A hybrid-computer vision model to predict lung cancer in diverse patient populations

Abdul J. Zakkar[1], Nazia Perwaiz[1], Weiheng Zhong[2], Alex Krule[1], Mason Burrage-Burton[1], Daniel Kim[1], Mehak Miglani[1], Vijeth Narra[1], Farah Yousef[1], V.K. Gadi[1], Mark C. Korpics[3], Sage J. Kim[4], Aly A. Khan[5], Yamilé Molina[4], Yang Dai[6], G. Elisabeta Marai[7], Hadi Meidani[2], Ryan H. Nguyen[1], Ameen A. Salahudeen[1*].

Departments of [1]Medicine, [3]Radiation Oncology, [6]Bioengineering, College of Medicine, University of Illinois at Chicago.

[2]Department of Civil & Environmental Engineering, Grainger College of Engineering, University of Illinois Urbana Champaign

[4]School of Public Health, University of Illinois at Chicago

[5]Department of Pathology, Biological Sciences Division, University of Chicago

[7]Department of Computer Science, College of Engineering, University of Illinois at Chicago

*Correspondence: ameen@uic.edu

ABSTRACT

Importance

Lung cancer disparities occur across minorities, namely Black populations, who face increased risks yet are screened at lower rates. Standards set by the United States Preventive Services Task Force (USPSTF) are derived from a predominantly White cohort: the National Lung Cancer Screening Trial (NLST), which exacerbates disparities in lung cancer screening (LCS) and diagnosis.

Objective

To evaluate individualized risk assessment using highly accurate risk models that integrate clinical and imaging-based risk factors for lung cancer prediction for improving LCS accuracy to reduce disparities among minoritized populations.

Design, Setting, and Participants

A retrospective real-world patient cohort from University of Illinois Health (UIH) using available LDCT scans (January 1, 2015 to March 16, 2024) was assembled. We then evaluated the performance of a ResNet-18 model trained on LDCTs from the predominantly white NLST cohort on the diverse UIH patient population, consisting of 65,106 patients, of which 8,823 identify as Black. Inclusion criteria of the UIH cohort utilized CPT codes, as well as ICD-9 and ICD-10 criteria for neoplasm of the bronchus or lung. The proposed hybrid model was assessed for its predictive accuracy across different racial groups and Body-Mass Index (BMI) categories.

Main Outcomes and Measures

The primary outcomes included the hybrid AI model's ability to improve lung cancer screening adherence, its effectiveness across diverse racial groups—highlighting disparities in performance between Black and White populations—and its performance in individuals with

varying BMI, particularly those with BMI ≥ 30. Secondary outcomes were the hybrid model's performance in terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) compared to traditional USPSTF guidelines.

Results

The hybrid AI model was trained using clinical and imaging data from the NLST cohort and tested on a diverse urban and suburban population in the Chicago metropolitan area (UIH cohort). The model, optimized to 7 clinical features, achieved ROC-AUC values of 0.64-0.67 in the NLST test set and 0.60-0.65 in the UIH cohort. The inclusion of ResNet-based image predictors significantly improved the model's performance, achieving ROC-AUC values of 0.78-0.91 and PR-AUC values of 0.25-0.33 in NLST. However, the hybrid model's performance deteriorated when applied to Black patients in the UIH cohort, with ROC-AUC values of 0.65-0.75, and to 0.67 in obese patients (BMI ≥ 30). Further investigation found the ResNet-18 model was the underlying cause of the disparate results with higher performance among White patients compared to Black patients UIH patients. Attempts to optimize the ResNet-18 outputs revealed a domain shift, where model optimization in Black patients resulted in deterioration in White patients, reflecting the limited representation of Black patients in the model's original training dataset. Model performance also deteriorated for individuals with a BMI ≥ 30 in both the NLST and UIH data sets.

Conclusions and Relevance

The hybrid AI model shows promise in providing personalized lung cancer risk predictions with improved accuracy compared to clinical risk models alone. However, biases in training data, particularly regarding race and BMI, limit its generalizability. Future work should focus on developing more inclusive training datasets and further validating the model in diverse prospective cohorts to enhance its applicability in reducing lung cancer disparities.

INTRODUCTION

Lung cancer treatment in the United States carries health disparities in minority populations and are especially pronounced in Black individuals[1]. Low-dose CT (LDCT) lung cancer screening shifts diagnosis to early-stage disease and is associated with reduced mortality[2, 3], but stringent eligibility criteria from the United States Preventive Services Task Force (USPSTF)[4] exacerbates disparities in Black populations[4, 6]. These criteria, specifically regarding age and duration of tobacco usage, were derived from risk/benefit analysis of the ~ 92% White National Lung Screening Trial (NLST) cohort. This results in exclusion of at-risk Black patients from screening. The eligibility to incidence ratio measures this disparity, and according to current USPSTF guidelines, White populations often have >100 eligible patients screened for one incidence of lung cancer, whereas in Black populations this is reduced to approximately 50 eligible patients for every diagnosed case of lung cancer[4]. We[7] and others[5,8,9] have identified that individualized selection using clinical feature-based risk models improved lung cancer screening eligibility and sensitivity as compared to USPSTF guidelines. Moreover, individualized lung cancer risk prediction from LDCT has been achieved using Convolutional Neural Networks (CNNs). Taken together, risk prediction models have the potential to guide screening selection, to follow up, to determine eligibility, and to mitigate lung cancer disparities in early detection and survival. However, these models have been trained on patient cohorts which underrepresent minorities. To advance these risk models for implementation, their generalizability must first be assessed in minority populations including Black patients.

In this retrospective ever-smoker cohort analysis, we evaluated the utility of clinical and imaging based lung cancer risk models in both the NLST cohort, and a diverse urban and suburban real-world patient population across multiple sites in the Chicago metropolitan area (University of Illinois Health, UIH) representing 65,106 patients including 8,823 Black patients. The Chicago metropolitan area is racially, socioeconomically, and geographically diverse and has regions

with higher rates of lung cancer, predominantly in areas populated by Black patients[10-12]. We find that a hybrid deep learning model using ResNet and Support Vector Machines (SVMs) from LDCT and clinical features, respectively, outperformed individual models in the NLST test set. When evaluated in the UIH cohort, we observed a racially disparate performance deterioration in the hybrid model on Black – but not White subjects. Further analysis uncovered the model deterioration occurred within the ResNet imaging component of the Hybrid model. Efforts to regularize model performance via Artificial Neural Networks (ANNs) uncovered underlying disparities where improvement of accuracy in Black patients coincided with decreased accuracy in White patients. However, the residual bias still affected the Black population's performance. This could potentially be due to a large imbalance of LDCTs in White versus Black patients in the ResNet training cohort. Overall, we show proof of concept that a hybrid model has the potential to identify individuals at risk of developing lung cancer with high accuracy in this retrospective cohort, but with disparate performance in Black patients. To mitigate against further lung cancer screening disparities, risk-based model performance must be benchmarked across all races and be trained in patient cohorts that have adequate representation of Black and other minoritized populations.

METHODS

Research analysis was carried out under approved UIC IRB protocols 2023-1321 and 2023-1377. We developed a hybrid AI predictive model to probe the biological basis of lung cancer, and thus assembled a real-world cohort of > 10,000 patients treated in urban and suburban settings in the city of Chicago as well as greater Cook County. This real-world cohort, including > 5,000 Black patients with history of smoking and/or lung cancer, was also filtered based on available chest imaging studies.

128 Collecting Data: NLST

129 Deidentified NLST data was accessed from the National Cancer Institute Cancer Data Access
130 System approved under UIC IRB Protocol 2023-1321.

131

132 Collecting Data: UIH Lung Cancer Screening

133 To assess the hybrid model in the UIH patient population, we assembled a retrospective cohort
134 of patients with a history of smoking under an IRB approved protocol. Inclusion criteria for this
135 cohort is listed in **Supplementary Table 1**. Briefly, LDCT screening patients were placed into the
136 cohort by CPT code 71271; and/or based on ICD codes for smoking cessation counseling and
137 cigarette smoking, and through documented history of smoking. ICD-9 code 305.1 and ICD-10
138 code F17 were used to select LDCT patients and LDCT associated DICOMS were chosen; In
139 addition to LDCT patients, we also expanded the real-world cohort to include patients with
140 smoking histories and CT imaging. A diagnosis of lung cancer was determined based on ICD
141 codes of neoplasm of bronchus or lung (ICD-9 162, ICD-10 C34). Those without a diagnosis of
142 lung cancer were assumed to have no lung cancer at the time of the preparation of this
143 manuscript. Median duration of follow-up between LDCT scans was 431 days and the median
144 number of LDCT scans per patient with at least one LDCT was 1 (the median is 5 if each
145 convolution kernel and reconstruction of a given LDCT exam is counted separately).
146 The composition of races and ethnicities in both the NLST cohort and the UIH cohort are
147 presented in (**Fig. 1A).** The UIH cohort had a diverse geographic distribution including urban,
148 semiurban, and suburban and patients received healthcare across multiple sites (**Fig. 1B**).

