Physics in Medicine & Biology

Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy

To cite this article before publication: Rongfang Wang *et al* 2019 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ab555e

Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy

Rongfang Wang[1,2], Yaochung Weng[1], Zhiguo Zhou[1], Liyuan Chen[1], Hongxia Hao[3], and Jing Wang[1,4]

1 Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75235, United States of America

2 School of Artificial Intelligence, Xidian University, Xi’an 710071, People’s Republic of China

3 School of Computer Science and Technology, Xidian University, Xi’an 710071, People’s Republic of China

4 Author to whom any correspondence should be addressed.

E-mail: Jing.Wang@utsouthwestern.edu

Abstract

Accurately predicting treatment outcome is crucial for creating personalized treatment plans and follow-up schedules. Electronic health records (EHRs) contain valuable patient-specific information that can be leveraged to improve outcome prediction. We propose a reliable multi-objective ensemble deep learning (MoEDL) method that uses features extracted from EHRs to predict high risk of treatment failure after radiotherapy in patients with lung cancer. The dataset used in this study contains EHRs of 814 patients who had not achieved disease-free status and 193 patients who were disease-free with at least one year follow-up time after lung cancer radiation therapy. The proposed MoEDL consists of three phases: 1) training with dynamic ensemble deep learning; 2) model selection with adaptive multi-objective optimization; and 3) testing with evidential reasoning (ER) fusion. Specifically, in the training phase, we employ deep perceptron networks as base learners to handle various issues with EHR data. The architecture and key hyper-parameters of each base learner are dynamically adjusted to increase the diversity of learners while reducing the time spent tuning hyper-parameters. Furthermore, we integrate the Snapshot Ensembles (SE) restarting strategy, multi-objective optimization, and ER fusion to improve the prediction robustness and accuracy of individual networks. The SE restarting strategy can yield multiple candidate models at no additional training cost in the training stage. The multi-objective model simultaneously considers sensitivity, specificity, and AUC as objective functions, overcoming the limitations of single-objective–based model selection. For the testing stage, we utilized an analytic ER rule to fuse the output scores from each optimal model to obtain reliable and robust predictive results. Our experimental results demonstrate that MoEDL can perform better than other conventional methods.

Keywords: Electronic health records; Lung cancer radiotherapy; Multi-objective optimization; Ensemble deep learning

1. Introduction

Lung cancer has become the leading cause of cancer-related mortality around the world. The International Agency for Research on Cancer, which is the World Health Organization's cancer research agency, reported 9.6 million deaths from cancer in 2018; lung cancer accounted for 1.76 million of these deaths, much more than any other kind of cancer [1]. Currently, radiation therapy is one of the primary definitive treatment modalities for patients with lung cancer. However, the survival rate after radiation therapy is still low, especially for patients with advanced stage lung cancer. Accurately predicting treatment outcomes for individual patients could enable personalized treatment plans and follow-up schedules, ultimately leading to better clinical outcomes. For example, physicians could design treatment plans differently and prescribe intensified treatments for a particular subset of patients at higher risk for treatment failure. Similarly, patients identified as very likely to fail could be followed up more closely with repeat chest or body imaging. Therefore, stratifying risk to determine which patients will achieve remission after radiotherapy and which patients will not is essential to achieving better treatment outcomes for lung cancer.

Electronic health records (EHR) systems contain valuable patient-specific information, such as demographics, diagnoses, laboratory test results, prescribed or administered medications, clinical notes, and other HIPAA-protected patient data [2]. Currently, EHR data are used not only to improve clinical efficiency, but also to assist in clinical decision making. Using EHR data to predict patient outcomes has been widely investigated and has achieved promising performance in various applications [3], such as predicting distant failure in lung cancer [4, 5], predicting response to analgesics [6], stratifying suicide risk [7], identifying osteoporosis [8], predicting clinical events [9], classifying diagnoses [10], and predicting unplanned readmission [11]. However, there are various challenges to using EHR data because of its heterogeneity, sparseness, and random errors, and the inconsistency of codes between institutions [10]. Deep learning, as an efficient tool for data analysis, can potentially overcome these limitations via automatic data-oriented feature extraction [12]. In various clinical applications, such as information extraction, representation learning, outcome prediction, and phenotyping, several deep learning methods have been developed and have demonstrated the ability to handle the unique characteristics of EHR data [13-16].

