ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

Subtype-Dependent Performance of Cox and Machine Learning Survival Models for Recurrence Prediction in Breast Cancer: Development and External Validation Using Public Clinical Data

Francis Mawutor Amuyao[1]

[1]Department of Statistics, Florida State University
117 N. Woodward Avenue, Tallahassee, FL 32306-4330, USA

Publication Date: 2026/07/13

Abstract: Breast cancer recurrence risk prediction informs adjuvant treatment decisions and follow-up planning. Molecular subtypes capture biologically distinct risk profiles, yet whether machine learning (ML) survival methods offer consistent advantages over Cox proportional hazards (PH) modelling across subtypes remains unclear. We analysed 1,964 breast cancer patients across five molecular subtypes. Penalised Cox PH, Random Survival Forest (RSF), and Gradient Boosting Survival (GBS) models were developed for recurrence-free survival (RFS) prediction using discrimination, calibration, decision curve analysis, and subtype-stratified SHAP explainability. External validation was performed in the independent METABRIC cohort (n=2,388). Overall C-indices were comparable: Cox 0.676, RSF 0.672, GBS 0.664. Subtype-stratified analysis revealed heterogeneity: RSF outperformed Cox in Triple-negative (0.646 vs 0.578) and Luminal B HER2 (0.656 vs 0.530) disease, while Cox was superior in HER2-positive disease (0.671 vs 0.595). SHAP analysis identified biologically coherent subtype-specific risk drivers. In METABRIC validation, RSF achieved C-index 0.743 versus Cox 0.658. Machine learning offers selective, subtype-dependent advantages over Cox modelling rather than uniform superiority.

Keywords: *Breast Cancer; Recurrence Prediction; Survival Analysis; Molecular Subtype; Random Survival Forest; SHAP; METABRIC; Cox Proportional Hazards; External Validation.*

How to Cite: Francis Mawutor Amuyao (2026) Subtype-Dependent Performance of Cox and Machine Learning Survival Models for Recurrence Prediction in Breast Cancer: Development and External Validation Using Public Clinical Data.
International Journal of Innovative Science and Research Technology, 11(6), 3298-3304.
https://doi.org/10.38124/ijisrt/26jun1500

I. INTRODUCTION

Breast cancer recurrence represents a primary driver of disease-related morbidity, mortality, and treatment burden. Accurate prediction of recurrence-free survival (RFS) informs decisions regarding adjuvant therapy, follow-up intensity, and patient counselling. Despite the widespread adoption of molecular subtype classification, survival prediction studies overwhelmingly report aggregate performance metrics without subtype stratification, obscuring clinically meaningful differences.

Cox proportional hazards (PH) regression remains dominant in oncology survival modelling due to its interpretability. However, machine learning survival methods such as Random Survival Forest (RSF) and Gradient Boosting Survival (GBS) can accommodate nonlinear effects and high-order feature interactions. Published evidence on their advantage over Cox models is inconsistent [4,5], and most studies omit calibration, clinical utility assessment, and external validation.

A further gap is the absence of interpretability analysis stratified by molecular subtype. SHAP (SHapley Additive exPlanations) values provide a principled framework for quantifying feature contributions within each subtype independently [6]. In this study, we develop and externally validate Cox PH, RSF, and GBS models following TRIPOD+AI reporting standards [1], with subtype-stratified SHAP analysis as the primary novel contribution.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

II. METHODS

➢ *Study Cohort and Outcome*

RFS data were available for 1,964 breast cancer patients across five molecular subtypes: Luminal A-like (n=972, 49.5%), Luminal B-like (n=425, 21.6%), Triple-negative (TNBC; n=320, 16.3%), HER2-positive (n=139, 7.1%), and Luminal B HER2 (n=108, 5.5%). The subtype distribution is shown in Fig. 1. External validation used the METABRIC cohort (n=2,388) obtained from cBioPortal. Missing values were imputed using column medians for continuous variables and column mode for categorical variables. Predictors included age at diagnosis, tumour size, tumour stage, histological grade, positive lymph node count, ER status, PR status, HER2 status, chemotherapy, hormone therapy, radiotherapy, treatment intensity, a composite clinical risk score, and an age-by-tumour-size interaction term.

