Genomic Machine Learning Model Predicts Radiation Therapy Benefit in Early-Stage Breast Cancer Patients with High Accuracy

Kimberly Badal (kim.badal@gmail.com)
University of the West Indies
Jerome E. Foster
University of the West Indies
Rajini Haraksingh
University of the West Indies
Melford John
University of the West Indies

Research Article

Keywords: Radiation therapy, early-stage breast cancer, recurrence, machine learning, genomic

Posted Date: February 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1209026/v2

Abstract

Background

Radiation therapy (RT) is frequently recommended for post-surgery treatment of early-stage breast cancer (BC) patients, though not all benefit. Clinical factors currently guide RT treatment decisions. At present, models to predict RT-benefit predominantly use statistical methods with modest performance. In this paper we present a high-accuracy genomic Machine Learning (ML) model to predict RT-benefit in early-stage BC patients. We also present a novel method for selecting genomic features for training ML algorithms.

Methods

Gene expression data from 463 early-stage BC patients treated with surgery and RT from the METABRIC cohort were obtained. Wilcoxon Rank Sum (Wilcoxon RS) test and Cox Proportional Hazards (Cox PH) were used to reduce the number of genes used to train eight ML algorithms. ML algorithms were trained on 80% of data using 10-fold cross validation and tested on 20% of data to assess performance in predicting relapse status.

Results

Genome-wide gene expression data was reduced by 96% using Wilcoxon RS and Cox PH to a 1,596 gene set and a 977 gene set. These gene sets were used to train eight ML algorithms resulting in models that ranged in performance accuracies from 54.01% to 95.6%. Highest accuracies were obtained using Support Vector Machine (SVM977–93.41%, SVM1596–95.6%) and Neural Networks algorithms (NN977 – 92.31%, NN1596 – 93.41%). In RT-untreated patients, accuracies of all models were 30% to 40% lower compared to RT-treated patients. SVM977 had the highest sensitivity of 91.09%. Members of the 977 set were enriched with genes involved in cell cycle and differentiation as well as genes associated with radiosensitivity and radioresistance.

Conclusion

This study presents a novel genomic feature selection approach that used Wilcoxon RS followed by Cox PH to reduce the number of genes from genome-wide gene expression data used for training ML algorithms by 96%. This approach led to an SVM model that used the expression values of 977 genes to predict RT-benefit in early-stage BC patients with 93.41% accuracy. This work demonstrates that ML models can be clinically useful for predicting cancer patient outcomes.

Introduction

Breast cancer (BC) is a major global public health concern. It represents 11.6% of all cancer cases and 6.6% of all cancer deaths globally, second after lung cancer (1). Radiation therapy (RT) has been shown to reduce 10-year recurrence risk for early-stage BC patients by 15% (2). RT also offers improvements in

overall and disease-free survival in patients presenting with distant metastases at diagnosis (3). Presently, RT recommendations are based on clinicopathological characteristics and patient age only (4). According to the 2019 European Society for Medical Oncology Clinical Practice Guidelines, RT is strongly recommended for all early-stage BC patients following breast-conserving surgery (BCS) to reduce locoregional recurrence (LRR) risk (4). An RT boost is recommended for patients who are at a high risk of recurrence such as those who were diagnosed under the age of 50 years, with grade three tumours, and with vascular invasion (4). Despite these recommendations 10% of node-negative BC patients will still have a recurrence after BCS and RT (2). Therefore, a more personalized approach to determine which patients should and should not have RT is warranted.

The role of RT is to remove occluded cancer deposits in the tumour area after resection. RT destroys cancer cells by directly causing DNA crosslinks, and single and double strand breaks, and indirectly through free radical production which in turn damages DNA leading to apoptosis (5, 6). Rapidly proliferating cancer cells have been shown to be more sensitive to DNA damage, repair more slowly, produce more double strand breaks in their own proliferation, and typically have mutations that cause a loss in DNA repair pathway redundancies that are seen in normal cells (5, 7). Tumours also have mechanisms of acquired radioresistance which include enhanced migration and invasion (8), subtype switching (8), repopulation during gaps in RT (9, 10), and redistribution of cells to more radioresistant G1 and S phases of the cell cycle (11).

Some studies have suggested that BC subtype may be predictive of recurrence after postmastectomy RT as basal-like and luminal tumours show the most reduction in locoregional recurrence (LRR) rate after RT (12, 13). However, clinical trials have confirmed that subtype is prognostic of recurrence rather than predictive of RT response (14–16). The 2021 St. Gallen Breast Cancer Conference also concluded that genomic signatures of intrinsic subtype currently approved for clinical practice such as Oncotype DX® and MammaPrint® cannot guide RT treatment decisions (17). Therefore, RT recommendations are currently based on clinicopathological characteristics and patient age only.

Advances in molecular medicine and precision oncology offer the opportunity to include genetic information in addition to clinicopathological and demographic information in RT-benefit prediction models. Clinicogenomic models predicting RT-benefit as percentage risk reduction in either LRR, disease-free or overall survival have been previously proposed (18–23). These models generally utilize gene sets selected from the literature shown to be associated with radiosensitivity or radioresistence and use traditional statistical approaches such as Cox Proportional Hazard (Cox PH) and linear regression to predict RT response. A recent study by Zhang *et al*. used clinical factors to develop a nomogram that predicted 5- and 10- year survival benefit from post mastectomy RT with 60–80% accuracy (18). In 2019, Sjostrom *et al.* developed a clinicogenomic classifier called the Adjuvant Radiotherapy Intensification Classifier (ARTIC), which used expression values of 27 genes and patient age to predict the need for RT intensification in early stage BC patients that was validated on clinical trial data (19). Patients with low ARTIC scores had a statistically significant 70% risk reduction in LRR with RT compared to no RT, while

patients with a high ARTIC score had a 30% risk reduction in LRR with RT compared to no RT which was not statistically significant (19).

Earlier work includes a 2012 study by Torres-Roca *et al*. where a radiosensitivity index (RSI) was developed which used the expression values of 10 genes and a linear regression model (20). This molecular signature was validated in BC cohorts where it stratified patients as radioresistant or radiosensitive, and where radiosensitive patients had an improved 5-year relapse-free survival compared to radioresistant patients (95% vs. 75%) that was not observed in RT-untreated patients (21). In 2014, Tramm *et al.* used expression values of seven genes to predict LRR post-mastectomy in high-risk BC patients treated with systemic therapy from the Danish Breast Cancer Cooperative Group randomized control trial. They classified patients as low- or high-risk of LRR and RT benefit was observed in high-risk patients (22). In 2018, Cui *et al.* developed a 34-gene radiosensitivity signature that also stratified patients into radiosensitive and radioresistant groups (23). Notably, there is no overlap between the genes used in some of the models proposed (21, 22, 24), possibly reflecting the complexity of the biological networks that govern RT response. A more comprehensive, hypothesis-independent approach to gene selection that considers the expression values of all genes in the human genome is yet to be published.

ML algorithms are advantageous as they can process large amounts of data, and build complex models to make predictions (25). As such, they have been successfully applied to a range of cancer prediction problems, with accuracies as high as 100% (25–28). Statistical tests such as Wilcoxon Rank Sum test (Wilcoxon RS) have been used to select genomic features for ML algorithms (29, 30). For example, Niméus-Malmström *et al.* selected 5,237 genes using Wilcoxon RS which were used as features in a Support Vector Machine (SVM) algorithm which was able to predict recurrence in oestrogen receptor-positive patients (Area Under Receiver Operating Characteristics curve (AUROC) – 0.91) (30). ML algorithms have yet to be applied to the prediction of RT-benefit and therefore present a unique opportunity to create an improved and novel model. In this paper we present a high-accuracy genomic ML model to predict RT-benefit in early-stage BC patients. We also present a novel method for selecting genomic features for training ML algorithms that incorporates Wilcoxon RS.

Methods

Description of datasets

Clinical and gene expression data from the METABRIC study (31) were downloaded from cBioPortal (32). Clinical data included BC type, stage, surgery, chemotherapy and recurrence status. Gene expression data was in the form of log-transformed Z-scores compared to the expression distribution of all samples for 24,368 genes obtained using an Illumina HumanHT-12 v3 Expression BeadChip microarray. Gene expression and clinical data were merged using patient ID. The cohort was limited to patients with stage one or two breast invasive ductal carcinoma, who were treated with surgery and RT, but not chemotherapy. The outcome to be predicted was relapse status which was the time from the date of diagnosis to the date of the first report of a new tumour event, which included LRR, distant metastasis or

death with tumour (33). In this dataset, referred to as *dataset one*, the event status was assumed to be known for all patients; that is, all patients who were coded as “Disease Free”, were assumed to not have a relapse and all patients who were coded as “Recurred/Progressed” were assumed to have a relapse during the follow-up. Therefore, patients whose recurrence status was unknown because they were lost to follow-up, or died of other causes before a relapse, i.e. right censored patients were coded as “DiseaseFree.”

To address the issue of patients with unknown relapse status due to right censoring, *dataset one* was further limited to patients that had complete follow-up for at least 15 years or had a recurrence or death within 15 years, whichever came first. A 15-year period was chosen as the majority of patients (73.58%), had a follow-up period of 15 years or less. This dataset is referred to as *dataset two*. Patients who had a recurrence within 15 years were coded as “Recurred/Progressed”, while patients who did not have a recurrence within 15 years were coded as “DiseaseFree.”