Deep learning approaches have a large number of hyper-parameters that must be tuned, and it would be a time-consuming process to do this manually. Recently, automated machine learning (AutoML) has been proposed as a new subcategory in machine learning; this technique aims to reduce the efforts put toward tuning hyper-parameters. One of AutoML's characteristics is that it constructs machine learning programs without human assistance but within limited computational budgets [17]. In recent years, AutoML has already been successfully applied in automated model selection [18, 19], neural architecture search [20, 21], and automated feature engineering [22, 23]. In addition, ensemble deep learning, which trains multiple base learners to solve the same problem, has proven to be more robust and accurate than individual networks [24]. However, traditional ensemble methods have two disadvantages: 1) training multiple deep networks is computationally expensive; and 2) multiple learners are typically constructed based on a single objective function, such as overall accuracy or area under the curve (AUC). When data are imbalanced, a single objective function may not be a good measure. To overcome these limitations, Huang *et al*. [25] proposed a Snapshot Ensembling (SE) technique which can decrease the training time for the entire ensemble to the time required to train a single traditional model by saving several local minima model when training a single neural network. Zhou *et al*. [5] proposed a multi-objective model based on the iterative multi-objective immune algorithm that simultaneously considers sensitivity and specificity as objective functions.

Motivated by AutoML [17] and EHR-driven deep learning approaches, we propose a reliable and dynamic multi-objective ensemble deep learning (MoEDL) method that exploits robust features from EHR data to overcome the aforementioned challenges to stratify patients at high risk of treatment failure after radiotherapy for lung cancer. In the proposed method, we employ deep perceptron networks as base learners to handle various issues with EHR data. The architecture and key hyper-parameters of each base learner are dynamically adjusted using the AutoML approach to increase the diversity of learners and reduce the time spent tuning hyper-parameters. Furthermore, we integrate SE restarting strategy, multi-objective optimization, and evidential reasoning (ER) fusion to improve the prediction robustness and accuracy of individual networks. The SE restarting strategy [25] can yield multiple candidate models at no additional training cost in the training stage. The multi-objective model [5] simultaneously considers sensitivity, specificity, and AUC as objective functions, overcoming the limitation of single-objective–based models. For the testing stage, we utilized an analytic ER rule [26] to fuse the output scores from individual optimal models to obtain reliable and robust predictive results. In this study, we compare our proposed method with a traditional deep perceptron network (DNN) [27], support vector machines (SVM) [28], and an improved group-based multi-objective model (GMO) [5]. Moreover, we analyze the importance of individual features by evaluating the AUC changes in MoEDL predictions based on the feature analysis method used in [29].

2. Materials and Methods

2.1 Dataset and preprocessing

The EHR data used in this study were extracted from a cancer patient registry at the Harold C. Simmons Comprehensive Cancer Center at the University of Texas Southwestern Medical Center. The dataset includes 1007 patients with lung cancer who received radiation therapy between 2004 and 2018 (male: 559 and female: 448; mean age: 65.74 ± 10.95 years; range: 32–92 years). Among these patients, 814 patients had not achieved disease-free status and 193 patients were disease-free (with at least 1 year follow up) after radiation therapy. "Disease-free" means that the patient had no evidence of disease, no disease recurrence, and no progression with at least one year after lung cancer radiotherapy. We felt that based on the available data, a threshold of 1-year for disease-free status was reasonable by excluding those who had short follow-up or survival time after radiotherapy. The outcome was represented in binary: "0" denoted disease-free, and "1" denoted that disease-free status had not been achieved. In total, 20 features per patient were extracted to develop the prediction model. These features include but are not limited to demographic parameters, tumor characteristics, and treatment schemes. Each feature vector contained numerical, ordinal, and nominal attributes, which are explained in Appendix A.1-3. Input normalization is shown to facilitate algorithm convergence [30]. In this work, each feature $\boldsymbol{x}$ is divided by the corresponding maximum of the training dataset $\boldsymbol{x}_{\text{train}}$ to make features on a similar scale, which can be expressed as:

$$\boldsymbol{x}' = \frac{\boldsymbol{x}}{\max(\boldsymbol{x}_{\text{train}})}. \quad (1)$$