➢ *Model Development*

The dataset was split into training (80%, n=1,571) and test (20%, n=393) sets using stratified sampling by event status. Three survival models were developed: a penalised Cox PH model (lifelines, L2 penaliser=0.1) [8], a Random Survival Forest (RSF; 300 trees, max depth 10) [4,5], and a Gradient Boosting Survival model (GBS; 200 estimators, learning rate 0.05) [5]. The proportional hazards assumption was assessed using the Grambsch-Therneau test [9]; violations were observed for ER status (p=0.002), PR status (p=0.015), and hormone treatment (p=0.020) but were deemed of limited practical consequence given sample size.

➢ *Performance Evaluation*

Discrimination was assessed using the concordance index (C-index) [10] with 1,000-resample bootstrap 95% confidence intervals. Subtype-specific C-indices were calculated with 500 bootstrap resamples per subtype. Calibration was assessed at 3year and 5-year RFS. Clinical utility was evaluated by decision curve analysis (DCA) [3] across threshold probabilities from 0% to 60%.

➢ *SHAP Analysis*

Subtype-stratified SHAP values were computed using the SHAP Permutation Explainer applied to RSF models trained independently within each molecular subtype [6]. Mean absolute SHAP values per feature per subtype were aggregated to characterise differential risk driver profiles. All analyses used Python 3.11 with pandas, numpy, lifelines, scikit-survival, and shap.

III. RESULTS

➢ *Cohort Characteristics*

Subtype-specific event rates and survival characteristics are summarised in Table 1. Luminal B HER2 had the highest RFS event rate (54.6%) and Luminal A-like the lowest (37.4%). Median RFS ranged from 45.3 months (HER2-positive) to 111.1 months (Luminal A-like). Kaplan-Meier RFS curves by subtype are shown in Fig. 2.

Table 1 Cohort Characteristics by Molecular Subtype

Subtype	n	RFS Event %	Median Age (yr)	Median RFS (mo)
Luminal A-like	972	37.4	63.5	111.1
Luminal B-like	425	40.5	65.0	100.8
Triple-negative	320	40.9	55.6	87.2
HER2-positive	139	50.4	53.2	45.3
Luminal B HER2	108	54.6	62.5	78.6

