Predicting Breast Cancer Stages for Improved Early Detection and Treatment

Alexis Luevanos[1]

[1] School of Computer Science and Engineering
[2] California State University, San Bernardino

Abstract

This experimental project, titled ”Predicting Breast Cancer Stages for Improved Early Detection and Treatment,” aims to enhance breast cancer diagnosis through machine learning techniques.

By predicting stages I, II, III, or IV, this project seeks to improve early detection rates, leading to personalized treatment plans that can potentially lower cancer-related mortality. The dataset, sourced from Kaggle, comprises 4,024 instances and 16 attributes detailing patients' tumor characteristics such as tumor size, hormone receptor status, and tumor grade. This balanced dataset supports multi-class classification tasks.

Machine learning algorithms applied include Random Forest, Support Vector Machine. With accurate stage predictions, this project intends to contribute to more effective treatments and higher survival rates among breast cancer patients

1. Introduction

Breast cancer is a common health concern worldwide, where early detection and precise staging are important for improving patient outcomes. Machine learning has recently become a useful tool in healthcare, with the potential to enhance diagnostic accuracy and support early intervention.

This project aims to apply machine learning to predict breast cancer stages (I, II, III, or IV) based on patient tumor characteristics, such as size, hormone receptor status, and grade. Accurate stage predictions can assist physicians in creating personalized treatment plans, reducing unnecessary procedures, and potentially lowering cancer-related mortality rates. By enhancing diagnostic capabilities, this project supports the health and well-being of breast cancer patients.

2. Dataset

The dataset was sourced from the SEER Program by the National Cancer Institute (NCI) [1]. It includes records of 4,024 female patients diagnosed between 2006 and 2010 with infiltrating duct and lobular carcinoma of the breast. Patients with missing tumor sizes, lymph node data, or survival times below one month were excluded to ensure data quality.

Table 1 summarizes the key attributes of the dataset, including demographic, clinical, and pathological information. Table 2 presents the class distributions for important attributes like tumor stages, hormone receptor statuses, and survival outcomes. This dataset provides a balanced and comprehensive foundation for building predictive models to classify breast cancer stages.

2.1. Dataset Attributes

Table 1. Attributes of the Breast Cancer Dataset

Attribute	**Data Type**	**Example Values**
Age	Numeric	45, 58, 67
Race	Categorical	White, Asian, Black
Marital Status	Categorical	M, S, D
Tumor Stage (T Stage)	Categorical	T1, T2, T3
Node Stage (N Stage)	Categorical	N0, N1, N2
6th Stage	Categorical	IIA, IIB, IIIA
Differentiate	Categorical	Moderate, Poorly, Well
Grade	Categorical	1, 2, 3
A Stage	Categorical	Regional, Distant
Tumor Size	Numeric (cm)	1.5, 2.3, 3.1
Estrogen Status	Categorical	Positive, Negative
Progesterone Status	Categorical	Positive, Negative
Regional Node Examined	Numeric	10, 20, 30
Regional Node Positive	Numeric	1, 2, 3
Survival Months	Numeric	12, 36, 60
Patient Status	Categorical	Alive, Deceased

Table 2. Attribute Class Distributions

Attribute	Class	Instances
Race	White	3,413
	Other	320
	Black	291
Marital Status	Married	2,643
	Single	615
	Divorced	486
	Widowed	235
	Separated	45
Tumor Stage (T Stage)	T1	1,603
	T2	1,786
	T3	533
	T4	102
Node Stage (N Stage)	N1	2,732
	N2	820
	N3	472
6th Stage	IIA	1,305
	IIB	1,130
	IIIA	1,050
	IIIC	472
	IIIB	67
Differentiate	Moderately Diff.	2,351
	Poorly Diff.	1,111
	Well Diff.	543
	Undifferentiated	19
Grade	1	543
	2	2,351
	3	1,111
	Anaplastic (Grade IV)	19
Estrogen Status	Positive	3,755
	Negative	269
Progesterone Status	Positive	3,326
	Negative	698
A Stage	Regional	3,932
	Distant	92
Patient Status	Alive	3,408
	Deceased	616

2.2. Data Pre-processing

2.2.1. Random Forest

- Handling Missing Values: Rows with incomplete entries were excluded to ensure data integrity. But no missing values overall.
- Encoding Categorical Variables: Label encoding was used to transform categorical features into numerical values, making them compatible with the Random Forest algorithm.
- Train-Test Split: The dataset was split into training (80%) and testing (20%) sets, ensuring that class distributions were preserved.
- Class Balancing: Class weights were applied within the model to address any imbalance in the dataset.

