Vrije Universiteit Amsterdam
Faculty of Science

Bioinformatics for Translation Medicine - CATS assignment

Classifying breast cancer subgroups with chromosomal aberration patterns

Will Harley, Aamir Hasan, Saarika Prathivadi Bhayankaram, Koen Rademaker and Joris Visser

Abstract

Motivation: Breast cancer can be divided into at least three distinct molecular subgroups for which prognosis and treatment outcomes differ significantly. Accurate subgroup diagnosis is therefore a prerequisite of effective breast cancer treatment. Previous research has shown that chromosomal copy number aberrations (CNAs) vary between molecular subgroups. Here we build and evaluate the performance of machine learning classification models using CNA data from 100 breast tumours labelled as one of three subgroups: HER2+, HR+ and Triple Negative. We evaluate two feature selection methods and three classification methods.

Results: We find a combination of Boruta for feature selection and a neural network classifier to be the best performing model with an overall accuracy of 0.895, and find 38 genomic regions to be important for classification between the three subgroups. In these regions we find several genes that have been found to be involved in the formation of tumours of specific subgroups and identify the HER2+ amplicon that is specific to HER2+ tumour formation and proliferation. The use of accurate machine learning classifiers to predict breast cancer subgroups from CNA data may lead to faster and more accurate diagnosis and better treatment outcomes.

Availability: Scripts and data are available at https://github.com/krademaker/CATS

1 Introduction

Chromosomal instability is one of the hallmarks of cancer. Chromosomal aberrations affect transcriptional activity and thus greatly affect the diagnosis, progression, and prediction of response to treatment in different tumors (Kloosterman and Hochstenbach, 2014). Cancer is classified in molecular distinct subgroups that differ significantly in prognosis and patient survival rates (Van't Veer *et al.*, 2002) and evidence has been found that patterns of chromosomal copy number aberrations (CNAs) differ clearly between subgroups, for example, in colorectal cancer (Van Den Broek *et al.*, 2015). These so-called *molecular signatures* of chromosomal copy number aberrations could potentially be used to build classifiers to improve diagnosis and thus treatment of cancer patients.

In the first instance, microarray gene expression profiles were used as a tool to identify subpopulations and to perform diagnostic and prognostic predictions in the field of cancer (Van't Veer *et al.*, 2002). However, the introduction of array comparative genomic hybridization (aCGH) opened doors for the development of DNA-based diagnostic and prognostic predictors by investigating DNA copy number changes on a genome-wide scale (Redon *et al.*, 2009). Methodologically, genomic experimental DNA and reference DNA are both hybridized on an array containing genomic sequences of the corresponding species. In the context of cancer, DNA copy number profiles are subjected to several pre-processing methods to improve further biological analysis regarding tumor progression, survival and treatment outcome. Pre-processing involves quality measures like median absolute deviation (MAD) for the exclusion of outliers; removal of artifacts caused by technical and biological parameters (e.g. density of G and C nucleotides, DNA quantity and quality); calling, which assigns discrete states ('loss'; < 2 copies, 'gain'; 2 - 4 copies or amplification; > 4 copies) to the regions; and segmentation, which clusters the chromosomal regions together, however assuming a homogeneous relationship (van de Wiel *et al.*, 2011).

From a machine learning point of view, these molecular signatures of cancer patients face the problem that there are many more features to discriminate between cancer patients than there are samples - the so-called curse of dimensionality. For training purposes and to reduce the risk of overfitting, dimension-reduction is required in the form of *feature selection*. Multiple methods exists to select important features: filter methods, which are univariate methods that rank features in terms of importance and are commonly statistical measures (e.g. Student's t-test, Chi-squared or Wilcoxon sum-rank test); wrapper methods, which

iteratively remove (backward selection) or add (forward selection) features that contribute to the predictive power for the predictor; and embedded methods, which are learning algorithms which select features during training. Suprisingly, (Haury *et al.*, 2011) found that filter methods that are more simple can outperform wrapper or embedded methods. However, Boruta, a wrapper method for feature selection, has proved to be a powerful approach to -omics data as shown in previous work (Degenhardt *et al.*, 2019) and is, therefore, justified for application to aCGH data. Additionally, 'simple' classification algorithms (e.g. nearest centroids or K-nearest neighbours) have been shown to perform above expectations in multiple gene expression datasets (Haury *et al.*, 2011, Wessels *et al.*, 2005). For this reason, the effect that different feature selection methods have on the overall performance of 'simple' classifiers has to be explored further. It remains relatively unclear which feature selection methods, forward or wrapper method, in combination with 'simple' classification algorithms based on chromosomal copy number aberrations give the highest performance for the prediction of cancer subgroups.