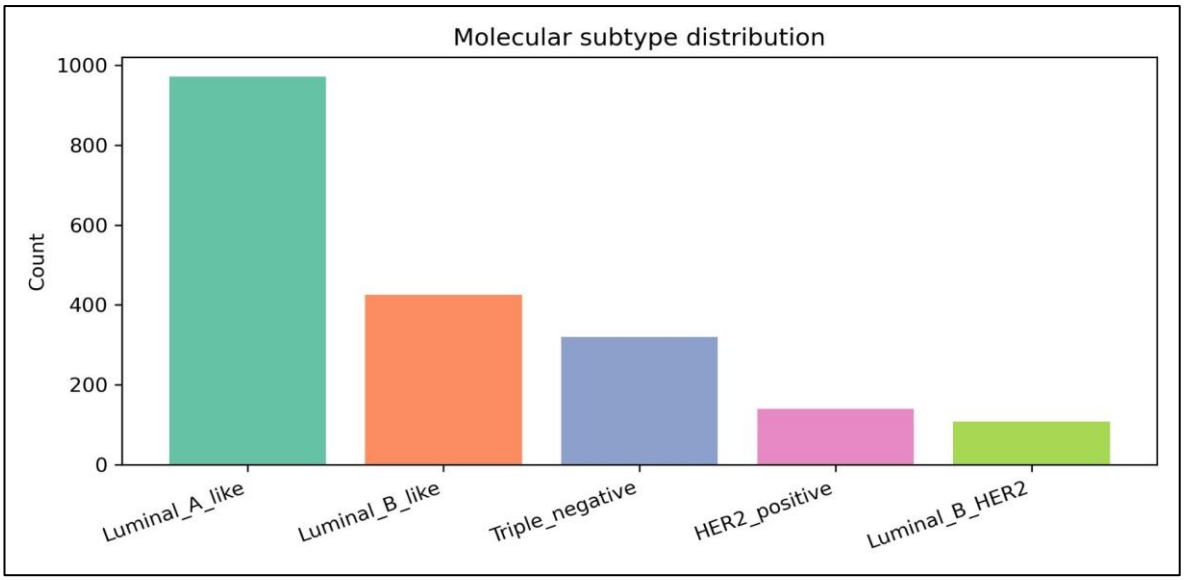

Fig 1 Molecular Subtype Distribution (n=1,964). Luminal A-like was the Most Prevalent Subtype.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

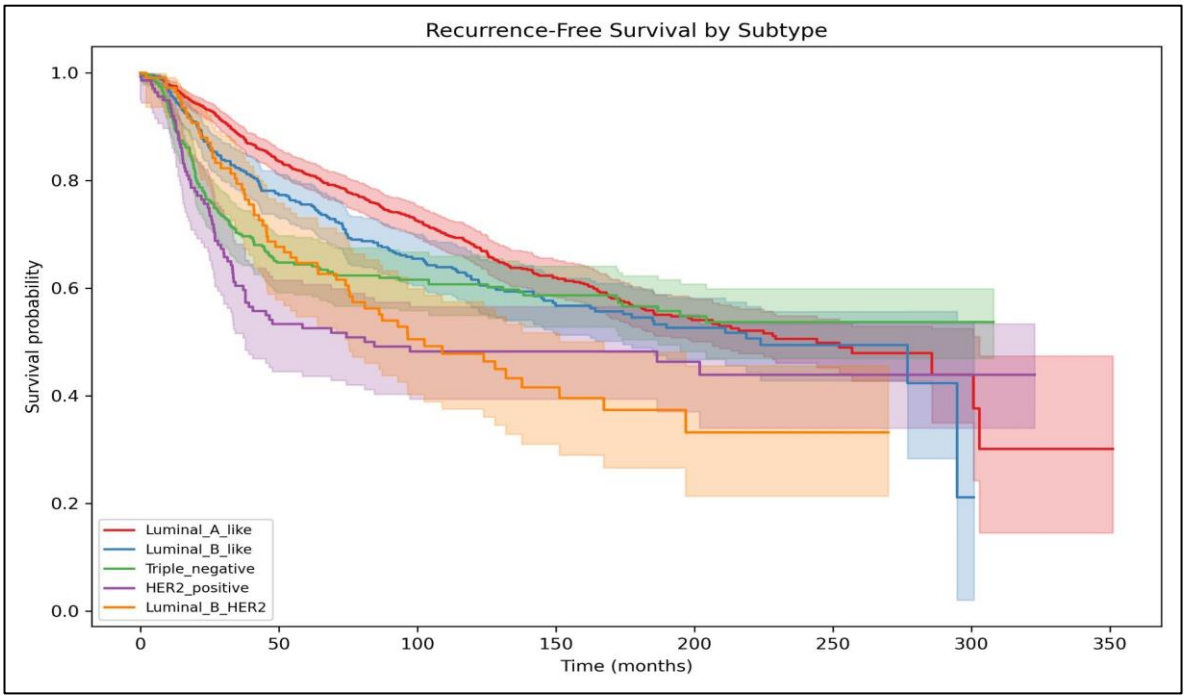

Fig 2 Kaplan-Meier RFS Curves by Molecular Subtype with 95% Pointwise CIs. Luminal A-Like Showed the Most Favourable Prognosis (Median RFS 111.1 Months); HER2 Positive and Luminal B HER2 Showed the Earliest Divergence.

