Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Bioinformatics for Translational Medicine

SVM breast cancer subtype classification on ArraryCGH data using different feature selection methods.

Myrthe van Baardwijk [1], Martin Banchero [1], Dennis Dekker [1] and Marina Diachenko [1]

[1] Master Bioinformatics and Systems Biology, Vrije Universiteit, Amsterdam, 1081 HV, the Netherlands.

Abstract

Motivation: Array-based comparative genomic hybridization (aCGH) is a commonly used tool to study copy number aberrations (CNAs) in cancer. The aCGH dataset obtained from different tumors and tumor subtypes are subject to supervised classification which can help to improve prognosis or diagnosis of cancer. Genomic intervals generated in these experiments may carry key information regarding transformations specific to a cancer subtype. However, some intervals can introduce noise, rendering the trained model prone to overfitting. Selection of a subset of discriminative features can therefore improve classification of aCGH data.
Results: In this study, we employ three different feature selection approaches for Support Vector Machine (SVM) breast cancer subtype classification. Each method, ANOVA, SVM-RFE, and L1 regularization, is benchmarked against a baseline classifier trained on all the features from the dataset. The performance of each method was found to be superior to the baseline. However, the results were inconclusive as to which of these was the best feature selection method with SVM.
Impact: Comparison of these feature selection method using SVM for this purpose was a novel approach. The research was relevant as no previous studies have shown which feature selection method is most suited for classification on aCGH data.

1 Introduction

According to the Global Cancer Observatory at the International Agency for Research on Cancer, breast cancer has the highest incidence compared to other cancer sites, with mortality rates being the second highest worldwide [12]. Among the three well-defined molecular breast cancer subtypes, Hormone Receptor-positive (HR+) remains the most prevalent whereas Triple Negative (TN) and Human Epidermal Growth factor Receptor 2 positive (HER2+) occur in 10-20% of the cases [18]. Moreover, each molecular subtype is treated with a different medical therapy it responds best to, with prognosis and survival depending on molecular features, clinical parameters, and the level of cell differentiation [4]. Therefore, early diagnosis and the accurate tumor subtype classification will promote well-timed and appropriate treatment and facilitate patient's survival.

Data-driven analytic studies complement breast cancer research in subtype-specific biomarker identification. As such, machine learning and data mining methods are effective tools for the detection of critical features from breast cancer datasets. One of the sources for these datasets is array-based comparative genomic hybridization (aCGH) which has been widely used to study cytogenetic profiles of breast cancer subtypes. In aCGH experiments, test and reference samples containing isolated genomic DNA with distinguishable fluorescent labels are hybridized to a microarray consisting of wells, each having DNA sequences or probes, complementary to a specific gene. Intensity log2-ratios of sample and reference DNA at a given location are used to infer DNA copy number abnormalities in the test sample genome relative to the reference [19] [25]. Normalized aCGH data which accounts for systematic variation of microarray experiments [26] is represented as a collection of status values (-1, 0, +1, and +2 for loss, normal, gain, and amplification, respectively) which are assigned to genomic intervals on chromosomes. One can think of such data as a high-dimensional feature space, with thousands of ordered genomic intervals on chromosomes as features and a relatively small number of samples. Therefore, an important processing task for this data is to determine a subset of discriminative features, while removing the noise that can obscure the true performance of the predictive model. These features may also help to uncover meaningful patterns in chromosomal aberrations, corresponding to specific cancerous transformations [22].

Feature selection along with cross-validation can substantially lessen the risk of overfitting and improve predictive accuracy.

Out of many machine learning algorithms, Support Vector Machine (SVM) has been widely employed in microarray data classification, and a lot of different feature selection methods have been adopted to deal with the large feature space mentioned above. For example, Rapaport et al. extended the fused Lasso algorithm for classification and feature selection of multi-class aCGH data from bladder and melanoma cancers to build a sparse linear SVM-based classifier which included prior biological knowledge about DNA copy number variations, accounting for the inherent correlation structure of the data [21]. Liu et al. developed a novel SVM-based method with a new non-linear kernel function which also accounted for the specificities of the aCGH data and showed a significant increase in the average accuracy over 12 datasets derived from the Progenetix database [14]. Among other SVM-based methods, Support Vector Machine Recursive Feature Elimination (SVM-RFE) used by Chai and Domeniconi [6] was tested on multi-class microarray gene expression data and showed the best performance in different kernel settings, which was robust in higher dimensions as compared to several correlation scores, Chi-squared, and Information Gain methods used for gene ranking feature selection.