In this paper, we aim to investigate 1) the use of CNA patterns for the classification of breast cancer subgroups and 2) the effect of a statistical filter method, Pearson's chi-squared test, and a wrapper method, Boruta, on the performance of classification algorithms. We evaluated three distinct classification methods: Nearest Shrunken Centroid (NSC), K-Nearest Neighbour (KNN) and Neural Network. KNN is considered a 'simple' algorithm according to (Haury *et al.*, 2011) and seems to outperform other more 'complex' algorithms. Additionally, NSC is a potentially useful algorithm in classifying subpopulations in high-dimensional datasets (Tibshirani *et al.*, 2003). Feed-forward neural networks with a single hidden (SLFN) layer have been found to perform significantly better in comparison to other popular and powerful classification methods (Huynh *et al.*, 2007). All of the above methods were trained on the data consisting of the chromosomal copy number aberrations of 100 breast cancer tumors divided into three subgroups based on pathological analysis. The performance of the feature selection method and classifier algorithm combination was evaluated using a nested 10-fold cross validation scheme, calculating accuracy for overall performance and the specificity & sensitivity for class-specific performance. The ultimate goal of this study is to produce a classification model that predicts the subgroup of unclassified breast cancer samples with the highest possible accuracy.

2 Methods

2.1 Exploratory analyses

Principal Component Analysis (PCA) was performed on the arrayCGH dataset to explore its structure and investigate the potential separability of subgroups. PCA was conducted with scaling applied, together with analysis of principal components and their explained variance, using the 'stats' package (version 3.6.3) for the statistical programming language R (version 3.6.3, R Core Team, 2020) Various clustering methods were investigated, including K-medoids and cosine distance (k = 3, pam() function from R package 'cluster', version 2.1.0), hierarchical clustering (Euclidean distance, average-linkage, dist() and hclust() functions from R package 'stats').

2.2 Cross-validation

A nested cross-validation scheme with a stratified 10-fold outer loop containing a stratified 10-fold inner loop was implemented to validate the model performance together with that of feature selection and hyperparameter tuning. Training and test datasets of the outer and inner loops were divided 70% / 30%, respectively, using createDataPartition()

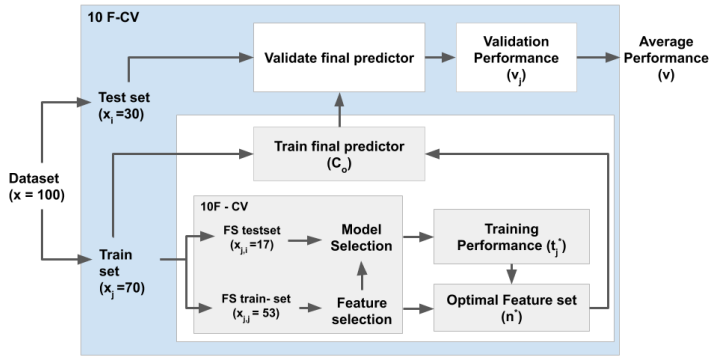

Fig. 1. Train-validation schematic depicted in a simplified format. The input is the breast cancer aCGH dataset and output is the final predictor model with average performance.

from R package 'caret' (version 6.0-86, Kuhn, 2008). The inner loop was used for feature selection with Pearson's chi-squared test or Boruta (Kursa and Rudnicki, 2010) and hyperparameter tuning (see below), the specific features and hyperparameters of the highest accuracy inner loop model were subsequently used in the outer loop. For each neural network, K-nearest neighbors (KNN) or Nearest Shrunken Centroids (NSC) classification model combined with either feature selection method, the final model was selected based on the highest accuracy score of 10 outer loop iterations.

