An Alternative Composite Score Using Easier to Access Data to Determine the Probability of Recurrence in HER2-ve Breast Cancer

Rishi V. Pai

Northview High School, 10625 Parsons Rd., Johns Creek, Georgia, 30097, USA; pai.rishi0709@gmail.com

ABSTRACT: Luminal breast cancers (BC) are the most common subtype, with human epidermal growth factor receptor 2 negative being the most prevalent. Currently, the genetic-based Oncotype DX score is the widely used metric for determining recurrence probability, but the 21-genes it utilizes are difficult and time-consuming to obtain data for. This study aimed to use machine learning classification to predict composite scores for recurrence probability based on easier-to-access data, selected from feature importance identified by the model(s). Four classification models were trained and tested to predict BC recurrence: random forest (RF), logistic regression (LR), gradient boosting (GBM), and decision tree (DT). The models were compared by their F1 score. The most significant variables from the best-performing model trained a Calibrated GBM (chosen as the final predictor as it had the highest F1 score at 0.24) to predict composite scores from 0-10. The predictive model returned scores with ~92% accuracy. Findings showed that the GBM is not reliable for solely binary recurrence prediction due to poor performance metrics, but it serves as a promising tool for predicting composite scores for recurrence probability. These scores could be utilized in clinical settings to determine intervention plans to improve patient prognosis.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Predictive Analytics, Breast Cancer, Prognosis.

Introduction

Luminal breast cancers (BC) account for approximately 65% of all BC cases, being the most common subtype.[1, 2] Human epidermal growth factor receptor 2 negative (HER2-negative) is the most common for early stage BC, as represented in Figure 1.[3, 4] The number of deaths from BC is estimated to rise to 2.9 million annually.[5] Thus, early diagnosis and effective treatment are crucial for prognosis efforts.[1, 6]

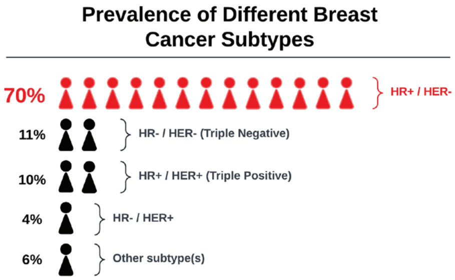


Figure 1: Prevalence of different BC subtypes. Data gathered from the National Cancer Institute SEER (Surveillance, Epidemiology, and End Results Program).[4] The visual presents the HER2-negative subtype as the most prevalent of all breast cancer cases (~70%).

Machine learning's (ML) ability to analyze large amounts of data makes it promising as a tool for predicting BC recurrence.[7, 9] Some known important factors include histological grade, tumor size, nodes, genomic score, and the Ki67 proliferation index.[3] Thus far, the recurrence of BC and its factors have not been thoroughly studied through machine learning techniques as a consequence of patient recurrence data rarely being available in accessible datasets.[10]

The widely used metric score for BC recurrence probability is the Oncotype DX score, a 21-gene recurrence tool that examines the activity of genes in the breast tumor tissue.[2, 11] The scores range from 0 to 100, with higher scores reflecting a higher probability of BC recurrence as well as the likelihood of benefitting from chemotherapy and hormonal therapy. Patients who have Oncotype DX scores above 26 benefit most from chemotherapy.[2] The score was developed mainly for estrogen receptor-positive (ER-positive) and HER2-negative BC, and it gives an accurate estimation of recurrence probability.[11]

Genetic patient data for the Oncotype DX score is generally difficult, tedious, or time-consuming to obtain, so an equally accurate recurrence probability score calculated from easier-to-access data would be more efficient. This study aims to develop a machine learning classification model to predict the recurrence of BC based on clinical, histological, immunohistochemical, molecular biology, and treatment patient data. A second classifier trained by data selected from feature importance identified by the original model(s) will be tested to predict accurate composite scores from 0 to 10 for recurrence probability.

Methods

All computations were performed on Python version 3.11.7 on JupyterLab 4.0.11. All machine learning models and techniques were provided by the *scikit-learn* package in Python. The research methods are outlined in Figure 2.

DOI: 10.36838/v7i8.7

Data Collection and Curation:

A publicly available and anonymous Oncotype DX patient dataset containing 321 patients was originally obtained from the Georges Francois Lecler Cancer Centre and North Trévenans County Hospital, both in France.[12] The dataset was accessed via data.world, which provided a convenient platform for data retrieval. The dataset was originally in French and was translated into English.