In general, feature selection methods can be grouped into different categories, such as filter and wrapper methods [10]. Filter methods select features independently, by determining their relevance using a certain scoring metric, such as ANOVA F-score or Student's t-test. Wrapper methods use a machine learning approach to iteratively combine features in order to get the best performance. Besides feature selection, regularization can also prevent overfitting. It does so by adding a penalty to the loss function for using many features. Therefore, this also results in a model with a reduced number of features. In this study, we try to address the problem of multi-class classification and feature selection of aCGH data and answer the following question: which feature selection method, ANOVA, RFE or L1 regularization, in combination with SVM, has the best performance on classifying multiple breast cancer subtypes?

2 Methods

2.1 Study design

2.1.1 Cross-validation scheme and benchmarking

For this study, a breast cancer aCGH dataset from patients with three different cancer subtypes was provided to us. ANOVA, recursive feature elimination, and L1 regularization were implemented along with hyperparameter tuning for building an optimal Support Vector Machine classifier for one-versus-rest breast cancer subtype classification. Data exploration and classification were performed utilizing Python with scikit-learn, matplotlib, numpy, and pandas packages. Predictor training and performance validation were included in an extensive cross-validation (CV) scheme with outer validation and inner training loops (Fig. 1). Considering the size of our dataset, we employed 5-fold CV in the outer loop, which ensured a larger proportion of the data for validation, and 4-fold CV in the inner loop, which also ensured an adequate proportion of the data for predictor optimization. In this scheme, the validation set in each of the outer loop iterations served as unseen data to validate the performance of the predictor. Stratified k-fold cross-validation was used to have all three class labels present in balance in each validation set.

The performance of different SVM classifiers was evaluated by reporting the accuracy of the outer CV folds on their validation sets. The outer CV was repeated 15 times, and the mean accuracy together with the standard error was calculated for each method. The performance of the classifier with each feature selection method was benchmarked against a baseline SVM classifier to which no feature selection method was applied. Such a classifier was suitable and allowed to infer the meaning of the results obtained by each feature selection method. Hyperparameters for the baseline were also optimized by using the GridSearchCV from scikit-learn and the cross-validation scheme described above. P-values of unpaired 2-sample Wilcoxon-Mann-Whitney test (wilcox.test in R) between each method and the baseline as well as between each pair of methods were reported. Multiple testing (n = 6) was corrected for using the Bonferroni correction.

In the end, the feature selection method to be used with SVM for the final class contest predictions was chosen based on the best average validation performance in the CV scheme described above. This feature selection method then was trained on the entire training dataset with 4-fold cross-validation to select for optimal features and hyperparameters to be used in the final model to estimate its generalization power on the independent test set.

2.1.2 Data preparation

The dataset used to perform this study contained 100 breast cancer samples with the three subtypes mentioned above, HER2 positive (HER2+), Triple negative (TN; ER-, PR- and HER2-) and Hormone receptor positive (HR+; ER+ and/or PR+, and HER2-), and was provided with the data pre-processed.

To obtain this dataset, the samples were analyzed on an aCGH platform containing 244000 probes per array to identify different chromosomal aberrations. The samples were compared against a reference sample to determine the log2-ratios of copy number aberration (CNA) per bin (analyzed region). In order to reduce the complexity of the data, the ratios were converted to loss, normal, gain or amplification of a chromosomal region. In the dataset, those rations were translated to calls, which are -1, 0, 1 or 2, respectively. While this reduces the complexity, the amount of features is still equal to the number of bins in the original data. To reduce this number, neighboring bins with the same CNA are taken together, creating regions with similar CNA for all samples. Thus, the final dataset that we worked with comprised 2834 regions and 100 observations. For this dataset, the presence of correlations between the neighboring regions mentioned earlier was explored and confirmed by the Spearman's correlation method (figures not included in the paper).

As part of pre-processing, the labels (the subtypes of cancer) were transformed from categorical to numerical values using the function LabelEncoder() from scikit-learn. The labels then were transformed as follows: HER2+ to 0, HR+ to 1 and TN to 2. This step was required due to the algorithm of machine learning used which receives numerical data as input.

2.1.3 Support Vector Machine

For testing the differences between the feature selection methods the same classifier method for all feature selection methods was used. In this study, for each feature selection method, a Support Vector Machine (SVM) supervised learning model was trained. An SVM finds an optimal decision boundary that maximizes the margin separating the classes and can be applied to linear as well as non-linear classification problems. As this research is based on classifying three classes, the standard two-class SVM cannot be used. This multi-class problem is approached by employing a one-versus-one method, in which a voting system is employed. The SVM assigns a classification based on two classes, a vote, and then repeats this for all combinations of classes. In the end, the instance is assigned the most voted class. In case of a tie, the method picks the classes based on the highest confidence level calculated by the binary classifier.

