Project Report
EAS508: Statistical Learning and Data Mining - 1

Instructor : Dr. Scott Broderick

Prediction of Recurrence of Breast Cancer Using Ensemble Learning Models

Abhiroop Ghosh
Kshitij Thakur
Rama Lahari Vedala
Shailesh Mahto

1. Abstract

This project focuses on predicting the recurrence of breast cancer through a systematic approach using machine learning approaches. Utilizing diverse patient data, including demographics, genetics, and treatment details, various ML models are trained and optimized.We begin with Exploratory Data Analysis (EDA) and rigorous Feature Selection to enhance model input. Subsequently, we employ various Machine Learning models and compare their performances. To optimize predictive accuracy, we implement a stacking technique, combining two or more models into an ensemble. This strategic amalgamation aims to harness the strengths of individual models, enhancing the overall predictive capacity for more robust recurrence predictions in breast cancer.

2. Introduction

Breast cancer is a malignant tumor that originates in the cells of the breast, typically in the ducts or lobules. It is the most common cancer among women globally. Research on breast cancer is essential to understand its complexities, improve treatment options, and ultimately save lives. Breast cancer affects millions, and ongoing research helps enhance prevention, early detection, and personalized treatment strategies. Early diagnosis can make the treatment easier and increase the chances of survival significantly.

Many individuals who have successfully overcome breast cancer remain at risk of recurrence. It has been studied that women with early-stage breast cancer commonly face local recurrence within the initial five years post-treatment. Typically, 7 percent to 11 percent of women in this category encounter a local recurrence within this timeframe.

3. Problem Statement

The primary objective of our project is to develop a novel ML model for the prediction of recurrence of breast cancer. When we use a robust ML model, the chances of prediction and diagnosis increase as compared to ONLY clinical medical examination of patients. Using ML models based on validated data can also help in identifying borderline cases. An important note is that the purpose of the ML models is NOT to replace the physicians and doctors working with a patient but only to aid them in the diagnosis and treatment plans.

4. Dataset

The dataset that we chose for this task is the ‘Breast Cancer Wisconsin (Prognostic)’ dataset which is available on the UCI repository.

The dataset consists of follow-up data of breast cancer patients where each row represents one case. It includes only those cases that exhibit invasive breast cancer and show no evidence of distant metastasis at the time of diagnosis.

The dataset has 33 features and 1 target variable of 193 samples, that is, 193 patients. The first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. The target variable is the outcome, whether or not the case showed recurrence or non-recurrence. The dataset has 1 missing value.

We also observed that the given dataset is imbalanced. Out of the 193 samples, 147 show non-recurrence, which is about 76% of the total cases.

