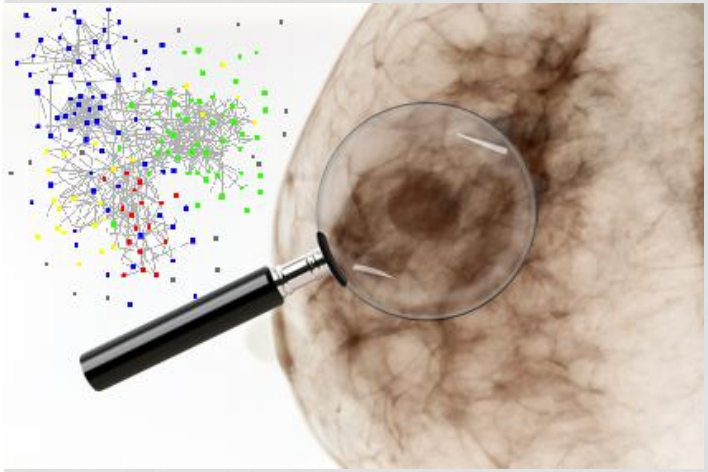
Improving the Mann-Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography

Noel Pérez[1], Miguel A. Guevara[2], Augusto Silva[2] and Isabel Ramos[3]

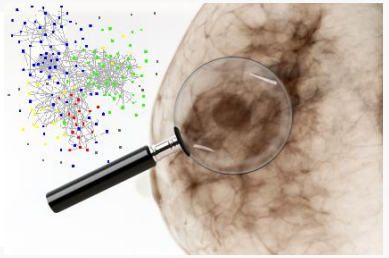


[1] Institute of Mechanical Engineering and Industrial Management (INEGI)
University of Porto, Porto, Portugal
noelperez@outlook.pt

[2] Institute of Electronics and Telematics Engineering of Aveiro (IEETA)
University of Aveiro, Aveiro, Portugal
{mguevaral, augusto.silva}@ua.pt

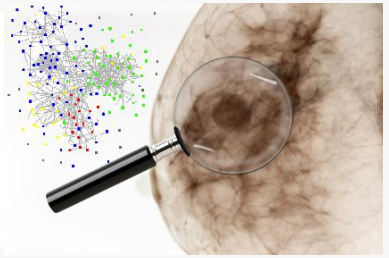
[3] Faculty of Medicine - Centro Hospitalar São João (FMUP-HSJ)
University of Porto, Porto, Portugal
radiologia.hs@mail.telepac.pt

OUTLINE



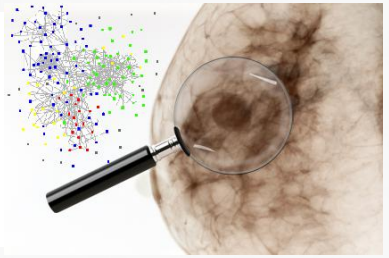
- Introduction
- Proposed Method
- Experimental Evaluation
- Results and Discussions
- Conclusions
- Future Work

INTRODUCTION



- Devijver and Kittler define feature selection as the problem of "extracting from the raw data the information which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability while enhancing the between-class pattern variability".
- Guyon and Elisseeff consider that feature selection addresses the problem of "finding the most compact and informative set of features, to improve the efficiency or data storage and processing".

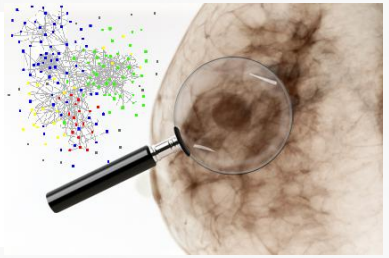
INTRODUCTION



- During the last decade parallel efforts from researchers in statistics, machine learning, and knowledge discovery have been focused on the problem of feature selection and its influence in machine learning classifiers.

- Feature selection lies at the center of these "efforts" with applications in the pharmaceutical and oil industry, speech and pattern recognition, biotechnology and many other emerging fields with significant impact in health systems for cancer detection/classification.

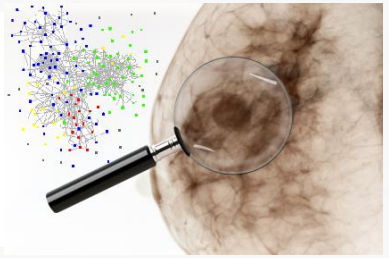
INTRODUCTION



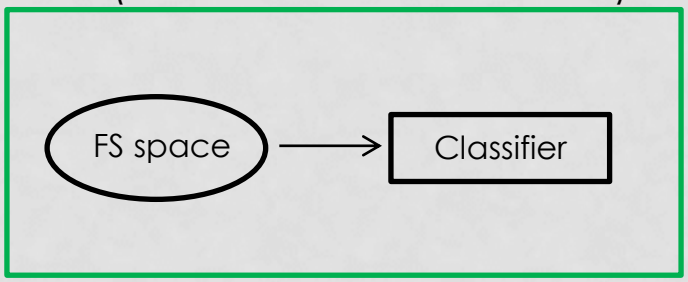
- The potential benefits include: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defining the curse of dimensionality to improve the predictions performance.

- The objectives are related: to avoid overfitting and improve model performance; to provide faster and more cost-effective models, and to gain a deeper insight into the underlying processes that generated the data.

