Ben-Gurion University of the Negev

Machine Learning and Data Mining

Course Project

Predicting Survival of Breast Cancer Patients Using Classification Algorithms

Lecturer:
Prof. Boaz Lerner

Students:
Orin Cohen

Shir Greif

Shira Jacob

27.02.2025

Abstract

Breast cancer remains a leading cause of cancer-related deaths worldwide, with survival outcomes influenced by numerous clinical and demographic factors. In this project, we leverage machine learning and data mining techniques to predict survival outcomes for breast cancer patients using a dataset derived from the SEER program. The dataset contains 4,024 records with 16 clinical and demographic features, enabling a comprehensive analysis of patient data.

We applied advanced machine learning algorithms, including Random Forest, XGBoost, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), to uncover patterns in the data and optimize predictive accuracy. Data preparation included rigorous handling of outliers, normalization, and feature engineering to enhance the models' performance. A Recursive Feature Elimination with Cross-Validation (RFECV) process was used to select the most relevant features, reducing redundancy and improving computational efficiency.

Evaluation metrics such as AUC-ROC, Recall, and F2 Score were utilized to address the challenges posed by the imbalanced dataset, where 84.7% of patients were classified as surviving and 15.3% as non-surviving. Among the models tested, XGBoost emerged as the most effective, achieving an AUC-ROC score of 0.8662 and outperforming other models with the highest Recall and F2 Score. This demonstrates its capability in managing class imbalance and accurately identifying high-risk patients. The findings from this project highlight XGBoost’s potential to aid personalized treatment planning and enhance patient outcomes in clinical settings.

keywords: Breast Cancer, Machine Learning, Binary Classification, Predictive Modeling, Imbalanced Dataset, Model Evaluation Metrics, Random Forest, XGBoost, SVM, MLP.

Table of Contents

List of Abbreviations

1. AUC-ROC – Area Under the Receiver Operating Characteristic Curve
2. ML – Machine Learning
3. MLP – Multi-Layer Perceptron
4. NCI – National Cancer Institute
5. RF – Random Forest
6. RFECV – Recursive Feature Elimination with Cross-Validation
7. SVM – Support Vector Machine
8. SVC – Support Vector Classification
9. SEER – Surveillance, Epidemiology, and End Results Program
10. VIF – Variance Inflation Factor
11. XGBoost – Extreme Gradient Boosting

Business understanding

Breast cancer is one of the most common cancers in women and a major cause of cancer-related deaths worldwide. It happens when cells in the breast grow uncontrollably, with common types being Ductal Carcinoma and Lobular Carcinoma. Although early detection and targeted treatments have helped improve survival rates, the disease's complexity still makes it hard to fully understand and effectively treat (Kibria et al., 2024).

In recent years, machine learning (ML) has provided new ways to study patterns in breast cancer. These algorithms can analyze large amounts of data to find hidden connections and improve predictions about patient outcomes. For example, using medical data, researchers have been able to predict survival rates more accurately and find important biomarkers that help in treatment planning (Zhang et al., 2024). Machine learning models like Support Vector Machines (SVM), Multilayer Perceptron (MLP), XGBoost, and Random Forest are essential tools for analyzing medical data. SVM is effective in classifying high-dimensional datasets, while MLP uncovers complex nonlinear patterns. XGBoost delivers high predictive accuracy, and Random Forest provides robust predictions and feature rankings. These models collectively enhance the understanding of clinical data, supporting better patient outcome predictions and treatment planning (Hirsch & Gilad-Bachrach, 2021; Xia et al., 2023).

In this project, advanced machine learning methods, as outlined above, will be applied specifically to the classification of breast cancer outcomes. These methods will be used to uncover patterns in the data and derive actionable insights. The goal is to develop reliable models that accurately classify survival outcomes, personalize treatments, and identify new risk factors associated with breast cancer.

