

Breast Cancer risk  
prediction website

EmpowerHER

DOLLADA SRISAI

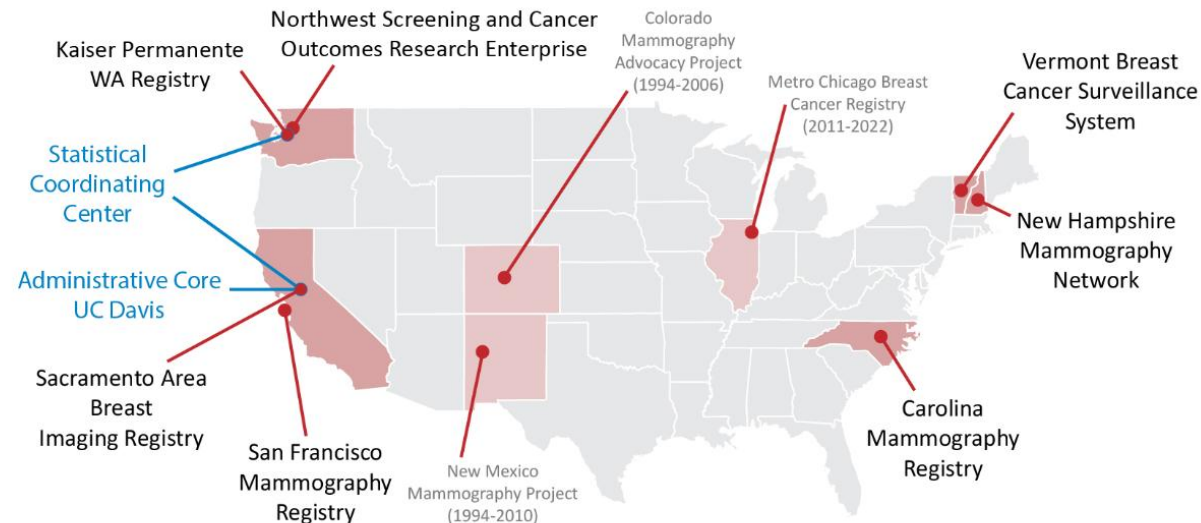
# Data source: Breast Cancer surveillance Consortium(BCSC)

Demographic, family history and clinical data was collected between 2005-2017 from Mammography registry in the USA



[HOME](#) [ABOUT THE BCSC](#) [BCSC BLOG](#) [STATISTICS](#) [DATASETS](#) [DOCUMENTATION](#) [BCSC TOOLS](#)

## RESEARCH SITES AND PRINCIPAL INVESTIGATORS



# Features

Variable Name	Description	Coding
year	Calendar year of observation	Numerical, 2005-2017
age_group_5_years	Age (years) in 5 year groups	1 = Age 18-29 2 = Age 30-34 3 = Age 35-39 4 = Age 40-44 5 = Age 45-49 6 = Age 50-54 7 = Age 55-59 8 = Age 60-64 9 = Age 65-69 10 = Age 70-74 11 = Age 75-79 12 = Age 80-84 13 = Age ≥85
race_eth	Race/ethnicity	1 = Non-Hispanic white 2 = Non-Hispanic black 3 = Asian/Pacific Islander 4 = Native American 5 = Hispanic 6 = Other/mixed 9 = Unknown
first_degree_hx	History of breast cancer in a first degree relative	0 = No 1 = Yes 9 = Unknown
age_menarche	Age (years) at menarche	0 = Age ≥14 1 = Age 12-13 2 = Age <12 9 = Unknown
age_first_birth	Age (years) at first birth	0 = Age < 20 1 = Age 20-24 2 = Age 25-29 3 = Age ≥30 4 = Nulliparous 9 = Unknown