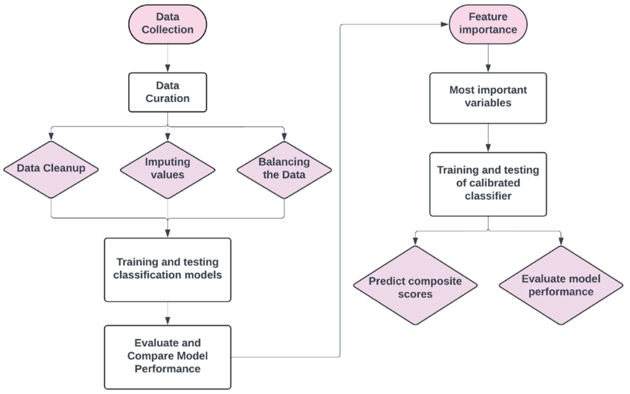

Data Collection
Data Curation
Data Cleanup
Imputing values
Balancing the Data
Training and testing classification models
Evaluate and Compare Model Performance
Feature importance
Most important variables
Training and testing of calibrated classifier
Predict composite scores
Evaluate model performance

Figure 2: Workflow of the research methods. The chart displays the process from sole BC recurrence prediction to the final composite score prediction methods.

The columns *NBlock* and *Block ODX* were removed as they had no significance on patient recurrence in the context of the data, since they were present for patient identification. The *oncogenetic consultation, mutation, HER2 IHC,* and *HER2 SISH* columns were removed as they had 70% or more missing information. Imputing data for these columns would not be a proper representation of patient data as there is not enough basis for estimates. The *Stage pTNM* column was removed because it had too many unique values to be significant enough for predicting BC recurrence. While mutation patterns are strong indicators of cancer recurrence, the data set used in this study had a weak basis for clinical staging correlation and would give biased results if the empty data were to be predicted. Finally, the *last contact date* and *last contact status* columns were removed. The date provides no information and does not have a recurrence metric. Pre-testing showed that these columns had little to no impact on changing recurrence predictions, as the most prominent variables consisted of clinical, histological, immunohistochemical, molecular biology, and treatment data native to the dataset. The few deceased patients in the dataset in the *last contact status* column were removed to keep only patients who survived from HER2-negative BC.

Any patients who had blank data for the *Recurrence* column were removed from the dataset. For the remaining data columns that had missing information, these values were filled in using mode imputation to ensure data consistency for model training. Additional data curation included removing secondary or tertiary values from cells that had more than one value and changing all classification columns with string/character data to numeric classification (0, 1, 2, etc)

Patient data for the *node type, histological type,* and *histological subtype* columns were represented with a range of values. These columns were expanded to better visualize the types of each that occurred for each patient. For example, a patient could have either a 0, 1, or 2 for the *node type* column, with each value representing 'no sentinel node or dissection' (column 1), 'sentinel node' (column 2), or 'axillary dissection' (column 3), respectively. The original column was split into three, with each patient recording either a 0 (does not have) or a 1 (does have). A patient who has only sentinel and axillary dissection nodes would have a 0 for column 1, and a 1 for columns 2 and 3.

The dataset was standardized for columns that had numeric values using *scikit-learn* pre-processing. This was done to ensure uniformity and consistency between data columns for model efficiency. After data curation, the dataset was split 65% for model training and 35% for model testing, resulting in 191 training patients and 104 testing patients.

The provided recurrence column had classes 0 and 1, with 0 representing no recurrence of BC and 1 representing recurrence. This column was highly imbalanced, with more than 90% of the patients recording class 0 (non-recurrence) and a much smaller proportion for class 1 (recurrence). To prevent overfitting for the models, the majority class was down-sampled, and the minority class was up-sampled to balance the target column for the training set. This resulted in a training dataset with a combined 92 rows: 50 for class 0 and 42 for class 1.

