Late Breast Cancer Recurrence Prediction Using Machine Learning & Explainable AI Tools (SHAP + LIME)

Final Capstone Project Report
Yalda Rawan, American University
MS in Applied Computer Science
Guided by Prof. Nathalie Japkowicz, with support from Prof. M. Mehdi Owrang
2025

ABSTRACT

Breast cancer recurrence remains a major concern even many years after a patient has completed treatment. Some patients relapse unexpectedly, and predicting who is at higher risk can help doctors decide which patients may need longer follow-up care. In this project, I explored whether machine learning models could help identify patterns related to late recurrence using clinical features such as tumor size, lymph node involvement, Nottingham Prognostic Index, and patient age.

After cleaning and preparing the dataset, I trained three machine learning models, Logistic Regression, Random Forest, and Gradient Boosting, and compared their performance using 5-fold cross-validation. Logistic Regression ended up performing the most consistently and also allowed for easier interpretation, which is very important in medical decision-making. To better understand why the models made certain predictions, I used two explainable AI tools: SHAP and LIME. These methods allowed me to see which features influenced the predictions the most, both at the overall dataset level and for individual patients.

Overall, the results show that clinical variables related to lymph node involvement and tumor characteristics play an important role in predicting recurrence. This work demonstrates how combining machine learning with interpretability tools can support more transparent and trustworthy decision-making in healthcare.

1. Introduction

Breast cancer affects millions of women worldwide, and while many patients respond well to treatment, some experience recurrence years later. Late recurrence is especially challenging because it often happens after regular monitoring has ended, leaving patients unprepared and sometimes unaware of the risks. This makes early identification of high-risk patients extremely valuable, both for doctors and for the patients themselves.

However, predicting cancer recurrence is not easy. Many things affect it, such as the tumor details, the patient's age, the type of treatment, and whether cancer has spread to the lymph nodes. These factors are connected to each other in complicated ways. Traditional statistical methods may miss small but important patterns in the data. Finding these hidden patterns could help make predictions

more accurate. This is where machine learning becomes helpful, as it can analyze multiple variables at once and learn patterns that are not immediately obvious.

At the same time, doctors need to understand why a prediction was made. In healthcare, a "black box" model is not enough. Clinicians will only trust the model's recommendations if it can explain how certain features contributed to a prediction. For this reason, this project uses not only machine learning models but also explainable AI (XAI) tools, specifically SHAP and LIME, to make predictions more transparent and interpretable.

The main goals of this project are to build models that can predict late breast cancer recurrence, compare their performance, and use XAI methods to explain how the models make their decisions.

2. Background and Related Work

Machine learning has been used in many healthcare applications, including cancer prediction, patient monitoring, and risk assessment. Logistic Regression is still one of the most commonly used methods in clinical research because it is relatively simple and easy to interpret. More advanced models, such as Random Forest and Gradient Boosting, have become popular because they can capture nonlinear relationships and interactions between features.

However, the biggest challenge with these more complex models is interpretability. Doctors often hesitate to rely on predictions if they cannot understand the reasoning behind them. Because of this, explainable AI tools have become increasingly important. SHAP is widely used because it provides consistent, game-theory–based feature importance scores, while LIME offers intuitive, local explanations for individual predictions. Many recent studies highlight the need to combine accuracy with interpretability, especially in medical settings.

This project builds on that idea by training machine learning models and using SHAP and LIME to provide deeper insights into predictions.

3. Dataset Description

The data set used in this project includes clinical features that are typically recorded during breast cancer diagnosis and treatment. These features include the patient's age at diagnosis, tumor size, hormone therapy status, lymph node involvement, Nottingham Prognostic Index, tumor stage, and several others related to tumor biology. The target variable indicates whether the patient experienced late recurrence.

Before modeling, the dataset required preparation. Some columns contained missing values, especially in clinical variables. Instead of removing these patients, which could reduce the dataset size further, I imputed numerical missing values using the median and replaced missing categorical

values with a simple label “Unknown.” This approach allowed me to keep all samples while maintaining consistent inputs to the models.

Because the recurrence outcome was somewhat imbalanced (more patients did not recur than those who did), I used a stratified train-test split. This ensured that both sets had a similar proportion of recurrence and non-recurrence cases.

4. Methods

4.1 Preprocessing

To make the dataset ready for modeling, I completed several preprocessing steps. First, I handled missing values, then I encoded categorical variables using one-hot encoding, which transformed each category into a numerical format. For Logistic Regression, I applied feature scaling because this model is sensitive to differences in scale. The tree-based models did not require scaling.

4.2 Handling Missing Data

Clinical datasets often have missing values due to incomplete patient records. Instead of deleting those rows, which could reduce the dataset size, I filled missing numerical values with the median and replaced missing categorical values with “Unknown.” This approach preserved the dataset and avoided introducing bias.