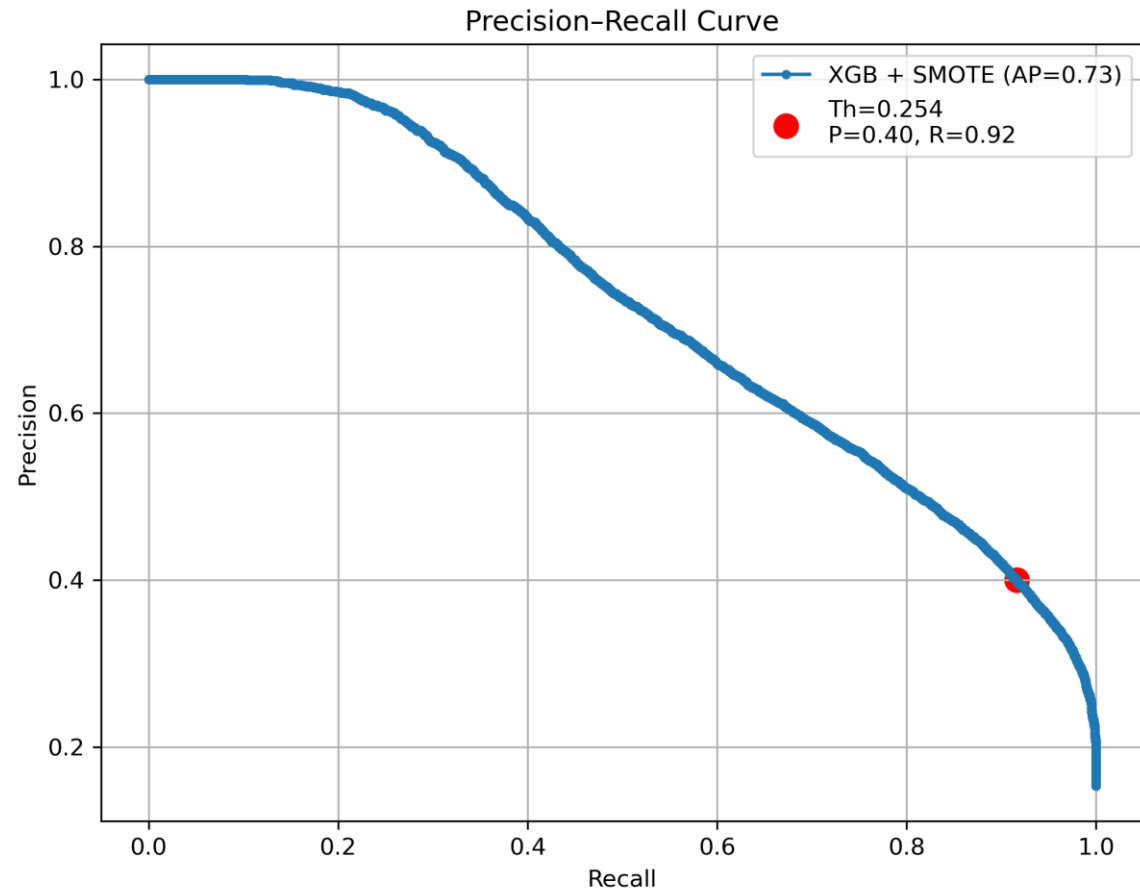
BIRADS_breast_density	BI-RADS breast density	1 = Almost entirely fat 2 = Scattered fibroglandular densities 3 = Heterogeneously dense 4 = Extremely dense 9 = Unknown or different measurement system
current_hrt	Use of hormone replacement therapy	0 = No 1 = Yes 9 = Unknown
menopaus	Menopausal status	1 = Pre- or peri-menopausal 2 = Post-menopausal 3 = Surgical menopause 9 = Unknown
bmi_group	Body mass index (kg/m <sup>2</sup> )	1 = 10-24.99 2 = 25-29.99 3 = 30-34.99 4 = 35 or more 9 = Unknown
biophx	Previous breast biopsy or aspiration	0 = No 1 = Yes 9 = Unknown
breast_cancer_history	Prior breast cancer diagnosis	0 = No 1 = Yes 9 = Unknown
count	Frequency count of this combination of covariates	Numerical

After removing rows with unknown, data from 345,064 women was used to train **XGBoost** model.