2.2.2. Support Vector Machine (SVM)

- Handling Missing Values: Rows with missing values were removed to maintain dataset quality. But overall no missing values.
- Encoding Categorical Variables: Label encoding was applied to transform categorical features into numerical values.
- Feature Scaling: Numerical features were standardized using StandardScaler to improve SVM performance.
- Train-Test Split: An 80%-20% train-test split was performed with stratification to maintain class distributions in the splits.
- Class Balancing: Balanced class weights were applied within the SVM model to account for class imbalances.

3. Methodologies

3.1. Random Forest

Random Forest is a machine learning algorithm designed to enhance prediction accuracy and reduce the risk of overfitting by combining the outputs of multiple decision trees. Each tree is constructed using bootstrap samples of the training data, where subsets of the dataset are selected with replacement. At each node, the algorithm selects a subset of features randomly and determines the best split based on a specified splitting criterion, ensuring diversity among the trees and improved generalization to unseen data [**?**, 2–4].

3.1.1. Equation Function

The splitting criterion used in this project was Gini impurity, which quantifies the likelihood of incorrect classification of a randomly chosen element within a node. Gini impurity is defined as:

$$G = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k represents the proportion of samples in class k, and K is the total number of classes. The algorithm selects the split that results in the largest reduction in Gini impurity, leading to increasingly homogeneous nodes with respect to the target variable. The process iterates until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples per leaf [3–5].

3.1.2. Implementation

For this project, Random Forest Classifier module from Scikit-learn [3] was employed. Default parameters were used, including: estimators = 100, specifying the number of decision trees in the forest.criterion = Gini, using Gini impurity to evaluate splits.

Additionally, the model computed the importance of the characteristics by aggregating the reduction in Gini impurity achieved by each characteristic in all trees. These insights helped identify key predictors in the dataset, such as tumor size and hormone receptor status. The Random Forest classifier demonstrated its effectiveness in capturing patterns within the data, enabling an accurate classification of stages of breast cancer [2].

3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm designed to classify data by identifying the hyperplane that maximizes the margin between different classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class, referred to as support vectors [4–6].

3.2.1. Equation Function

For datasets that are not linearly separable, SVM employs kernel functions to map input features into a higher-dimensional space, where linear separation becomes feasible. The decision function for SVM is represented mathematically as:

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right),$$

where $K(x_i, x)$ denotes the kernel function, α_i are coefficients learned during training, y_i are the class labels, and b is the bias term. The Radial Basis Function (RBF) kernel was selected in this project due to its ability to model complex, non-linear relationships effectively [7] [6] [4].

3.2.2. Implementation

The implementation of SVM was carried out using the `SVC` module from Scikit-learn [7]. Default hyperparameters were utilized, including:C = 1.0, which controls the trade-off between maximizing the margin and minimizing classification errors.gamma = 'scale', which automatically scales the kernel coefficient as 1 / (number of features).

These default settings provided a robust baseline for classifying breast cancer stages, effectively balancing model flexibility and generalization [6] .

4. Experiments

4.1. Model Selection

Two machine learning models were selected for this project: Random Forest and Support Vector Machine (SVM). Random Forest was chosen for its interpretability and ability to handle both categorical and numerical data, while SVM was selected for its strong performance on non-linear data patterns.