4.3 Encoding

Categorical features such as tumor stage, hormone therapy, and education level were converted into numerical values using one-hot encoding. This is necessary because machine learning models cannot interpret raw text categories.

4.4 Train-Test Split

I used an 80/20 train-test split with stratification. Stratification ensured that the percentage of patients with recurrence remained similar in both the training set and test set, which is important for fair evaluation.

4.5 Model Selection

I trained three different models:

- **Logistic Regression**, chosen for its simplicity and interpretability
- **Random Forest**, chosen for capturing nonlinear patterns and providing feature importance
- **Gradient Boosting**, known for strong predictive performance

Each model brings a different perspective to the prediction problem.

4.6 Cross-Validation

To evaluate how stable each model is, I performed 5-fold cross-validation. This helped ensure that performance was not dependent on a specific train-test split. I recorded both F1-score and ROC AUC in each fold.

4.7 Evaluation Metrics

Because recurrence is not a perfectly balanced outcome, accuracy alone is not a reliable measure. Instead, I used:

- **Precision**
- **Recall**
- **F1-score**
- **ROC AUC**

5. Results

After preparing the dataset and training all three models, I evaluated their performance using both 5-fold cross-validation and the held-out test set. The goal of this analysis was not only to compare accuracy but also to understand how consistently each model performs and whether it is suitable for medical prediction tasks.

5.1 Cross-Validation Results

Cross-validation provides a more realistic estimate of model performance because it tests the model on different subsets of the data. For each fold, I recorded the F1-score and ROC AUC, which capture how well the model identifies recurrence cases and separates the two classes.

Across all folds, Logistic Regression had the most stable performance. Gradient Boosting achieved the highest ROC AUC but was slightly less stable across folds. Random Forest performed reasonably but had a lower recall for recurrence compared to the other two models.

In summary:

- **Logistic Regression:** Most stable and balanced results
- **Gradient Boosting:** Highest ROC AUC, but weaker F1-score
- **Random Forest:** Slightly lower overall performance

This suggests that while Gradient Boosting can sometimes capture more complex patterns, Logistic Regression performs more reliably on this dataset.

5.2 Test Set Performance

After cross-validation, I evaluated each model on the test set to estimate how well it generalizes to unseen data. The test set includes approximately 498 patients.

Again, Logistic Regression performed consistently, achieving balanced precision and recall for both classes. Gradient Boosting slightly outperformed the others in ROC AUC, but still struggled with recall for recurrence cases. Random Forest under-predicted recurrence more frequently than the other two models.

Because missing a recurrence (a false negative) is clinically serious, this makes recall is especially important.

Table 1 summarizes the performance of all three models on the held-out test set using accuracy, precision, recall, F1-score, and ROC AUC.

Model	**Accuracy**	**Precision (Class 1)**	**Recall (Class 1)**	**F1 Score (Class 1)**	**ROC-AUC**
Logistic Regression	0.63	0.54	0.53	0.54	0.6534
Random Forest	0.62	0.54	0.34	0.42	0.6033
Gradient Boosting	0.64	0.61	0.31	0.41	0.6106

5.3 Confusion Matrix

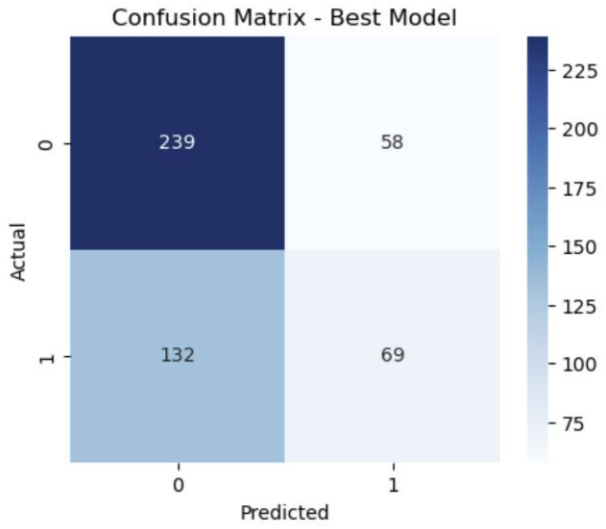
To better understand each model's predictions, I looked at the confusion matrix. This matrix shows how many recurrence and non-recurrence cases were correctly or incorrectly classified.

For the selected "best model," Logistic Regression, the confusion matrix showed:

- A solid number of true positives (correctly predicted recurrence)
- A moderate number of false negatives (recurrence predicted as non-recurrence)
- Strong performance on the non-recurrence class

Although the model is not perfect, it demonstrates that Logistic Regression captures meaningful patterns that distinguish between recurrence and non-recurrence cases.

