Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241083

Survival Stacking Ensemble Model for Lung Cancer Risk Prediction

Eduardo ALONSO[a,b,1], Xabier CALLE[a], Ibai GURRUTXAGA[b], Andoni BERISTAIN[a]
[a] *Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia - San Sebastián, Spain*
[b] *Department of Computer Architecture and Technology, University of the Basque Country (UPV/EHU), Donostia - San Sebastián, Spain*
ORCiD ID: Eduardo Alonso https://orcid.org/0009-0003-3984-549X, Xabier Calle https://orcid.org/0000-0001-5689-7433, Andoni Beristain https://orcid.org/0000-0002-5452-2141, Ibai Gurrutxaga https://orcid.org/0000-0003-1830-1058

Abstract. The most well-established risk factor for lung cancer (LC) is smoking, responsible for approximately 85% of cases. The Lung Cancer Risk Assessment Tool (LCRAT) is a key advancement in this field, which predicts individual risk based on factors like smoking habits, demographic details, personal and family medical history, and environmental exposures. This paper proposes a model with fewer features that improves state of the art performance, using a simplified stacking ensemble, making it more accessible and easier to implement in routine healthcare practice. The data used in this work were derived from two cohorts in the United States: The National Lung Screening Trial (NLST) and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Both our model and LCRAT achieve an AUC of 0.799 and 0.782 on test respectively. In terms of percentage of positives, in the 50% of the population, both detect 0.766 and 0.754 of the cases. The ensemble of different survival models enhances robustness by mitigating the weakness of individual models and directly impacts the efficiency of the model, increasing the efficiency and generalizability.

Keywords. Cancer, risk factors, machine learning, ensemble models

1. Introduction

Lung cancer (LC), a predominant cause of cancer-related mortality globally, presents a significant public health challenge due to its high incidence and poor prognosis. The most well-established risk factor for LC is tobacco smoking, responsible for approximately 85% of cases [1].

In recent decades, extensive research efforts have been dedicated to combating LC. Among the various strategies developed, screening programs have emerged as a crucial tool in reducing LC mortality by enabling early detection of the disease. These programs are designed to identify individuals at high risk based on associated risk factors (family history, smoking, age…), allowing for timely intervention and treatment. A notable example is the Lung Cancer Risk Assessment Tool (LCRAT) [2], which employs a Cox model to provide individual-level risk assessments on smokers over 50 years or patients

[1] Corresponding Author: Eduardo Alonso; E-mail: ealonso@vicomtech.org.

with previous respiratory conditions, called high risk population. This approach has significantly enhanced the early detection of LC cases, thereby improving patient outcomes and highlighting the importance of targeted screening in the fight against LC.

While the LCRAT has significantly advanced the early detection of LC, it is not without its limitations. One major drawback is its limited predictive accuracy, as it may not accurately identify many high-risk individuals, leading to false positives or false negatives. Additionally, LCRAT's effectiveness is hindered by its reliance on input data quality; inaccurate or incomplete data can compromise its risk assessments. To address these limitations, recent advancements in survival analysis and machine learning [5] offer an opportunity to leverage more sophisticated and modern survival models to select the most important known risk factors. This work aims to develop a stacking survival ensemble approach for high risk population enhancing predictive accuracy, better handle diverse and intricate risk factors, and improve the overall quality and reliability of LC risk assessments using a reduced number of predictive features.

2. Methods

2.1. Data Sources

In the conducted experiment, data was utilized from two large-scale cohorts in the United States: The National Lung Screening Trial (NLST) [3] and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial [4]. Participants were recruited from multiple centers and data were collected through structured questionnaires and follow-ups, following the appropriate clinical trial protocols validated by the corresponding ethics committee. The NLST was a randomized trial involving over 53,000 smokers aged 55 to 74 years with at least 30 pack-years smoked from 2002-2004. It aimed to assess if low-dose computed tomography could reduce LC mortality compared to standard chest X-rays. From 1993-2001 PLCO trial was another randomized study with about 155,000 smoker participants aged 55 to 74 years. It evaluated the impact of specific cancer screening tests on cancer-related mortality. Additional insights into LC screening efficacy and patient demographics were provided by the PLCO data. Table 1 reports the statistical characteristics of the cohorts.

Table 1. Arm cohorts statistical characteristics. 'cig_years' represents the years smoking, 'cigpd_f' the cigarettes per day, 'cig_stop' the years since stopped smoking, 'lung_fh_cnt' the number of first degree familiars with history of LC.