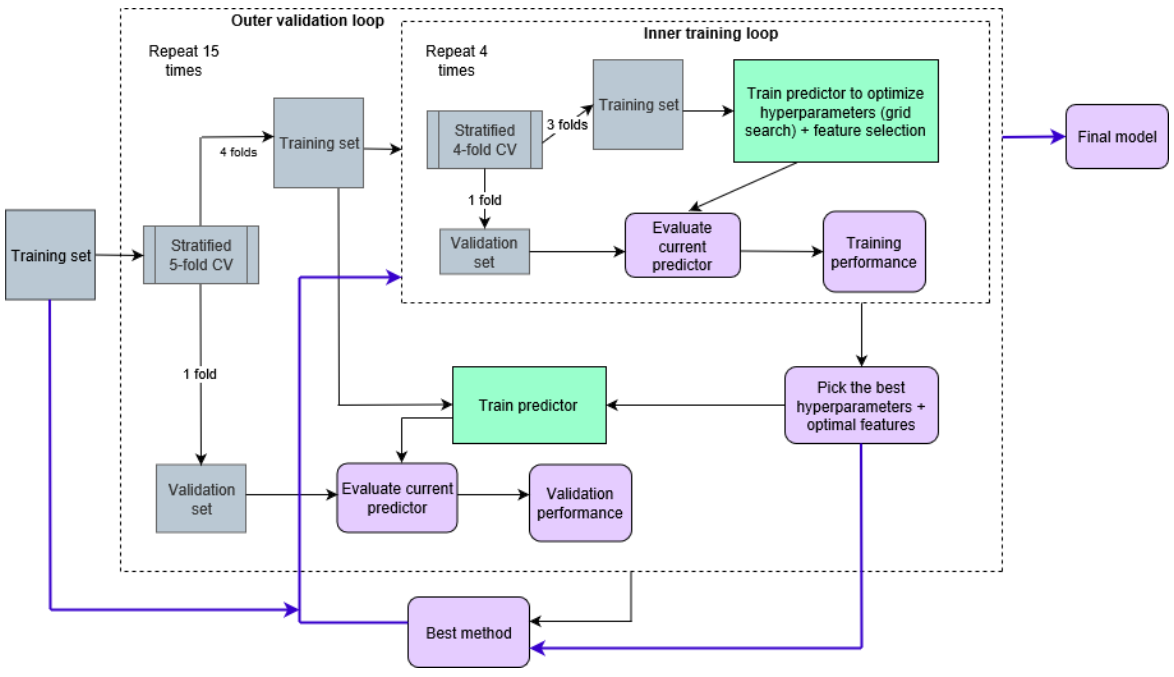

Fig. 1: Cross-validation scheme designed for training and validating the performance of the predictors. Black arrows show the initial workflow for training and validating the performance for each feature selection method. Blue arrows indicate the workflow after the final feature selection method has been chosen to be trained for the class contest.

2.1.4 Hyperparameter selection

For training an SVM classifier, different parameters needed to be tuned in order to get the best performance for the current dataset. Important parameters were the kernel type and the penalty parameter C. When the radial basis function (RBF), polynomial or sigmoid kernel were used, the kernel coefficient also needed to be adjusted, as well as the degree of this kernel in case of the polynomial function. The gamma parameter was set on 'auto' in order to save computation time. Different search methods can be applied to find the best hyperparameters for the classifier, such as Random Search (RS) or Grid Search (GS). In this study, Grid Search was used to select the best hyperparameters in the inner CV loop. All kernel types were evaluated for ANOVA and L1 regularization. For the SVM-RFE method, only the linear kernel could be used. Therefore, kernel selection and kernel specific parameters were excluded from the Grid Search of SVM-RFE. For C, the penalty parameter of the error term, the values 0.1, 1, 10 and 100 are explored. For L1, also the penalty parameter λ was taken into the hyperparameter selection for optimization

2.2 ANOVA feature selection

ANOVA feature selection is a univariate filter method. The ANOVA F-value is computed for each of the possible features. This F-value, also called the Fisher ratio, is the mean sum of squares of the differences between groups ($MS_{between}$) divided by the mean sum of squares of the differences within groups (MS_{within}) [24]. $MS_{between}$ is calculated as follows:

$$MS_{between} = \frac{K \sum_{(} kx_{i.} - x_{..})^2}{df_{between}} \tag{1}$$

In which K denotes the number of classes, $x_{i.}$ the sample mean of the i-th class and $x_{..}$ the overall mean of the dataset. MS_{within} is calculated as follows:

$$MS_{within} = \frac{\sum_i \sum_{(} kx_{ik} - x_{i.})^2}{df_{within}} \tag{2}$$