Model Training and Testing for Binary BC Recurrence Prediction:

Four classification models were used initially to test BC recurrence prediction: random forest (RF), logistic regression (LR), gradient boosting (GBM), and decision tree (DT). These models were chosen for their previous usage in BC studies involving diagnosis and prognosis analysis, and primarily for their ability to record feature importance, which was considered for predicting the composite scores later in the study. Each of the models was evaluated with *F1 score (harmonic mean of precision and recall), accuracy (proportion of total correct predictions), precision (proportion of predicted positive instances that are true positives), recall (sensitivity; proportion of true positive instances correctly identified), and specificity (the proportion of true negative instances correctly identified).* The models were compared by their respective F1 scores. Receiver operating characteristic (ROC) curves and confusion matrices were created for each model to compare efficacy.

Random Forest (RF). The random forest model is a type of supervised learning, where the model uses patterns in the dataset to make predictions based on labeled data.[13] Random forest is an expansion on decision trees;[13, 14] As the method may suggest, random forest models typically perform better than decision tree algorithms, as it is a multi-faceted approach.[14] Decision trees split data into subsets at each node by choosing the feature that best separates the data, and repeat this process recursively until reaching a final prediction.[13] In a random forest, each "tree" votes on a prediction, and the class with the highest number of votes is the final prediction.[13-15] Random forest is typically favored for medical studies and has shown sufficient results.[14,16] The random forest algorithm is presented in Figure 3.

Logistic Regression (LR). The logistic regression algorithm finds the relationship between features and outcome probability through a sigmoidal curve. Simply, logistic regression returns the likelihood of an outcome when given individual features.[13,17] The model produces single numerical values from 0 to 1 from numeric features.[13]

Gradient Boosting (GBM). A gradient boosting classifier sequentially builds an ensemble of weaker models, typically decision trees, where each new model is trained to correct errors from the previous ones;[18] Gradient boosting gradually improves overall predictive accuracy by fitting consecutive models.[18] Boosting is an improvement on simpler ensemble techniques like decision trees or random forests by iteratively training new models from prior mistakes.

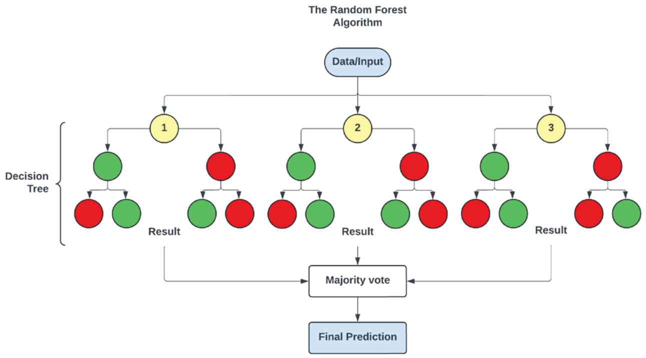


Figure 3: Visual of the random forest algorithm. The algorithm is an expansion of decision tree classification, seen in the diagram as a culmination of several subtrees. The visual presents how the algorithm uses "votes" from each subtree to reach a conclusion.

Each model is able to return the variables that had the most impact in predicting the target column (feature importance). It can be thought of as determining the "decision-making power" for each data variable.

Predicting the Recurrence Probability Composite Scores:

The same 92-row training dataset for binary recurrence prediction was used for training and testing the model to generate composite scores, with an 80:20 training/testing split. However, the model was trained only with the top 21 features, or the features that had a significant numeric score. Significant numeric scores were regarded as scores that had a positive value, meaning that they contributed to a notable percentage of the prediction. For example, a score with a value of 0.15 represents a 15% contribution to the total importance of the recurrence/non-recurrence prediction. The classifier for composite score prediction was trained with the same algorithm(s) as those of the model for binary BC prediction. While the variable range was limited in this regard, through feature importance selection, the model still utilized the provided data to identify patterns between patient data and their assigned Oncotype DX score.

Out of the top 21, variables that had no real impact on patient prognosis were removed, such as diagnosis year and birth year. The model that was used for score prediction was a calibrated GBM, as it had the highest F1 score compared to the other classifiers when predicting recurrence. * **Explained in Results.**

Regular classifiers will return raw scores while a calibrated one adjusts scores to provide more accurate probabilities, meaning that the predictions more accurately reflect the true likelihood of an outcome. Calibration techniques provide a more proper reflection of recurrence probability through the composite scores. The recurrence target column remained in the training set but was removed in the testing set for the model to predict scores purely from patient data.

Result and Discussion

All charts/graphs were created using the matplotlib package in Python.