Technical specifications

All analyses were performed in RStudio Version 1.3.959 using a computer with the following specifications: MacOS 10.14.6, 1.6 GHz Intel Core i5 CPU, 8 GB RAM. All ML models were developed using the Classification and Regression Training (caret) package (34). Sample code used for analyses is included in Additional file 1.

Gene set selection and machine learning algorithm training

The set of gene expression values for 24,368 genes was reduced using Cox PH and Wilcoxon RS to form three gene sets. First, Wilcoxon RS was used to determine which genes were differentially expressed between patients with and without a relapse at a significance level of $p < 0.05$. This led to the first gene set referred to as the Wilcoxon set. Cox PH was used to determine which genes affected recurrence risk at a significance level of $p < 0.05$. This led to a second gene set referred to as the Cox PH set. Lastly, Wilcoxon RS followed by Cox PH were used sequentially to reduce the number of genes which led to the third gene set referred to as the Wilcoxon-Cox set.

Eight machine learning (ML) models were chosen based on their extensive use in cancer prediction research (25, 27). These models were: Artificial Neural Networks (NN) (35), Linear Support Vector Machine (SVM) (36), K Nearest Neighbours (KNN) (37), Extreme Gradient Boosting (XGBoost) (38), Naïve Bayes (NB) (39), Decision Trees (DT) (40), Random Forest (RF) (41), and Logistic Regression (LR) (42). Each model was trained on 80% of the data randomly selected with 10-fold cross validation using gene sets selected using Wilcoxon RS and/or Cox PH as features, and then tested on the remaining 20% of data.

Other techniques were used to further reduce the three gene sets to determine the effect of smaller sets on prediction accuracy. First, the Wilcoxon set and the Cox PH set were reduced to the top 1000, 500, 100 and 50 genes with the lowest p values. Second, genes in the Wilcoxon-Cox set were reduced using hazard

ratios (HR) to form three sets: (i) genes with a HR>1; (ii) genes with a HR<1; and (iii) genes with HRs in the first and third tertiles. The subtype variable was also added to the Wilcoxon-Cox set to determine if it improved model performance. Each of these reduced sets were also used to train the ML algorithms. Third, recursive feature elimination (RFE) was used to investigate whether smaller subsets of genes would improve classification accuracies when used for algorithm training.

Five sets of genes selected from the set of genes with a Wilcoxon RS p value greater than 0.05 were also used for training and accuracy compared. The gene set was ordered by increasing Wilcoxon RS p value and two subsets of 1,596 genes with a p value greater than 0.05 corresponding to differentially expressed genes with ranked numbers 1597-3192 and 3193-4788 were selected. Another three sets of 1,596 genes chosen at random from genes that had a Wilcoxon RS p value greater than 0.05 were selected. Each of these five sets of insignificant genes were used to train SVM and NN algorithms.

A curated list of 64 radiogenes (Supplementary Table S2) from three publications on RT-benefit (21, 22, 24) was also used to train an SVM algorithm to determine whether genes of biological relevance selected from the literature would be valuable for training ML algorithms.

Model comparison and testing in other clinical populations

The four ML models with the highest accuracies were tested on patients who did not have RT to determine whether they were specific to RT-treated patients. These four models were further compared on computing time on a test set of 90 patients, AUROC, sensitivity and specificity. Of the four models, the best performing model was further tested on ER+, ER-, and chemotherapy-treated patients.

Hyperparameter tuning and gene set enrichment analysis

The hyperparameter of the best performing model was tuned using a manual grid search. The genes in the final model were characterized using Gene Set Enrichment Analysis (GSEA) with Gene Ontology terms for “biological process”, “cellular compartment” and “molecular function.” The overlap of this model’s gene set with a list of 64 genes previously used in RT-benefit predictive models (22, 24, 43) and 723 cancer driver genes from the Catalogue for Somatic Mutations in Cancer (COSMIC) v93 (44) were calculated. The stringApp (45) in Cytoscape (46) was used to visualize protein to protein interactions in the final gene set using an evidence threshold of 0.6.

Results

Demographic and clinicopathological characteristics of training cohort

After limiting the METBRIC cohort to patients who were stages one or two, who had surgery and RT but no chemotherapy and had gene expression data, this cohort had 463 patients (Figure 1A) of which 36.5%

(n=169) had a recurrence (Figure 1B). The median follow-up time was 10.68 years (range 0.21 – 29.25 years).

The clinical profile of patients in the RT-treated cohort was similar to that of the entire METABRIC cohort as previously described (33). In RT-treated patients the average age at diagnosis was 63.17 years (Table 1). The majority of patients' tumours were classified as luminal A (38.78%, n=185) or B (31.44%, n=150), were grade three (50.11%, n=236), stage two (55.56%, n=265), or had no lymph nodes examined as positive (62.72%, n=291) (Table 1). The mean tumour size was 21.93 mm. The majority of patients had hormone therapy (65.83%, n=314) or BCS (70.23%, n=335) (Table 1).

Gene sets used for ML algorithm training

The first stage of feature selection led to three gene sets: (i) the Wilcoxon set which consisted of expression values for 1,596 genes; (ii) the Cox PH set which consisted of expression values for 1,768 genes; and (iii) the Wilcoxon-Cox set which consisted of expression values for 977 genes (Figure 1B). Each of these gene sets were used to train ML algorithms. These gene sets were further reduced in a second stage using p values and recursive feature elimination as shown in Figure 1B.

Table 1
Distribution of clinical and demographic characteristics in radiation therapy-treated and -untreated patients in the METABRIC cohort

Variable	RT-treated No. (%) (n=477)	RT-untreated No. (%) (n=312)	p value[b]
Age	Mean 63.17 years	Mean 63.41 years	0.570[c]
PAM50 subtype			
Basal	37 (7.76)	24 (7.69)	0.258
HER2	34 (7.12)	27 (8.65)	
Luminal A	185 (38.78)	143 (45.83)	
Luminal B	150 (31.44)	72 (23.08)	
Normal	25 (5.24)	18 (5.77)	
Other[a]	46 (9.64)	28 (5.87)	
Grade			
I	35 (7.54)	34 (11.53)	0.137
I	200 (42.46)	125 (42.37)	
III	236 (50.11)	136 (46.10)	
Not reported	6 (1.29)	17 (5.45)	
Tumour size	Mean 21.93 mm	Mean 23.80mm	0.004[c]
Hormone therapy			
Yes	314 (65.83)	179 (57.37)	0.990
No	163 (34.17)	133 (42.62)	
Number of lymph nodes examined positive			
0	291 (62.72)	219 (71.80)	0.018
1	76 (16.38)	48 (15.73)	
2	35 (7.54)	12 (3.93)	
>=3	62 (13.36)	26 (8.52)	
Not reported	13 (2.72)	7 (2.24%)	
Stage			
1	212 (44.44)	135 (43.26)	0.745

Variable	RT-treated No. (%) (n=477)	RT-untreated No. (%) (n=312)	p value[b]
2	265 (55.56)	177 (56.73)	
Surgery			
Breast conserving	335 (70.23)	40 (12.82)	<0.001
Mastectomy	142 (26.79)	272 (87.18)	

SVM and NN machine learning algorithms trained using the Wilcoxon set, Cox PH set, and the Wilcoxon-Cox set consistently perform with high accuracy

The Wilcoxon set of 1,596 genes when used to train the eight ML algorithms resulted in classification accuracies that ranged from 54.01–95.60% (Figure 2). The models with the highest accuracies along with their associated confidence intervals were: SVM (95.6% [89.13-98.79]), NN (93.41% [86.2-97.54]), and XGBoost (79.12% [69.24-87.15]) (Figure 2). When the 1,000 genes with the lowest p values from the Wilcoxon set were used to train each of the eight ML algorithms a decrease in classification accuracies were observed across all eight models compared to when the full Wilcoxon set was used for training (Figure 2). SVM (87.91% [79.4 – 93.81]) and NN (85.71% [76.81-92.17]) maintained the highest accuracy with the reduced set of 1000 genes. When the 500, 100, and 50 genes with the lowest p values from the Wilcoxon Set were used for training each of the eight ML algorithms a general trend of decreasing accuracy with smaller gene sets was observed (Figure 2). These accuracies ranged from 52.17–84.62%.

The Cox PH set consisting of 1,768 genes when used to train the eight ML algorithms resulted in classification accuracies that ranged from 56.94–94.51% (Supplementary Figure S1). SVM (94.51% [87.64-98.19]), NN (91.21% [83.41-96.13]) and XGBoost (65.93% [55.25-75.55]) had the highest accuracies (Supplementary Figure S1). When the top 1000, 500, 100 and 50 genes with the lowest p values from the Cox PH set were used to train each of the eight ML algorithms, a decrease in accuracy across all models was observed. SVM and NN maintained the highest accuracy when the top 1000 genes (SVM 90.11% [82.05-95.38], NN 84.62% [75.54-91.33]) and 500 genes (SVM 85.71% [76.81-92.17%], NN 85.71% [76.81-92.17]) were used for training (Supplementary Figure S1). When the top 100 and top 50 genes were used for training, all of the eight models had lower classification accuracies ranging from 58.24–72.53% (Supplementary Figure S1).

