Machine Learning-Based Multi-Class Prediction of Breast Cancer Tumor Stage Using Clinical and Pathological Features: A METABRIC Cohort Study

Muhammad Adeel Abid, Amna Siddique, Khadija Younas
Department of Data Science & AI (Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan).
*__Correspondence:__ adeel.abid@kfueit.edu.pk

Citation | Abid. M. A, Siddique. A, Younas. K, "Machine Learning-Based Multi-Class Prediction of Breast Cancer Tumor Stage Using Clinical and Pathological Features: A Metabric Cohort Study",IJIST, Vol. 8 Issue. 7 pp 3038-3049, September 2026
DOI:
Received| Sep 02, 2026 **Revised|** Sep 21, 2026 **Accepted|** Sep 24, 2026 **Published|** Sep 27, 2026.

Breast cancer is the most frequently diagnosed cancer among women worldwide, and the tumor stage recorded at diagnosis strongly determines treatment planning, prognosis and survival. Unlike most prior METABRIC-based studies that predict molecular subtype or survival, this research benchmarks thirteen classifiers with four resampling strategies, hyperparameter optimization, cross-validation and SHAP-based explainability specifically for multi-class tumor-stage prediction from routinely collected clinical variables. Clinical and pathological records of 2,509 METABRIC patients are used, retaining 17 attributes and a four-level stage label. Stage 3 and 4 are merged into a single Advanced class to counter sparsity, and thirteen classifiers are trained on an 80:20 stratified split, compared against SMOTE, ADASYN, Borderline-SMOTE and random under sampling, then tuned via grid search and validated with 5-fold cross-validation. Random Forest with ADASYN gave the strongest pipeline, reaching 83.9% accuracy and macro-F1 of 0.770, versus 0.753 without resampling. The tuned model achieved a macro-AUC of 0.943, while a neural network underperformed the tree ensembles (macro-F1 = 0.75). SHAP identified positive lymph nodes, tumor size and the Nottingham Prognostic Index as the leading predictors, consistent with established clinical staging criteria. A resampled, tuned and explained ensemble model can support, but not replace, clinical tumor staging, pending further validation.

Keywords: Breast Cancer, Metabric, Multi-Class Classification, Tumor Stage Prediction, Shap.

Introduction:

Breast cancer is the most frequently diagnosed cancer in women and remains one of the major contributors to cancer-related death globally. In 2022, an estimated 2.3 million new cases and approximately 670,000 deaths due to breast cancer are reported around the world [1]. Given that treatment options and outcomes vary greatly between early and late stages of the disease, stage at diagnosis is the most vital piece of information that a medical team has. It affects the decisions regarding surgery, the application of adjuvant chemotherapy or hormonal therapy, and the nature of post-treatment monitoring. However, the manual staging process itself relies on a combination of multiple histopathological and clinical variables that might not be consistently logged in different cohorts, generating a need for advanced methods of data processing that would facilitate the process of assessing stage-related risk based on parameters that are consistently tracked and recorded in clinical records.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, comprising long-term clinical and molecular records for 2,509 breast cancer patients drawn from tumor banks in the United Kingdom and Canada, has become a widely used benchmark for breast cancer machine learning research since its original publication [2]. Existing machine learning work on METABRIC has, however, concentrated mainly on predicting overall survival, molecular subtype, chemotherapy or hormone-therapy response and metastasis risk. Tegaw and Asfaw used the same 2,509-patient cohort together with SHAP-based explainability to compare survival outcomes between mastectomy and breast-conserving surgery [3], while [4] combined text mining of electronic medical records with several classifiers to detect breast cancer metastasis from blood-profile data [4]. Overall, the subject of predicting the actual tumor stage based on clinical and pathological variables has been largely overlooked, meaning that although the stage is part of the dataset, it is more relevant for treatment strategies compared to subtype or overall survival. This label has another specific challenge relating to its very uneven distribution across disease stages: extremely few patients in Stage 3 or Stage 4 compared to much bigger numbers in Stages 1 and 2. Ignoring this problem can lead to classifiers being biased towards majority classes resulting in misleading accuracy measurements despite much lower ability of classifiers to identify advanced illnesses.