Evaluation of the Models for Recurrence Prediction:

The five recorded performance metrics for each model are presented in Table 1. In this context, classes are predicted by the model (true positives - recurrence and true negatives – non-recurrence), so the F1 score is a preferred metric rather than accuracy to compare model performance, as it is the numeric mean of precision and recall.

Table 1: Performance metrics of the four classifiers to predict BC recurrence. The table shows that the Gradient Boosting (GBM) classifier resulted in the highest F1 Score (indicated in bold), but did not perform sufficiently in other metrics.

Model	F1 Score	Accuracy	Precision	Recall	Specificity
Random Forest	0.12	0.89	0.25	0.11	0.97
Logistic Regression	0.19	0.75	0.13	0.33	0.79
Gradient Boosting *	**0.24**	0.75	0.16	0.44	0.78
Decision Tree	0.08	0.76	0.06	0.11	0.84

Apart from accuracy and specificity, all models returned relatively low performance metrics (values closer to 1.0 indicate better performance in the field). Random forest showed the highest accuracy (0.89), the highest specificity (0.97), and the highest precision (0.25) compared to the other models, but had a low F1 score (0.12). As briefly mentioned in the methods, the Gradient Boosting Classifier was chosen as the final recurrence predictor as it had the highest F1 score (0.24).

The ROC curve was plotted for the Gradient Boosting Classifier, as shown in Figure 4. The curve shows model performance by plotting the true positive rate (y-axis) compared to the false positive rate (x-axis).

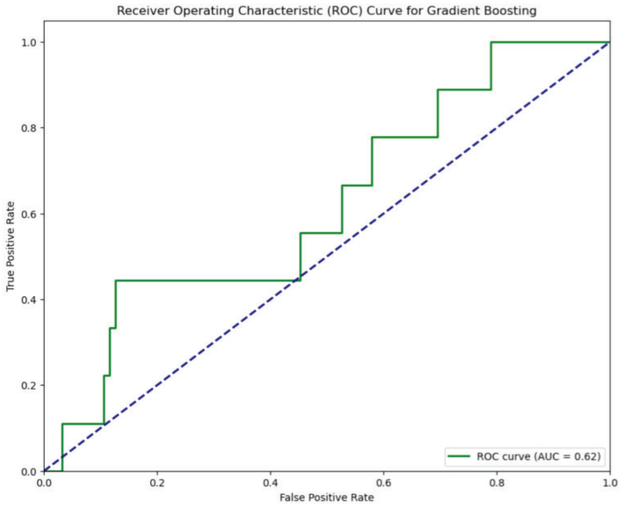
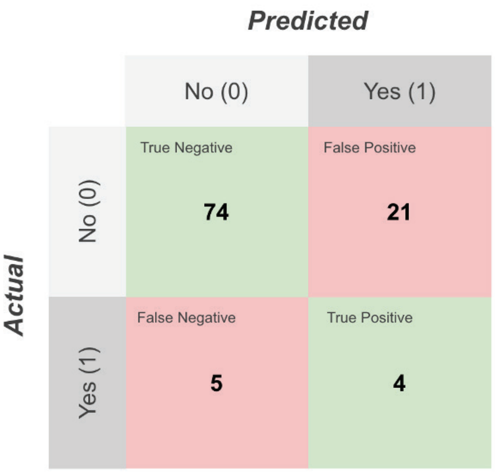


Figure 4: The ROC curve for the GBM compared to the area under the curve (AUC). The figure shows that the model's correlation between false and true positive predictions is similar to that of random guessing, indicated by its close distance to the AUC line.

The Area under the curve (AUC; blue-dotted line) is representative of random guessing. Ideally, peak model performance would result in an ROC curve that is closest to the top left corner of the graph, somewhat like a logarithmic function. Figure 4 shows that although the ROC curve for the model is higher than the AUC curve, it does not necessarily achieve high performance. It reflects only a slightly better performance than random guessing.

The confusion matrix is another method to evaluate the performance of a classification model by displaying the number of true/false positives and negatives predicted by the model for the testing dataset. The confusion matrix for the Gradient Boosting Classifier is outlined in Table 2.

Table 2: Confusion matrix for the testing dataset for the Gradient Boosting Classifier. The table shows strong results for the GBM classifier's negative prediction capability. Conversely, the GBM classifier is unreliable for predicting positive instances.