INTRODUCTION



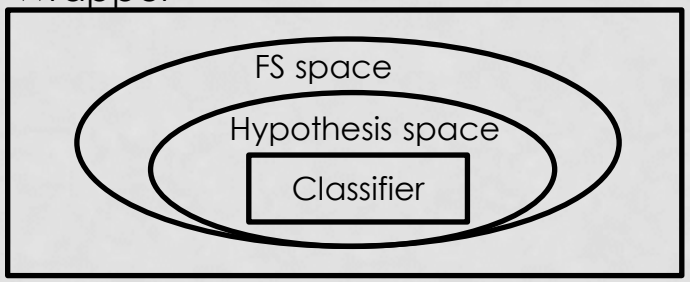
Filter (Univariate and Multivariate)



Advantages
- Fast
- Scalable
- Independent of classifier

Disadvantages
- Ignores feature dependencies

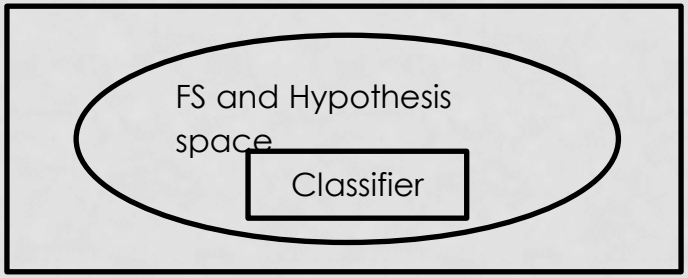
Wrapper



- Interacts with the classifier
- Models feature dependencies

- Risk of data over fitting
- More prone to getting stuck in a local optimum
- Classifier dependent selection

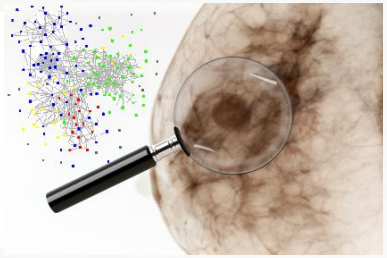
Embedded



- Interacts with the classifier
- Better computational complexity than wrapper
- Models feature dependencies

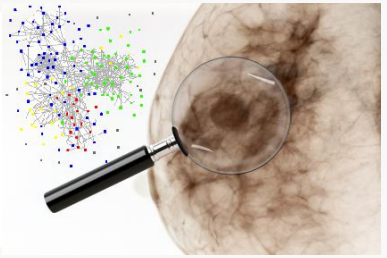
- Classifier dependent selection

INTRODUCTION



- Univariate filter methods, such as chi-square (CHI2) discretization, t-test, information gain (IG) and gain ratio, present two main disadvantages:
- (1) ignoring the dependencies among features and
- (2) assuming a given distribution (Gaussian in most cases) from which the samples (observations) have been collected. In addition, to assume a Gaussian distribution includes the difficulties to validate distributional assumptions because of small sample sizes.
- Multivariate filters methods such as: correlation based-feature selection, Markov blanket filter, fast correlation based-feature selection, ReliefF overcome the problem of ignoring feature dependencies introducing redundancy analysis (models feature dependencies) at some degree, but the improvements are not always significant: domains with large numbers of input variables suffer from the curse of dimensionality and multivariate methods may overfit the data. Also, they are slower and less scalable than univariate methods.

INTRODUCTION



- We considered developing the uFilter feature selection method based on the Mann–Whitney U-test, in a first approach, to be applied in binary classification problems. The uFilter algorithm is framed in the univariate filter paradigm since it requires only the computation of *n* scores and sorting them. Therefore, its time execution (faster) and complexity (lower) are beneficial when is compared to wrapper or embedded methods.

- the uFilter method is an innovative feature selection method for ranking relevant features that assess the relevance of features by computing the separability between class-data distribution of each feature.

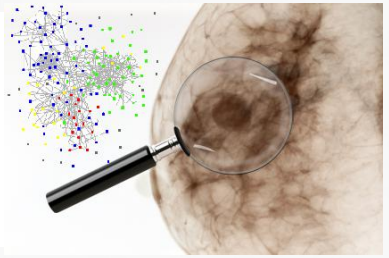
- It solves some difficulties remaining on previous methods, such as:

1. it is effective in ranking relevant features independently of the samples sizes (tolerant to unbalanced training data).

2. it does not need any type of data normalization.

3. it presents a low risk of data overfitting and does not incur the high computational cost of conducting a search through the space of feature subsets as in the wrapper or embedded methods.

PROPOSED METHOD



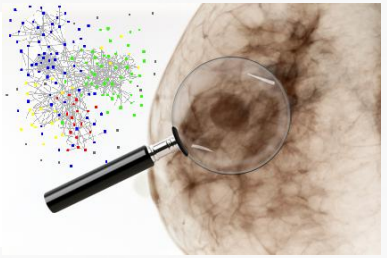
- Foundation

- The Mann–Whitney U-test is a non-parametric method used to test whether two independent samples of observations are drawn from the same or identical distributions. U-test is based on the idea that the particular pattern exhibited when *m* number of *X* random variables and *n* number of *Y* random variables are arranged together in increasing order of magnitude provides information about the relationship between their parent populations.

- Hypothesis evaluated:

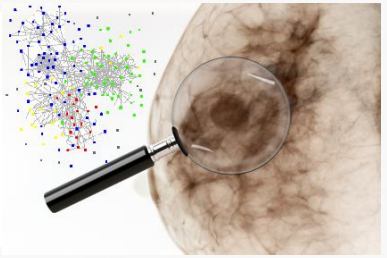
- Do two independent samples represent two populations with different median values (or different distributions with respect to the rank-orderings of the scores in the two underlying population distributions)?