Ensemble tree-based learners such as Random Forest [5] and gradient-boosted trees [6] are, in general, strong performers on small-to-medium tabular datasets of the kind found in clinical research, and recent large-scale benchmarking work has shown that they continue to match or outperform deep neural networks on tabular data of this scale [7]. At the same time, growing emphasis on transparency in clinical machine learning means that predictive performance alone is no longer considered sufficient; a model intended to support clinical decisions should also be able to explain which variables drove a given prediction, which is the role that SHapley Additive exPlanations (SHAP) plays in this study [8].

The overall workflow followed in this study begins with acquisition and cleaning of the METABRIC clinical/pathological subset, followed by class merging and imputation, an 80:20 stratified train-test split, baseline training of thirteen classifiers, systematic comparison of four class-imbalance correction techniques, hyperparameter tuning and cross-validation of the best pipeline, a supplementary deep-learning baseline, a sensitivity analysis across different train-test split ratios, and finally SHAP-based explanation of the selected model. Each stage is reported quantitatively in the Results and Discussion section so that the effect of every design choice on predictive performance can be traced.

Objectives of the Study:

The specific objectives are as follows:

To develop and test thirteen classical and ensemble classifiers for three-class tumor-stage classification using clinical and pathological attributes from the METABRIC cohort.

To assess the effectiveness of four resampling methods for minority class (advanced stage) detection.

To tune and cross validate the best performing pipeline and compare it with a feedforward neural network; and

To apply SHAP to check that the model's predictions align with known clinical staging criteria.

Novelty Statement:

The novelty of this research is to combine resampling comparison, hyperparameter tuning, cross validation, and SHAP explainability for multi-class tumor stage prediction on the METABRIC cohort. It relies only on clinical and pathological data, not genomic data. This combination has not been reported before for stage prediction on this cohort.

Material and Methods:

Investigation Site:

This study has no physical field site. The data comes from METABRIC. This is a cohort of breast cancer patients treated in the United Kingdom and Canada [2]. The dataset is public on Kaggle [9]. It has 2509 patient records. Each record has 34 variables. These cover clinical checks, pathology reports, treatment history and molecular tests. This study kept only variables a clinician would have at diagnosis. Molecular variables and outcome variables recorded after treatment are also dropped. This includes survival time, survival status and relapse status. The goal is to build a model that reflects real clinical practice. This also makes the model useful in hospitals without routine genomic testing.

Material and Methods:

Seventeen clinical and pathological attributes are used as predictors. These are age at diagnosis, type of breast surgery, cancer type, tumor cellularity, chemotherapy status, estrogen receptor status, HER2 status, histologic grade, hormone therapy status, menopausal state, tumor laterality, positive lymph nodes, Nottingham Prognostic Index, progesterone receptor status, radiotherapy status, tumor size and histologic subtype. Tumor stage is the target label. Column names are cleaned into a simple lowercase format. Records with a missing stage label are removed. Records with a missing stage label were removed first, leaving 1,788 records; 24 records with an invalid Stage-0 label were then also removed, leaving 1,764 usable records across Stage 1–4. While stages 3 and 4 are not seen that often in the data, it is also important to note that the very reasons for merging these two stages are of clinical significance. Stage 4 has distant metastasis, while Stage 3 is locally advanced that generally requires a systemic therapy (chemotherapy, targeted therapy, or hormonal treatment) and not a surgical treatment like in the case of Stage 1 and 2 with more localized disease that is usually treated surgically. In a model focusing on classification of diseases, the difference between early disease and an advanced disease that calls for systemic treatment makes it possible to build the necessary diagnosis. However, the classification created cannot communicate in the same way the disease integrity within the advanced category where Stage 3 is still considered a disease that can be treated with an intention to cure via multimodal treatment while Stage 4 (with distant metastasis) will require mostly palliative treatment.

Imputation statistics (median for numeric, mode for categorical variables) are computed on the training split only and then applied to the test split to avoid data leakage. Categorical variables are one-hot encoded, avoiding the artificial ordinal relationships that label encoding would impose on nominal predictors. Numeric variables are standardized using parameters fitted on the training data. The split ratio is 80 to 20. Stratified sampling kept class proportions equal in both sets. The classes stayed imbalanced after merging. Stage 1 had 630 records. Stage 2 had 979 records. Advanced had 155 records. Four resampling methods are tested on the training data only. These are SMOTE, ADASYN, Borderline SMOTE and random under sampling [10][11]. SMOTE creates synthetic samples between minority points. ADASYN focuses on harder minority cases. Borderline SMOTE creates samples near the

class boundary. Random under sampling removes majority samples. Every resampling method is tested with every classifier. All models are tested on the real, untouched test set.