The top five algorithms with the highest accuracies were chosen for further training. They were SVM, NN, RF, KNN, and XGBoost. The Wilcoxon-Cox set of 977 genes when used for training resulted in a similar accuracy profile as when the Wilcoxon set and the Cox PH set were used for training (Figure 3). The SVM (93.41% [86.2– 97.54]) and NN (92.31% [84.79 – 96.85]) models continued to have the highest accuracy (Figure 3). The BC subtype variable when included as an additional feature to the Wilcoxon-Cox set of 977 genes resulted in similar accuracy for SVM (94.51% [87.64 – 98.19]), and a reduced accuracy for NN (80.22% [70.55 – 87.84]) (Figure 3).

The SVM algorithm was chosen for RFE as it had consistently high accuracy across the various gene sets used for training. RFE with subsets of the Wilcoxon set and Wilcoxon-Cox set showed a linear decreasing relationship between model accuracy and the number of genes in the subset. The full Wilcoxon set and Wilcoxon-Cox set of genes resulted in the highest classification accuracies of approximately 90% while gene sets with less than 200 genes resulted in the lowest classification accuracies of less than 60% (Supplementary Figure S2).

Decreased model accuracy was observed when the Wilcoxon RS p value threshold for gene set selection exceeded 0.05

The two algorithms that gave the models with highest accuracies – SVM and NN – were chosen for further training. Two subsets of 1,596 genes with a p value greater than 0.05 corresponding to DEGs ranked numbers 1597-3192 and 3193-4788 in the Wilcoxon RS test results were used for training. When these gene sets were used there was a decrease in accuracy of approximately 10–40% for both the SVM and NN models (Figure 4). When random sets of 1,596 insignificant genes ($p>0.05$) were used for training, there was a further decrease in accuracies which ranged from 46.15–67.47%. In summary, there was a trend of decreased accuracy when the gene sets used for training had a Wilcoxon RS test p value that exceeded 0.05 (Figure 4).

SVM and NN model accuracy was 30-40% higher for RT-treated versus untreated patients

There were no statistically significant differences ($p>0.05$) in age, PAM50 subtype, grade, stage and hormone therapy treatment status between the RT-treated and untreated cohorts (Table 1). RT-untreated patients had larger tumours compared to RT-untreated patients (23.80 mm vs. 21.93 mm, $p=0.004$). A larger proportion of RT-untreated patients had a mastectomy compared to RT-treated patients (87.18% vs. 26.79%, $p <0.001$). In RT-untreated patients there was a larger proportion of patients with no lymph nodes examined as positive ($p=0.018$) (Table 1).

The SVM and NN models trained using the Wilcoxon set and the Wilcoxon-Cox set of 1,596 and 977 genes respectively were tested on RT-untreated patients. The accuracies and associated confidence intervals for each model when applied to RT-untreated patients were as follows: SVM1596, 59.02% [53.27-64.59]; SVM 977, 54.33% [48.65-60.11]; NN1596, 52.79% [47.02-58.5]; NN977, 52.70% [47.02-58.5] (Figure 5). Thus, accuracies were lower by approximately 30–40% for RT-untreated compared to RT-treated patients for all models (Figure 5).

Computational Time, Sensitivity and Specificity of ML Models

The four models with the highest accuracies: SVM1596, SVM977, NN1596 and NN977 were compared on computing time, sensitivity and specificity. Each model took less than one second to predict relapse status on a test set of 90 patients (Supplementary Figure S3). All models had an AUROC value greater

than 0.94 (Supplementary Figure S4). SVM977 had the highest sensitivity (91.09%), but the lowest specificity (78.95%) compared to all other models, while SVM1596 had the highest specificity (92.79%) but the lowest sensitivity (81.55%) (Supplementary Table S1). The NN models had sensitivities and specificities between 85.29–89.23% (Supplementary Table S1). SVM977 was chosen for further analysis using GSEA and hyperparameter tuning as it had the highest sensitivity. Varying the values of the cost hyperparameter of the SVM algorithm with manual grid search showed no change in accuracy (Figure S5). Therefore, a cost value of one was chosen.

SVM model trained with the Wilcoxon-Cox set shows high performance independent of ER status

When the SVM model trained with the Wilcoxon-Cox set was tested on ER+ or ER- patients only, the accuracy, sensitivity and specificity were high ranging between 95.45% and 99.32% (Table 2). When this model was tested on chemotherapy treated patients, the accuracy decreased to 64.02% (Table 2). The sensitivity and specificity also decreased to 52.27% and 74.26% respectively.

Table 2
Accuracy, sensitivity and specificity of the SVM model with the Wilcoxon-Cox set when applied to different patient groups

Patient group	No. of patients	Accuracy (%)	Sensitivity (%)	Specificity (%)
Chemotherapy-treated	189	64.02 [56.74-70.86]	52.27	74.26
ER+	392	98.98 [97.41 – 99.72]	99.32	98.78
ER-	71	97.18 [90.19 – 99.66]	95.45	97.96

Wilcoxon-Cox gene set is enriched for known radiogenes and cancer driver genes

GSEA using Gene Ontology terms revealed that the most significant terms ($p<0.01$) in the 977 gene set mapped to biological processes related to multicellular organism development (n = 204 genes), mitotic cell cycle (n = 114 genes), cell cycle (n = 153 genes) and cell differentiation (n = 153 genes). When GSEA was performed for terms related to which cellular compartment the proteins of these genes operate in, it was found that the most highly represented cellular regions were the nucleoplasm (n = 194 genes), chromosome centromeric region (n = 39 genes) and centrosome (n = 22 genes). Lastly, the most significantly annotated terms related to molecular function were microtubule binding (n = 29 genes), cell adhesion molecule binding (n = 22 genes) and ATP binding (n = 72 genes).

Sixty-four of the 68 radiogenes (Supplementary Table S2) had mRNA expression data in the METABRIC dataset, representing 0.26% of the dataset. Sixteen of the 64 curated radiogenes (Supplementary Table S2) were in the Wilcoxon-Cox set of 977 genes representing 1.6% of the set. This means that there was an approximately 6.2 times enrichment of known radiogenes in the Wilcoxon-Cox set. Most of these 16

radiogenes are involved processes related to cell division e.g. *MKI67*, *SPC25*, and *PRC1* (Supplementary Table S3).

Fifty-five of the 723 genes from the COSMIC database overlapped with the Wilcoxon-Cox set. Assuming that all COSMIC genes are represented in the METBRIC dataset, this corresponds to an approximately 1.4 times enrichment of cancer driver genes in the Wilcoxon-Cox set. Cytoscape protein-protein network analysis revealed a highly interconnected network with interactions between almost all genes in the Wilcoxon-Cox set (Supplementary Figure S6).

SVM algorithm trained using 61 known radiogenes shows decreased performance

The SVM algorithm trained with 61 known radiogenes present in the METABRIC dataset had an accuracy of 54.61% [46.34-62.69], sensitivity of 24.56% and specificity of 72.63%.

Performance of an SVM model using gene sets from the cohort with complete 15-year follow up shows 7-9% decreased performance compared to full cohort

Of the 463 patients in the initial cohort, 252 (54.28%) had complete 15-year follow-up. Of the 252 patients in *dataset two* with complete 15-year follow-up, 60.32% (n=152) had a recurrence within 15 years while 39.68% (n=100) did not have a recurrence within 15 years. Of the 100 patients that did not have a recurrence within 15 years, 14.00% (n=14) went on to have a recurrence after 15 years.

When the Wilcoxon-Cox set of 977 genes from *dataset one*, were selected from *dataset two*, and used to train an SVM algorithm the resulting model had lower accuracy of 85.71% [72.76-94.06%], lower sensitivity of 70.59% but higher specificity of 93.75% than when *dataset one* was used for training (Table 3). When Wilcoxon RS ($p<0.05$) followed by Cox PH ($p<0.05$) were applied to *dataset two*, a new Wilcoxon-Cox set of 1,044 genes was selected. When this set of 1,044 genes was used to train an SVM algorithm, the resulting model had lower accuracy of 87.76% [75.23-95.37%], lower sensitivity of 70.59%, but higher specificity of 93.75% (Table 3). There were 316 genes that overlapped between the set of 977 and 1,044 genes.

Table 3
Comparison of the accuracy, sensitivity and specificity of SVM algorithms trained with the Wilcoxon-Cox gene sets selected from datasets one and two

Gene set	Accuracy (%) [95% CI]	Sensitivity	Specificity
Wilcoxon-Cox set of 977 genes (*dataset one*)	94.51% [87.64 – 98.19]	91.09%	78.95%
Wilcoxon-Cox set of 977 genes (*dataset two*)	85.71% [72.76 – 94.06%]	70.59%	93.75%
Wilcoxon-Cox set of 1,044 genes (*dataset two*)	87.76% [75.23 – 95.37%]	70.59%	96.88%

Discussion

This study demonstrates that ML can be used to develop highly accurate, sensitive, and specific models to predict RT-benefit in early-stage BC patients. We present a high-performance SVM model (93.41% accuracy, 91.09% sensitivity, and 78.95% specificity) that can predict RT-benefit in early-stage BC patients independent of subtype. Here, RT-benefit was defined as relapse-free status following surgery and RT. The accuracy of this model (93.41%) represents an improvement from that of the best previously reported model (80%) in predicting RT benefit (18). This model used an SVM algorithm and expression values of a set of 977 genes referred to as the Wilcoxon-Cox set to predict RT-benefit. This study also presents a novel genomic feature selection approach that reduced the number of genes from genome-wide gene expression data by 96% using Wilcoxon RS followed by Cox PH. This feature selection method contrasts previous studies that selected genes with known functions from the literature or *in vitro* experiments to build RT-benefit models (18–23). To our knowledge, this is the first study to apply ML algorithms to predicting RT-benefit with consideration of all genes in the human genome.