2.3 Feature selection

Pearson's chi-squared test was used as statistical feature selection method and Boruta (Kursa and Rudnicki, 2010) was used as wrapper feature selection method. Chi-squared tests were conducted on all 2,834 genomic regions individually using the function chisq.test() of the R package 'stats'. Contingency tables were constructed of clinical subgroups versus chromosomal aberration values. A Bonferroni-correction was applied to a significance threshold of 0.05 to correct for multiple testing, the corrected threshold being $2834 * 0.05$. The effect of various significance thresholds on the number of selected features and the training accuracy was investigated in order to select the most adequate threshold. The nested cross-validation scheme (previously described) was used with a 1-fold outer loop and 10-fold inner loop to investigate the effect of different thresholds (0.05, 0.01, 0.001) with additional Bonferroni-correction on the training set accuracy as well as the validation set accuracy and number of selected features. Boruta was implemented with the function Boruta() from the R package 'Boruta' (version 6.0.0). Random forest models were trained on the original dataset together with 'shadow' versions of randomized features to asses the importance of individual features for model performance. Feature importances were determined by two-sided hypothesis tests of equality of Z-scores between original and shadow version, where features were considered important if their Z-score was higher than that of the shadow versions (Kursa and Rudnicki, 2010, Kursa, 2014). Undecided or "tentative" features were corrected for with the TentativeRoughFix() function using default parameters.

2.4 Classification models and hyperparameter tuning

Neural network, KNN and NSC classification models were developed to classify breast cancer subgroup status for samples based on chromosomal aberration values (previously described). All models were implemented with the function train() from the R package 'caret'. Neural network models were trained with the value 'nnet' for the 'method' parameter of

the function train(), which employed a single hidden layer feed-forward network (SLFN). Hyperparameters were tuned by passing combinations of 'decay' (ranging 5 to 7) and 'size' (0.1, 0.01, 0.001) to the 'tuneGrid' parameter of train(). KNN models were trained with the value 'knn' for the 'method' parameter of the function train(). Hyperparameters were tuned by passing values of 'k' (ranging 1 to 10) to the 'tuneGrid' parameter of train(). NSC models were trained with the value 'pam' for the 'method' parameter of the function train(). Hyperparameters were tuned by passing values of 'threshold' (ranging 1 to 8) to the 'tuneGrid' parameter of train().

2.5 Hierarchical clustering

Hierarchical clustering of feature subsets from specific classification models was conducted using the function heatmaply() from the R package 'heatmaply' (version 1.1.0) using average-linkage (UPGMA) clustering by passing 'average' to the parameter 'clust_method'.

2.6 Permutation analysis

Permutations of the best-performing neural network model training set were conducted to assess the likelihood of obtaining that accuracy score by chance under similar conditions. Samples were first divided in 70% training and 30% test data that was representative for clinical subgroup distributions (createDataPartition() function of 'caret'). Second, feature sets of identical size to the best-performing model (38 features) were randomly sampled with replacement from the full feature space ($n = 2,834$). Third, neural networks were trained with the train() function from 'caret' using permuted training sets and identical hyperparameters to the best performing model (size = 5, decay = 0.01, MaxNWts = $10 * (size * (n.o.features + 2) + size + 1)$, maxit = 100). Fourth, accuracy scores were obtained from comparing model predictions with the actual subgroup status of test samples. In total 1,000 permutations were conducted, the Pvalue for the best-performing model's accuracy was calculated as the number of observations with equal or higher accuracy than the original (0.895).

3 Results

3.1 Exploratory analyses

In order to obtain an overview of the dataset structure, Principal Component Analysis (PCA) was performed and chromosomal CNAs were visualized, ordered by clinical subgroup. However, PCA results were not suitable to detect any separation of subgroups once projected on the first two Principal Components (combined explained variance of only 22.21%). Moreover, visualization of chromosomal CNAs failed to point out obvious discriminatory features for subgroups. Altogether, these findings indicated that further feature selection was needed before classification model training.

3.2 Benchmark of classifiers

3.2.1 Nested cross-validation

We employed a nested 10-fold stratified cross-validation scheme, with inner loop for feature selection / hyperparameter selection and outer loop for model training, both loops split the samples to 70% training and 30% testing (see Methods for full details).

