# Predictive Modeling for Invasive Ductal Carcinoma Recurrence


## Rohan Kaushikan, Sindhu Ghanta



International Electronic Conference on Cancer 2024


## Introduction

Invasive ductal carcinoma (IDC) constitutes approximately 80% of breast cancer cases in the United States. After treatment, there is a 3-15% chance that IDC will recur.

Forecasting the recurrence of a patient's IDC as early as possible is critical for long term survival [2].

Previously, histopathological analyses were used to diagnose cancers and assess prognosis. These methods are often unable to integrate multiple data types or handle large amounts of genomic data.

## Engineering Goal and Objective

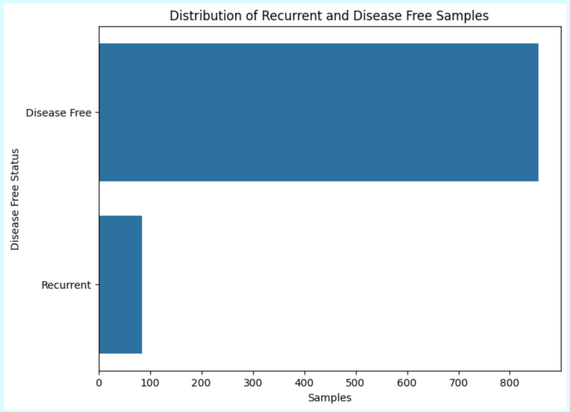
We hypothesized that mRNA expression data could be used to predict IDC recurrence at least one decade in advance.

Our engineering goal was to build and train machine learning models on mRNA expression data to predict IDC recurrence with high accuracy, precision, and recall.

## Dataset

Patient genomic and clinical samples collected in TCGA's Breast Invasive Carcinoma study were downloaded from cBioPortal [3].

The dataset contained 1084 patient samples, of which 856 samples were non-recurrent, 84 were recurrent, and 144 were not categorized. Genomic data was profiled about 11 years, on average, before IDC recurred.

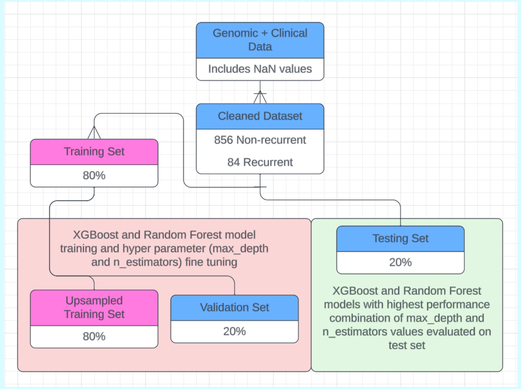



| Gene | Protein Change | Chromosome | Mutation Type | mRNA Expr. |
|---|---|---|---|---|
| CHEK2 | ***Q69**** | 22 | **Nonsense** | 60% |
| H3C2 | ***E106Q*** | 6 | **Missense** | 98% |
| TP53 | ***S183**** | 17 | **Nonsense** | 4% |
| TET2 | ***E754**** | 4 | **Nonsense** | 34% |
| PTPRD | ***L1535**** | 9 | **Nonsense** | 49% |
| NFKB2 | ***Q155**** | 10 | **Nonsense** | 42% |
| HOXC13 | ***M1?*** | 12 | **Nonstart** | 59% |
| ZMYM2 | ***E1174Q*** | 13 | **Missense** | 12% |
| MYO18A | ***I840M*** | 17 | **Missense** | 97% |

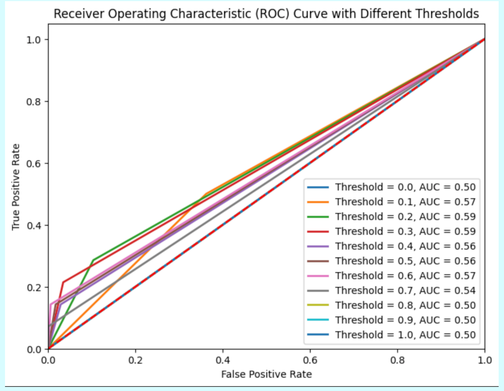
## Methodology

XGBoost and Random forest, both tree based models, are efficient on datasets with a small number of features such as the one used in this study. Additionally, both models have a Feature Importance Analysis tool that offers interpretability of the model's prediction process.

Differential mRNA expression analysis: T-tests used to calculate t-values for each gene's mRNA expression between cohorts. The 10 with lowest p-values (all <0.05) were selected as features. Rows with null values were removed. An XGBoost and Random Forest model were trained and used for prediction.




While they are both based on an ensemble of decision trees, XGBoost learns sequentially while Random Forest trains decision trees independently.