149

150 Generating Ground Truth

151 The ground truth is represented by six values (1 or 0) indicating the presence or absence of a
152 lung cancer diagnosis 1 to 6 years after a CT scan was performed. Using the Python programming
153 language, a data frame was created which represents the ground truth associated with each

154 LDCT in both the NLST and UIH data sets. The time of each CT event was compared to the time
155 of the lung cancer diagnosis event of the associated patient. If the difference is less than or equal
156 to a given number of years, it is labeled as a positive sample (1), otherwise it is labeled as a
157 negative sample (0). However, if the sample has no known lung cancer diagnosis event, and the
158 number of years exceeds the number of days to last follow-up event (or in the case of UIH data,
159 the date 3/16/2024), then the sample is removed due to a lack of information in determining the
160 ground truth. Over years 1-6, 1888, 3128, 3828, 4056, 4476, and 4884 samples (out of 5,980)
161 were removed from the UIH cohort respectively. The NLST data set includes the number of days
162 to the last known cancer status of each patient, and this was used in a similar manner to remove
163 4, 35, 159, 280, 461, and 1082 samples (out of 11,980) from the NLST data set for each respective
164 year.
165
166
167
168 Train and Test set selection in NLST
169 A training set from the NLST was created using the same split as previously described[14] with each
170 sample representing an individual patient. In addition, data from patients in the chest X-ray arm
171 of NLST was also incorporated. The SVM models which incorporate a Sybil score were also
172 trained using the same training set, however each sample represents an individual CT scan rather
173 than a patient.
174
175 Test set selection in UIH
176 Data from patients at UIH consisted of various Digital Imaging and Communications in Medicine
177 (DICOM) types, a digital format for storing CT scans and associated metadata. First, LDCTs were
178 selected using codes associated with each DICOM, provided by the UIC Center for Clinical and
179 Translational Sciences.

180

181 Prediction with $PLCO_{m2012}$

182 $PLCO_{m2012}$ is a logistic regression model for lung cancer risk inference and was evaluated on
183 NLST data. The $PLCO_{m2012}$ model was published in 2012, and was created using a set of 11
184 chosen features to predict the probability of lung cancer diagnosis for a given patient within 6
185 years: age, race, education, Body-Mass Index (BMI), presence of Chronic Obstructive Pulmonary
186 Disease (COPD), personal history of cancer, family history of lung cancer, smoking status,
187 cigarettes per day, duration of smoking, and duration of quitting. All 11 features are directly
188 represented or can be extracted from the data in the NLST. The $PLCO_{m2012}$ logistic regression
189 model is represented in an R package created by the authors of $PLCO_{m2012}$[16]. This was translated
190 to Python code to facilitate incorporation with the remainder of the code base. When using
191 $PLCO_{m2012}$ for prediction, each sample represents an individual patient. The model was used to
192 generate predictions for probability of lung cancer diagnosis within 6 years only (it was not used
193 to evaluate performance within 1 to 5 years) as intended by its authors (**Supp. Fig 3**), and these
194 predictions were compared with the ground truth at 6 years, and performance was evaluated
195 across White versus Black patients.

196

197 SVM Prediction

198 Clinical data elements from the NLST were evaluated for inclusion in a SVM using the Scikit-
199 Learn Python library[13]. 11 features were retained after recursive feature elimination using linear
200 kernel. We started with a large set of initial features from those available in NLST, as well as some
201 generated features which were calculated from existing ones, such as BMI. Recursive feature
202 elimination showed best results when using 11-13 features. We chose to optimize the SVM to 11
203 features since this is the fewest number of features which resulted in optimal performance. Six
204 unique SVMs were then trained to make predictions 1-6 years post CT scan, with each sample
205 being an individual patient. Each of the 6 SVM models was then evaluated on the test sets using

206 the corresponding ground truths. Additional SVM models were created using a truncated set of
207 features to allow comparison between UIH and NLST cohorts. This truncated set contains 7
208 features which are well-represented in both data sets; the 4 removed features were poorly
209 represented in the UIH cohort. The training and prediction pipeline of SVM method is given in **Fig.**
210 **2A**.
211
212 ResNet Prediction
213 DICOM data files were evaluated by a ResNet-18 CNN model, trained on the Imagenet database
214 with parameters locked and further trained upon the NLST cohort as previously described[14]. The
215 ResNet-18 model, Sybil, was run with default parameters on the same test set used by the authors
216 of the model which contains data from 2,328 patients and 11,980 images, and the resulting output
217 from each LDCT image was a set of 6 values indicating the probability that lung cancer was
218 diagnosed within 1 to 6 years from the date of the scan. The ResNet was used in a similar manner
219 on data from UIH across 695 patients with 4,092 LDCT images. Each prediction score is then
220 compared with the ground truth values to determine the model's performance, broken down into
221 6 performance evaluations by years 1 to 6.
222
223 SVM + ResNet Hybrid Model Prediction
224 For the hybrid model, SVM learning was used to train six unique models for years 1-6 post imaging
225 including the corresponding ResNet inferred risk for that given year (**Fig. 2B)**. Creating the
226 ResNet + SVM models involved the same process as the initial SVM, with two main differences.
227 Each ResNet + SVM model incorporates the ResNet risk inference of the current year and that
228 the training and testing samples are individual DICOMs associated with the clinical features of
229 each subject ID at baseline study entry. The SVM + Resnet hybrid models thus utilized the same
230 subject ID for training and testing the ResNet model.
231

232 ResNet+ANN Hybrid Model Prediction

233 A simple fully connected artificial neural network (ANN) was constructed to incorporate race and
234 age along with a ResNet prediction score, specifically the 1-year prediction score generated by
235 the ResNet. There are thus 3 inputs: a numerical age, a binary value (0 or 1) indicating if the
236 patient is Black, and LDCT-based score generated by the ResNet representing the probability of
237 cancer diagnosis (Fig. 2C). Patients listed as “more than one race” in the NLST cohort were
238 evaluated as not Black. This category was not present in the UIH cohort. This hybrid model has
239 6 outputs, each output representing the probability of lung cancer diagnosis by year N (1 to 6).
240 There are 2 fully connected hidden layers. 100 independent ANNs were trained using randomly
241 initialized parameters, and the top 10 performers on a validation set were used to create an
242 ensemble to determine if race data may improve the predictive performance of the original ResNet
243 model.

244

245

246 Evaluating Model Performance

247 Model performance was evaluated on completely held out test sets within the NLST cohort as
248 well as the real-world UIH cohort. Statistical comparisons were performed using paired and
249 unpaired DeLong tests to compare receiver operating characteristic (ROC) area-under-curve
250 values. When models are tested on a given test set, both a precision-recall (PR) curve and a ROC
251 curve are generated using the Scikit-Learn Python library. Then, a prediction cutoff value is
252 selected such that the sensitivity is as close to 0.8 as possible to capture 80% of lung cancer
253 diagnosis events based on proposed sensitivities utilized in the International Lung Screening Trial
254 for the $PLCO_{m2012}$ model. The remainder of the statistics (i.e., specificity, PPV, NPV) were
255 calculated based on this cutoff value.

256

257

258 Comparison of models

259 To determine whether there was a significant difference in the performance of a given model on
260 a similar population, DeLong's test was utilized. Unpaired analysis was performed as previously
261 described[15].

262

263 **RESULTS**

264 NLST and UIH cohorts evaluated in this study and racial and ethnic breakdown initially comprised
265 of 53,452 patients in NLST and 11,654 patients in UIH (Table 1). We sought to develop a hybrid
266 predictive model to combine the long-term predictive potential of clinical features as well as near-
267 term predictive potential of radiographic features from CT. To develop a hybrid lung cancer
268 prediction model, we used the NLST as a training data set inclusive of CT scans and Chest X-
269 Ray arms amounting to 48,628 total patients[2] for the training of the SVM. A SVM model was
270 chosen owing to its interpretability of features and weights and allowed for multi-year risk
271 prognostication from baseline characteristics of participants at entry (**Supp Fig. 1**). The SVM was
272 also selected as it is effective at classifying two classes (i.e. patients with no lung cancer diagnosis
273 versus patients with lung cancer diagnosis) where ground-truth was measured by the time from
274 study entry to cancer incidence within the window of that year's specific model (e.g. SVM_1 predicts
275 the probability of lung cancer within one year or less, SVM_2 within two years etc.). We first focused
276 on and initiated the SVM models with 65 features and utilized recursive feature elimination to
277 optimize the model to 11 features (**Supp Fig. 2, Supp. Table 2**). These 11 features overlapped
278 with many previous features identified in the $PLCO_{m2012}$ 6-year risk model but were predictive of
279 multi-year risk as opposed to just 6-year risk (**Supp. Table 2**)[16]. We then further optimized the
280 model for real-world application by eliminating features not captured in electronic health records
281 within routine clinical care since data was not available in > 50% of patients. The excluded
282 features included former vs current smoker, whether patients have family history of lung cancer,
283 patients' education whether higher than bachelor's or not, and the maximum number of years

exposed to one of the following at work: asbestos, chemical, sand blasting, coal, foundry. We found that the SVM model could be reduced to 7 clinical features well represented in EHR of the UIH cohort while still retaining predictive performance (**Fig. 3A, Supp Fig. 2**). Performance of the 7-feature multiyear SVM model was first evaluated in a held-out test set from the NLST cohort with ROC-AUC values of 0.64-0.67 similar to the 6-year lung cancer risk when compared to the $PLCO_{m2012}$ model (**Supp. Fig 2**). We then applied the SVM to the UIH cohort of patients and found similar performance with ROC AUC 0.60-0.65 (**Fig. 3B**).

