This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

Lung Cancer Prediction using Curriculum Learning based Deep Neural Networks

Jackson Zhou, Shahadat Uddin, Simon K. Poon, University of Sydney, Australia

Mohammad Ali Moni, The University of New South Wales, Sydney, Australia

Matloob Khushi, University of Sydney, Australia; University of Suffolk, UK

***Abstract*—The high incidence and low survival rate of lung and bronchus cancers contribute to their high death count, and necessitate the development of pre-emptive computer-based prediction models using socio-demographic factors. The five year relative survival rate of small cell lung cancer in particular (6%) is almost four times less than that of non small cell lung cancer (23%), though no predictive models have been developed for it so far. This study aimed to expand on the results of previous lung cancer prediction studies and develop innovative models for general and small cell lung cancer prediction. Several machine learning models were implemented including decision trees, random forests, logistic regression classifiers, multilayer perceptron classifiers in addition to a novel curriculum learning based deep neural network. All models were evaluated in this study using data from the National Cancer Institute's Prostate, Lung, Colorectal and Ovarian Cancer screening trial, and performance was measured using the area under the receiver operator characteristic curve (AUROC). Random forest models were found to give the best performances in lung cancer prediction (bootstrap optimism corrected (BOC) AUROC = 0.927), outperforming previous logistic regression based models (BOC AUROC = 0.859). Additionally, curriculum learning based neural networks were shown to outperform all other model types for small cell lung cancer prediction in particular (AUROCs were 0.873 and 0.882 for two different feature sets). To conclude, high-performance models were developed for lung cancer and small cell lung cancer prediction, which could help improve non-invasive lung cancer prediction models in a clinical setting.**

***Index Terms*—Curriculum learning, lung cancer prediction, machine learning, small cell lung cancer prediction.**

I. Introduction

REDUCTIONS in smoking have seen a downturn in lung cancer mortality rates over the past two decades. Even so, lung and bronchus cancers still remain as one of the more deadly forms of cancer, with a five year relative survival rate of only 19%. In the U.S. alone, it has been estimated that 235,760 new cases of lung cancer will have been diagnosed in 2021, resulting in an estimated 131,880 deaths [1].

The high prevalence and low survival rate of lung cancer has necessitated research into its early detection and diagnosis, when it can be more easily treated – it has been shown that pre-emptive screenings for lung cancer using low-dose computed tomography (CT) decrease the lung cancer mortality rate by 20% compared to screenings using standard chest radiographs [2], for which no significant effects have been observed [3]. In addition to physical screenings, there have been consistent efforts in the building of computer-based predictive models for lung cancer incidence. These models allow for the effective implementation of a lung cancer screening system, where individuals at a high risk are able to be identified by a computer and given priority for physical screens.

In 2009, Spitz et al. [4] used data for 1851 lung cancer patients and 2001 matching control subjects to build multivariate logistic models for the prediction of lung cancer in never, former and current smokers. The predictors used included total pack-years smoked, family history of cancer and levels of exposure to dust, fumes, asbestos and pesticides. The respective areas under the receiver operating characteristic curve (AUROC) were found to be 0.57, 0.63 and 0.58, indicating modest discriminatory ability for the models.

The analysis of a more comprehensive dataset by Tammemagi et al. [5] in 2009 yielded increased model performance – data for 77,465 participants originating from the intervention arm of the National Cancer Institute's Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial was used to build logistic regression models for the prediction of lung cancer incidence, given that an individual has already received an abnormal suspicious result from a chest radiograph screening. Important predictors of these true positives were found to include factors such as lower education, a greater number of pack-years smoked, and a family history of lung cancer. The model was found to have high discrimination, with an AUROC of 0.864.

These high performing predictive models for true positives were followed up by Tammemagi et al. [6] in 2011, where data for 70,962 participants originating from the control arm of the PLCO Cancer Screening Trial was used to create more general logistic regression models for overall lung cancer incidence. The predictors used included general demographical information such as age, education, and a family history of cancer, and also included detailed information on the smoking habits of a participant. Both the model for the entire cohort as well as the model for a subset of ever-smokers were found to have high discriminatory ability, with AUROCs of 0.857 and 0.805 respectively, outperforming the previous study by Spitz et al.

Though the models presented in these studies as well as many others [7], [8], [9] offer promising results in the field of lung cancer prediction using socio-demographic factors, only a small number of model types were considered in these studies, whereas a comparison of multiple prediction techniques may have yielding a higher-performing model. In addition, these studies did not consider the prediction of the type of lung cancer an individual might be diagnosed with. In this context, an area of research of particular importance would be the prediction of small cell lung cancer (SCLC) incidence for individuals already identified as being high-risk for lung cancer. This importance stems from the pronounced difference in survival rates between non-small cell lung cancers

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

(NSCLCs) and small cell lung cancers – patients diagnosed with a NSCLC are almost 4 times as likely to survive (23% five year relative survival rate) compared to those diagnosed with small cell lung cancers (6% five year relative survival rate) [1]. With such a large discrepancy, effective prediction of small cell lung cancer incidence would be of great benefit to individuals already identified as being high-risk for lung cancer overall – increased monitoring would be able to be focused on individuals predicted to be at further risk for small cell lung cancer, facilitating early diagnoses and increased survival rates.