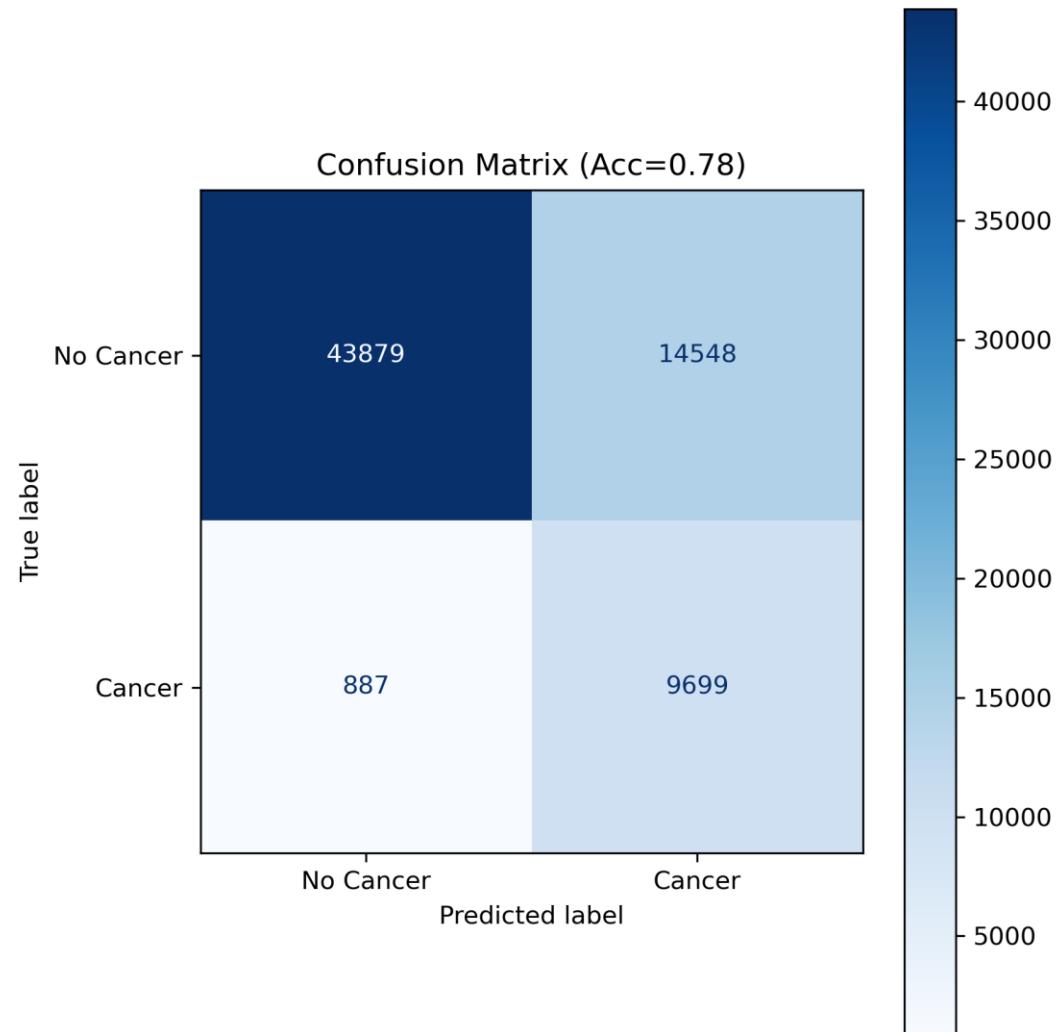
# The model is designed to minimize false negatives ~~even~~ if it means accepting more false positives.