The preliminary challenge in model development was finding a publicly available dataset with complete clinical and gene expression data. The dataset also needed to be balanced in the outcome variable as unbalanced datasets can lead to a misleadingly high prediction accuracy. The dataset also needed to be sufficiently large as small datasets can lead to model overfitting (47) and lack of precision (48) for some ML algorithms. The METABRIC dataset was chosen because of its large cohort size (2,509 patients), balance in outcome (approx. 40% of patients had a recurrence), extensive longitudinal follow-up data (approx. 30 years) and availability of genome-wide gene expression data (24,368 genes).

There was significant variability in the follow-up time for patients (range 0.21 – 29.25 years). This meant that some patients were right censored, i.e. the patient's follow-up ended before relapse occurred. While Cox PH models account for right censorship, classification ML algorithms do not. Two strategies were used to address this issue of right censorship in the data: 1) patients who were lost to follow-up were assumed to have had no relapse in *dataset one*; 2) the cohort was limited to patients who were followed for at least 15 years, or had a recurrence or death within 15 years, whichever came first in *dataset two*. The assumption in the first method allowed all the data for the 463 patients to be retained for training. However, it had a disadvantage in that it treated patients with an unknown relapse status as no relapse, which may and may not be true for each patient. The second method made no assumption about the patient's relapse status as it limited the cohort to only patients who had complete follow-up. The disadvantage of this method is that it led to loss of 46% of the data which would have introduced some degree of exclusion bias into the dataset. The majority of analysis here used *dataset one*, due to its larger sample size. *Dataset two* was used for comparison in order to determine whether or not the ML methodology used would apply across datasets with different limitations.

Controls on BC type, stage, chemotherapy, surgery and RT status were implemented to define the clinical population which was early-stage BC patients who were treated with BCS and RT. Previous work modelling RT-benefit also controlled for BC subtype by building separate models for ER+ and ER- patients

(30). The rationale was that ER+ and ER- tumours are distinct in their gene expression profiles which are associated with the differences in outcomes observed by subtype (30). This study did not control for BC subtype for two reasons: first, such controls would create a more homogenous patient group eliminating key differences in expression profiles that the ML algorithm can utilize to make a classification; and second, a model that works irrespective of BC subtype would be easier to implement rather than two distinct models. For these same reasons, the cohort was not limited to patients who had hormone therapy (HT). HT is recommended for BC patients who are HR+ therefore, limiting to patients who either did or did not have HT would also limit the dataset to patients who were either HR+ or HR- subtype. The SVM model with the Wilcoxon-Cox set of 977 genes was shown to have high accuracy in ER+ and ER- patients demonstrating that the model is independent of subtype.

A key challenge in this study was to reduce the number of genes as the use of the entire set of 24,368 genes as the use of too many features can result in an overfitted model for some ML algorithms (49). To achieve this, a novel filter feature selection approach was developed that used Wilcoxon RS followed by Cox PH which reduced the number of genes by 96% to the Wilcoxon-Cox set of 977 genes. Wilcoxon RS was previously used to determine differentially expressed genes (DEGs) in BC datasets (29, 30). The application of Wilcoxon RS followed by Cox PH for genomic feature selection has not been reported. This novel approach reduced the dataset substantially by selecting a set of DEGs that also affected recurrence risk. This approach was also better than selecting known genes of biological relevance for training as when 64 radiogenes were used for training, the resulting SVM model had poor accuracy of 54.61%. Therefore, considering all genes, in a hypothesis-independent manner appears to be a better approach than selecting known genes for training.

A clear relationship between model accuracy and the number of genes selected using Wilcoxon RS and Cox PH was observed. When smaller gene sets of the top 1000, 500, 100 and 50 gene with the lowest p values were used for training, there was an overall decline in accuracy across all eight ML algorithms. Therefore, a larger number of genes was needed for higher accuracy. There was also a relationship between the significance threshold of 0.05 for gene selection and model accuracy. When gene sets with a p value greater than 0.05 were selected there was also a decline in accuracy for both the SVM and NN algorithms. The lowest performance (~55%) was seen when insignificant genes with a p value greater than 0.05 were randomly selected. These results demonstrate the importance of considering the significance threshold in genomic feature selection using Wilcoxon RS and Cox PH.

The top four models presented use SVM or NN with either the Wilcoxon set of 1,596 genes or the Wilcoxon-Cox set of 977 genes. Given that SVM and NN are the most consistently used algorithms in BC prediction research, this result further corroborates the utility and consistent performance of these models (25). The consistently high accuracy of SVM suggests that the genomic features selected by Wilcoxon RS and Cox PH are sufficiently separate in high dimensional space to determine an optimal hyperplane with a large margin. It also suggests that this feature selection approach was able to reduce noise in the feature space and overlap between classes. SVM with a radial or polynomial kernel function was also

investigated, however this did not improve accuracy (data not shown), therefore a linear hyperplane was sufficient for this problem.

The lower performance of the majority of ML algorithms chosen (RF, DT, XGBoost, KNN, NB and LR) may be attributed to their underlying assumptions or their inability to model complex relationships. For example, NB and LR assume independence among predictors. This assumption would not hold with gene expression data where the expression pattern of one gene is often directly or indirectly dependent on the expression of another. LR is also generally not able to model complex relationships and is traditionally used to model a linearly separable classification problem. KNN is known to underperform with high dimensional data where all the vectors are almost equidistant making it difficult to determine clusters using distance metrics. DTs are also known to underperform as single trees are unstable and tend to overfit the data.

Addition of the subtype variable to the Wilcoxon-Cox set did not improve accuracy of the SVM model and decreased accuracy of the NN model. For the SVM model, the subtype variable was not a support vector and therefore did not influence the position of the linear hyperplane separating those who did have a recurrence from those who did not. In summary, subtype was an unnecessary feature for the models presented.

It is significant that the ML models demonstrated better prediction accuracies for RT-treated patients compared to untreated-patients. The top four models (SVM977, SVM1596, NN977, NN1596) all performed poorly when applied to RT-untreated patients, with prediction accuracies of 50–60%. Notably, patients in the RT-untreated cohort had larger tumours, were more likely to have a mastectomy, and to have no lymph nodes examined as positive. Therefore, biological differences between the tumours of patients in the RT-treated and untreated cohorts likely resulted in differences in gene expression profiles between the cohorts, which subsequently impacted the SVM model performance. A similar trend of poor accuracy (64.02%) was also observed when the SVM977 model was tested on data for chemotherapy-treated patients. Taken together, these results are promising in supporting the validity of the SVM977 model in predicting relapse in early-stage, surgery and RT-treated, chemotherapy-untreated BC patients. Future work would involve further controlling for treatment factors such the type of surgery, and control for the extent disease progression by selecting patients with no lymph node metastasis in the training cohort.

Comparison of the four models with the highest accuracy (SVM977, SVM1596, NN977 and NN1596) revealed small differences in AUROC values (1-2%), and even smaller differences in computational time (<1 second) that would not be noticeable to the end-user. Therefore, a model was not chosen based on these characteristics. Sensitivity or the number of true positives was more important than the specificity or the proportion of true negatives. That is, it is more important to correctly predict recurrence in RT-treated patients as they can be given the opportunity for RT-intensification or sensitization as a clinical intervention to reduce the risk of recurrence. A RT boost has been shown to significantly reduce the risk of LRR but with an increased risk of moderate to severe fibrosis (50). Patients who are correctly identified as

having no recurrence (specificity) can continue with standard of care or have RT omission. Careful consideration of false positives is needed as these patients would be overtreated. Thus, the RT treatment course would require a risk-benefit discussion between the treating radiation oncologist and the patient. In summary, SVM977 is the best model because it had the highest sensitivity among all models.

Characterization of the Wilcoxon-Cox 977 gene set using GSEA revealed that many of these genes are involved in cell cycle and division and operate in the nucleoplasm. This was expected as it is well known that uncontrolled cell division is a hallmark of cancer (51). Further, previous work in BC cell lines found that the expression levels 51 genes that were correlated with radiosensitivity were enriched for genes involved in cell cycle arrest (24). This is also consistent with research that has shown that RT-resistance mechanisms are involved in repopulation and redistribution of cells to more radioresistant G1 and S phases of the cell cycle (10, 11). The 977 gene set was also enriched (6.2 times) with radiogenes which further demonstrated that the feature selection approach was able to select for known genes of biological relevance. These results suggest that it is likely the compounded effect of several hundred genes in highly interconnected networks involved in cell division and redistribution of cells in the cell cycle, that drives recurrence after RT.