Initially, Pearson's chi-squared test for differences between clinical subgroups was conducted per feature, with a Bonferroni-adjusted threshold P-value of $0.05 * 2,834$. However, still at several hundred significant features for most models, we tested more stringent P-values (0.01, 0.001). Initially, Pearson's chi-squared test for differences between clinical subgroups was conducted per feature. We performed a test for the significance threshold by nested CV with a 1-fold outer and 10-fold inner loop to check for the effect on the accuracy of the training as well as the validation and the amount of selected features. It was observed that the amount of features selected decreases upon lowering the significance threshold for all classifiers (NSC: 586, 195, 36; KNN: 421, 206, 28; NNet: 421, 207, 36). For NSC, upon lowering the significance threshold (0.05, 0.01, 0.001) a relatively stable training accuracy (0.70, 0.69, 0.69) and increasing validation accuracy (0.54, 0.64, 0.68) was observed. For KNN, an increasing trend was observed for both the training accuracy (0.58, 0.62, 0.67) and the validation accuracy (0.57, 0.57, 0.64). For Neural Network, the training accuracy remained relatively stable (0.76, 0.77, 0.75) whereas the validation accuracy fluctuated (0.75, 0.61, 0.71). Considering all classifiers, the significance threshold of 0.001 was selected for feature selection with Pearson's Chi-squared test.

Table 1. Metrics of (subgroup-specific) performance for cross-validated models

Model	Feature selection	Accuracy	Subgroup	Specificity	Sensitivity	F1
NN	Chi-squared	0.821	HER2+	1	1	1
			HR+	0.882	0.727	0.762
			TN	0.850	0.750	0.706
NN	Boruta	0.895	HER2+	1	1	1
			HR+	0.917	0.857	0.857
			TN	0.923	0.833	0.833
KNN	Chi-squared	0.821	HER2+	1	1	1
			HR+	0.882	0.727	0.762
			TN	0.850	0.750	0.706
KNN	Boruta	0.857	HER2+	1	1	1
			HR+	0.889	0.8	0.8
			TN	0.894	0.778	0.778
NSC	Chi-squared	0.786	HER2+	0.950	1	0.941
			HR+	0.810	0.857	0.706
			TN	0.933	0.615	0.727
NSC	Boruta	0.750	HER2+	1	1	1
			HR+	0.789	0.667	0.632
			TN	0.833	0.6	0.632

NN, neural network; KNN, K-nearest neighbors; NSC, nearest shrunken centroids; TN, Triple Negative.
Selected models had the highest accuracy score for equivalent models in 10-fold stratified cross-validation.

3.2.2 Performance of classifiers

We benchmarked neural networks, K-nearest neighbors (KNN) and nearest shrunken centroids (NSC) as classifiers with chi-squared test and Boruta (Kursa and Rudnicki, 2010) as feature selection methods, using the accuracy as performance metric. This revealed that a neural network (hyperparameters: decay = 0.01, size = 5) with 38 Boruta-selected features performed best at classifying clinical subgroups with an accuracy of 0.895. A comprehensive overview of all models is displayed in Table 1.

Overall, the neural networks performed better than KNN models, which subsequently outperformed all NSC models. Chi-squared feature selection in neural network and KNN models had in fact identical accuracy and other performance metrics. Boruta outperformed chi-squared as feature selection method for both neural network and KNN models, albeit with a difference of only 0.055 on average, while Boruta failed to outperform the chi-squared model of NSC. One consistency in all neural network and KNN models were perfect HER2+ specific performance metrics. Finally, while the F1-score was consistently lower in the triple negative subgroup of neural network and KNN models, the best-performing model did reach a F1-score of 0.833 for this apparently more challenging classification subject.

3.3 Evaluation of best model

Based on the promising model accuracy score and subgroup-specific metrics, HER2+ in particular, we set out to further evaluate the model and detect potential flaws.

3.3.1 Clustering of genomic regions

The relationships between the best model features in all samples were assessed with average linkage (UPGMA) clustering of chromosomal aberration values as shown in Figure 2. This was visualized for the features from the best performing Boruta - neural network model (Figure 2A) and the features, which were also retrieved with Boruta, overlapping between the best models of all classification algorithms (Figure 2B). Amplification events were overall sparse yet dominated the HER2+ subgroup in a cluster-outgroup region of chromosome 17 (Figure 2A). Besides single occurrences for chromosomes 1, 16 and 22, most chromosomes with multiple occurrences were directly clustered together (3, 5, 10 and 12). It was in chromosomes 6 and 17 that regions were not all directly linked together. The former appeared to be divided by more chromosomal losses for HR+ samples in certain regions, while the latter included the amplification outgroup for HER2+ and a region with more copy losses for HR+ samples. The amplification outgroup was located on a chromosome 17q12 subregion (feature: 2185, hg18: chr17:35076296-35282086) and included *HER2*/*ERBB2* for which amplification is a hallmark in the HER2+ subgroup (Harbeck *et al.*, 2019). All classification algorithms confirmed the amplification outgroup 17q12 to be able to classify HER2+ samples (Figure 2B). For the remainder of the shared features, clustering of the chromosomes occurred in an ordered fashion. Here, it can be observed that chromosomal gains for HR+ occur relatively more often at chromosome 12q21.31 subregion (feature: 1678, chr12:84542006-85443011 & feature: 1679, chr12:85450052-85962613) and relatively more chromosomal gains occur for TN at chromosome 22q13.2 subregion (feature: 2752, chr22:41307174-41912419).