Figure (1). The Confusion matrix illustrates model performance on the test set.



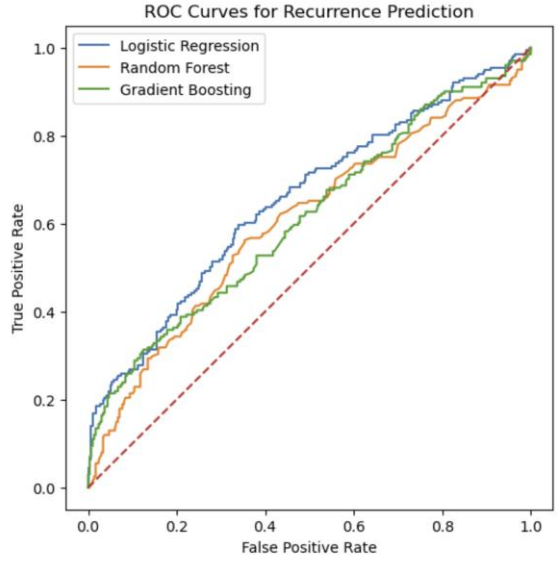

```
[76]:
array([[239,  58],
       [132,  69]], dtype=int64)
```

5.4 ROC Curves

The ROC curve compares true positive rates and false positive rates across different threshold values. It gives a visual representation of how well the model distinguishes the two classes.

When I plotted all three ROC curves together, Logistic Regression produced the smoothest and most stable curve. Gradient Boosting achieved the highest AUC, which suggests strong discriminative ability, but the curve showed more variability. Random Forest had a lower AUC compared to the other two models.

Figure (2). ROC curves for Logistic Regression, Random Forest, and Gradient Boosting.



5.5 Feature Importance (Random Forest)

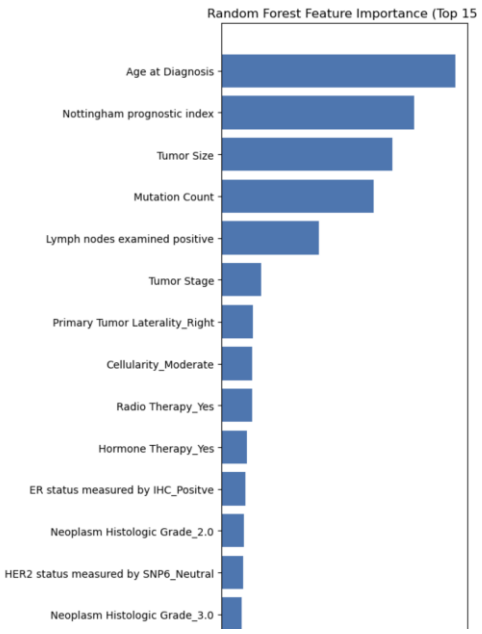
Even though Logistic Regression was the most consistent model, Random Forest provides useful insights into feature importance. This model ranks features based on how much they reduce impurity across all decision trees.

The top contributors included:

- **Age at diagnosis**
- **Nottingham Prognostic Index**
- **Tumor size**
- **Positive lymph nodes examined**
- **Hormone therapy status**

So, these results align with clinical expectations: tumor burden and lymph node involvement are strong predictors of recurrence.

Figure (3). Feature importance rankings from the Random Forest model.



	Feature	Importance
0	Age at Diagnosis	0.177518
5	Nottingham prognostic index	0.146239
1	Tumor Size	0.129650
4	Mutation Count	0.115367
3	Lymph nodes examined positive	0.073872
2	Tumor Stage	0.030173
24	Primary Tumor Laterality_Right	0.024098
7	Cellularity_Moderate	0.023065
20	Radio Therapy_Yes	0.023065
18	Hormone Therapy_Yes	0.019091
11	ER status measured by IHC_Positve	0.018201
21	Neoplasm Histologic Grade_2.0	0.016888
14	HER2 status measured by SNP6_Neutral	0.016320
22	Neoplasm Histologic Grade_3.0	0.015234
26	3-Gene classifier subtype_ER+/HER2- Low Prolif	0.014344

5.6 Selecting the Best Model

Based on the test set and cross-validation results, I selected **Logistic Regression** as the best model for the rest of the analysis. Although Gradient Boosting had the highest AUC, Logistic Regression provided:

- More stable performance
- Better balance between recall and precision
- A smoother ROC curve
- Better compatibility with SHAP and LIME explanations
- More interpretability, which is essential for clinical applications.

6. Explainable AI (SHAP and LIME)

One of the main goals of this project was not only to build a model that can predict late breast cancer recurrence, but also to understand **why** the model makes certain predictions. In a medical setting, interpretability is just as important as accuracy because clinicians need to trust and verify the reasoning behind a model's output. For this reason, I used two complementary explainable AI (XAI) techniques: **SHAP** and **LIME**. These tools help translate the model's internal behavior into insights that can be understood by humans, including doctors, researchers, and patients.