4.2. Training

The dataset was preprocessed with categorical encoding, numerical scaling, and stratified splitting to maintain balanced class distributions. For this project, the 6th stage variable was chosen as the target to train the machine learning models, carry out predictions, and test their performance. This variable, based on the AJCC Cancer Staging Manual (6th Edition) guidelines [8] classifies breast cancer into stages I-IV, forming the basis for accurate staging and clinically relevant predictions. Class weighting was applied to address class imbalances.Models were trained on 80% of the data and tested on 20%, with 5-fold cross-validation ensuring robust evaluation.

Table 3 shows that Random Forest achieved an average accuracy of 99.90%, slightly outperforming SVM, which achieved 99.68%. These results highlight the strong predictive performance of both models, with Random Forest demonstrating superior consistency and robustness.

Model	Fold	Accuracy
Random Forest	1	1.0000
Random Forest	2	1.0000
Random Forest	3	0.9963
Random Forest	4	0.9988
Random Forest	5	1.0000
Random Forest	**Mean**	**0.9990**
SVM	1	0.9963
SVM	2	0.9988
SVM	3	0.9938
SVM	4	0.9975
SVM	5	0.9975
SVM	**Mean**	**0.9968**

Table 3. Cross-Validation Results for Random Forest and SVM Models

4.3. Evaluation

4.3.1. Random Forest

The Random Forest classifier achieved perfect classification across all classes, as shown in Table 4. The precision, recall, and F1-scores for all classes were 1.00, indicating that the model successfully identified every instance correctly. The confusion matrix further confirms this, showing no misclassifications. The model achieved an overall accuracy of 100% and an ROC-AUC score of 0.9999, underscoring its exceptional predictive capabilities.

Table 4. Classification Report and Confusion Matrix for Random Forest Classifier

Class	**Prec**	**Recall**	**F1-Score**	**Support**
IIA	1.00	1.00	1.00	275
IIB	1.00	1.00	1.00	222
IIIA	1.00	1.00	1.00	220
IIIB	1.00	1.00	1.00	5
IIIC	1.00	1.00	1.00	83
Accuracy	1.00 (805 total)			
Macro Avg	1.00	1.00	1.00	805
Weighted Avg	1.00	1.00	1.00	805

Confusion Matrix				
275	0	0	0	0
0	222	0	0	0
0	0	220	0	0
0	0	0	5	0
0	0	0	0	83

ROC-AUC Score: 0.9999

4.3.2. Support Vector Machine (SVM)

The Support Vector Machine (SVM) classifier also demonstrated high performance, as summarized in Table 5. While the precision and recall were perfect for most classes, slight misclassifications occurred for classes 3 and 4. The model achieved an overall accuracy of 99.75% and an ROC-AUC score of 0.9999.

The classification report shows: - Classes 0, 1, and 2 had perfect precision, recall, and F1-scores. - Class 3 had a recall of 0.92, leading to an F1-score of 0.96, which reflects minor misclassification. - Class 4 achieved a recall of 0.99 and an F1-score of 0.99, slightly lower than perfect classification.

The confusion matrix, also shown in Table 5, illustrates these misclassifications, where one instance of class 3 was classified as class 4, and one instance of class 4 was classified as class 3.

Table 5. Classification Report and Confusion Matrix for SVM Classifier

Class	**Prec**	**Recall**	**F1-Score**	**Support**
0	1.00	1.00	1.00	261
1	1.00	1.00	1.00	226
2	0.99	1.00	1.00	210
3	1.00	0.92	0.96	13
4	1.00	0.99	0.99	95
Accuracy	0.9975 (805 total)			
Macro Avg	1.00	0.98	0.99	805
Weighted Avg	1.00	1.00	1.00	805

Confusion Matrix				
261	0	0	0	0
0	226	0	0	0
0	0	210	0	0
0	0	1	12	0
0	0	1	0	94

ROC-AUC Score: 0.9999

4.4. Comparison of Models

The performance of the Random Forest and Support Vector Machine (SVM) classifiers is summarized in Table 6. Both models demonstrated excellent results, with Random Forest achieving perfect accuracy of 100% and SVM slightly lower at 99.75%. The weighted F1-scores for both models were 1.00, indicating their strong ability to correctly classify instances across all classes. However, the macro F1-score for SVM was slightly lower at 0.98, reflecting its misclassifications in less frequent classes such as 3 and 4. The ROC-AUC scores for both models were almost perfect, with Random Forest scoring 1.00 and SVM scoring 0.9999, highlighting they can distinguish classes.