Feature/Arm	**NLST CT**	**NLST X-Ray**	**PLCO Control**	**PLCO Radio**
N	26627	26621	40064	40590
age	61.42 ± 5.02	61.41 ± 5.01	62.45 ± 5.31	62.38 ± 5.28
cig_years	39.83 ± 7.34	39.86 ± 7.33	27.76 ± 13.81	27.59 ± 13.85
cigpd_f	28.47 ± 11.44	28.42 ± 11.51	19.5 ± 13.69	19.26 ± 13.52
cig_stop	3.75 ± 5.0	3.74 ± 5.0	16.1 ± 13.47	16.2 ± 13.46
lung_fh_cnt	0.24 ± 0.52	0.24 ± 0.51	0.12 ± 0.37	0.13 ± 0.37
bmi	27.89 ± 5.03	27.9 ± 5.07	27.35 ± 4.83	27.39 ± 4.88
sex (male)	15725 (59.1%)	15698 (58.9%)	23210 (57.9%)	23701 (58.4%)
Positives	1079 (4.0%)	964 (3.6%)	1604 (4.0%)	1705 (4.2%)

2.2. Model Architecture

To predict time-dependent LC risk, we designed a stacked ensemble model tailored to include LC-related risk factors as input data. The ensemble employs a dual-phase

strategy. In the first phase, multiple individual survival analysis models, including Cox Proportional Hazards (CoxPH), CoxNet, Extra Survival Trees, Gradient Boosting Survival Analysis, and Survival Support Vector Machine (SVM), independently produce predictions based on the input data. In the second phase, these individual predictions are fed as input variables to a final CoxPH model to predict the time-dependent risk of developing LC. In this stacked ensemble approach, each base model first predicts the risk for each sample independently. These individual risk predictions are then used as input features for training the meta-model. Specifically, the meta-model learns to combine these risk predictions to produce a final, refined risk prediction. This method allows the meta-model to leverage the strengths and unique insights of each base model.

2.3. Training Workflow

The training workflow for our models involves a detailed and systematic process to ensure robust and reliable performance. The models are trained using the PLCO dataset, while validation is performed with the NLST dataset. This approach leverages the strengths of both datasets and ensures that our models generalize well across different cohorts. Within the PLCO dataset, individuals from the control arm are used for training the models, while individuals who underwent radiographic screening are used for testing.

During the training workflow, a preprocessing pipeline with several data transformation steps and a final estimator is built. This pipeline is used during the training and validation steps, ensuring that every transformation is consistently applied to both the validation and prediction data. All transformations have been carried out following the same procedures as LCRAT in order to be compatible with them. Numerical data imputation is performed using the mean value of the training set, while for categorical data, the most frequent value is used. The next step is the standardization of numerical features and the categorical encoding, and finally the AI model that is going to be used.

The training process was executed through a 5-fold cross-validation (CV) strategy. Throughout each iteration of CV, rigorous hyperparameter optimization was performed using evolutionary algorithms. These algorithms explore the hyperparameter space based on heuristic scores, seeking optimal configurations that minimize bias and variance in the model. This optimization process helps for fine-tuning the models to the dataset's characteristics, enhancing their adaptability and performance across different subsets.

2.4. Feature selection

Feature selection (FS) is a crucial technique for improving model performance, reducing overfitting, and enhancing interpretability. In this work we have executed various feature selection techniques, including Boruta, Lasso, and XGB. The main idea is to remove those features that do not contribute much information to the models, thereby achieving higher performance as a result and making the models easier to implement in clinical practice.

3. Results

Our objective on this work focuses on achieving a similar performance, or even better, than the LCRAT reference model by reducing the number of variables required for prediction. The models trained focus on high-risk populations and can estimate risk at different time states. We present the results obtained for the 3-year forecast, helping to

prioritize those patients most likely to develop lung cancer in the near future, ensuring timely detection and potential early treatment. The FS strategy described in Section 4 of the methodology indicated the removal of the variables race, education, number of packs smoked per year, and history of emphysema.