Under the contexts of limited model diversity and the absence of exploration into SCLC prediction in the aforementioned lung cancer prediction studies, there were two main aims for this study:

1) Build and identify high-performing models for lung cancer prediction through the consideration of various machine learning methods, and compare the performances of these models to that of the model proposed by Tammemagi et al. in their 2011 study [6]
2) Build and identify high-performing models for small cell lung cancer through the consideration of various machine learning methods

We focused on curriculum learning based neural networks in particular, a novel machine learning method which has seen little application in the context of cancer prediction using socio-demographic factors [15], [16]. The key concept in curriculum learning based neural networks is to train a neural network with instances sorted in order of increasing difficulty, with the motivation being that learning in any context is much more efficient when easier ideas are presented first, followed by harder ideas that build upon the easier ones. The performance of a curriculum learning based neural network was compared to that of standard machine learning methods, in order to determine which method provided the best performance. The standard machine learning methods considered in this study included (1) decision trees, (2) random forests, (3) logistic regression and (4) standard neural networks (specifically, multilayer perceptron classifiers).

As with Tammemagi et al., the models in this study were built and evaluated using participant data from the National Cancer Institute's PLCO Cancer Screening Trial, a clinical trial which aimed to assess whether screening exams for PLCO cancers reduce mortalities from these cancers.

II. Methods

A. Study Population

All models in this study were built and evaluated using data from the National Cancer Institute's PLCO Cancer Screening Trial. The PLCO trial was a randomised and controlled prospective trial based in the U.S., consisting of screening tests for prostate, lung, colorectal and ovarian cancers for participants in its intervention arm. The PLCO trial aimed to assess the effectiveness of pre-emptive screenings in reducing mortality rates for PLCO cancers through comparing mortalities of participants in this intervention arm to that of a control arm (where participants were not offered regular screenings for PLCO cancers).

Recruitment of possible participants commenced across 9 U.S. centres in November 1993, with 154,897 applications by the end of enrolment in July 2001. During the initial recruitment period, an applicant was required to be 60-74 years old to be eligible for the study, with this requirement being relaxed to 55-74 years old starting from 1996. An applicant was considered ineligible for the study if they had experienced a history of PLCO cancers, or if they were currently receiving treatment for any cancer, excluding basal or squamous cell skin cancer. Further details regarding PLCO trial eligibility are provided online by the National Cancer Institute [10].

Eligible participants were randomised in equal proportions to the control and intervention arms. Regarding lung cancer screening specifically, participants in the intervention arm of the PLCO trial were invited to receive up to four annual posteroanterior chest X-rays to screen for possible lung cancer, while participants in the control arm received usual care.

To ensure comparability between this study and the 2011 study by Tammemagi et al., only data for participants in the control arm ($N = 77453$) was used to build the lung cancer prediction models. However, data for participants in both the control and intervention arms ($N = 154,897$) were used to build the SCLC prediction models, so as to utilise all participant data – as there have been no studies focusing on the prediction of SCLC incidence as of yet, comparability was not an issue.

B. Implementation Details

Feature selection, preprocessing and model building were conducted in Python version 3.7.1, using a combination of the *pandas* (data manipulation), *matplotlib* (plotting) and *scikit-learn* (machine learning) libraries. Additional plotting was also performed in R version 3.6.1, using the *ggplot2* package.

C. Feature Selection

The comprehensiveness of the PLCO trials provided a selection of 205 potential predictor variables. The main predictors of interest were participant responses to the baseline questionnaire, which was administered during a participant's randomisation into either the control or intervention arm.

For the sake of comparability, predictor variables for the overall lung cancer predictive models were matched to those from the 2011 study conducted by Tammemagi et al. [6]. There were 12 predictors in total: the numeric predictors consisted of age at trial entry, total number of years spent smoking and total number of pack-years smoked. The categorical predictors consisted of gender, ethnicity, socio-economic status (as measured by education level), current smoking status, BMI range, and histories relating to lung cancer in the family, bronchitis, emphysema and X-ray screenings.

In the small cell lung cancer predictive models, two sets of features were chosen. The first set, $\mathcal{R}$, was obtained through sorting features by their importance weights in a random forest model, and selecting the 12 best predictors (that is, those with the highest importance weights). The second set of

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

predictors, $\mathcal{G}$, was similarly obtained through sorting features by their importance weights in a gradient boosting machine, and selecting the 12 best predictors. Note that 12 was chosen specifically to match the number of predictors in the overall lung cancer models.