Data understanding

The dataset we will work on in this project contains medical records of patients with Infiltrating Breast Carcinoma. The set is taken from the Kaggle website and is based on the November 2017 update of the National Cancer Institute (NCI) SEER program. The dataset consists of 4,024 records, each record describing a patient, and for each patient, 16 clinical and demographic characteristics were collected. There are 15 explanatory variables, 5 of which are continuous and 10 are categorical (3 of which are binary). The dependent variable is a binary variable describing whether the patient survived or not.

Details in
Table 1:

Feature	Type	Explanation	Value Space
Age	Continuous	Age of the patient at the time of diagnosis	[30,69]
Race	Categorical	Patient's race	White, Black, Other
Marital Status	Categorical	Patient's marital status	Married, Divorced, Single, Widowed, Separated
T Stage	Categorical	The size of the tumor and its spread	**T1** - Small tumor, **T2** - Medium-sized tumor, **T3** - Large tumor, **T4** - Spread to adjacent tissues
N Stage	Categorical	The degree of involvement of nearby lymph nodes	**N1, N2 , N3**- The higher the number (Ni), the more extensive the lymph node involvement.

6th Stage	Categorical	The total stage of the disease - combines the stages T and N and determines the severity of the disease	**IIA** - Medium-sized tumor with minimal lymph node involvement **IIB** - Larger tumor and/or limited lymph node involvement **IIIA** - Significant lymph node involvement without distant spread **IIIB** - Significant lymph node involvement with local spread to the chest wall/skin **IIIC** - Extensive lymph node involvement near the collarbone without distant spread
differentiate	Categorical	The degree of cell differentiation - the degree to which the tumor cells resemble normal cells. The lower the differentiation, the more aggressive the tumor is.	Undifferentiated, Poorly differentiated, Moderately differentiated, Well differentiated
Grade	Categorical	The grade of the tumor - rates the cancer cells between 1 and 3, where anaplastic grade (Grade IV) is the most aggressive.	1, 2, 3, anaplastic; Grade IV
A Stage	Categorical	Describes the development of the disease.	Regional - Tumor has spread only to nearby areas, such as adjacent tissues Distant- The disease has spread to distant organs such as the lungs, liver, or bones
Tumor Size	Continuous	The size of the tumor, measured in millimeters.	[1,140]
Estrogen Status	Categorical	Refers to the expression of estrogen receptors (ER) in tumor cells.	Positive, Negative
Progesterone Status	Categorical	Refers to the expression of progesterone receptors (PR) in tumor cells.	Positive, Negative
Regional Node Examined	Continuous	Number of lymph nodes examined	[1,61]
Regional Node Positive	Continuous	Number of lymph nodes affected	[1,46]
Survival Months	Continuous	Months of survival	[1,107]
Status	Categorical	Survival status	Alive, Dead

Table 1: Features explanation, type, and space

As part of the data analysis process, we generated box plots, histograms (Appendix 1), and descriptive statistics (Appendix 2) for all quantitative variables to examine their distribution and identify potential outliers. Outliers were identified based on the interquartile range (IQR) method, in which values outside the range Q1–1.5IQR to Q3+1.5IQR are defined as outliers. However, to maintain a sample that is as broad and representative as possible, we decided to widen the range for defining outliers (using a coefficient of 3 instead of 1.5), thereby reducing the number of values classified as outliers. This approach balances identifying true outliers with preserving the natural variability of the variables, especially in medical contexts, where extreme values may contain important and meaningful information. Outlier handling will be addressed during the data preparation phase.

Next, we conducted a general-level analysis of the data to identify trends and characteristics that may be relevant for training the model. As part of this process, we examined the correlations between continuous variables to detect high correlations that might indicate redundancy, potentially leading to excessive model complexity and reduced accuracy. As shown in Figure 1, there are no significant correlations between the variables, except for the correlation between Regional Node Examined and Regional Node Positive, which has a value of 0.41. This correlation is