		Predicted	
		No (0)	Yes (1)
Actual	No (0)	True Negative 74	False Positive 21
	Yes (1)	False Negative 5	True Positive 4

The testing dataset consisted of 104 rows but was imbalanced for the recurrence column due to the original Oncotype DX dataset, so there was a considerably higher number of negative predictions. Out of 79 cases of negative instances (non-recurrence), the model was able to predict 74 accurately, with the other 5 predictions returning false negatives. However, out of 25 that did have recurrence, the model was only able to predict 4 accurately, with the other 21 returning false positives. These results, displayed in the confusion matrix, show that the model is accurate at predicting negative instances (~94% correct predictions) but poor at predicting positive instances (16% correct predictions). These metrics correlate to the model's low scores for specificity (true negative rate) and sensitivity (true positive rate) from Table 1. Thus, the model is highly accurate for predicting non-recurrence in patients but is not very accurate at predicting recurrence.

Composite Score Results:

The feature importance for the GBM is presented in Figure 5.

The five most important features for recurrence prediction identified by the model were **diagnosis year, surgery, oncotype score, progesterone receptor (PR) Allred score, and PR percentage,** with numeric scores of ~0.25, ~0.17, ~0.10, and ~0.06, respectively.

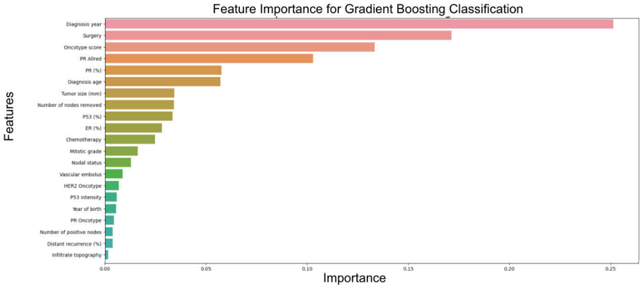
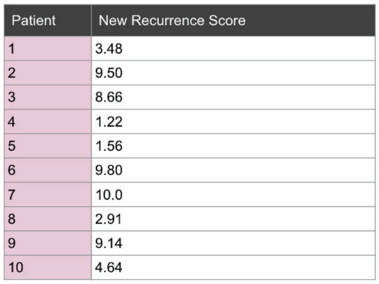


Figure 5: A bar graph that visualizes and compares the feature importances for the Gradient Boosting Classifier, showing which data columns (variables) are considered the most important for predicting BC recurrence. Features that did not have significant values were excluded from the analysis.

Examples of composite scores for patients in the testing dataset are presented in Table 3. The calibrated Gradient Boosting Classifier was able to accurately predict composite scores to reflect the probability of BC recurrence (~92% accuracy). A score greater than 5 and closer to 10 indicates a higher probability of recurrence. Conversely, a score less than 5 and closer to 0 indicates a lower probability of recurrence, a similar system to Oncotype DX. The calibrated model was trained with all the top 21 variables from the feature importance graph (Figure 5), excluding *diagnosis year, oncotype score,* and *year of birth*. Diagnosis year and year of birth were insignificant variables as the "diagnosis age" was already provided for each patient. Specifically, the diagnosis year and year of birth columns were values that had little correlation to the recurrence predictions. The distant recurrence percentage was included as a variable in composite score training to serve as a proxy for metastatic potential. The metric reflects the likelihood of the tumor spreading to distant organs over time and helps the training model better understand the biological behavior and aggressiveness of the tumor. The distant recurrence percentage is linked to molecular biology, which allows the model to base predictions regarding metastatic risk.

Table 3: A sample of 10 patients from the testing dataset with new predicted recurrence probability scores. The scores were rounded to the nearest two decimal places. The composite scores were predicted with ~92% accuracy to the corresponding original Oncotype-DX score for the patient.



Patient	New Recurrence Score
1	3.48
2	9.50
3	8.66
4	1.22
5	1.56
6	9.80
7	10.0
8	2.91
9	9.14
10	4.64

Discussion:

This study aimed to analyze prognostic factors for BC recurrence and use those factors to develop a model to predict composite scores to reflect recurrence probability as an alternative to Oncotype DX.