➢ *Model Discrimination*

Overall discrimination was comparable across models (Table 2), with overlapping bootstrap confidence intervals indicating no statistically meaningful aggregate difference. The clinical risk score alone achieved a C-index of 0.665, nearequivalent to the full ML models.

Table 2 Overall Model Discrimination with 95% Bootstrap CIs

Model	C-index	95% CI
Cox PH	0.676	0.629–0.725
Random Survival Forest	0.672	0.626–0.721
Gradient Boosting Survival	0.664	0.615–0.712
Clinical Risk Score	0.665	0.617–0.714

Subtype-stratified analysis revealed heterogeneity masked at the aggregate level (Table 3, Fig. 3). RSF outperformed Cox by 0.068 C-index units in Triple-negative disease and by 0.126 in Luminal B HER2 disease. Cox was superior in HER2-positive disease (0.671 vs RSF 0.595), where smaller sample size likely limits tree-based model stability.

Table 3: Subtype-Specific C-Index with 95% Bootstrap CIs

Subtype	Cox (95% CI)		RSF (95% CI)	
Luminal A-like	0.635 0.702)	(0.560–	0.625 0.691)	(0.549–
Luminal B-like	0.603 0.720)	(0.481–	0.610 0.714)	(0.506–
Triple-negative	0.578 0.680)	(0.480–	0.646 0.735)	(0.552–
HER2-positive	0.671 0.825)	(0.505–	0.595 0.762)	(0.398–
Luminal B HER2 0.530 0.709)		(0.347–	0.656 0.815)	(0.477–

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

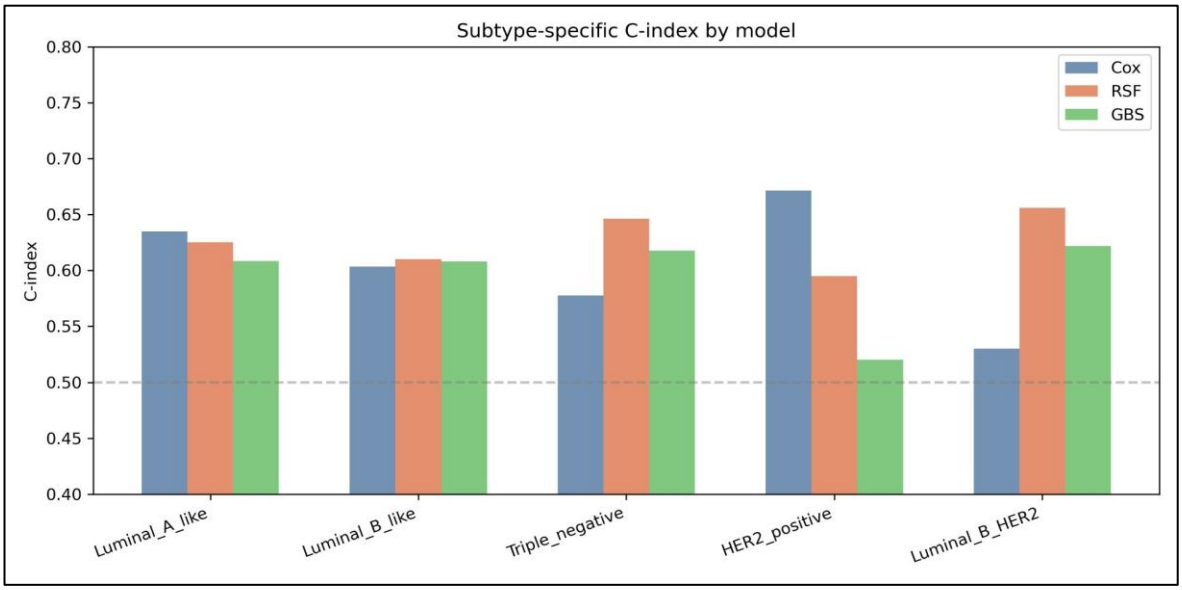

Fig 3 Subtype-Specific C-Index by Model. RSF Outperformed Cox in Triple-Negative and Luminal B HER2 Disease. Cox was Superior in HER2-Positive Disease.

