NN with TF-IDF for content order. Quality and speed of the arrangement was absolutely the reason for this system. It discovers protests that are comparable in view of the Euclidean separation and TF-IDF computes the weight for each term in each record. Both *k*-NN and TF-IDF implanted together, demonstrate great together, gave great outcomes and affirmed the underlying desires. Structure is performed on various classifications of reports and the testing is performed. Amid testing, grouping gives exact outcomes because of *k*-NN calculation. This blend gives better outcomes and need to redesign and need to enhance the system for better and high exactness comes about.

Singh and Prakash ([SP14]) reported that consideration of scientists has been gotten by content classification in last ten (10) years with the ascent in electronic substance of the reports. Finding a specific report from the web or any vast storehouse, content or record arrangement is the most reasonable assignment. Some better framework and upgraded machine learning classifiers are requested to do the undertaking of report order. The study composed a multi-operator based framework which contains some product cross breed specialists that recovers classification of an archive and relates with each other in order to take ultimate conclusion about the classification and information is then bolstered to the machine learning classifier in order to upgrade the execution.

In 2014, Lin ([Lin14]) assessed the vitality cost of various classifiers and lessened the vitality cost by parallelization, attempting to distinguish the classifier that performs best on the two parts of adequacy and productivity. The study proposes a novel lexical way to deal with content arrangement in the bio-medicinal space. LK-NN (Lexical *k*-NN) calculation has been proposed, in which lexemes (or tokens) are utilized to speak to the therapeutic archives. These tokens are utilized to characterize abstracts by coordinating them with the standard rundown of watchwords indicated as Work (Medicinal Subject Headings). This naturally arranges diary articles of therapeutic space into particular classes. The study utilized gathering of therapeutic reports, called Ohsumed, as the test information for assessing the proposed approach. The research demonstrates that LK-NN beats the customary *k*-NN calculation as far as standard Fmeasure.

Ganiz, Tutkan and Akyokus ([GTA15]) presented another classifier for printed information, called Administered Meaning Classifier (SMC) was proposed. The new SMC classifier utilizes an importance measure, which depends on Helmholtz standard from Gestalt Hypothesis. In SMC, seriousness of the terms with regards to classes are figured and utilized for arrangement of the archive.

Perez and Marwala ([PM12]) reported that there is a need to isolate numerous accumulations of archives into comparative ones through bunching. Determining number of bunches is required in existing parceling calculations and the yield is absolutely reliant on given information. The study presents an Enhanced Report Grouping (ERG) calculation by creating a number of bunches for any content records and uses cosine comparability measures to put comparative archives in appropriate bunches. Trial comes about demonstrated that precision of proposed calculation is high contrasted with existing calculation as far as F-Measure and time multifaceted nature.

3. METHODOLOGY

In this study, the method used consists of two different phase that is, Feature extraction and Classification. The dataset is loaded as an input to the feature extraction module (shown is figure 1) where ICA is used to remove redundant and irrelevant features. The output of the Feature extraction module (shown in figure 2) serves as input to the classification module where *k*-NN algorithm is used as a classifier and the final output is displayed.

The entire experiment performed is explained in stages.

Stage 1: All the data in colon cancer dataset are classified using *k*-NN algorithm.

Stage 2: ICA is use for feature extraction on the colon cancer dataset, it reduces the data as a sub-optimal of the whole dataset, and classification was then carried out using *k*-NN.

Stage 3: The result of stage 1 and stage 2 are compared based on accuracy, sensitivity, specificity and precision.