Out of the four classification models that were tested (random forest, logistic regression, gradient boosting, and decision tree), the Gradient Boosting Classifier was chosen as the final model as it had the highest F1 score (0.24). This model had a relatively high overall accuracy (0.75) for predicting BC recurrence. It is important to note that the model's performance in predicting positive instances (having recurrence) is quite poor, as indicated by the low precision and recall (0.16, and 0.44, respectively), as well as the disparities of the confusion matrix (Table 2) true/false positive prediction for the testing dataset. The high specificity of the model (0.78) and the confusion matrix again show that the model is sufficient at identifying negative instances. The discrepancy between positive and negative case predictions is likely due to the imbalance of the target column in the dataset.

The top 21 features of highest importance in descending order include: *diagnosis year, surgery, oncotype score, PR Allred, PR %, diagnosis age, tumor size (mm), number of nodes removed, P53 %, ER (estrogen receptor) %, chemotherapy, mitotic grade, nodal status, vascular embolus, HER2 oncotype, P53 intensity, year of birth, PR oncotype, number of positive nodes, distant recurrence %, and infiltrate topography*. These factors include clinical, histological, immunohistochemical, molecular biology, and treatment data. The results fairly align with previous BC prognostic studies that showed tumor size as one of the most important variables for recurrence prediction.[3]

While the Gradient Boosting Classifier might not be a reliable tool for solely binary prediction of BC recurrence, it appears as a promising method for predicting the probability scores of recurrence (~92% accuracy). The GBM was inconsistent with recurrence/non-recurrence predictions in the first step of the methodology, but proved sufficient in its ability to correlate data points to previously calculated Oncotype DX scores. This evaluation allowed it to identify what data constitutes a specific score, ensuring accuracy in the newly generated composite scores. The predicted scores are a fairly accurate alternative to the widely used Oncotype DX score, as they utilize more easily accessible patient data for calculating recurrence probability rather than the 21-genes used for calculating the Oncotype DX score.

■ Conclusion

In this study, four classification machine learning models were trained and tested with recurrence data from HER2-negative BC patients in France to determine the most important factors for recurrence and use them to predict an efficient composite score based on easier-to-obtain data to reflect a patient's chances of BC recurrence.

The Gradient Boosting Classifier overall is not a great method for predicting BC recurrence in hospital settings. While not sufficient for solely recurrence prediction, the GBM is practical for predicting composite scores. However, the model does show promise, but requires refining for recurrence prediction. This predictive score could be utilized in hospital settings to determine correct intervention plans to improve patient prognosis.

Future Work:

The Oncotype DX dataset used in this study was fairly limited with the number of patients, but the key downside was the imbalance of the recurrence data. More than 90% of the patients recorded non-recurrence, while a few recorded recurrences, which required artificial balancing of the data. While still an accurate estimation of recurrence data, this placed a bias on negative instance prediction compared to positive instance prediction for the testing datasets. Training and testing a model with a larger BC recurrence dataset with a larger scope, with more true and balanced prognosis data, should be done to yield higher performance results for classification models.

Additionally, common clinical practice is to use one dataset for model training and a separate test for model validation. Regarding breast cancer recurrence, there is a scarcity of publicly available, anonymous datasets with true recurrence data, which is the sole reason that breast cancer recurrence is an under-studied field of oncology. To compensate, the dataset used in this study was split to create separate sets for training and testing, with the testing dataset having no predicted/imputed data to mitigate bias for model validation. Thus, using a larger and more diverse data set is a prospective endeavor to boost performance for the models used in this study.

Applying the methodology and composite score prediction techniques used in this study for other cancers and diseases, such as cardiovascular disease, may help to better understand recurrence probability predictors.

■ Acknowledgments

I would like to thank Jagdeep Podichetty, PhD, at the Critical Path Institute for his guidance and support throughout the research process.

■ References

1. Gilchrist, J. Current Management and Future Perspectives of Hormone Receptor-Positive HER2-Negative Advanced Breast

DOI: 10.36838/v7i8.7

Cancer. *Seminars in Oncology Nursing*. **2024**, *40* (1). DOI: 10.1016/j.soncn.2023.151547