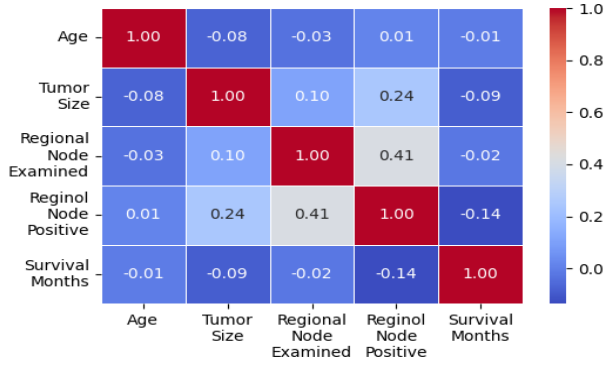


Figure 1: Correlation between continuouns variavles

logical, as having more lymph nodes increases the likelihood of more nodes being affected. In the next steps, we will consider creating a new, more informative feature to capture the relationship between these two variables.

Through our visual analysis, we derived important insights into the relationships between the variables. In Figure 2, we observed that among patients who died compared to those who survived, the variation in values is wider, with higher averages. This may suggest that the number of lymph nodes affected is a significant factor influencing clinical outcomes Figure 3 shows that earlier clinical stages (IIA, IIB, IIIA) are common at the Regional level, while advanced stages (IIIC, IIIB) dominate at the Distant level, demonstrating the progression associated with metastatic spread. Figure 4 shows that as lymph node involvement (N Stage) progresses, clinical stages of the disease also advance, with all N3 patients classified as stage IIIC, the most advanced stage. A similar trend is observed in Appendix 3, where larger tumors (T3, T4) correspond to an increased prevalence of advanced stages, such as IIIA, IIIB, and IIIC. These findings align with the fact that N Stage and T Stage are direct components of the 6th Stage, with advanced stages in these components indicating higher disease severity.

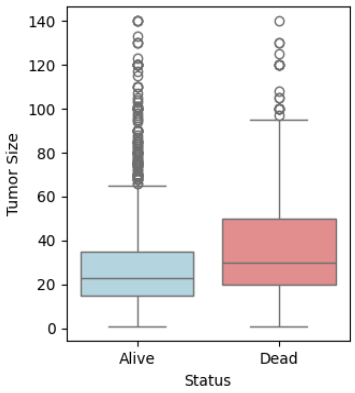


Figure 3: Boxplot of tumor size by survival status

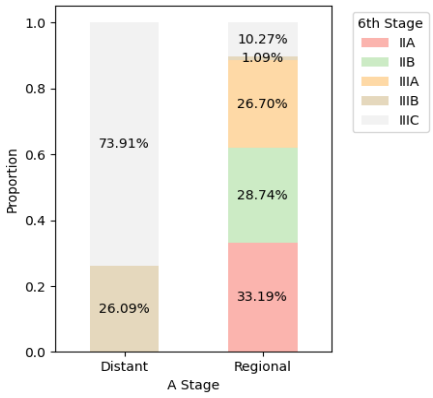


Figure 4: Stacked bar plot for A stage and 6th stage

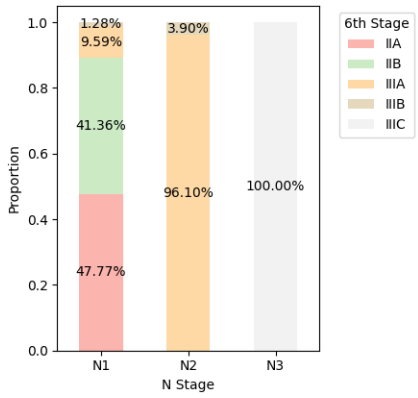


Figure 2: Stacked bar plot for N stage and 6th stage

Next, we performed a multicollinearity test to identify strong relationships between explanatory variables, which can make it challenging for the model to distinguish variable effects and may reduce its accuracy and stability. We used the Variance Inflation Factor (VIF) to evaluate correlations among continuous variables. VIF values between 1 and 5 indicate moderate and acceptable correlation, while values above 10 suggest significant issues. In our analysis, all VIF values were close to 1 (see Appendix 4), confirming no multicollinearity. Therefore, we proceeded with model development without further adjustments.