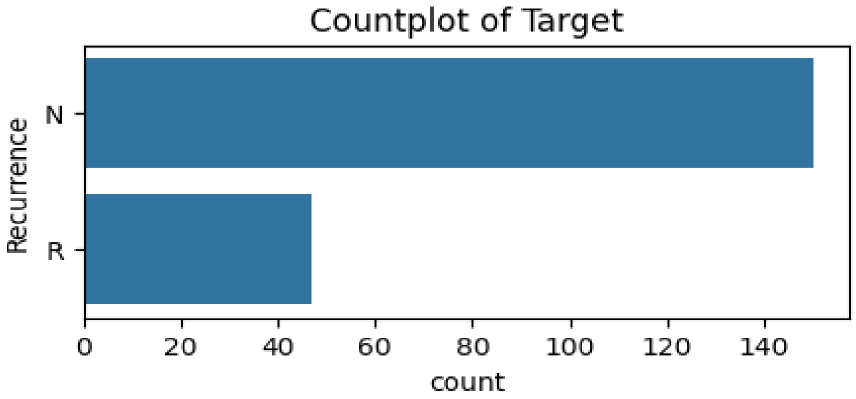
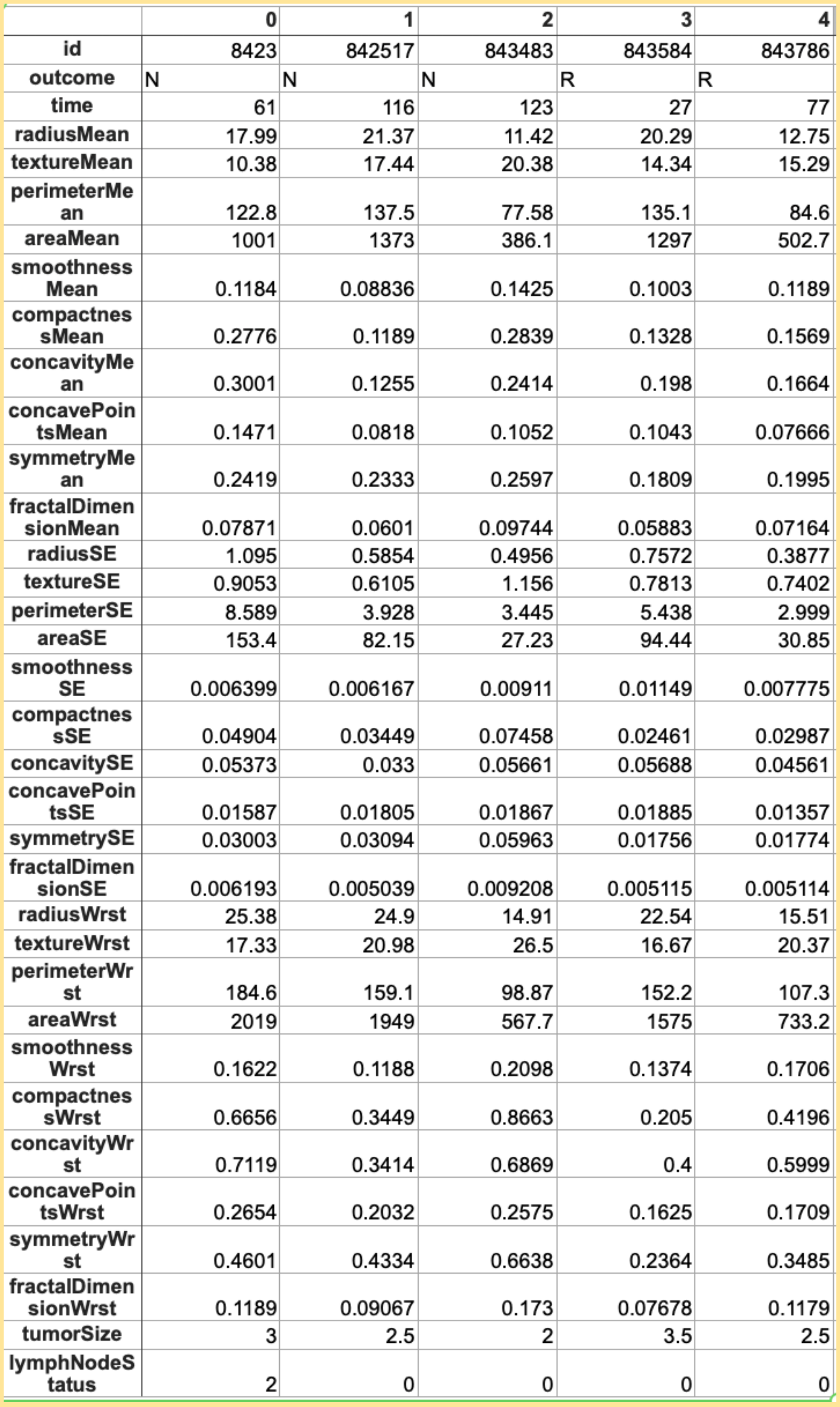


Image 1: Plot showing the count of Recurrence and Non-Recurrence cases

List of all the features used for building the model is represented in Table1 as follows -



