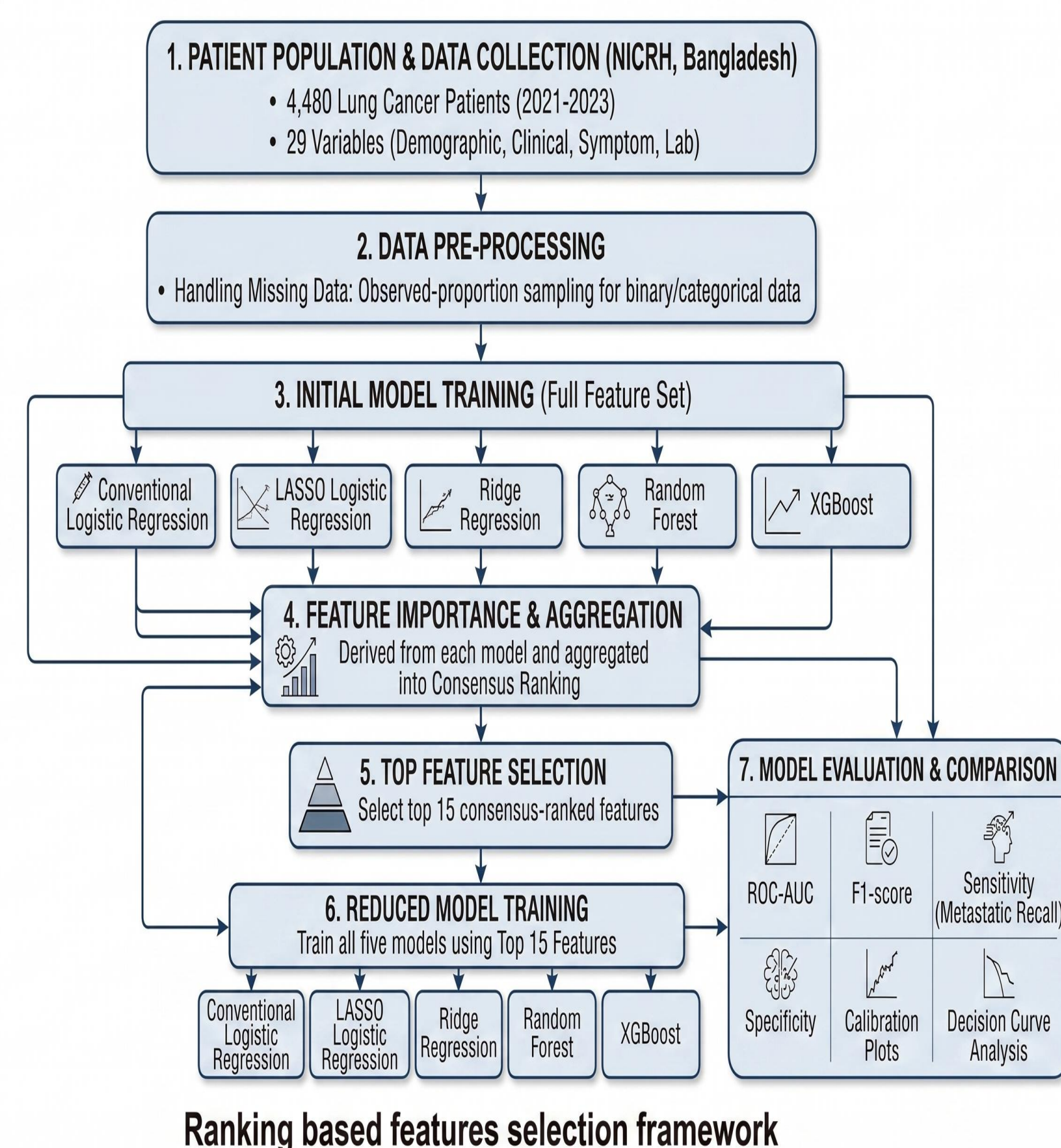


Presenting Author: M.R. Islam¹, S.M. Siddiqua², S. Sayin³, S.B. Al Ayub¹, H. Ahsan⁴¹Directorate General of Health Services, Dhaka/BD, ²Unity through population, Dhaka/BD, ³Sahlgrenska University Hospital, Gothenburg/SE, ⁴University of Chicago, Chicago/IL/USA**Correspondence:** muhammadrafiqulislam80@gmail.com

INTRODUCTION

In low- and middle-income countries (LMICs), lung cancer is frequently diagnosed at advanced or metastatic stages due to limited early detection infrastructure. Accurate risk stratification using real-world clinical and laboratory data is essential but challenging. Conventional logistic regression provides interpretability but limited predictive performance in high-dimensional settings, whereas penalized regression and machine-learning models can capture complex, non-linear relationships. Comparative evaluation of these approaches in LMIC metastatic cohorts remains scarce. We assessed regression-based and machine-learning models to predict metastatic lung cancer risk and identify key determinants, emphasizing the trade-off between predictive accuracy and clinical interpretability.

METHODS



DISCUSSION AND CONCLUSION

Consensus-ranked feature selection successfully streamlined clinical variables while preserving vital non-linear interactions in tree-based architectures. In this real-world LMIC cohort, Random Forest optimized for true-positive detection (Sensitivity = 0.592, F1-score = 0.560), effectively minimizing missed metastatic cases in resource-constrained settings. In contrast, penalized regression models (LASSO Specificity = 0.853) prioritized rule-out capability and interpretability, though at the expense of sensitivity. This dichotomy highlights that non-linear ensemble models are best suited for frontline case capture, whereas linear models offer transparent rule-out utility.

Key messages

Model Superiority: Tree-based models outperform linear regressions for metastatic case detection, with 15-feature Random Forest delivering the best balanced recall and full XGBoost achieving peak discrimination (ROC-AUC = 0.663).

•Dimensionality Reduction: Reducing 29 variables to top 15 consensus features maintains diagnostic power in ensemble architectures while simplifying data collection.

•Clinical Utility: The framework provides a practical, low-burden risk-stratification tool tailored to the diagnostic challenges faced by resource-limited oncology centers.

RESULTS

Across the evaluated configurations, the highest overall discrimination was achieved by the full-feature XGBoost model (ROC-AUC = 0.663), closely followed by the 15-feature Random Forest (ROC-AUC = 0.656). Following consensus-based feature selection, the reduced 15-feature Random Forest model emerged as the most clinically robust performer for metastatic detection. It achieved the highest overall harmonic balance (F1-score = 0.560) and the top true-positive capture rate (Sensitivity = 0.592, identifying 1,089 metastatic cases), reflecting a 1.5% relative gain in sensitivity over its full-feature version. While the reduced 15-feature XGBoost demonstrated slight performance attenuation (ROC-AUC –0.9%, F1-score –3.2%, sensitivity –4.0% to 0.407), penalized regression methods suffered much steeper trade-offs in clinical recall. Both LASSO and Ridge saw marked declines in sensitivity (LASSO: –10.9% to 0.280; Ridge: –9.7% to 0.307) and F1-scores, missing substantial proportions of positive cases. Conversely, feature parsimony significantly favored rule-out capability in linear models: the top 15 LASSO model yielded the highest true-negative rate across all paradigms (Specificity = 0.853, +2.5% increase), tightly followed by Ridge (Specificity = 0.839).