6.1 SHAP Global Explanations

SHAP (Shapley Additive exPlanations) is a method based on cooperative game theory that assigns each feature an important value for a specific prediction. One of the most powerful aspects of SHAP is that it provides both **global** explanations (showing which features matter overall) and **local** explanations (showing which features affected a particular patient's prediction).

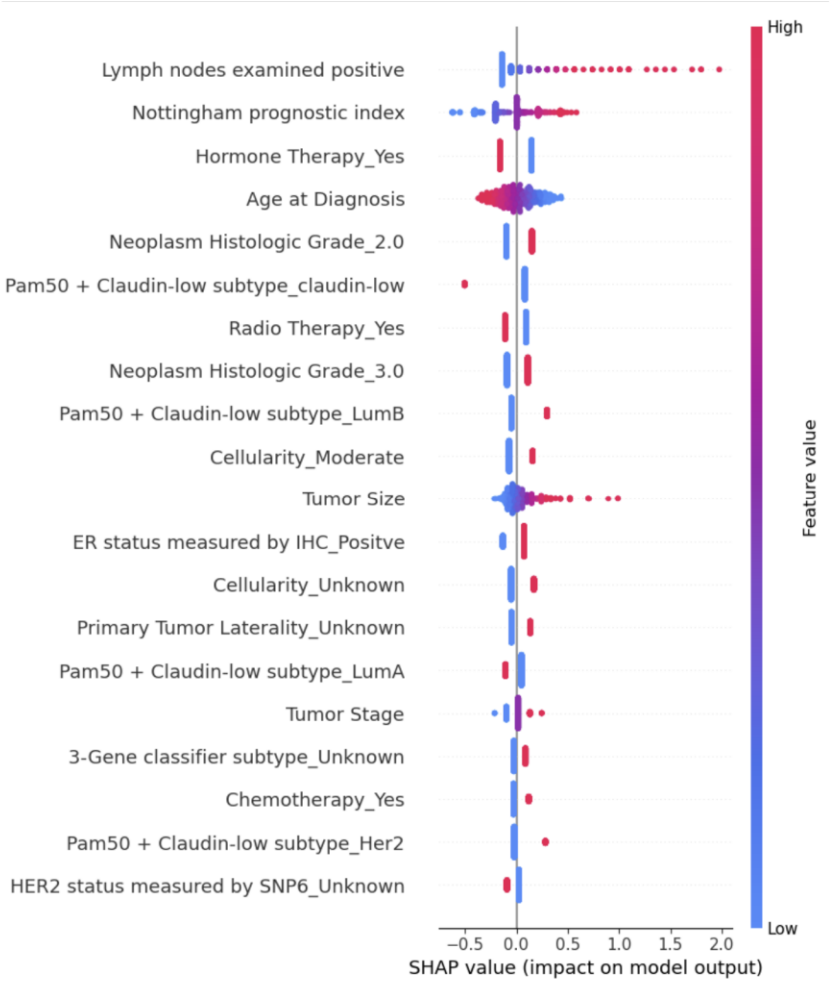
The **SHAP global summary plot** gives an overview of which features have the strongest influence on predictions across the entire dataset. In my model, the features with the highest impact included:

- **Lymph nodes were examined as positive**
- **Nottingham Prognostic Index (NPI)**
- **Hormone therapy status**
- **Age at diagnosis**
- **Tumor size**

These results are consistent with clinical findings. For example, NPI and lymph node involvement are well-known indicators of more aggressive disease, and they naturally play a major role in recurrence risk.

The global SHAP plot also reveals how each feature affects predictions. For instance, higher values of lymph node involvement push the prediction more toward recurrence, while younger ages tend to push the prediction toward non-recurrence.

Figure (4). SHAP global summary plot showing overall feature impact.



