

# **Biased Support Vector Machine and Kernel Methods for Tumor Classification (Using Microarray Gene Expression Data)**

**Srinivas Mukkamala  
Andrew H Sung  
Krishna Yendrapalli  
Ram Basnet**

**Dep. Computer Science, New Mexico Tech, Socorro, NM, USA**

# Data

- Leukemia, Lymphoma and Prostate cancer data sets are obtained from broad institute
- Colon cancer data set is obtained from Princeton gene expression project

# Data

- Leukemia data set has (37 training samples and 38 testing samples)
- Lymphoma data set has (40 training samples and 39 testing samples)
- Prostate data set has (52 training samples and 52 testing samples)
- Colon data set has (32 training samples and 32 testing samples)

# Leukemia Data

- Leukemia data set comes from a study of gene expression in two types of acute Leukemia: 48 acute lymphoblastic Leukemia (ALL) samples and 25 acute myeloblastic Leukemia (AML) samples

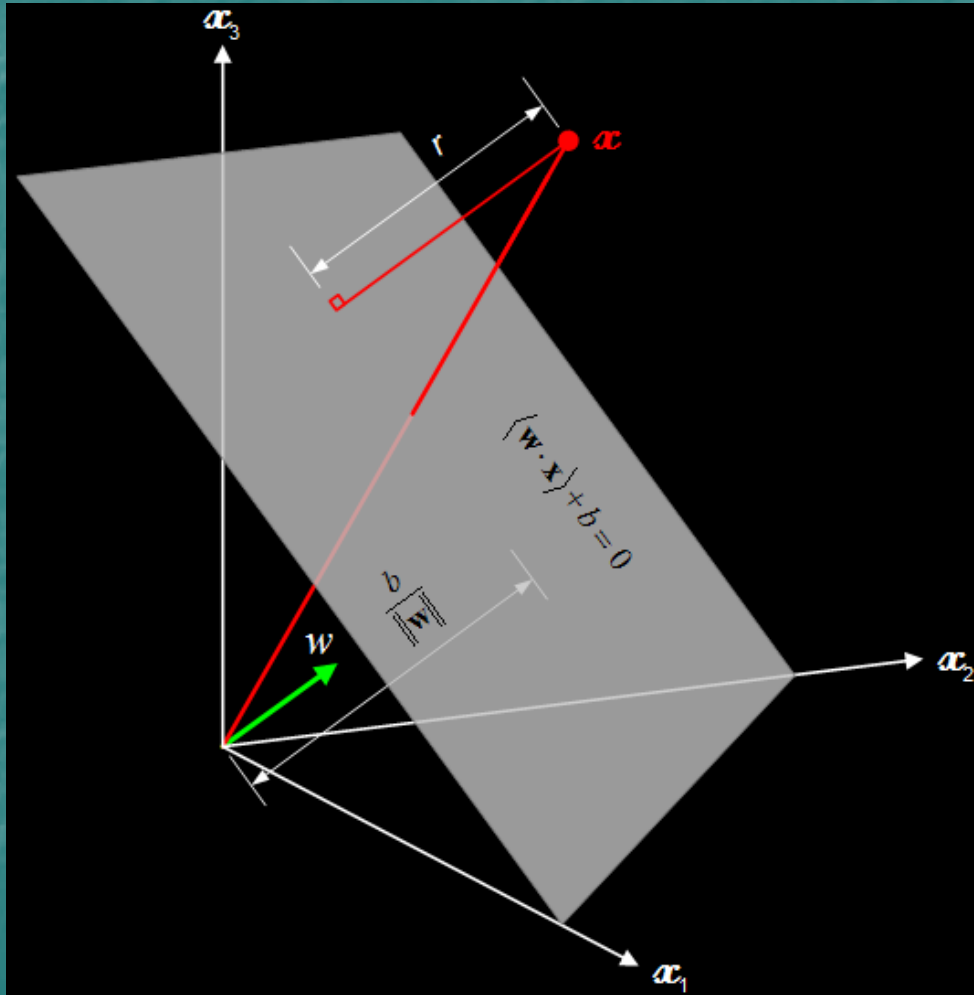
# Lymphoma Data

- Lymphoma data set consists of 58 diffuse large B-cell lymphoma (DLBCL) samples and 19 follicular lymphoma (FL) samples. It was studied in [16]. The data file, lymphoma\_8\_lbc\_fsc2\_rn.res, and the class label file, lymphoma\_8\_lbc\_fsc2.cls are used in our experiments for identifying DLBCL and FL