Thirteen classifiers are trained. These include Logistic Regression, K Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest [5], Extra Trees, Gradient Boosting, AdaBoost, XGBoost [6], LightGBM, CatBoost and a Multilayer Perceptron. A feedforward neural network was constructed utilizing TensorFlow and Keras. Its architecture was composed of 2 hidden layers of 64 and 32 neurons (with ReLU activation) and a dropout rate of 0.3 after the hidden layers followed by a softmax output layer. It was trained with the Adam optimizer, categorical cross-entropy loss, and a batch size of 32 for a maximum of 100 epochs, while utilizing early stopping (patience = 10 epochs, monitoring validation loss) and a 20% validation split. All train-test splits, resampling procedures, and model initializations used a fixed random seed (random_state = 42) to ensure reproducibility.

The best classifier and resampling pair are chosen by macro F1 score. This pair is tuned using grid search. Stratified 5-fold cross validation is used during tuning. Macro F1 is the scoring target. The tuned model is checked again with 5-fold cross validation to test its stability. SHAP is used to explain the model's predictions. This step showed that the features contributing most to predictions were clinically plausible, consistent with established staging criteria, rather than the model relying on spurious associations [8]. Model performance is measured with several metrics. These are accuracy, precision, recall, F1 score for each class, macro F1, and macro and weighted AUC. Macro F1 is the main metric for choosing the best model. It treats the small Advanced class equally with the larger classes.

Result and Discussion:

The initial distribution of the classes, which is shown in Figure 1, demonstrates the rather small representation of Stage 3 and Stage 4 within the dataset. It led to the combination of these two stages into one Advanced (Stage 3/4) class because of which the final distribution consisted of 630 records classified as the first stage, 979 records classified with the second stage, and 155 records classified with the advanced stage.

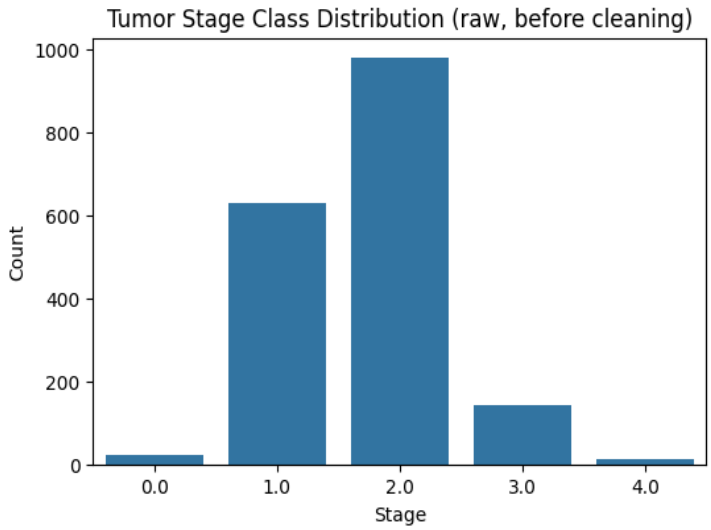


Figure 1. Distribution of tumor-stage records in the cleaned cohort (prior to merging Stage 3 and Stage 4).

Table 1 summarizes the eighteen retained clinical/pathological features together with their missingness in the raw 2,509-patient cohort; missingness ranged from below 1% for age at diagnosis to close to 29% for the tumor-stage label itself, which motivated the median/mode imputation and record-removal strategy described in the Methods section. Once the records with missing stage are removed and those recorded with Stage 1-4 are preserved, the sample consisted of 1764 records.

Table 1. Clinical and pathological features used as predictors, with missingness in the raw cohort.