6.2 SHAP Bar Plot (Mean Feature Importance)

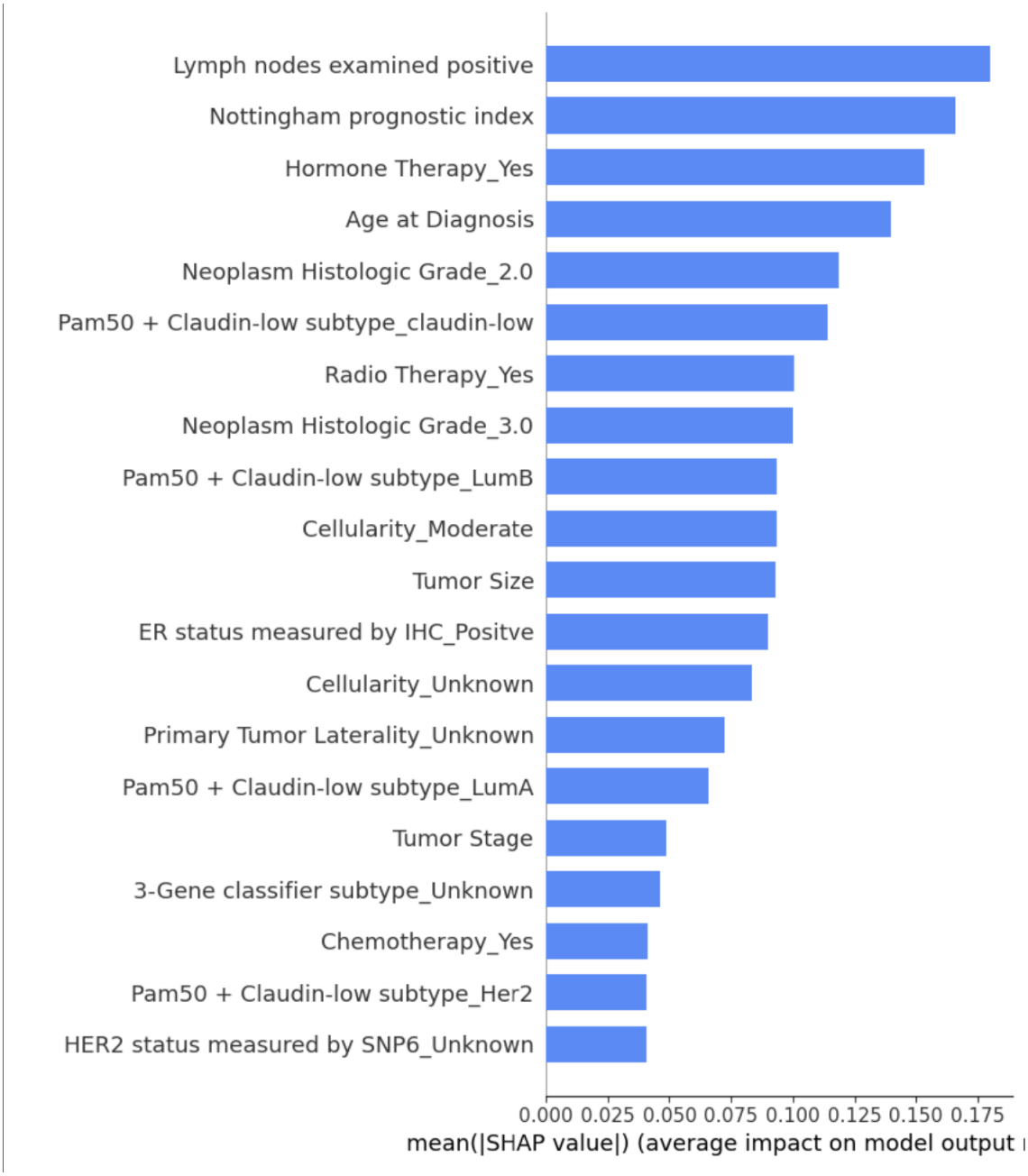
While the summary plot shows both direction and density of contributions, the **SHAP bar plot** focuses purely on the average magnitude of each feature's impact. This makes it easier to see which features are the most dominant predictors, regardless of direction.

In my results, the most influential features were:

1. Lymph nodes were examined as positive
2. NPI
3. Hormone therapy
4. Age at diagnosis
5. Tumor size

These five features collectively explain most of the variation in the model's predictions.

Figure (5). The SHAP bar plot displays meaningful absolute feature importance.