In which x_{ik} is the k-th observation in the i-th class. Features that differ the most between groups will result in the highest F-values. Using these F-values, different subsets of features are selected with the SelectKBest method from scikit-learn. The value k, the number of features to use, is one of the hyperparameters to tune in the inner CV rounds and is set to vary between 5 and 80 with steps of 5. In the outer CV rounds, the training sets are used to generate a model using the best set of features and hyperparameters from the inner CV round. If in the inner CV round different values for k result in the same score, the lowest value for k is selected.

2.3 RFE-SVM

Recursive Feature Elimination-Support Vector Machine (RFE-SVM) is one of the most popular wrapper method used for feature selection. This method uses the performance of one particular machine learning algorithm to evaluate the subset of selected features. In this case, the Support Vector Classifier (SVC) was employed. RFE uses the weights of the classifier to generate a ranking [9]. The features with the highest ranking are those that are eliminated last. However, this does not mean that these features are the most relevant by themselves; they are important within the selected subset of features. The pseudocode for the RFE-SVM is shown in Listing 1, following the scheme used by Sanz et al. [23].

Starting with all the features, the SVC is generated. The feature with the lowest weight is left out, and this procedure is iterated. This process is repeated until all the features are eliminated. In the recursive step, a different number of features can be chosen to be left out. In this study, 1 feature was eliminated in each step. This is more computationally expensive but also more accurate [9]. The ranking of features follows

the order in which the features were eliminated. The optimal number of features, k, is then based on the best score for a certain subset of features and is optimized during the Grid Search. The same process is repeated for all k-folds.

Data: k^* features and n instances in dataset.
Output: List of features in order of importance.
Hyperparameter selection for SVM model;
Train SVM model;
Let:
  $k \leftarrow k^*$;
  while $k \geq 0$ **;**
    $SVM_k \leftarrow$ SVM using hyperparameters for k features
    $w_k \leftarrow$ weight vector given by $SVM_k(w_{k1}, ..., w_{kk})$;
    $rank.criteria \leftarrow$ following order $(w_{k1^2}, ..., w_{kk^2})$;
    $low.rank.criteria \leftarrow$ feature in the last position;
    in $rank.criteria$;
    Remove low.rank.criteria from **Data;**;
    $Rank_k \leftarrow low.rank.criteria$;
    k$\leftarrow$ k$-$1;
  end;
  $Rank_1 \leftarrow$ feature in **Data;** $\notin (Rank_2, ..., Rank_k*)$;
  $return$ $(Rank_1$,...,$Rank_k*)$

Listing 1: Example of algorithm Recursive Feature Elimination SVM (RFE SVM)

2.4 L1 Regularization

L1 regularization is a widely used feature selection method. It adds weight to features, so more important features will get a higher weight, while less important features will get a small weight assigned. This helps to reduce overfitting of the model, as noise in the data is filtered out. L1 feature selection results in sparse models, meaning that a feature can be assigned a weight of 0 and will, therefore, not contribute to the model. The regularization is displayed in the Equation 3, where the regularization term λ controls the weight added to the feature. The loss function calculates the sum of squared differences between the actual values y_i and the predictions x_i, and then penalizes the weights by a factor λ. The penalty is applied to the sum of the weights θ_i.

$$L(x, y) = \sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n}|\theta_i| \quad (3)$$

Using L1 regularization, the most important features can be extracted and used in the classification method to predict class labels for new data. In this paper, we used the scikit-learn package feature_selection to extract the features for a linear model penalized with L1 regularization.

2.5 Gene Ontology analysis

For each feature selection method, the 10 highest weighing features of the highest scoring method were further analyzed. Biomart, a data-mining tool from Ensemble, was used to report the genes within these features from their chromosome number, together with start and end regions of the array. For these genes, gene ontologies were collected from the GO database to estimate their relevance to our breast-cancer classification problem. If information from gene ontologies was limited, other databases like OMIM were consulted.