Feature	Type	Missing (n)	Missing (%)
age_at_diagnosis	Numeric	11	0.4
type_of_breast_surgery	Categorical	554	22.1
cancer_type_detailed	Categorical	0	0.0
Cellularity	Categorical	592	23.6
Chemotherapy	Categorical	529	21.1
er_status	Categorical	40	1.6
her2_status	Categorical	529	21.1
neoplasm_histologic_grade	Numeric	121	4.8
hormone_therapy	Categorical	529	21.1
inferred_menopausal_state	Categorical	529	21.1
primary_tumor_laterality	Categorical	639	25.5
lymph_nodes_examined_positive	Numeric	266	10.6
nottingham_prognostic_index	Numeric	222	8.9
pr_status	Categorical	529	21.1
radio_therapy	Categorical	529	21.1
tumor_size	Numeric	149	5.9
tumor_other_histologic_subtype	Categorical	135	5.4
tumor_stage	Numeric (target)	721	28.7

As demonstrated in Table 2, a variety of baseline classifiers are evaluated using the original unbalanced dataset to determine both accuracy and macro-F1 measures. The results indicated Random Forest produced the most accurate identification of samples (accuracy = 85.6%, macro-F1 = 0.753), followed closely by the Extra Trees and Decision Tree methods, with KNN and Naive Bayes having the lowest performances. In spite of generally favorable results for all models, class-wise examination shows that the recall on the advanced-stage class is the weakest element among all baseline classifiers with values between 0.23 (SVM) and 0.65 (Decision Tree), indicating the accuracy paradox in imbalanced clinical classification problem where a classifier could deliver high accuracy overall, while performing poorly in recognizing clinically significant minority outcome.

Table 2. Baseline performance of selected classifiers on the original (unbalanced) training data, evaluated on the held-out test set.

Model	Accuracy	Macro F1
Random Forest	0.856	0.753
Extra Trees	0.802	0.704
Logistic Regression	0.810	0.703
Decision Tree	0.802	0.724
SVM	0.824	0.710
KNN	0.731	0.658
Naive Bayes	0.102	0.080
Model	Accuracy	Macro F1

The best-performing combinations of the different model-resampling methods that are evaluated in this study can be seen in Table 3, with Random Forest using ADASYN being the combination with the best results overall (85.3% accuracy, macro-F1=0.785), which is equal to 5.4 percentage points between the macro-F1 of the same model but with no resampling. The model-resampling selection is done based on the macro-F1 comparisons on the same held-out test set and not on a separate validation set. This is a significant limitation in the study; an ideal approach would involve using a separate validation set or a nested cross-

validation for model selection, with the held-out test data being used only once in the final objective evaluation.

Table 3. Top ten model-resampling combinations ranked by macro-F1 on the test set.

Model + Resampling	Accuracy	Macro F1
Random Forest + ADASYN	0.853	0.785
Random Forest + Borderline-SMOTE	0.841	0.779
MLP + ADASYN	0.830	0.777
XGBoost + ADASYN	0.841	0.762
LightGBM + SMOTE	0.853	0.759
XGBoost + Borderline-SMOTE	0.836	0.758
Random Forest + SMOTE	0.841	0.758
CatBoost + SMOTE	0.844	0.753
MLP + SMOTE	0.822	0.752
LightGBM + ADASYN	0.844	0.751

The figures in the confusion matrix of this setup (Figure 2) allow for estimating that the recall for the Advanced stage went up to almost 55% (17 out of 31 test cases have been correctly classified), which is significantly greater than the corresponding baseline figure for Random Forest that is found to be equal to 35%, while the amount of sacrifice in Stage 1/2 precision has been negligible. The use of oversampling methods such as SMOTE, ADASYN, and Borderline-SMOTE paired with tree ensembles appeared to be very useful across the top ten combinations as there are no top-performing combinations using random under sampling, which may be explained by the expectation that removing the records of the majority class leads to loss of important information in cases when the records of the majority class are not noisy.