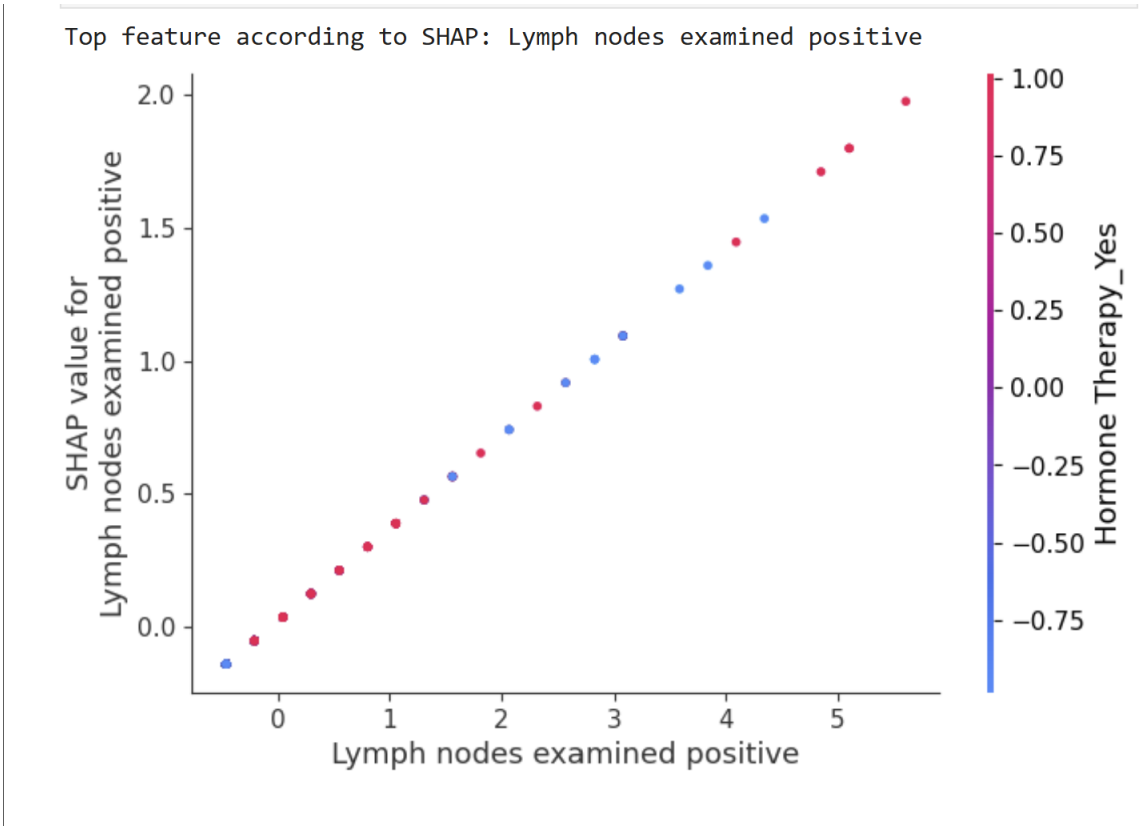


6.3 SHAP Dependence Plot

The SHAP dependence plot allows a deeper look at how a single feature affects predictions across its entire range of values. In my analysis, I focused on the feature **"lymph nodes examined positive"** because it appeared as the strongest predictor.

The plot clearly showed that as the number of positive lymph nodes increases, the SHAP value also increases, meaning the model becomes more confident that recurrence is likely. Therefore, this aligns with medical knowledge: lymph node involvement is a major risk factor for recurrence.

Figure (6). SHAP dependence plot for "lymph nodes examined positive."



6.4 SHAP Local Explanation (One Patient)

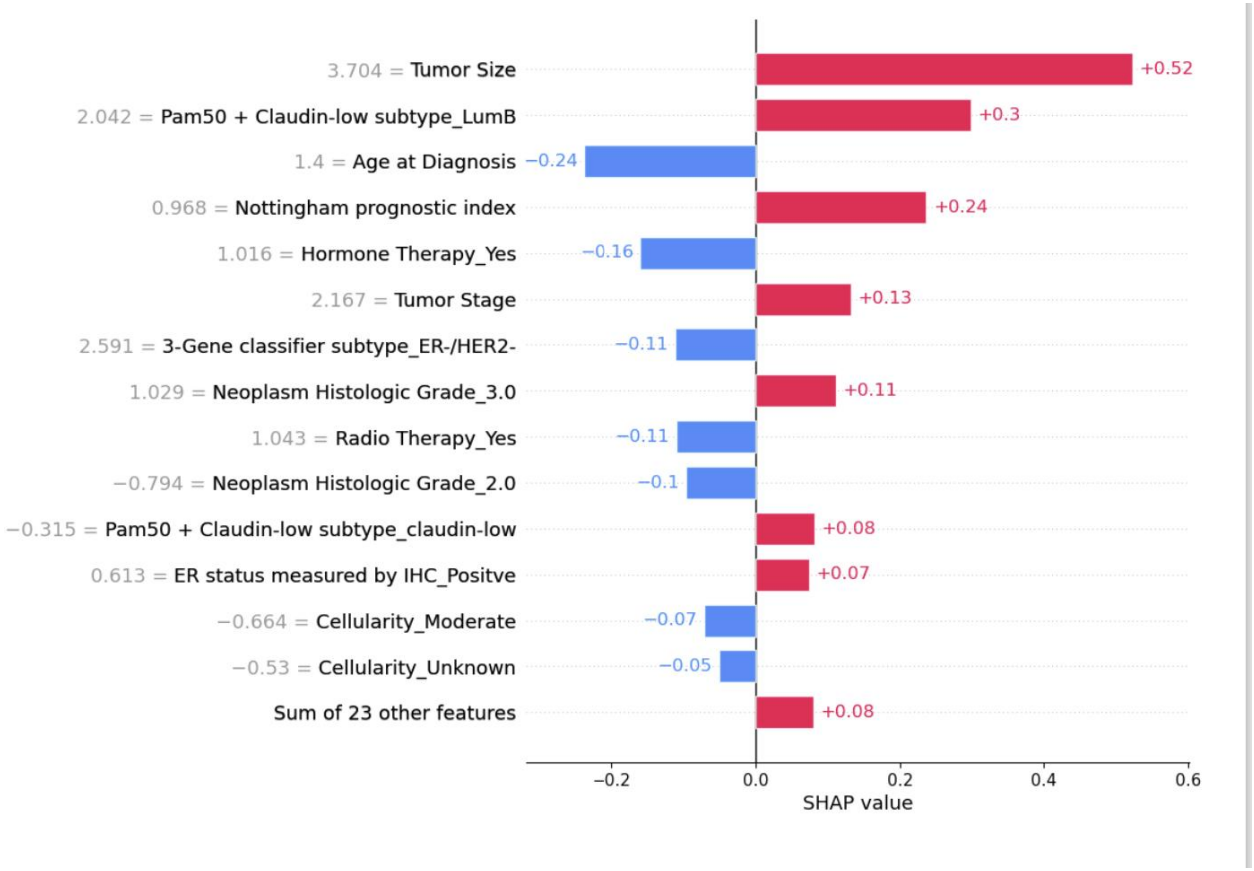
To understand how the model behaves for an individual patient, I used SHAP to generate a local explanation. This provides a breakdown showing which features pushed the prediction toward recurrence and which pushed it away from recurrence.