PROPOSED METHOD



- The overall procedure for carrying the U-test:
- 1. Arrange all the *N* observations (scores) in order of magnitude (irrespective of group membership).
- 2. All *N* scores are assigned a rank.
- 3. The ranks must be adjusted when there are tied scores present in the data.
- 4. The sum of the ranks for each of the groups is computed: $\sum R_x$ and $\sum R_y$
- 5. The values Ux and Uy are computed employing: $U_x = n_x n_y + [n_x(n_x+1)/2] - \sum R_x$ and $U_y = n_x n_y + [n_y(n_y+1)/2] - \sum R_y$
- 6. Calculate $U = min(U_x, U_y)$. The smaller of the two values U_x versus U_y is designated as the obtained *U* statistic.
- 7. Use statistical tables for the Mann-Whitney U-test to find the probability of observing a value of *U* or lower than the tabled critical value at the prespecified level of significance.
- 8. Interpretation of the test results (accept or reject the null hypothesis).

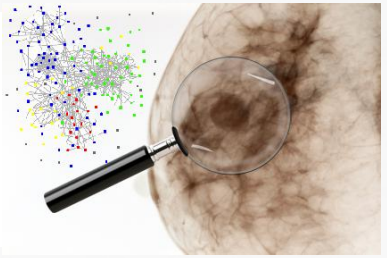
PROPOSED METHOD



Algorithm 1: uFilter

1. Let F a set of features and F_i the i^{th} –feature under analysis, $i: 1..t; t$ = total of features
2. Let $F_i = \{I_{c,1}, I_{c,2}, \dots, I_{c,t}\}$ where $I_{c,j}$ is an instance, $j: 1..n; n$ = total of instances and c is the class value (B or M)
3. For each F_i
   a. Initial weight of the feature $w_i = 0$
   b. $\text{Sort}(F_i, \text{'ascendant'})$
   c. Perform the tie analysis of resultant in b:
      Range $R = \text{avg}(\text{position of tied elements})$
   d. Compute the range summatory of benign and malignant instances $S_B = \sum_{j=1}^{T_B} R_j$ and $S_M = \sum_{j=1}^{T_M} R_j$, where T_B and T_M are the totals of benign and malignant instances
   e. Compute u-values:
      $uB = n_B n_M + \frac{n_B(n+1)}{2} - S_B$ and $uM = n_B n_M + \frac{n_M(n_M+1)}{2} - S_M$
   f. Compute z-values:
      $z_B = \frac{uB - \bar{u}}{\sigma_u}$ and $z_M = \frac{uM - \bar{u}}{\sigma_u}$ where, $\bar{u}$ is the mean and the standard deviation
      $\sigma_u = \sqrt{\frac{n_B n_M}{n(n-1)}\left(\frac{n^3 - n}{12} - \sum_i^k \frac{l_i^3 - l_i}{12}\right)}$; k is the total of range where had tied elements and l_i means the total of tied elements within the range k.
   g. Updating the weight of the feature $w_i = |z_B - z_M|$
4. End for
5. Output ranking $\text{Sort}(w, \text{'descendant'})$

EXPERIMENTAL EVALUATION



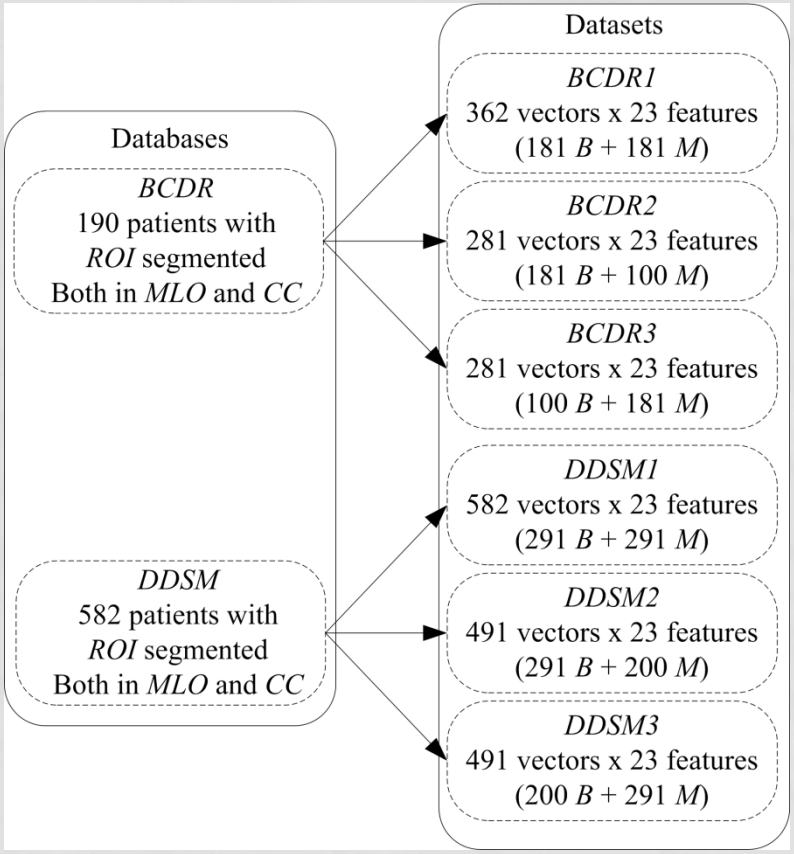
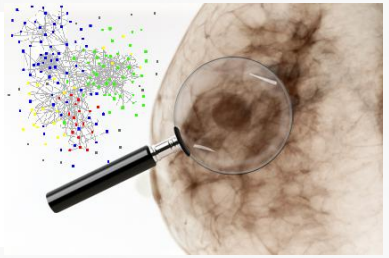


Fig. 1. Datasets creation; B and M represent benign and malignant class instances.