	0	1	2	3	4
id	8423	842517	843483	843584	843786
outcome	N	N	N	R	R
time	61	116	123	27	77
radiusMean	17.99	21.37	11.42	20.29	12.75
textureMean	10.38	17.44	20.38	14.34	15.29
perimeterMean	122.8	137.5	77.58	135.1	84.6
areaMean	1001	1373	386.1	1297	502.7
smoothnessMean	0.1184	0.08836	0.1425	0.1003	0.1189
compactnessMean	0.2776	0.1189	0.2839	0.1328	0.1569
concavityMean	0.3001	0.1255	0.2414	0.198	0.1664
concavePointsMean	0.1471	0.0818	0.1052	0.1043	0.07666
symmetryMean	0.2419	0.2333	0.2597	0.1809	0.1995
fractalDimensionMean	0.07871	0.0601	0.09744	0.05883	0.07164
radiusSE	1.095	0.5854	0.4956	0.7572	0.3877
textureSE	0.9053	0.6105	1.156	0.7813	0.7402
perimeterSE	8.589	3.928	3.445	5.438	2.999
areaSE	153.4	82.15	27.23	94.44	30.85
smoothnessSE	0.006399	0.006167	0.00911	0.01149	0.007775
compactnessSE	0.04904	0.03449	0.07458	0.02461	0.02987
concavitySE	0.05373	0.033	0.05661	0.05688	0.04561
concavePointsSE	0.01587	0.01805	0.01867	0.01885	0.01357
symmetrySE	0.03003	0.03094	0.05963	0.01756	0.01774
fractalDimensionSE	0.006193	0.005039	0.009208	0.005115	0.005114
radiusWrst	25.38	24.9	14.91	22.54	15.51
textureWrst	17.33	20.98	26.5	16.67	20.37
perimeterWrst	184.6	159.1	98.87	152.2	107.3
areaWrst	2019	1949	567.7	1575	733.2
smoothnessWrst	0.1622	0.1188	0.2098	0.1374	0.1706
compactnessWrst	0.6656	0.3449	0.8663	0.205	0.4196
concavityWrst	0.7119	0.3414	0.6869	0.4	0.5999
concavePointsWrst	0.2654	0.2032	0.2575	0.1625	0.1709
symmetryWrst	0.4601	0.4334	0.6638	0.2364	0.3485
fractalDimensionWrst	0.1189	0.09067	0.173	0.07678	0.1179
tumorSize	3	2.5	2	3.5	2.5
lymphNodeStatus	2	0	0	0	0

Table 1: List of all the features

5. Methodology

5.1 Data Preprocessing

a. Missing Values - The dataset shows 4 missing values for ‘LymphNodeStatus’ for patients ‘844359, 854253, 877500, 947204’

a. We can use data imputation techniques like mean imputation, knn or linear regression imputation. Or we can simply remove the missing data. In our case, we chose to remove the rows with the missing data because there were only 4 values and removing the entire row preserves the statistical properties of the original data.

b. One-hot encoding - We labeled the recurrent cases as 1 and non-recurrent cases as 0.

5.2 Feature Selection

We plotted the correlation matrix and also utilized coefficients of a Logistic Regression model to sort features by importance and eliminate features not contributing to the model. Using these techniques, we brought down the number of features from 33 to 18.