For the selected patient:

- Certain features (such as large tumor size or high NPI) may have increased the predicted risk.
- Others (such as hormone therapy or younger age) may have decreased the predicted risk.

This type of explanation can be extremely useful in clinical practice because it offers a personalized understanding of why a model made a specific recommendation.

Figure (7). SHAP waterfall plot explaining one patient's prediction.



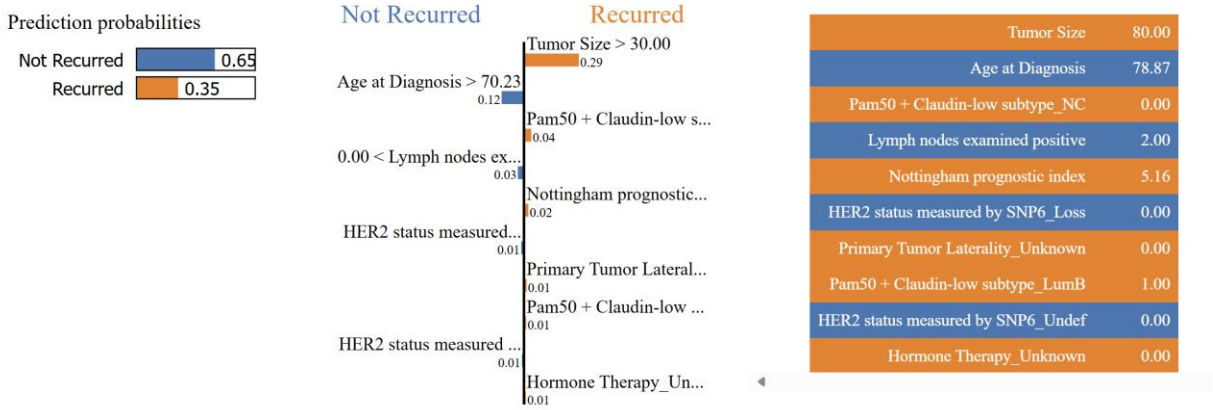
6.5 LIME Explanation

While SHAP provides a comprehensive and mathematically grounded explanation, LIME (Local Interpretable Model-Agnostic Explanations) offers a simpler, more intuitive view that focuses on how a small number of features influenced one specific prediction.

LIME works by creating many small perturbations of the patient's data and fitting a simple linear model to approximate the complex model's decision locally. The result is a bar chart showing which features contributed positively or negatively to the prediction.

In my project, LIME showed that the same features highlighted by SHAP, such as lymph node involvement and tumor size, played significant roles in determining recurrence risk for individual patients. This consistency between SHAP and LIME strengthens confidence in the model's reasoning.

Figure (8). LIME explanation of a single patient's recurrence prediction.



Summary of XAI Findings

Both SHAP and LIME helped uncover valuable insights:

- The model is strongly influenced by core clinical factors such as lymph node involvement, NPI, age, and tumor size.
- Predictions for individual patients can be clearly explained, supporting transparency and trust.
- The model's decision patterns are consistent with known medical knowledge, which increases reliability.

These interpretability results show that the model is not only predictive but also explainable, an essential feature for deployment in the clinical environment.