Following initial model optimization of clinical risk features, we then incorporated an image-based lung risk predictor of the ResNet-18 CNN image classifier with additional training on LDCT images from NLST, Sybil[14], which infers lung cancer risk probabilities at multiple years of follow up. The hybrid risk model was trained with the previously identified 7 clinical features in the SVM but with inclusion of the inference of the multiyear risk from the image based ResNet model from 65,161 LDCT study DICOM files. The Hybrid model demonstrated substantially improved accuracy with ROC-AUC values Y1-Y6 all significantly improved (p-values < .001, **Supp. Table 4**). We then benchmarked the hybrid model on the NLST test set compared to SVM models trained with clinical risk features alone which resulted in a substantial improvement of predictive potential (**Fig 3C**; **Table 2;** ROC-AUC ranging from 0.78-0.91; PR-AUC values 0.25-0.33). Positive predictive values and negative predictive values at a sensitivity of 80% (corresponding to a $PLCO_{m2012}$ 6-year risk of 1.6%)[16] and found an order of magnitude improvement in positive and negative predictive values when compared to the $PLCO_{m2012}$ risk predictive model (**Supp. Table 3**)[16].

We then evaluated whether our hybrid model's predictive ability could be reproduced in the diverse UIH cohort. We observed a substantial decrease in model performance (**Fig 3D**; ROC AUC 0.68-0.80). Given the similar performance of the clinical feature SVM model between the NLST and UIH cohorts, we focused on the ResNet component of the Hybrid model. To confirm

310 the ResNet model performance was reproducible, we re-ran the model on individual CT scans
311 from the NLST that were not used for the ResNet model training and confirmed findings as
312 previously reported[14] (**Fig. 4A**). We then evaluated baseline LDCT scans from the UIH cohort and
313 identified reduced performance in the UIH cohort with Year 1 ROC-AUC values of 0.8 versus 0.92
314 in NLST, for example (**Fig. 4B**). We then asked whether model performance deterioration[17] was
315 due to a technical difference of the LDCT scans (e.g. modern CT scanners and image
316 reconstruction in UIH versus NLST) or due to the demographic differences such as race.
317 Evaluating performance of the ResNet model on White patients in the UIH showed no differences
318 in performance (**Fig. 4C**) between the NLST and the UIH cohorts. However, when assessed in
319 Black patients, there was marked deterioration in model performance with Y1-Y6 ROC AUC
320 values 0.65-0.75 (**Fig. 4D**). We also evaluated other potential factors of model deterioration and
321 found differences in BMI where model performance had significant deterioration in obese patients
322 (BMI ≥ 30) (**Supp. Table 4**) where in the NLST cohort Y1 ROC AUC was 0.94 in BMI < 30 versus
323 0.88 in BMI ≥ 30 subjects. In the UIH cohort, we observed a more pronounced model deterioration
324 where Y1 ROC AUC was 0.90 BMI < 30 versus 0.67 in BMI ≥ 30 subjects (**Supp. Table 4**).
325 Notably, we found no correlation between race and BMI to suggest that these clinical features
326 were linked to model deterioration.
327
328 Because of the racial disparity in model performance, we subsequently inquired whether this
329 could be mitigated by incorporating clinical information along with LDCT images. We utilized a
330 simple two-layer ANN given ease of training to train a large number of unique models. We then
331 used the NLST training cohort for ANN model training, 10% of the UIH cohort for testing, and
332 completely held out 90% of the UIH cohort for validation. The ANN models were trained with the
333 following three inputs: ResNet predicted lung cancer risk, subject's age, and subject's race. 100
334 ANN models were trained with Year 1 ROC AUC performance ranging from 0.898 to 0.908 in the
335 NLST test set (**Fig 5A**). We then tested on 10% of the UIH cohort (UIH validation set) and utilized

336 the top 10 performing ANNs to be evaluated in the held out UIH test cohort. The results were
337 mixed, where in the overall UIH cohort Y1 ROC AUC values increased to 0.83 (**Fig 5B**) as
338 compared to the ResNet model alone (**Fig 4B**). When evaluating this difference across races, we
339 observed that ROC AUC values modestly increased by 0.01 to 0.02 across Years 1-6 (**Fig 5D;**
340 **Table 3**) in Black subjects as compared to the ResNet model alone (**Fig 4D**). However, in White
341 subjects, there was a substantial decrease in performance in Y2-6 ROC-AUC values by 0.10 (**Fig**
342 **5C**) as compared to the ResNet model alone (**Fig 4C**). While domain shift - which refers to the
343 situation where the distribution of the training data differs from the distribution of the test data - is
344 a significant factor in the performance discrepancies observed, the bias in the training data and
345 model generalization also play critical roles. The observed results indicate that while the new
346 model has improved overall performance, ongoing disparities reflect the lingering impact of the
347 initial training data biases and the complexity of the population-specific factors influencing lung
348 cancer risk predictions.
349
350
351 **DISCUSSION**
352 Lung cancer disparities are extensively documented in minority populations with respect to
353 survival outcomes, clinical management, and screening. Recognizing these disparities as a
354 potential means to uncover biological factors, we developed a hybrid AI model that integrates
355 extracted clinical and lung parenchymal imaging features for risk inference. This novel approach
356 not only addresses the limitations of traditional screening strategies but also enhances sensitivity
357 and specificity for diverse populations. While LDCT screening has proven effective in reducing
358 mortality through early detection and improved survival[2,18], existing guidelines have been shown
359 to perpetuate inequities. The USPSTF LDCT criteria[18], primarily based on the nearly 95% White
360 NLST study population[2] has been shown to contribute to disparities where Black populations
361 screening eligible population carries twice the risk of lung cancer as compared to White

populations[4,5]. Alternative strategies have been proposed including utilizing clinical risk models to determine screening eligibility[5,7]. While these strategies demonstrate increased sensitivity as compared to USPSTF criteria, they have several limitations including patient selection only at long-term risk levels > 1-1.5% lung cancer risk, lack of risk adjustment from LDCT findings, and limited PPV and NPV values. When the hybrid model was evaluated in the diverse real-world cohort within the Chicago metropolitan area, we noted the hybrid model demonstrated higher performance among populations identifying as White as compared to Black, potentially reflecting bias from a training set of a >90% White population. Bias in AI models is an increasing concern, particularly among institutions serving diverse patient populations and may further exacerbate health disparities[19,20]. However many initiatives are aiming to reduce bias in AI development including AIM-AHEAD[21,22] and specifically in lung cancer screening, the National Cancer Institute's CANDID-4AI[23]. Because of the existing disparities in lung cancer including the higher risk of lung cancer observed among Black patients who smoke, implementation of AI models will require representative data in training sets and inclusion of Black populations and other minorities in validation cohorts and studies. Furthermore, the domain shift we observed suggests that lung parenchymal features in White LDCT images have different features and latent representations than in Black individuals. The biological basis of this might be related to the lung parenchyma in Black individuals who smoke, which requires further research and radiomic, histologic and cellular characterizations. In addition, we observed model deterioration in individuals with a BMI ≥ 30. While the underlying mechanism for this deterioration in both the NLST and UIH cohorts was outside of the scope of this study, we speculate that latent features within lung parenchyma may be inadequately visualized due to a biological obesity paradox that confounds our hazard risk models[24,25]. Alternatively, chest wall thickness from subcutaneous adipose tissue may interfere with ray penetration, or cause compression of lung parenchyma due to obesity hypoventilation.

387 This study has limitations, chiefly those of the retrospective nature of the UIH cohort which is
388 subject to bias and data missingness. We are aiming to address these limitations with a
389 prospective study on new LDCT participants across multiple sites. The hybrid model also utilizes
390 a CNN architecture trained on the NLST cohort which has limited representation of Black and
391 other minoritized populations. Current efforts such as the CANDID-4AI initiative[23] to develop more
392 inclusive training sets may eliminate these biases and overcome domain shifts as we observed
393 in our assessment of the ResNet model[14] where performance in White subjects decreased when
394 the model was optimized for Black subjects. Having an inclusive and representative training set
395 will allow ResNet and more advanced deep learning architectures with race labeling to understand
396 underlying structures of racial disparities in lung cancer incidence and incorporate these features
397 and enhance model predictive ability and generalizability.
398
399 Collectively, we demonstrate an integrated approach combining imaging features from human
400 lung parenchyma and clinical characteristics that elucidated biological contributors to lung cancer.
401 The risk prediction model is capable of predicting lung cancer in two distinct patient cohorts,
402 though it suffers from deterioration in Black populations and individuals with a BMI ≥ 30. Future
403 studies should be conducted aimed at understanding the biological explanation underlying
404 increased lung cancer incidence in Black populations through investigation of the lung
405 parenchyma in Black individuals who smoke or in tumor adjacent normal tissue. Understanding
406 these biological factors would increase our understanding of lung cancer, open up new avenues
407 for prevention, and potentially enable biological based strategies that move beyond USPSTF
408 criteria to create an inclusive and individualized paradigm for lung cancer screening.
409
410 **ACKNOWLEDGEMENTS**
411 We wish to thank the UIC Center for Clinical Translational Sciences, the UI Cancer Center Data
412 Integrated Shared Resource, for supporting this research. A.Z. was supported by an AI.Health4all

postdoctoral fellowship, M.B. was supported by a Winn-CIPP fellowship, A.A.S. received in kind support as a National Cancer Institute, Center for Cancer Health Equity Early Investigator Advancement Program Scholar,

FIGURE LEGENDS

Figure 1. Patient demographics and distribution **A)** Composition of races and ethnicities in the National Lung Screening Trial (NLST) cohort and the University of Illinois at Chicago (UIC) lung cancer screening cohort. **B)** Distribution of patient count by zip code in UIC cohort.