• The Breast Cancer Digital Repository (BCDR) is a wide ranging annotated Portuguese Breast Cancer database, with 1734 anonymous patient cases from medical historical archives supplied by Faculty of Medicine - Centro Hospitalar de São João at University of Porto, Portugal. The BCDR supplies several datasets for scientific purposes (Availaible on http://bcdr.inegi.up.pt), we used the BCDR-F01 distribution for a total of 362 features vectors.

• The Digital Database for Screening Mammography (DDSM) is composed by 2620 patient cases divided into three categories: normal cases (12 volumes), cancer cases (15 volumes) and benign cases (14 volumes). We considered only two volumes of cancer and benign cases (random selection) for a total of 582 features vectors.

EXPERIMENTAL EVALUATION



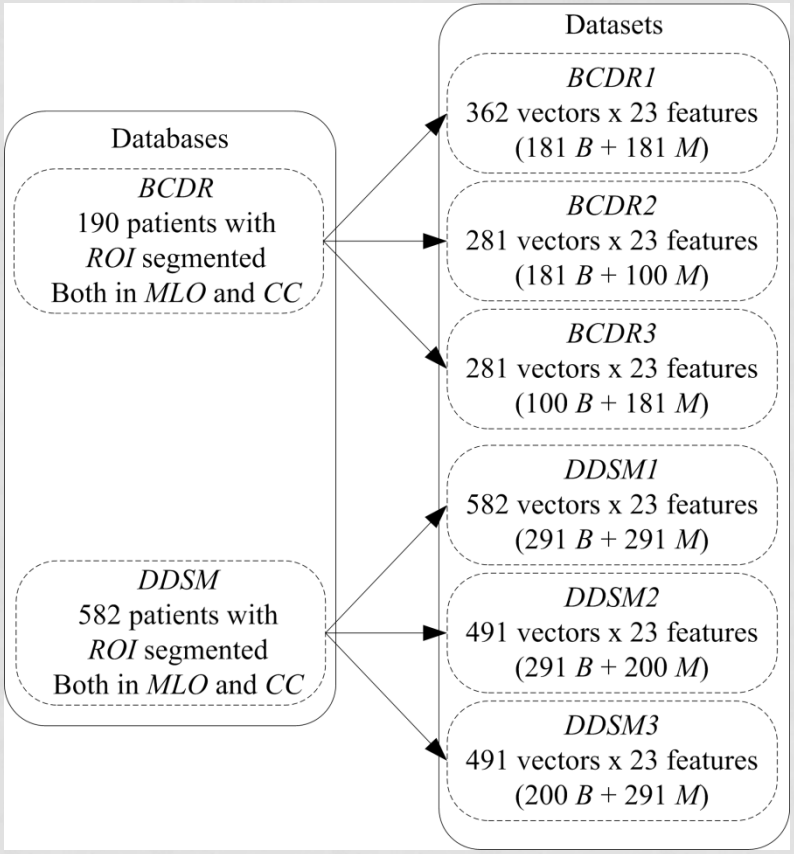


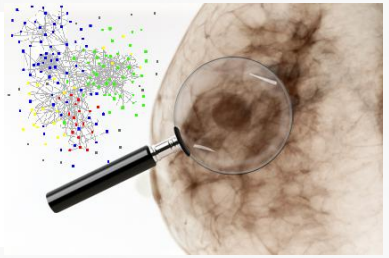
Fig. 1. Datasets creation; B and M represent benign and malignant class instances.

• A set of 23 image-based descriptors (features) were extracted from the BCDR and DDSM databases to be used in this work. Selected descriptors included intensity statistics, shape and texture features, computed from segmented calcifications and masses in both MLO and CC mammography views.

• Conformable to the number of patient cases of used databases, it were created six datasets containing calcifications and masses lesions with different configurations:

- BCDR1 and DDSM1 balanced datasets (same quantity of benign and malignant instances).
- BCDR2 and DDSM2 unbalanced datasets containing more benign than malignant instances.
- BCDR3 and DDSM3 unbalanced datasets holding more malignant than benign instances.

EXPERIMENTAL EVALUATION



The overall procedure for the uFilter evaluation involves five main steps:

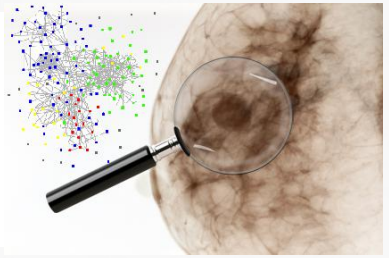
•Applying the classical Mann–Whitney U-test (U-test), the new proposed uFilter method and four well known feature selection methods: CHI2 discretization (CHI2), Information Gain (IG), One Rule (1Rule) and Relief to the six previously formed breast cancer datasets.

• Creating several ranked subset of features using increasing quantities of features. The top *N* features of each ranking (resultant from the previous step) were used for feeding different classifiers, with *N* varying from 5 to the total number of features of the dataset, with increments of 5.

• Classifying the generated ranked subset of features using FFBP neural network, SVM, LDA and NB classifiers for a comparative analysis of AUC scores. All comparisons were using the Wilcoxon statistical test to assess the meaningfulness of differences between classification schemes.

• Selecting the best classification scheme on datasets (BCDR1,BCDR2, BCDR3, DDSM1, DDSM2 and DDSM3), and thus the best subset of features.