3 Results

3.1 Performance of the models

Mean accuracy and standard error of the SVM classifiers in the outer CV rounds are shown in Fig. 2. From this figure, it can be seen that there are no significant differences between the methods; this was assessed using a nonparametric unpaired 2-sample Wilcoxon-Mann-Whitney test (with a "two.sided" alternative hypothesis), which showed adjusted p-values of 1 between each pair of the methods (RFE and L1, ANOVA and L1, and ANOVA and RFE). However, all methods performed significantly better than the baseline.

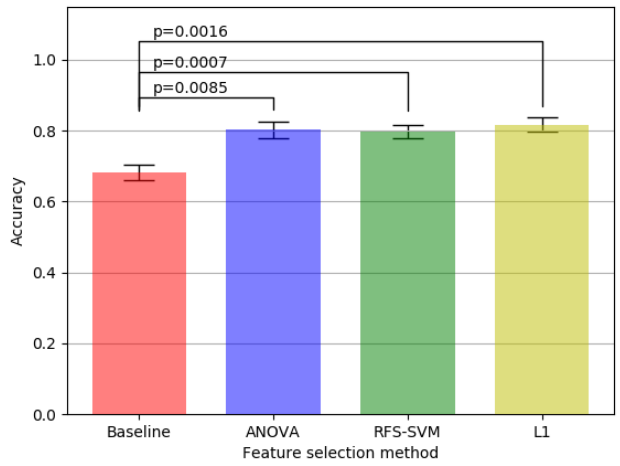

Fig. 2: Mean accuracy and standard error of the 15 classifiers of the outer CV rounds for each method. Standard error of the baseline classifier does not overlap with any of the other methods, implying there might be a significant improvement from feature selection. The standard errors of all feature selection methods overlap, suggesting that there is no significant difference between the methods. Accuracy is plotted on the y-axis; feature selection methods and the baseline are shown on the x-axis. P-values obtained on an unpaired 2-sample Wilcoxon-Mann-Whitney test for each method with the baseline are shown on the top; alternative hypotheses were specified with a character string "greater" for wilcox.test in R.

Overall, the results obtained in this paper are comparable to the ones found in the literature. For example, in the study conducted by Rapaport et al. [21], L1-SVM on aCGH data from bladder and melanoma tumors showed 88% and 76% accuracy rates, respectively. In the paper by Ramaswamy et al. [20], RFE-SVM for multi-class cancer classification of tumor gene expression data yielded 78% prediction accuracy. ANOVA-SVM was found to be used by Arowolo et al. for binary colon cancer classification of microarray gene expression data, improving the overall accuracy of SVM from 79% to 86.7% [3].

3.2 Gene ontology analysis

In ANOVA, six of the best scoring features originated from chromosome 17 and three from chromosome 12. Besides that, features derived from the same chromosome were often found in the neighboring arrays. In CGH data, neighboring arrays are often strongly correlated [14], and since univariate feature selection methods consider each feature independently, there is a high probability of choosing redundant features. This is also reflected in the features chosen by ANOVA. In RFE-SVM, the GO analysis was performed using the optimal subset of 22 features. Genes of interest that have been found were CSNK2A1 and LAMA5, which are implicated in phosphorylation of PTEN and tumor progression, respectively [1, 8]. For L1, the 10 highest features (out of 57) were mostly related to cancer gene transcription and mammary development. Most of these regions were found in chromosomes 17, 16 and 12, similar to ANOVA.

Interesting genes within the features selected by ANOVA, RFE, and L1 are listed in Table 1. Some of these genes are well-known to be implicated

Table 1. Genes of interest found within the feature regions used by the SVM classifiers, together with their most important ontologies.

Method	Chr	Gene	Ontologies
ANOVA, RFE, L1	17	ERBB2	MAPK cascade, protein phosphorylation, tyrosine kinase activity.
ANOVA	17	GRB7	RNA binding.
ANOVA	17	TCAP	Somitogenesis, muscle contraction.
ANOVA	17	MAP3K14	MAPK cascade activation.
ANOVA	17	BRCA1	Ubiquitin ligase, double-strand break repair.
ANVOA	17	PSME3	MAPK cascade, protein polyubiquitination, p53 binding.
RFE	20	CSNK2A1	Casein kinase II complex.
RFE	20	LAMA5	Cell migration, differentiation, signaling, metastasis.
L1	12	E2F7	Transcription activity.
L1	12	PPFIA2	Mammary gland development, protein binding
L1	12	PAWR	Tumor suppressor, apoptosis
L1	16	CALB2	calcium ion binding, biomarker different cancers.

in breast cancer (BRCA1, ERBB2) [17, 16]; they are reported to be amplified in previous aCGH studies (GRB7, TCAP) [7] or found within the same ontology. One remarkable finding of the gene ontology analysis was three genes of the MAPK cascade, namely ERBB2, MAP3K14, and PSME3. ERBB2, also called HER2, activates this signaling pathway that is used to transfer signals from outside of the cell to the nucleus. In many cancers, defects in this pathway result in the uncontrolled cell growth [11]. Therefore, genes of this pathway could serve as potential biomarkers.