➢ *Calibration and Decision Curve Analysis*

At 3-year RFS, both models showed adequate calibration in the low-to-mid predicted risk range. At 5-year RFS, RSF demonstrated more consistent calibration across the full risk range, with Cox underestimating observed event probability in the highest risk decile. DCA confirmed positive net benefit for both RSF and Cox over treat-all and treat-none strategies across threshold probabilities of approximately 15–55% at 3 years and 20–55% at 5 years (Fig. 4), with RSF showing marginally higher net benefit throughout.

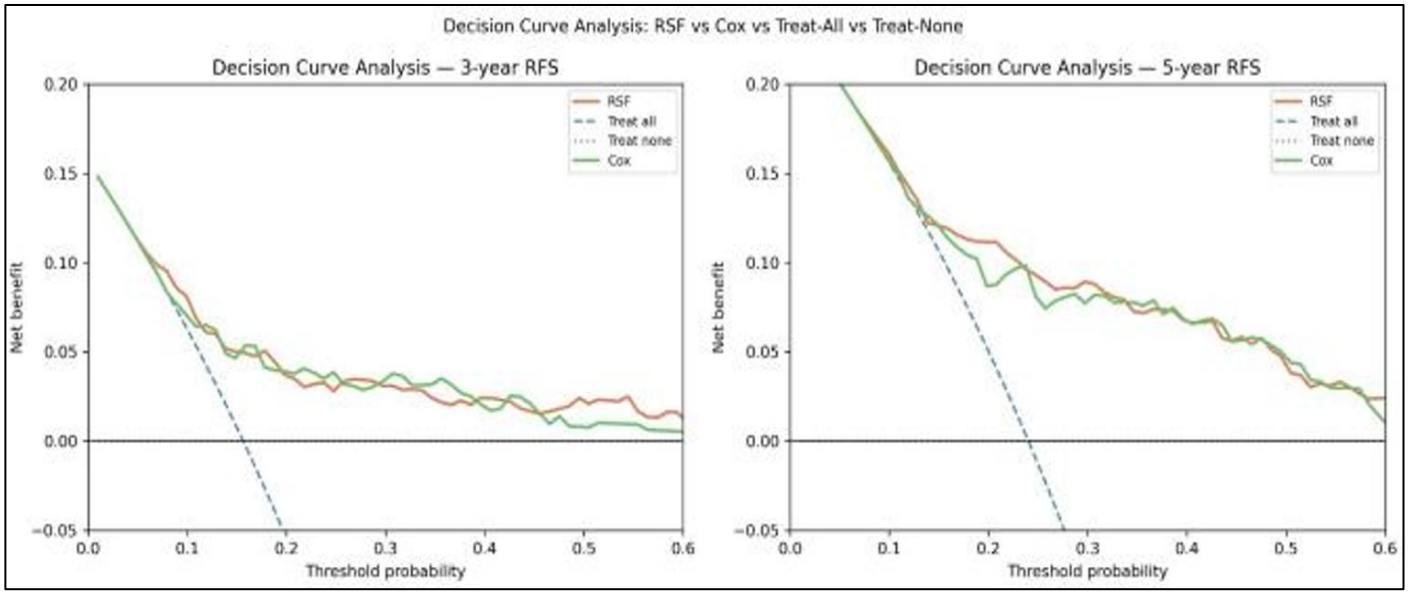

Fig 4 Decision Curve Analysis at 3-Year (Left) and 5-Year (Right) RFS. Both RSF and Cox Demonstrated Positive Net Benefit Above Treat-All and Treat-None Strategies.

➢ *Subtype-Stratified SHAP Analysis*

Subtype-stratified SHAP analysis constitutes the primary novel contribution of this study. The SHAP heatmap (Fig. 5) reveals biologically coherent and distinctly different risk driver profiles across subtypes.