EXPERIMENTAL EVALUATION



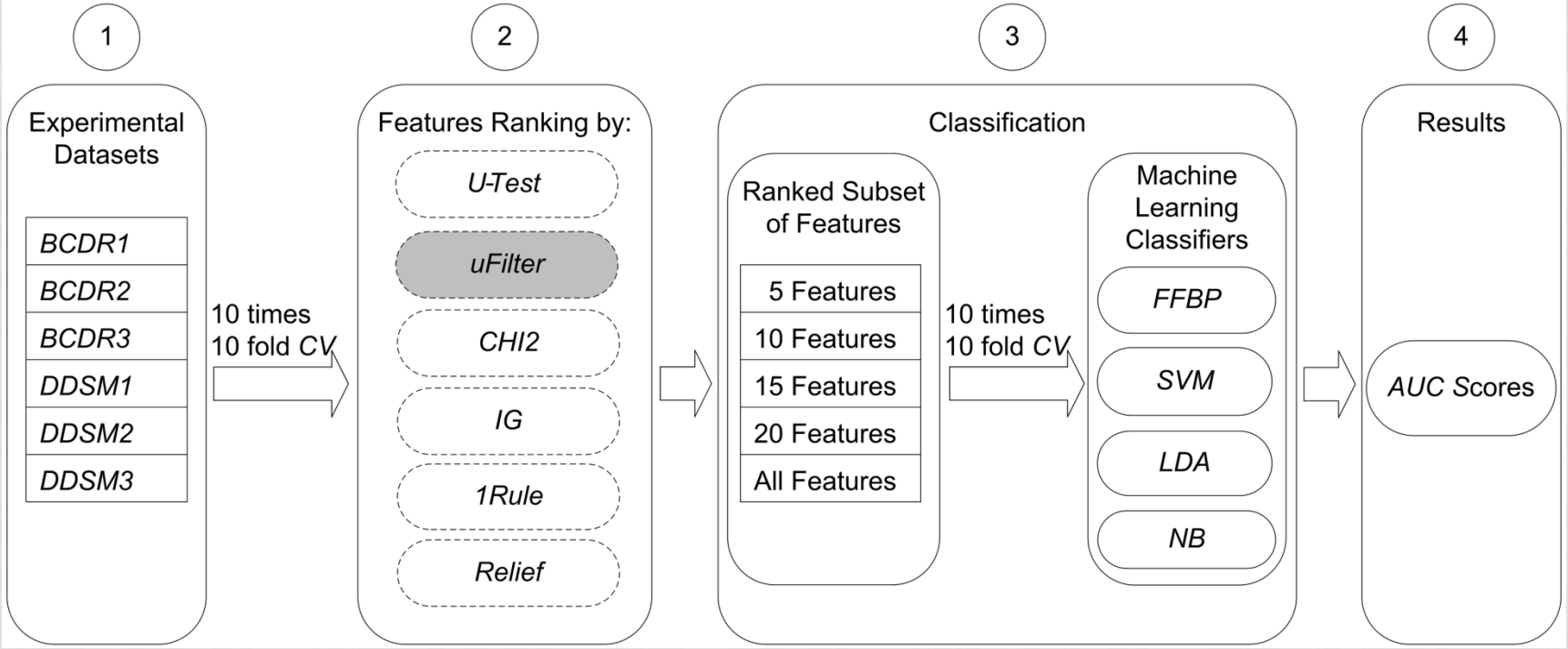
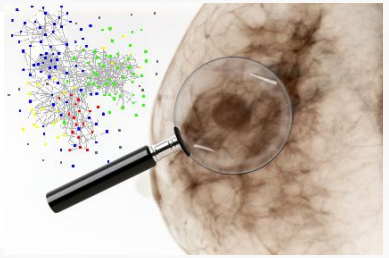
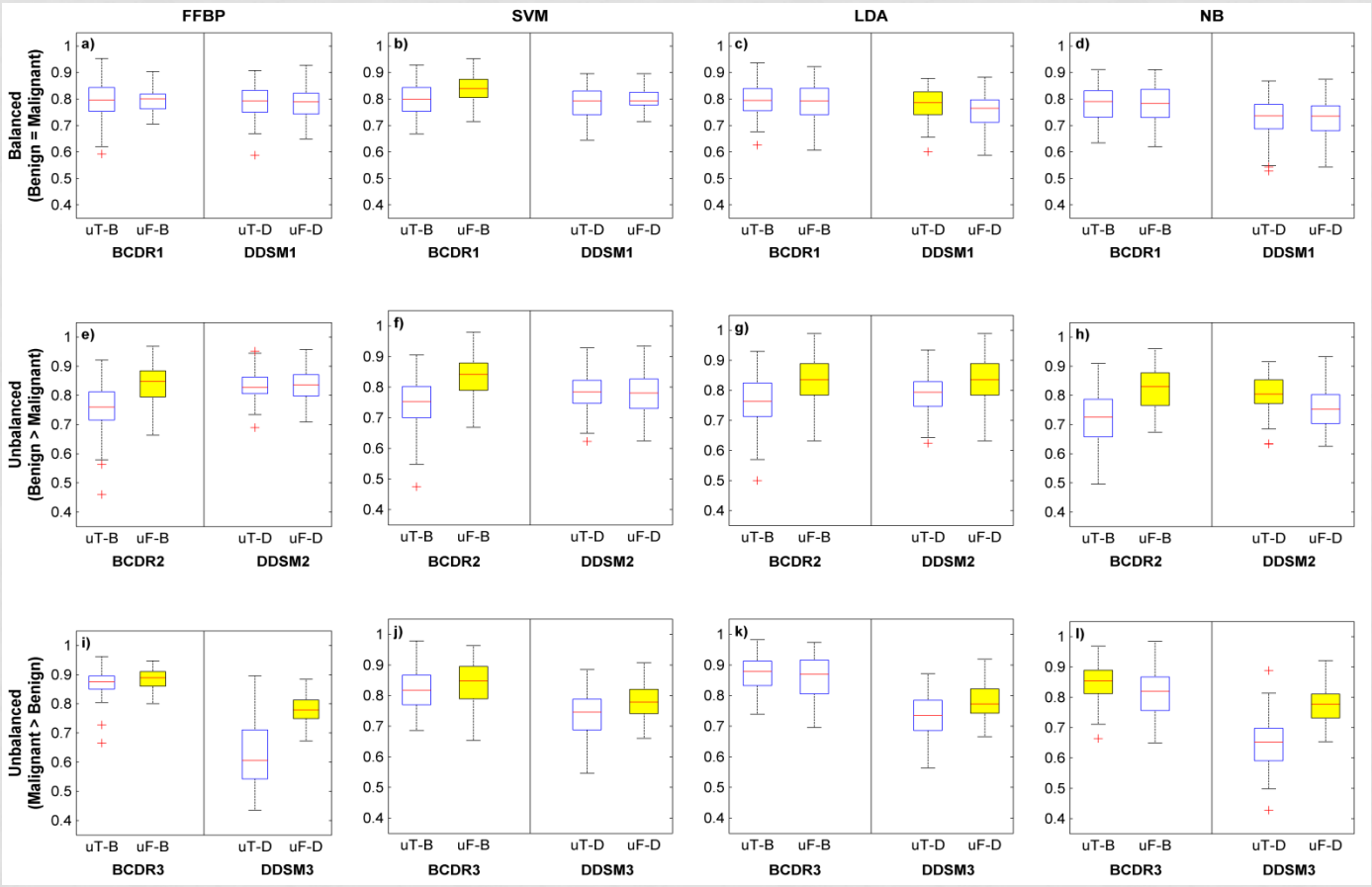