Data preparation and representation

Outlier Handling- Outliers were identified using the IQR method, testing two thresholds: a stricter multiplier of 3 and a more lenient multiplier of 1.5. To ensure robustness, we selected the stricter threshold (3*IQR), capping extreme values at the upper or lower bounds. This approach preserves most of the original data while minimizing the impact of extreme values that could bias the model's learning. Outliers were most observed in variables such as Tumor Size, Regional Node Examined, and Regional Node Positive, primarily in the majority class ("Alive"). For the minority class ("Dead"), outliers were considered critical as they often represent unique patterns essential for classification. No outliers were

detected for other quantitative variables, such as Age and Survival Months, under the chosen threshold (Table 2).

	Feature	Dead_Outliers	Alive_Outliers	Total_Outliers	Outliers_Percentage
0	Age	0	0	0	0.0%
1	Tumor Size	12	27	39	0.97%
2	Regional Node Examined	2	6	8	0.2%
3	Reginol Node Positive	63	76	139	3.45%
4	Survival Months	0	0	0	0.0%

*Table 2: Outlier analysis of quantitative variables based on the defined 3*IQR threshold*

Data Balancing- In our dataset, the target variable indicates whether a patient died from the disease or remained alive. The original distribution is highly imbalanced, with 84.7% in the majority class (alive) and 15.3% in the minority class (dead), approximately a 1:6.5 ratio. To address this, we experimented with multiple downsampling ratios, including 1:3 and 1:1. After evaluation, we found that maintaining a 2:1 ratio between the majority and minority classes provided the best performance while minimizing distortion of the original data distribution. This approach improves the model's ability to identify the minority class while preserving accuracy and reliability, which are essential in medical data. During the evaluation, we will consider that the data remains partially imbalanced for realistic representation.
Normalization- At this stage, we adjusted the variables to create a uniform scale in the model, ensuring efficient learning and preventing disproportionate influence from variables with different scales. We applied the Min-Max method to all quantitative variables, transforming the data to a scale of [0,1].
Discretization- For the age variable, values ranged from [30,69] before normalization. We categorized age groups as Young for ages [30,45], Middle-age for ages [46,59], and Elderly for ages [60,69]. Discretization was performed after normalization, assigning the categories the values 0, 1, and 2, respectively.
Data Transformation- Categorical variables without an ordinal relationship (e.g., Race and Marital Status) were converted to dummy variables using one-hot encoding to ensure all levels were included in the learning process and to allow the model to identify significant levels. Categorical variables with a logical order (e.g., N Stage or 6th Stage) were converted to consecutive numerical values to preserve their ordinal relationships. Binary categorical variables were encoded as 0 and 1, with the target variable Survival Status coded as 0 for "Alive" and 1 for "Dead."
Feature Engineering- A correlation was identified between Regional Node Examined and Regional Node Positive, prompting the creation of a new feature representing the ratio of positive lymph nodes to those examined in addition to the original features. This ratio captures the dependency between the variables and is expected to provide more informative input while reducing redundancy.

Feature Selection

To address our binary classification problem, we utilized Recursive Feature Elimination with Cross-Validation (RFECV) to select the most relevant features for our model. Using a Random Forest classifier and the AUC-ROC as the evaluation metric, RFECV iteratively removed the least important features based on stratified 5-fold cross-validation. The optimal set of features was identified (Shown in Appendix 5), and a new dataset was created containing only these selected features alongside the target variable, ensuring a more focused and efficient model training process.

Modeling

In this section, we evaluate the performance of Random Forest (RF), XGBoost, Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) models and present their results. The final models were assessed using metrics such as AUC-ROC, recall and F2 score, chosen to provide a balanced evaluation of precision and recall, addressing the challenges of the imbalanced dataset.