In Luminal A-like disease, clinical risk score (mean |SHAP| 10.32) and tumour size (10.05) were co-dominant, reflecting the hormone-driven, stage-dependent recurrence biology of this subtype. In Triple-negative disease, chemotherapy emerged as a distinct contributor (1.47) absent in luminal subtypes, consistent with cytotoxic treatment response being a principal prognostic axis when hormone receptor targets are absent. In Luminal B-like disease, lymph node count (5.52) and tumour size (5.34) were closely matched as the leading individual predictors. In Luminal B HER2 disease, PR status (0.88) emerged as a distinctive contributor absent or negligible in all other subtypes, reflecting the mixed hormonal and HER2amplified biology of this subgroup. ER status and HER2 status showed zero SHAP

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

contribution across all subtypes, consistent with their role in subtype definition rather than within-subtype risk stratification.

➢ *External Validation in METABRIC*

In the independent METABRIC cohort (n=2,388; RFS event rate 39.9%), RSF demonstrated strong generalisation with C-index 0.743 (95% CI 0.729–0.757), substantially outperforming Cox (0.658, 95% CI 0.640–0.676; Fig. 6, Table 4). Non-overlapping confidence intervals confirm the robustness of this difference. The RSF advantage was consistent across all available subtypes and exceeded the internal advantage by approximately 0.08 C-index units, suggesting RSF captures more population-conserved patterns of recurrence risk.

Table 4 External Validation in METABRIC (n=2,388)

Model	C-index	95% CI
Cox PH	0.658	0.640–0.676
RSF	0.743	0.729–0.757
RSF subtype C-indices:		
Luminal A-like	0.758	0.731–0.786
Luminal B-like	0.706	0.674–0.740
Triple-negative	0.776	0.744–0.807
HER2-positive	0.743	0.708–0.787

