**Supplementary Files for** Information Theory-based Feature Selection: Minimum Distribution Similarity with Removed Redundancy

**Supplementary Files**

**Content**

**Figures**

[Figure S1. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances less than 1000. 2](#_Toc31806649)

[Figure S2. Comparison of Acc achieved by different choice of bin number on datasets with total number of instances more than 1000. 3](#_Toc31806650)

[Figure S3. The average accuracies of 10-fold validation with different choices of redundancy remove threshold and different classifiers on different datasets. 4](#_Toc31806651)

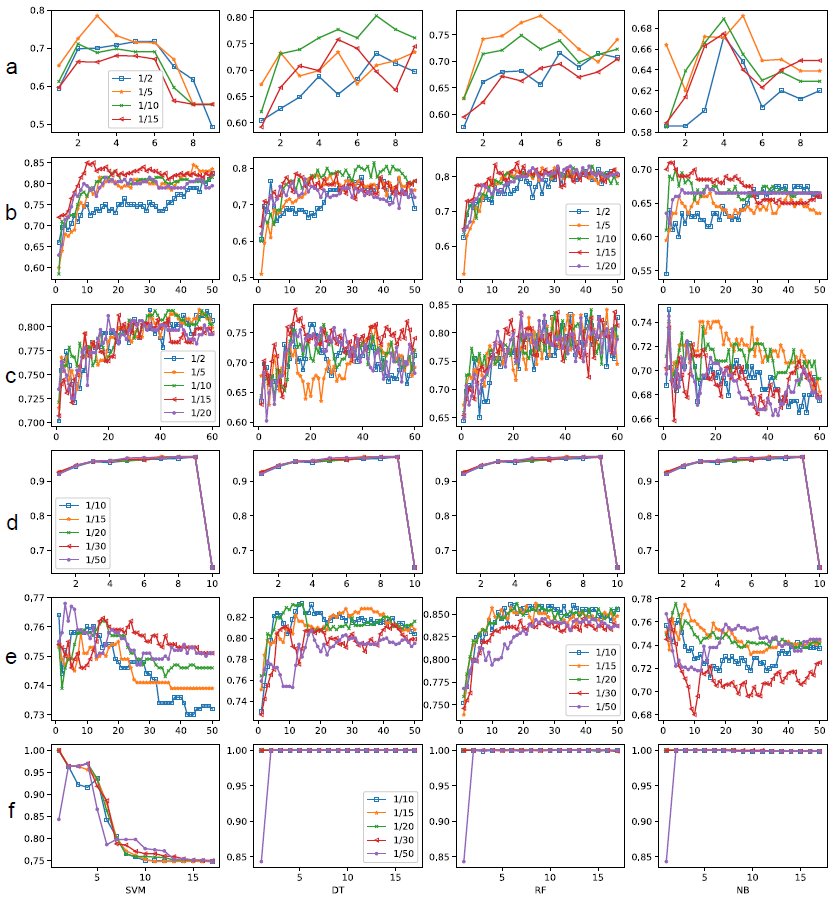
[Figure S4. The average Acc of different feature subset sizes for different feature selection methods with classifier SVM, DT, RF and NB. 5](#_Toc31806652)