For example, if in the training dataset, the range of feature "Age" is [32, 92], then the age is divided by 92. We applied this normalization strategy for both training and testing samples. We augmented the data using the Synthetic Minority Over-sampling Technique (SMOTE) [31] to balance the training samples.

2.2 MoEDL model development

The framework of MoEDL is illustrated in Figure 1. MoEDL consists of three phases: 1) training with dynamic ensemble deep learning; 2) model selection with adaptive multi-objective optimization; and 3) testing with evidential reasoning (ER) fusion. The details of each stage are described in the following.

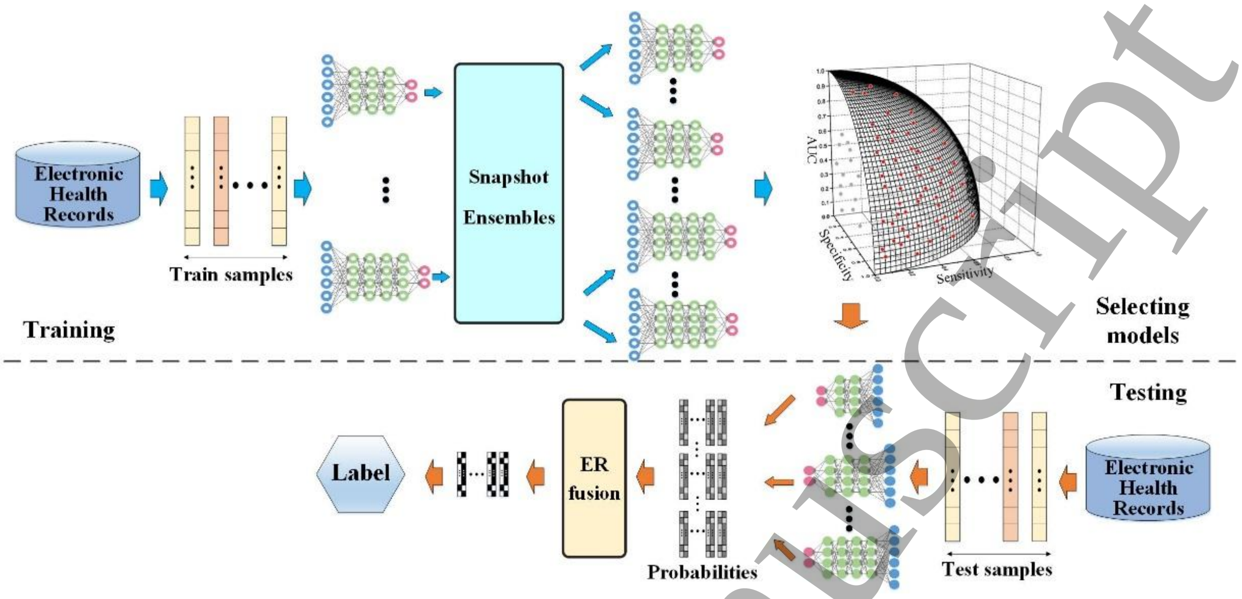


Figure 1: Illustration of the proposed MoEDL framework.