D. Data Preprocessing

Data imbalance was an issue in the PLCO trial data, where the number of participants diagnosed with lung cancer during the trial was much lower than the number of participants who did not receive a lung cancer diagnosis. This was handled differently between the lung cancer and small cell lung cancer models:

- The Synthetic Minority Oversampling Technique (SMOTE) was used to rebalance the data for the small cell lung cancer prediction models, to potentially improve model performance
- SMOTE was not used to rebalance the data for the lung cancer prediction models, to ensure consistency between the 2011 study by Tammemagi et al. [6]

For both the general lung cancer and small cell lung cancer predictive models, numerical data such as age and total years smoked were scaled when training the neural networks, in order to improve the performance of these models; values in numerical columns were scaled to zero mean and unit variance.

Basic filtering was also performed on the control arm data, where participants who either did not return their baseline questionnaire, had a history of lung cancer at trial entry, or had associated missing values for any predictor variables were excluded in the analysis. These filter conditions are based on the filters Tammemagi et al. used in their 2011 study [6].

E. Model Building

Our implementation of a curriculum learning based neural network was based on Hacohen and Weinshall's [11] design in their 2019 curriculum learning study, where they aimed to classify images in the CIFAR-10 and CIFAR-100 datasets. The training process of the curriculum classifier can be divided into the following steps:

1) A multilayer perceptron (a standard neural network) is trained on the input data, with this network then being used to generate probability estimates for lung cancer incidence in the input data.
2) The difficulty of all rows in the input data is computed – the difficulty D_i of a single row/instance i refers to the difficulty in classifying it into a particular class, and is estimated to be
$$D_i = 0.5 - |p_i - 0.5| \tag{1}$$
where p_i is the probability of lung cancer incidence for instance i.
3) The instances in the input data are sorted in increasing order of difficulty.
4) For i in $1, 2, \cdots, B$ (where B is the specified number of curriculum batches), compute $g(i)$, where $g : \mathbb{N}^+ \to [N]$ is the pacing function, as described in Hacohen and Weinshall's study [11] (note that N is the total number of instances in the input data). This study utilised a fixed exponential pacing function, of the form:
$$g(i) = \min\left(\mathrm{SP} \cdot \mathrm{I}^{\lfloor i/\mathrm{SL} \rfloor}, 1\right) \cdot N \tag{2}$$
where $\mathrm{SP} \in (0, 1)$ is a predetermined starting percentage, $\mathrm{I} > 1$ is a predetermined increase amount per step, and $\mathrm{SL} \in \mathbb{N}^+$ is the predetermined step length.
5) Each curriculum batch number is associated with a subset of the input data, such that the difficulty (as computed in step 2) of the instances in each curriculum batch increases as the curriculum batch number increases. In particular, for each curriculum batch number $1, 2, \cdots, B$, a sample of size s_b (where s_b is the predefined curriculum batch size) is taken without replacement from the first $g(i)$ instances of the sorted input data from step 3. This sample represents the curriculum batch.
6) A multilayer perception is trained on these curriculum batches in order of increasing difficulty. That is, curriculum batch number 1 is used as the first mini-batch in the neural network's gradient descent algorithm, curriculum batch number 2 is used as the second mini-batch, and so on until all curriculum batches have been exhausted. This trained neural network represents the trained curriculum classifier.

After the training process, future predictions for a trained curriculum classifier are reduced to the predictions of the trained neural network from step 6.

In both the lung cancer and small cell lung cancer models, the model types considered were: (1) a decision tree classifier, (2) a random forest classifier, (3) a logistic regression classifier, (4) a standard neural network classifier (based on a multilayer perceptron), (5) a soft voting ensemble classifier consisting of the random forest, logistic regression and standard neural network models (R-L-N ensemble), and (6) a curriculum classifier, as described previously.

As seen in the 2011 study by Tammemagi et al. [6], lung cancer models in this study were trained and evaluated on the entire study population, where a bootstrap method with 50 resamplings was used to correct for any optimism induced by having identical training and evaluation sets. In contrast, cross validation (4 folds) was used for the evaluation of the small cell lung cancer models. The evaluation metric used for all models was the AUROC score.

Hyper-parameter optimisation with respect to AUROC was performed on all models excluding the voting ensemble classifier, in order to improve model performance.