When *dataset two* was used to develop a model to predict RT-benefit, the SVM model had approximately 7-9% lower accuracy, 20% lower sensitivity, but 15-18% higher specificity than when *dataset one* was used for training. This change in performance is likely due to the smaller training dataset used (limited to patients who had complete 15-year follow-up), also reflected in the wider confidence interval. However, the overall performance profile of this model was good, demonstrating that the methodology used for ML model development was valid using both datasets. Wilcoxon RS followed by Cox PH selected for a set of 1,044 genes in *dataset two*, of which 316 genes overlapped with the 977 gene set. Therefore, the genes selected for training using the proposed feature selection methodology is not fixed and depends on the patients in the cohort used. Given the genomic heterogeneity that has been shown to occur between and within BC subtypes (31, 52), and between different ancestral populations (53), it would be expected that the gene sets selected would vary with the cohort used.

This study had some limitations. First, the outcome used was relapse-free status. A more direct outcome for measuring RT-benefit would be ipsilateral LRR which was unavailable in the METABRIC dataset. Therefore, this study could not differentiate those patients who had recurrence of the same primary BC versus those who had a new primary. No information on the status of resection margins was available which is a known factor affecting recurrence risk. Further, information on RT-fields and dosages were absent to determine if the RT given was a commonly used dosage. This study also could not limit the cohort to patients who had a lumpectomy as the majority of patients had a mastectomy (~80%) while few had a lumpectomy (~20%). This is likely because the METABRIC cohort consists of patients who were diagnosed between 1977 and 2005 and since then there has been a shift toward breast conservation for early-stage BC patients (4). This study was also unable to test the SVM977 model in another BC cohort as the BC datasets available for public use were not adequately clinically annotated or sufficiently large for ML training. Further, inconsistent gene naming conventions resulted in an inability to

select the Wilcoxon-Cox set of 977 genes in other dataset. Future work would involve the application of the methodology used here to another BC cohort, preferably in the setting of a prospective randomized controlled trial as the gold standard (54).

Conclusion

We presented a methodology that can be used to develop ML models to predict RT-benefit in early-stage BC patients. The methodology incorporates a novel genomic filter feature selection approach that used Wilcoxon RS followed by Cox PH to reduce the set of genes from genome-wide gene expression data by 96%. This methodology resulted in a high-performance SVM model (93.41% accuracy, 91.09% sensitivity, and 78.95% specificity) that predicted RT-benefit in early-stage BC patients, independent of subtype, using the expression values for a set of 977 genes. The achievement of high accuracy demonstrates the potential of ML to address important problems of clinical interest. Our methodology can be applied to develop ML models that can be used to differentiate those patients who will go on to have a recurrence despite RT.

Declarations

Ethical Approval and Consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

The METABRIC dataset analysed during the current study is available in the cBioPortal repository, https://www.cbioportal.org/study/summary?id=brca_metabric

Competing interests

The authors have no competing interests to disclose.

Funding

This work was not funded by any grant.

Authors' Contributions

KB did the analysis and wrote the manuscript. MJ, RH and JF were major contributors in conceptual development and in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank Ms. Aliya Mohammed for her contributions in gene set characterization.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. Epub 2018/09/13. doi: 10.3322/caac.21492. PubMed PMID: 30207593.
2. Early Breast Cancer Trialists' Collaborative G, Darby S, McGale P, Correa C, Taylor C, Arriagada R, et al. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet. 2011;378(9804):1707–16. Epub 2011/10/25. doi: 10.1016/S0140-6736(11)61629-2. PubMed PMID: 22019144; PubMed Central PMCID: PMCPMC3254252.
3. Ly BH, Nguyen NP, Vinh-Hung V, Rapiti E, Vlastos G. Loco-regional treatment in metastatic breast cancer patients: is there a survival benefit? Breast Cancer Res Treat. 2010;119(3):537-45. Epub 2009/10/31. doi: 10.1007/s10549-009-0610-z. PubMed PMID: 19876731.
4. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-updagger. Ann Oncol. 2019;30(8):1194–220. Epub 2019/06/05. doi: 10.1093/annonc/mdz173. PubMed PMID: 31161190.
5. Baskar R, Dai J, Wenlong N, Yeo R, Yeoh KW. Biological response of cancer cells to radiation treatment. Front Mol Biosci. 2014;1:24. Epub 2014/01/01. doi: 10.3389/fmolb.2014.00024. PubMed PMID: 25988165; PubMed Central PMCID: PMCPMC4429645.
6. Huang RX, Zhou PK. DNA damage response signaling pathways and targets for radiotherapy sensitization in cancer. Signal Transduct Target Ther. 2020;5(1):60. Epub 2020/05/02. doi: 10.1038/s41392-020-0150-x. PubMed PMID: 32355263; PubMed Central PMCID: PMCPMC7192953.
7. Moding EJ, Kastan MB, Kirsch DG. Strategies for optimizing the response of cancer and normal tissues to radiation. Nat Rev Drug Discov. 2013;12(7):526–42. Epub 2013/07/03. doi: 10.1038/nrd4003. PubMed PMID: 23812271; PubMed Central PMCID: PMCPMC3906736.
8. Gray M, Turnbull AK, Ward C, Meehan J, Martinez-Perez C, Bonello M, et al. Development and characterisation of acquired radioresistant breast cancer cell lines. Radiat Oncol. 2019;14(1):64. Epub 2019/04/17. doi: 10.1186/s13014-019-1268-2. PubMed PMID: 30987655; PubMed Central PMCID: PMCPMC6466735.
9. Phillips TM, McBride WH, Pajonk F. The response of CD24(-/low)/CD44+ breast cancer-initiating cells to radiation. J Natl Cancer Inst. 2006;98(24):1777–85. Epub 2006/12/21. doi: 10.1093/jnci/djj495. PubMed PMID: 17179479.
10. Pajonk F, Vlashi E, McBride WH. Radiation resistance of cancer stem cells: the 4 R's of radiobiology revisited. Stem Cells. 2010;28(4):639–48. Epub 2010/02/06. doi: 10.1002/stem.318. PubMed PMID:

20135685; PubMed Central PMCID: PMCPMC2940232.

11. Sato K, Shimokawa T, Imai T. Difference in Acquired Radioresistance Induction Between Repeated Photon and Particle Irradiation. Front Oncol. 2019;9:1213. Epub 2019/12/05. doi: 10.3389/fonc.2019.01213. PubMed PMID: 31799186; PubMed Central PMCID: PMCPMC6863406.
12. Kyndi M, Sorensen FB, Knudsen H, Overgaard M, Nielsen HM, Overgaard J, et al. Estrogen receptor, progesterone receptor, HER-2, and response to postmastectomy radiotherapy in high-risk breast cancer: the Danish Breast Cancer Cooperative Group. J Clin Oncol. 2008;26(9):1419–26. Epub 2008/02/21. doi: 10.1200/JCO.2007.14.5565. PubMed PMID: 18285604.
13. Wu SG, He ZY, Li Q, Li FY, Lin Q, Lin HX, et al. Predictive value of breast cancer molecular subtypes in Chinese patients with four or more positive nodes after postmastectomy radiotherapy. Breast. 2012;21(5):657–61. Epub 2012/07/28. doi: 10.1016/j.breast.2012.07.004. PubMed PMID: 22835918.
14. Liu FF, Shi W, Done SJ, Miller N, Pintilie M, Voduc D, et al. Identification of a Low-Risk Luminal A Breast Cancer Cohort That May Not Benefit From Breast Radiotherapy. J Clin Oncol. 2015;33(18):2035–40. Epub 2015/05/13. doi: 10.1200/JCO.2014.57.7999. PubMed PMID: 25964246.
15. Sjostrom M, Lundstedt D, Hartman L, Holmberg E, Killander F, Kovacs A, et al. Response to Radiotherapy After Breast-Conserving Surgery in Different Breast Cancer Subtypes in the Swedish Breast Cancer Group 91 Radiotherapy Randomized Clinical Trial. J Clin Oncol. 2017;35(28):3222–9. Epub 2017/08/02. doi: 10.1200/JCO.2017.72.7263. PubMed PMID: 28759347.
16. Goodwin PM. Breast Cancer Biologic Subtype Does Not Predict Radiotherapy Benefit. Oncology Times. 2019;41:33. doi: 10.1097/01.cot.0000558227.52108.74.
17. Thomssen C, Balic M, Harbeck N, Gnant M. St. Gallen/Vienna 2021: A Brief Summary of the Consensus Discussion on Customizing Therapies for Women with Early Breast Cancer. Breast Care (Basel). 2021;16(2):135-43. Epub 2021/05/19. doi: 10.1159/000516114. PubMed PMID: 34002112; PubMed Central PMCID: PMCPMC8089428.
18. Zhang N, Zhang J, Zhang H, Liu Y, Zhao W, Wang L, et al. Individualized Prediction of Survival Benefit from Postmastectomy Radiotherapy for Patients with Breast Cancer with One to Three Positive Axillary Lymph Nodes. The Oncologist. 2019;24:1286–93. doi: 10.1634/theoncologist.2019-0124.
19. Sjöström M, Chang SL, Fishbane N, Davicioni E, Zhao SG, Hartman L, et al. Clinicogenomic Radiotherapy Classifier Predicting the Need for Intensified Locoregional Treatment After Breast-Conserving Surgery for Early-Stage Breast Cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2019;37:3340–9. doi: 10.1200/JCO.19.00761. PubMed PMID: 31618132.
20. Torres-Roca JF. A molecular assay of tumor radiosensitivity: a roadmap towards biology-based personalized radiation therapy. Per Med. 2012;9(5):547–57. Epub 2012/10/30. doi: 10.2217/pme.12.55. PubMed PMID: 23105945; PubMed Central PMCID: PMCPMC3480204.

