PREDICTING BREAST CANCER SURVIVAL USING MACHINE LEARNING MODELS

OBJECTIVE

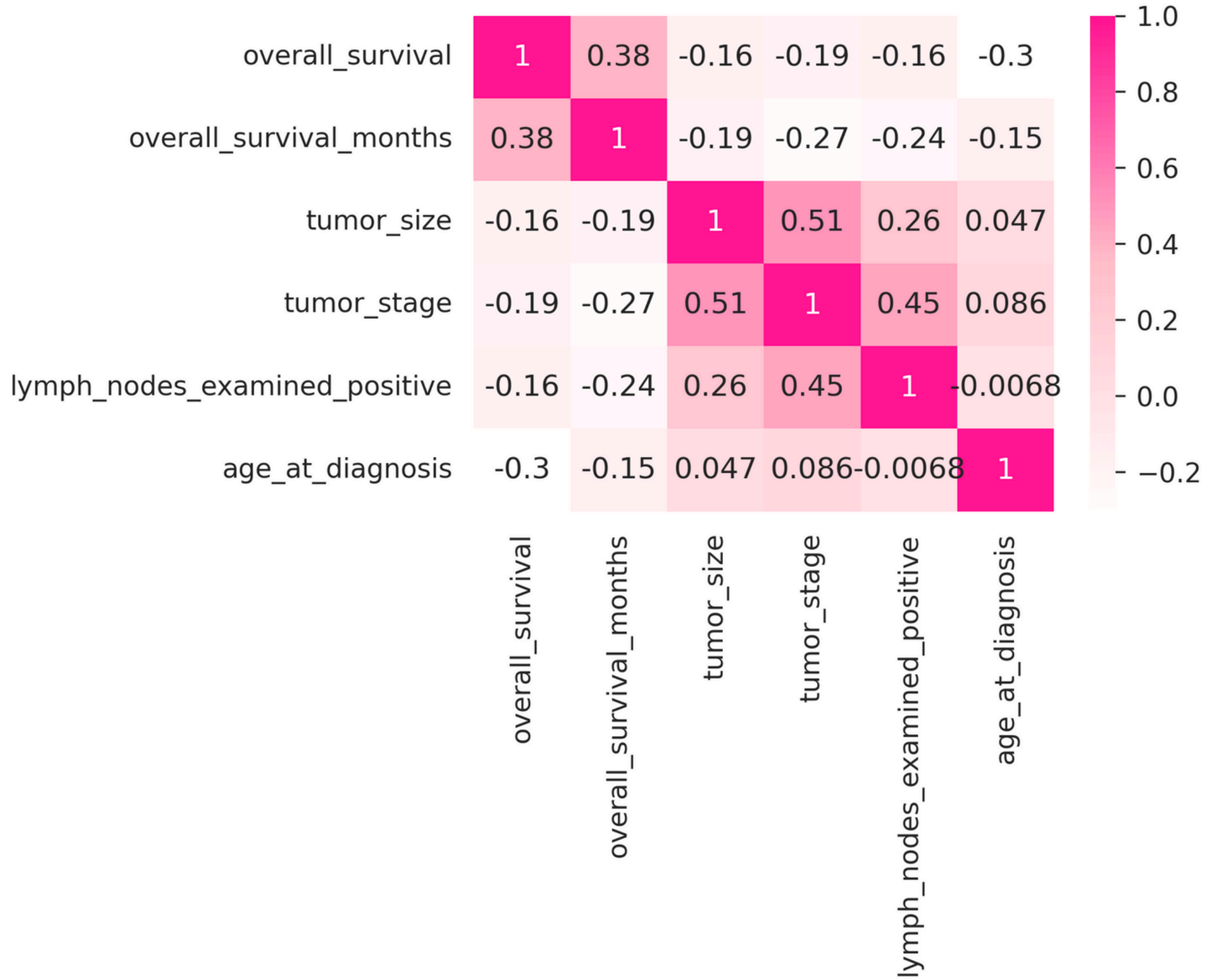
Dataset Overview:

