ARTIFICIAL INTELLIGENCE

Machine Learning Frameworks to Predict Neoadjuvant Chemotherapy Response in Breast Cancer Using Clinical and Pathological Features

Nicholas Meti, MD[1,2]; Khadijeh Saednia, MSc[3]; Andrew Lagree, MSc[4]; Sami Tabbarah, MSc[4]; Majid Mohebpour, PhD[4]; Alex Kiss, PhD[5]; Fang-I Lu, MD[6]; Elzbieta Slodkowska, MD[6]; Sonal Gandhi, MD, MSc[1,7]; Katarzyna Joanna Jerzak, MD, MSc[1,7]; Lauren Fleshner, BSc (c)[4]; Ethan Law, BSc (c)[4]; Ali Sadeghi-Naini, PhD[2,3,4]; and William T. Tran, MRT(T), MSc, PhD[2,4,8]

PURPOSE Neoadjuvant chemotherapy (NAC) is used to treat locally advanced breast cancer (LABC) and high-risk early breast cancer (BC). Pathological complete response (pCR) has prognostic value depending on BC subtype. Rates of pCR, however, can be variable. Predictive modeling is desirable to help identify patients early who may have suboptimal NAC response. Here, we test and compare the predictive performances of machine learning (ML) prediction models to a standard statistical model, using clinical and pathological data.

METHODS Clinical and pathological variables were collected in 431 patients, including tumor size, patient demographics, histological characteristics, molecular status, and staging information. A standard multivariable logistic regression (MLR) was developed and compared with five ML models: k-nearest neighbor classifier, random forest (RF) classifier, naive Bayes algorithm, support vector machine, and multilayer perceptron model. Model performances were measured using a receiver operating characteristic (ROC) analysis and statistically compared.

RESULTS MLR predictors of NAC response included: estrogen receptor (ER) status, human epidermal growth factor-2 (HER2) status, tumor size, and Nottingham grade. The strongest MLR predictors of pCR included HER2+ versus HER2– BC (odds ratio [OR], 0.13; 95% CI, 0.07 to 0.23; $P < .001$) and Nottingham grade G3 versus G1-2 (G1-2: OR, 0.36; 95% CI, 0.20 to 0.65; $P < .001$). The area under the curve (AUC) for the MLR was AUC = 0.64. Among the various ML models, an RF classifier performed best, with an AUC = 0.88, sensitivity of 70.7%, and specificity of 84.6%, and included the following variables: menopausal status, ER status, HER2 status, Nottingham grade, tumor size, nodal status, and presence of inflammatory BC.

CONCLUSION Modeling performances varied between standard versus ML classification methods. RF ML classifiers demonstrated the best predictive performance among all models.

JCO Clin Cancer Inform 5:66-80. © 2021 by American Society of Clinical Oncology

INTRODUCTION

Neoadjuvant (preoperative) chemotherapy (NAC) is used to treat locally advanced breast cancer (LABC), defined by tumor size (ie, T3 or T4), skin, or chest wall involvement, and may demonstrate fixed regional lymph nodes.[1] In LABC, NAC can help facilitate tumor downstaging for resection or breast conservation and potentially treat micrometastatic disease in high-risk subtypes earlier. The desired end point is, in part, a complete eradication of invasive disease, known as pathological complete response (pCR). NAC is also increasingly administered to patients with high-risk, early-stage breast cancer (BC) with triple-negative or human epidermal growth factor-2 (HER2+) subtypes.[2] In these two patient subtypes, the presence of residual disease at surgery (ie, non-pCR) predicts for benefit from escalation of therapy with additional systemic chemotherapy in the postoperative period.[3] However, rates of pCR vary greatly (7%-75%) by receptor status and other initial features.[4] Overall, significant residual disease burden after NAC, particularly in lymph nodes, is associated with poor outcomes for most BC subtypes, although the correlation with pCR and long-term outcomes is more robust for specific subtypes (eg, hormone negative or HER2+). Nonetheless, rates of pCR are important to optimize overall, and predicting patient likelihood of achieving pCR is an important area of research.[5]

There have been substantial efforts to develop robust prediction models that can help forecast the likelihood of pCR versus non-pCR patients a priori[6-8]; this is important to help better select different types of NAC, clinical trials, or alternative treatment modalities for patients who are not responding or unlikely to respond to planned NAC. This can not only help increase pCR rates but also prevent unnecessary toxicities in patients from starting or continuing NAC that is unlikely to work. Modeling frameworks exploit either canonical

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on October 21, 2020 and published at ascopubs.org/journal/cci on January 13, 2021: DOI https://doi.org/10.1200/CCI.20.00078

CONTEXT

Key Objective

To evaluate and compare statistical and machine learning (ML) models using clinical and pathological variables, to predict pathological complete response (pCR) in patients treated with neoadjuvant chemotherapy (NAC).

Knowledge Generated

Clinical and pathological variables are predictive of NAC response when used in ML models. The results of this study reveal variability in predictive performances among several ML classifiers to predict the likelihood of pCR in NAC-treated patients. We demonstrate that random forest ML classifiers yield the greatest accuracies among all models tested.

Relevance

As NAC continues to play a central role in the management of patients with locally advanced breast cancer (BC) and high-risk BC, there is a need to predict NAC response upfront, to improve our ability to preselect patients for tailored therapies. Emerging research in computational oncology has the potential to improve our point-of-care clinical decision making in oncology.

statistical approaches or artificial intelligence (AI) using machine learning (ML) classification algorithms. Prediction models have been explored using data extracted from breast imaging, genomics, or clinical and pathological (ie, clinicopathological) information. Indeed, the use of clinicopathological variables is readily accessible from patient-care documentation platforms, which is convenient for the clinician. Clinicopathological parameters, including patient age, Nottingham grade, histological type, molecular status, and tumor size, are known potential predictors of NAC response and have been used to develop nomograms that predict chemotherapy response outcomes.[9,10] Several studies have used multivariable logistic regression (MLR) models to predict pCR and the associated survival outcomes.[11-13]

Despite thorough investigations on the use of clinicopathological biomarkers using standard statistical modeling, there is, nevertheless, immense interest in using ML techniques for biomarker analysis in BC. Major advantages of ML over standard statistical models such as logistic regression (LR) include improved capabilities for handling high-dimensional data sets, overcoming distribution biases, and addressing problems with missing datapoints.[14] Although ML tools are not new, most recent efforts have been focused on addressing diagnostic classification problems in BC, risk (epidemiological) modeling, and radiomic analysis obtained from emerging quantitative breast imaging.[15-17] However, ML techniques remain elusive for routine clinical implementation. This may be due to the diversity of available ML algorithms with variances in tuning classifiers, selecting optimal hyperparameters, and retaining feature sets (ie, dimensionality reduction) that best represent the target population. To the best of our knowledge, there are limited reports comparing predictive ML algorithms to standard statistical frameworks within the context of BC and NAC response. Deeper insight into the underlying data patterns and relationships from ML may subsequently enhance prediction tasks and lead to improvements in developing robust decision-support tools in the oncology clinic.

Here, we evaluated and compared multiple classification algorithms; the predictive performance of a standard statistical model (MLR) was challenged by five ML classifiers. ML models were selected on the basis of their popularity in previous classification studies.[18-20] Diagnostic clinicopathological variables (features) were collected from 431 patients with BC who received NAC. We report and contrast the performance of ML classifiers to predict pCR versus non-pCR after NAC a priori.

METHODS

Patient Population and Data Collection

This was a retrospective single-institution study. The study was approved by the institutional review board prior to data collection. Data were extracted from 431 patients with biopsy-confirmed BC who were treated with anthracycline- and taxane-based NAC between 2013 and 2018. The study cohort was not a consecutive sample; patients were excluded from analysis for the following reasons: (1) progressive disease requiring salvage therapies (eg, neoadjuvant chemoradiation), (2) incomplete clinicopathological data reporting (ie, diagnostic workup performed at an outside institution), (3) nonstandard polychemotherapy or trial agents, (4) treatment noncompliance or incomplete course of NAC. Only women with unifocal unilateral BC and nonmetastatic (distant) disease at the time of presentation were included in the study cohort. All data were collected from the institutional electronic medical record.