A. Training with dynamic ensemble deep learning

In the training stage, we adopt the multilayer perceptron (MLP) network as the base learner in a parallel style. The rectified linear units (ReLU) and softmax functions are employed as non-linear activation functions in multiple hidden layers and output layers, respectively. In MoEDL, a set of models are constructed for the ensemble learning to obtain final predictive results. Motivated by the AutoML [17], we propose a dynamic ensemble deep learning strategy to generate multiple base learners. The architecture and key hyper-parameters of each base learner are dynamically adjusted based on the given configuration matrix **B**. As the performance of MLP is usually determined by the number of hidden layers, the number of nodes in each hidden layer, the learning rate, and the loss function, we define the configuration matrix $\mathbf{B} \in R^{N \times 4}$ as

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \beta_{1,4} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} & \beta_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{N,1} & \beta_{N,2} & \beta_{N,3} & \beta_{N,4} \end{bmatrix} \tag{2}$$

where N is the number of base learners, $\boldsymbol{\beta}_{i,1}$ denotes the number of hidden layers, $\boldsymbol{\beta}_{i,2}$ denotes the number of nodes in each hidden layer, $\boldsymbol{\beta}_{i,3}$ denotes the learning rate, and $\boldsymbol{\beta}_{i,4}$ denotes the loss function. The parameters $\beta_{i,1}, \beta_{i,2}, \beta_{i,3}$ and $\beta_{i,4}$ are randomly generated in their corresponding search spaces. The four hyper-parameters mentioned above were randomly combined from their corresponding search spaces and N base learners are obtained that are corresponding to N different combinations of the model parameters. This dynamic ensemble strategy not only increases the diversity of base learners, but also reduces the time spent on tuning hyper-parameters.

Furthermore, we exploit a Snapshot Ensembles (SE) restarting strategy [25] to reduce the training cost while increasing the diversity of candidate models. As discussed in the SE method, multiple local minima lie along the optimization path of a model, an optimization process will visit several local minima before converging to a final solution. By taking snapshots of weights and biases at these various local minima and ensemble the predictions of these models at the test time, this approach is termed as Snapshot Ensembles. There are four main steps of the SE training stage: 1) the total epochs in the whole training process are divided into S cycles; 2) in each cycle, the network is restarted with a large initial learning rate α_0, because the large learning rate provides the model enough energy to escape from a critical point; 3) a cyclic cosine annealing schedule [32] was adopt to lower the learning rate at a very fast pace to encourage the model to converge towards the multiple local minima after as few as epochs; 4) this process was repeated several times to obtain multiple convergences, which can be expressed as

$$\alpha(e) = \frac{\alpha_0}{2}(\cos(\frac{\pi \operatorname{mod}(e-1, \lceil E/S \rceil}{\lceil E/S \rceil} + 1), e = 1 \cdots E \, . \tag{3}$$

where E is the number of maximal epoch, S is the number of cycle, α_0 is the initial learning rate .

To increase the diversity of candidate models, we preserve four models based on the validation set after each cycle: 1) the model with the highest accuracy; 2) the model with the largest AUC value; 3) the model with the smallest loss function value; and 4) the terminal model at the last iteration of the current cycle. In total, we obtain $N \times S \times 4$ models utilizing only the training cost for N base learners. Finally, we delete repeated models to get a candidate model set $\mathbf{MC} = \{\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_C\}$, where C is the number of candidate models.

B. Model selection with adaptive multi-objective optimization

In the model selection stage of the proposed MoEDL, we need to calculate each model's fitness value f_c based a fitness function to select and combine feasible models to obtain final prediction results. Although the AUC provides a better result than overall accuracy, as it takes both sensitivity and specificity into account, it weights omission (falsely predicted positive fraction) and commission (falsely predicted negative fraction) errors equally [33]. However, different clinical applications will emphasize different criteria. For example, in this study, we are interested in identifying high-risk patients who have never reached disease-free status and who ought to receive intensified treatments and more frequent follow-ups. Thus, the models selected are biased toward high sensitivities. To obtain more reliable and robust prediction results, we propose an adaptive multi-objective fitness function that simultaneously considers the tradeoff in sensitivity, specificity, and AUC, which can be expressed as

$$f_c = \begin{cases} \lambda \mu_c^{-\rho} + (1-\lambda)\eta_c, & \text{if } \upsilon_{\text{sen}}^c > T_{\text{sen}} \; \& \; \upsilon_{\text{spe}}^c > T_{\text{spe}} \\ 0 & \textit{otherwise} \end{cases}, \quad c = 1 \cdots C . \tag{4}$$