**Tables**

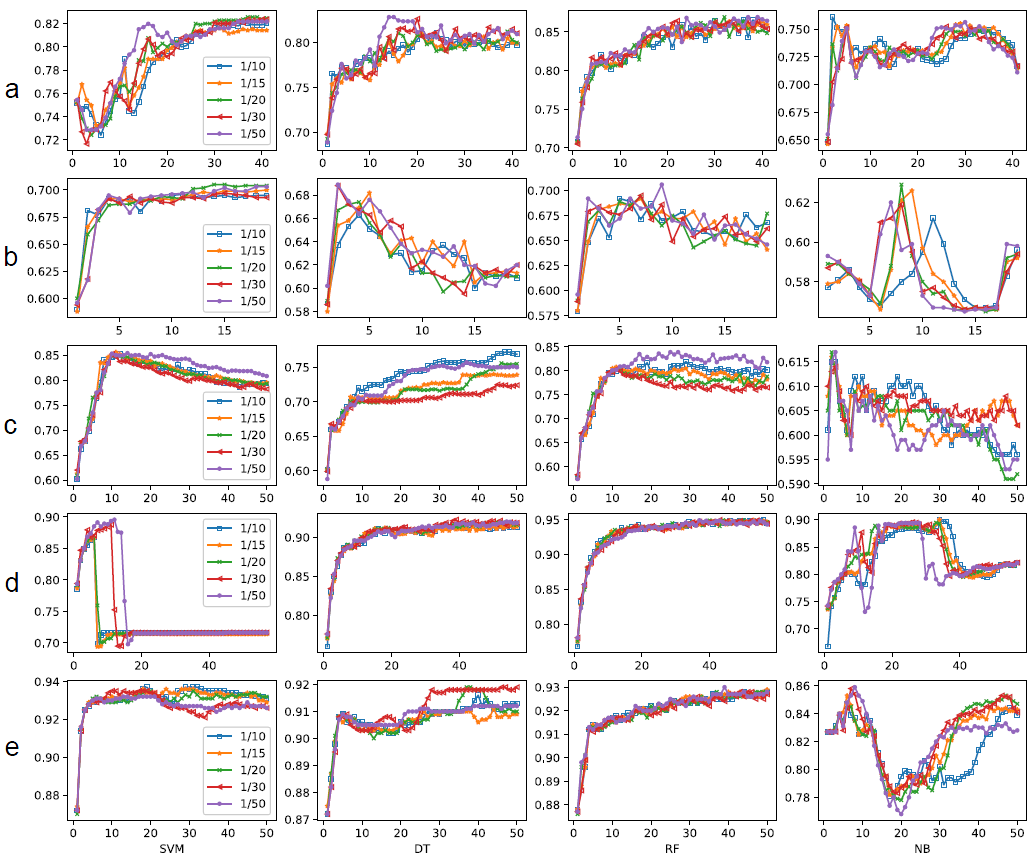
[Table S1. Counts of times for best Acc achieved by different choice of bin number. 6](#_Toc31806653)

[Table S2. The number of features left after removing redundancy with different choice of redundancy remove threshold. 7](#_Toc31806654)

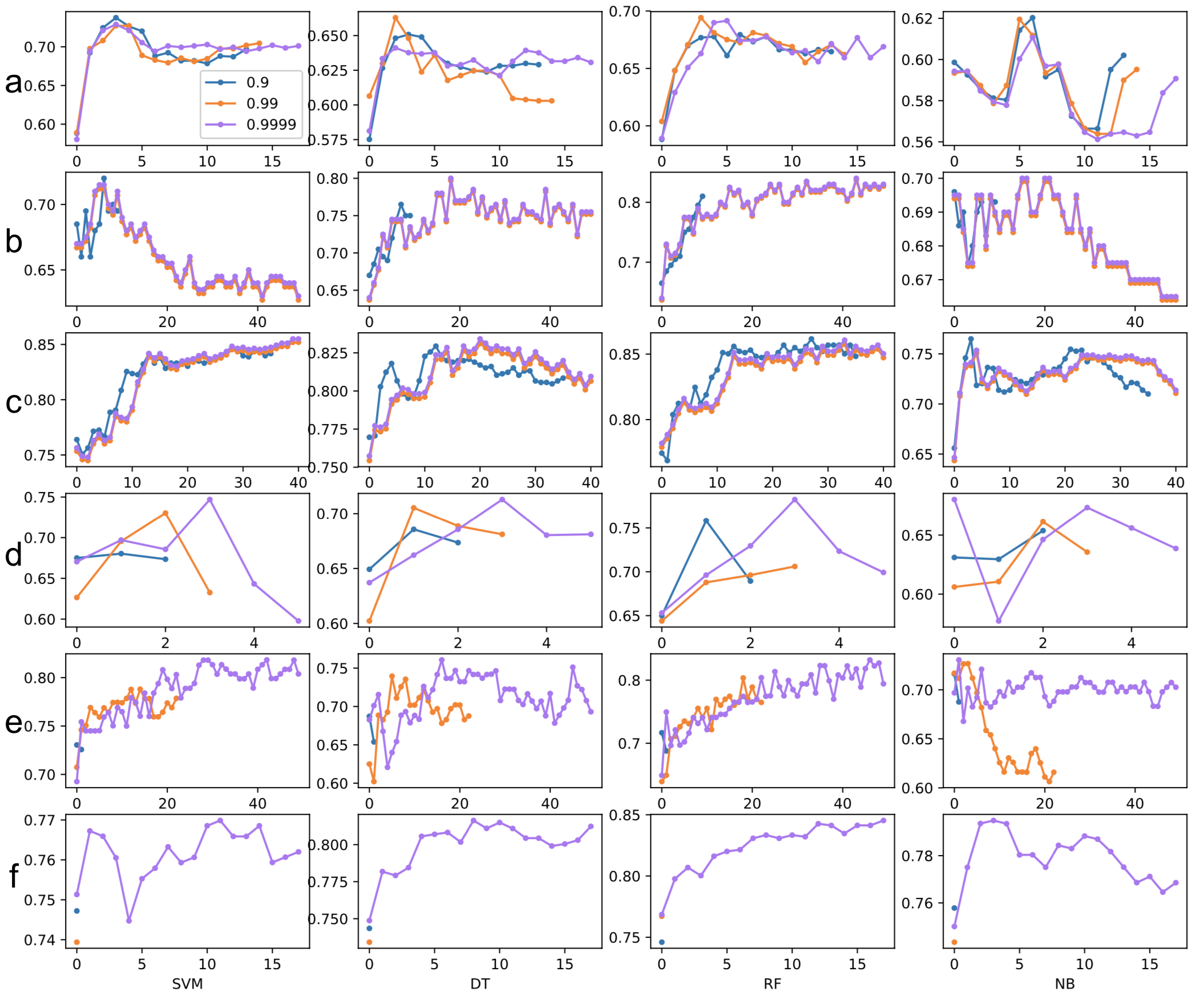
# **Figure S1.** **Comparison of Acc achieved by different choice of bin number on datasets with total number of instances less than 1000.** Dataset (a) Breast Cancer Coimbra, (b) Arcene, (c) Sonar, (d) Breast Cancer Wisconsin, (e) Parkinson and (f) Audit.