- 31 clinical attributes
- 331 gene m-RNA level z-scores
- 175 gene mutations

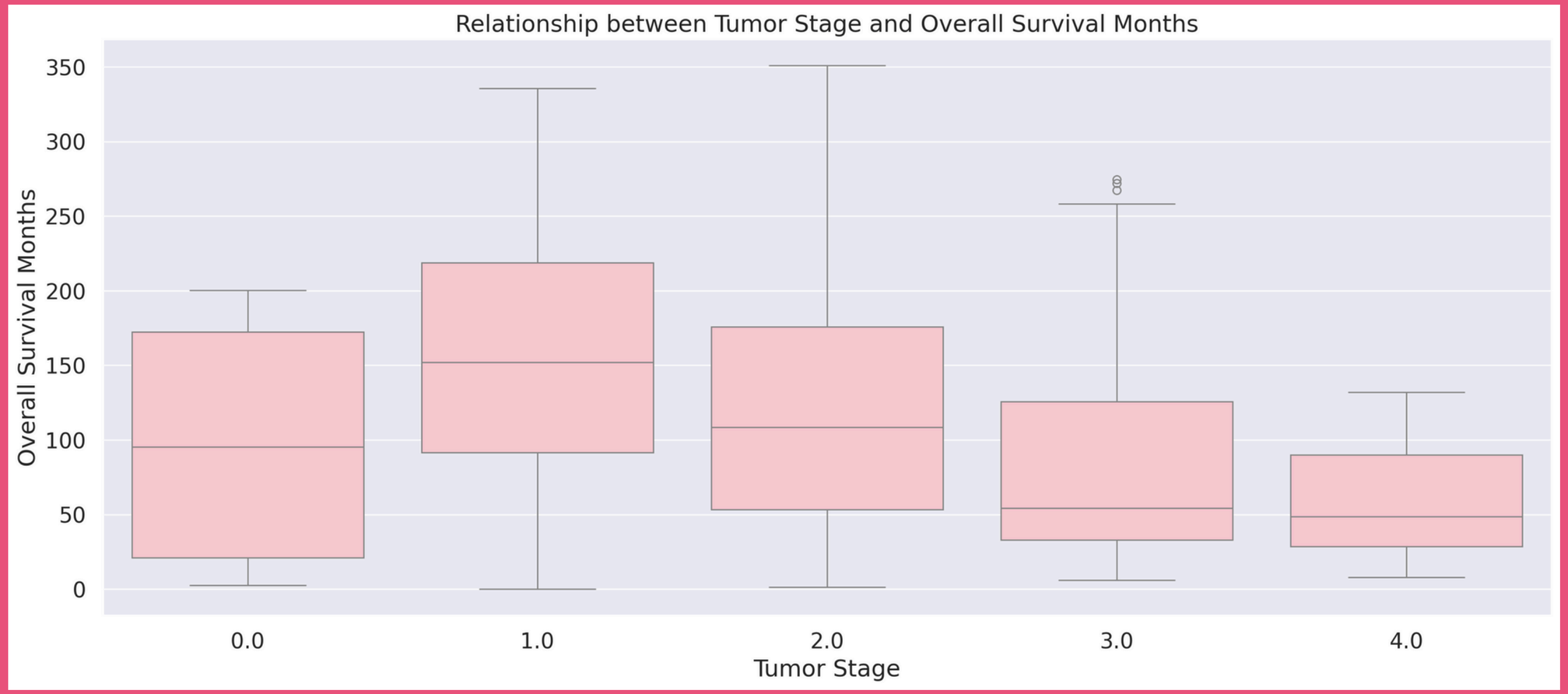
Key Clinical Features:

- Tumor size
- Tumor stage (0-4)
- Lymph nodes tested positive
- Breast cancer type
- Age