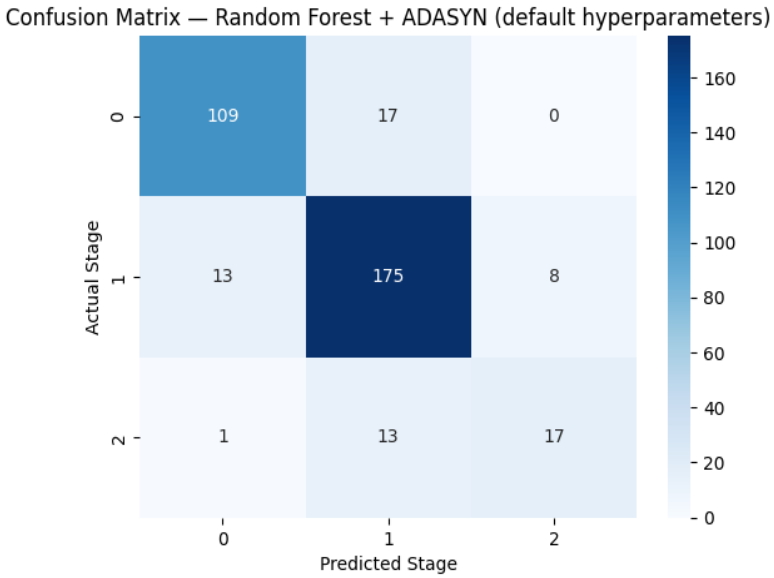


Figure 2. Confusion matrix for Random Forest + ADASYN on the held-out test set (0 = Stage 1, 1 = Stage 2, 2 = Advanced 3/4).

A grid search is carried out on the number of estimators, maximum depth, and minimum samples per split, leading to n_estimators = 200, max_depth = 10, and min_samples_split = 2 for the Random Forest + ADASYN pipeline. As shown in Table 4, the tuned settings gave a slightly lower accuracy (0.839 vs. 0.853) and macro-F1 (0.77 vs. 0.79) than the default configuration. Despite this small drop, the tuned model is retained as the final pipeline used for all subsequent analysis, including cross-validation, ROC/AUC, the NPI sensitivity check, and SHAP.

Table 4. Effect of grid-search hyperparameter tuning on the Random Forest + ADASYN pipeline.

Configuration	Precision (macro)	Recall (macro)	F1 (macro)	Accuracy
RF + ADASYN (default params)	0.81	0.77	0.79	0.853
RF + ADASYN (grid-search tuned)	0.77	0.77	0.77	0.839

Hyperparameter optimization is performed by using GridSearchCV with 5-fold stratified cross-validation and macro F1 as the scoring metric (Table 5). The tune is refitted independently in every cross-validation fold to avoid leakage. The configuration providing the best score is presented in the column "Selected Value".

Table 5. Grid-search hyperparameter space for the Random Forest + ADASYN pipeline.

Hyperparameter	Values tested	Selected value
n_estimators	100, 200, 300	200
max_depth	5, 10, 15, None (unlimited)	10
min_samples_split	2, 5, 10	2

The mean macro-F1 score calculated through stratified 5-fold cross-validation on the modified approach using the modified training dataset is 0.916 with a standard deviation of 0.009, thus confirming that the model performed successfully. This score is higher than 0.80 obtained in the initial test dataset, and the difference can be explained by the fact that the sample data are generated prior to performing the cross-validation instead of generating these samples independently within each fold; thus, the synthetic samples produced through the Borderline-SMOTE algorithm might be the same both in the training and evaluation fold, which results in getting optimistic assessments for the performance of the model. Thus, it is recommended to interpret the value-mentioned herein as merely an indication of the reliability of the model, not as the unbiased evaluation of the potential performance of the model within real-world conditions. Therefore, future research is advised to perform the resampling separately in all the cross-validation folds.

For evaluating how consistent the obtained performance is in view of the small Advanced-stage class (n = 155), the tuned pipeline is assessed again with the use of repeated stratified 5-fold cross-validation (10 repeats, 50 folds total) on the training data with the resampling refit happening within every fold separately in table 6. The average value of macro F1 is found to be 0.756 across all folds (95% CI ± 0.007; std. dev. 0.026 across 50 folds).

Table 6. Robustness of the tuned Random Forest + ADASYN model: single test-split estimate versus repeated cross-validation.

Evaluation	Macro F1	95% CI	Std. dev.
Single held-out test split (tuned model)	0.77	—	—
Repeated stratified 5-fold CV × 10 repeats (leak-free, 50 folds)	0.756	±0.007	0.026

The one-vs-all ROC curve that is generated for the tuned model on the test data, as set out in Figure 3 below, indicates that Stage 1 has strong discrimination (AUC = 0.976), which slowly decreases for Stage 2 (AUC = 0.928) and the Advanced category (AUC = 0.924). With a macro-average AUC value of 0.943 and a weighted-average AUC value of 0.945, it suggests that although the Advanced category is still the hardest to classify, one way of handling it would still be adjusting the classification threshold as opposed to relying on resampling.