3.1 Description of Dataset

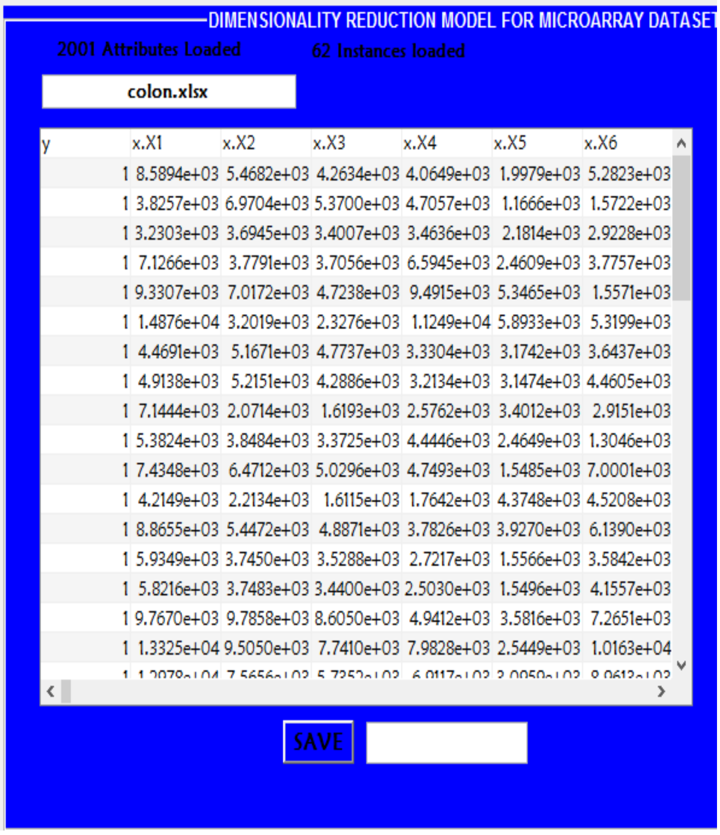
This study uses Colon cancer dataset by Alon. It contains DNA microarray gene expression data with 2000 features and 62 samples. MATLAB 2015a was used as implementing tools.

4. RESULT AND DISCUSSION

Based on the evaluation metrics used, ICA method performs better compare to the KNN method for the microarray analysis of colon cancer as shown in table 1. The result is also presented graphically in figure 3.

Table 1: Comparative Analyses of the Performance Metrics Used

Performance Metrics	ICA based method	*k*-NN based method
Accuracy (%)	88.7	77.4
Sensitivity (%)	*100*	74.1
Specificity (%)	*72.7*	36.4
Precision (%)	86.9	74.1