Standard-of-care clinical and pathological features were collected for all patients and there were no missing data for any sample. Data included patient age (> 18 years), menopausal status (pre/post), clinical tumor size (largest radiologically reported dimension in millimeters), clinical nodal status (TNM; confirmed by fine-needle aspiration), Nottingham grade (G1/G2/G3), presence or absence of

inflammatory cancer (defined as breast carcinoma with dermal lymphatic invasion), and histological type (ductal *v* lobular). Estrogen receptor (ER%) status, progesterone receptor (PR%) status, HER2 status (+/–) were also included in the feature set (ER, PR, and HER2 by immunohistochemistry [IHC], and equivocal HER2 status confirmed by in situ hybridization [ISH or fluorescent in situ hybridization]). Other IHC markers, such as Ki-67 and p53 expression, were not included in the analysis since they were not routinely assessed as part of the standard of care. The neoadjuvant drug regimen was recorded for analysis (either dose-dense doxorubicin, cyclophosphamide, and paclitaxel [ddAC-T], or fluorouracil, epirubicin, cyclophosphamide, and docetaxel [FEC-D]). All patients with HER2+ BCs were given trastuzumab anti-HER2 therapy during taxane chemotherapy, which was included in this data set. Other anti-HER2 targeted drug therapies such as pertuzumab were not included in the analysis, as this was not offered as standard first-line therapy at the treating institution and within the publicly funded health system.

All cases were evaluated using the American Joint Committee on Cancer (AJCC) TNM staging guideline, eighth edition,[21] where applicable. To enhance the interpretation of features, continuous and ordinal variables such as age, ER%, PR%, tumor size, and nodal status (ie, primary markers) were also converted into composite biomarkers as dichotomous categorical features. These included: age; ≥ 50 *v* < 50, ER +/– (+ defined as $> 1\%$ expression), PR +/– (+ defined as $> 1\%$ expression), AJCC T (tumor)-stage, and nodal status +/–. Parallel analysis was conducted on the two features sets. Specifically, feature set 1 included continuous, categorical, and ordinal variables, whereas feature set 2 only included categorical, dichotomous, or ordinal data. The feature sets and experiments conducted are summarized in the Appendix Figure A1.

For ground truth labeling, each patient case was classified into a discrete class label (ie, pCR *v* non-pCR). A standard assessment method using the residual cancer burden index (RCBI)[22] was employed. An RCBI score of 0 (ie, pCR) was defined as the absence of residual invasive and nodal disease.[22] Patients who demonstrated residual disease were classified as non-pCR (ie, RCBI > 0). All pathology reviews (pretreatment histopathology and post-NAC synoptic pathology) were evaluated by board-certified breast pathologists and as part of the patient's standard of care. Similarly, radiological reporting was conducted at the time of diagnosis by board-certified breast radiologists.

Statistical Analysis

Variables were summarized using descriptive statistics. Data variances were calculated and reported as standard deviations. The data distribution was tested for normality violations using a Shapiro-Wilk test. Individual parameters were compared between pCR versus non-pCR groups. For categorical variables, a Pearson χ^2 test was conducted to test for significant differences between groups. Continuous predictive variables were tested using an independent two-tailed *t*-test for normally distributed samples. The significance level was set to an alpha of .05.

This investigation aimed to compare a standard statistical model to ML classifiers. To achieve this, an MLR model (ie, baseline model) was developed for each feature set. Features were first tested for multicollinearity. To reduce redundant features, clinical and pathological parameters that demonstrated a correlation coefficient of $r > 0.7$ were identified within the feature set; the feature with the strongest correlation to the classification variable (pCR/non-pCR) was retained for MLR modeling. Clinicopathological variables were selected for MLR modeling on the basis of a forward stepwise approach, using an entry probability of .05 and a retention probability of .10. Multivariable estimates of the odds ratio (OR), the 95% CI, and significance levels were calculated for features. Statistical analyses were conducted using SPSS Ver.24 (IBM Corp, Chicago, IL).

Machine Learning Classifiers

Machine learning models were developed as comparative models. First, data were partitioned into a training set (75%) and an independent test set (25%). There were 323 patients used for training and validation of the models, and 108 patients were selected as an independent test set (ie, unseen data set). Sufficient class representations were ensured within each fold of the cross-validation during the training phase (ie, ensuring equal balance between pCR and non-pCR within each fold) by upsampling using SMOTE (synthetic minority over-sampling technique).[23] This was to minimize performance biases during model training with the unbalanced data set applied in this study. A sequential forward feature selection (SFFS) approach was used to select the most discriminative features for subsequent modeling. The average accuracy in a five-fold cross-validation on the training data was used as the criteria of SFFS. For all prediction models, we used Harrell's rule to reduce risk of overfitting models,[24] based on the class with the least samples in the training set (ie, pCR; n = 79 in the training set). This rule was applied to mitigate model overfitting and address the peaking phenomenon or classification performance plateauing. Thus, a maximum of eight features were permitted for inclusion into the final model prior to testing. Five ML models were used, which included, (1) k-nearest neighbor (k-NN), (2) random forest (RF), (3) naive Bayes (NB), (4) support vector machine (SVM), and (5) multilayer perceptron. The modeling frameworks (eg, hyperparameters and tuning) are outlined in the Appendix Table A1.

Model Performance Measures

Performance evaluation for all models included the receiver operating characteristic (ROC) analysis on the test set, and the area under the curve (AUC), prediction accuracy

(Acc%), sensitivity (Sn%), specificity (Sp%), positive predictive value (PPV%), and negative predictive value (NPV %) were reported. Models' performances were compared; tests of significant differences between AUCs were conducted and *P* values were reported, using methods previously described by Hanley and McNeil.[25] A *P* value of $< .05$ was used to indicate significant differences between ROC curves.

RESULTS

Patient Population Characteristics

The Appendix Table A2 and Appendix Figure A2 summarizes clinical, treatment, and pathological details of the study cohort. There were 431 patients included in this retrospective study; 105 patients (24.3%) derived a pCR versus 326 patients (75.6%) who exhibited residual disease (non-pCR) following treatment. There were insignificant differences in univariate analysis between patients who achieved a pCR versus those who did not for the following variables: age, menopausal status, NAC, presence of inflammatory BC, and tumor laterality (not included in modeling analysis).

Overall, predictors for pCR/non-pCR among all models included receptor status (ER, PR, and HER2), tumor grade (Nottingham), tumor size, and nodal status. We also grouped patients into molecular subtypes (ie, luminal A, luminal B, HER2+, and triple-negative BC [TNBC]) based on previous methods,[2] as Ki-67 expression data were not available for analysis. Molecular subtypes were only used for descriptive purposes; this was important to assess if there were sufficient group representations compared with clinical observations as molecular subtypes have varying response rates.[5] The luminal B subtype constituted the largest proportion (40.3%) of the cohort (Appendix Fig 2C).

MLR Model

The MLR served as a baseline statistical model. In multivariable analysis, significant predictors (primary and composite biomarkers) included (1) ER status, (2) HER2 status, (3) Nottingham grade, and (4) presenting tumor size. The MLR demonstrated that an increase in ER positivity resulted in decreased pCR rates; for ER as a continuous variable (ER%), OR = 0.98; 95% CI, 0.97 to 0.99 ($P < .001$) and for ER as a categorical variable (ER +/–), OR = 0.19; 95% CI, 0.11 to 0.34 ($P < .001$). Lower Nottingham grade (ie, grade 1-2) showed an increased probability of non-pCR; the OR = 0.36; 95% CI, 0.20 to 0.65 in the MLR containing continuous variables (ie, feature set 1) and the OR = 0.34; 95% CI, 0.19 to 0.61 in the MLR model with composite markers (ie, feature set 2). A summary of all clinicopathological features and associated ORs is presented in Figure 1.