Fig. 2. Applied experimental workflow; CV means cross-validation.

In the last step of the experiment, we determined the feature relevance analysis using a two-step procedure involving (1) selecting the best subset of features for each dataset, and (2) performing a redundancy analysis based on the Pearson correlation, to determine and eliminate redundant features from relevant ones, and thus to produce the final subset of features.

RESULTS AND DISCUSSIONS

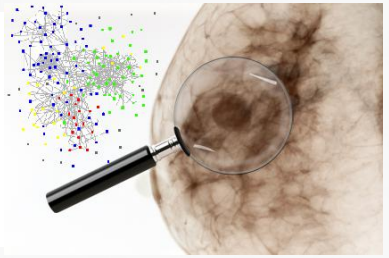


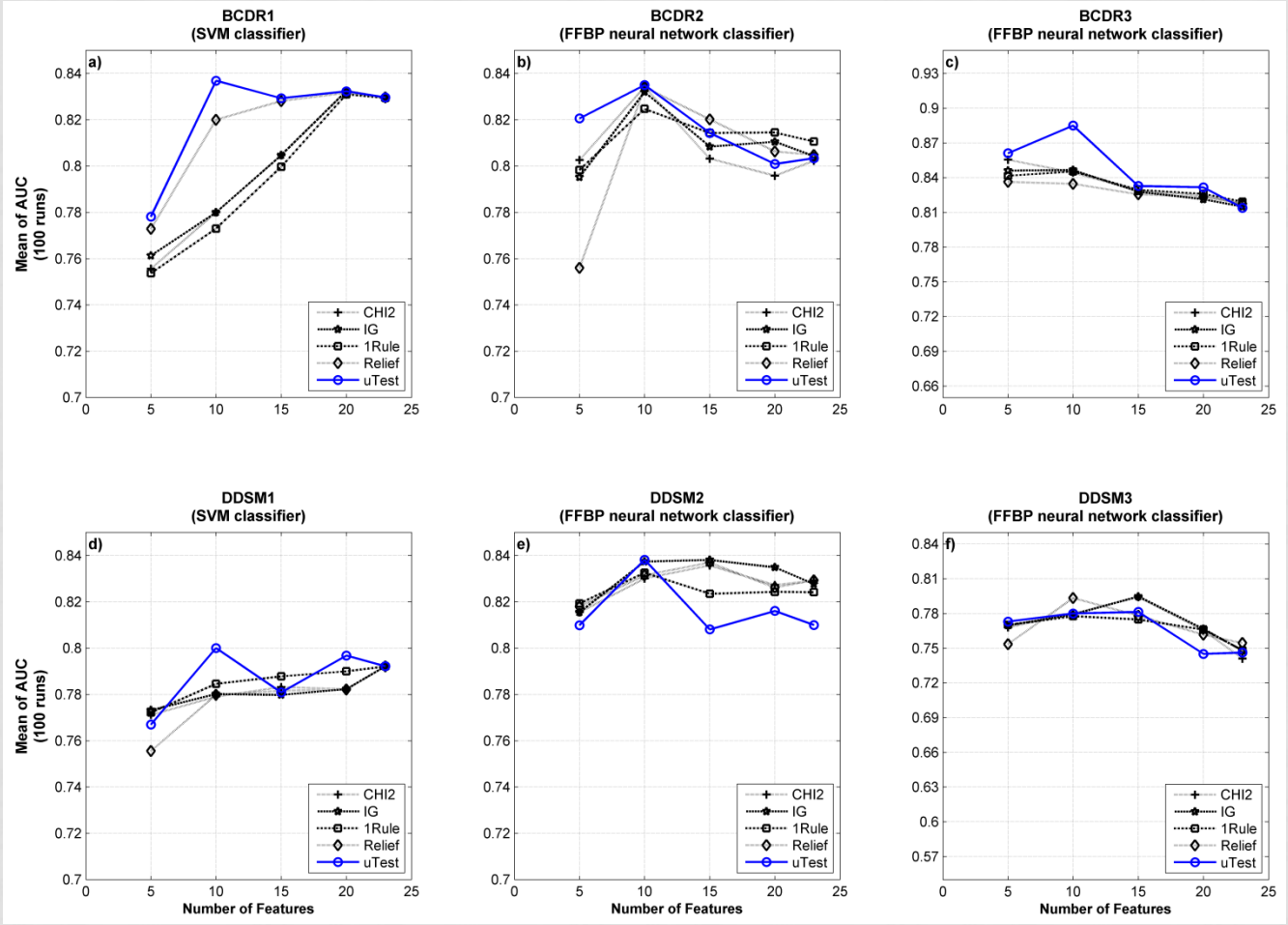


A total of 48 ranked subsets of image-based features were analyzed by feeding four machine learning classifiers and the straightforward statistical comparison based on the mean of AUC performances over 100 runs highlighted interesting results for balanced and unbalanced datasets (see Fig. 3).

Fig. 3. Head-to-head comparison between uFilter (uF) and U-test (uT) methods using the top 10 features of each ranking. Filled box represents significant difference ($p < 0.05$) in the AUC performance.

RESULTS AND DISCUSSIONS

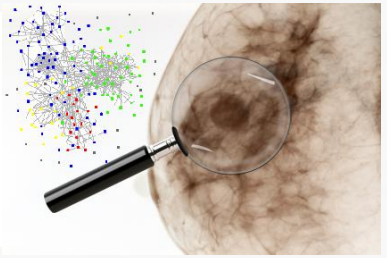




A total of 720 ranked subsets of image-based features were analyzed and the straightforward statistical comparison based on the mean of AUC performances over 100 runs highlighted interesting results for balanced and unbalanced datasets(see Fig. 3).

Fig. 4. Behavior of the best classification schemes when increasing the number of features on each dataset.

RESULTS AND DISCUSSIONS

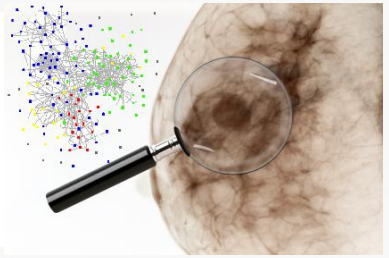


Feature relevance analysis

We achieved this goal using a two-step procedure involving:

1. Selecting the best subset of features for each dataset.
2. Performing the redundancy analysis based on the correlation of Pearson to determine and eliminate redundant features from relevant ones, and thus to produce a final optimal subset of features.

