Impact of Feature Selection on Support Vector Machine Using Microarray Gene Expression Data

Choudhury Muhammad Mufassil Wahid, A B M Shawkat Ali, Kevin Tickle
School of Computing Sciences, CQUniversity
QLD 4702, Australia
e-mail:{c.wahid, s.ali, k.tickle}@cqu.edu.au

Abstract—**Recent researches have investigated the impact of feature selection methods on the performance of support vector machine (SVM) and claimed that no feature selection methods improve it in high dimension. However, they have based this argument on their experiments with simulated data. We have taken this claim as a research issue and investigated different feature selection methods on the real time micro array gene expression data. Our research outcome indicates that feature selection methods do have a positive impact on the performance of SVM in classifying micro array gene expression data.**

Keywords- Support Vector Machine, Feature selection, Cancer Classification, Microarray Gene expression data

I. INTRODUCTION

Support Vector Machine (SVM) [1, 2] has recently gained wide popularity among machine learning community due to its robust mathematical basis and high prediction performance. It has been successfully applied to the wide variety of problems including (but not limited to) classification, regression and others [3].

DNA micro array technology is one area where the application of SVM has found broad contemporary acceptance. The technology is used to measure changes in expression levels of genes. This expression of the genetic information occurs in two stages: transcription stage and translation stage. In transcription, DNA molecules are transcribed into mRNA while in translation stage, mRNA is translated to amino acid sequences of the corresponding proteins. DNA micro array analysis provides access to thousands of genes at once by recording expression levels simultaneously. It has been shown that gene expression changes are related with different types of cancers [4]. Cancer classification using gene expression data is a non trivial task due to the very nature of the gene expression data. The expression data has very high dimensionality, usually in the order of thousands to tens of thousands of genes. The situation is more complicated with the number of sample sizes- usually below hundred. The high dimensionality of the features and the low population size usually cause over-fitting of the classifier. A term - the *curse of dimensionality*, is coined to refer to this situation. Computational expenses also impose important limitations. Another key issue is, due to not all genes being related to the cancer, it is difficult to extract biologically meaningful genes.

Problem of feature selection is, hence, an important issue in this research context. Feature selection process comprises selecting relevant features and eliminating irrelevant features from data, and training classifiers on the reduced dimensionality. It has been shown that, in many applications, feature selection process improves a classifier's prediction capability [5]. Nilsson et al [6], however, claims that feature selection has no impact on SVM classifiers in high dimension. But with their study being limited to a set of simulated datasets, the claim is questionable. Considering the importance of feature selection in DNA micro array classification, we investigate in this paper the impact of feature selection on SVM classifier's performance for DNA micro array classification.

The rest of the paper is organized as follows: Section II provides a brief introduction to SVM followed by which, section III defines the cancer classification problem. A brief literature review on SVM and feature selection is provided in section IV. Description of the experimental setup is given in section V which includes biomedical data set description and description of employed feature selection methods. In section VI results and discussions are provided and finally we draw the conclusion.

II. SUPPORT VECTOR MACHINE

The Support vector machine is a classification algorithm rooted in statistical learning theory. Considering a binary classification problem, the SVM constructs a maximum margin hyper-plane that separates the positive tuples from the negative ones. The data points that are closest to the hyper-plane are called support vectors. Only these support vectors define the separating hyper-plane and are used during the classification process. For non-linearly separable data, SVM maps the input data to a high-dimensional feature space in which the data are supposed to be linearly separable. Special functions, termed as kernel functions, are incorporated into the training and classification for data mapped to the high dimensional space [4].

If ($\boldsymbol{x}_i$, y_i) denotes the set of n training points where, $\boldsymbol{x}_i$ are the input data vectors and y_i denotes the class in a two class problem (for $i=1..n$), then the two classes are linearly separable by a hyper-plane (w ,b) if the following condition holds:

978-0-7695-3944-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMV.2009.46

$$\begin{cases} w^T x_i + b >= 0 \;\; for \;\; y_i = +1 \\ w^T x_i + b <= 0 \;\; for \, y_i = -1 \end{cases}$$

Here, w is a vector in the direction of normal to the hyper-plane from origin and b is a scalar called bias.

Let $\phi : I \subseteq R \rightarrow F \subseteq R$ be a mapping from input space I to the feature space F and $< x_1, x_2 >$ denotes the dot product of vectors x_1 and x_2. Then the max-margin hyper-plane is defined as the plane that has the value,

$$\gamma = \min_{i=1}^{n} y_i < w, \phi(t_i) > -b$$

is maximized.