The proposed fitness function includes two parameters: 1) non-dominated sorting μ_c and 2) adaptive focused weighting η_c. For each candidate model, we calculate sensitivity, specificity, and AUC using the validation set and obtain the solution $\mathbf{x}_c = \left[\upsilon_{\text{sen}}^c, \upsilon_{\text{spe}}^c, \upsilon_{\text{auc}}^c\right]^T$. All candidate models are sorted in descending order using the fast non-dominated sorting approach [34] according to $\mathbf{x}_c$. The non-dominated order of each model is then generated as $\{\mu_c \mid \mu_c \in [1,2,3,...]\}$, where a smaller μ_c denotes that the model is closer to the Pareto-optimal front. All the models which locate at the Pareto-optimal front have $\mu_c = 1$. If we select model solely based on the μ_c, it can potentially lead to selected models with extremely imbalanced sensitivity and specificity. Although using the thresholds T_{sen} and T_{spe} can exclude the extremely imbalanced models, some other potential good models that are not exactly located at but close to the Pareto-optimal front can't be selected. The parameter η_c is introduced to handle this issue:

$$\eta_c = \mathbf{w}\mathbf{x}_c = \left[w_{\text{sen}}, w_{\text{spe}}, w_{auc}\right] \cdot \left[\upsilon_{\text{sen}}^c, \upsilon_{\text{spe}}^c, \upsilon_{auc}^c\right]^T, c = 1 \cdots C . \tag{5}$$

where $\mathbf{w}$ is the weight vector, and it indicates the preferences of the selected model; and C is the number of candidate models. Through combining the η_c, those models with $\mu_c \neq 1$ but with better sensitivity, specificity, and AUC, can also have a chance to be selected. Moreover, $\mathbf{w}$ can selectively emphasize different criteria depending on the clinical applications by setting $w_{\text{sen}}, w_{\text{spe}}$ or w_{auc} to a larger value.

Finally, based on each candidate model's fitness value f_c, we select the top K optimal models as the optimal model set $\mathbf{MK} = \{\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_K\}$ from candidate model set $\mathbf{MC}$ to allow for ensemble learning and to obtain final predictive results.

C. Testing with evidential reasoning fusion

In the testing stage, each test sample is fed into K trained optimal models to get K pairs of predicted probabilities, $\boldsymbol{p} = [p_k^1, p_k^2] \in R^{K\times 2}, k = 1 \cdots K$, where K is the number of optimal models. We then employ the analytic Evidential Reasoning (ER) rule [26], a generic evidence-based multi-criteria decision analysis tool, to fuse $\boldsymbol{p}$. The fused probability $\{ \boldsymbol{p}^* = [p^{1*}, p^{2*}] \}$ is obtained for all testing samples, where $p^{1*} + p^{2*} = 1$, p^{1*}, p^{2*} are the output probabilities for attaining post-treatment "disease-free" and "not disease-free," respectively. The analytic ER rule is described in Eqs. (5) and (6):

$$p^{i*} = \frac{\sigma\left[\prod_{k=1}^{K}\left(\frac{f_k p_k^i}{1+f_k-r_k} + \frac{1-r_k}{1+f_k-r_k}\right) - \prod_{k=1}^{K}\left(\frac{1-r_k}{1+f_k-r_k}\right)\right]}{1-\sigma\prod_{k=1}^{K}\left(\frac{1-r_k}{1+f_k-r_k}\right)}, i = 1, 2 \tag{6}$$

$$\sigma = \left[\prod_{k=1}^{K}\frac{f_k p_k^1 + 1 - r_k}{1+f_k-r_k} + \prod_{k=1}^{K}\frac{f_k p_k^2 + 1 - r_k}{1+f_k-r_k} - \prod_{k=1}^{K}\left(\frac{1-r_k}{1+f_k-r_k}\right)\right]^{-1} \tag{7}$$

where f_k is the fitness value for each Pareto-optimal model by Eq. (5), and r_k is the reliability of each Pareto-optimal model. Zhou et al. have provided a detailed definition of reliability [26]. The output of each test sample is determined by max($\boldsymbol{p}^*$).