21. Eschrich SA, Fulp WJ, Pawitan Y, Foekens JA, Smid M, Martens JW, et al. Validation of a radiosensitivity molecular signature in breast cancer. Clin Cancer Res. 2012;18(18):5134–43. Epub 2012/07/27. doi: 10.1158/1078-0432.CCR-12-0891. PubMed PMID: 22832933; PubMed Central PMCID: PMCPMC3993974.
22. Tramm T, Mohammed H, Myhre S, Kyndi M, Alsner J, Borresen-Dale AL, et al. Development and validation of a gene profile predicting benefit of postmastectomy radiotherapy in patients with high-risk breast cancer: a study of gene expression in the DBCG82bc cohort. Clin Cancer Res. 2014;20(20):5272–80. Epub 2014/08/26. doi: 10.1158/1078-0432.CCR-14-0458. PubMed PMID: 25149560.
23. Cui Y, Li B, Pollom EL, Horst KC, Li R. Integrating Radiosensitivity and Immune Gene Signatures for Predicting Benefit of Radiotherapy in Breast Cancer. Clin Cancer Res. 2018;24(19):4754–62. Epub 2018/06/21. doi: 10.1158/1078-0432.CCR-18-0825. PubMed PMID: 29921729; PubMed Central PMCID: PMCPMC6168425.
24. Speers C, Zhao S, Liu M, Bartelink H, Pierce LJ, Feng FY. Development and validation of a novel radiosensitivity signature in human breast cancer. Clinical Cancer Research. 2015;21:3667–77. doi: 10.1158/1078-0432.CCR-14-2898.
25. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17. Epub 2015/03/10. doi: 10.1016/j.csbj.2014.11.005. PubMed PMID: 25750696; PubMed Central PMCID: PMCPMC4348437.
26. Gupta S, Tran T, Luo W, Phung D, Kennedy RL, Broad A, et al. Machine-learning prediction of cancer survival: A retrospective study using electronic administrative records and a cancer registry. BMJ Open. 2014;4:4007. doi: 10.1136/bmjopen-2013-004007. PubMed PMID: 24643167.
27. Ramroach S, Joshi A, John M. Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers. Mol Omics. 2020;16(2):113–25. Epub 2020/02/26. doi: 10.1039/c9mo00198k. PubMed PMID: 32095794.
28. Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer. 2012;15(2):230–8. Epub 2012/07/19. doi: 10.4048/jbc.2012.15.2.230. PubMed PMID: 22807942; PubMed Central PMCID: PMCPMC3395748.
29. Liao C, Li S, Luo Z, editors. Gene Selection Using Wilcoxon Rank Sum Test and Support Vector Machine for Cancer Classification2007; Berlin, Heidelberg: Springer Berlin Heidelberg.
30. Nimeus-Malmstrom E, Krogh M, Malmstrom P, Strand C, Fredriksson I, Karlsson P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence after breast-conservation surgery, with or without postoperative radiotherapy. Breast Cancer Res. 2008;10(2):R34. Epub 2008/04/24. doi: 10.1186/bcr1997. PubMed PMID: 18430221; PubMed Central PMCID: PMCPMC2397536.
31. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun.

2016;7:11479. Epub 2016/05/11. doi: 10.1038/ncomms11479. PubMed PMID: 27161491; PubMed Central PMCID: PMCPMC4866047.

32. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4. Epub 2012/05/17. doi: 10.1158/2159-8290.CD-12-0095. PubMed PMID: 22588877; PubMed Central PMCID: PMCPMC3956037.
33. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52. Epub 2012/04/24. doi: 10.1038/nature10983. PubMed PMID: 22522925; PubMed Central PMCID: PMCPMC3440846.
34. Kuhn M. Building Predictive Models in R Using the caret Package. 2008. 2008;28(5):26. Epub 2008-09-23. doi: 10.18637/jss.v028.i05.
35. Jain AK, Jianchang M, Mohiuddin KM. Artificial neural networks: a tutorial. Computer. 1996;29(3):31–44. doi: 10.1109/2.485891.
36. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications. 1998;13(4):18–28. doi: 10.1109/5254.708428.
37. Mucherino A, Papajorgji PJ, Pardalos PM. k-Nearest Neighbor Classification. Data Mining in Agriculture. New York, NY: Springer New York; 2009. p. 83–106.
38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
39. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Mach Learn. 1997;29(2–3):131–63. doi: 10.1023/a:1007465528199.
40. Quinlan JR. Induction of Decision Trees. Mach Learn. 1986;1(1):81–106. doi: 10.1023/a:1022643204877.
41. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. doi: 10.1023/a:1010933404324.
42. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer; 2009.
43. Eschrich SA, Pramana J, Zhang H, Zhao H, Boulware D, Lee JH, et al. A Gene Expression Model of Intrinsic Tumor Radiosensitivity: Prediction of Response and Prognosis After Chemoradiation. International Journal of Radiation Oncology Biology Physics. 2009;75:489–96. doi: 10.1016/j.ijrobp.2009.06.014.
44. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941-D7. Epub 2018/10/30. doi: 10.1093/nar/gky1015. PubMed PMID: 30371878; PubMed Central PMCID: PMCPMC6323903.
45. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. 2019;18(2):623–32. Epub 2018/11/20. doi:

10.1021/acs.jproteome.8b00702. PubMed PMID: 30450911; PubMed Central PMCID: PMCPMC6800166.

46. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. Epub 2003/11/05. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMCPMC403769.
47. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nature Methods. 2016;13(9):703–4. doi: 10.1038/nmeth.3968.
48. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage. 2018;180(Pt A):68–77. Epub 2017/06/29. doi: 10.1016/j.neuroimage.2017.06.061. PubMed PMID: 28655633.
49. Ying X. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 2019;1168:022022. doi: 10.1088/1742-6596/1168/2/022022.
50. Bartelink H, Maingon P, Poortmans P, Weltens C, Fourquet A, Jager J, et al. Whole-breast irradiation with or without a boost for patients treated with breast-conserving surgery for early breast cancer: 20-year follow-up of a randomised phase 3 trial. Lancet Oncol. 2015;16(1):47–56. Epub 2014/12/17. doi: 10.1016/S1470-2045(14)71156-8. PubMed PMID: 25500422.
51. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. Epub 2011/03/08. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.
52. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70. Epub 2012/09/25. doi: 10.1038/nature11412. PubMed PMID: 23000897; PubMed Central PMCID: PMCPMC3465532.
53. Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol. 2017;3(12):1654–62. Epub 2017/05/05. doi: 10.1001/jamaoncol.2017.0595. PubMed PMID: 28472234; PubMed Central PMCID: PMCPMC5671371.
54. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. Epub 2019/10/31. doi: 10.1186/s12916-019-1426-2. PubMed PMID: 31665002; PubMed Central PMCID: PMCPMC6821018.