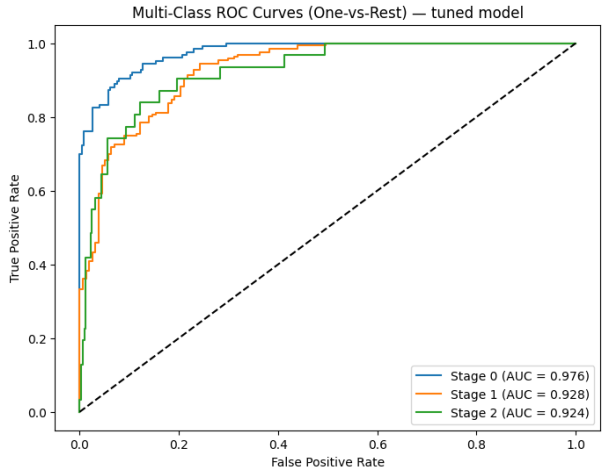


Figure 3. One-vs-rest ROC curves for the tuned Random Forest + ADASYN model on the test set.

Since model evaluations are dependent on the chosen train-test distribution, accuracy and macro-F1 outcomes are recalculated for Random Forest + SMOTE across four split ratios (Table 7). Both metrics stay within a narrow range across all four splits, with macro-F1 between 0.758 and 0.772 and accuracy between 0.831 and 0.843. Accuracy peaks at the 70:30 split (0.843), while macro-F1 is highest at 90:10 (0.772). Given how little the results vary across ratios, allocating more than 20% of the data to the test set brings no clear benefit, which supports the 80:20 ratio used for the main analysis. Figure 4 shows the model performance on different train-test split ratios..

Table 7. Accuracy and macro-F1 of Random Forest + SMOTE across different train-test split ratios.

Train:Test	Accuracy	Macro F1
90:10	0.831	0.772
80:20	0.841	0.758
70:30	0.843	0.763
60:40	0.841	0.761



Figure 4. Model performance across different train-test split ratios.

The feed-forward neural network achieved an accuracy of 80% with a macro-F1 score of 0.75 in the test data but did not outperform the optimized Random Forest pipeline on any of the metrics reported. This phenomenon is explained by the connection between the recent large-scale benchmarking studies where trees-based ensembles are found to perform as well as or better than deep neural networks on tabular datasets of the size used in this particular

case, especially considering the fact that a few thousand training instances are used instead of a few tens of thousands [7].

Table 8 summarizes and presents results obtained from thirteen classifiers including baseline macro-F1, resampled macro-F1, accuracy, macro-AUC and mean cross-validation F1 for the best-performing resampling technique for each individual model. Overall, Random Forest showed superior results in resampled macro-F1, accuracy and mean F1 of cross-validation, while Gradient Boosting achieved the highest value of macro-AUC (0.951). CatBoost, XGBoost and LightGBM demonstrated comparable results in the second tier. In contrast, Naive Bayes and KNN showed the worst and most consistent performance throughout the experiment, indicating their insufficient ability of modeling non-linear dependencies between clinical variables, which are successfully used by tree-based models.

Table 8. Full comparison of all thirteen classifiers after resampling with the best-performing method for each model, evaluated on the held-out test set.

Model	Baseline F1	Resampled F1	Accuracy	Macro AUC	CV Mean F1
Random Forest	0.753	0.785	0.853	0.943	0.749
MLP	0.758	0.777	0.830	0.924	0.691
XGBoost	0.709	0.762	0.841	0.944	0.754
LightGBM	0.724	0.751	0.844	0.944	0.753
Gradient Boosting	0.759	0.747	0.830	0.948	0.759
CatBoost	0.749	0.741	0.833	0.951	0.744
Extra Trees	0.704	0.722	0.802	0.907	0.663
SVM	0.710	0.718	0.768	0.931	0.680
Decision Tree	0.724	0.693	0.776	0.795	0.700
AdaBoost	0.766	0.677	0.742	0.882	0.707
Logistic Regression	0.703	0.640	0.686	0.911	0.658
KNN	0.658	0.637	0.671	0.850	0.610
Naive Bayes	0.080	0.080	0.102	0.707	0.081