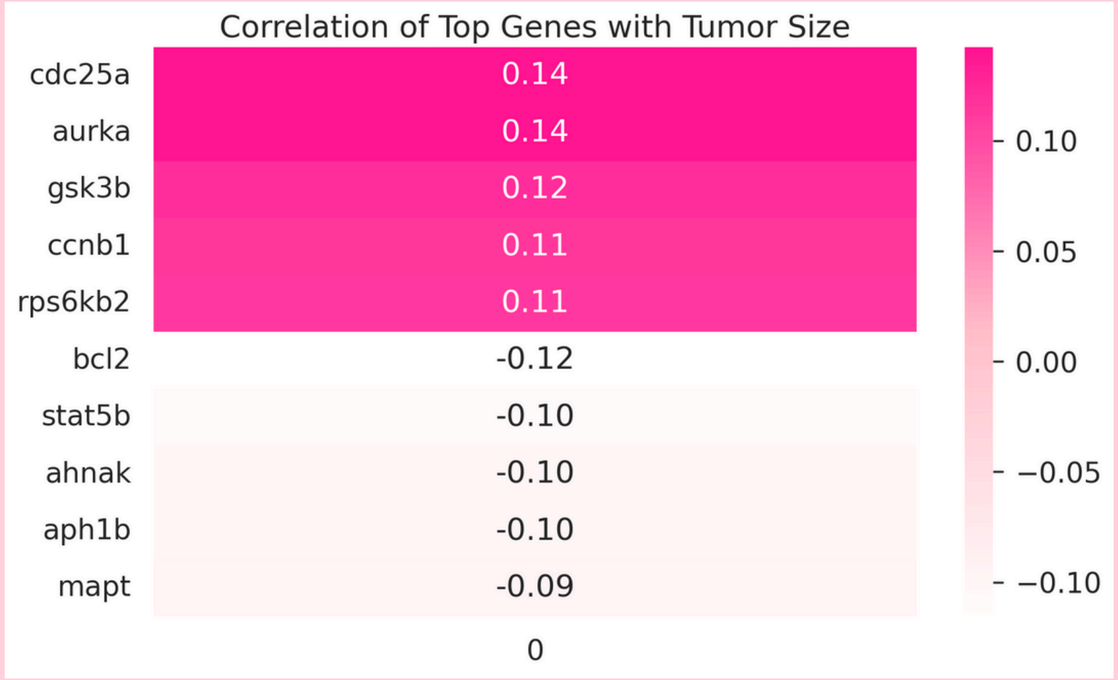
Objective: Build a machine learning model to predict overall survival (died vs. survived) using clinical and gene expression data

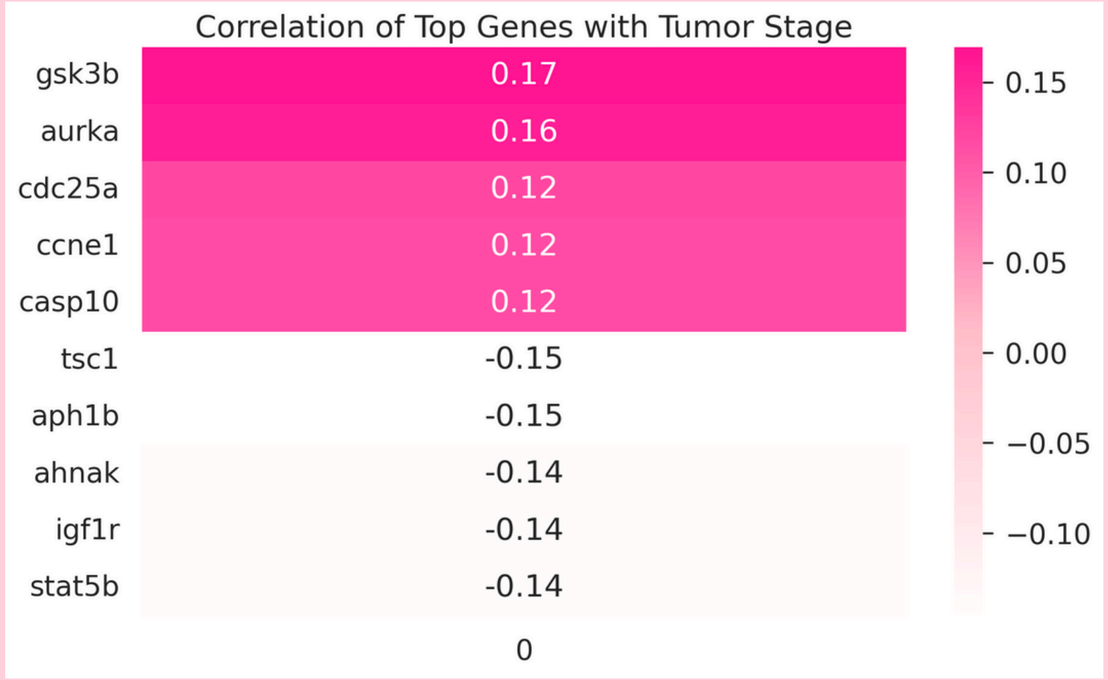


EARLY TUMOR STAGES HAVE HIGHER SURVIVAL TIME



HOW DO GENE EXPRESSION LEVELS CORRELATE WITH TUMOR AGGRESSIVENESS?