The MLR used all four clinicopathological markers in model construction and showed moderate predictive performances

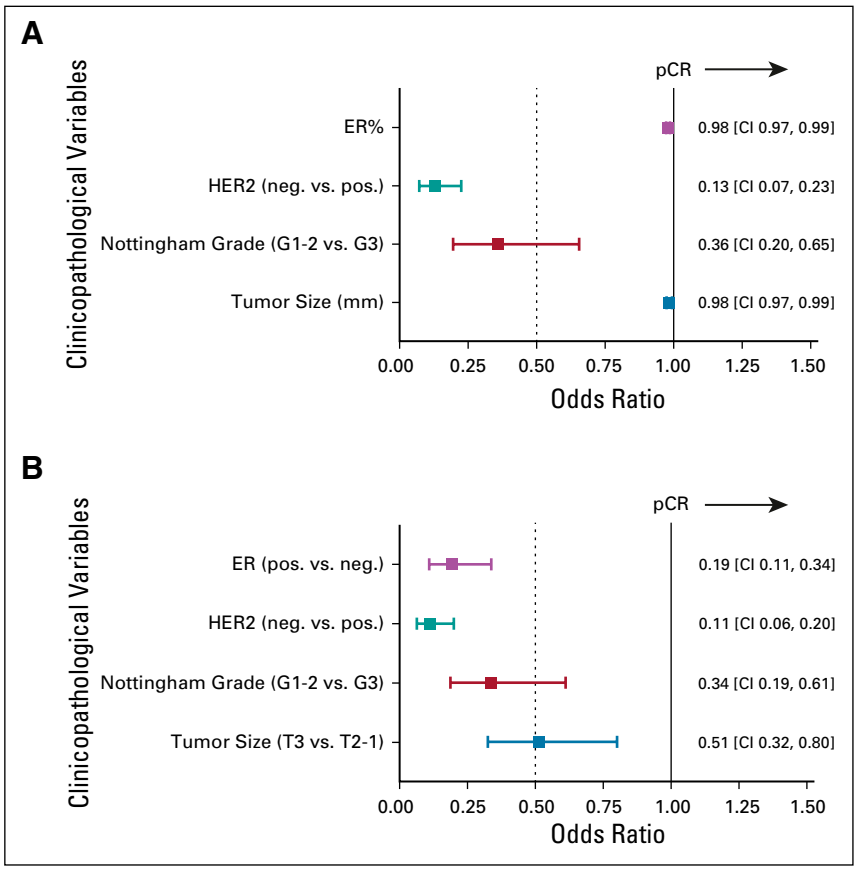

FIG 1. Odds ratio (OR) estimates (multivariable logistic regression). (A) A multivariable logistic regression model was constructed using continuous data for ER status and tumor size. A forward stepwise analysis was conducted to yield significant features in the model. The ORs demonstrate that an increase in ER%, HER2-negativity, Nottingham grade, and tumor size (mm) are associated with a decreased likelihood of achieving a pCR. (B) Parallel model development using categorical data for ER status and tumor size (AJCC). OR with 95% CI shown for each clinicopathological variable. Multivariable *P*-values were statistically significant ($P < .01$). AJCC, American Joint Committee on Cancer; ER, estrogen receptor; HER2, human epidermal growth factor-2; pCR, pathological complete response.

in the test set. The models that contained continuous predictors versus the MLR with composite markers demonstrated similar AUC values. The MLR model for continuous variables had an AUC of 0.63 (Sn% of 79.8%, Sp% of 57.9%, PPV% = 89.9%, and NPV% = 37.9%). The MLR model that constituted composite markers demonstrated an AUC of 0.64 (Sn% of 79.4%, Sp% of 81.2%, PPV% = 97.5, NPV% = 31.0%, and accuracy of 79.6% in the test data set). A summary of predictive performances is presented in Tables 1-2. The corresponding ROC curves are presented in Figures 2-3. A comparison test of models' performances demonstrated an insignificant difference between MLR ROCs (P = .980). Therefore, there was no significant impact on the predictive performances of the MLR model when using continuous versus categorical composite variables (Fig 4).

Classification Performance of ML Models

Different ML models were investigated to assess their performance and generalization ability on never-seen data in terms of response prediction using the clinical and pathological features. Specifically, five ML classification models were constructed to challenge the predictive MLR, in terms of comparing classification performances and significant biomarkers yielded. To address the issue of feature selection stability, all experiments were repeated six times to mitigate random effects for feature reduction and selection. Repeated experiments were also conducted to enhance reliability in the ML model training and model performances. Overall, there were varying levels of prediction accuracies among ML models. Using continuous clinicopathological variables, an accuracy between 58.3% and 74.1% was achieved on the independent test set (Table 1). Among the ML models, the k-NN classifier performed the worst, as measured by the AUC (AUC = 0.70), identifying menopausal status and histological type as meaningful predictors of pCR through cross-validation during model training. Conversely, the RF classifier showed the best classification test accuracy of 74.1% (AUC = 0.88) and included the following features: menopausal status, ER (%), HER2 (+/–), Nottingham grade (G1-3), tumor size (mm), nodal status (N0-N3), and inflammatory BC (+/–).

TABLE 1. Prediction Models and Performance Measures Based on Continuous, Dichotomous, and Ordinal Variables (Feature Set 1)

		Training Set	Test Set					
Model	**Feature Sets**	**Mean Acc (%) ± SD**	**Acc (%)**	**Sn (%)**	**Sp (%)**	**PPV (%)**	**NPV (%)**	**AUC**
LR	ER (%)	81.1 ± 3.1	75.9	79.8	57.9	89.9	37.9	0.63
	HER2 (+/–)							
	Nottingham grade							
	Tumor size (mm)							
k-NN	Menopausal status	79.3 ± 0.1	58.3	56.1	65.4	83.6	32.0	0.70
	Histological type							
RF	Menopausal status	84.9 ± 0.1	74.1	70.7	84.6	93.6	47.9	0.88
	ER (%)							
	HER2 (+/–)							
	Nottingham grade							
	Tumor size (mm)							
	Nodal status (TNM)							
	IBC (+/–)							
NB	ER (%)	80.6 ± 0.1	64.8	57.3	88.5	94.0	39.7	0.80
	HER2 (+/–)							
	Tumor size (mm)							
	Nodal status (TNM)							
SVM	NAC	75.4 ± 0.1	68.5	64.6	80.8	91.4	42.0	0.73
	HER2 (+/–)							
MLP	ER (%)	81.0 ± 0.1	67.6	62.2	84.6	92.7	41.5	0.79
	HER2 (+/–)							
	Tumor size (mm)							

Standard deviations shown for accuracies obtained from the training data set, based on a five-fold cross-validation.

Abbreviations: Acc (%), prediction accuracy; AUC, area under the curve; ER, estrogen receptor; HER2, human epidermal growth factor-2; IBC, inflammatory breast cancer; k-NN, k-nearest neighbor; LR, logistic regression; MLP, multilayer perceptron; NB, naive Bayes; NPV, negative predictive value; PR, progesterone receptor; RF, random forest; SD, standard deviation; Sn (%), sensitivity; Sp (%), specificity; SVM, support vector machine.