RESULTS AND DISCUSSIONS

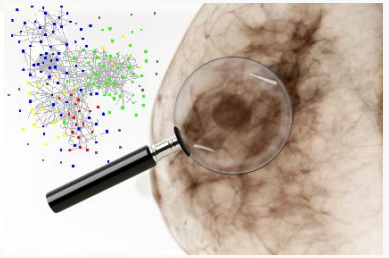


Dataset	Best subset of features	Redundant features	c-Pearson	p-Value (a=0.05)	Weakly relevant	Strongly relevant
BCDR1	f_4, f_{12}, f_{15}, f_{21}, f_7, f_{10}, f_3, f_6, f_{18}, f_8	$f_{21}=f_4$ $f_{10}=f_7,f_3$ $f_3=f_7$ $f_{18}=f_6$	0.79 0.96, -0.92 -0.84 -0.62	p<0.01	f_4, $f_{15}^{(+)}$, f_7, f_6, f_8	f_{12}
BCDR2	f_{14}, f_{22}, f_{21}, f_4, f_{12}, f_{15}, f_6, f_{13}, f_{11}, f_8	$f_{14}=f_{22},f_{13}$, f_{11} $f_{21}=f_4$ $f_{13}=f_{22}$ $f_8=f_6$	0.99, 0.56, 0.56 0.89 0.55 0.75	p<0.01	f_{22}, $f_{12}^{(+)}$, $f_{15}^{(+)}$, f_6, f_{11}	f_4
BCDR3	f_7, f_{10}, f_3, f_4, f_{12}, f_{18}, f_{15}, f_{22}, f_{19}, f_{13}	$f_{10}=f_7,f_3$, f_{22} $f_3=f_7,f_{22}$ $f_{18}=f_{12}$ $f_{13}=f_7,f_{10},f_3,f_{22}$	0.97,-0.94,0.56 -0.85,-0.62 -0.75 0.50, 0.57,-0.62, 0.99	p<0.01	f_7, $f_4^{(+)}$, f_{12}, $f_{15}^{(+)}$, f_{22}	f_{19}
DDSM1	f_9, f_{16}, f_{19}, f_{23}, f_4, f_{12}, f_{21}, f_6, f_{10}, f_{15}	$f_{23}=f_9,f_{16},f_{19}$ $f_{21}=f_4$ $f_6=f_4$, f_{15} $f_{15}=f_{12}$ $f_{16}=f_{19}$	0.85, 0.94, 0.94 0.93 0.56,-0.71 -0.79 0.99	p<0.01	$f_9^{(+)}$, f_4, f_{12}, $f_{10}^{(+)}$, f_{19}	-
DDSM2	f_7, f_{19}, f_{16}, f_{23}, f_9, f_3, f_1, f_5, f_{12}, f_8	$f_{23}=f_{19},f_{16},f_9,f_{12},f_8$ $f_9=f_{19},f_{16},f_8$ $f_{12}=f_9$	0.97,0.98,0.89,0.71,0.51 0.92,0.92,0.61 0.68	p<0.01	f_{19}, f_{16}, $f_3^{(+)}$, $f_1^{(+)}$, $f_5^{(+)}$, f_8	f_7
DDSM3	f_9, f_4, f_{21}, f_{23}, f_{16}, f_{10}, f_{19}, f_{12}, f_{18}, f_6	$f_{21}=f_9$ $f_{23}=f_9,f_{16},f_{19}$ $f_{16}=f_9,f_{19}$ $f_{12}=f_9,f_4,f_{18},f_6$	0.84 0.78,0.92,0.91 0.85,0.99 0.60,0.56,0.57,0.76	p<0.01	f_9, $f_{10}^{(+)}$, f_{19}, f_{18}, f_6	f_4