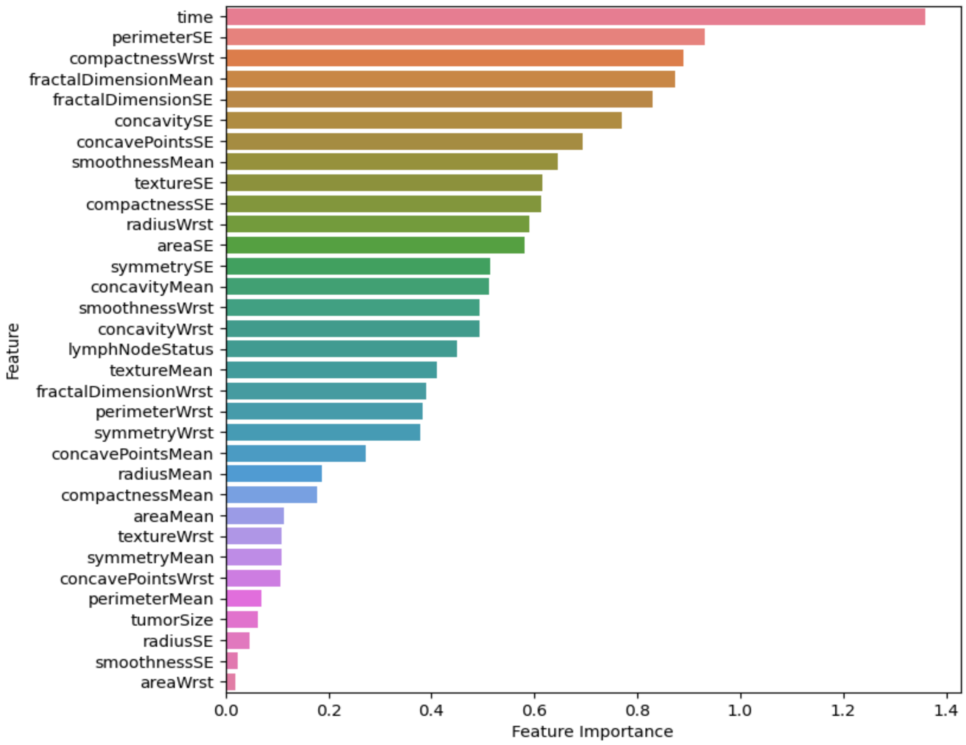


Image 2: This plot shows the coefficients for the Logistic Regression models

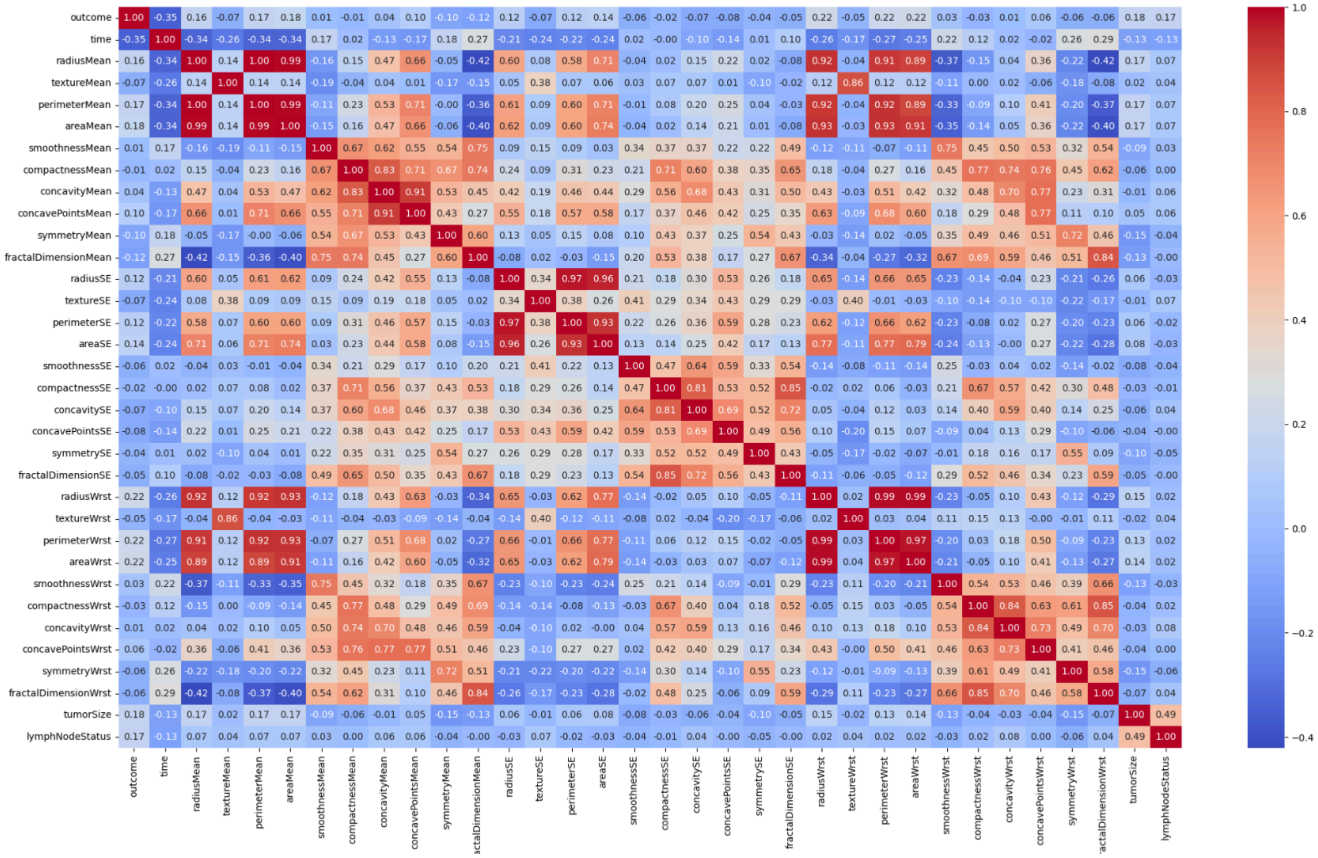


Image 3: Correlation Heatmap for all features

5.3 Model Building

After we got our desired set of features, we used a stratified test-train split to split the data into training and testing sets instead of a randomized split because of the imbalance in our data. Stratified split maintains the ratio of the target variable for training the model which results in reducing the bias and preserving the class distribution. We then proceeded to build multiple classification models to predict the outcome of a case as recurrent or non-recurrent.

The metrics that we used to determine the quality of the model are the train and test accuracy, sensitivity, and specificity. Train and test metrics were compared to judge the fit of the model. If the train metrics are significantly higher than the test metrics, then the model is said to be overfit. Also, given the criticality of the domain, we decided that not only the accuracy, but the sensitivity of the model also plays an important role. Ideally, the model should have a very high true positive rate and accuracy as well. Since our dataset is imbalanced, having 76%