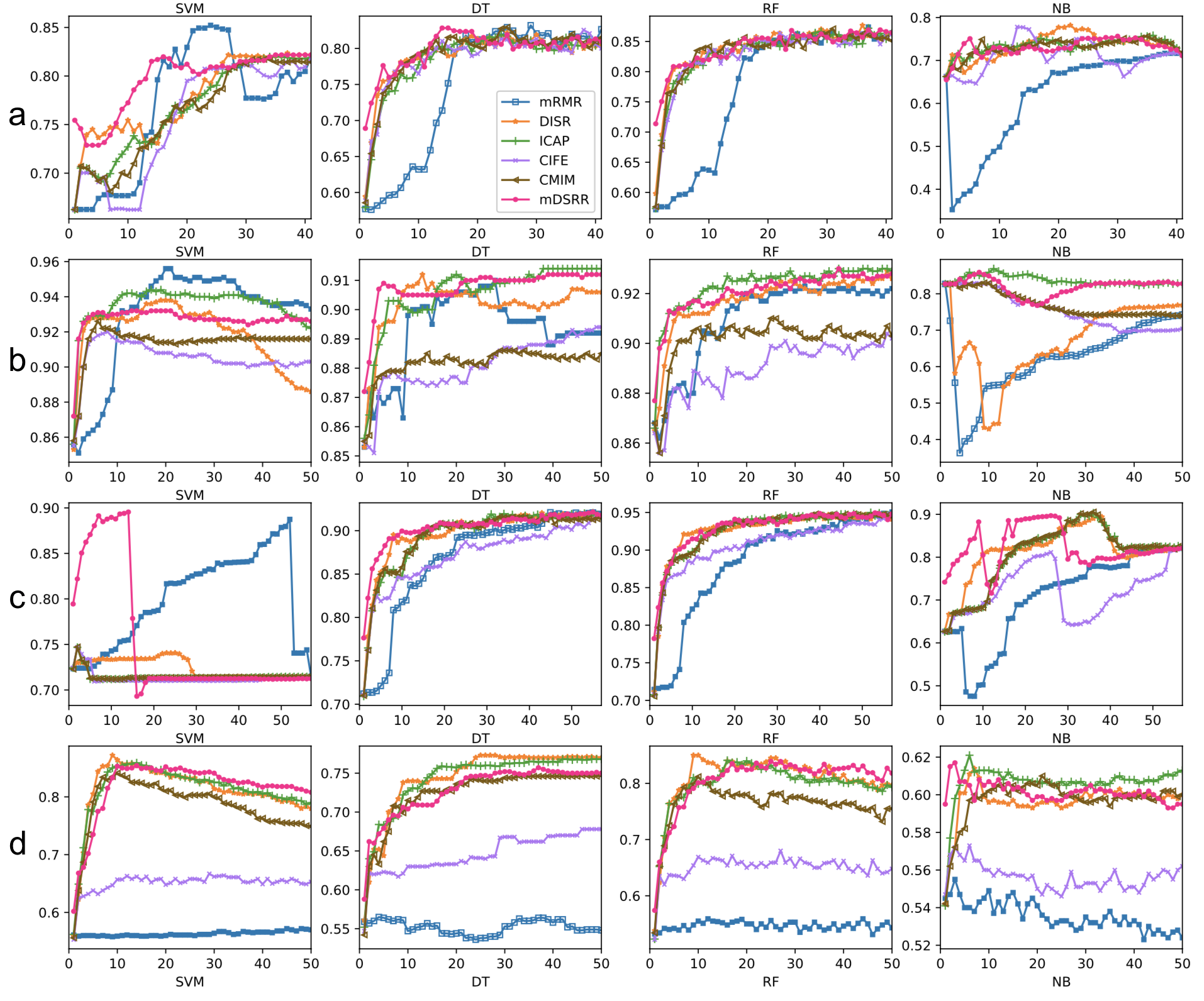
# **Figure S2.** Comparison of Acc achieved by different choice of bin number on datasets with total number of instances more than 1000. Dataset (a) Biodegradation, (b) Diabetic Retinopathy Debrecen, (c) Madelon, (d) Spambase and (e) Musk.