The mean absolute SHAP values reveal that the identified lymph nodes, tumor size, and Nottingham Prognostic Index are the strongest contributors to predicting the tumor stage. Each bar is divided by classes allowing prediction contributions of particular variables to Stage 1, Stage 2, and Advanced stage to be revealed. The types of breast surgery, chemotherapy, and radiotherapy provided lower but similar contributions across the three classes. The findings suggested the earlier waterfall explanation and emphasized the absence of a unique variable being, overall, the leading one. Figure 5 shows the global SHAP features importance across tumor stage classes.

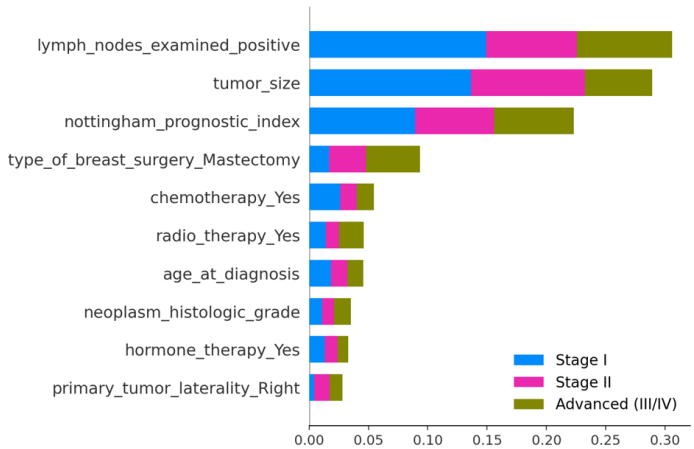


Figure 5. Global SHAP features importance across tumor stage classes.

Figure 6 portrays an example of a SHAP waterfall explanation for a test patient. It explains how each individual feature impacted the overall output of the model. We analyzed the entire dataset and found that the most important predictors of tumor stage are the number of positive lymph nodes examined, tumor size, and the Nottingham Prognostic Index. This finding is in line with clinical practice, since the Nottingham Prognostic Index is itself derived from tumor size, histological grade, and lymph-node status, which explains why these related variables dominate the SHAP ranking. Since the NPI derives from the tumor size, histological grade, and lymph node involvement all of which are also included as separate features, the NPI is removed in remodeling the model. By removing NPI from the model the macro F1 scores dropped from 0.770 to 0.727 while the accuracy dropped from 0.839 to 0.819. Table 9 shows the sensitivity analysis of model performance with and without NPI on test set.

Table 9. Sensitivity analysis of model performance with and without the Nottingham Prognostic Index (NPI) as a predictor, evaluated on the held-out test set.

Feature set	Accuracy	Macro F1	Macro F1 vs. with-NPI
With NPI (all 17 predictors)	0.839	0.77	—
Without NPI (16 predictors)	0.819	0.727	-0.043

The agreement between the model's learned feature importances and established clinical staging criteria is a step toward validating models intended for clinical decision support.