3.3.2 Permutation analysis

To assess the likelihood of obtaining the model's accuracy with these specific features due to chance, we performed a permutation analysis. New feature sets of identical size (38 features) were sampled with replacement from the entire feature set, neural networks were trained and tested under identical conditions and hyper parameters, and the accuracy was assessed ($n = 1000$). None of the permuted models achieved a similar or higher accuracy than the original model, thus a P-value of $P \approx 0.001$ was assigned to the original model (Figure 2). These results strongly indicate that the model accuracy can not be attributed to random effects.

3.3.3 Identification of genes and biomarkers

We investigated the specific genomic regions that served as model features in the best-performing neural network to find biological factors that underpin the model. To accomplish this, genes that were (partially) located within the hg18 reference genome coordinate ranges of model features were obtained and investigated for associations with breast cancer in literature.

The most striking results were the genes identified in the 17q12 subregion, which displayed a distinct amplification in all HER2+ samples in Figure 2B, and included *HER2*/*ERBB2* that is characteristic for HER2+ patients (Harbeck *et al.*, 2019). Beyond this single gene, we hypothesized that the 17q12 subregion could be the *HER2* amplicon (Sahlberg *et al.*, 2013). We found several genes that provided additional evidence for this hypothesis, including *PNMT*, *GRB7*, *STARD3*, *TCAP* and *PERLD1*, that have been associated with the HER2 amplicon (Sahlberg *et al.*, 2013). In addition, all these genes except for *PERLD1* were also found to be