Figures

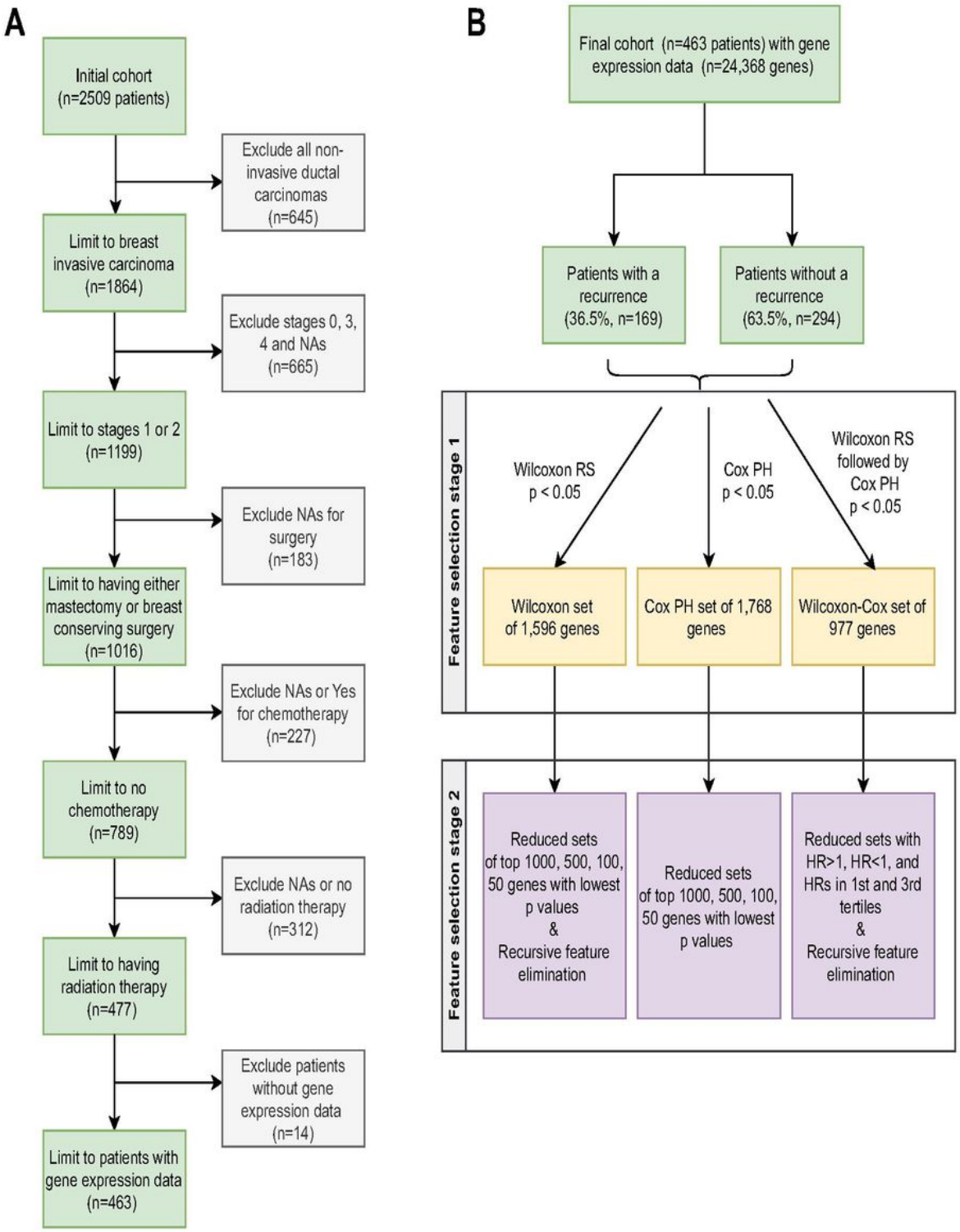

Figure 1

Flowcharts of A) patient selection criteria for the METBRIC cohort; and B) the feature selection methods used to select the genes to train ML algorithms to predict relapse status

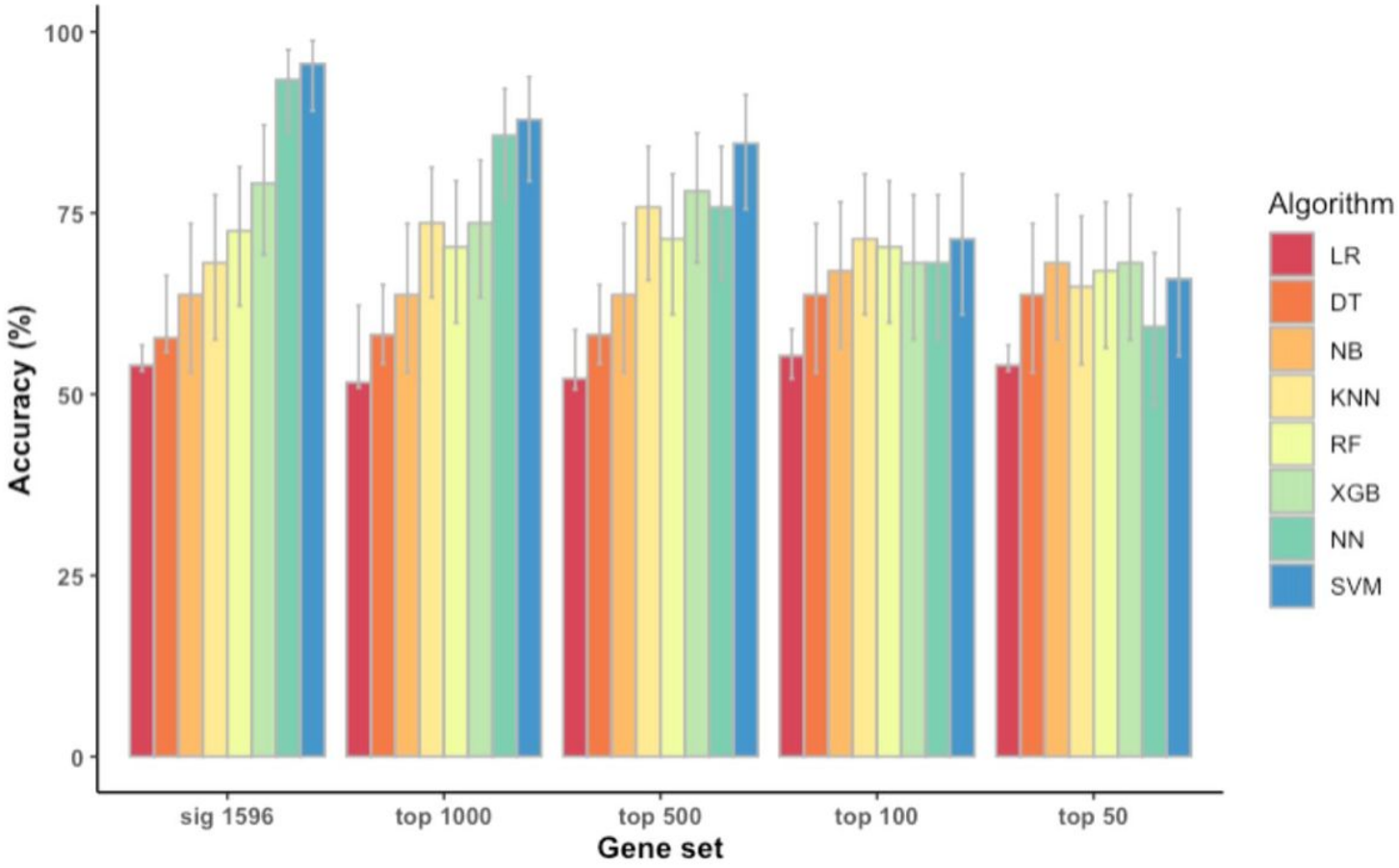

Figure 2

Classification accuracies of eight machine learning algorithms trained using the full Wilcoxon set of 1,596 genes and the top 1000, 500, 100 and 50 genes with the lowest p values from this set.

Logistic regression (LR); Decision trees (DT); Naïve bayes (NB); K nearest neighbours (KNN); Random forest (RF); XGBoost (XGB); Neural Networks (NN); Support Vector Machine (SVM)

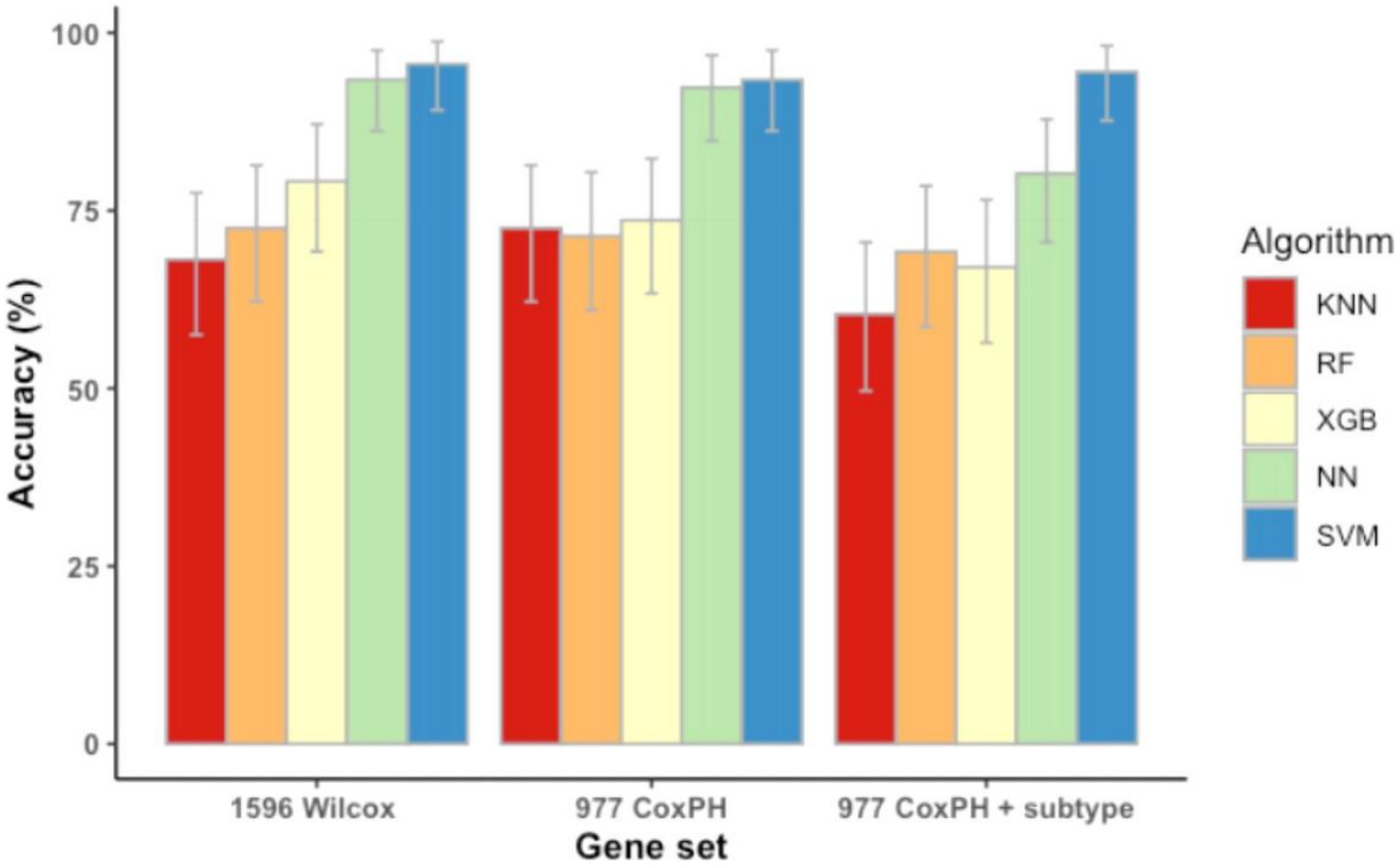

Figure 3

Classification accuracies of the top five machine learning algorithms trained using the Wilcoxon set of 1,596 genes and the Wilcoxon-Cox set of 977 genes with and without the subtype variable included.