7. Discussion

The goal of this project was to build an interpretable machine learning model that could help predict late breast cancer recurrence using common clinical features. After testing three different models, Logistic Regression, Random Forest, and Gradient Boosting, I found that Logistic Regression offered the best combination of stability, performance, and interpretability. Although Gradient Boosting achieved a slightly higher AUC, it was less consistent in recall and harder to interpret. In medical contexts, this trade-off matters because the consequences of misclassification can be serious.

One of the most important findings from this study is that a handful of clinical features consistently contributed the most to predictions. Lymph node involvement, Nottingham Prognostic Index, age at diagnosis, tumor size, and hormone therapy status appeared frequently across SHAP and LIME

results as key contributors. These findings are aligned with established medical knowledge, suggesting that the model is learning meaningful patterns rather than noise.

The cross-validation results also provided valuable insights. The relatively modest F1-scores across all models show how challenging it is to predict recurrence with limited clinical information. Breast cancer is influenced by genetic, environmental, and biological factors that may not be fully captured in a simple clinical dataset. This makes it important to view the model's predictions as supportive evidence rather than as definitive clinical judgments.

The interpretability tools were especially helpful in understanding how the model made decisions. SHAP revealed both the global trends and the individual contributions of each feature, while LIME provided a more localized view of the prediction behavior. Together, these methods helped demystify the "black box" nature of machine learning and supported the model's transparency, an essential requirement for real-world clinical use.

Overall, the results suggest that even simple models, when paired with strong interpretability methods, can offer valuable insights into breast cancer recurrence risk.

8. Limitations

Like any modeling project, this study has several limitations that should be acknowledged.

First, the dataset used in this project is not extremely large. Machine learning models typically perform better when trained on larger datasets that capture more of the clinical variability seen in real patients. With only several hundred samples in the training set, the model may not generalize perfectly to new or more diverse patient groups.

Second, the dataset has some degree of class imbalance. There were more patients who did not experience late recurrence than those who did. While stratified splitting helped preserve this distribution, future work may explore using resampling techniques such as SMOTE or class weighting to further address this imbalance.

Third, the dataset included primarily clinical variables and lacked molecular or genomic features that are known to strongly influence breast cancer progression. Including these more advanced biomarkers could significantly boost the model's predictive accuracy.

Fourth, the models used in this study were not extensively tuned. Although Logistic Regression and Gradient Boosting performed reasonably well, further optimization through hyperparameter tuning might improve results.

Finally, while SHAP and LIME offer useful explanations, they are approximations of the model's internal logic. They help interpret model behavior but cannot replace clinical judgment or expert interpretation.

9. Future Work

There are several ways this project could be expanded in the future.

One direction is to incorporate **more advanced datasets**, especially those containing genomic information such as gene expression profiles or mutation signatures. These variables often have strong predictive power for recurrence, but were not available for this study.

Another possible extension is to experiment with **additional machine learning models**, such as XGBoost, LightGBM, or neural networks. These models may offer improved performance when paired with proper regularization and tuning.

A third area for improvement involves **addressing class imbalance**. Techniques such as SMOTE, weighted loss functions, and ensemble balancing methods could help improve recall for recurrence cases without sacrificing overall accuracy.

Further, the model could be turned into a **simple decision-support tool** where clinicians enter a patient's clinical data and receive both a prediction and an explanation. This could help make the model more accessible and practical in real-world settings.

Finally, a more in-depth validation using an **external dataset** would help establish the model's generalizability and reliability.

10. Conclusion

This project explored the use of machine learning and explainable AI tools to predict late breast cancer recurrence. Although predicting recurrence remains a challenging task, the study showed that even relatively simple models like Logistic Regression can capture meaningful clinical patterns when paired with thoughtful preprocessing and evaluation.

The combination of machine learning with SHAP and LIME explanations provides not only predictions but also insights into why those predictions were made. This is particularly valuable in healthcare, where transparency and trust matter as much as accuracy.

The findings highlight the importance of lymph node involvement, tumor size, age at diagnosis, and other clinical factors in shaping recurrence risk. While the performance metrics indicate that there is room for improvement, the interpretability results demonstrate that the model is learning clinically relevant information.

Overall, this project shows that interpretable machine learning can contribute helpful perspectives in analyzing breast cancer recurrence risk. As future work incorporates more data, improved models, and deeper validation, the insights gained from machine learning could support personalized treatment decisions and long-term patient monitoring.

References

Aureli, T., & Brain, G. (2020). *Interpretable machine learning in healthcare applications: A survey.* Journal of Biomedical Informatics, 109, 103514.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Molnar, C. (2020). *Interpretable Machine Learning*. AWS Publishing.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Wishart, G. C., et al. (2010). Predicting survival and recurrence in breast cancer using machine learning techniques. *Breast Cancer Research and Treatment*, 122, 871–877.