Here the value $(< w, \phi(t_i) > -b)$ is the distance between the point t_i and the hyper-plane in the feature space. It has been shown [3] that for the maximum hyper-plane, the vector w is given by,

$$w = \sum_{i=1}^{n} \alpha_i y_i (x_i)$$

here, $A = \{\alpha_1, \alpha_2, ..., \alpha_n\}$ are positive Lagrange Multipliers that maximize the following dual optimization problem,

$$\sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j < \phi(x_i), \phi(x_j) >$$

subject to the condition,

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \; \alpha_i > 0$$

The classification of testing sample (s, y) is, then, given by,

$$Class(s) = sign(y(< w_0, \phi(s) > - b_0)$$

where w_0 and b_0 denote the max-margin hyper-plane. In a detection problem like cancer identification, if Class(s) is positive, it means that *s* is correctly classified (the presence of cancer has been detected correctly). Otherwise *s* is incorrectly classified.

III. Cancer Classification Problem

Cancer classification problem has been defined and investigated in [4].

Let $X_1, X_2, ..., X_m$ be random variables for genes $G_1, G_2, ..., G_m$ respectively. Let $t = \{t.X_1, t.X_2, . . ., t.X_m\}$ be a size *m* tuple of expression values for *m* genes. Let $T = \{(t_1, y_1), (t_2, y_2), ..., (t_n, y_n)\}$ denoting a training set of *n* tuples, where $y_i \in \{+1, -1\}$ is the class label of tuple t_i, Let $S = \{s_1, s_2, ..., s_l\}$ denoting a test set of l tuples. A classifier is a function class that returns a class prediction for sample *s*.

Classification accuracy is defined as the number of correct predictions made by the classifier trained on the training tuples. Cancer classification problem is to find a classification function Class that gives maximal classification accuracy on *S*.

IV. SVM And Feature Selection

Nilsson et al [6] have verified the impact of feature selection on SVM performance using simulated data. They also have studied on micro array data but have not revealed the nature of their data. They have chosen five feature selection methods namely Pearson Correlation (PC), SVM Naive Weight Rank (WR), Recursive Feature Elimination (RFE), Linear Programming SVM (LPSVM) and Approximation of the zeRO-norm Minimization (AROM). They have taken SVM risk functional as the performance metric and claimed that no feature selection method improves SVM performance.

Alladi et al [7] have investigated performance of SVM, linear regression and neural network on colon tumor data sets after performing feature selection. They have selected 10 and 50 features by t-statistic feature selection method and achieved maximum of 85% accuracy on SVM with RBF kernel.

Guyon et al [8] introduced a novel feature selection method namely recursive feature elimination (RFE) and done experiments on colon tumor and leukemia gene expression dataset. With the colon cancer dataset using 4 genes their method achieved 98% accuracy.

Weston et al [9] introduced a novel feature selection method where to minimize bounds on the leave-one-out error is the selection criteria for the features. . They also compared their algorithm with three other feature selection methods namely Pearson correlation coefficients, the Fisher criterion score and the Kolmogorov-Smirnov test on micro array data along with other real time data sets. Using 20 genes 0 error is made by their algorithm for leukemia data sets.

V. Experimental Setup

After data normalization we have performed 10-fold cross validation on input datasets and average accuracy is taken into account. The 10-fold cross validation is usually performed when the population size is less than one thousand. The missing values are replaced with zero. We have confined our attention to binary class problem. Here we describe the nature of the data set used and the feature selection methods that are employed.

A. Biomedical Data Description

The high-dimensional bio medical data sets have been taken from the Kent Ridge biomedical data repository [10]. The datasets reveal the very nature of the typical micro array gene expression data, namely, high dimensionality and low population size. For example the breast cancer data set included 24481 genes while the population size is only 97. Lung cancer, ovarian cancer and prostate cancer all have number of genes more than 12 thousands with the sample size below 250. The minimum number of genes is included in the Colon Cancer data set. But even in this dataset, number of genes is 2000 while population size is only 62. Table 1 provides a summary of the datasets descriptions.

The Breast Cancer dataset includes data from patients who had developed distance metastases within 5 years labeled as "relapse"; the 'non-relapse' labeled data are taken from the patient who remained healthy from initial diagnosis after five years.