III. Results

A. Lung Cancer Models

The control arm of the PLCO data was initially filtered for participants who returned their baseline questionnaire, had no previous history of lung cancer, and had no associated missing values. Models were then trained using this filtered dataset (N = 68, 147). Of the participants in the filtered dataset, 1487 (2.2%) were diagnosed with lung cancer during the trial, while 66,670 (97.8%) remained free of lung cancer over the

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

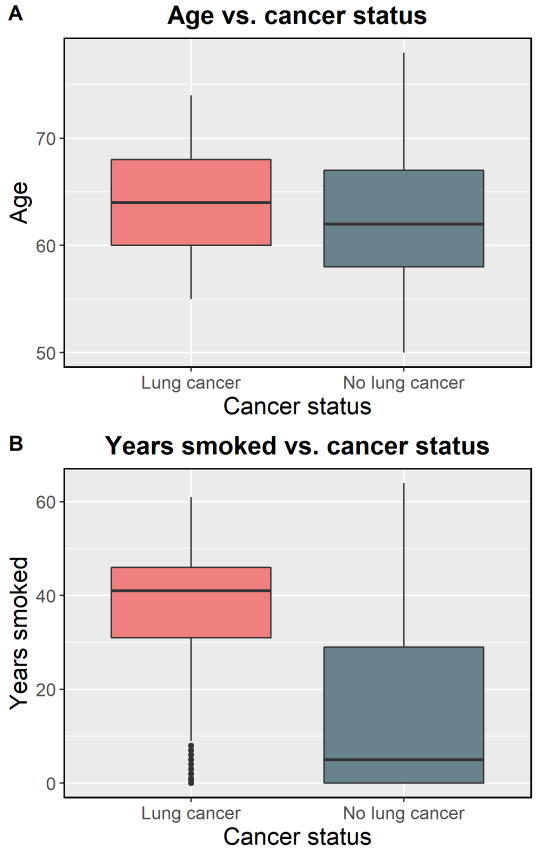

Fig. 1. Box plots indicating the distributions of age and smoking duration in the control arm of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

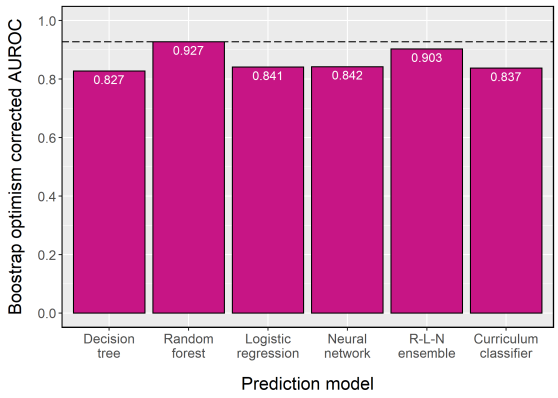

Fig. 2. Bar plot of bootstrap optimism corrected AUROCs for various model types in predicting lung cancer in the control arm of the PLCO Cancer Screening Trial.

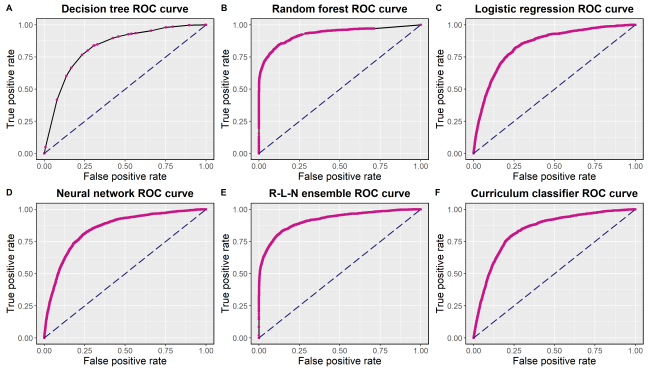

Fig. 3. ROC curves (with no adjustment for optimism) for various model types in predicting lung cancer in the control arm of the PLCO Cancer Screening Trial.

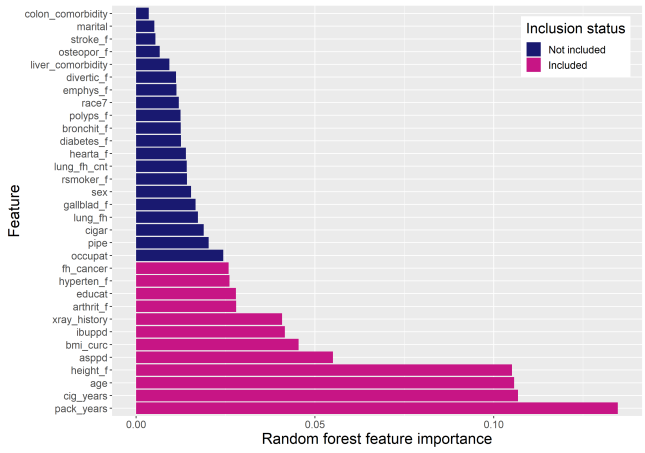

Fig. 4. Bar plot of feature importances for small cell lung cancer prediction in the PLCO Cancer Screening Trial, using a random forest model.

trial period. The mean age of participants diagnosed with lung cancer was 64.3 years, while corresponding mean age for those who were not was 62.6 years (Figure 1). As was expected, the mean for both number of years smoked and number of pack-years smoked was higher for participants diagnosed with lung cancer compared to those who were not. The cancer status of a patient was found to have a noticeable relationship between all categorical variables, with the exception of ethnicity.