3.3 Single biomarker for classification

The single best region found by ANOVA (F-score 155.29), which was also top-selected by L1 and RFE, was the one located on chromosome 17 with the start and end positions of 35076296 and 35282086, respectively. As mentioned above, this interval includes 6 genes one of which was found to be involved in the MAPK pathway (ERBB2). To assess the discriminative power of this region in SVM, the baseline SVM classifier was trained using the original dataset containing only that one region in the same cross-validation scheme. This resulted in the mean of 0.66 and the standard deviation of 0.035 over 15 outer loop iterations (standard error of 0.009). Thus, the performance of the baseline SVM with that single feature was quite similar to the performance of the baseline SVM with all the features (mean 0.68 and standard deviation 0.08).

ERBB2/HER2 is known to be used as a biomarker and target of therapy for patients with HER2 breast cancer subtype. This could also explain why the performance of the baseline based exclusively on this region was comparable to the performance of the baseline on all the regions. In essence, this region makes the model very sensitive to detecting the HER2 subtype. It is expected that for a balanced dataset, which was the case in this study, such a model would be able to predict all the HER2 samples correctly (giving 0.33 accuracy) but predict randomly for the ER+ and Triple Negative samples (giving 0.165 + 0.165), resulting in an accuracy rate of 0.66.

Interestingly, the same region was demarcated from the rest upon principal component analysis (PCA) which was performed separately for each subtype (Fig. 3). Genomic intervals were considered as samples and normalized counts of losses, normal, gains, and amplifications per interval over all the samples belonging to the same subtype as features (see the full description in the caption). With these results, it may also be interesting to see how the exclusion of that genomic region from the dataset influences classification and feature selection by the methods described above. This will be elaborated on in the discussion section.

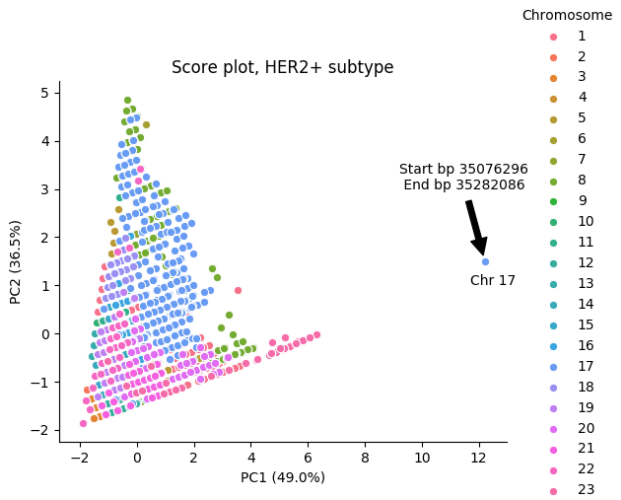

Fig. 3: Principal component analysis (PCA) score plot for the first two PCs performed for HER2+ subtype. PC 1 and PC 2 with their explained variances in parentheses are shown on the x-axis and y-axis, respectively. Each colored dot is a genomic region. To obtain this plot, the original dataset was grouped by the subtype, and the number of losses (-1), normal (0), gains (1), and amplifications (2) were calculated for each of the 2834 genomic intervals over all the samples for that subtype. This resulted in 3 separate data frames (one for each subtype), with genomic regions as rows and -1, 0, 1, and 2 as columns. Each column was first normalized by the total count for that column, and then mean centering and standard deviation scaling was performed prior to the PCA decomposition. The genomic region containing the ERBB2 gene is annotated with the black arrow. In addition, the spatial correlation structure between the genomic intervals can be noticed. Score plots for HR+ and TN subtypes are not shown.

4 Discussion

In this paper, we have compared three different feature selection methods used in combination with SVM for multi-class breast cancer subtype classification of pre-processed aCGH data against a baseline SVM classifier. In a cross-validation setting, all methods, ANOVA, RFE, and L1-regularization, were considerably outperforming the baseline SVM, showing an increase in the mean performance accuracy by up to 10-11%. Although the best mean performance was achieved by the L1-SVM model (81.72% accuracy), which was subsequently trained for building the final predictor, all three methods can be said to have performed on the same level. This introduced some ambiguity into the determination of the best feature selection method with SVM, which was the subject of our research. It is clear that this could not be decided based on those results alone.