BEST PERFORMING MODELS TESTED USING XGBOOST

XGBoost Model 1: Clinical Features

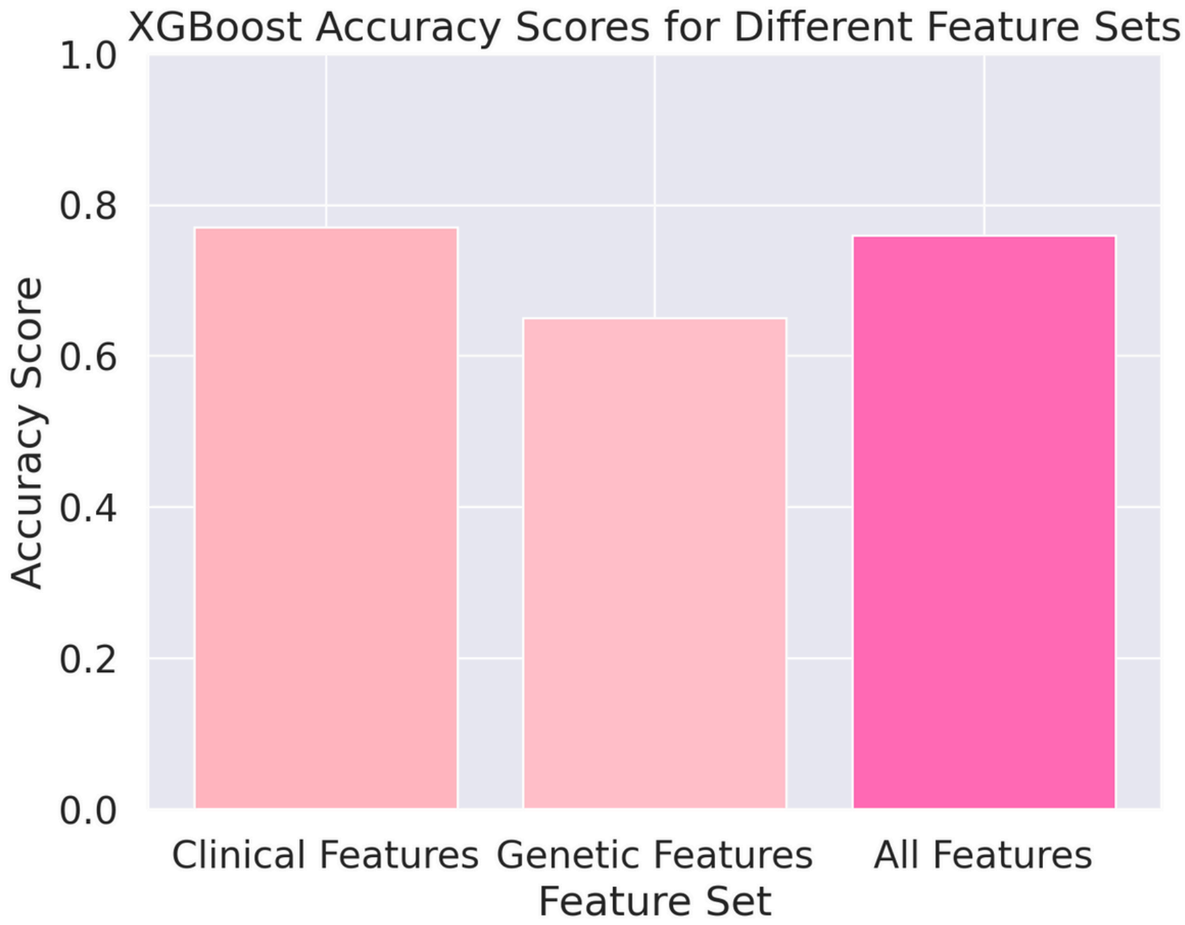
- **Best performing model**
- Accuracy score: 77%
- AUC-ROC score: 0.86 (Strong separation between surivivors & non-survivors)