non-recurrent cases, if a model simply predicts 0 for all cases, it will have a 76% accuracy and if it predicts 1 for all cases, it will have 100% sensitivity. Therefore having a good balance of both is very important. Additionally, different models will perform differently and therefore some may have a higher accuracy than others, some may have higher sensitivity and some may overfit or underfit.

Due to that, we decided to stack multiple models to make an ensemble that tries to capture the best aspects of different models to yield better results. We used soft voting to make ensemble models because it consists of combining the predicted probabilities assigned by each classifier to classes rather than the actual prediction of their class. These probabilities are averaged or weighted, and the class with the highest mean is selected as the final predictor. Finally, we used our ensemble models to predict and classify the sample as a recurrent case or a non-recurrent case.

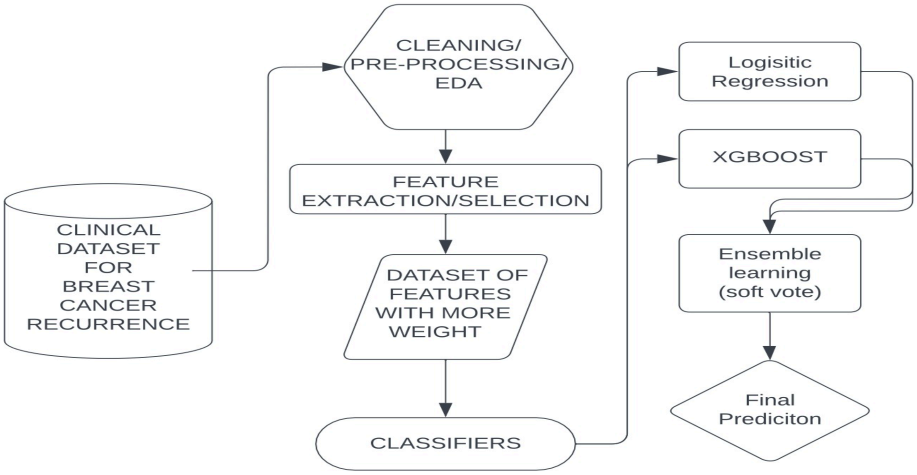

CLINICAL DATASET FOR BREAST CANCER RECURRENCE
CLEANING/ PRE-PROCESSING/ EDA
FEATURE EXTRACTION/SELECTION
DATASET OF FEATURES WITH MORE WEIGHT
CLASSIFIERS
Logisitic Regression
XGBOOST
Ensemble learning (soft vote)
Final Prediciton

Image 4: Flowchart

6. Results

Table 2 shows the results that we got for the models that we built. We picked standard classification models to predict the outcome for recurrence.