Strong relationships between the predictors and lung cancer incidence allowed for favourable performance in all model types, with each model type achieving an AUROC of at least 0.8 (Figure 2). The random forest model offered the best performance, with a bootstrap optimism corrected AUROC of 0.927, in contrast to the decision tree, which gave the lowest performance with a corrected AUROC of 0.827. Interestingly, the performance of the R-L-N ensemble, which included the random forest, was lower than the random forest's individual performance. Additionally, the standard neural network outperformed the curriculum learning based neural network in this instance (Figure 2). Receiver operator characteristic (ROC) curves for each model type are presented in Figure 3.

B. Small Cell Lung Cancer Models

1) Models Based on $\mathcal{R}$: Predictors were sorted with respect to their importance weights in a random forest model, and the 12 features with the highest weights were selected for inclusion in $\mathcal{R}$ (see Figure 4).

In total, 4 numeric predictor variables and 8 categorical predictor variables were selected. The numerical predictors were age at trial entry, total number of years spent smoking, total number of pack-years smoked and height in inches at trial entry. The categorical predictors were socio-economic

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

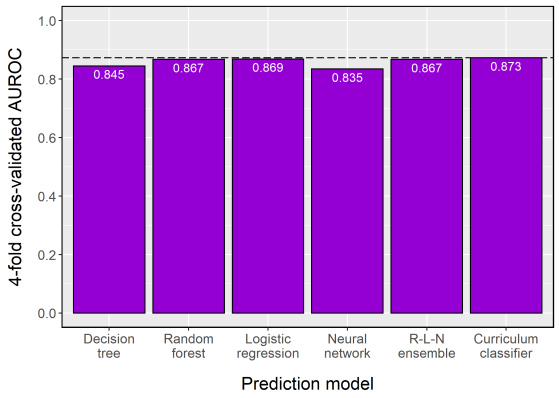

Fig. 5. Bar plot of 4-fold cross-validated AUROCs for various model types in predicting small cell lung cancer in the PLCO Cancer Screening Trial ($\mathcal{R}$ used as predictors).

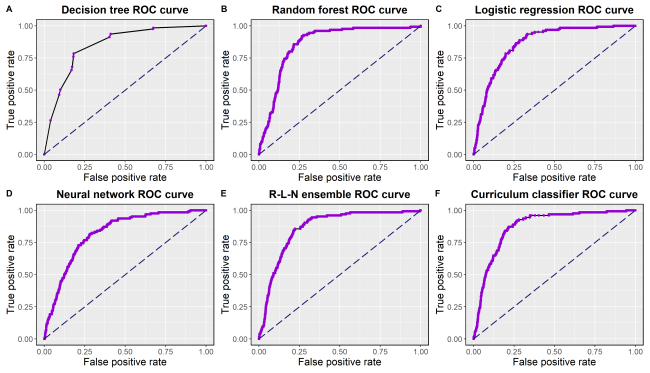

Fig. 6. ROC curves (using holdout evaluation with a testing proportion of 0.25) for various model types in predicting small cell lung cancer in the PLCO Cancer Screening Trial ($\mathcal{R}$ used as predictors).

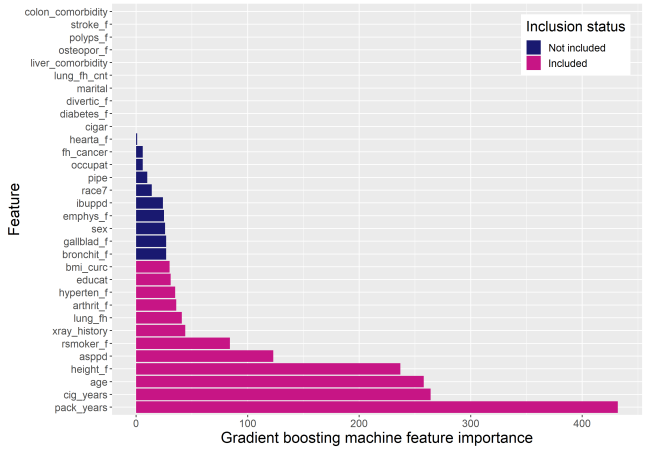

Fig. 7. Bar plot of feature importances for small cell lung cancer prediction in the PLCO Cancer Screening Trial, using a gradient boosting machine.

status (as measured by education), BMI range at trial entry, histories for arthritis, hypertension, cancer in the family and X-ray screenings, in addition to intake frequency of aspirin and ibuprofen in the year prior to trial entry. Filtering for participants who had associated missing values for any of these predictors yielded N = 137,687 participants for the model building. Of these participants, 431 were diagnosed with SCLC during the trial and 137,256 were either not diagnosed with lung cancer, or were diagnosed with NSCLC during the trial. The mean age of participants diagnosed with SCLC was 64.1 years, while the mean for those who were not was 62.6 years. Furthermore, the mean for both number of years smoked and number of pack-years smoked was higher for participants diagnosed with small cancer compared to those who were not. Noticeable differences in distribution was seen across participants diagnosed with small cell and participants who were not for socio-economic status, arthritis and X-ray screening histories, and aspirin and ibuprofen intake frequencies.