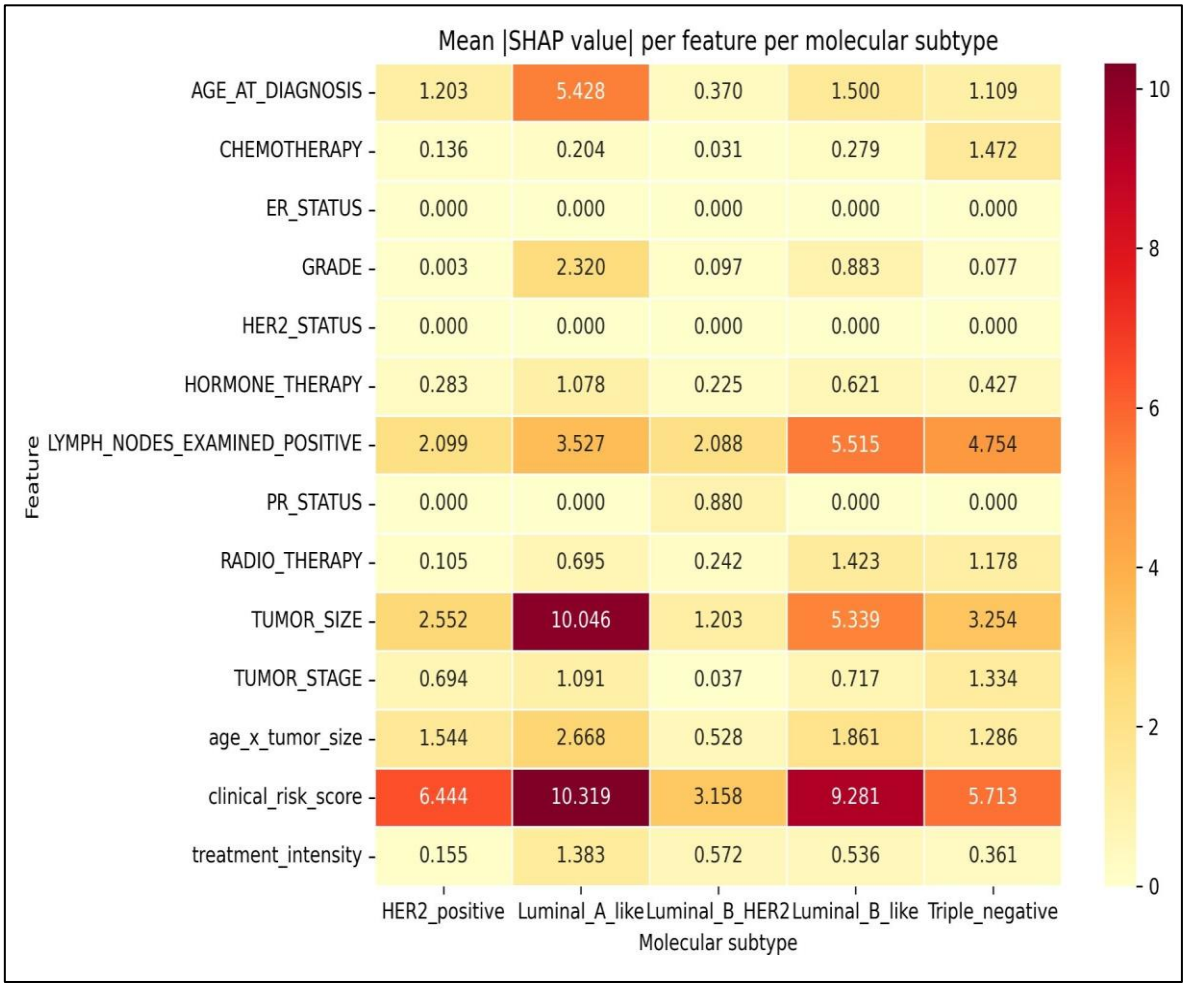

Fig 5 Subtype-Stratified SHAP Feature Importance Heatmap. Mean Absolute SHAP Values Per Feature Per Molecular Subtype Reveal Distinct Risk Driver Patterns: Tumour Burden Dominates in Luminal A; Chemotherapy is Distinctively Elevated in Triple-Negative; PR Status is Uniquely Prominent in Luminal B HER2.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

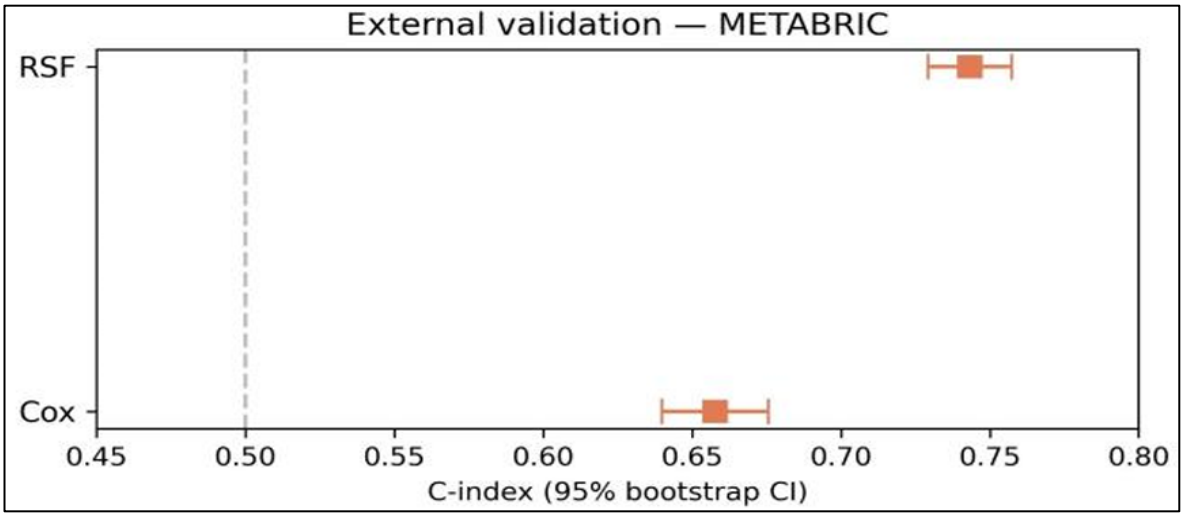

Fig 6 External Validation in METABRIC: C-Index with 95% Bootstrap CIs. RSF (0.743) Substantially Outperformed Cox (0.658) with Non-Overlapping CIs.