DIMENSIONALITY REDUCTION MODEL FOR MICROARRAY DATASET

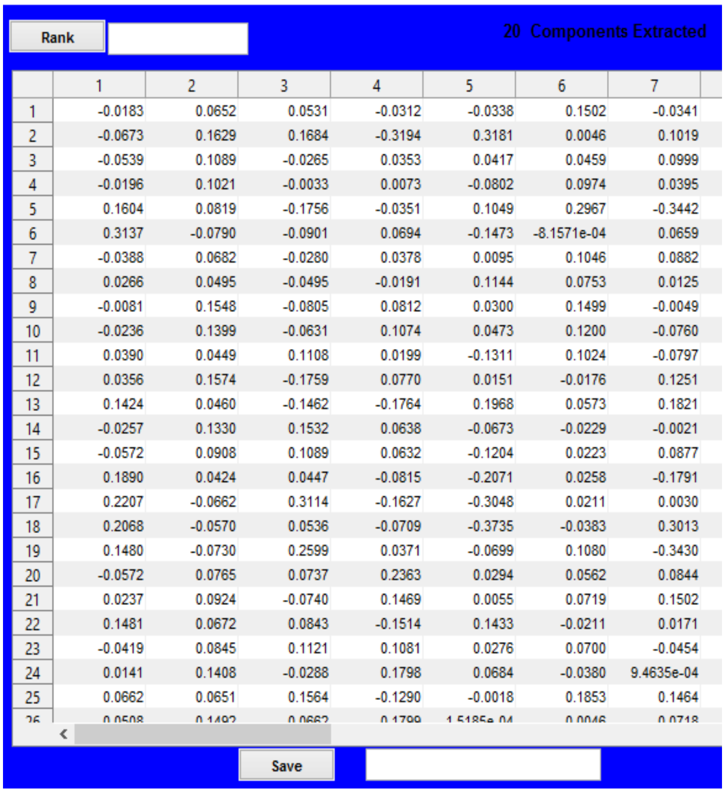
2001 Attributes Loaded

62 Instances loaded

colon.xlsx

y	x.X1	x.X2	x.X3	x.X4	x.X5	x.X6
1	8.5894e+03	5.4682e+03	4.2634e+03	4.0649e+03	1.9979e+03	5.2823e+03
1	3.8257e+03	6.9704e+03	5.3700e+03	4.7057e+03	1.1666e+03	1.5722e+03
1	3.2303e+03	3.6945e+03	3.4007e+03	3.4636e+03	2.1814e+03	2.9228e+03
1	7.1266e+03	3.7791e+03	3.7056e+03	6.5945e+03	2.4609e+03	3.7757e+03
1	9.3307e+03	7.0172e+03	4.7238e+03	9.4915e+03	5.3465e+03	1.5571e+03
1	1.4876e+04	3.2019e+03	2.3276e+03	1.1249e+04	5.8933e+03	5.3199e+03
1	4.4691e+03	5.1671e+03	4.7737e+03	3.3304e+03	3.1742e+03	3.6437e+03
1	4.9138e+03	5.2151e+03	4.2886e+03	3.2134e+03	3.1474e+03	4.4605e+03
1	7.1444e+03	2.0714e+03	1.6193e+03	2.5762e+03	3.4012e+03	2.9151e+03
1	5.3824e+03	3.8484e+03	3.3725e+03	4.4446e+03	2.4649e+03	1.3046e+03
1	7.4348e+03	6.4712e+03	5.0296e+03	4.7493e+03	1.5485e+03	7.0001e+03
1	4.2149e+03	2.2134e+03	1.6115e+03	1.7642e+03	4.3748e+03	4.5208e+03
1	8.8655e+03	5.4472e+03	4.8871e+03	3.7826e+03	3.9270e+03	6.1390e+03
1	5.9349e+03	3.7450e+03	3.5288e+03	2.7217e+03	1.5566e+03	3.5842e+03
1	5.8216e+03	3.7483e+03	3.4400e+03	2.5030e+03	1.5496e+03	4.1557e+03
1	9.7670e+03	9.7858e+03	8.6050e+03	4.9412e+03	3.5816e+03	7.2651e+03
1	1.3325e+04	9.5050e+03	7.7410e+03	7.9828e+03	2.5449e+03	1.0163e+04

Figure 1: Colon cancer dataset loaded for feature extraction process



	1	2	3	4	5	6	7
1	-0.0183	0.0652	0.0531	-0.0312	-0.0338	0.1502	-0.0341
2	-0.0673	0.1629	0.1684	-0.3194	0.3181	0.0046	0.1019
3	-0.0539	0.1089	-0.0265	0.0353	0.0417	0.0459	0.0999
4	-0.0196	0.1021	-0.0033	0.0073	-0.0802	0.0974	0.0395
5	0.1604	0.0819	-0.1756	-0.0351	0.1049	0.2967	-0.3442
6	0.3137	-0.0790	-0.0901	0.0694	-0.1473	-8.1571e-04	0.0659
7	-0.0388	0.0682	-0.0280	0.0378	0.0095	0.1046	0.0882
8	0.0266	0.0495	-0.0495	-0.0191	0.1144	0.0753	0.0125
9	-0.0081	0.1548	-0.0805	0.0812	0.0300	0.1499	-0.0049
10	-0.0236	0.1399	-0.0631	0.1074	0.0473	0.1200	-0.0760
11	0.0390	0.0449	0.1108	0.0199	-0.1311	0.1024	-0.0797
12	0.0356	0.1574	-0.1759	0.0770	0.0151	-0.0176	0.1251
13	0.1424	0.0460	-0.1462	-0.1764	0.1968	0.0573	0.1821
14	-0.0257	0.1330	0.1532	0.0638	-0.0673	-0.0229	-0.0021
15	-0.0572	0.0908	0.1089	0.0632	-0.1204	0.0223	0.0877
16	0.1890	0.0424	0.0447	-0.0815	-0.2071	0.0258	-0.1791
17	0.2207	-0.0662	0.3114	-0.1627	-0.3048	0.0211	0.0030
18	0.2068	-0.0570	0.0536	-0.0709	-0.3735	-0.0383	0.3013
19	0.1480	-0.0730	0.2599	0.0371	-0.0699	0.1080	-0.3430
20	-0.0572	0.0765	0.0737	0.2363	0.0294	0.0562	0.0844
21	0.0237	0.0924	-0.0740	0.1469	0.0055	0.0719	0.1502
22	0.1481	0.0672	0.0843	-0.1514	0.1433	-0.0211	0.0171
23	-0.0419	0.0845	0.1121	0.1081	0.0276	0.0700	-0.0454
24	0.0141	0.1408	-0.0288	0.1798	0.0684	-0.0380	9.4635e-04
25	0.0662	0.0651	0.1564	-0.1290	-0.0018	0.1853	0.1464
26	0.0508	0.1492	0.0662	0.1799	1.5185e-04	0.0046	0.0718