TABLE 2. Prediction Models and Performance Measures Based on Categorical, Dichotomous, and Ordinal Variables (ie, Composite Markers)

		Training Set	Test Set					
Model	**Feature Sets**	**Mean Acc (%) ± SD**	**Acc (%)**	**Sn (%)**	**Sp (%)**	**PPV (%)**	**NPV (%)**	**AUC**
LR	ER (+/–)	79.5 ± 2.4	79.6	79.4	81.2	97.5	31.0	0.64
	HER2 (+/–)							
	Nottingham grade							
	Tumor size (TNM)							
k-NN	PR (+/–)	83.6 ± 0.1	59.3	58.5	61.5	82.8	32.0	0.69
	HER2 (+/–)							
	Histological type							
	IBC (+/–)							
RF	ER (+/–)	85.3 ± 0.1	77.8	79.3	73.0	90.3	52.8	0.85
	PR (+/–)							
	HER2 (+/–)							
	Nottingham grade							
	Tumor size (TNM)							
	Nodal status (+/–)							
NB	PR (+/–)	78.5 ± 0.1	70.4	68.3	76.9	90.3	43.5	0.79
	HER2 (+/–)							
	Tumor size (TNM)							
	Nottingham grade							
SVM	ER (+/–)	80.6 ± 0.1	64.8	58.5	84.6	92.3	39.3	0.81
	HER2 (+/–)							
	Nodal status (+/–)							
MLP	ER (+/–)	80.0 ± 0.1	63.0	53.7	92.3	95.7	38.7	0.84
	HER2 (+/–)							
	Histological type							
	Nottingham grade							
	IBC (+/–)							

Models were constructed based on feature set 2; that is, composite markers set. All models were trained using a five-fold cross-validation. Standard deviations are shown for results from the five folds of the training set.

Abbreviations: Acc (%), prediction accuracy; AUC, area under the curve; ER, estrogen receptor; HER2, human epidermal growth factor-2; IBC, inflammatory breast cancer; k-NN, k-nearest neighbor; LR, logistic regression; MLP, multilayer perceptron; NB, naive Bayes; NPV, negative predictive value; PR, progesterone receptor; RF, random forest; SD, standard deviation; Sn (%), sensitivity; Sp (%), specificity; SVM, support vector machine.

Using composite categorical variables demonstrated similar trends. The k-NN model showed the worst classification predictions (accuracy of 59.3%, AUC = 0.69), whereas the RF algorithm showed an improvement in accuracy of 77.8% and an AUC = 0.85 (Table 2).

Statistical Comparison of Models

Statistically significant differences between AUCs of all models were evaluated. A summary of results is presented in Figure 4. In experiments using feature set 1, the ML model that demonstrated the most significantly different AUC to the MLR was the RF ($P = .024$). For feature set 2 (ie, composite variables), the RF model demonstrated the most significant difference to the MLR AUC ($P = .014$). A comparison between parallel models showed that the most significant difference was found between the AUCs of the MLR using feature set 1 and the RF using feature set 2 ($P = .008$). The associated predictive features for each model are reported in Tables 1-2.

DISCUSSION

In this study, we analyzed clinical and pathological data of 431 patients with BC and compared an MLR model against five ML classification algorithms in predicting pCR following NAC. The study aimed to compare modeling performances and classification frameworks to predict response to NAC among patients with BC a priori. We report that select ML classifiers, such as the RF classifier, can outperform MLR predictive modeling performances, as seen in the increases in AUCs (eg, $AUC_{MLR} = 0.63$ *v* $AUC_{RF} = 0.88$ using feature set 1; $P < .024$).

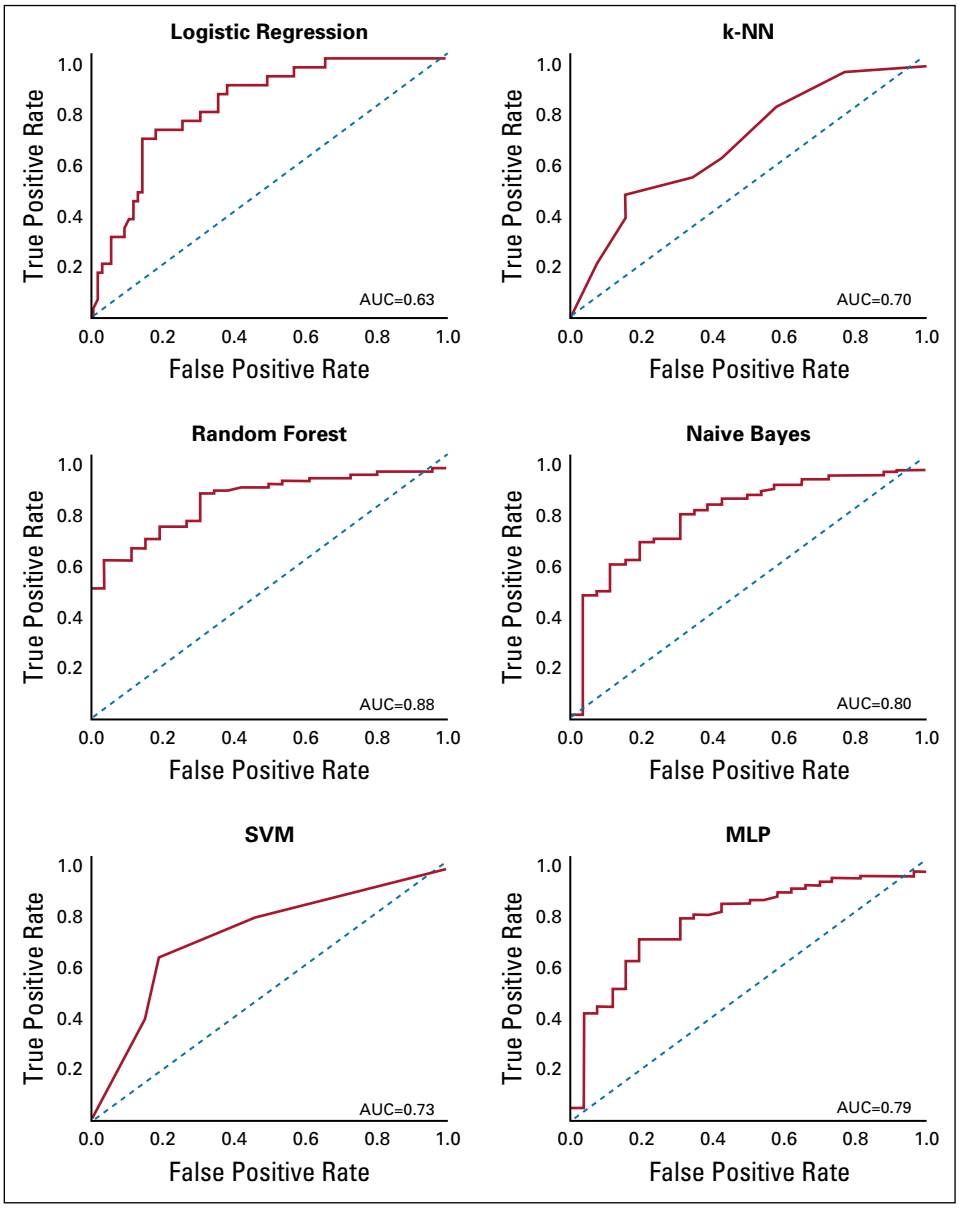

FIG 2. Representative ROC curves. Statistical and machine learning prediction models were constructed based on continuous, categorical, and ordinal variables (feature set 1). k-NN, k-nearest neighbor; MLP, multilayer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

In terms of clinical relevance, AI-driven prediction models can potentially help to better select NAC options for patients by determining the likelihood of NAC success before patients start therapy. The information provided by such models, if validated, can help improve patient treatment algorithms beyond standard clinical and pathological features' modeling. The results of this study suggest that there are variances in model predictivity, differences in predictive variables, as well as the number of features that can influence classification performances. The significant MLR predictors included ER status, HER2 status, Nottingham grade, and presenting tumor size. These parameters are also consistent with many of the ML models presented, although features such as ER status and nodal status were commonly selected during feature reduction for ML modeling. It is important to mention that statistically significant variables are not necessarily predictive variables.[26] This is explained by the different frameworks used for feature selection and model construction. For example, k-NN, SVM, and NB models construct nonlinear decision boundaries based on the multidimensional space; by contrast, the MLR model is used to derive a linear decision boundary based on the logit function and MLR performance is often dependent on larger sample populations for optimal Acc %.[26,27] The results also demonstrate that among all the ML models applied, RF as a well-known ensemble classifier resulted in the best generalization on the independent test set.