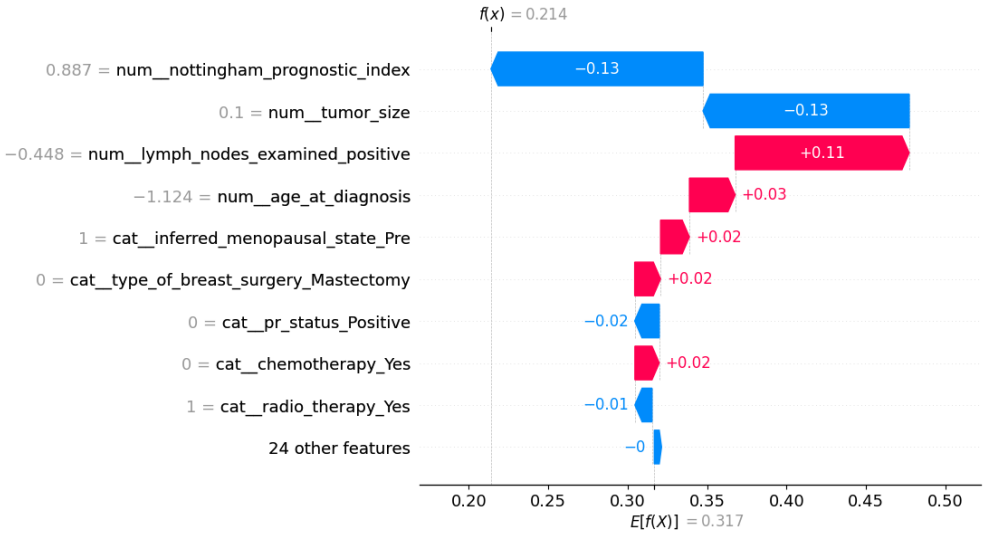


Figure 6. SHAP waterfall explanation for an individual test-set prediction, showing the contribution of each feature relative to the average model output.

A direct comparison between earlier studies based on METABRIC is difficult due to the difference in prediction targets; the survival prediction study by Ahmad et al. on a similarly-sized subset of METABRIC shows accuracy in the 70%–78% range obtained from seven classifiers while the study focusing on subtype classification on combined METABRIC predictors reports the accuracy of 71% [10]. The 83.9% accuracy and 0.770 macro-F1 score on three-class tumor-stage classification reported in this research are generally comparable or slightly higher than those previously measured; however, since the tasks are not strictly comparable, this comparison should only be taken as a rough sanity check and not as proving anything about the performance superiority.

Conclusion:

In this research, we developed and methodically compared thirteen classifiers to predict multiple breast cancer stages based on clinical and pathological parameters of the METABRIC cohort. A Random Forest classifier trained on ADASYN-resampled data gave the strongest pipeline. After tuning, it reached 83.9% accuracy. Macro F1 is 0.770. Macro

average AUC is 0.943. It clearly beats the untuned baselines and the feedforward neural network.

Resampling was beneficial for enhancing the detection of the Advanced stage class, which is an important class for clinical reasons. The improved performance of the Advanced stage class did not come to the detriment of performance of the majority classes. However, one limitation of this approach is that merging Stage 3 and Stage 4 means that the important distinction between curable and metastatic disease is lost. Future work with more patients in the Advanced stage may be able to model these two stages separately. SHAP analysis showed the model relied on nodal status, tumor size and the Nottingham Prognostic Index. These are already central to standard clinical staging. This gives the model some clinical plausibility.

These findings suggest a resampled, tuned and explained model like this could be useful. After external validation, it could act as a decision support tool. It could flag patients likely to have advanced stage disease for closer review. It should not replace histopathological staging.

Acknowledgement:

The authors thank the Department of Data Science & AI, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), for providing the resources and support needed to carry out this research.

Author's Contribution:

Muhammad Adeel Abid is responsible for conceptualization, methodology, and Analysis. Amna and Khadija are responsible for data collection and manuscript preparation.

Conflict of Interest:

The authors declare that there are no conflicts of interest.

Project details:

Not Applicable.

Refrences:

[1] Freddie Bray, Jacques Ferlay, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30207593/

[2] C. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nat. 2012 4867403*, vol. 486, no. 7403, pp. 346–352, Apr. 2012, doi: 10.1038/nature10983.

[3] B. B. Asfaw and E. M. Tegaw, "Explainable machine learning to compare the overall survival status between patients receiving mastectomy and breast conserving surgeries," *Sci. Reports 2025 151*, vol. 15, no. 1, pp. 10700-, Mar. 2025, doi: 10.1038/s41598-025-91064-2.

[4] M. Botlagunta *et al.*, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Sci. Reports 2023 131*, vol. 13, no. 1, pp. 485-, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[5] Leo Breiman, "Random Forests," *Mach. Learn.*, vol. 45, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[7] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," Jul. 2022, Accessed: Sep. 10, 2026. [Online]. Available: https://arxiv.org/pdf/2207.08815

[8] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Sep. 04, 2026. [Online]. Available: https://arxiv.org/pdf/1705.07874

[9] "Breast Cancer (METABRIC)." Accessed: Sep. 10, 2026. [Online]. Available: https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[11] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/J.INS.2019.07.070.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.