Figure 2. Pipelines of Machine Learning algorithms. **A)** Pipeline for implementation a Support Vector Machine (SVM) to predict lung cancer risk using NLST clinical and demographic data. **B)** Pipeline for the proposed Hybrid SVM model (SVM + ResNet18) combining LDCT scans with NLST clinical and demographic data to predict lung cancer risk. **C)** Pipeline for the proposed Hybrid model (ANN + ResNet18) utilizing NLST demographic data and LDCT scans to predict lung cancer risk.

Figure 3. PR and ROC-AUC curves for SVM model (A,B) and Hybrid models (C,D). **A)** SVM performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** SVM performance in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **C)** Hybrid AI model performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** Hybrid AI model performance in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). Numerical values in all panels are AUC values for each corresponding year.

Figure 4. PR and ROC-AUC curves for the ResNet model alone in NLST and UIH Cohorts. **A)** ResNet performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** ResNet model performance in UIH test set measured by PR-AUC curve (left) and ROC-

AUC curve (right). **C)** ResNet model performance in White subjects in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** ResNet model performance in the African American subjects in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). Numerical values in all panels are AUC values for each corresponding year.

Figure 5. PR and ROC-AUC curves for the hybrid models (ANN+ResNet) in NLST and UIH Cohorts. **A)** Hybrid ANN performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** Hybrid ANN model performance in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **C)** Hybrid ANN model performance in White subjects in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** Hybrid ANN model performance in African American subjects in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right).

TABLE LEGENDS

Table 1. The racial and ethnic breakdown of NLST and UIH cohorts evaluated in this study

Table 2. Comparative performance analysis over six years of NLST data across all Races, Sexes, and BMIs: PLCOm2012 vs. SVM vs. Sybil vs. Hybrid (SVM + ResNet) vs Hybrid (ANN + ResNet) – Asterisks indicate top results.

Table 3. Performance comparison of Sybil and enhanced models across White and Black races in UIH dataset: AUC scores over six years (Asterisks indicate best results)

465 **REFERENCES**

466 1. Soneji, S., Tanner, N.T., Silvestri, G.A., Lathan, C.S., and Black, W. (2017). Racial and
467 Ethnic Disparities in Early-Stage Lung Cancer Survival. Chest *152*, 587-597.
468 10.1016/j.chest.2017.03.059.

469 2. National Lung Screening Trial Research, T., Aberle, D.R., Adams, A.M., Berg, C.D.,
470 Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., and
471 Sicks, J.D. (2011). Reduced lung-cancer mortality with low-dose computed tomographic
472 screening. N Engl J Med *365*, 395-409. 10.1056/NEJMoa1102873.

473 3. Bandi, P., Star, J., Ashad-Bishop, K., Kratzer, T., Smith, R., and Jemal, A. (2024). Lung
474 Cancer Screening in the US, 2022. JAMA Intern Med. 10.1001/jamainternmed.2024.1655.

475 4. Pinsky, P.F., Lau, Y.K., and Doubeni, C.A. (2021). Potential Disparities by Sex and Race
476 or Ethnicity in Lung Cancer Screening Eligibility Rates. Chest *160*, 341-350.
477 10.1016/j.chest.2021.01.070.

478 5. Choi, E., Ding, V.Y., Luo, S.J., Ten Haaf, K., Wu, J.T., Aredo, J.V., Wilkens, L.R.,
479 Freedman, N.D., Backhus, L.M., Leung, A.N., et al. (2023). Risk Model-Based Lung Cancer
480 Screening and Racial and Ethnic Disparities in the US. JAMA Oncol.
481 10.1001/jamaoncol.2023.4447.

482 6. Aredo, J.V., Choi, E., Ding, V.Y., Tammemagi, M.C., Ten Haaf, K., Luo, S.J., Freedman,
483 N.D., Wilkens, L.R., Le Marchand, L., Wakelee, H.A., et al. (2022). Racial and Ethnic Disparities
484 in Lung Cancer Screening by the 2021 USPSTF Guidelines Versus Risk-Based Criteria: The
485 Multiethnic Cohort Study. JNCI Cancer Spectr *6*. 10.1093/jncics/pkac033.

486 7. Pasquinelli, M.M., Tammemagi, M.C., Kovitz, K.L., Durham, M.L., Deliu, Z., Rygalski, K.,
487 Liu, L., Koshy, M., Finn, P., and Feldman, L.E. (2020). Risk Prediction Model Versus United
488 States Preventive Services Task Force Lung Cancer Screening Eligibility Criteria: Reducing
489 Race Disparities. J Thorac Oncol *15*, 1738-1747. 10.1016/j.jtho.2020.08.006.

490 8. Field, J.K., Vulkan, D., Davies, M.P.A., Duffy, S.W., and Gabe, R. (2021). Liverpool Lung
491 Project lung cancer risk stratification model: calibration and prospective validation. Thorax *76*,
492 161-168. 10.1136/thoraxjnl-2020-215158.
493 9. Katki, H.A., Kovalchik, S.A., Berg, C.D., Cheung, L.C., and Chaturvedi, A.K. (2016).
494 Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer
495 Screening. JAMA *315*, 2300-2311. 10.1001/jama.2016.6255.
496 10. Byrne, C.A., Gomez, S.L., Kim, S., Oddo, V.M., Koh, T.J., and Fantuzzi, G. (2022).
497 Disparities in inflammation between non-Hispanic black and white individuals with lung cancer in
498 the Greater Chicago Metropolitan area. Front Immunol *13*, 1008674.
499 10.3389/fimmu.2022.1008674.
500 11. Kim, S.J., Kery, C., An, J., Rineer, J., Bobashev, G., and Matthews, A.K. (2024).
501 Racial/Ethnic disparities in exposure to neighborhood violence and lung cancer risk in Chicago.
502 Soc Sci Med *340*, 116448. 10.1016/j.socscimed.2023.116448.
503 12. Watson, K.S., Hulbert, A., Henderson, V., Chukwudozie, I.B., Aponte-Soto, L., Lerner,
504 L., Martinez, E., Kim, S., and Winn, R.A. (2019). Lung Cancer Screening and Epigenetics in
505 African Americans: The Role of the Socioecological Framework. Front Oncol *9*, 87.
506 10.3389/fonc.2019.00087.
507 13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel,
508 M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in
509 Python. the Journal of machine Learning research *12*, 2825-2830.
510 14. Mikhael, P.G., Wohlwend, J., Yala, A., Karstens, L., Xiang, J., Takigami, A.K.,
511 Bourgouin, P.P., Chan, P., Mrah, S., Amayri, W., et al. (2023). Sybil: A Validated Deep Learning
512 Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed
513 Tomography. J Clin Oncol *41*, 2191-2200. 10.1200/JCO.22.01345.
514 15. Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver
515 operating characteristic (ROC) curve. Radiology *143*, 29-36. 10.1148/radiology.143.1.7063747.

516 16. Tammemagi, M.C., Katki, H.A., Hocking, W.G., Church, T.R., Caporaso, N., Kvale, P.A.,
517 Chaturvedi, A.K., Silvestri, G.A., Riley, T.L., Commins, J., and Berg, C.D. (2013). Selection
518 criteria for lung-cancer screening. N Engl J Med *368*, 728-736. 10.1056/NEJMoa1211776.
519 17. Zhang, A., Xing, L., Zou, J., and Wu, J.C. (2022). Shifting machine learning for
520 healthcare from development to deployment and from models to data. Nat Biomed Eng *6*, 1330-
521 1345. 10.1038/s41551-022-00898-y.
522 18. Force, U.S.P.S.T., Krist, A.H., Davidson, K.W., Mangione, C.M., Barry, M.J., Cabana,
523 M., Caughey, A.B., Davis, E.M., Donahue, K.E., Doubeni, C.A., et al. (2021). Screening for Lung
524 Cancer: US Preventive Services Task Force Recommendation Statement. JAMA *325*, 962-970.
525 10.1001/jama.2021.1117.
526 19. Crowe, B., and Rodriguez, J.A. (2024). Identifying and Addressing Bias in Artificial
527 Intelligence. JAMA Netw Open *7*, e2425955. 10.1001/jamanetworkopen.2024.25955.
528 20. Ratwani, R.M., Sutton, K., and Galarraga, J.E. (2024). Addressing AI Algorithmic Bias in
529 Health Care. JAMA. 10.1001/jama.2024.13486.
530 21. Hendricks-Sturrup, R., Simmons, M., Anders, S., Aneni, K., Wright Clayton, E., Coco, J.,
531 Collins, B., Heitman, E., Hussain, S., Joshi, K., et al. (2023). Developing Ethics and Equity
532 Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity
533 in AI and Machine Learning: Modified Delphi Approach. JMIR AI *2*, e52888. 10.2196/52888.
534 22. Vishwanatha, J.K., Christian, A., Sambamoorthi, U., Thompson, E.L., Stinson, K., and
535 Syed, T.A. (2023). Community perspectives on AI/ML and health equity: AIM-AHEAD
536 nationwide stakeholder listening sessions. PLOS Digit Health *2*, e0000288.
537 10.1371/journal.pdig.0000288.
538 23. Institute, N.C. (2022). Leveraging Artificial Intelligence to Improve Accuracy of Lung
539 Cancer Screening was originally published by the National Cancer Institute.
540 https://prevention.cancer.gov/news-and-events/blog/leveraging-artificial-intelligence-improve-
541 accuracy-lung-cancer-screening.