Colon Tumor dataset contains samples collected from colon-cancer patients with biopsies from tumors as well as healthy parts of the colons of the same patients.

There are several kinds of classifications about Diffuse Large B-Cell Lymphoma (DLBCL) datasets. The problem includes differentiating between DLBCL versus FL (Follicular Lymphoma), cured versus fatal, germinal versus activated and alive versus dead samples.

Leukemia dataset consists of bone marrow samples of over 7129 probes from 6817 human genes and contains samples of AML (Acute myeloid leukemia) and ALL (Acute lymphoblastic leukemia) classes.

Lung Cancer dataset provides classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. The other set include samples where patients had experienced relapse of their tumor either locally or as a distant metastasis (labeled as "relapse"). Some samples are from patients who remain disease-free based on both clinical and radiological testing (labeled as "non-relapse").

The ovarian cancer dataset is to identify proteomic patterns in serum that distinguish ovarian cancer from non-cancer.

Prostate Cancer dataset includes samples to be classified as Tumor versus Normal. The other dataset included samples from patients with respect to recurrence following surgery with patients having relapsed and patients remained relapse free ("non-relapse") for at least 4 years.

Central Nervous System dataset is the patient's outcome prediction for central nervous system embryonic tumor. *Survivors* are patients who are alive after treatment while the *failures* are those who died after their disease.

B. Feature Selection

We have investigated three feature selection methods namely Kernel PCA, Greedy Kernel PCA and Generalized discriminant Analysis. Brief descriptions of these are provided below:

Kernel PCA: Kernel principal component analysis [11, 12] termed as KPCA is the non linear extension of the linear PCA. Here the input training vectors $T_x = \{x_1,...,x_l\},\ x_i\, \varepsilon\, X \subseteq R$ are mapped by $\phi : X \rightarrow F$ to a higher dimensional feature space. The kernel PCA trains the kernel data projection $z = A^T k(x) + b$ such that the reconstruction error $\varepsilon_{KMS}(A,b) = \frac{1}{l}\sum_{i=1}^{l}\left\|\phi(x_i) - \tilde{\phi}(x_i)\right\|$ is minimized.

Greedy Kernel PCA: It is termed as GKPCA, which is an efficient algorithm to compute the Kernel PCA [13] . The goal is to train the kernel data projection $z = A^T k_s(x) + b$ where A, b are the parameters and $k_s = [k(x,s_1),...,k(x,s_l)]^T$ are kernel functions centered in the vectors $T_s = \{s_1,...,s_d\}$. The vector set T_s is a subset of the training data. The objective of the GKPCA is to minimize the reconstruction error while the size of the subset T_s is kept small.

GDA: The generalized Discriminant Analysis [14] termed as GDA is the non-linear extension of the linear discriminant Analysis. Input training vectors $T = \{(x_1,y_1),...,(x_l,y_l)\}, x_i\, \varepsilon\, X \subseteq R,\ y\, \varepsilon\, Y = \{+1,-1\}$ are mapped by $\phi : X \rightarrow F$ to a higher dimensional feature space. The linear discriminant analysis is applied on the mapped data. The resulting kernel data projection is defined as $z = A^T k(x) + b$. Parameters (A, b) are trained to increase between-class-scatter and decrease the within-class-scatter of the extracted data.

We have taken SVM accuracy as the performance metric. Linear kernel, polynomial kernel with degree two and three and rbf kernel with fixed sigma value was taken as the kernel mappings. In this experiment there was no intention to check for the optimal value of the kernel parameters. We have extracted 2, 4 and 10 features at a time. The choice of the number of extracted features was arbitrary but common practice in the literature, e.g. [8]. For all these, the non linear feature extraction methods, namely KPCA, GKPCA and GDA were adopted in SVM before starting classification task.

VI. Results And Discussion

Colon Tumor, Leukemia, Lung Cancer (Harvard, Michigan, Ontario), Ovarian Cancer, Prostate Cancer, and Central Nervous System datasets clearly show significant improvement after applying all the three feature selection methods (Fig. 1(a)). Particularly we have found hundred percent accuracy on lung cancer (Harvard, Michigan) and ovarian cancer (NCI PBSII) data sets with linear and polynomial kernels (Fig. 1(c) and Fig. 1(d)). Only Breast Cancer and some DLBCL (Diffuse Large B-Cell Lymphoma) datasets have not shown any significant improvement on classification accuracy (Fig. 1(e)). In some cases the rbf kernel did not perform comparatively well (Fig. 1(b)).