Model	Test Accuracy	Test Sensitivity	Test Specificity	Train Accuracy	Train Sensitivity	Train Specificity
XGBoost	89.74%	66.67%	96.67%	100.00%	100.00%	100.00%
Logistic Regression	87.18%	55.56%	96.67%	85.71%	51.35%	96.58%
Naive Bayes	69.23%	66.67%	70.00%	73.38%	56.76%	78.63%
MLP	84.62%	44.44%	96.67%	100.00%	100.00%	100.00%
Decision Tree	76.92%	44.44%	86.67%	100.00%	100.00%	100.00%
KNN	76.92%	44.44%	86.67%	85.71%	54.05%	95.73%
RF	84.62%	33.33%	100.00%	100.00%	100.00%	100.00%
SVC	84.62%	33.33%	100.00%	82.47%	29.73%	99.15%
GBM	79.49%	33.33%	93.33%	100.00%	100.00%	100.00%
Adaboost	74.36%	33.33%	86.67%	100.00%	100.00%	100.00%

Table 2: Metrics for individual models

We stacked these models to make an ensemble learner. From this table, we can observe that XGBoost gives us the best test accuracy and test sensitivity. However, looking at the train accuracy we can say that it is slightly overfitting. Also, we can see that Logistic Regression may not have an equally high testing accuracy or sensitivity, it does not overfit at all and still gives the second-highest accuracy. Based on these observations, our first stacked model was an ensemble of XGboost and Logistic Regression. We aimed to achieve even higher testing metrics while trying to not overfit the model. Similarly, we tried a few more combinations of models to make our ensemble. Finally after making a few more ensemble models, we got the final metrics for the new models.

Table 3 shows the results for the new models.

Model	Test Accuracy	Test Sensitivity	Test Specificity	Train Accuracy	Train Sensitivity	Train Specificity
LR + XGBoost	92.31%	66.67%	100.00%	98.70%	94.59%	100.00%
LR + XGBoost + MLP	89.74%	66.67%	96.67%	100.00%	100.00%	100.00%
XGBoost + MLP	87.17%	55.55%	96.67%	100.00%	100.00%	100.00%
LR + MLP	84.61%	44.44%	96.67%	98.70%	94.59%	100.00%
RF + GBM + SVC	87.17%	44.44%	100.00%	100.00%	100.00%	100.00%

Table 3: Metrics for combined models

7. Discussion

As mentioned earlier, our dataset was imbalanced having 76% Non-Recurring Cases. To deal with this, we can do undersampling of the majority class or use techniques like SMOTE. Both undersampling and SMOTE contribute to improved model generalization by balancing class distribution, mitigating the impact of class imbalance, and fostering more accurate and robust predictions across all classes.

It was observed that ensemble models tend to outperform individual models. This is because stacking and ensemble models combine the strengths of diverse individual models, leveraging their complementary abilities to capture different aspects of the data. This synergistic approach enhances predictive performance by mitigating the weaknesses of individual models and yielding a more robust and accurate overall prediction.

For the models that we built, the best model that we got was the ensemble of XGBoost and Logistic Regression having an accuracy of **92.31%** and a sensitivity of **66.67%**.

To further improve the model, we can try to combine information from different types of data. The original aim of our project was to combine clinical data and molecular data to get better results. Different types of data may yield different results for the same sample set. The results from the new data can be combined with the current data to get even better predictions. Given different types of data, we can build more models to further improve the results and make the model more reliable. For example, let’s say we had the genetic data available for the same set of patients, we can use that data to build another model which may have slightly different results.