# SVMs

- Supervised learning problems
  - Classification
    - Regression
- Two key ideas
  - Learn separating hyperplane with maximum margin
  - Expand input into high-dimensional space

# Conceptual Simplicity



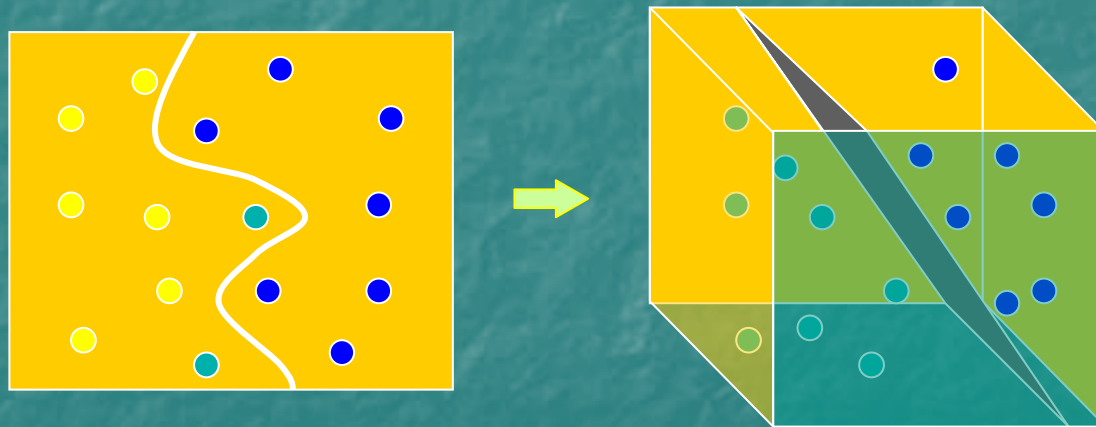
An SVM model defines a hyperplane in the feature space in terms of coefficients  $(\mathbf{w})$   $\frac{b}{\|\mathbf{w}\|}$  and a bias term  $(b)$

Prediction:

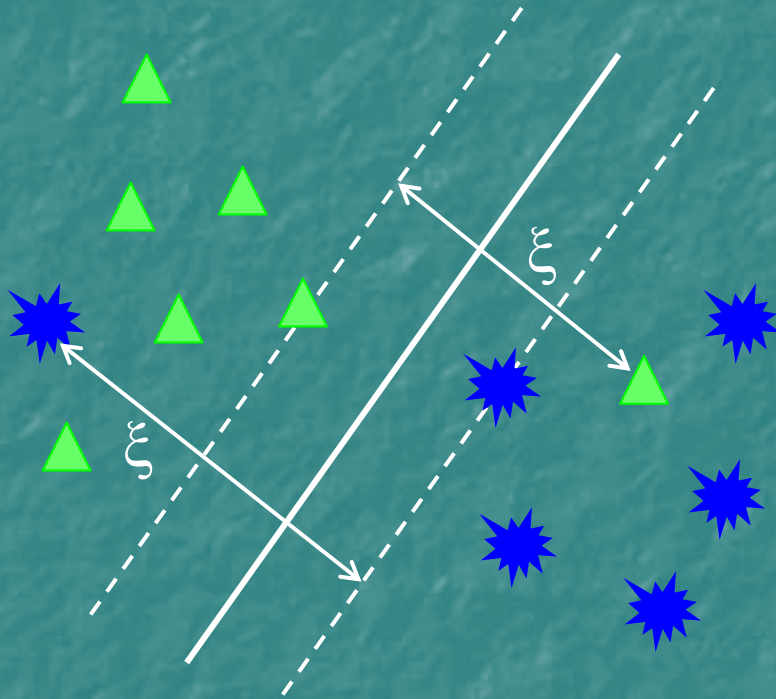
$$f = \text{sign}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b)$$

# Kernel Classifiers

- Transform data via non-linear mapping to an inner product feature space
  - Gaussian, polynomial kernels
- Train a linear machine in the new feature space