3. Experimental Setup and Results

To evaluate the effectiveness of the MoEDL method, we compared it with a traditional deep perceptron neural network (DNN) [27], support vector machines (SVM) [28], and an improved group-based multi-objective model (GMO) [5]. The evaluation criteria were sensitivity, specificity, accuracy, and AUC. We used five-fold cross-validation to evaluate the performance of all methods. For all experiments, each algorithm runs five times independently, and the average results are given.

3.1 Parameter setup

For MoEDL, we set the number of base learners as N=50 and the maximal training epoch as E=1000, where the whole training process that includes E epochs are divided into S=5 cycles; we set the number of optimal models K at 20. The search space configuration set **B** was described as: 1) the number of hidden layers: $\boldsymbol{\beta}_{i,1} \in [3, 4, 5]$; 2) the number of nodes in each hidden layer: $\boldsymbol{\beta}_{i,2} \in [8, 10, 12]$; 3) the learning rate: $\boldsymbol{\beta}_{i,3} \in [5\times 10^{-5}, 1\times 10^{-4}, 5\times 10^{-4}]$;

and 4) the loss function: $\boldsymbol{\beta}_{i,4} \in$[FL; CE; MSE], where “FL” denotes Focal loss [35], “CE” denotes Cross-Entropy loss and “MSE” denotes Mean Squared Error loss. Three loss functions are respectively defined as:

$$\mathrm{FL}(p_t)=\frac{1}{n}\sum_{i=1}^{n}-\alpha(1-p_t^i)^{\gamma}\log(p_t^i)\,. \tag{8}$$

$$\mathrm{CE}(p_t)=\frac{1}{n}\sum_{i=1}^{n}-\log(p_t^i)\,. \tag{9}$$

$$\mathrm{MSE}(p)=\frac{1}{n}\sum_{i=1}^{n}(y^i-p^i)^2\,. \tag{10}$$

$$p_t=\begin{cases} p & if\ y=1 \\ 1-p & otherwise \end{cases}. \tag{11}$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0,1]$ is the model’s estimated probability for the class with label y = 1. For Focal Loss, we set γ=2，α=0.5, which are recommended settings in reference [35]. In configuration set **B**, all the hyper-parameter levels are selected based on empirical values. We determined the above parameters by a number of trials to ensure that each model achieved optimal performance. For SVM, the parameters were set to the default value. For the DNN, the number of hidden layers, the number of nodes in each hidden layer, and the learning rate were empirically set to 4, 10, and 1×10^{-4} , respectively. The maximal training epoch set as 1000. The loss functions included CE, MSE, and FL. For GMO, the parameters were set according to the original work [5], with the exception of the number of optimal models and the number of base learner, which we set to 20 and 50; for ease of comparison, we made these consistent with the numbers for MoEDL.

3.2 Influence of weights in MoEDL

To investigate the impact of adaptive focused weighting on the model’s prediction performance, we evaluated the performance of MoEDL with different weights for sensitivity, specificity, accuracy, and AUC. Figure 2 illustrates the influence of the sensitivity, as we changed the sensitivity weighting w_{sen} from 1 to 10 while keeping the weightings for specificity and AUC, w_{spe} and w_{auc}, at 1. These results demonstrate that, without additional training, we can emphasize different criteria in the final predictive model to meet the preferences or requirements of different practical applications by changing only the weighting values. The weighting $\mathbf{w}=\left[w_{\mathrm{sen}}, w_{\mathrm{spe}}, w_{\mathrm{auc}}\right]=[3,1,1]$, achieves a good balance between sensitivity and specificity.