Random Forest

Random Forest, a non-parametric ensemble learning method, was chosen for our breast cancer survival classification problem due to its ability to handle complex, non-linear relationships, imbalanced data, and outliers. By combining multiple decision trees and averaging their predictions, it enhances accuracy, generalization, and robustness. To optimize the model, we applied RandomizedSearchCV with a predefined parameter grid containing 1,134 potential hyperparameter combinations. However, to ensure computational efficiency, only 60 diverse combinations were evaluated using 10-fold cross-validation, and the best-performing model was selected. Table 3 represents the Hyperparameters tuning:

Hyperparameter	Tested values	Reason for Selection
n_estimators	[100, 300, 500]	Represents the number of trees. We chose this moderate range to test both small and large values across combinations.
max_depth	np.arange (8, 15, 1)	Refers to the maximum tree depth. Large depth may lead to overfitting, while shallow trees might not predict well. We selected a diverse range to find the optimal depth.
criterion	['gini', 'entropy']	The splitting criterion. We tested two measures: Entropy (measuring data disorder) and Gini (measuring misclassification probability). Correct splits are critical.
max_features	[3, 5, 7]	Number of features considered in the tree. Fewer features simplify the model and reduce overfitting, while more features capture complex patterns. This range covers both cases.
min_samples_split	[2, 5, 10]	Minimum samples required to split an internal node. The chosen range ensures flexibility in avoiding unnecessary splits that could cause overfitting.
min_samples_leaf	[1, 2, 4]	Ensures the minimum number of samples per leaf is not too small. We included small values (1) for flexibility and larger ones (2, 4) to reduce noise.

Table 3: Hyperparameters tuning for Random Forest

XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful ensemble method that builds decision trees sequentially, optimizing for speed and accuracy. Its ability to handle complex, non-linear relationships and imbalanced data makes it well-suited for our breast cancer classification task. XGBoost's advanced regularization ensures robustness across diverse datasets. We utilized RandomizedSearchCV with a parameter grid of 6,480 combinations. To maintain efficiency, 30 combinations were evaluated using 10-fold cross-validation, selecting the best-performing model. Table 4 summarizes the hyperparameter tuning process.

Hyperparameter	Values Tested	Reason for Selection
learning_rate	[0.001, 0.002, 0.01, 0.1]	Controls the contribution of each tree. Lower values allow for more gradual learning, reducing the risk of overfitting.
max_depth	[1, 2, 3, 4, 5, 6]	Controls the depth of each tree. Lower values help prevent overfitting, while higher values allow learning complex patterns.
n_estimators	[50, 100]	Determines the number of trees. A balance between computational cost and model complexity.
subsample	[0.6, 0.8, 1.0]	Specifies the fraction of data samples used for training each tree. Helps reduce overfitting and improve generalization.
colsample_bytree	[0.6, 0.8, 1.0]	Specifies the fraction of features used per tree. Adding randomness improves generalization and reduces overfitting.

gamma	[0, 1, 5]	Minimum loss reduction is required for further splits. Larger values prevent overfitting by discouraging unnecessary splits.
Scale_pos_weight	[1, 2,3, 5, 10]	Addresses class imbalance by assigning higher weight to the minority class, improving the model's ability to detect positive cases without being biased towards the majority class.

Table 4: Hyperparameters tuning for XGBoost

SVM (SVC)

Support Vector Machine (SVM), implemented via SVC, is a robust algorithm that creates precise decision boundaries using hyperplanes, making it highly suitable for a breast cancer classification task. Its ability to handle complex, high-dimensional data ensures accurate classification. To optimize performance, we used RandomizedSearchCV with a hyperparameter grid of 144 combinations, evaluating 50 using 10-fold cross-validation. This approach balanced efficiency and accuracy, selecting the best model for robust and generalizable predictions. Table 5 summarizes the hyperparameter tuning process.

Hyperparameter	Values Tested	Reason for Selection
C	[0.01, 0.1, 1, 10, 100,1000]	Controls the trade-off between achieving a low error on the training set and minimizing the margin's complexity. A wide range ensures capturing both underfitting and overfitting scenarios.
kernel	['linear', 'rbf', 'poly']	Specifies the kernel function. Different kernels allow the model to adapt to various patterns and complexities in the data.
gamma	[0.0001, 0.001, 0.01, 0.1]	Determines the influence of a single training example. Lower values consider far points, while higher values focus on closer neighbors. Suitable for tuning model complexity.
tol	[1e-4, 1e-3]	Sets the tolerance for stopping criteria during optimization. Smaller values ensure precise convergence but increase training time, while larger values speed up training at the cost of precision.