2. Durrani, S.; Al-Mushawa, F.; Heena, H.; Wani, T.; Al-Qahtani, A. Relationship of Oncotype Dx score with tumor grade, size, nodal status, proliferative marker Ki67 and Nottingham Prognostic Index in early breast cancer tumors in Saudi Population. *Annals of Diagnostic Pathology*. **2021**, *51*. DOI: 10.1016/j.anndiagpath.2020.151674
3. Zambelli, A.; Gallerani, E.; Garrone, O.; Pedersini, R.; Rota Caremoli, E.; Sagrada, P.; Sala, E.; Cazzaniga, ME. Working tables on Hormone Receptor Positive (HR+), Human Epidermal growth factor Receptor 2 negative (HER2-) early stage breast cancer: Defining high risk of recurrence. *Critical Reviews in Oncology/Hematology*. **2023**, *191*. DOI: 10.1016/j.critrevonc.2023.104104
4. NIH SEER. *Cancer Stat Facts: Female Breast Cancer Subtypes*. https://seer.cancer.gov/statfacts/html/breast-subtypes.html (accessed 2024-07-20)
5. Sharma, A.; Goyal, D.; Mohana, R. An ensemble learning-based framework for breast cancer prediction. *Decision Analytics Journal*. **2024**, *10*. DOI: 10.1016/j.dajour.2023.100372
6. Zuo, D.; Yang, L.; Jin, Y.; Qi, H.; Liu, Y.; Ren, L. Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*. **2023**. DOI: 10.1186/s12911-023-02377-z
7. González-Casto, L.; Chávez, M.; Duflot, P.; Bleret, V.; Martin, A. G.; Zobel, M.; Nateqi, J.; Lin, S.; Pazos-Arias, J. J.; Del Fiol, G.; López-Nores, M. Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. *MDPI*. **2023**, *15* (10). DOI: 10.3390/cancers15102741
8. Liu, Y.; Fu, Y.; Peng, Y.; Ming, J. Clinical decision support tool for breast cancer recurrence prediction using SHAP value in cooperative game theory. *Heliyon*. **2024**, *10* (2). DOI: 10.1016/j.heliyon.2024.e24876
9. Jin, Y.; Lan, A.; Dai, Y.; Jiang, L.; Li, S. Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *European Journal of Medical Research*. **2023**, *28* (394). DOI: 10.1186/s40001-023-01361-7
10. Henriquez Abreu, P.; Seoane Santos, M.; Henriques Abreu, M.; Andrade, B.; Castro Silva, D. Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. *ACM Computing Surveys*. **2016**, *49* (3), 1-40. DOI: 10.1145/2988544
11. Loggie, J.; Barnes, P. J.; Carter, M. D.; Rayson, D.; Bethune, G. C. Is Oncotype DX testing informative for breast cancers with low ER expression? A retrospective review from a biomarker testing referral center. *The Breast*. **2024**, *75*. DOI: 10.1016/j.breast.2024.103715
12. Zemouri, R.; Omri, N.; Morello, B.; Devalland, C.; Arnould, L.; Zerhouni, N.; Fnaiech, F. Constructive Deep Neural Network for Breast Cancer Diagnosis. *IFAC-PapersOnLine*. **2018**, *51* (27). DOI: 10.1016/j.ifacol.2018.11.660
13. Choi, R. Y.; Coyner, A. S.; Kalapathy-Cramer, J.; Chiang, M. F.; Peter Campbell, J. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Translational Vision Science & Technology*. **2020**, *9* (2). DOI: 10.1167/tvst.9.2.14
14. Minnoor, M.; Baths, V. Diagnosis of Breast Cancer using Random Forests. *Procedia Computer Science*. **2023**, *218*, 429-437. DOI: 10.1016/j.procs.2023.01.025
15. Macauley, B. O.; Aribisala, B. S.; Akande, S. A., Akinnuwesi, B. A.; Olabanjo, O. A. Breast cancer risk prediction in African women using Random Forest Classifier. *Cancer Treatment and Research Communications*. **2021**, *28*. DOI: 10.1016/j.ctarc.2021.100396
16. Ganggayah, M. D.; Taib, N. A.; Har, Y. C.; Lio, P.; Dhillon, S. K. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making*. **2019**, *19* (48). DOI: 10.1186/s12911-019-0801-4
17. Sperandei, S. Understanding logistic regression analysis. *Biochemia Medica (Zagreb)*. **2014**, *21* (1), 12-18. DOI: 1.11613/BM.2014.003
18. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. **2013**, *7* (21). DOI: 10.3389/fnbot.2013.00021

■ Author

Rishi Pai is currently a sophomore at Northview High School in Johns Creek, Georgia, USA. He has a deep passion for machine learning, data science, robotics, and materials science/engineering, and hopes to pursue further research projects in these fields.