Table 1 Summary of the redundancy analysis.

$^{(+)}$Weakly relevant features but non-redundant; c-Pearson is the value of correlation of Pearson; p-Value means whether the correlation value is significantly different from zero (i.e. are correlated).

RESULTS AND DISCUSSIONS

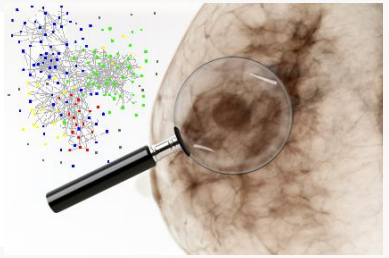


Dataset	Best subset of features	AUC	Weakly + Strongly	AUC	Wilcoxon (a=0.05)
BCDR1	f_4, f_{12}, f_{15}, f_{21}, f_7, f_{10}, f_3, f_6, f_{18}, f_8	0.839	f_4, $f_{15}^{(+)}$, f_7, f_6, f_8, f_{12}	0.8315	p=0.811
BCDR2	f_{14}, f_{22}, f_{21}, f_4, f_{12}, f_{15}, f_6, f_{13}, f_{11}, f_8	0.835	f_{22}, $f_{12}^{(+)}$, $f_{15}^{(+)}$, f_6, f_{11}, f_4	0.8413	p=0.841
BCDR3	f_7, f_{10}, f_3, f_4, f_{12}, f_{18}, f_{15}, f_{22}, f_{19}, f_{13}	0.885	f_7, $f_4^{(+)}$, f_{12}, $f_{15}^{(+)}$, f_{22}, f_{19}	0.8821	p=0.918
DDSM1	f_9, f_{16}, f_{19}, f_{23}, f_4, f_{12}, f_{21}, f_6, f_{10}, f_{15}	0.8004	$f_9^{(+)}$, f_4, f_{12}, $f_{10}^{(+)}$, f_{19}	0.8001	p=0.982
DDSM2	f_7, f_{19}, f_{16}, f_{23}, f_9, f_3, f_1, f_5, f_{12}, f_8	0.8382	f_{19}, f_{16}, $f_3^{(+)}$, $f_1^{(+)}$, $f_5^{(+)}$, f_8, f_7	0.8435	p=0.757
DDSM3	f_9, f_4, f_{21}, f_{23}, f_{16}, f_{10}, f_{19}, f_{12}, f_{18}, f_6	0.7806	f_9, $f_{10}^{(+)}$, f_{19}, f_{18}, f_6, f_4	0.7759	p=0.685

Table 2 AUC-based statistical comparison between the best and optimal subset of features.

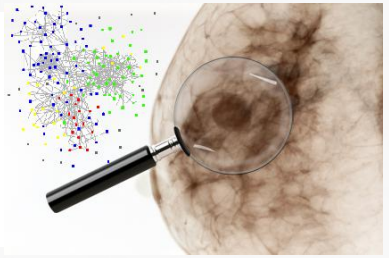
$^{(+)}$Weakly relevant features but non-redundant.

CONCLUSIONS



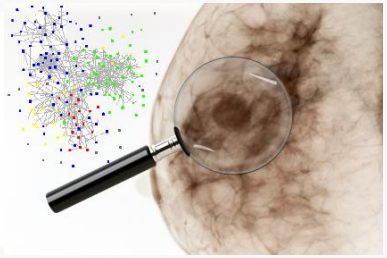
1. A head-to-head comparison proved that the uFilter method significantly outperformed the U-Test method for almost all of the classification schemes. It was superior in 50%; tied in a 37.5% and lost in a 12.5% of the 24 comparative scenarios.
2. Moreover, a global comparison against other four well known feature selection methods (CHI2 discretization, IG, 1Rule and Relief) demonstrated that uFilter statistically outperformed the remaining methods on several datasets (BCDR1, DDSM1 and BCDR3), and it was statistically similar on the BCDR2, DDSM2 and DDSM3 datasets while requiring less number of features.
3. The uFilter method revealed competitive and appealing cost-effectiveness results on selecting relevant features, as a support tool for breast cancer CADx methods especially in unbalanced datasets contexts.
4. Finally, the redundancy analysis as a complementary step to the uFilter method provided us an effective way for finding optimal subsets of features.

FUTURE WORK



Future work will be aimed to:

1. Increasing the number of features in benchmarking breast cancer datasets.
2. Exploring the performance of uFilter in other knowledge domains.
3. Extending uFilter allowing it to be used on multiclass classification problems.



- Thanks for your attention !!

Noel Pérez Pérez, Miguel A. Guevara López, Augusto Silva, Isabel Ramos, "Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography", Artificial Intelligence in Medicine, 2015, vol. 63, no. 1, pp. 19-31, http://dx.doi.org/10.1016/j.artmed.2014.12.004.