542 24. Cespedes Feliciano, E.M., Kroenke, C.H., and Caan, B.J. (2018). The Plausibility of the
543 Obesity Paradox in Cancer-Response-Reply to Point. Cancer Res *78*, 1904-1905.
544 10.1158/0008-5472.CAN-17-3590.
545 25. Park, Y., Peterson, L.L., and Colditz, G.A. (2018). The Plausibility of Obesity Paradox in
546 Cancer-Point. Cancer Res *78*, 1898-1903. 10.1158/0008-5472.CAN-17-3043.

Figure 1

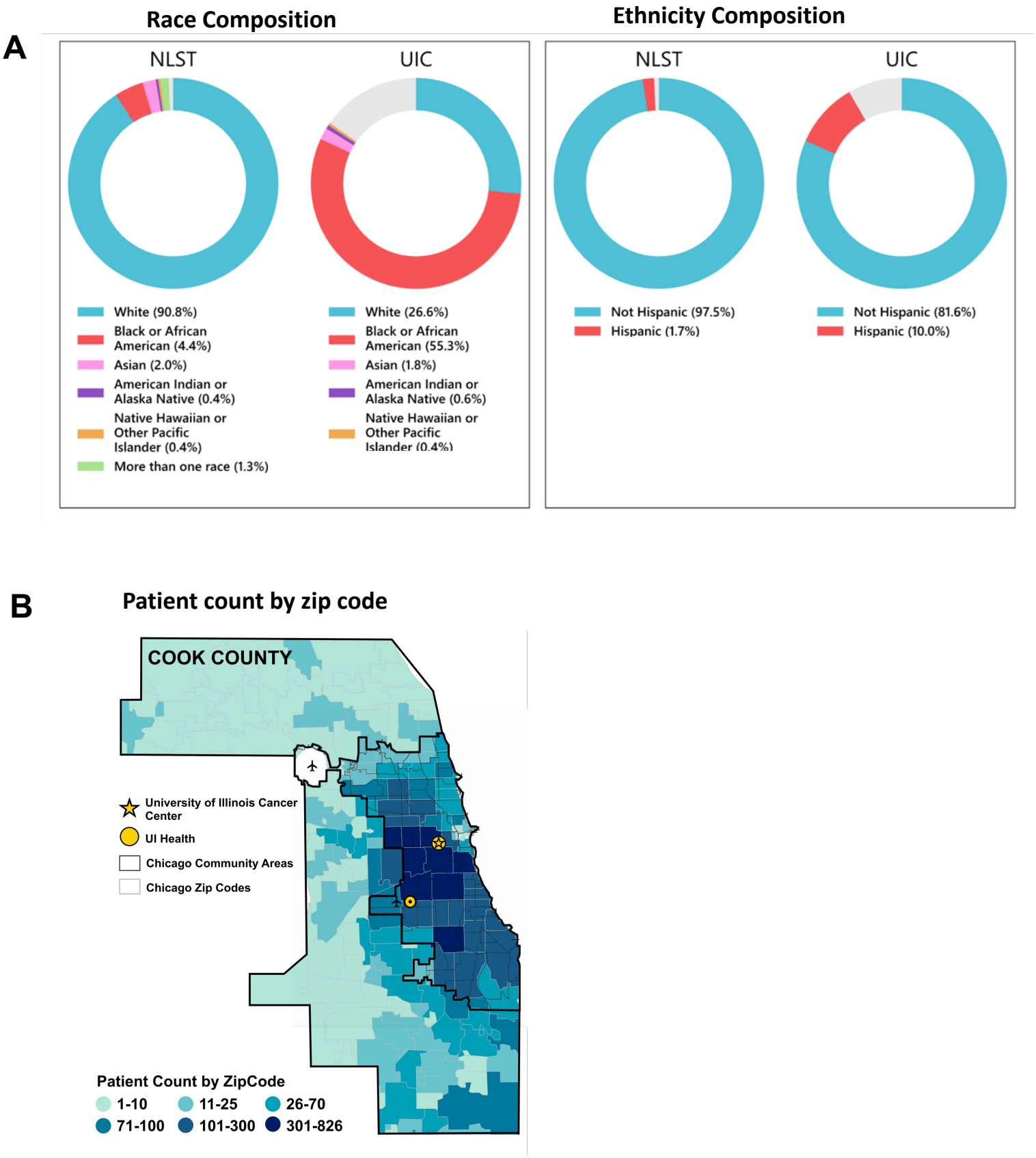


Figure 1. Patient demographics and distribution **A)** Composition of races and ethnicities in the National Lung Screening Trial (NLST) cohort and the University of Illinois at Chicago (UIC) lung cancer screening cohort. **B)** Distribution of patient count by zip code in UIC cohort

Figure 2

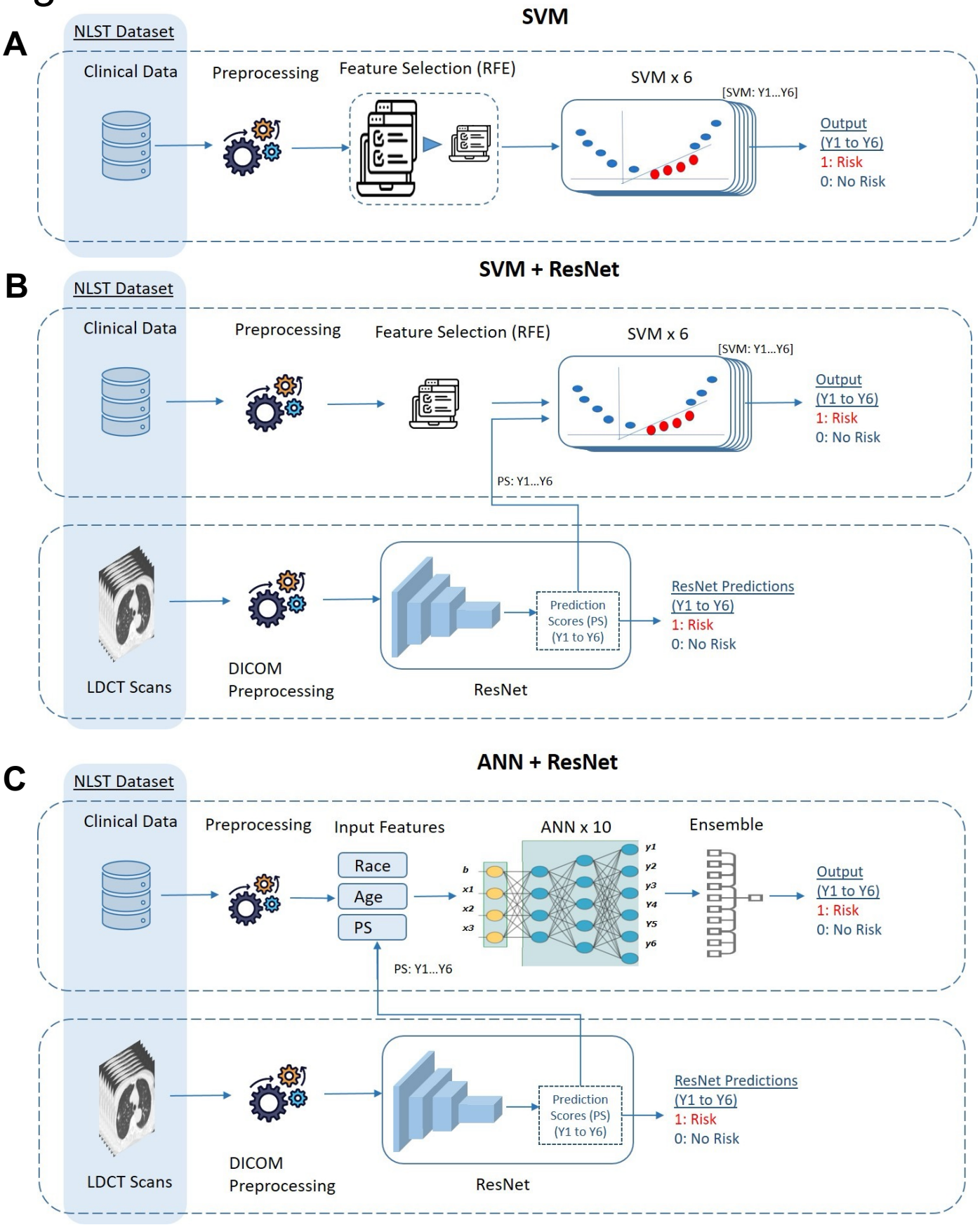


Figure 2. Pipelines of Machine Learning algorithms. **A)** Pipeline for implementation a Support Vector Machine (SVM) to predict lung cancer risk using NLST clinical and demographic data. **B)** Pipeline for the proposed Hybrid SVM model (SVM + ResNet18) combining LDCT scans with NLST clinical and demographic data to predict lung cancer risk. **C)** Pipeline for the proposed Hybrid model (ANN + ResNet18) utilizing NLST demographic data and LDCT scans to predict lung cancer risk.