Table 5: Hyperparameters tuning for SVM

MLP

Multi-Layer Perceptron (MLP) is a powerful deep learning algorithm designed to model complex, non-linear patterns through its interconnected layers of neurons. Its flexibility and capacity to handle high-dimensional data makes it particularly suitable for our breast cancer classification task. Using RandomizedSearchCV, we sampled 50 hyperparameter configurations from a space of 192 possible combinations, evaluating them with 10-fold cross-validation to balance efficiency and accuracy. Table 6 summarizes the tuning process.

Hyperparameter	Values Tested	Reason for Selection
hidden_layer_sizes	[(100, 50), (150, 100, 50)]	Specifies the number of neurons in each layer. Larger networks can capture more complex patterns, while smaller ones prevent overfitting.
activation	['relu', 'tanh', 'logistic']	Determines the activation function. relu is commonly used for efficiency, tanh and logistic for non-linear mappings
alpha	[0.001, 0.01, 0.1]	Regularization term to prevent overfitting. Higher values reduce model complexity by penalizing large weights.
learning_rate	['constant', 'adaptive', 'invscaling']	constant maintains a fixed learning rate, adaptive adjusts the rate based on performance, and invscaling gradually decreases the learning rate over iterations to fine-tune the model in later stages.
max_iter	[2000, 4000, 8000]	Specifies the maximum number of iterations for optimization. Higher values allow the model to converge fully.

Table 6: Hyperparameters tuning for MLP

Evaluation

Table 7 presents a summary of the model results, comparing their performance across key metrics such as AUC-ROC, F2 Score, and Recall. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) was chosen as the primary evaluation metric for its ability to assess the model's capacity to distinguish between classes across all thresholds. This makes it particularly effective for imbalanced datasets, like ours, as it balances sensitivity (True Positive Rate) and specificity (False Positive Rate), ensuring robust performance across various decision thresholds. Recall, which measures the model's ability to correctly identify all true positive cases, is critical in reducing false negatives and identifying high-risk patients. F2 Score, a variation of the F1 Score, places greater emphasis on Recall over Precision, prioritizing the correct identification of positive cases - an essential factor in medical contexts where missing a positive diagnosis can have severe consequences. Full results for training and test are available in Appendix 6.

Model	AUC-ROC	Recall	F2 Score
Random Forest	0.8595	0.7372	0.7335
XGBoost	0.8662	0.8813	0.7715
SVM	0.862	0.7372	0.719
MLP	0.867	0.661	0.6842

Table 7: Summary of the models' results

Discussion and Conclusions

In this project, we evaluated four machine learning models- Random Forest, XGBoost, SVM, and MLP- using AUC-ROC as the primary metric due to dataset imbalance. MLP achieved the highest AUC-ROC (0.867), slightly outperforming XGBoost (0.866), but XGBoost demonstrated superior Recall (0.8813) and F2 Score (0.7715), making it more reliable for identifying minority class instances. Random Forest and SVM performed similarly, with AUC-ROC scores of 0.8595 and 0.862 and identical Recall (0.737), but had lower F2 Scores (0.7335 and 0.719), indicating limitations in minority detection. Despite its higher AUC-ROC, MLP struggled with Recall (0.661) and F2 Score (0.6842), reflecting challenges with class imbalance. Unlike other models that handle imbalance through mechanisms like class weighting and localized optimization, MLP relies on gradient-based optimization, often favoring the majority class, making it less reliable in minimizing false negatives- critical in medical applications.

Feature importance analysis for XGBoost and Random Forest identified SurvivalMonthsNormalized and RegionalNodeRatio as key predictors, offering clinical insights. Additional features, such as 6thStageMapped and TumorSizeNormalized, had smaller contributions. The tree-based nature of these models provides interpretable insights, making them valuable in medical decision-making.