In Fig. 1 the results are shown graphically. For space limitations we have not included all of the charts but showing only selected ones. Detailed results (data and charts) could be available upon request.

VII. Conclusion

From the experiment on real time micro array gene expression data we can draw the conclusion that feature selection methods have a positive impact on SVM performance. But the challenges still remain. Particularly the experiment raises some questions that are needed to be answered and subject to further investigation. Among them some are like: how to find biologically relevant features in other words how to find out those genes that are directly related to cancer and finally how to select optimum value for the kernel parameters. Further research need to be carried out to answer these questions. Future work can be carried out by adding more feature selection methods on other microarray gene expression data.

Acknowledgement

For analyzing the classification accuracy we have utilized the SVM implementation of Statistical Pattern Recognition toolbox for Matlab. The toolbox is freely available software in the internet [15]. We have also made use of the nonlinear feature selection methods implemented in the toolbox.

References

[1] B. Boser, I. Guyon, and V. Vapnik, "A training Algorithm for Optimal Margin Classifiers," Proceedings of 5th Annual ACM Workshop on Computational Learning Theory, pp. 144-152, 1992.

[2] V. Vapnik, "Statistical Learning Theory," Wiley, New York, NY, 1998.

[3] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and other kernel based learning methods," Cambridge University Press., 2004.

[4] L. Ying and H. Jiawei, "Cancer classification using gene expression data," Information Systems, vol. 28, pp. 243-268, 2003.

[5] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[6] R. Nilsson, J. M. Pena, J. Bjorkegren, and J. Tegner, "Evaluating Feature Selection for SVMs in High Dimensions," Proceedings of the 17th european conference on machine learning, pp. 719-726, 2006.

[7] S. M. Alladi, S. Shantosh, V. Ravi, and U. S. Murthy, "Colon Cancer Prediction with genetic profiles using intelligent Techniques," Bioinformation, 2008.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, 2002.

[9] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," Advances in Neural Information Processing Systems, vol. 13, 2000.

[10] "Kent Ridge Bio-medical Dataset," http://datam.i2r.a-star.edu.sg/datasets/krbd/.

[11] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear Component Analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, pp. 1299-1319, 1998.

[12] B. Scholkopf and A. Smola, "Learning With Kernels," 2002.

[13] V. Franc and V. Hlavac, "Greedy Algorithm for a training set reduction in the kernel methods," In N. Petkov and M. A. Westenberg, editors, Computer Analysis of Images and Patterns, pp. 426-433, 2003.

[14] G. Baudat and F. Anouar, "Generalized Discriminant Analysis using a Kernel Approach," Neural Computation, vol. 12, pp. 2385-2404, 2000.

[15] V. Franc and V. Hlavac, "Statistical Pattern Recognition Toolbox for Matlab," Research Reports of CMP, Czech Technical University in Prague, 2004.

TABLE 1: KENT RIDGE BIOMEDICAL DATASETS DESCRIPTION

Data Set	No. of Genes	No. of Instances	Classes	
Breast Cancer	24481	97	Relapse=+1	Non-relapse=-1
Colon Tumor	2000	62	Normal=+1	Tumor=-1
Leukemia ALL-AML	7129	72	ALL=+1	AML=-1
Lung Cancer Harvard 2	12533	181	MPM=+1	ADCA=-1
Lung Cancer Michigan	7129	96	Normal=+1	Tumor=-1
Lung Cancer Ontario	2880	39	Relapse=+1	Non-relapse=-1
DLBCL Harvard (Tumor)	6817	77	DLBCL=+1	FL=-1
DLBCL Harvard (Outcome)	6817	58	Cured=+1	Fatal=-1
DLBCL NIH	7399	240	Alive=+1	Dead=-1
DLBCL Stanford	4026	47	Germinal=+1	Activated=-1
Ovarian Cancer (NCI PBSII)	15154	253	Normal=+1	Cancer=-1
Prostate Cancer	12600	136	Normal=+1	Tumor=-1
Prostate Cancer (Outcome)	12600	21	Relapse=+1	Non-relapse=-1
Central Nervous System	7129	60	survivors=+1	Failures=-1