We also assessed the biological relevance of the genomic intervals found by these methods. Most of the top-ranked intervals contained genes that are reasonably related to breast cancer. The methods differed in how importance was assigned to the features, nonetheless we expected to find an

overlap within the top features selected by all three methods. As such, the region containing the ERBB2 gene was top selected by each method, and it showed to be the most important feature for discriminating the HER2+ subtype.

Using the baseline classifier we obtained an accuracy of 0.66, only using this region. It would be interesting to see whether this feature could classify all the HER2+ correctly. If this is true, this region could be used to filter out the HER2+ samples and train the data to classify between the TN and the HR+ classes. This could lead to a higher accuracy of the models, as then the features for which HER2+ was similar to either TN or HR+ could be more discriminative.

It should be noted that several limitations of our study may have influenced the results and interpretations of the findings. From the methodological point of view, one of the limitations was imposed by the aCGH data used for this research. Specifically, the representation of the data as a collection of statuses, which are inferred from log2 intensity ratios, might result in information loss. This is especially the case if contamination with healthy tissue has occurred or if heterogeneity was present. According to Rapaport et al. [21], using log2-ratios overcomes this difficulty. Also, as carefully noticed by the authors, the transformation of ratios into statuses could affect the possible subtle signals which might have the power to distinguish between the cancer types. Another challenge hidden in the data was its inherent correlation structure mentioned in the introduction. Neighboring genomic intervals are highly correlated, and the weights of the intervals cannot be interpreted as direct individual interactions of these intervals with classification [21] [14].

The relatively small size of the dataset and the implemented cross-validation scheme could have also introduced bias into the results. In order to determine the best feature selection method, the data was split into 5 folds (see cross-validation scheme in Fig. 1). Next, the hyperparameters were selected by 4-fold cross-validation, resulting in only 60 samples per hyperparameter selection iteration to train on, and 20 samples for both of the validation test sets (inner and outer loop validation). While SVMs work generally well in 'large n small p' cases, the number of samples in this study was still quite low.

The SVM showed to be robust for high-dimensional data [2]. This makes this type of algorithm suitable to work with biological data, genomic data in particular. One disadvantage is the potential overfitting during the model selection, this during the hyperparameter selection. [5]. In addition, as stated by Liu et al. [14], the linear kernel might not be the most appropriate method to use with RFE. Furthermore, the implementation of a function to perform RFE with a non-linear kernel was not possible within a reasonable time, as a higher level of computational power was required. Nonetheless, the accuracy of the classifier using RFE-SVM could be improved using a non-linear kernel.

Gene ontology analysis was carried out on the ten best scoring features for each method, performing GO database searches manually. However, in most other studies, gene ontology analysis is performed on all features, using automated software like GOTM [27]. Therefore, genes of interest might be missed by our approach. Besides that, features that are found in multiple methods might also be missed when not found within the top ten. However, due to lack of experience, this was not explored. Apart from that, both ANOVA and RFE found the same genomic intervals that contained no genes. Recently, it has become more apparent that non-coding regions can also play an important role in cancer [13]. Therefore, these regions might be overlooked by gene ontology analysis, as this focuses on protein-coding sequences.

In the current study, features were selected based on a scoring metric (ANOVA), by recursively eliminating or by adding regularization. However, performance might be improved by adding prior knowledge about the genes related to these breast cancer subtypes. For example, Meric-Bernstam et al. reported that alterations in TP53 were found in most patients with TN, while only in a minority of patients with HR+ [15]. Using this type of expert knowledge to select important genomic regions might result in better feature subsets. However, due to lack of time, this was not explored. Moreover, classifying solely on features already known to be implicated in breast cancer subtypes would not result in the discovery of new biomarkers.

5 Conclusion

This study assessed the usage of ANOVA feature selection, recursive feature elimination, and L1 regularization, in combination with SVM, for classification of breast cancer subtypes using aCGH data. We found that all feature selection methods performed similarly on the current dataset. A baseline classifier was made using all features to compare the effect of feature selection to no feature selection on the accuracy. All methods performed significantly better than the baseline classifier. Therefore, ANOVA, RFE, and L1 regularization are all suited to increase classification performance. However, the current study is inconclusive as to which of these is the best method.

References

[1]I. U. Ali, L. M. Schriml, and M. Dean. Mutational Spectra of PTEN/MMAC1 Gene: a Tumor Suppressor With Lipid Phosphatase Activity. *JNCI Journal of the National Cancer Institute*, 91(22):1922–1932, nov 1999.