IV. DISCUSSION

This study presents a TRIPOD+AI-compliant comparison of Cox PH, RSF, and GBS survival models for breast cancer RFS prediction, distinguishing itself through subtype-stratified evaluation, SHAP-based explainability, and external validation in METABRIC.

At the aggregate level, all three models performed comparably, consistent with published comparative studies. The nearequivalence of the four-variable clinical risk score with full ML models is practically significant: more complex models provide no appreciable discriminatory benefit over transparent composites built from routinely collected staging variables.

The subtype-stratified analysis reveals what aggregate comparisons obscure. RSF gains of 0.068 and 0.126 C-index units over Cox in Triple-negative and Luminal B HER2 disease are practically meaningful. These biologically aggressive subtypes exhibit complex, potentially nonlinear interactions between treatment response and anatomical burden that Cox's linearity and proportional hazards assumptions may inadequately represent. The Cox advantage in HER2-positive disease, where sample size is smallest, further supports subtype-aware model selection rather than a one-model-fits-all approach.

The SHAP analysis is the primary novel contribution. Subtype-specific patterns are biologically coherent: chemotherapy's distinctive prominence in TNBC directly reflects the absence of hormone receptor targets; PR status emerging uniquely in Luminal B HER2 disease reflects its mixed hormonal and HER2-amplified biology. These findings provide mechanistic insight not available from aggregate feature importance.

The enlarged RSF advantage in external METABRIC validation (0.085 C-index externally vs 0.004 internally) suggests RSF captures population-conserved recurrence patterns beyond what Cox's linear predictor generalises. Limitations include median imputation for 25.9% tumour stage missingness, proportional hazards violations for three predictors, singlecohort external validation, and the absence of genomic predictors.

V. CONCLUSIONS

Machine learning survival methods provide selective, subtype-dependent advantages over Cox proportional hazards modelling. RSF shows meaningful gains in Triple-negative and Luminal B HER2 disease while Cox is competitive in HER2-positive disease. Subtype-stratified SHAP analysis reveals biologically coherent differential risk drivers entirely obscured in aggregate comparisons. Strong external validation in METABRIC supports generalisation of these findings. Future work should incorporate genomic signatures and pursue validation across geographically diverse cohorts.

ACKNOWLEDGMENT

The author thanks Isaac Tosin Adisa, Sherif Mohammed, and Francis Okyere for constructive discussions. Development data are from the publicly available SEER database. METABRIC data are available via cBioPortal (https:// www.cbioportal.org).

REFERENCES

[1]. G.S. Collins, P. Dhiman, C.L.A. Navarro, et al., "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI)," *BMJ*, vol. 375, p. n2334, 2021.
[2]. B. Van Calster, D.J. McLernon, M. van Smeden, et al., "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, no. 1, p. 230, 2019.
[3]. A.J. Vickers, E.B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1500

[4]. H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.
[5]. S. Polsterl, "scikit-survival: A library for time-to-event analysis built on top of scikit-learn," *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020.
[6]. S.M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[7]. C. Curtis, S.P. Shah, S.-F. Chin, et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346– 352, 2012.
[8]. C. Davidson-Pilon, "lifelines: survival analysis in Python," *Journal of Open Source Software*, vol. 4, no. 40, p. 1317, 2019.
[9]. P.M. Grambsch, T.M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994.
[10]. F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, R.A. Rosati, "Evaluating the yield of medical tests," *JAMA*, vol. 247, no. 18, pp. 2543–2546, 1982.