As with the overall lung cancer models, favourable performance was seen in all model types, with each model type achieving an AUROC of at least 0.8 (Figure 5). The curriculum learning based neural network achieved the highest performance, with a 4-fold cross validated AUROC of 0.873, in contrast to the standard neural network model, which gave the lowest performance with a 4-fold cross validated AUROC of 0.815. The R-L-N ensemble was seen to have a similar performance compared to a standalone random forest (Figure 5). ROC curves for each mode type are presented in Figure 6.

2) Models Based on $\mathcal{G}$: Predictors were sorted with respect to their importance weights in a gradient boosting machine, and the 12 features with the highest weights were selected for inclusion in $\mathcal{G}$ (see Figure 7).

Similarly to predictor set $\mathcal{R}$, 4 numeric predictor variables and 8 categorical predictor variables were selected. The numerical predictors were age at trial entry, total number of years spent smoking, total number of pack-years smoked and height in inches at trial entry. The categorical predictors were socio-economic status (as measured by education), regular smoking status, BMI range at trial entry, histories for arthritis, hypertension, lung cancer in the family and X-ray screenings, in addition to intake frequency of aspirin in the year prior to trial entry. The majority of predictors in $\mathcal{G}$ were identical to those in $\mathcal{R}$. The differences were that $\mathcal{G}$ used lung cancer family history specifically in place of general cancer family history, and used regular smoking status instead of ibuprofen intake frequency. Filtering for participants who had associated missing values for any of these predictors yielded N = 137,764 participants for the model building. Of these participants, 434 were diagnosed with SCLC during the trial and 137,330 were either not diagnosed with lung cancer, or were diagnosed with NSCLC during the trial. As with the analysis using $\mathcal{R}$, the mean age of participants diagnosed with SCLC was 64.1 years, while the mean for those who were not was 62.6 years. Furthermore, the mean for both number of years smoked and number of pack-years smoked was higher for participants diagnosed with small cancer compared to those who were not. Noticeable differences in distribution was seen across participants diagnosed with small cell and participants who were not for socio-economic status, arthritis history, lung cancer family history, regular smoking history, X-ray screening history, and aspirin intake frequency.

As with the previous models, favourable performance was

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICDH52753.2021.00013, 2021 IEEE International Conference on Digital Health, ICDH 2021

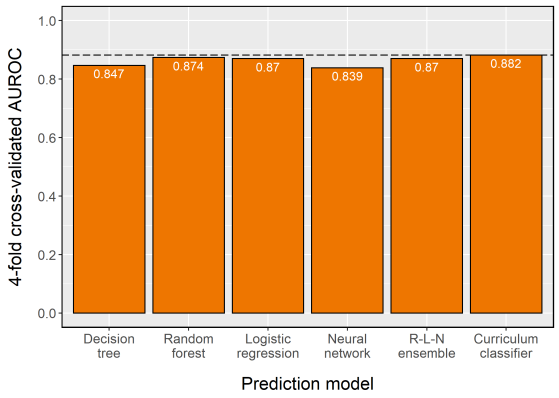

Fig. 8. Bar plot of 4-fold cross-validated AUROCs for various model types in predicting small cell lung cancer in the PLCO Cancer Screening Trial ($\mathcal{G}$ used as predictors).

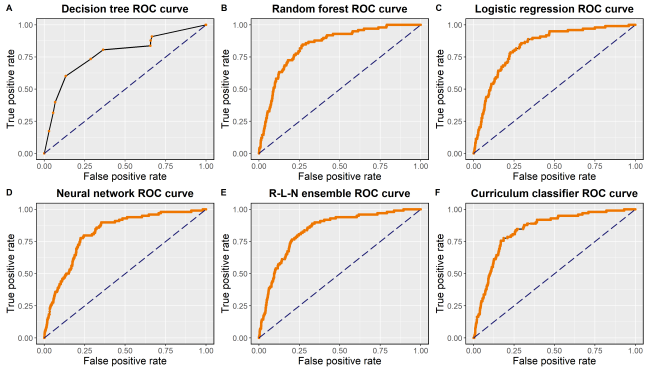

Fig. 9. ROC curves (using holdout evaluation with a testing proportion of 0.25) for various model types in predicting small cell lung cancer in the PLCO Cancer Screening Trial ($\mathcal{G}$ used as predictors).