[2]C. F. Aliferis, D. Hardin, and P. P. Massion. Machine Learning Models For Lung Cancer Classification Using Array Comparative Genomic Hybridization. Technical report.

[3]M. O. Arowolo, S. O. Abdulsalam, Y. K. Saheed, and M. D. Salawu. A Feature Selection Based on One-Way-Anova for Microarray Data Classification. *Al-Hikmah Journal of Pure Applied Sciences*, 3:30–35, 2016.

[4]A. Bergamaschi, Y. Kim, P. Wang, T. SA¸rlie, T. Hernandez-Boussard, P. Lonning, R. Tibshirani, A. BÃ¸rresen-Dale, and J. Pollack. Distinct patterns of dna copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer*, 45(11):1033–1040, Nov 2006.

[5]G. C. Cawley and N. L. C. Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Technical report, 2010.

[6]H. Chai and C. Domeniconi. An evaluation of gene selection methods for multi-class microarray data classification. *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics,*, pages 3–10, 2004.

[7]S. Chin, A. Teschendorff, J. Marioni, Y. Wang, N. Barbosa-Morais, N. Thorne, J. Costa, S. Pinder, M. van de Wiel, A. Green, I. Ellis, P. Porter, S. Tavare, J. Brenton, B. Ylstra, and C. Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, 8(10):R215, 2007.

[8]A. Gordon-Weeks, S. Y. Lim, A. Yuzhalin, S. Lucotti, J. A. F. Vermeer, K. Jones, J. Chen, and R. J. Muschel. Tumour-Derived Laminin α5 (LAMA5) Promotes Colorectal Liver Metastasis Growth, Branching Angiogenesis and Notch Pathway Inhibition. *Cancers*, 11(5):630, may 2019.

[9]I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/3):389–422, 2002.

[10]A. Haury, P. Gestraud, and J. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12):e28210, 2011.

[11]R. Hilger, M. Scheulen, and D. Strumberg. The Ras-Raf-MEK-ERK pathway in the treatment of cancer. *Onkologie*, 25(6):743193, Dec 2002.

[12]G. C. O. IARC. Estimated age-standardized incidence and mortality rates (world) in 2018, worldwide, both sexes, all ages, 2019.

[13]E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, 17(2):93–108, Feb 2016. [DOI:10.1038/nrg.2015.17] [PubMed:25954001].

[14]J. Liu, S. Ranka, and T. Kahveci. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, 24(13):i86–95, Jul 2008.

[15]F. Meric-Bernstam, G. Frampton, J. Ferrer-Lozano, R. Yelensky, J. Perez-Fidalgo, Y. Wang, G. Palmer, J. Ross, V. Miller, X. Su, P. Eroles, J. Barrera, O. Burgues, A. Lluch, X. Zheng, A. Sahin, P. Stephens, G. Mills, M. Cronin,

and A. Gonzalez-Angulo. Concordance of genomic alterations between primary and recurrent breast cancer. *Mol. Cancer Ther.*, 13(5):1382–1389, May 2014.

[16]Z. Mitri, T. Constantine, and R. O'Regan. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother Res Pract*, 2012:743193, 2012.

[17]P. O'Donovan and D. Livingston. BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis*, 31(6):961–967, Jun 2010.

[18]S. G. K. Organization. Molecular subtypes of breast cancer, 2019.

[19]D. Pinkel and D. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37:S11–17, 2005.

[20]S. Ramaswamy, P. Tamayo, R. R., S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, L. M., E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26):15149–54, December 2001.

[21]F. Rapaport, E. Barillot, and V. J-P. Classification of arraycgh data using fused svm. *Bioinformatics*, 24(13):i375âŁ“i382, Jul 2008.

[22]M. Renan. How many mutations are required for tumorigenesis? implications from human cancer data. *Mol Carcinog*, 7(3):139–146, 1993.

[23]H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, 19(1):432, dec 2018.

[24]L. Stahle and S. Wold. Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*, 6(4):259–272, Nov 1989.

[25]M. van de Wiel, F. Picard, W. van Wieringen, and B. Ylstra. Preprocessing and downstream analysis of microarray dna copy number profiles. *Brief Bioinform*, 12(1):10–21, 2011.

[26]Y. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.

[27]B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16, Feb 2004.

6 Participation

Table 2. Participation of the different group members per task

Task	Dennis	Marina	Martin	Myrthe
Writing	20%	30%	20%	30%
Coding	28%	16%	28%	28%
Literature scanning	10%	40%	10%	20%