# **Figure S3.** The average accuracies of 10-fold validation with different choices of redundancy remove threshold and different classifiers on different datasets. (a) Diabetic Retinopathy Debrecen, (b) Arcene, (c) Biodegradation, (d) Breast Cancer Coimbra, (e) Sonar and (f) Parkinson.



# **Figure S4.** The average Acc of different feature subset sizes for different feature selection methods with classifier SVM, DT, RF and NB on dataset (a) Biodegradation, (b) Musk, (c) Spambase and (d) Madelon.



# **Table S1**. Counts of times for best Acc achieved by different choice of bin number.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| datasets | 1/2 (percentage of total instance number) | 1/5 (percentage of total instance number) | 1/10 (percentage of total instance number) | 1/15 (percentage of total instance number) | 1/20 (percentage of total instance number) |
| Breast Cancer Coimbra |  | 3 | 1 |  |  |
| Arcene |  |  | 1 | 3 |  |
| Sonar | 1 | 1 | 1 | 1 |  |
| datasets | 1/10 (percentage of total instance number) | 1/15 (percentage of total instance number) | 1/20 (percentage of total instance number) | 1/30 (percentage of total instance number) | 1/50 (percentage of total instance number) |
| Breast Cancer Wisconsin\* | - | - | - | - | - |
| Parkinson | 1 |  | 2 |  | 1 |
| Audit \*\* |  |  | 1 | 1 |  |
| Biodegradation | 1 |  | 1 |  | 2 |
| Diabetic Retinopathy Debrecen |  |  | 2 |  | 2 |
| Madelon | 1 | 1 |  |  | 2 |
| Spambase \*\*\* |  | 1 |  |  | 1 |
| Musk | 1 |  |  | 1 | 2 |

\*the performances of different bin number choices for breast cancer Wisconsin dataset are not distinguishable.

\*\*the choices of bin number as 1/10, 1/15, 1/20 and 1/50 work equally on DT, RF and NB classifiers, 1/20 and 1/30 perform similarly on SVM classifier.

\*\*\* the performances of different bin number choices with DT and RF classifiers are not distinguishable.

# **Table S2**. The number of features left after removing redundancy with different choice of redundancy remove threshold.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset name | Number of features\* | | Threshold 0.9 | | Threshold 0.99 | | Threshold 0.9999 | |
| Arcene | 50 | 10 | | 50 | | 50 | |
| Biodegradation | 41 | 36 | | 41 | | 41 | |
| Breast Cancer Wisconsin | 10 | 10 | | 10 | | 10 | |
| Breast Cancer Coimbra | 9 | 3 | | 4 | | 6 | |
| Diabetic Retinopathy Debrecen | 19 | 14 | | 15 | | 18 | |
| -Madelon | 50 | 50 | | 50 | | 50 | |
| -Musk | 50 | 50 | | 50 | | 50 | |
| Parkinson | 50 | 1 | | 1 | | 18 | |
| Sonar | 50 | 1,2 | | 23 | | 50 | |
| Spambase | 50 | 50 | | 50 | | 50 | |

\*For dataset like Biodegradation, Breast Cancer Wisconsin, Breast Cancer Coimbra and Diabetic Retinopathy Debrecen, the number of features represents their original feature set size; for other datasets, the number of features are taken as 50 and the redundant feature are removed from the first 50 ranked features.

The result of Dataset Audit is not counted as the first feature found by mDSRR realise 100% accuracy.

The number of features left in the Table are the smallest value in 10-fold validation.