In comparison to other reports, MLR models have been well described in previous studies. Rouzier et al[9] constructed an MLR model to estimate the probability of residual cancer

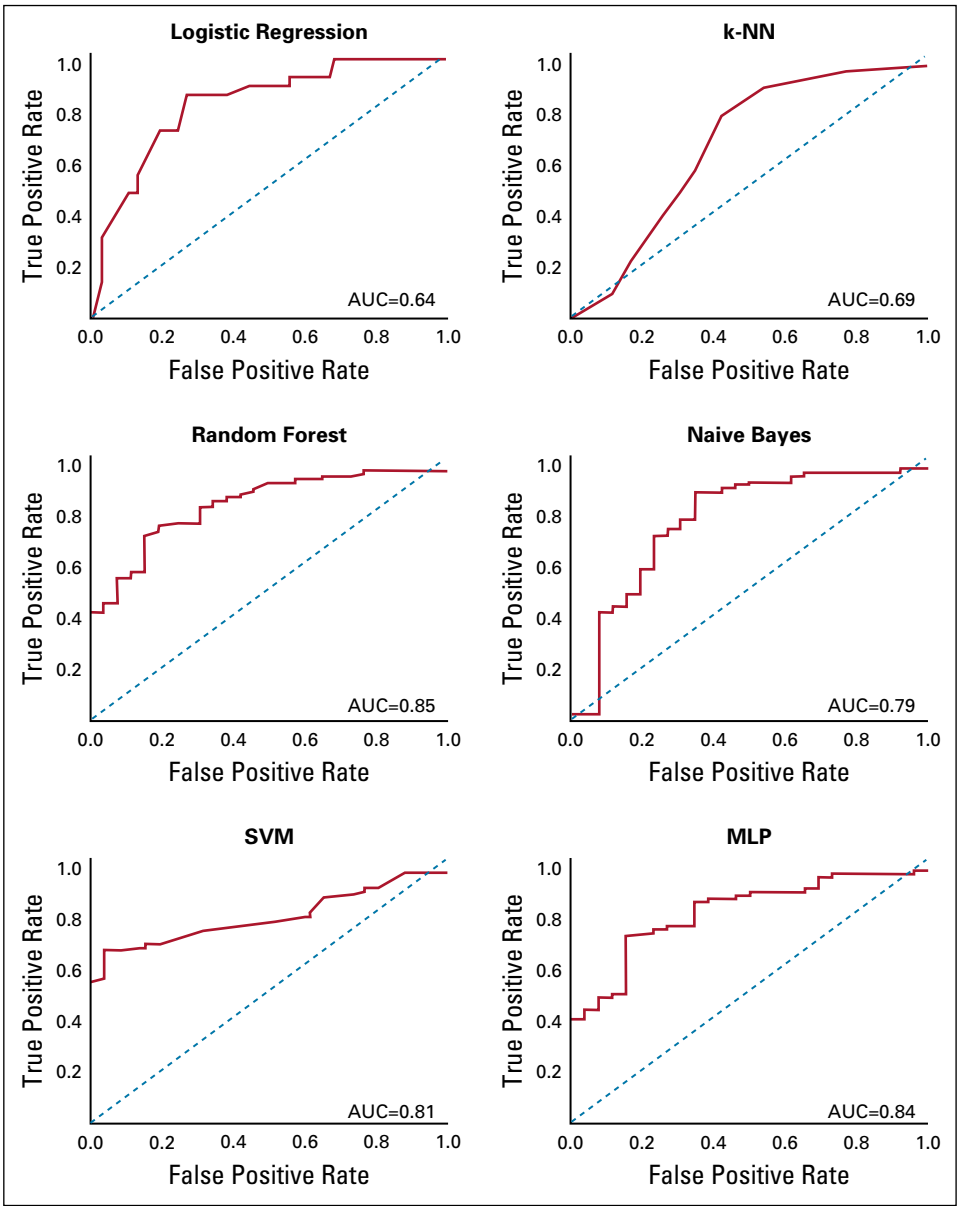

FIG 3. Representative ROC curves. Models were constructed based on categorical and ordinal variables (feature set 2; ie, composite markers). k-NN, k-nearest neighbor; MLP, multilayer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

after NAC. Clinicopathological features in the LR model discriminated for residual disease (< 3 cm) with good predictive performance; the concordance index was 0.71 ($P < 10^{-10}$).[9] Another study by Takada et al[14] used an ML framework to compute probabilities for achieving a pCR. An alternating decision tree (ADTree) classification algorithm was trained on data collected from 150 patients and constituted 28 clinicopathological variables. The final model comprised a subset of those features (f = 15), which were selected based on ADTrees (n = 19) that yielded optimal AUC after cross-validation.[14] The model's performance, tested on an external validation cohort, demonstrated an AUC of 0.787 ($P < .0001$).[14] The ADTree model was also compared with a multivariate LR model; the ML technique performed better in predicting pCR. These studies underscore the opportunities for integrated ML and prediction modeling in breast oncology; as well, they highlight the differences in predictivity between canonical statistical models versus ML models. The results of this study show comparable results and yield predictors akin to those in previous reports.

Limitations of this study include the retrospective study design, which may exclude other potentially relevant variables for modeling that were not included in this study. This may cause biases in the data set, which can alter model generalizations. For example, we did not explore other relevant clinical biomarkers such as Ki-67 and did not conduct subgroup analysis based on molecular subtypes, which could have enhanced prediction performance. Measurement biases can also impact estimates of the models' performances,[28] specifically interuser variability in reporting radiological and pathological parameters. Other