# SVM Soft Margin Optimization: Non-Separable Case



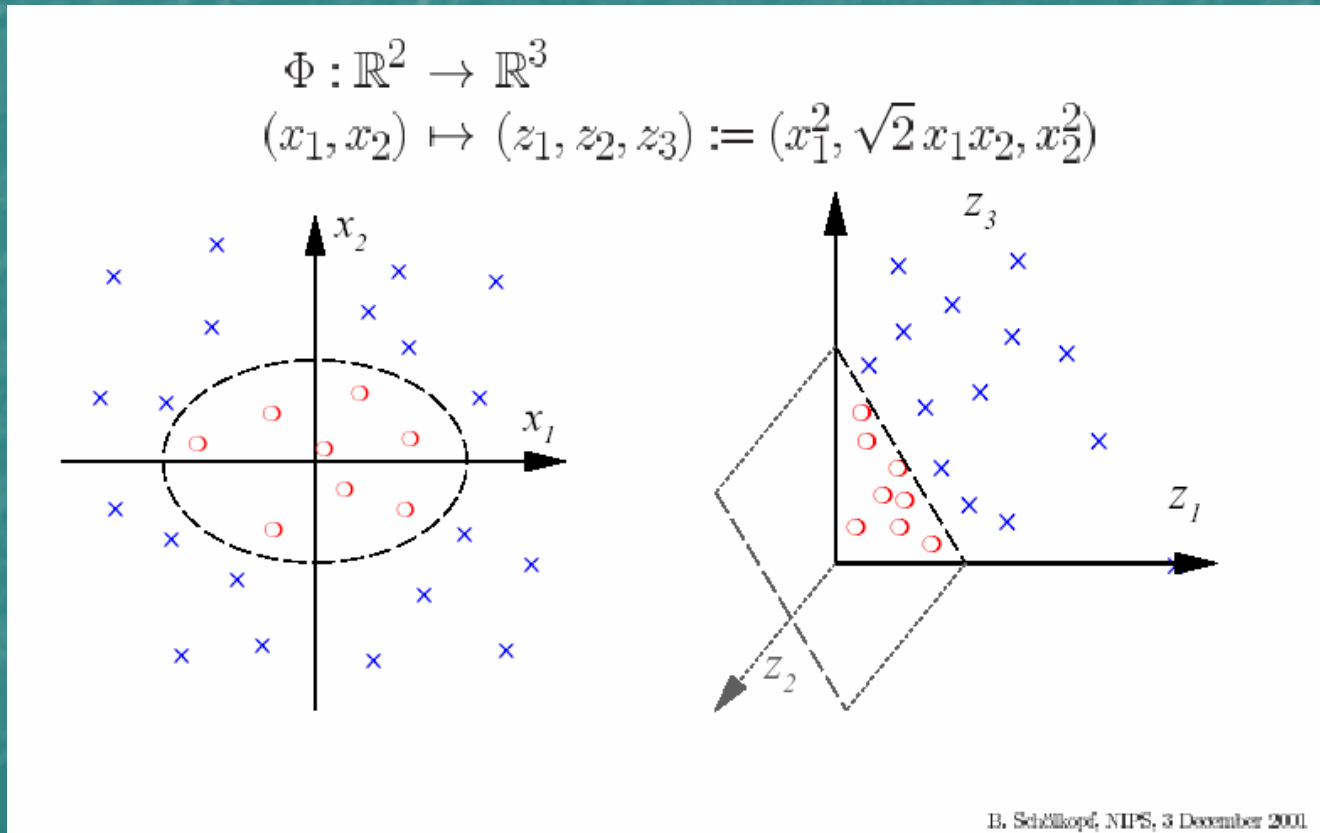
Capacity parameter  $C$  trades off complexity and empirical risk

$$L_p(\mathbf{w}) = \frac{1}{2} \langle \mathbf{w} \cdot \mathbf{w} \rangle + C \sum \xi^k$$