Although prediction of breast cancer recurrence, is an important area for improving patient outcomes, faces several inherent limitations. Major among these is the lack of complete and diverse datasets necessary for robust sampling. The insufficient availability of data limits the potential opportunity and generalization possibilities. Considering data for various ethnicities across multiple geographical locations is a challenge that would require collaboration. Limited our research to items written in English, due to the possibility of related publications in this area of study existing in other languages, which leads to linguistic bias. There could be vital information being missed due to this restraint.

The main limitation of this idea was the sheer lack of data for the same set of patients.
Cancer tests, such as imaging tests (such as X-rays, CT scans, MRIs), physical examinations, blood tests (such as tumor markers or genetic tests), and molecular tests (such as genomic sequencing) are diverse in terms of complexity and cost more. We will never have enough data to build models using all the data as no patient would be undergoing all the tests available. Apart from that, fields like precision medicine are quite new and are not favored by everyone since it gives a lot of unnecessary results which can not be acted upon. For example, people may come to know that they are more susceptible to a disease due to a specific genetic mutation but they can not act upon it, leading to unnecessary anxiety.

Future advancements in breast cancer prognosis entail integrating diverse data modalities (molecular, clinical, imaging) for enhanced predictions. Updating models with time-dependent covariates (treatment changes, lifestyle factors) ensures dynamic, accurate prognostication. Leveraging publicly available molecular data will amplify predictive power. Innovations in machine learning (ML) techniques, like attention mechanisms, promise refined precision. Studies suggest that thermography could potentially detect changes in breast tissue heat patterns associated with cancer recurrence. Thermography is a way to examine patients using infrared rays instead of radiation and exposure to ionizing radiation is a known risk factor for developing cancer. The envisioned combination of molecular and clinical datasets promises enhanced predictive power once more publicly available molecular data becomes accessible. Pursuing these directions will enhance breast cancer prognosis models, empowering clinicians with precise tools for improved patient care and decision-making.

8. Conclusion

In conclusion, breast cancer recurrence poses a pervasive threat, underscoring the importance of leveraging Machine Learning (ML) for accurate diagnosis and prognostication. ML methodologies offer promise in identifying recurrence risks and aiding physicians in personalized patient care strategies. Future advancements hold the potential for faster, more precise models as data availability diversifies. Integrating attention mechanisms within deep learning marks a notable stride toward enhanced precision. Encouraging future avenues, researchers should explore incorporating gene sequencing data and amalgamating diverse datasets for comprehensive predictions.

These efforts could refine predictive models, advancing personalized medicine and improving outcomes for breast cancer patients. Additionally, merging computer vision which uses image segmentation to highlight important areas that can be used to identify features that clinical data cannot capture. As technology evolves and datasets expand, the convergence of innovative ML techniques with comprehensive, multi-dimensional data sources holds the key to more precise, timely, and individualized strategies in combating breast cancer recurrence, ultimately striving for improved patient care and outcomes.

References

- González-Castro L, Chávez M, Duflot P, Bleret V, Martin AG, Zobel M, Nateqi J, Lin S, Pazos-Arias JJ, Del Fiol G, López-Nores M. Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. Cancers (Basel). 2023 May 13;15(10):2741. doi: 10.3390/cancers15102741. PMID: 37345078; PMCID: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10216131/.

- Tasci, Erdal, Ying Zhuge, Harpreet Kaur, Kevin Camphausen and Andra Valentina Krauze. “Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics.” International Journal of Molecular Sciences 23 (2022): n. Pag. | https://www.semanticscholar.org/reader/992bf4c0b92ef251644ac2854dd1baacd7e42dc5

- Wolberg,William, Street,W., and Mangasarian,Olvi. (1995). Breast Cancer Wisconsin (Prognostic). UCI Machine Learning Repository | https://doi.org/10.24432/C5GK50.

- K. Chakradeo, S. Vyawahare and P. Pawar, "Breast Cancer Recurrence Prediction using Machine Learning," 2019 IEEE Conference, Allahabad, India, 2019, pp. 1-7, doi: 10.1109/CICT48419.2019.9066248 | Breast Cancer Recurrence Prediction using Machine Learning

- Susan G. Komen for the cure | Breast Cancer Foundation | https://www.komen.org/breast-cancer/treatment/recurrence/