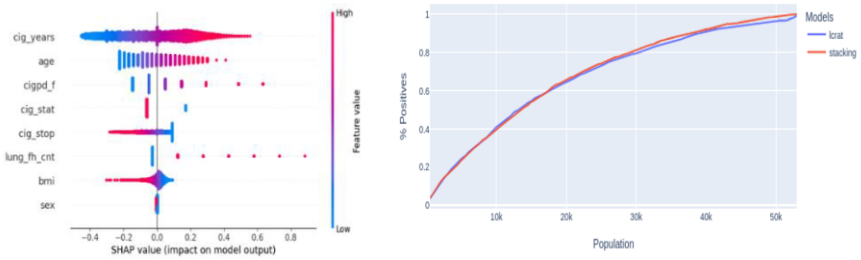

Figure 1. (a) Percentage of positive cases detected among population and **(b)** SVM shap values

To develop our LC screening model, we trained a stacking ensemble algorithm using prioritized risk factors identified through the aforementioned feature selection strategy. We compared its performance against LCRAT for 3-year predictions using the same datasets (train: PLCO control, test: PLCO radiography) and validation: NLST). Our model consistently outperformed LCRAT in ROC-AUC scores: 0.789 vs. 0.781 (Train), 0.799 vs. 0.782 (Test), and 0.698 vs. 0.697 (Validation). Additionally, we evaluated both models by analyzing the percentage of positive cases within the at-risk population in the NLST validation cohort (Figure 1a). Initially similar, our model showed slightly higher detection rates across different risk strata: 0.339 vs. 0.338 (first 15%), 0.766 vs. 0.754 (50%), and 0.92 vs. 0.90 (75%). Overall, these results underscore our model's superior predictive performance in various evaluation metrics.

Finally, we applied a feature importance algorithm to one of the base models to evaluate the impact of input variables on predictions. Our analysis revealed that variables such as the number of years smoking, age, number of cigarettes smoked, and number of first-degree relatives with cancer increase the risk, whereas years of smoking cessation decrease it. Sex showed minimal impact, with values centered around zero indicating low influence (Figure 1b).

4. Discussion

The burden of LC on healthcare systems and individuals is undeniable. Early screening protocols, including contributions from models like the LCRAT [2], have eased this burden to some extent. Our work represents a step forward in optimizing such models. We propose that by reducing the number of features and integrating modern ensemble stacking techniques, we can enhance their performance, extend their applicability and utility in clinical settings.

The rationale for excluding certain risk factors, indicated by the FS technique, may be explained by confounding or spurious correlations. In the case of education, there is a recognized correlation between lower education levels and higher LC incidence. However, the true causal factor is likely lower socioeconomic status, which is often associated with higher pollution exposure and poorer diets, both known risk factors for LC [6]. Including education level in the risk model might not effectively capture these

underlying environmental factors and could confound the results. Therefore, excluding education may improve the model's accuracy by avoiding misleading correlations.

Our proposed stacking ensemble architecture demonstrates efficacy comparable to or surpassing that of the LCRAT, underscoring the capability of ensemble approaches to effectively understand relationships between variables and optimize predictive performance. While the increase in predicting positive cases may appear marginal, even a modest improvement, such as 1%, can yield significant clinical benefits and economic savings by reducing the need for additional costly tests.

5. Conclusions

We have developed a stacked survival ensemble LC screening model that improves upon the widely used LCRAT model in two key aspects. Firstly, our model enhances the detection of positive cases, leading to earlier identification and enabling prompt intervention and treatment. Early detection is crucial because lung cancer has a better prognosis when caught early, allowing for prompt treatment interventions. These treatments are more effective in the early stages, which can significantly reduce mortality rates. Secondly, the model streamlines patient data collection by minimizing required variables, addressing potential uncertainties in patient reporting.

Acknowledgements

This work has been founded by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement no 101096473. The authors also acknowledge the National Cancer Institute for granting access to the data from the National Lung Screening Trial (NLST) and the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO).

References

[1] Chang JT et al. Cigarette smoking reduction and health risks: A systematic review and meta-analysis. Nicotine Tob Res. 2021;23(4):635–42. doi: 10.1093/ntr/ntaa156

[2] Katki HA et al. Development and validation of risk models to select ever-smokers for CT lung cancer screening. JAMA. 2016;315(21):2300. doi: 10.1001/jama.2016.6255.

[3] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395–409. doi: 10.1056/NEJMoa1102873

[4] Zhu CS et al. The prostate, lung, colorectal, and ovarian cancer screening trial and its associated research resource. J Natl Cancer Inst. 2013;105(22):1684–93. doi: 10.1093/jnci/djt281

[5] Stepanek L et al. A machine-learning approach to survival time-event predicting: Initial analyses using stomach cancer data.2020 EHB. IEEE; 2020. p. 1–4.

[6] Pampel FC, Krueger PM, Denney JT. Socioeconomic disparities in health behaviors. Annu Rev Sociol. 2010;36:349–70. doi: 10.1146/annurev.soc.012809.102529