Table 6. Model Evaluation Results

Metric	**Random Forest**	**SVM**
Accuracy	100%	99.75%
Macro F1-Score	1.00	0.98
Weighted F1-Score	1.00	1.00
ROC-AUC	1.00	0.9999
Training Time	Faster	Slower

The graph in Figure 1 illustrates the accuracy trends of the Random Forest and SVM classifiers across varying training set sizes. Random Forest consistently achieved perfect accuracy (100%) across all training set sizes, demonstrating its robustness and ability to generalize well to unseen data. In contrast, the SVM classifier showed a gradual improvement in accuracy as the training set size

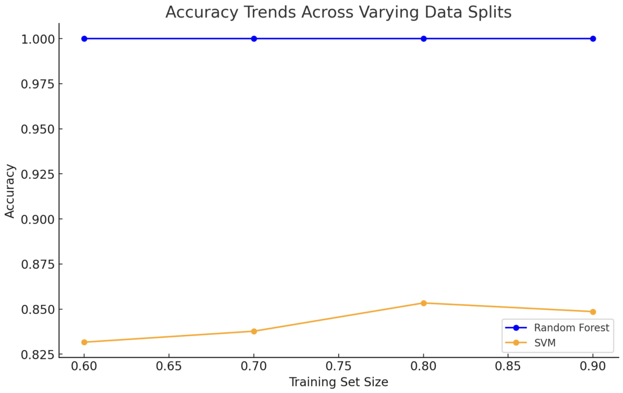

Figure 1. Accuracy Trends Across Varying Data Splits for Random Forest and SVM Classifiers

increased, reaching its peak performance with a training set size of 80%. However, the accuracy of SVM remained slightly below that of Random Forest at all training sizes, highlighting its sensitivity to data availability and reliance on scaled feature inputs. These observations emphasize the superior generalization capability of Random Forest and its suitability for this dataset compared to SVM.

4.5. Conclusion

This project shows the significant role of machine learning in advancing breast cancer diagnosis and staging, providing a foundation for early detection and personalized treatment planning. This project achieved highly accurate predictions of breast cancer stages (I, II, III, or IV). Random Forest excelled with perfect accuracy and faster training times, making it particularly suitable for clinical settings where efficiency is crucial. SVM, on the other hand, demonstrated robustness in handling complex patterns, highlighting its strength in datasets. Random Forest and SVM were both effective for predicting breast cancer stages. Random Forest achieved perfect accuracy and was faster to train, making it ideal for clinical applications. SVM demonstrated robust handling of complex patterns. Future work includes expanding the dataset and exploring deep learning models for further improvement.

Acknowledgments

I would like to express my gratitude to Dr. Zhang for their guidance and support throughout this project

References

[1] Surveillance E, End Results (SEER) Program NCI. Seer cancer statistics. Dataset provided by the SEER Program, updated November 2017, 2017. URL https://seer.cancer.gov/. Accessed December 3, 2024.

[2] Breiman L. Random forests. Machine Learning 2001; 45(1):5–32.

[3] Scikit-learn Team. Random forest classifier scikit-learn documentation. Scikit learn Machine Learning in Python 2024; URL https://scikit-learn.org/stable/. Accessed December 3, 2024.

[4] Mitchell TM. Machine Learning. McGraw-Hill, 1997. Referred to as [Mitchell].

[5] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. Free online version. Referred to as [Shwartz&David].

[6] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273–297.

[7] Scikit-learn Team. Support vector machines scikit-learn documentation. Scikit learn Machine Learning in Python 2024; URL https://scikit-learn.org/stable/. Accessed December 3, 2024.

[8] Surveillance E, Program ERS. Ajcc stage (6th edition), n.d. URL https://seer.cancer.gov/seerstat/variables/seer/ajcc- /6th/. Accessed: 2024-12-03.