XGBoost Model 2: Genetic Features

- Accuracy score: 65%
- AUC-ROC score: 0.68

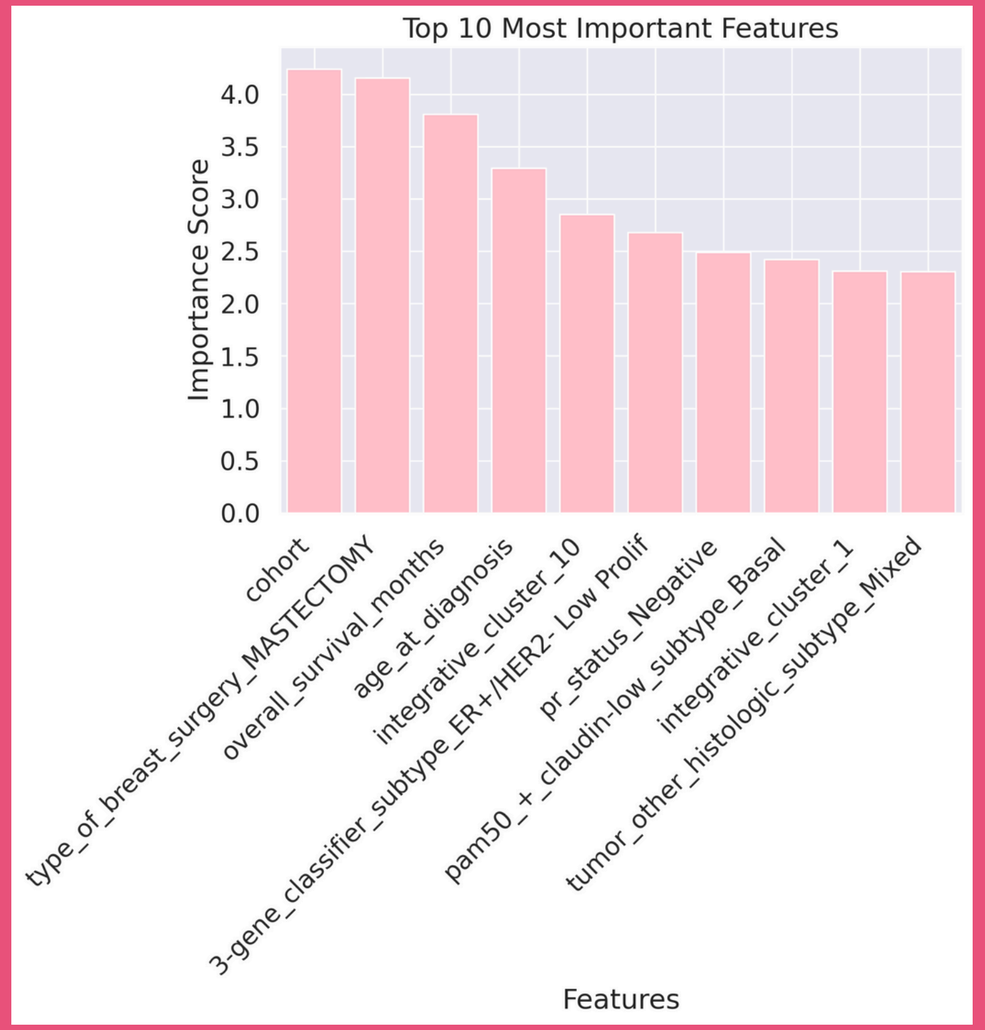
XGBoost Model 3: All Features

- Accuracy score: 76%
- AUC-ROC score: 0.85



TOP 10 MOST IMPORTANT CLINICAL FEATURES AFFECTING SURVIVAL

- Cohort
- Type of Breast Surgery
- Overall Surivival Months
- Age at Diagnosis
- Integrative Cluster
- 3-gene classifier subtype: ER+/HER2 Low Profile
- PR status Negative
- Pam 50 + Claudin low subtype basal
- Integrative cluster 1
- Tumor other histologic subtype mixed



UTILITY OF THE MODEL

Theses insights and findings can support breast cancer research and treatment:

- Helps identify high-risk patients based on gene expression and clinical data
- Can assisst healthcare providers to create personalized treatment planning
- Provides insights into biomarkers associated with survivial outcomes

REFERENCES:

- https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric
- https://pmc.ncbi.nlm.nih.gov/articles/PMC5461908/
- https://www.nature.com/articles/s41523-018-0056-8