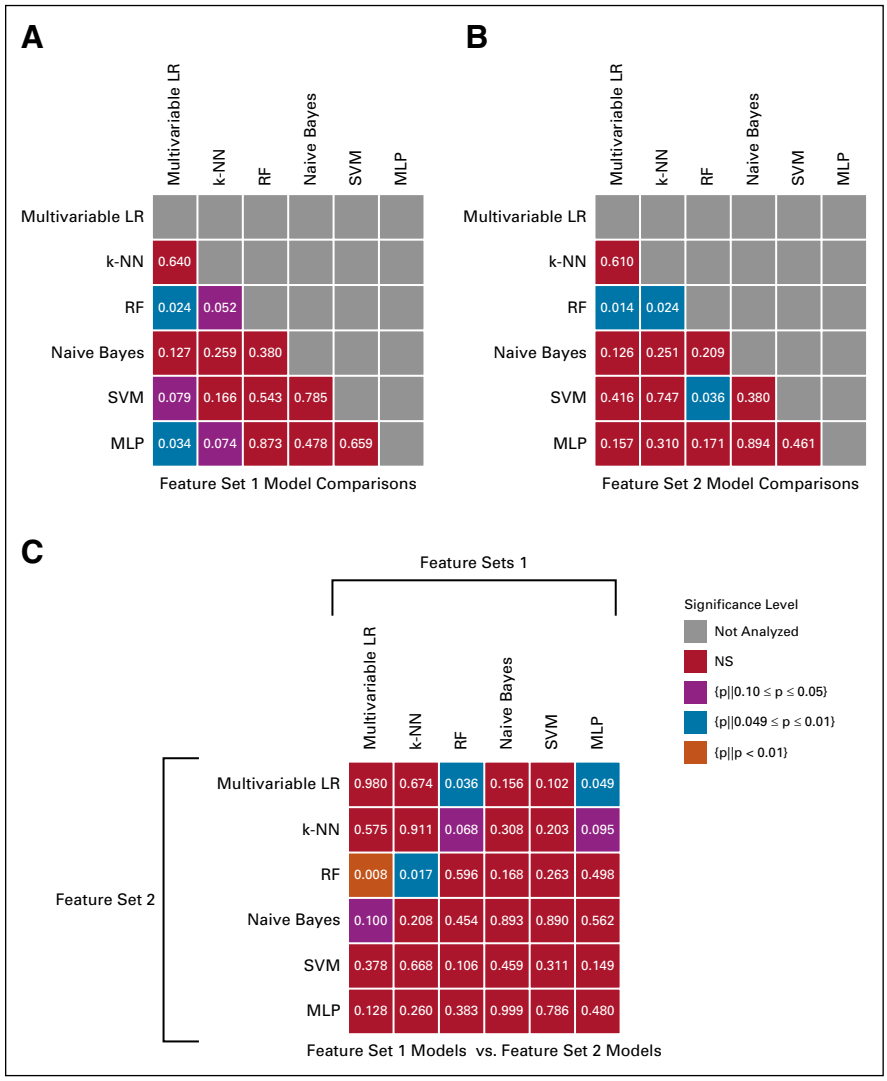

FIG 4. Statistical comparison of ROCs. Heatmap figures demonstrate the varying levels of significant differences between models. (A) Comparison of models using feature set 1. (B) Comparison of models using feature set 2. (C) Comparison of models between feature set 1 and feature set 2. A $P < .05$ demonstrates a significant difference between ROCs. k-NN, k-nearest neighbor; MLP, multilayer perceptron; NS, not significant; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine.

biases include incomplete information regarding the study population, since it is well known that deriving a pCR is a multifactorial problem.[28] For example, nonclinical factors that contribute to the risk of exhibiting residual disease after NAC, such as socioeconomic status and race play a role in treatment response.[29,30] However, these data were not available in this retrospective data set, and excluding patient-related and population-based factors may have impacted model predictions.[28]

For future work, studying continuous outcome measures (eg, RCBI scores in non-pCR patients), subgroup analysis (ie, basal-like *v* non–basal-like TNBC), and collection of multiomics variables will potentially enhance AI-based decision support. Biomarker-guided studies that are driven by robust predictive modeling will improve computational tools to better select NAC treatment regimens, including experimental algorithms. Emerging research is directed at developing state-of-the-art ensemble models that combine multiple ML classifiers. These frameworks are poised to improve Acc%, and can also incorporate complex models such as deep neural network (DNN) architectures.[31] However, DNN-based algorithms generally expend greater computational resources and require more training data compared with the ensemble classifier presented in this study (ie, RF). Nevertheless, the major advantages of such models include increased predictive power by using weighted functions, mitigating data biases, and can help reduce overfitting.[31]

As the neoadjuvant treatment of BC continues to evolve in scope and impact, the utility of ML prediction modeling tools will help improve clinical decision making. Specifically, ML models are poised to enhance response-guided treatments. If further validated, ML prediction models can be implemented into open-source user-based interfaces to calculate the individualized probabilities of pCR versus non-pCR. Subsequently, this could allow clinicians to tailor treatment approaches according to the patient's risk factors. Overall, this work will contribute to leveraging new technologies that will help bring the balance of efficacy and toxicity management from NAC to the next horizon.

In conclusion, NAC continues to play a central role in the management of patients with LABC and high-risk BC. In the growing era of computational oncology, robust predictive modeling has the potential to enhance patient care through advancements in computational modeling and robust data science.

AFFILIATIONS

[1]Division of Medical Oncology, Department of Medicine, University of Toronto, ON, Canada
[2]Temerty Centre for AI Research and Education in Medicine, University of Toronto, ON, Toronto, Canada
[3]Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON, Canada
[4]Department of Radiation Oncology, Sunnybrook Health Sciences Center, Toronto, ON, Canada
[5]Institute of Clinical Evaluative Sciences, Sunnybrook Health Sciences Centre, Toronto, ON, Canada
[6]Department of Laboratory Medicine and Molecular Diagnostics, Sunnybrook Health Sciences Centre, Toronto, ON, Canada
[7]Division of Medical Oncology, Sunnybrook Health Sciences Center, Toronto, ON, Canada
[8]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

CORRESPONDING AUTHOR

William T. Tran, MRT(T), MSc, PhD, Department of Radiation Oncology, University of Toronto, Radiation Therapist Clinician Scientist, Sunnybrook Health Sciences Centre, Department of Radiation Oncology, 2075 Bayview Ave (Room TB-97), Toronto, ON, Canada M4N 3M5; e-mail: william.tran@sunnybrook.ca.

EQUAL CONTRIBUTION

N.M. and K.S. contributed equally to this work.

SUPPORT

W.T.T. received grant funding from the Tri-Council (CIHR) Government of Canada's New Frontiers in Research Fund (NFRF, Grant # NFRFE-2019-00193), the Terry Fox Research Institute (TFRI, Grant #1083), and the Women's Golf Health Classic Foundation Fund. ASN laboratory is funded by the TFRI (Grant # 1083), NFRF (Grant #: NFRFE-2019-00193), and by the Natural Sciences and Engineering Research Council (NSERC, Grant #: RGPIN-2016-06472 and CRDPJ507521-16).

AUTHOR CONTRIBUTIONS

Conception and design: Nicholas Meti, Ali Sadeghi-Naini, William T. Tran
Collection and assembly of data: Nicholas Meti, Sami Tabbarah, Fang-I Lu, Elzbieta Slodkowska, Katarzyna Joanna Jerzak, Lauren Fleshner, Ethan Law, William T. Tran
Data analysis and interpretation: Nicholas Meti, Khadijeh Saednia, Andrew Lagree, Majid Mohebpour, Alex Kiss, Sonal Gandhi, Ethan Law, Ali Sadeghi-Naini, William T. Tran
Provision of study materials or patients: Fang-I Lu
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Sonal Gandhi
Consulting or Advisory Role: Roche, Lilly, Novartis, Exact Sciences, Agendia
Speakers' Bureau: Novartis, Knight
Katarzyna Joanna Jerzak
Honoraria: Roche, Novartis, Apobiologix, Pfizer, Purdue Pharma, Genomic Health, Eisai, AstraZeneca, Knight Therapeutics, Lilly
Consulting or Advisory Role: Roche, Novartis, Pfizer, Apobiologix, Purdue Pharma, Genomic Health, Eisai, Lilly, AstraZeneca, Knight Therapeutics
Research Funding: Lilly, AstraZeneca
Patents, Royalties, Other Intellectual Property: I am the lead inventor on a patent for the use of dronedarone and its structural derivatives as cancer therapies.

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Ms Jan Stewart, Dr Calvin Law, Dr Martin Yaffe, Ms Yulia Yerofeyeva, Ms Angela Leahey, and the Breast Site Group at Sunnybrook Health Sciences Centre for supporting this research.

REFERENCES

1. Lee MC, Newman LA: Management of patients with locally advanced breast cancer. Surg Clin North Am 87:379-398, 2007
2. Von Minckwitz G, Untch M, Blohmer JW, et al: Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol 30:1796-1804, 2012
3. Masuda N, Lee SJ, Ohtani S, et al: Adjuvant capecitabine for breast cancer after preoperative chemotherapy. N Engl J Med 376:2147-2159, 2017
4. Spring LM, Fell G, Arfe A, et al: Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: A comprehensive meta-analysis. Clin Cancer Res 26:2838-2848, 2020
5. Cortazar P, Zhang L, Untch M, et al: Pathological complete response and long-term clinical benefit in breast cancer: The CTNeoBC pooled analysis. Lancet 384, 164-172, 2014
6. Cain EH, Saha A, Harowicz MR, et al: Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: A study using an independent validation set. Breast Cancer Res Treat 173:455-463, 2019
7. Bhardwaj R, Hooda N: Prediction of pathological complete response after neoadjuvant chemotherapy for breast cancer using ensemble machine learning. Informatics Med Unlocked 16:100219, 2019