Figure 2: Result of extracted features using ICA

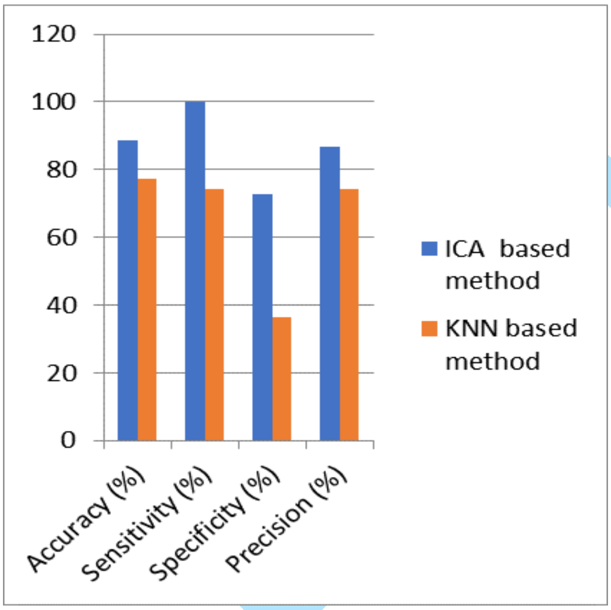


Figure 3: Graphical representation of ICA based method and KNN performance.

5. SUMMARY AND CONCLUSION

This study presents an extraction approach to identify informative features in the bioinformatics field. The ICA principle is use to extract features from high dimensional data and imbalanced data. The approach using feature extraction facilities exploits the features of ICA as a feature extraction and works efficiently on colon cancer data in the field. Using the universal computation of *k*-NN and the evolutionary ability of ICA to extract informative genes from a specific microarray dataset. The result of this study reveal that the ICA method combined with KNN algorithm performs better compare to when only *k*-NN is used by 11.3%, 25.2%, 36.3% and 12.8% in terms of accuracy, sensitivity, specificity and precision respectively. In future, there is need to improve the ICA technique in order to handle data with multiclass more efficiently, reproduce identical sets of extracted features and reuse the selection results.

REFERENCES

[APM11] **Ali A., Paul J., Madhu G. -** *Dimension reduction of microarray Data Based on Local Principal Component*. World Academy of Science, Engineering and technology, (IJCEACIE), Vol. 5, No. 5, pp. 65-71, 2011.

[CH67] **Cover T. M., Hart P. E. -** *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, 13, pp. 21–27, 1967.

[C+14] **Ching S. T., Wai S. T., Mohd S. M., Weng H. C., Safaai D., Zuraini A. S. -** *A Review of Feature Extraction Software for Microarray Gene Expression Data*. BioMed Research International. Pp. 1-12, 2014.

[DH73] **Duda R. O., Hart P. E. -** *Pattern classification and scene analysis*, New York: Wiley, 1973.

[FH51] **Fix E., Hodges J. -** *Discriminatory analysis – Nonparametric discrimination: Consistency properties*. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[GTA15] **Ganiz M. C., Tutkan T., Akyokus S. -** *A Novel Classifier Based on Meaning for Text Classification ICA*, European Journal of Human Genetics, Vol. 13, 2015.

[Lin14] **Lin H. -** *Research on energy-efficient text classification*. 2nd International Conference on Information Technology and Electronic Commerce (ICITEC 2014)

[LSC11] **Liu Q., Sung A.H., Chen Z. -** *Gene selection and classification for cancer microarray data based on machine learning and similarity measures*. BioMed Genomics, vol. 12, No. 5, 2011.

[PM12] **Perez M., Marwala T. -** *Microarray data feature selection using hybrid genetic algorithm simulated annealing*, in Proceedings of the IEEE 27th Convention of Electrical and Electronics Engineers in Israel (IEEEI '12), pp. 1–5, 2012.

[RT15] **Jindal R., Taneja S. -** *A Lexical Approach for Text Categorization of Medical Documents*. Proceeding National Academic. Science, USAVol.96, pp.6745–6750, 2015.

[SP14] **Singh S., Prakash C. -** *Document Categorization in Multi-Agent Environment with Enhanced Machine Learning Classifier*, 2014 Seventh International Conference on Contemporary Computing (IC3), 2014.

[ST10] **Suguna N. N., Thanushkodi K. -** *An Improved k-NN Classification Using GA*. International Journal of Computer Science, Vol. 7, No. 4, pp. 18-21, 2010.

[TMD14] **Trstenjak B., Mikac S., Donko D. -** *KNN with TF-IDF Based Framework for Text Categorization*. Massive Computing Series, Springer, Heidelberg, Germany, pp. 227, 2014

[X+05] **Xue W. Z., Yee L. Y., Dong W., Feng C., Antoine D. -** *Molecular Diagnosis of Human Cancer Type by Gene Expression Profiles and ICA*, European Journal of Human Genetics, Vol. 13, pp. 1301-1311, 2005.