subject to  $y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$

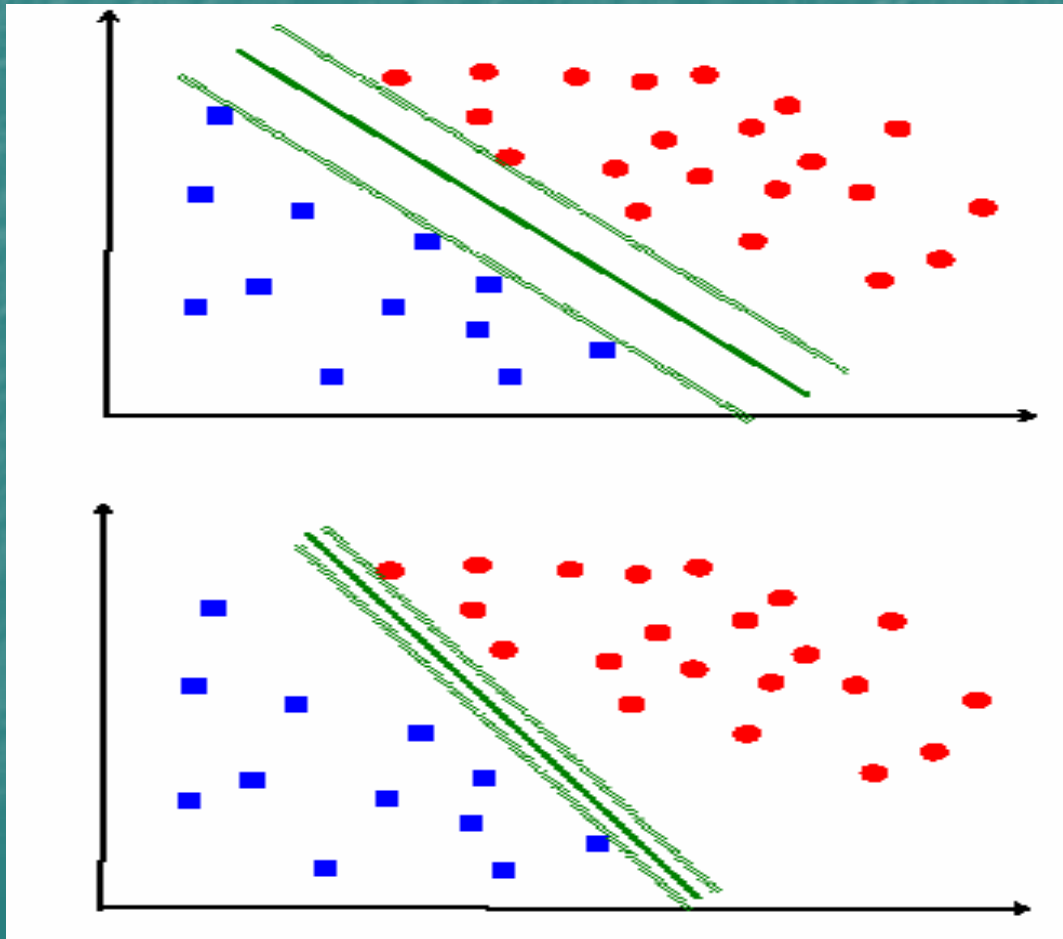
# High Dimension Mapping

- Linear inseparable problem becomes linear separable in higher dimension space

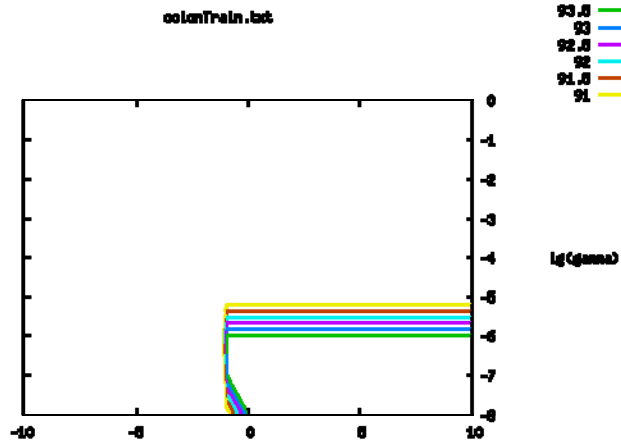


# Optimal Separating Hyperplane

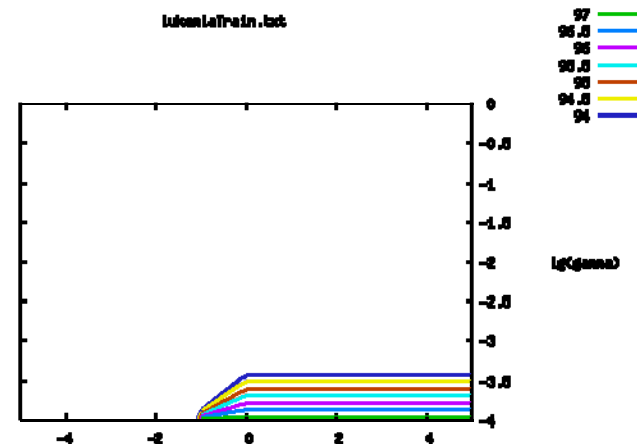
- The optimal separating hyper plane is the one which has the maximal margin