Figure 3

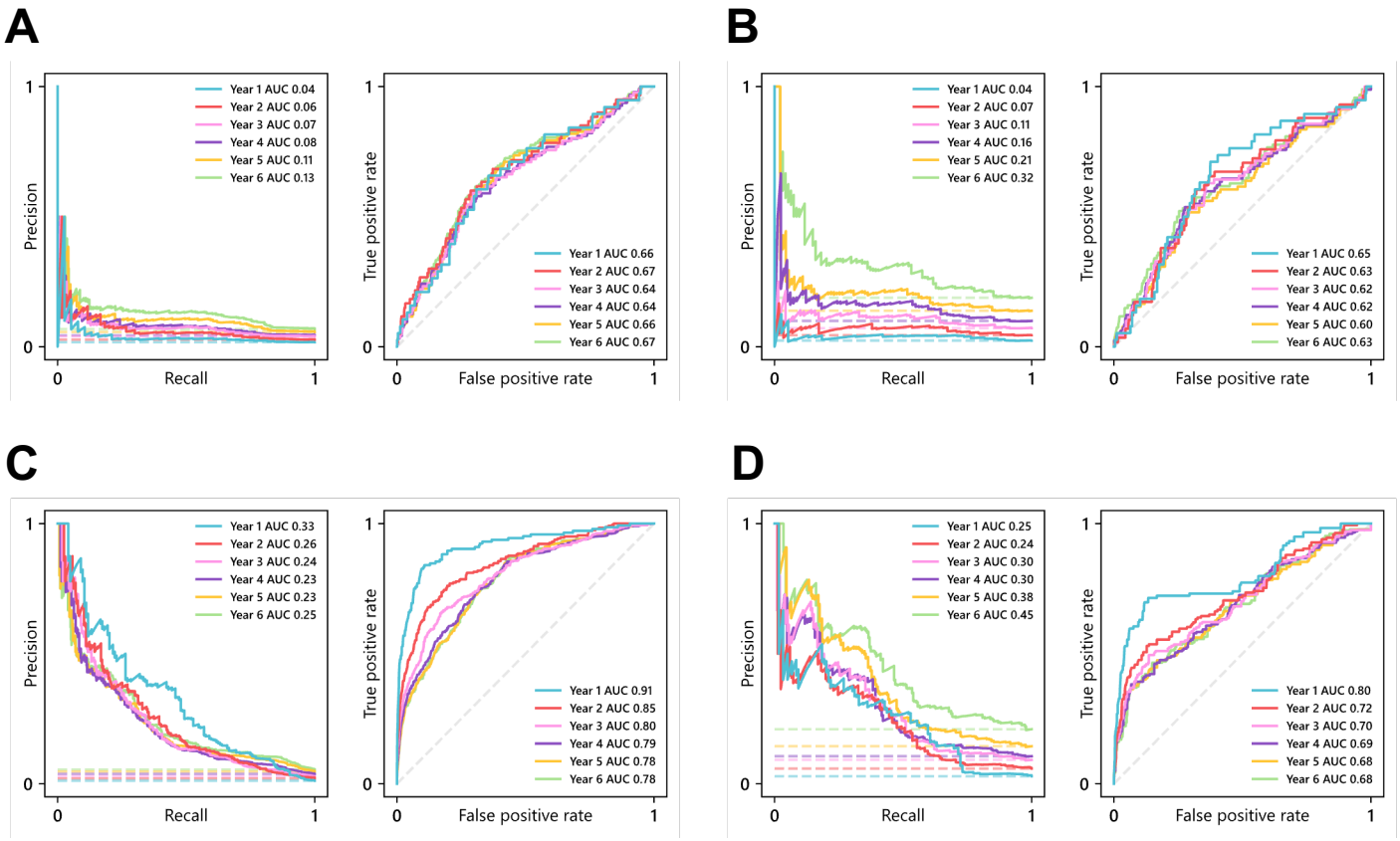


Figure 3. PR and ROC-AUC curves for SVM model (A,B) and Hybrid models (C,D). **A)** SVM performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** SVM performance in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **C)** Hybrid AI model performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** Hybrid AI model performance in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). Numerical values in all panels are AUC values for each corresponding year.

Figure 4

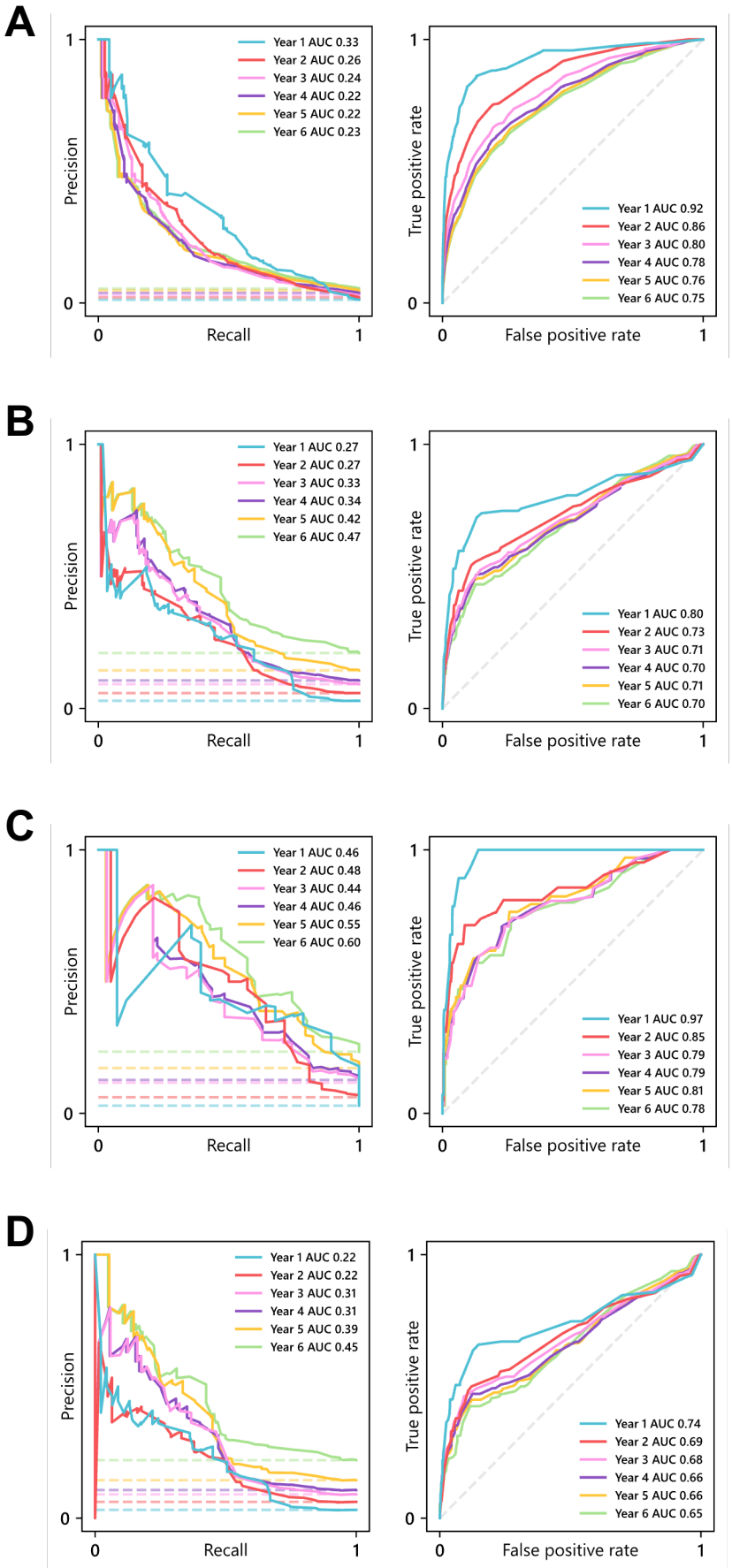


Figure 4. PR and ROC-AUC curves for the ResNet model alone in NLST and UIH Cohorts. **A)** ResNet performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** ResNet model performance in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **C)** ResNet model performance in White subjects in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** ResNet model performance in the African American subjects in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). Numerical values in all panels are AUC values for each corresponding year.