In conclusion, XGBoost's superior Recall and F2 Score make it the most suitable model for breast cancer classification. While Random Forest and SVM also performed well, they were less effective in detecting minority class instances, making XGBoost the best choice for minimizing false negatives in a medical context.

Future work could focus on expanding the dataset with diverse sources, such as international medical databases, to improve model generalizability. Incorporating unstructured data, like medical images and clinical notes, could provide deeper insights, while transitioning to a predictive model may better support early diagnosis and personalized treatment planning.

References

Kibria, M. K., Sifat, I. K., Hossen, M. B., Hassan, F., Mosharaf, M. P., & Hassan, M. Z. (2024). Identification of Bacterial Key Genera Associated with Breast Cancer Using Machine Learning Techniques. *The Microbe*, 100228.

Zhang, Y., An, W., Wang, C., Liu, X., Zhang, Q., Zhang, Y., & Cheng, S. (2024). Novel models based on machine learning to predict the prognosis of metaplastic breast cancer. *The Breast*, 103858.

Hirsch, R., & Gilad-Bachrach, R. (2021, July). Trees with attention for set prediction tasks. In *International Conference on Machine Learning* (pp. 4250-4261). PMLR.

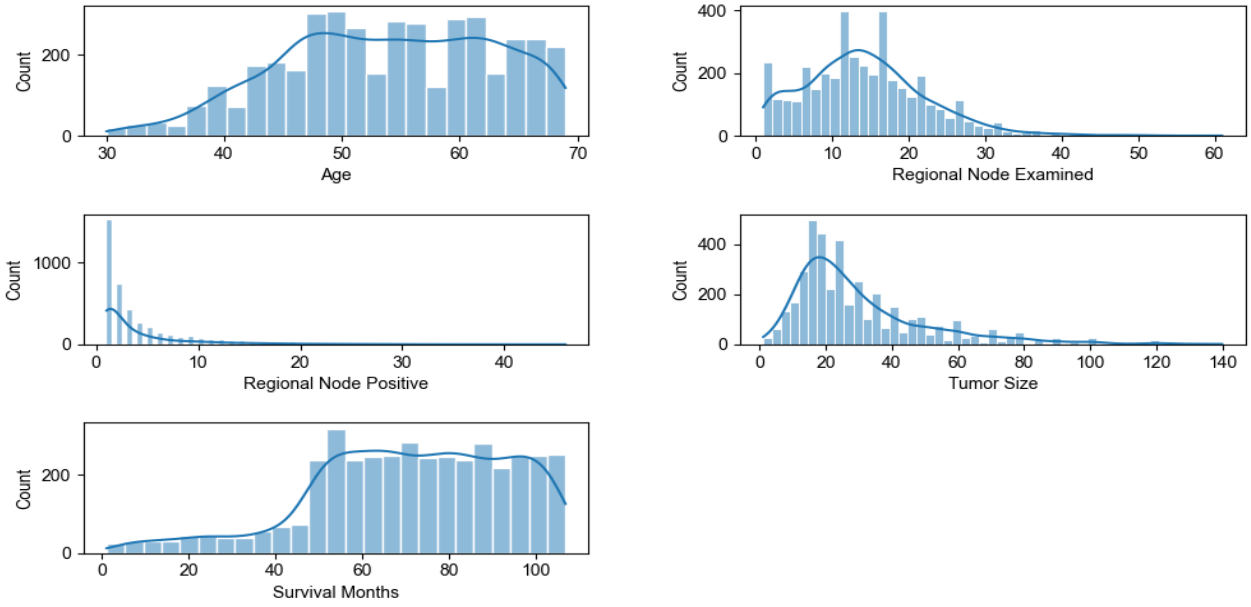
Xia, J., Zhang, L., Zhu, X., Liu, Y., Gao, Z., Hu, B., ... & Li, S. Z. (2023). Understanding the limitations of deep models for molecular property prediction: Insights and solutions. *Advances in Neural Information Processing Systems*, *36*, 64774-64792.