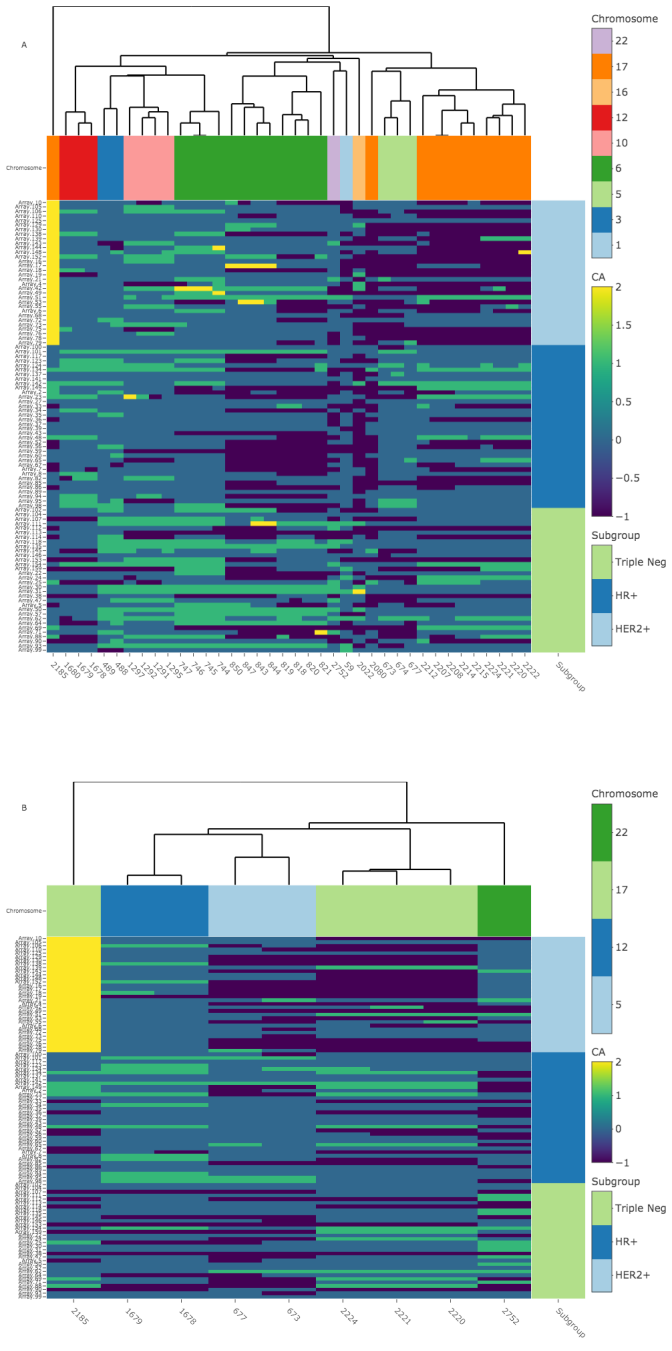

Fig. 2. Heatmap of chromosomal abberation patterns of all breast cancer samples A) including only the features from best performing NNet model and B) including features overlapping between the best models of all classification algorithms. The columns depict the chromosomal regions, whereas the rows show all the breast cancer samples, ordered based on subtype ("light blue" equals HER2+, "blue" equals HR+ and "green" equals Triple Negative). The filled tiles represent the different chromosomal aberration states: amplification (+2), gain (+1) and loss (-1). For chromosomal regions, the chromosome number has been shown and have been clustered by hierarchical clustering.

important for HER2+ in another breast cancer subtype classification study (Pan *et al.*, 2019).

In a subregion of 17q21.31 (feature: 2208, hg18: chr17:38077519-38710043) of the same chromosome we found additional genes with known associations or implications with breast cancer. First and foremost, we found the tumor suppressor and breast cancer hallmark gene *BRCA1*

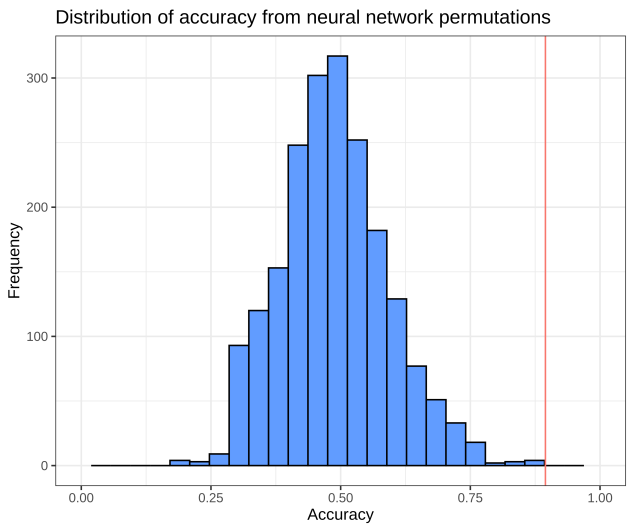

Fig. 3. Histogram of accuracy scores for 1000 permuted neural network models (in blue) and the original accuracy (0.895, red line) derived from best-performing neural network model for breast cancer subgroup classification. Accuracy scores of permuted models appear to be normally distributed, none are equal to or are large than the original accuracy score.

(Harbeck *et al.*, 2019). Adjacent to *BRCA1* lies *NBR2* and it has been found to encode for a non-coding RNA with a role in tumor development suppression (Xiao *et al.*, 2016) . *AOC3* has previously been implicated in metastatic breast cancer(Cha *et al.*, 2018). In another study, *BECN1* was identified as an oncogene with specific involvement in triple negative breast cancer (Wu *et al.*, 2018). Finally, the 17q12.31 subregion included *CCR10* which may have a key regulatory role in the cell invasion and migration of breast cancer (yu Lin *et al.*, 2017).

One additional discovery on chromosome 17 was *WNT3*, located in a subregion of 17q21.32 (feature: 2220, hg18: chr17:42161364-42296514), for which previous studies have linked it together with the WNT pathway to HER2-overexpressing cells (Wu *et al.*, 2017). Furthermore, *RASSF9* was found in a subregion of chromosome 12q21 (feature: 1678, hg18: chr12:84542006-85443011) and it has been found to be related to breast tumour initiation and propagation in previous studies (Li *et al.*, 2018). Finally, we found *CYB5R3* (NADH-Cytochrome B5 Reductase 3) in a subregion of chromosome 22q13 (feature: 2752, hg18: chr22:41307174-41912419) that has been shown to drive metastasis in triple negative (oestrogen receptor negative) breast cancer (Lund *et al.*, 2015). Another study (Blanke *et al.*, 2014) found that a particular variant of this gene, *I1M+6T*, is more common in breast cancer patients than in controls.