Logistic regression (LR); Decision trees (DT); Naïve bayes (NB); K nearest neighbours (KNN); Random forest (RF); XGBoost (XGB); Neural Networks (NN); Support Vector Machine (SVM)

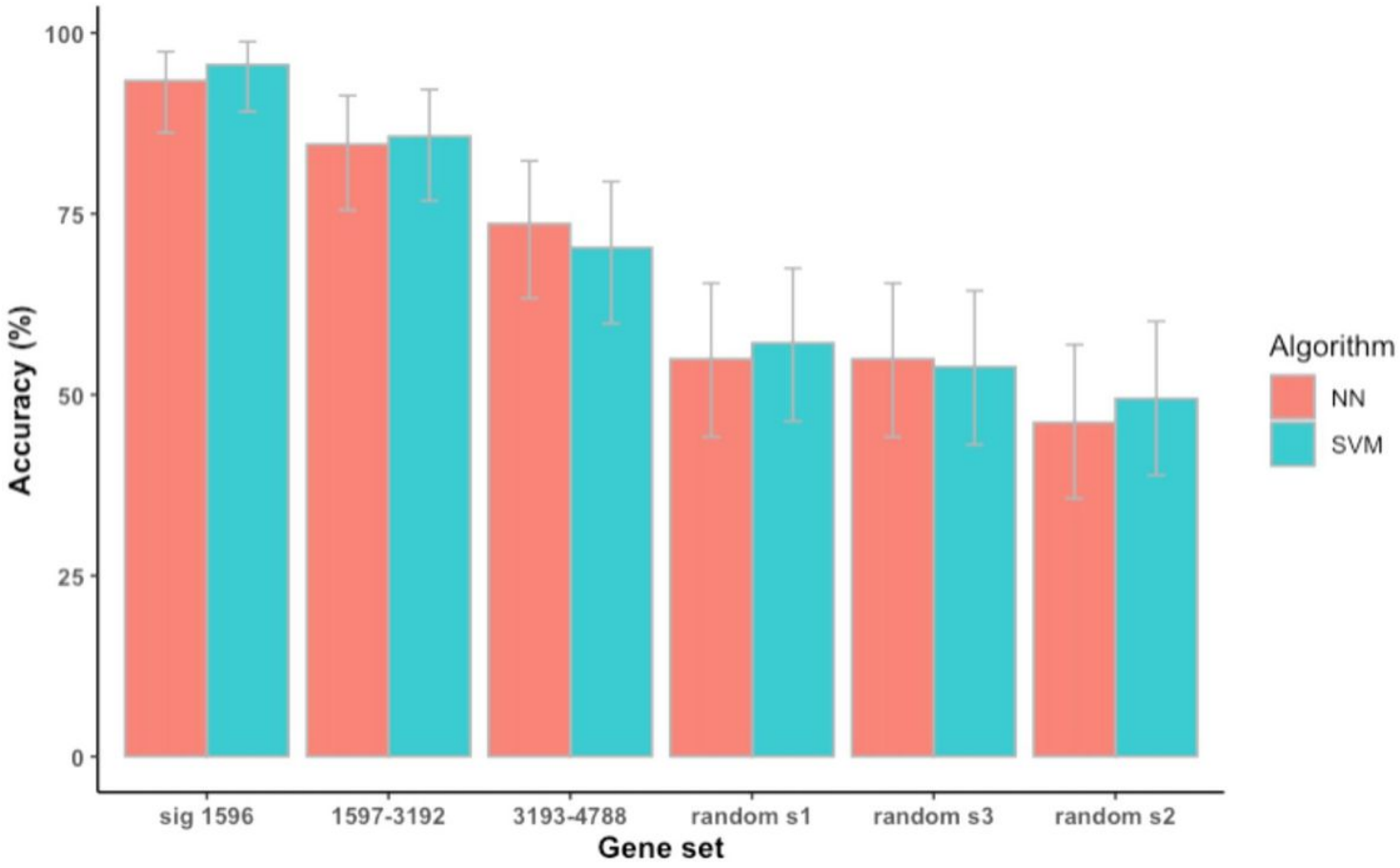

Figure 4

Classification accuracies when the Wilcoxon set of 1,596 genes and subsets of insignificant genes were used for training.

Logistic regression (LR); Decision trees (DT); Naïve bayes (NB); K nearest neighbours (KNN); Random forest (RF); XGBoost (XGB); Neural Networks (NN); Support Vector Machine (SVM); s1-s3, subsets 1-3

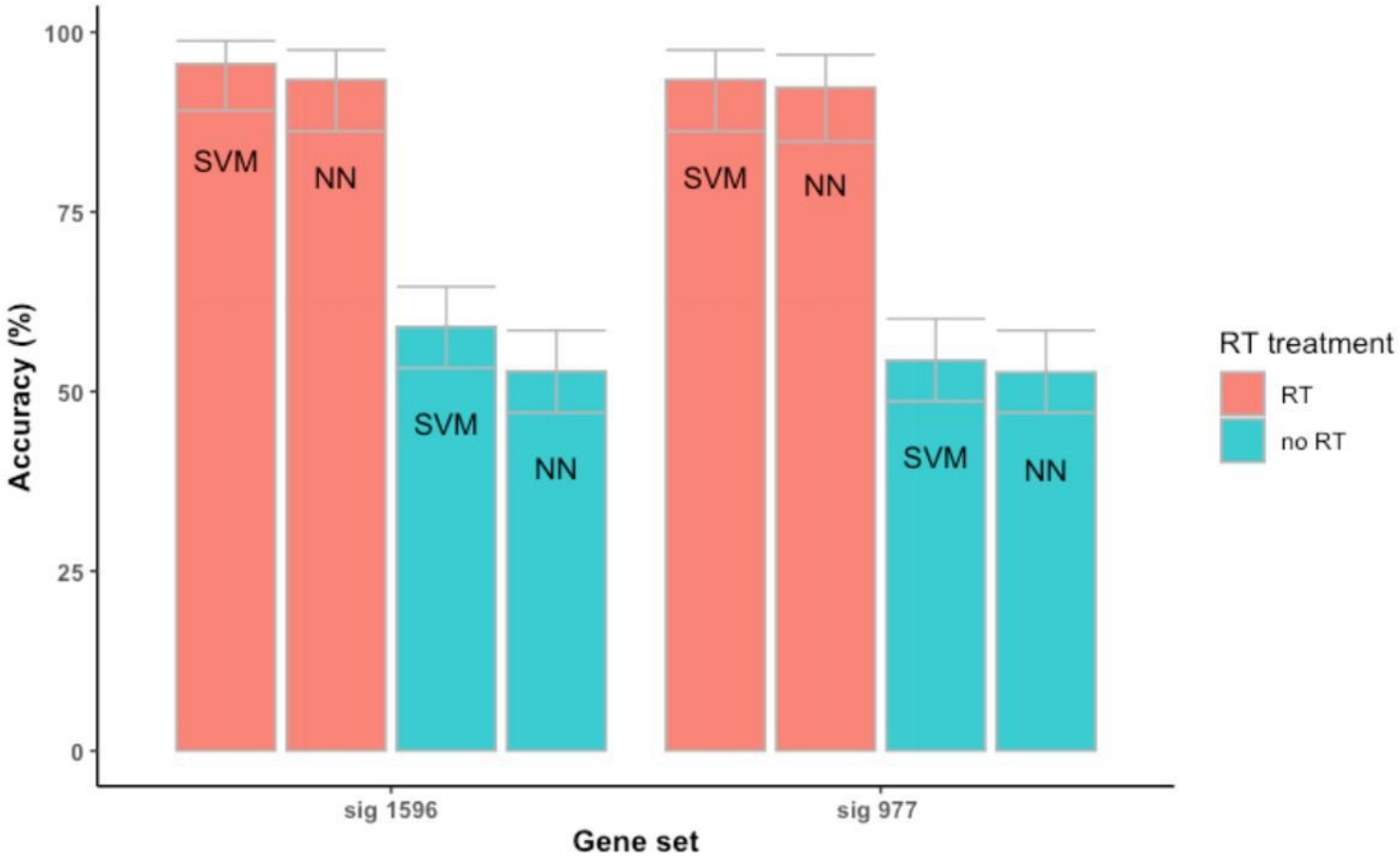

Figure 5

Classification accuracy comparison of the SVM and NN models in RT-treated and untreated patients using the Wilcoxon set and the Wilcoxon-Cox set.

SVM, support vector machine; NN, neural networks

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1SupplemntaryRCode.docx
- Additionalfile2Supplementarytablesandfigures.docx