Appendices

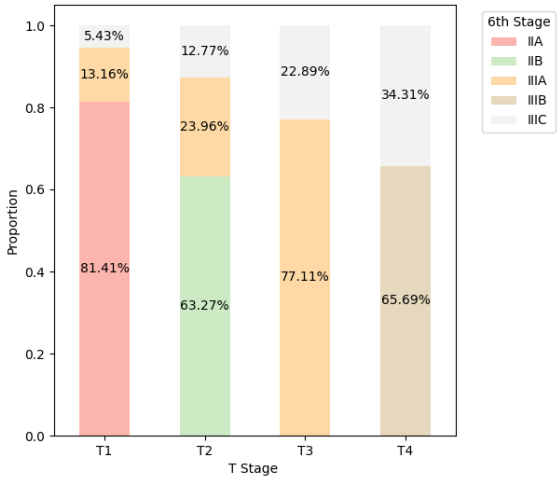
Appendix 1: *Descriptive statistics for quantitive variables*

	Age	Tumor Size	Regional Node Examined	Regional Node Positive	Survival Months
count	4024.0	4024.0	4024.0	4024.0	4024.0
mean	53.97	30.47	14.36	4.16	71.3
std	8.96	21.12	8.1	5.11	22.92
min	30.0	1.0	1.0	1.0	1.0
25%	47.0	16.0	9.0	1.0	56.0
50%	54.0	25.0	14.0	2.0	73.0
75%	61.0	38.0	19.0	5.0	90.0
max	69.0	140.0	61.0	46.0	107.0

Appendix 2: *Quantitive variables' distributions*



Appendix 3: *stacked bar plot for T stage and 6th stage*



Appendix 4*: Variance inflation factor (VIF)*

Variance Inflation Factor (VIF):

	Variable	VIF
1	Age	1.008951
2	Tumor Size	1.073227
3	Regional Node Examined	1.207765
4	Reginol Node Positive	1.287229
5	Survival Months	1.023430

***Appendix 5**: Feature selection*

```
Optimal number of features: 10
Selected features: Index(['Progesterone Status', 'Race_White', 'Marital Status_Married',
       '6th Stage_Mapped', 'differentiate_Mapped', 'Grade_Mapped',
       'Tumor_Size_Normalized', 'Survival_Months_Normalized', 'Age_Binned',
       'Regional_Node_Ratio'],
      dtype='object')
```

***Appendix 6:** Final models' results*

```
      Model Name    Metric  Train Set  Test Set
0  Random Forest   AUC-ROC   0.971972  0.859531
1  Random Forest    Recall   0.859438  0.737288
2  Random Forest  F2 Score   0.860821  0.733558
Best parameters: {'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 3, 'max_depth': np.int64(13), 'criterion': 'gini'}
```

```
  Model Name    Metric  Train Set  Test Set
0    XGBoost   AUC-ROC   0.887439  0.866240
1    XGBoost    Recall   0.913655  0.881356
2    XGBoost  F2 Score   0.801621  0.771513
Best parameters: {'subsample': 0.6, 'scale_pos_weight': 5, 'n_estimators': 100, 'max_depth': 1, 'learning_rate': 0.1, 'gamma': 1, 'colsample_bytree': 0.8}
```

```
  Model Name    Metric  Train Set  Test Set
0        SVC   AUC-ROC   0.872383  0.862019
1        SVC    Recall   0.755020  0.737288
2        SVC  F2 Score   0.742496  0.719008
Best parameters: {'tol': 0.0001, 'kernel': 'rbf', 'gamma': 0.01, 'C': 10}
```

```
  Model Name    Metric  Train Set  Test Set
0        MLP   AUC-ROC   0.875078  0.867030
1        MLP    Recall   0.660643  0.661017
2        MLP  F2 Score   0.680877  0.684211
Best parameters: {'max_iter': 2000, 'learning_rate': 'adaptive', 'hidden_layer_sizes': (100, 50), 'alpha': 0.01, 'activation': 'tanh'}
```

***Appendix 7:** Feature importance of Random Forest and XGBoost*