8. Gass P, Lux MP, Rauh C, et al: Prediction of pathological complete response and prognosis in patients with neoadjuvant treatment for triple-negative breast cancer. BMC Cancer 18:1051, 2018
9. Rouzier R, Pusztai L, Garbay JR, et al: Development and validation of nomograms for predicting residual tumor size and the probability of successful conservative surgery with neoadjuvant chemotherapy for breast cancer. Cancer 107:1459-1466, 2006
10. Lee, JK, Coutant C, Kim YC, et al: Prospective comparison of clinical and genomic multivariate predictors of response to neoadjuvant chemotherapy in breast cancer. Clin Cancer Res 16:711-718, 2010
11. Nwaogu IY, Fayanju OM, Jeffe DB, et al: Predictors of pathological complete response to neoadjuvant chemotherapy in stage II and III breast cancer: The impact of chemotherapeutic regimen. Mol Clin Oncol 3:1117-1122, 2015
12. Shien T, Akashi-Tanaka S, Miyakawa K, et al: Clinicopathological features of tumors as predictors of the efficacy of primary neoadjuvant chemotherapy for operable breast cancer. World J Surg 33:44-51, 2009
13. Keskin S, Muslumanoglu M, Saip P, et al: Clinical and pathological features of breast cancer associated with the pathological complete response to anthracycline-based neoadjuvant chemotherapy. Oncology 81:30-38, 2011
14. Takada M, Sugimoto M, Ohno S, et al: Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique. Breast Cancer Res Treat 134:661-670, 2012
15. McKinney SM, Sieniek M, Godbole V, et al: International evaluation of an AI system for breast cancer screening. Nature 577:89-94, 2020
16. Stark GF, Hart GR, Nartowt BJ, et al: Predicting breast cancer risk using personal health data and machine learning models. PLoS One 14:e0226765, 2019
17. Tahmassebi A, Wengert GJ, Helbich TH, et al: Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. Invest Radiol 54:110-117, 2019
18. Amrane M, Oukid S, Gagaoua I, et al: Breast cancer classification using machine learning, Presented at 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) 1-4. IEEE, Istanbul, Turkey, 2018.
19. Bazazeh D, Shubair R: Comparative study of machine learning algorithms for breast cancer detection and diagnosis, Presented at 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) 1-4. IEEE, Istanbul, Turkey, 2016.
20. Mojarad SA, Dlay SS, Woo WL, et al: Breast cancer prediction and cross validation using multilayer perceptron neural networks, Presented at 2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010) 760-764. IEEE, Newcastle, United Kingdom, 2010.
21. Giuliano AE, Edge SB, Hortobagyi GN: Eighth Edition of the AJCC Cancer Staging Manual: Breast cancer. Ann Surg Oncol 25:1783-1785, 2018
22. Symmans WF, Peintinger F, Hatzis C, et al: Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J Clin Oncol 25:4414-4422, 2007
23. Chawla NV, Bowyer KW, Hall LO, et al: SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321-357, 2002
24. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Stat Med 3:143-152, 1984
25. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36, 1982
26. Lo A, Chernoff H, Zheng T, et al: Why significant variables aren't automatically good predictors. Proc Natl Acad Sci U S A 112:13892-13897, 2015
27. Bewick V, Cheek L, Ball J: Statistics review 14: Logistic regression. Crit Care 9:112-118, 2005
28. Gianfrancesco MA, Tamang S, Yazdany J, et al: Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178:1544, 2018
29. Caudle AS, Gonzalez-Angulo AM, Hunt KK, et al: Predictors of tumor progression during neoadjuvant chemotherapy in breast cancer. J Clin Oncol 28:1821-1828, 2010
30. Killelea BK, Yang VQ, Wang SY, et al: Racial differences in the use and outcome of neoadjuvant chemotherapy for breast cancer: Results from the national cancer data base. J Clin Oncol 33:4267-4276, 2015
31. Sagi O, Rokach L: Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov 8:e1249, 2018

APPENDIX

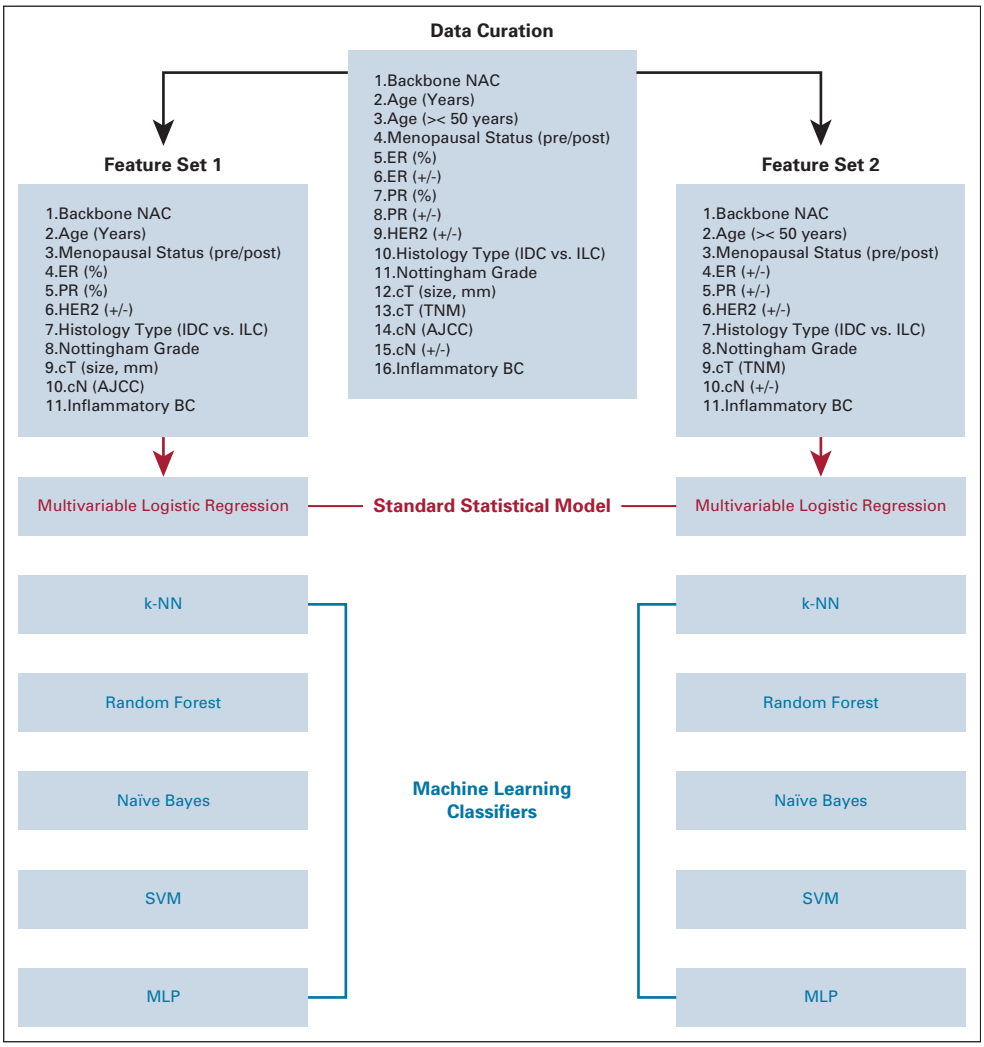

FIG A1. Modeling and experimental design: clinicopathological features were collected from all patients and converted into dichotomous features to run parallel analysis. Features included: age, ER, PR, clinical tumor size, and node status. Prediction models included a multivariable logistic regression (standard statistical model) and machine learning models including k-NN, random forest, naive Bayes, SVM classifier, and MLP. AJCC, American Joint Committee on Cancer; BC, breast cancer; ER, estrogen receptor; HER2, human epidermal growth factor 2; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; k-NN, k-nearest neighbor; MLP, multilayer perceptron; NAC, neoadjuvant chemotherapy; PR, progesterone receptor; SVM, support vector machine.