seen in all model types, with each model type achieving an AUROC of at least 0.8 (Figure 8). Slightly higher model performances were seen in $\mathcal{G}$ compared to $\mathcal{R}$. The curriculum learning based neural network achieved the highest performance, with a 4-fold cross validated AUROC of 0.882, in contrast to the standard neural network model, which gave the lowest performance with a 4-fold cross validated AUROC of 0.839. As with the overall lung cancer models, the R-L-N ensemble was seen to have lower performance compared to a standalone random forest (Figure 8). ROC curves for each mode type are presented in Figure 9.

IV. Discussion

The first aim for this study was to build and identify high-performing predictive models for lung cancer incidence, through the consideration of a range of machine learning model types. A number of previous studies have focused on utilising logistic regression based models for prediction of lung cancer, to varying degrees of success [4], [6], [5]. In this study, the logistic regression model for lung cancer incidence was seen to give high predictive performance, with a bootstrap optimism corrected AUROC of 0.841. This was comparable to the performance of the model presented by Tammemagi et al. in their 2011 study of the PLCO cohort, where an AUROC of 0.859 was established. In contrast, the random forest based model was found to have a notably higher performance – a bootstrap optimism corrected AUROC of 0.927 was attained, outperforming all other model types. The increased performance of the random forest based model may be explained by the complexity of the data – if the relationship between lung cancer incidence and the predictor variables used does not obey a linear equation, then a random forest, a non-linear model, will have better performance compared to logistic regression, a linear model. Equivalently, the decision boundaries for lung cancer incidence may be too complex to be modelled by logistic regression, whereas a random forest has a more flexible decision boundary, allowing for increased performance. The general applicability of random forests over logistic regression has been shown by Couronné, Probst and Boulesteix in their 2018 study [12], where random forests were found to outperform logistic regression in terms of accuracy in 69% of the 243 datasets tested. While the results of this study seem to agree with those by Couronné, Probst and Boulesteix, it should be noted that different metrics were used in the evaluation of models between the studies. In any case, this study was able to identify high-performance models for lung cancer incidence, outperforming the previous logistic regression models using the same data presented by Tammemagi et al. in their 2011 study [6]. Furthermore, the viability of random forests over logistic regression in the context of lung cancer prediction is demonstrated.

The second aim for this study was to build and identify high-performing models for SCLC incidence, through the consideration of the same machine learning model types as before. While artificial neural network ensembles have been used in the classification of lung cancer type from image data [13], there has been limited research into the prediction of SCLC using demographical information. For both sets of predictors, all model types were seen to give favourable performance, with the curriculum learning based neural network showing the highest performance across both $\mathcal{R}$ and $\mathcal{G}$. In both cases, the curriculum learning based neural network was seen to outperform the standard neural network by a large margin. Explanations for an increase in performance are provided by Hacohen and Weinshall in their 2019 study [11], where it was shown theoretically that the use of a curriculum changes the optimisation landscape for a neural network during its training process, allowing for easier convergence to the optimal set of neural network parameters. Empirical evidence for the benefits of using a curriculum were shown in the same study, where curriculum learning based neural networks were seen to have higher accuracy compared to 'vanilla' neural networks in the classification of image data from the Canadian Institute for Advanced Research (the CIFAR datasets). Similar results have been found by Bengio et al. in 2009 [14] in the classification of geometrical shapes, where average error rates for curriculum based neural networks were found to be much lower compared to standard neural networks.

It should be noted that while the curriculum classifier performed well in the prediction of SCLC, performance was sub-par relative to other model types (in particular the standard neural network) in the prediction of lung cancer overall. A

possible explanation would be that a curriculum is limited in effectiveness to large datasets for lung cancer prediction. In particular, the SCLC models were built on much more instances compared to the overall lung cancer models. Additionally, the hyper-parameters chosen for the curriculum classifier may have only been effective in the prediction of SCLC.

Interestingly, model performances for SCLC prediction were worse compared to those for overall lung cancer, even though SCLC incidences represented a subset of lung cancer incidences. This may be explained by the smaller amount of training instances for SCLC compared to lung cancer, reducing the effectiveness of models built. Moreover, there could be increased variation in predictor variables in SCLC patients compared to lung cancer patients overall, resulting in reduced performance when trying to predict specifically for SCLC.