Figure 5

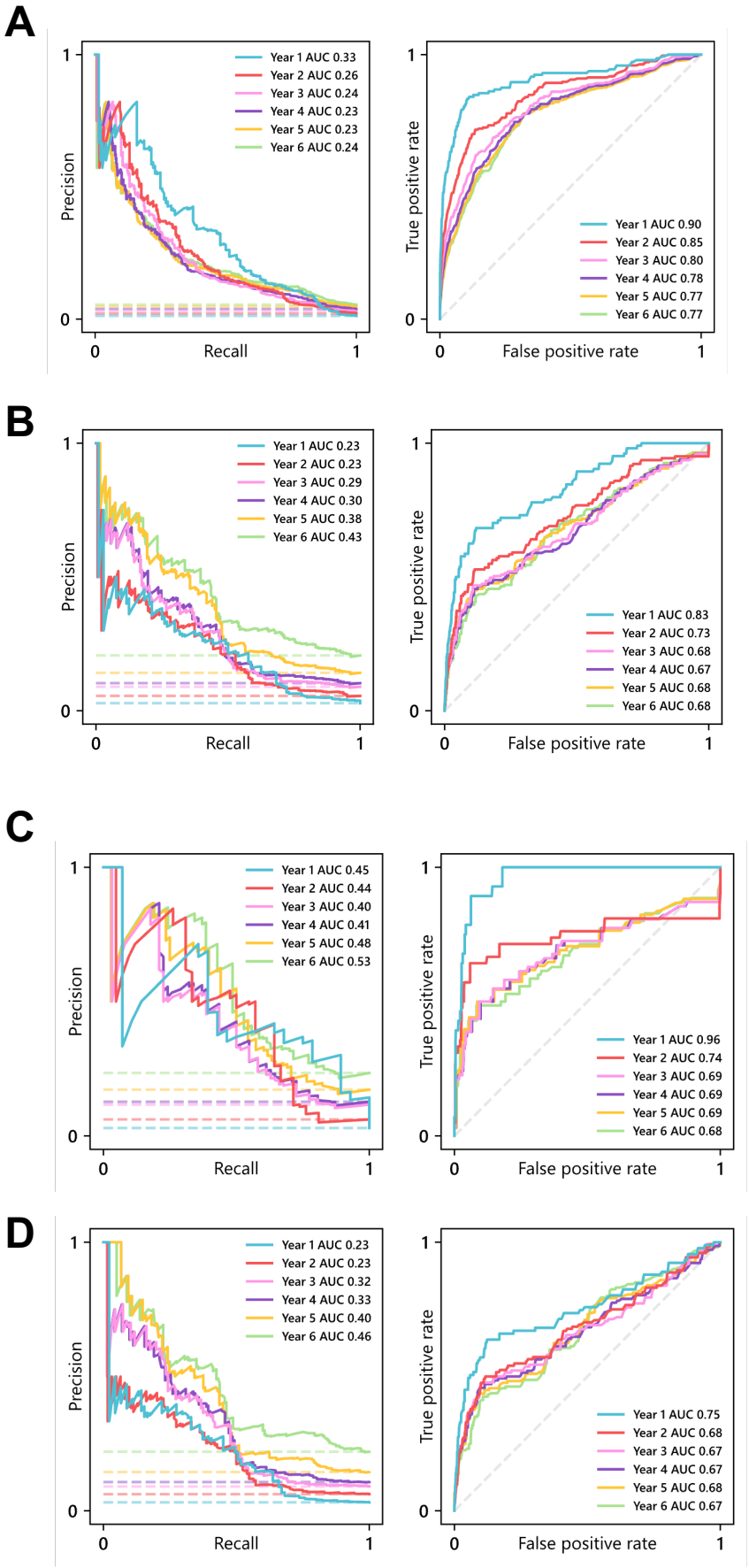


Figure 5. PR and ROC-AUC curves for the hybrid models (ANN+ResNet) in NLST and UIH Cohorts. **A)** Hybrid ANN performance in NLST test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **B)** Hybrid ANN model performance in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **C)** Hybrid ANN model performance in White subjects in the UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right). **D)** Hybrid ANN model performance in African American subjects in UIH test set measured by PR-AUC curve (left) and ROC-AUC curve (right).

Table 1

	NLST	UIC
Race		
White	48,549	3,099
Black or African American	2,376	6,447
Asian	1,095	210
American Indian or Alaska Native	190	68
Native Hawaiian or Other Pacific Islander	193	47
Other	1,049	1,783
Ethnicity		
Not Hispanic	52,118	9,514
Hispanic	935	1,170
No Information	399	970
Lung cancer incidence	3.2%	8.1%

Table 1. The racial and ethnic breakdown of NLST and UIH cohorts evaluated in this study

Table 2

Year	PLCOm2012	SVM11	Sybil	Hybrid SVM11 + ResNet	Hybrid ANN + ResNet
1	-	0.659	* 0.919	0.915	0.904
2	-	0.666	* 0.856	0.847	0.846
3	-	0.647	* 0.804	0.801	* 0.804
4	-	0.645	0.775	0.782	* 0.783
5	-	0.657	0.762	* 0.779	0.771
6	0.671	0.671	0.752	* 0.782	0.771

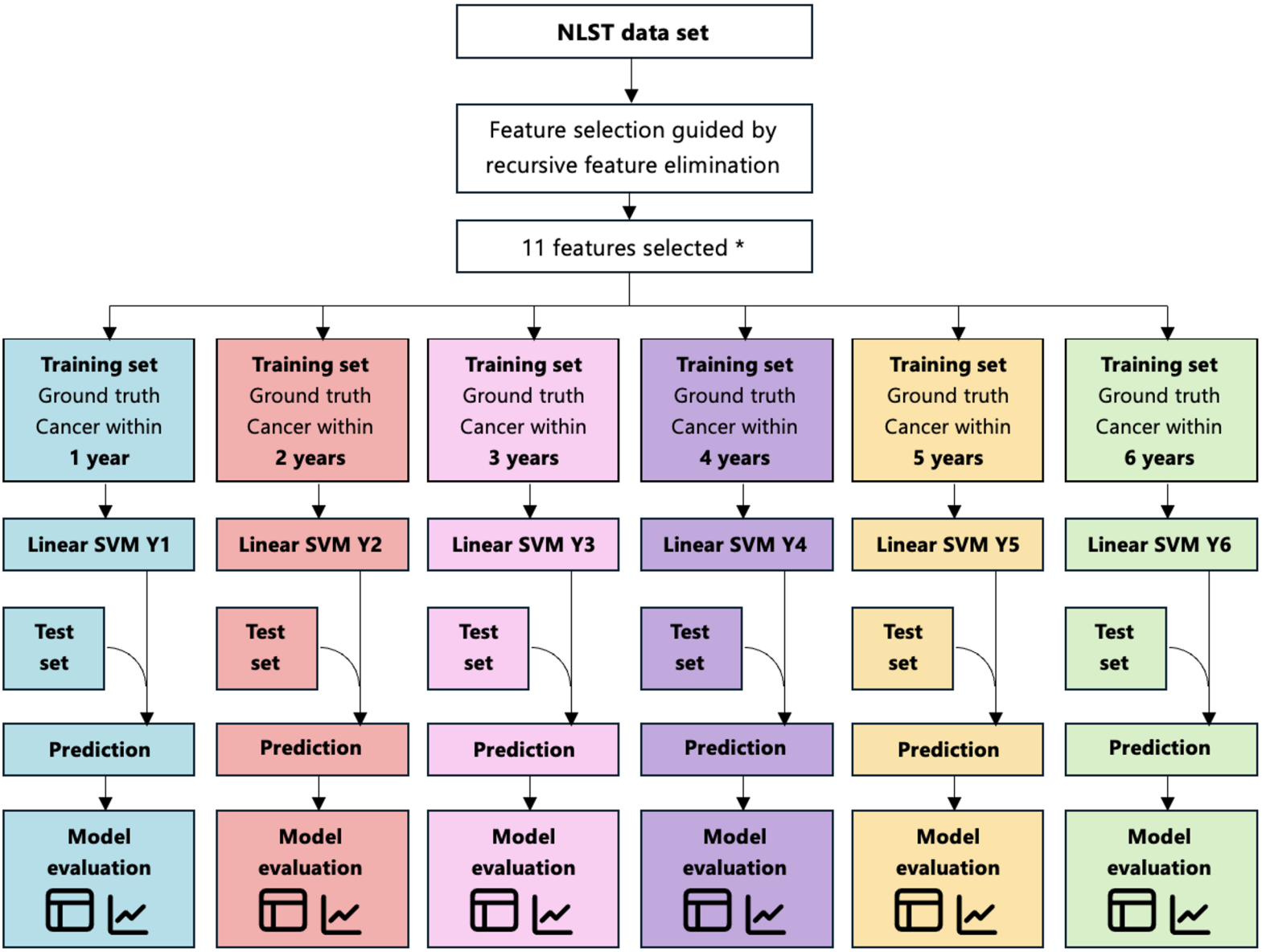
Table 2. Comparative performance analysis over six years of NLST data across all Races, Sexes, and BMIs: PLCOm2012 vs. SVM vs. Sybil vs. Hybrid (SVM + ResNet) vs Hybrid (ANN + ResNet) – Asterisks indicate top results

Table. 3

	White			Black		
Year	Sybil	Hybrid SVM + ResNet	Hybrid ANN + ResNet	Sybil	Hybrid SVM + ResNet	Hybrid ANN + ResNet
1	0.970	* 0.971	0.963	0.737	0.710	* 0.754
2	* 0.846	0.835	0.742	* 0.687	0.661	0.684
3	* 0.789	0.788	0.693	* 0.677	0.642	0.671
4	* 0.791	0.773	0.692	0.657	0.634	* 0.671
5	* 0.806	0.751	0.694	0.657	0.633	* 0.676
6	* 0.782	0.742	0.683	0.653	0.63	* 0.673

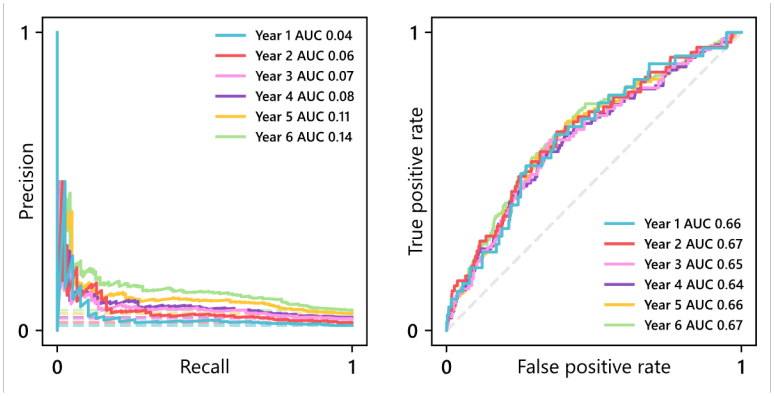
Table 3. Performance comparison of Sybil and enhanced models across White and Black races in UIH dataset: AUC scores over six years (Asterisks indicate best results)