In the end, we investigated the potential of individual model features to serve as biomarker for the clinical subgroups. Earlier on we have shown that a subregion of chromosome 17q12 could completely discriminate HER2+ patients from HR+/triple negative by a distinct amplification pattern (Figure 2) that was reflected by HER2+-specific sensitivity, specificity and F1-scores of 1 for HER2+ in most models, including the best-performing model. Therefore we nominated this feature to serve as biomarker for HER2+ patients, whereas distinct biomarkers for HR+ or triple negative patients were not apparent.

4 Discussion

Potential issues with our methodology that could have affected our results include that we have only tested a limited number of classification methods. In particular there are many other types of neural network classifiers besides our SLFN model that may prove to be better-suited to classifying breast cancer subgroups using CNA data. To that end, ensemble classifiers have been experimentally proven to give better accuracy than individual classifiers on most datasets (Hsieh *et al.*, 2012). Testing an ensemble machine learning model with simple classification algorithms on such data could form the basis of future research.

Another complicating factor in our three-way classification is that some subgroups are more difficult to identify than others. In particular, the triple negative subgroup is known to be very difficult to identify (Bianchini *et al.*, 2016) . Our results follow similar trends, as we observe that the classifier performs less on the triple negative breast cancer subgroup (Table 1.).

Additionally, we cannot distinguish in the dataset between driver mutations & aberrations and those that have arisen in the already-formed tumour. This could potentially introduce residual noise into the classifier. In a more general sense, the availability of 100 tumour samples could be limiting the classification performance. With a larger number of samples, the molecular signature of triple negative samples could potentially be exposed and thus classified. In addition, it could possibly detect previously characterized heterogeneity within the HER2+ subgroup. A subregion of chromosome 17q12 was found to display a characteristic amplification pattern in HER2+ patients that allowed for the complete discrimination of this subgroup against others. Biological factors underpinning the HER2+ subgroup were found in the *HER2* gene and amplicon. Altogether this region provided an accurate biomarker for HER2+ patients.

Regarding the research question, the potential of CNA patterns for classification of breast cancer subgroups has partially been achieved. For HER2+ a biomarker was identified that completely discriminated this subgroup from other subgroups, whereas similar specific biomarkers for the other subgroups were not found. The second part involved the effect of a statistical filter method, Pearson's chi-squared test and random forest-based wrapper method, Boruta. Boruta outperformed the chi-squared test in all neural networks and KNNs. According to previous research, random forest-based classifiers are more robust against increased error due to a large number of noisy features (Fortino *et al.* [2014]).

It is clear that our classifier distinguishes the HER2+ subgroup from HR+ and triple negative with high precision. Our classifier could therefore aid the speed and accuracy of HER2+ diagnosis in particular. Further elucidation of chromosomal aberration patterns specific to HR+ and triple negative subgroups are still needed to improve the classification of these subgroups, which could potentially speed up the access to personalized treatments for these patients.

References

Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E., and Gianni, L. (2016). Triple-negative breast cancer: Challenges and opportunities of a heterogeneous disease.

Blanke, K. L., Sacco, J. C., Millikan, R. C., Olshan, A. F., Luo, J., and Trepanier, L. A. (2014). Polymorphisms in the carcinogen detoxification genes CYB5A and CYB5R3 and breast cancer risk in African American women. *Cancer Causes and Control*.

Cha, Y. J., Jung, W. H., and Koo, J. S. (2018). Site-specific expression of amine oxidases in breast cancer metastases. *Tumor Biology*, **40**(5), 101042831877682.

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, **20**(2), 492–503.

Fortino, V., Kinaret, P., Fyhrquist, N., Alenius, H., and Greco, D. (2014). A robust and accurate method for feature selection and prioritization from multi-class OMICs data. *PLoS ONE*, **9**(9), e107801.

Harbeck, N., PenaFult-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., and Cardoso, F. (2019). Breast cancer. *Nature Reviews Disease Primers*, **5**(1).