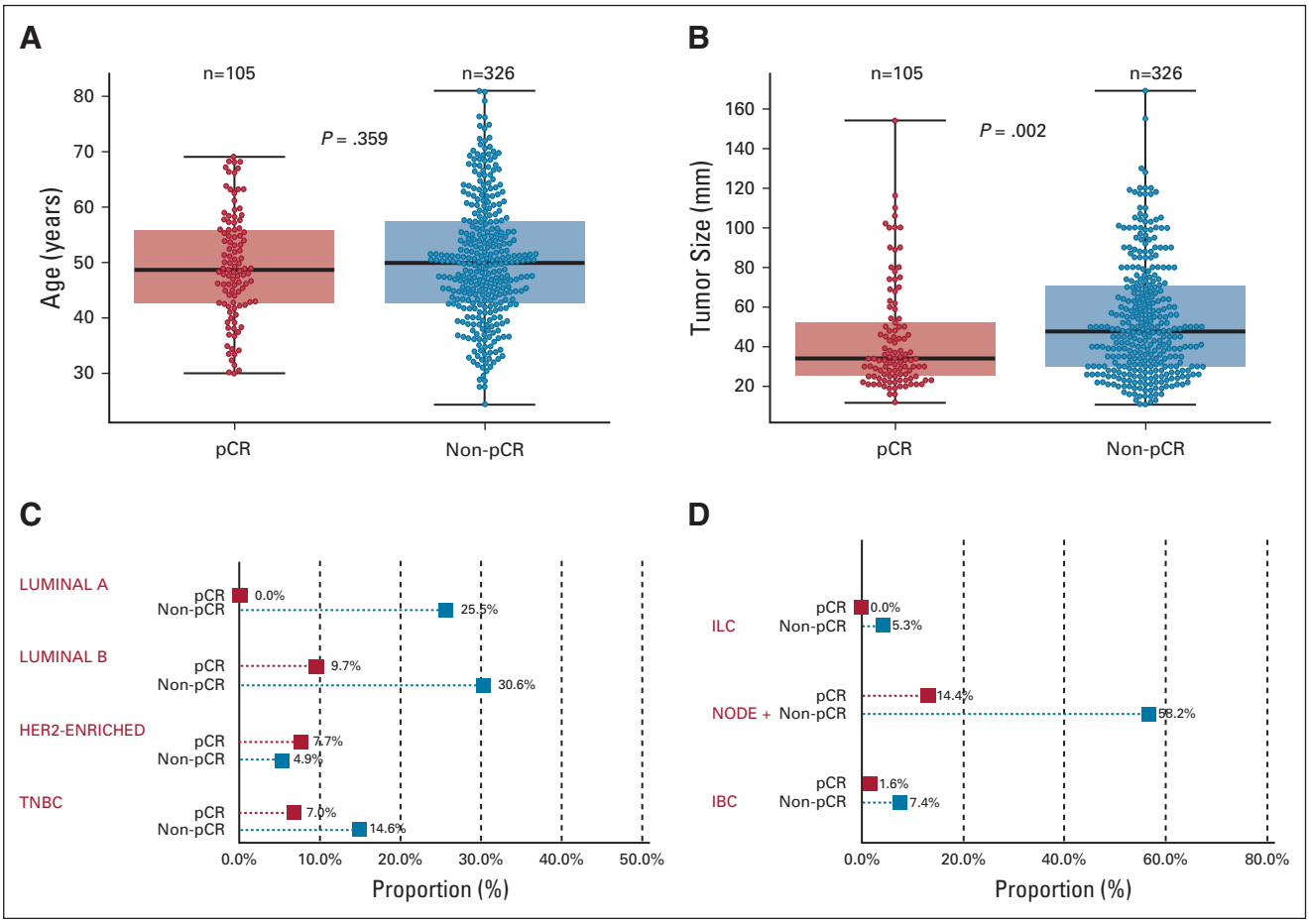

FIG A2. Clinicopathological features and data distribution. Patients were included in this analysis included women older than age 18 years and demonstrating high-risk breast cancer. (A-B) Box and whisker plots with swarm plot showing distribution of patients' age and presenting tumor sizes. (C) Relative proportion of molecular subtype. Ki-67 markers were not available for patients; therefore, molecular subtypes were defined by receptor status and grade. (D) Relative proportion (to entire cohort) based on risk factors such as ILC, node-positive disease, and invasive IBC. Proportion values are relative to the entire analysis cohort (n = 431). HER2, human epidermal growth factor 2; IBC, inflammatory breast cancer; ILC, invasive lobular carcinoma; pCR, pathological complete response; TNBC, triple-negative breast cancer.

TABLE A1. Machine Learning Classification

No.	ML Model	Classification Framework Used
1.	k-NN	• $k = 7$ neighbors
2.	Random forest	• Number of estimators: 50 • Maximum depth: 4
3.	Naive Bayes	• Gaussian naive Bayes model
4.	SVM	• Kernel: linear
5.	MLP	• Number of hidden layers: 2 • Number of neurons in each hidden layer: 50 • Activation function: ReLU • Number of iterations: 100

NOTE: Five machine learning models were used to challenge the standard statistical model. Machine learning classifiers were selected for experimentation based on high predictive performances in previous reports. Tuning parameters and hyperparameters are listed for each classification framework.

Abbreviations: MLP, multilayer perceptron; ReLU, Rectified Linear Unit; SVM, support vector machine.

TABLE A2. Summary of Clinicopathological Features

	Analysis Cohort (n = 431)		
Clinicopathological Characteristic	**n (%)**	**n (%)**	***P***
Patient demographics	pCR (n = 105)	non-pCR (n = 326)	
Median age (y)	48.6	49.9	*0.35857*
Menopausal status			
Pre- and/or perimenopausal	53 (51%)	182 (56%)	*0.33814*
Postmenopausal	52 (49%)	144 (44%)	
Neoadjuvant chemotherapy			
AC-T	63 (60%)	196 (60%)	*0.98219*
FEC-D	42 (40%)	130 (40%)	
Trastuzumab anti-HER2 therapy	73 (70%)	95 (29%)	***< 0.00001***
Tumor laterality			
Left	51 (49%)	145 (44%)	*0.46386*
Right	54 (51%)	181 (56%)	
Histology			
Invasive ductal carcinoma	105 (100%)	303 (93%)	***0.00515***
Invasive lobular carcinoma	0 (0%)	23 (7%)	
Receptor status			
ER-positive	40 (38%)	240 (74%)	***< 0.00001***
PR-positive	31 (30%)	207 (63%)	***< 0.00001***
HER2-positive	73 (70%)	95 (29%)	***< 0.00001***
Nottingham grade			
1	3 (3%)	19 (6%)	***< 0.00001***
2	23 (22%)	159 (49%)	
3	79 (75%)	148 (45%)	
Tumor size			
Mean tumor size (mm; ± SD)	43.8 ± 26.8	53.2 ± 28.8	***0.00263***
Clinical T stage			
1 (≤ 20 mm)	8 (8%)	24 (7%)	***0.00475***
2 (> 20 to ≤ 50 mm)	69 (66%)	158 (48%)	
3 (> 50 mm)	28 (26%)	144 (44%)	
4 (any size with extension)	0 (0%)	0 (0%)	
Clinical N stage			
0 (no lymph nodes)	43 (41%)	75 (23%)	***0.00053***
1 (1-3 lymph nodes)	59 (56%)	211 (65%)	
2 (4-9 lymph nodes)	3 (3%)	32 (10%)	
3 (≥ 10 lymph nodes)	0 (0%)	8 (2%)	
Other presenting clinical information			
Inflammatory breast cancer	7 (7%)	32 (10%)	*0.32792*

NOTE: There were 431 patients included in the analysis cohort. Retrospective data collection included all women treated with anthracycline- and taxane-based chemotherapy. Pretreatment pathological characteristics were collected, and post-treatment responses to NAC were collected from patient electronic medical records.

Percentages are relative to the group classification. Bold values denote statistical significance.

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor 2; PR, progesterone receptor; SD, standard deviation.