Supplementary Figure 1



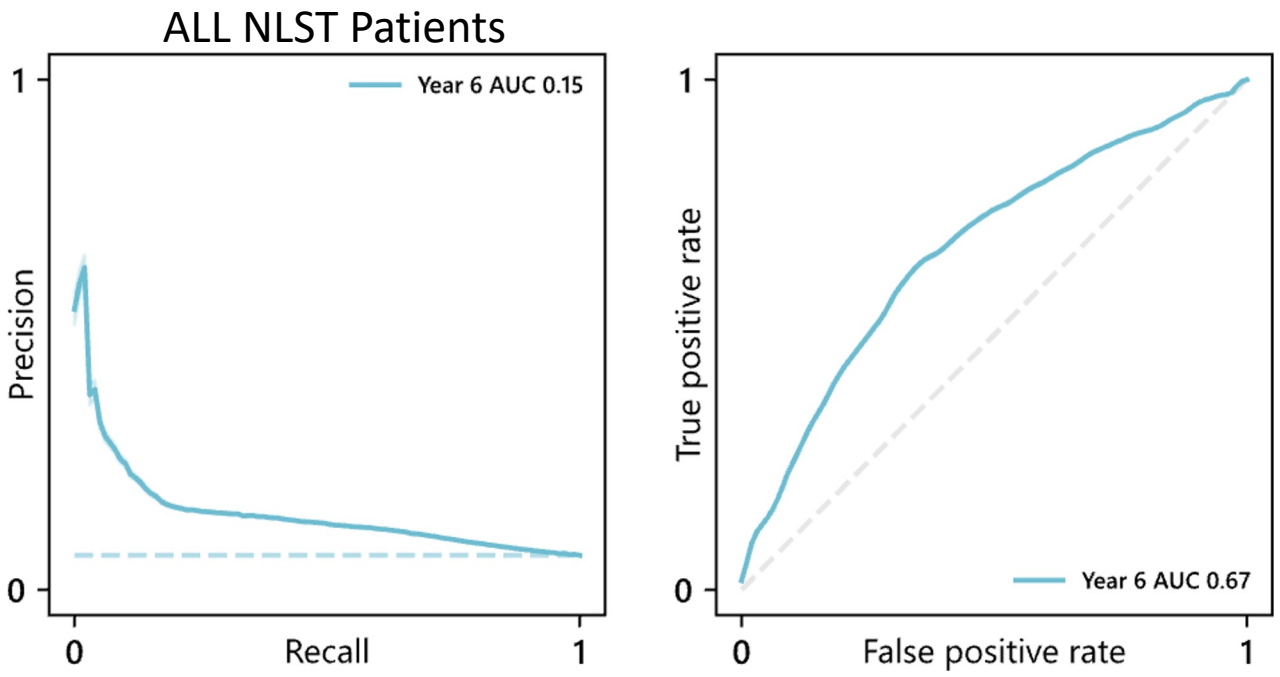
Supplementary Figure 1. Schematic of multi-year SVM training and evaluation

Supplementary Figure 2



Supplementary Figure 2. 11 Feature SVM model and corresponding PR-AUC (left) and ROC-AUC (right) plots in the NLST test set.

Supplementary Figure 3



Supplementary Figure 3. $PLCO_{m2012}$ model prediction of lung cancer at year 6 within the NLST test set. PR-AUC curve (left) with AUC 0.154, ROC-AUC curve (right) with AUC of 0.671.

Supplementary Table 1

SN	Criteria
1	Duration: January 1, 2015 till Present
2	Age: 50 to 80 years
3	Good health condition to undergo screening
4	Patients must have a prior CT scan of the thorax between zero and six years of the date of visit for lung cancer screening
5	All non-melanoma skin cancers ICD 9 173.X /ICD 10 C44 are eligible
6	No prior history of hematologic malignancy

Supplementary Table 1. The inclusion criteria for participants in the retrospective cohort of UIH

Supplementary Table 2

Feature	Coefficient
Smoking years	0.551
Age	0.309
Cigarettes per day	0.283
BMI	-0.229
Divorced or widowed	0.140
Smoking status	0.132
Education: Bachelor's and above	-0.127
Race: Black or African American	0.120
Family history of lung cancer	0.102
Ethnicity	-0.09
Work exposure *	0.06

Supplementary Table 2. 11 Feature SVM model features and weights.

Supplementary Table 3

a.

Year	ROC AUC	PR AUC	Sensitivity	Specificity	PPV	NPV	Proportion	N cases	N total
Year 1	0.914	0.329	0.799	0.91	0.097	0.997	0.012	144	11976
Year 2	0.848	0.257	0.798	0.731	0.058	0.994	0.02	243	11945
Year 3	0.804	0.238	0.801	0.622	0.059	0.991	0.029	342	11821
Year 4	0.786	0.226	0.8	0.619	0.077	0.987	0.038	444	11700
Year 5	0.781	0.232	0.801	0.618	0.095	0.984	0.047	547	11519
Year 6	0.781	0.246	0.8	0.608	0.105	0.981	0.055	595	10898

b.

Year	ROC AUC	PR AUC	Sensitivity	Specificity	PPV	NPV	Proportion	N cases	N total
Year 1	0.797	0.254	0.798	0.438	0.041	0.986	0.029	119	4092
Year 2	0.724	0.244	0.795	0.39	0.075	0.969	0.058	166	2852
Year 3	0.696	0.3	0.798	0.4	0.119	0.951	0.092	198	2152
Year 4	0.686	0.303	0.799	0.418	0.14	0.946	0.106	204	1924
Year 5	0.681	0.377	0.806	0.361	0.176	0.917	0.144	217	1504
Year 6	0.684	0.446	0.804	0.412	0.267	0.888	0.21	230	1096

Supplementary Table 3. Statistics representing the performance of the Hybrid (SVM+ResNet) model when predicting lung cancer in the NLST (a) and UIC (b) test sets.

Supplementary Table 4

Cohort	Model	CT Type	Race	Sex	BMI				
NLST	Sybil	LDCT	All	Both	High				
Year	**ROC AUC**	**PR AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Proportion**	**N cases**	**N total**
Year 1	0.879	0.229	0.786	0.834	0.053	0.997	0.012	42	3579
Year 2	0.815	0.181	0.806	0.631	0.04	0.994	0.019	67	3567
Year 3	0.757	0.173	0.798	0.559	0.049	0.99	0.028	99	3545
Year 4	0.765	0.189	0.805	0.563	0.06	0.988	0.033	118	3532
Year 5	0.775	0.219	0.818	0.556	0.068	0.987	0.038	132	3450
Year 6	0.789	0.253	0.771	0.643	0.089	0.984	0.043	140	3243
Cohort	**Model**	**CT Type**	**Race**	**Sex**	**BMI**				
NLST	Sybil	LDCT	All	Both	Low				
Year	**ROC AUC**	**PR AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Proportion**	**N cases**	**N total**
Year 1	0.937	0.369	0.784	0.921	0.109	0.997	0.012	102	8397
Year 2	0.871	0.288	0.801	0.796	0.078	0.995	0.021	176	8378
Year 3	0.824	0.262	0.815	0.626	0.062	0.991	0.029	243	8276
Year 4	0.778	0.234	0.779	0.594	0.074	0.985	0.04	326	8168
Year 5	0.756	0.226	0.798	0.532	0.085	0.98	0.051	415	8069
Year 6	0.738	0.228	0.822	0.444	0.086	0.975	0.059	455	7655
Cohort	**Model**	**CT Type**	**Race**	**Sex**	**BMI**				
UIC	Sybil	LDCT	All	Both	High				
Year	**ROC AUC**	**PR AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Proportion**	**N cases**	**N total**
Year 1	0.667	0.217	0.778	0.261	0.032	0.974	0.031	54	1766
Year 2	0.589	0.228	0.764	0.121	0.05	0.894	0.057	72	1266
Year 3	0.573	0.303	0.769	0.159	0.073	0.889	0.079	78	986
Year 4	0.563	0.312	0.802	0.122	0.086	0.857	0.093	81	870
Year 5	0.557	0.326	0.812	0.13	0.116	0.83	0.124	85	687
Year 6	0.56	0.368	0.812	0.147	0.168	0.787	0.175	85	486
Cohort	**Model**	**CT Type**	**Race**	**Sex**	**BMI**				
UIC	Sybil	LDCT	All	Both	Low				
Year	**ROC AUC**	**PR AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Proportion**	**N cases**	**N total**
Year 1	0.897	0.305	0.769	0.885	0.162	0.993	0.028	65	2326
Year 2	0.833	0.3	0.851	0.607	0.12	0.985	0.059	94	1586
Year 3	0.807	0.381	0.808	0.623	0.198	0.966	0.103	120	1166
Year 4	0.792	0.388	0.789	0.611	0.211	0.956	0.117	123	1054
Year 5	0.807	0.517	0.803	0.641	0.301	0.944	0.162	132	817
Year 6	0.784	0.561	0.807	0.587	0.379	0.907	0.238	145	610

Supplementary Table 4. Statistics representing the performance of the Hybrid model when predicting lung cancer in the NLST and UIC test sets across BMI categories.