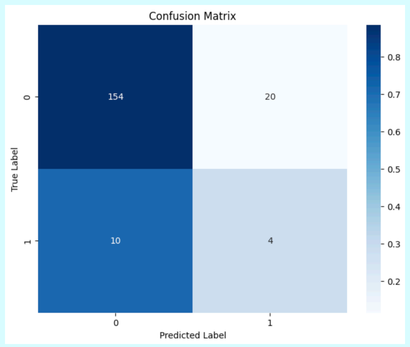


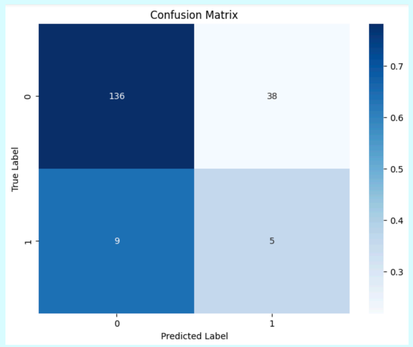

Random Forest model's performance on validation set for threshold values between 0 and 1 through ROC curves.

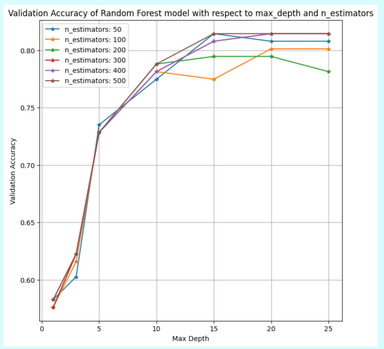
## Results

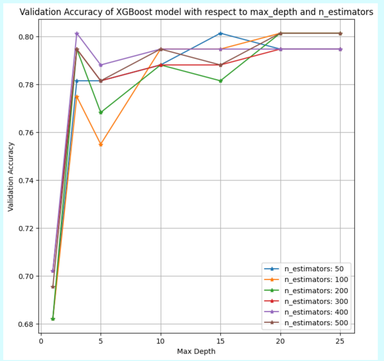
The XGBoost model performed best on the testing dataset with hyper parameters max_depth = 3 and n_estimators = 400 (accuracy = 0.84, recall = 0.29, precision = 0.89). The prediction threshold used was 0.5.

The Random Forest model performed best on the testing dataset with hyper parameters max_depth = 5 and n_estimators = 50 (accuracy = 0.75, recall = 0.36, precision = 0.78). The prediction threshold used was 0.3.




Random Forest and XGBoost performance on validation set graphed for each max_depth and n_estimators value combination. The graphs indicate that increasing the complexity of each decision tree in the random forest model results in higher accuracy. They also indicate that for our dataset, which only included 10 features, a larger number of decision trees allowed better performance of the XGBoost model.

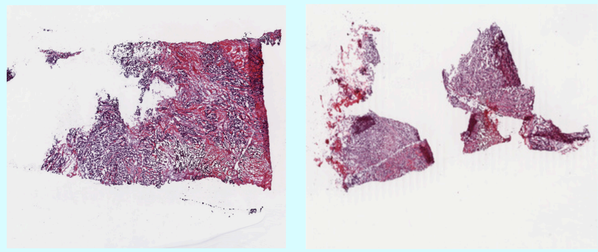
## Discussion and Conclusion

The XGBoost model with hyper parameters n_estimators = 400 and max_depth = 3 had the best performance on the testing dataset, with accuracy 0.84, recall 0.29, and precision 0.89. This indicates that a higher number of simpler decision trees allows for best model performance on our dataset. Our results are promising in that they indicate that by reducing skew in the dataset or by including more samples, model performance can be improved. While we were able to predict IDC recurrence using machine learning with high precision, the recall was much lower than desirable, especially for a medical setting. However, identifying ~29% of patient cancers as recurrent using data collected 11 years prior to recurrence while maintaining high precision speaks to the potential of our model.

Ultimately, we aim to develop a model that can consistently and accurately predict IDC recurrence early, leading to improved outcomes for IDC patients.

Prediction using our current dataset will likely yield similar results, but we believe that incorporation of clinical and imaging data will drastically improve model performance [4], eventually resulting in a desirably low false negative rate.

## Future Research



Sample disease free (left) and cancer recurrent (right) patient tissue slide images from TCGA's BRCA study

Our future work includes the incorporation of tissue slide images for IDC recurrence prediction. This would result in a multimodal approach integrating multiple data types.

After inclusion of more genes as features, we also aim to use XGBoost's and Random Forest's Feature Importance Analysis to identify genes that are closely associated with IDC recurrence.

## Selected References


1. Halima, A., et al. Intraoperative Radiation Therapy for Early-Stage Breast Cancer: Updated Outcomes from a Single-Institution Experience. *Annals of surgical oncology*, *31*(2), 931–935. https://doi.org/10.1245/s10434-023-14448-6
2. Visser, L. L., et al. (2019). Predictors of an Invasive Breast Cancer Recurrence after DCIS: A Systematic Review and Meta-analyses. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *28*(5), 835–845. https://doi.org/10.1158/1055-9965.EPI-18-0976
3. *CBioPortal for Cancer Genomics*. (n.d.-b). https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018
4. Ratan, R., et al. (2023). Factors determining the requirement of surgical intervention and prognosis in cases of traumatic bifrontal contusions: A prospective observational study. *Surgical neurology international*, *14*, 438. https://doi.org/10.25259/SNI_754_2023