Imputing Breast Cancer Stage in a Large EHR Dataset: Light Gradient Boosting Machine Algorithm and Explainable Artificial Intelligence

MSR81

Julia A O'Rourke, PhD[1,2], Jin Yu, PhD[2], Ellen Stein, MS[1], Zuzanna Drebert, PhD[1], Marley Boyd, MS[1], Mike Temple, MD[1], E. Susan Amirian, PhD[1]

[1]TriNetX, LLC, Cambridge, MA, USA. [2]Northeastern University, Artificial Intelligence Department, Boston, MA, USA.

BACKGROUND & OBJECTIVES

Electronic health records (EHR) data are often missing cancer staging. Advanced machine learning builds accurate but uninterpretable models; Explainable AI deciphers the logic behind these models.

In this study, Light Gradient Boosting Machine (LightGBM) imputed breast cancer stage at initial diagnosis, and SHAP (SHapley Additive Explanations) explored feature importance of the underlying model.

METHODS

TriNetX harmonizes de-identified patient data from 69 US healthcare organizations, largely comprised of academic medical centers (~75%). Using structured EHR data from the Dataworks USA Network, female breast cancer patients with index dates (stage/diagnosis date) between 2000 and 2024 were selected. In addition, selected patients were required to have least 10 encounters within the 9-month time window (1 month prior and 8 months after index).

The LightGBM model was trained with >400 demographics, diagnoses, procedures, medication, and sentence embedding features (extracted from the code-based notes).

RESULTS

De-identified data from 460,616 women with breast cancer were utilized (30,473 with stage). Age and race distributions of patients with known stage differed from those with undocumented stage. Patients with unknown stage were older (mean 61.5 vs. 58.8 years), had a lower proportion of Black women (9.8% vs. 19.2%), and a greater proportion of White women (72.1% vs. 67.1%).

After optimization, the predictive model reached 88% accuracy (the base model had 67% accuracy).

SHAP analysis revealed that the LightGBM assigned patients to higher stages based on the presence of increasingly complex diagnostic and treatment codes:

- in situ carcinoma diagnosis and the absence of complex interventions (stage 0);
- partial mastectomies, sentinel node biopsies, and receptor status testing (stage 1);
- systemic therapies, continuing nodal testing, and more diverse imaging (stage 2);
- secondary lymph node involvement, additional imaging complexity, and more systemic treatments (stage 3);
- extensive imaging, testing and greater procedure complexity (stage 4).

CONCLUSION

SHAP analysis revealed the expected diagnostic and treatment patterns. Advanced ML algorithms can impute missing stage at diagnosis with acceptable accuracy.

Figure 1. Modeling Approach

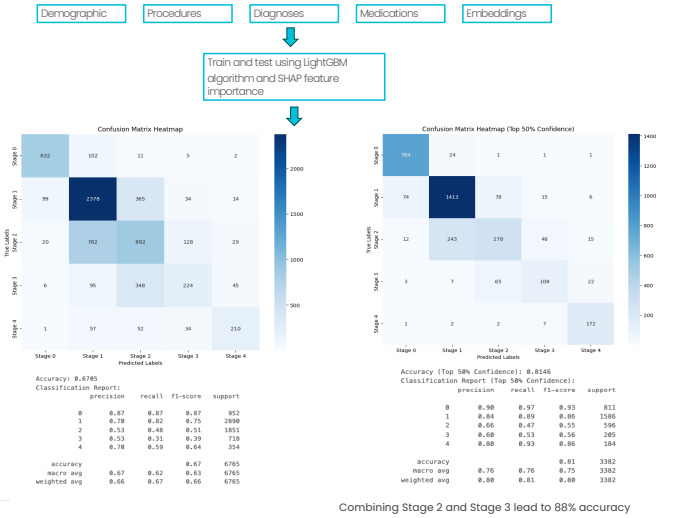
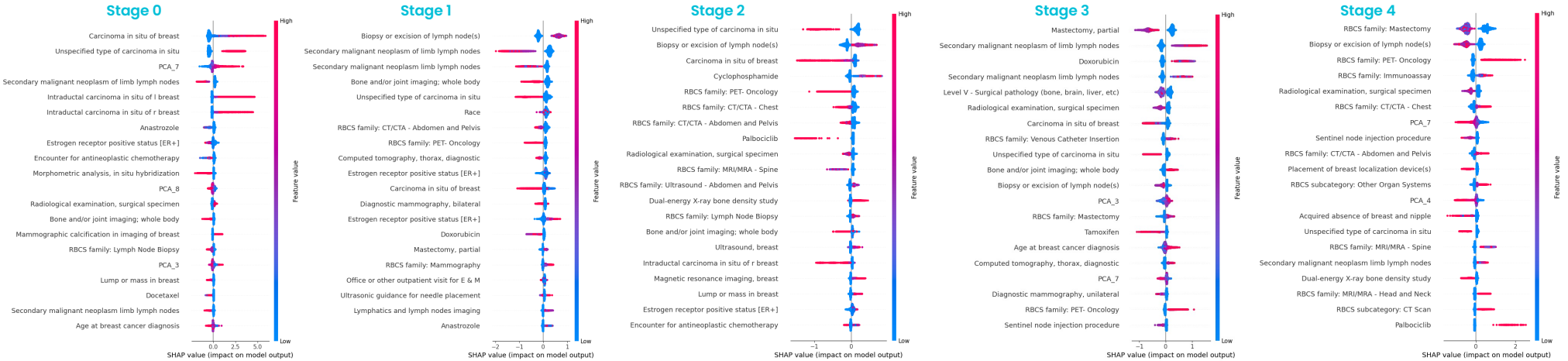


Figure 2. Beeswarm plot with SHAP values for the top 20 features for each stage based on SHAP mean feature importance.



PCA – principal components based on code descriptions transformed to text embeddings

ISPOR 2025 | Montréal, QC, Canada

Design services were provided by Kelly Marsocci and were funded by TriNetX, LLC.

Download Poster PDF