## Key Model Performance Metrics

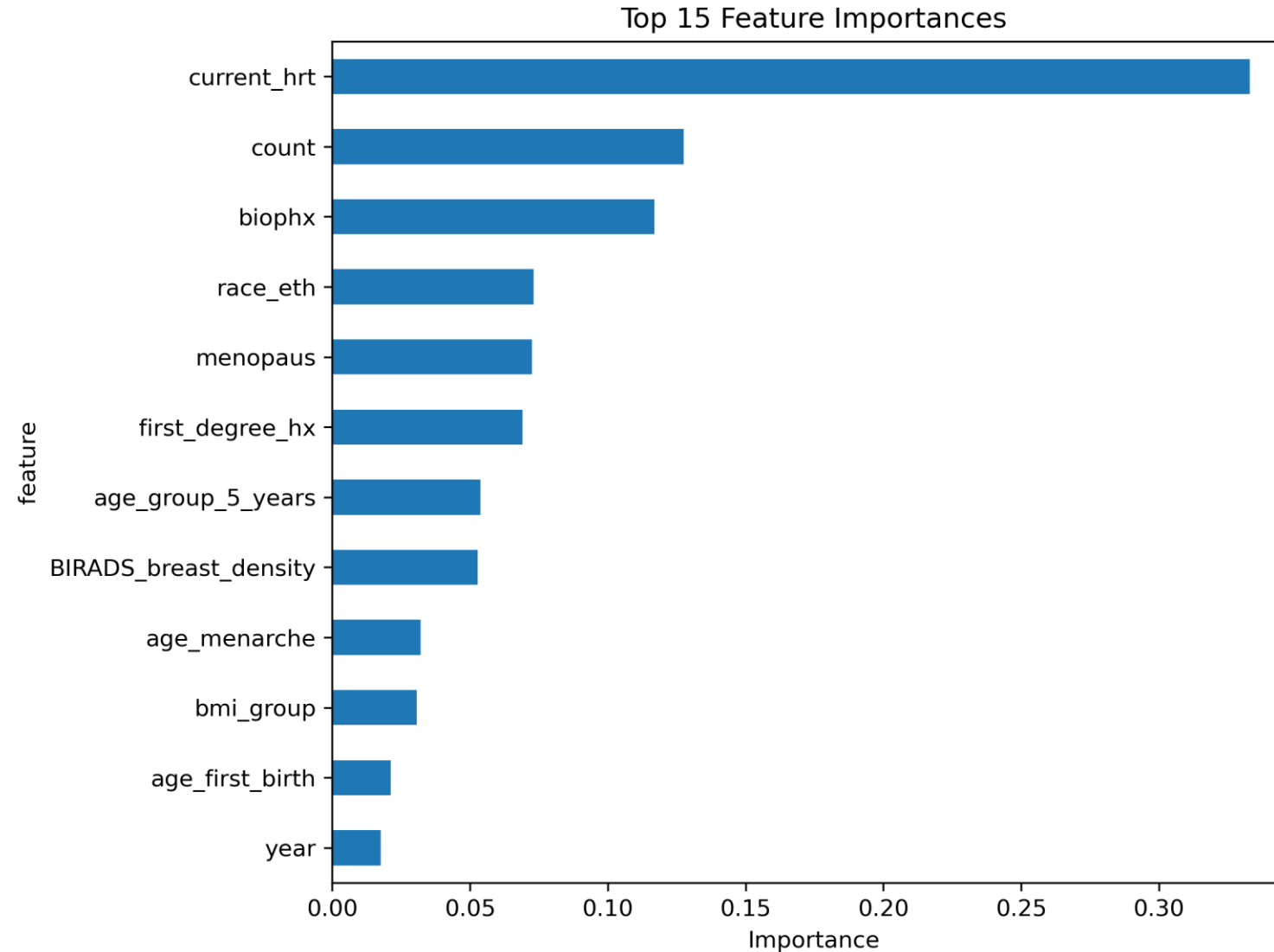
- Precision: 0.4
- Recall: 0.916 to
- Brier score: 0.1012
- MCC 0.5466
- ROC AUC: 0.92
- Accuracy: 0.776



78% of all predictions (the sum of true negatives and true positives) were correct



current hormone-replacement therapy use is by far the strongest predictor



# Core Functions of EmpowerHER

---