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

Hsieh, S.-L., Hsieh, S.-H., Cheng, P.-H., Chen, C.-H., Hsu, K.-P., Lee, I.-S., Wang, Z., and Lai, F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of medical systems*, **36**(5), 2841–2847.

Huynh, H. T., Kim, J.-J., and Won, Y. (2007). Dna microarray classification with compact single hidden-layer feedforward neural networks. In *2007 Frontiers in the Convergence of Bioscience and Information Technologies*, pages 193–198. IEEE.

Kloosterman, W. P. and Hochstenbach, R. (2014). Deciphering the pathogenic consequences of chromosomal aberrations in human genetic disease. *Molecular Cytogenetics*, **7**(1).

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, **28**(5), 1–26.

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods. *BMC bioinformatics*, **15**(1), 8.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*.

Li, B., Chen, P., Wang, J., Wang, L., Ren, M., Zhang, R., and He, J. (2018). MicroRNA-1254 exerts oncogenic effects by directly targeting RASSF9 in human breast cancer. *International Journal of Oncology*.

Lund, R. R., Leth-Larsen, R., Di Caterino, T., Terp, M. G., Nissen, J., Lænkholm, A. V., Jensen, O. N., and Ditzel, H. J. (2015). NADH-cytochrome b5 reductase 3 promotes colonization and metastasis formation and is a prognostic marker of disease-free and overall survival in estrogen receptor-negative breast cancer. *Molecular and Cellular Proteomics*.

Pan, X., Hu, X. H., Zhang, Y. H., Chen, L., Zhu, L. C., Wan, S. B., Huang, T., and Cai, Y. D. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Molecular Genetics and Genomics*.

R Core Team (2020). R: A language and environment for statistical computing.

Redon, R., Fitzgerald, T., and Carter, N. P. (2009). Comparative genomic hybridization: DNA labeling, hybridization and detection. pages 267–278.

Sahlberg, K. K., Hongisto, V., Edgren, H., Mäkelä, R., Hellström, K., Due, E. U., Moen Vollan, H. K., Sahlberg, N., Wolf, M., Børresen-Dale, A. L., Perälä, M., and Kallioniemi, O. (2013). The HER2 amplicon includes several genes required for the growth and survival of HER2 positive breast cancer cells. *Molecular Oncology*.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117.

van de Wiel, M. A., Picard, F., van Wieringen, W. N., and Ylstra, B. (2011). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Briefings in Bioinformatics*, **12**(1), 10–21.

Van Den Broek, E., Dijkstra, M. J., Krijgsman, O., Sie, D., Haan, J. C., Traets, J. J., Van De Wiel, M. A., Nagtegaal, I. D., Punt, C. J., Carvalho, B., Ylstra, B., Abeln, S., Meijer, G. A., and Fijneman, R. J. (2015). High prevalence and clinical relevance of genes affected by chromosomal breaks in colorectal cancer. *PLoS ONE*, **10**(9), 1–14.

Van't Veer, L. J., Dai, H., Van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

Wessels, L. F., Reinders, M. J., Hart, A. A., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**(19), 3755–3762.

Wu, C.-L., Zhang, S., Lin, L., Gao, S.-S., Fu, K.-F., Liu, X.-D., Liu, Y., Zhou, L.-J., and Zhou, P.-K. (2018). BECN1-knockout impairs tumor growth, migration and invasion by suppressing the cell cycle and partially suppressing the epithelial-mesenchymal transition of human triple-negative breast cancer cells. *International Journal of Oncology*.

Wu, Y., Tran, T., Dwabe, S., Sarkissyan, M., Kim, J., Nava, M., Clayton, S., Pietras, R., Farias-Eisner, R., and Vadgama, J. V. (2017). A83-01 inhibits TGF-β-induced upregulation of Wnt3 and epithelial to mesenchymal transition in HER2-overexpressing breast cancer cells. *Breast Cancer Research and Treatment*.

Xiao, Z.-D., Liu, X., Zhuang, L., and Gan, B. (2016). NBR2: A former junk gene emerges as a key player in tumor suppression. *Molecular & Cellular Oncology*, **3**(4), e1187322.

yu Lin, H., ming Sun, S., feng Lu, X., ying Chen, P., fa Chen, C., quan Liang, W., and yan Peng, C. (2017). CCR10 activation stimulates the invasion and migration of breast cancer cells through the ERK1/2/MMP-7 signaling pathway. *International Immunopharmacology*, **51**, 124–130.