Additionally, although high-performing predictive models for SCLC were built, it is possible that these models only used indicators of lung cancer to predict SCLC incidences. However, the increased performance of these models compared to the overall lung cancer models from Tammemagi et al. in their 2011 study [6] suggests that the models are able to predict for SCLC specifically. A reasonable alternative approach might be to restrict SCLC prediction to participants who were diagnosed with lung cancer specifically, although this needs to be performed on a much larger dataset, where the reduction in the size of the study population still leaves a substantial number of instances for the models to be trained on. In summary, this study was also able to identify potentially high-performing models for SCLC incidence, demonstrating the utility of curriculum based neural networks as models when using sufficiently large datasets.

This study shared many limitations with those of Tammemagi et al. in their 2011 study [6], due to the similar datasets used. Although high-performing models were built for lung cancer and SCLC incidence, external generalisability for these models to the overall population may be limited by the demographics of the PLCO participants – it should be noted that the PLCO trials only accepted participants between 55 and 75 years of age. Furthermore, it is likely that PLCO trial participants were of higher socio-economic status and were more health-conscious compared to the overall population. These differences can decrease the performance of the models built in this study, when used in a more general context. Furthermore, as with Tammemagi et al., internal validation was used over external validation, which may have resulted in inflated model performances. It is recommended that future studies carry out the same analysis conducted in this study using an independent dataset, so as to verify the results of this study. Additionally, although it was shown that curriculum learning outperformed the remaining machine learning techniques for SCLC prediction, this result is only true for the range of hyperparameters considered for each model type. A broader search of hyperparameters for each model is needed to verify these results more generally.

V. Conclusion

Our study aimed to use a variety of machine learning techniques to (1) build and identify high-performing models for lung cancer prediction and (2) build and identify high-performing models for SCLC prediction. These aims were accomplished, with high-performance models for both lung cancer and SCLC incidence being built using the PLCO dataset. Lung cancer models in this study were shown to outperform previous models based on the PLCO dataset, and it was shown that curriculum learning was a favourable alternative to regular machine learning methods in the case of SCLC prediction. It is recommended that future studies generalise these results through using a comparable external dataset, and by considering a broader range of hyperparameters for each model during the training process.

Acknowledgment

The authors would like to thank the National Cancer Institute for access to the data collected over the PLCO cancer screening trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by the National Cancer Institute.

References

[1] American Cancer Society, "Cancer Facts & Figures 2021", American Cancer Society, Atlanta, 2021.

[2] The National Lung Screening Trial Research Team, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening", *New England Journal of Medicine*, vol. 365, no. 5, pp. 395-409, 2011.

[3] M. Oken et al., "Screening by Chest Radiograph and Lung Cancer Mortality", *JAMA*, vol. 306, no. 17, pp. 1865-1873, 2011.

[4] M. Spitz et al., "A Risk Model for Prediction of Lung Cancer", *JNCI Journal of the National Cancer Institute*, vol. 99, no. 9, pp. 715-726, 2007.

[5] M. Tammemagi et al., "Prediction of True Positive Lung Cancers in Individuals with Abnormal Suspicious Chest Radiographs—A Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Study", *Journal of Thoracic Oncology*, vol. 4, no. 6, pp. 710-721, 2009.

[6] C. Tammemagi et al., "Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation", *JNCI: Journal of the National Cancer Institute*, vol. 103, no. 13, pp. 1058-1068, 2011.

[7] A. Cassidy et al., "The LLP risk model: an individual risk prediction model for lung cancer", *British Journal of Cancer*, vol. 98, no. 2, pp. 270-276, 2007.

[8] C. Hoggart et al., "A Risk Model for Lung Cancer Incidence", *Cancer Prevention Research*, vol. 5, no. 6, pp. 834-846, 2012.

[9] S. Park et al., "Individualized Risk Prediction Model for Lung Cancer in Korean Men", *PLoS ONE*, vol. 8, no. 2, pp. 1-9, 2013.

[10] "Trial Summary", *National Cancer Institute Cancer Data Access System*, 2019. [Online]. Available: https://cdas.cancer.gov/learn/plco/trial-summary/. [Accessed: 07- Oct- 2019].

[11] G. Hacohen and D. Weinshall, "On The Power of Curriculum Learning in Training Deep Networks", in *International Conference on Machine Learning*, 2019.

[12] R. Couronné, P. Probst and A. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment", *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[13] Z. Zhou, Y. Jiang, Y. Yang and S. Chen, "Lung cancer cell identification based on artificial neural network ensembles", *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25-36, 2002.

[14] Y. Bengio, J. Louradour, R. Collobert and J. Weston, "Curriculum learning", in *International Conference on Machine Learning*, 2009, pp. 41-48.

[15] Li, Huiyu, Xiabi Liu, Said Boumaraf, Weihua Liu, Xiaopeng Gong, and Xiaohong Ma. "A New Three-stage Curriculum Learning Approach for Deep Network Based Liver Tumor Segmentation." In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-6. IEEE, 2020.

[16] Wei, Jerry, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown et al. "Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2473-2483. 2021.