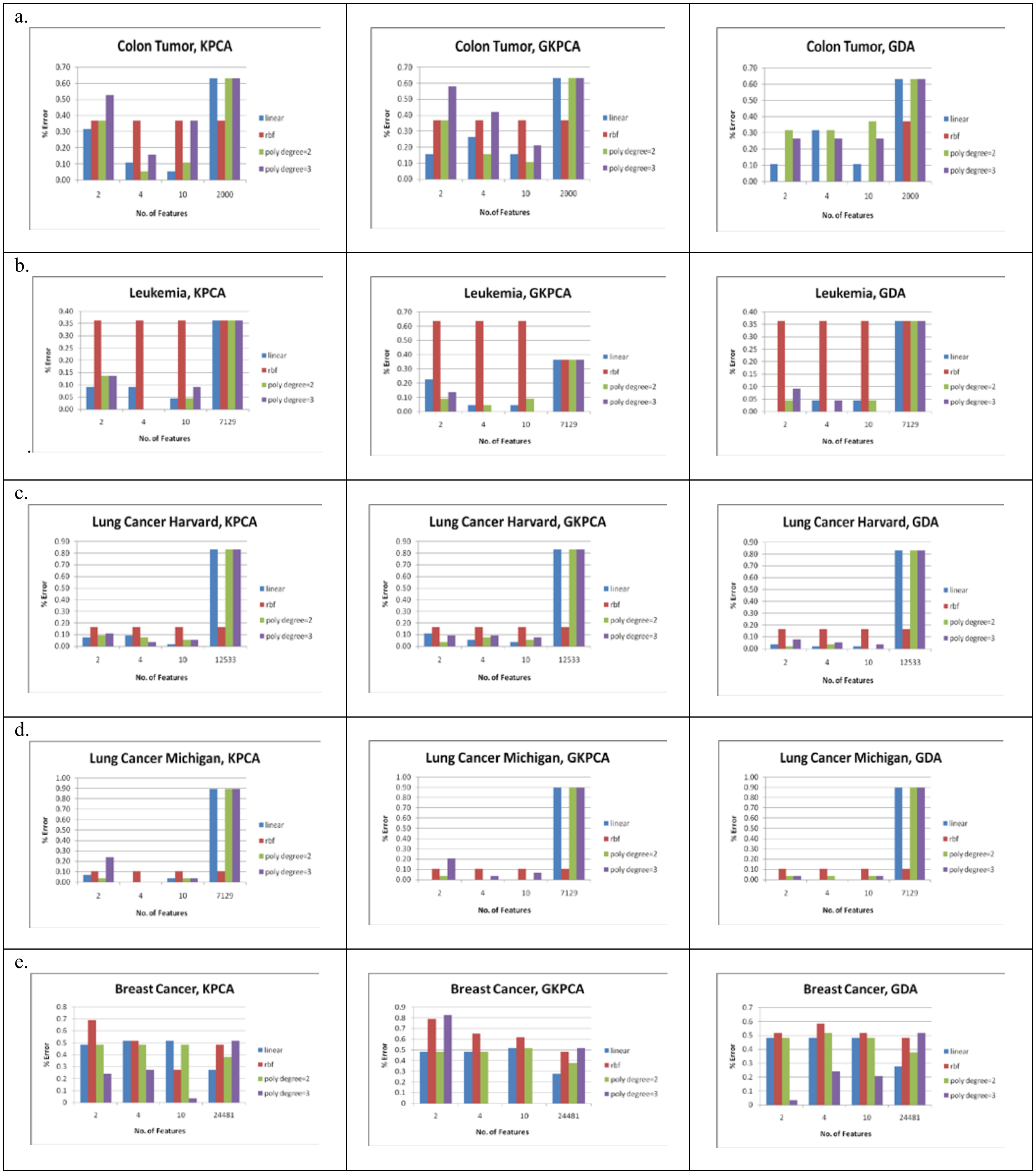

Figure 1: Performance of SVM before and after applying feature selection. In each graph the rightmost columns represent SVM performance (% Error) on the datasets with original dimensions. The other columns at the left represent SVM performance after selecting 2, 4 and 10 features. The columns of charts represent SVM Performance applying KPCA, GKPCA and GDA respectively. (a) Colon Tumor Dataset, (b) Leukemia Dataset, (c) Lung Cancer (Harvard) (d) Lung Cancer (Michigan) Dataset and (e) Breast Cancer Dataset