Breast Cancer Prognosis: Leveraging Cancer Registry Data for Survival Prediction

Parry Nall, Varun Kalidindi, Marietta K Saldias, Joshua J. Levy
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

Dartmouth Health

ABSTRACT

Background

- Breast cancer is the second leading cause of cancer-related deaths among women worldwide. Current models often classify survival status but rarely predict time-to-event, limiting timely intervention opportunities.

Methods

- Using a SEER-derived public clinical registry, we trained and compared an ensemble survival learner and a gradient-boosted time-to-event model.

Results

- The gradient-boosted model achieved better model performance, particularly for short-term survival prediction. SHAP analysis identified lymph node involvement and regional disease extent as the strongest predictive features.

Conclusion

- Time-to-event machine learning models can provide more precise survival estimates, enabling earlier, personalized treatment decisions and potentially improving clinical outcomes.

INTRODUCTION

Studies show that 1 out of 8 women will get breast cancer in their lifetime, making early identification of patients at higher risk of poor outcomes crucial for improving survival rates.

Why is Machine Learning useful?

ML can analyze vast amounts of data, identify patterns, and predict outcomes with greater accuracy than traditional methods: all in all, leading to faster diagnoses, personalized treatment plans, and improved patient care.

What is Survival Prediction?

Survival Prediction, a type of machine learning, estimates not just if a patient has breast cancer, but how much life they may have (time-to-event). Unlike regular classification, it uses both the event status and the time to event, helping doctors understand a patient's prognosis more accurately. In our project, we're using clinical data and models like Survival XGBoost to predict breast cancer survival over time.

Benefits

- **Early Intervention:** Without accurate survival prediction tools, clinicians may miss early interventions for high-risk patients.
- **Supporting Psychological and Life Planning:** With accurate modeling, patients and families have time to plan and make life decisions.
- **Facilitates Personalized Medicine:** By being prepared with a timeline, clinicians can make suitable decisions for their patients.

Current Approaches

Researchers built statistical and machine learning models to predict 10-year breast cancer mortality from national health data. They achieved a high predictive accuracy from traditional regression methods. However, their train models only estimate the probability of survival within 10 years, not how long a patient might survive.

METHODS

Data Collection

- Seer-derived public clinical registry (n = 4,024), cohort used for all analyses.

Feature engineering

- Encode tnm staging, receptor status, regional node findings, tumor grade, tumor size, age, and survival months; one-hot encode categorical variables where appropriate.

Train/test split

- Random 80/20 split, stratified by event status to preserve outcome balance during training and evaluation.

Models compared

- **Ensemble survival learner (Random Survival Forest):** Ensemble of survival decision trees using bootstrapped samples and random feature selection. Captures complex non-linear effects and interactions.
- **Gradient-boosted time-to-event model (XGBoost with AFT loss):** Gradient boosting method that optimizes the accelerated failure time loss to model log survival time. Allows flexible non-linear relationships between predictors and survival outcomes.

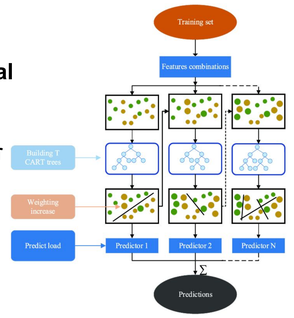

Figure 1: Diagram of XGBoost modeling

Evaluation

- Concordance index and area under the receiver operating characteristic curve (time-dependent auc) computed for short-term, medium-term, and long-term horizons to assess time-varying discrimination.

Interpretability

- Shap analysis to report global feature importance and produce patient-level explanation plots that show the direction and magnitude of predictor effects.

RESULTS

	C Index
Ensemble Learner	0.5769
Gradient Boosted	**0.7500**

Table 1: C Index Values

Model Evaluation

- **C Index:** Measures how well a model predicts the order of events, with 1 being perfect and 0.5 meaning random guessing.
- **ROC:** A curve showing the trade-off between true positive rate and false positive rate at different thresholds.
- **AUC:** The area under the ROC curve, indicating overall model accuracy; 1 is perfect, 0.5 is random.
- **SHAP Values:** Numbers that show how much each feature influences a model's prediction for an individual case.

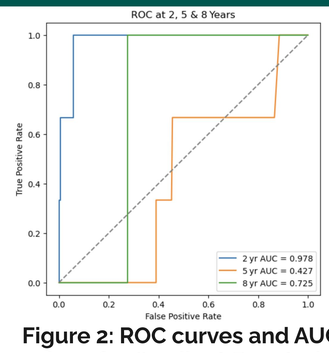

Figure 2: ROC curves and AUC values for Gradient Boosted

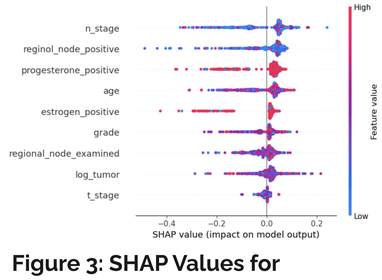

Figure 3: SHAP Values for Gradient Boosted

DISCUSSION

Model Performance and Key Predictors

- Gradient-boosted AFT model outperformed ensemble survival learner, especially for short-term survival prediction (C-index 0.7500 vs. 0.5769)
- Patient-level time-to-event predictions showed strong discrimination at 2 years (AUC 0.978), weak discrimination at 5 years (AUC 0.427), and moderate discrimination at 8 years (AUC 0.725).
- SHAP interpretability highlighted lymph node involvement and regional disease extent as the most influential features, with tumor size and local stage contributing less.

Data and Evaluation Insights

- Patient-level time-to-event predictions demonstrated strong discrimination across short-term, medium-term, and long-term horizons.
- Limitations include single-registry SEER dataset (n = 4,024), variable follow-up, and no external validation.
- Default parameter settings may have restricted maximum achievable performance.

FUTURE WORK

- Conduct external validation and recalibration on independent datasets to ensure generalizability and improve long-term prediction accuracy.
- Expand model inputs to capture more comprehensive patient information (treatment, comorbidities, sociodemographics) and integrate multimodal data, which can strengthen model robustness across time horizons.
- Optimize hyperparameters systematically to maximize performance and develop an accessible application for practical use by patients and clinicians.

CONCLUSION

Clinical Potential

- AI-based time-to-event models can deliver interpretable, individualized survival predictions to guide earlier interventions and informed treatment decisions.
- Gradient-boosted AFT framework demonstrated superior short-term prognostic accuracy, enabling more effective triage and patient prioritization.

Path to Impact

- Validation across diverse, prospective cohorts will be critical to ensure reliability and fairness in real-world settings.
- Seamless integration into clinical workflows could enhance oncology decision-making, streamline resource allocation, and improve patient outcomes.

References

edit.