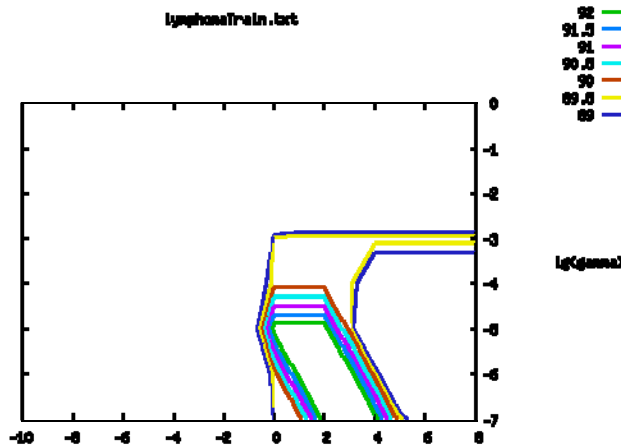
# SVM Model Files



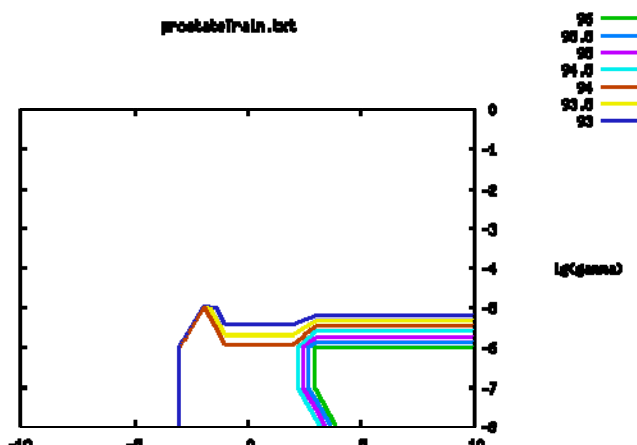
Colon Cancer



Leukemia Cancer

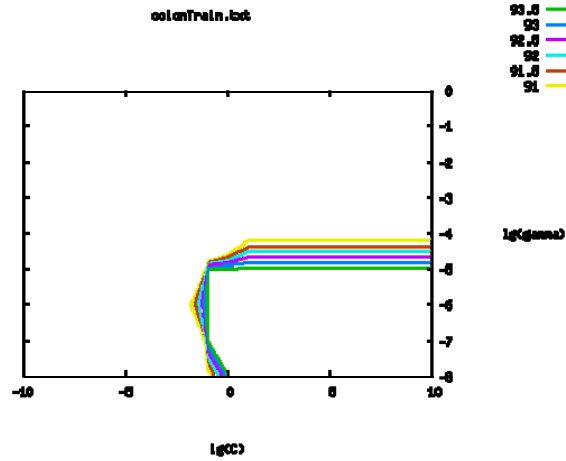


Lymphoma Cancer

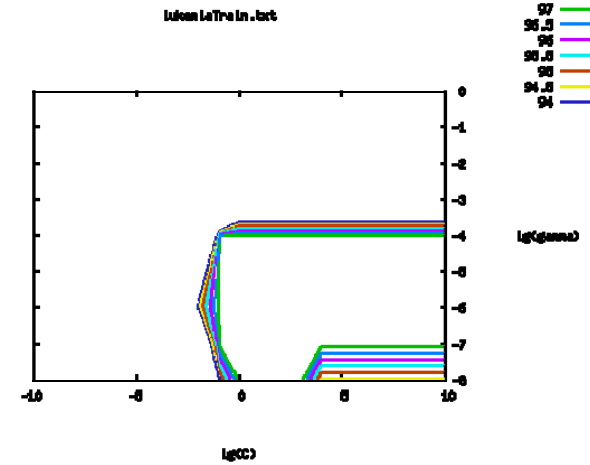


Prostate Cancer

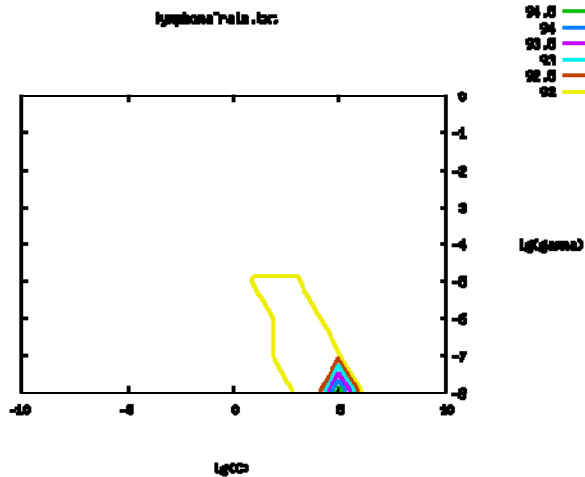
# B-SVM Model Files



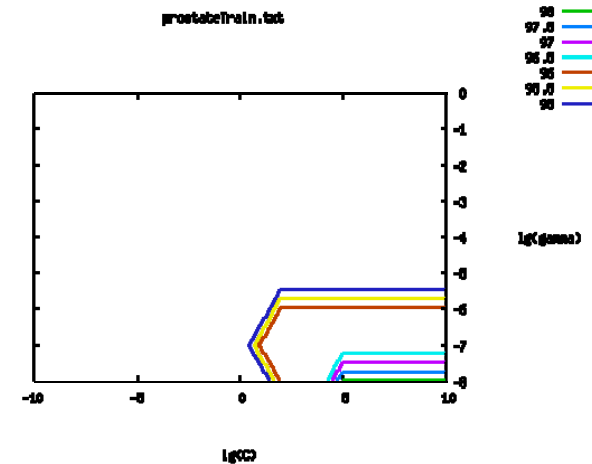
Colon Cancer



Leukemia Cancer

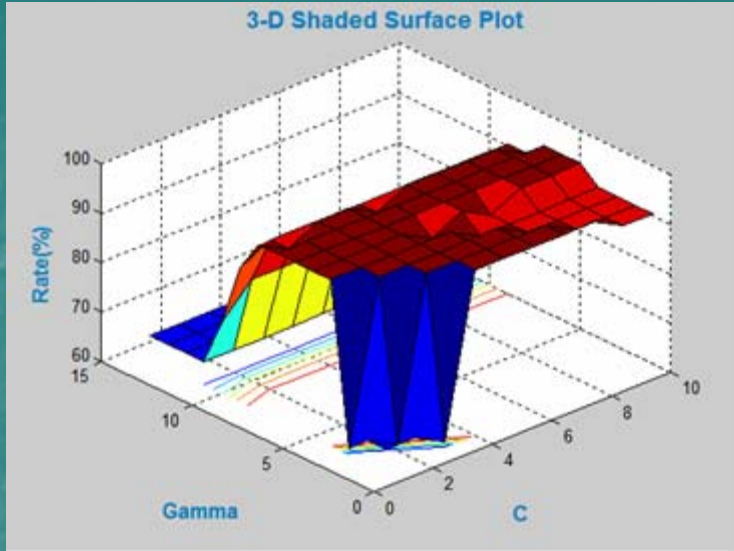


Lymphoma Cancer

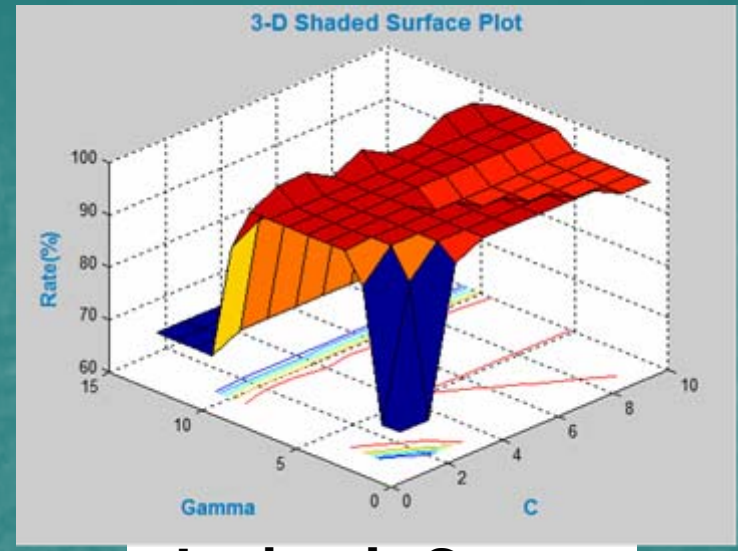


Prostate Cancer

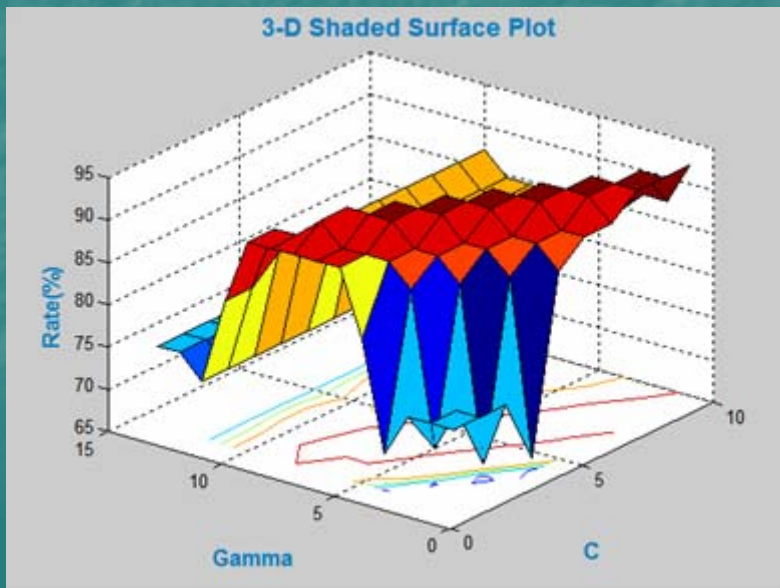
# LOOMS



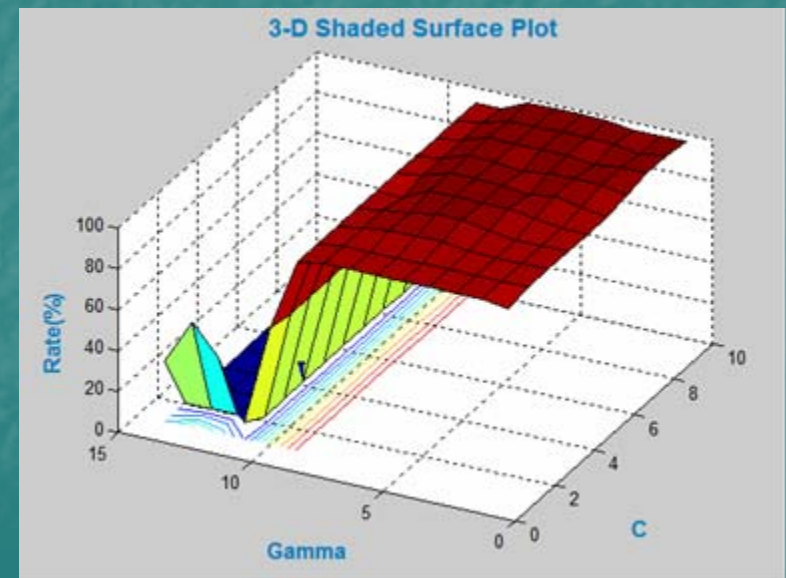
**Colon Cancer**



**Leukemia Cancer**



**Lymphoma Cancer**



**Prostate Cancer**

# Summary and Conclusions

- The Kernel classifiers used in this paper showed comparable or better performance in some cases when compared to the ones reported [artificial neural networks, clustering, expert systems, etc] in the literature using the same datasets
- Our results demonstrate the potential of using kernel machines in diagnosis of malignancy of a tumor
- As a future work we plan to use large datasets of patients. As more inputs are added, feature selection will have to follow a more stringent scrutiny
- Unbalanced samples should be avoided if we can!

# Acknowledgements

- **ICASA (a division of New Mexico Tech)**

# References

- Chapelle, O., Vapnik, V: Model selection for support vector machines. Advances in Neural Information Processing Systems 12, (1999)
- Cherkassy, V: Model complexity control and statistical learning theory. Journal of natural computing 1: (2002) 109–133
- Cristianini, N., Taylor, J.S: Support Vector Machines and Other Kernel-based Learning Algorithms. Cambridge, UK: Cambridge University Press, (2000)
- Chang, C.C, Lin, C.J: LIBSVM: a library for support vector machines. Department of Computer Science and Information Engineering, National Taiwan University, (2001)