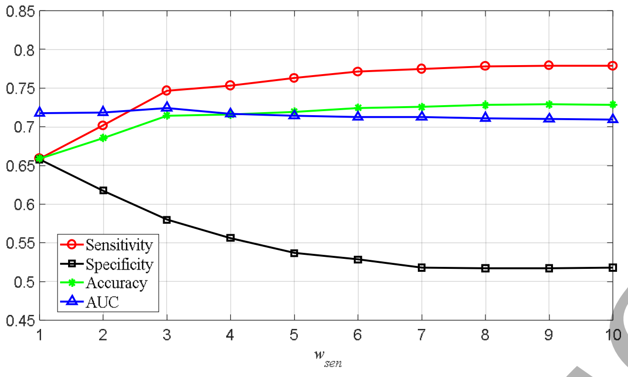
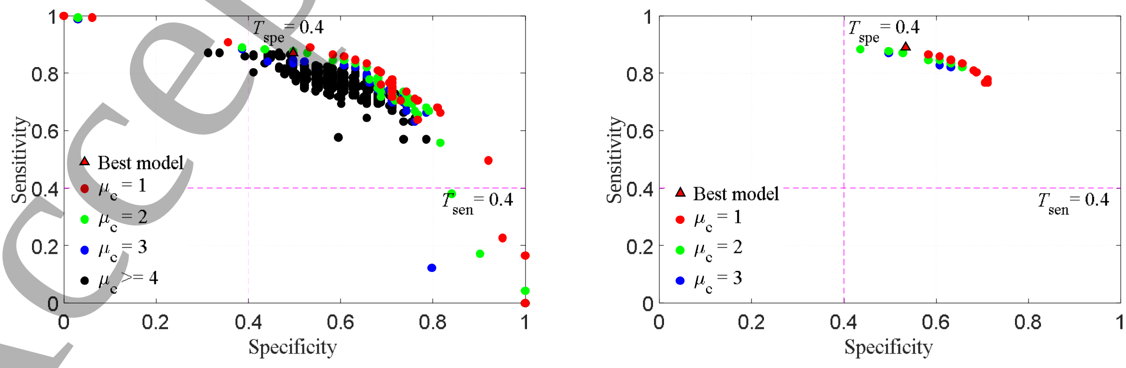


Figure 2: Comparison of MoEDL with different w_{sen}.

3.3 Illustration of diversity in MoEDL

As MoEDL is an ensemble learning method that requires constructing multiple models, the diversity of the candidate model greatly affects the final prediction. In Figure 3(a), we visualized the candidate model set of MoEDL to illustrate its diversity, where $\mu_c \in [1,2,3,\cdots]$ is the non-dominated order value determined by the fast non-dominated sorting approach, and the smaller μ_c denotes the model that is closer to the Pareto-optimal front. The best model is the one with the highest fitness value by Eq. (4). As shown in Figure 3(a), all candidate models are evenly distributed at the pareto front. For all Pareto-optimal models (μ_c =1), the means of the sensitivity, specificity, and AUC are 0.72, 0.69, and 0.75 respectively. These illustrate that our dynamic ensemble deep learning strategy and SE restarting strategy effectively improve diversity.

Figure 3(b) shows the optimal model set selected from the candidate model set in Figure 3(a). In MoEDL, we not only excluded the models with extremely imbalanced sensitivity and specificity by the thresholds T_{sen} and T_{spe}, respectively, but we also combined the non-dominated sorting and adaptive focused weighting. These strategies make optimal models more balanced and reliable. The mean values for the sensitivity, specificity, and AUC of the optimal models are 0.84, 0.61, and 0.77, respectively.



(a) Candidate model set (b) Optimal model set

Figure 3: Illustration of diversity in MoEDL.

3.4 Comparison with other models

A. Comparison with DNN

We compared the performance of MoEDL to the DNN coupled with three different loss functions: Mean Squared Error ($\mathrm{DNN_{MSE}}$), Cross-Entropy ($\mathrm{DNN_{CE}}$) and Focal-Loss ($\mathrm{DNN_{FL}}$). The maximal training epoch for all algorithms was set as 1000. From the results in Table 1, DNN with focal loss has slightly higher AUC as compared to DNN with MSE or CE losses. Meanwhile, MoEDL with dynamic parameter selection achieves better performance than all DNN-based methods measured by AUC.

Table 1: The loss function comparison results of DNN and MoEDL.

Method	Sensitivity	Specificity	Accuracy	AUC
$\mathrm{DNN_{MSE}}$	0.6518±0.0412	0.6228±0.0364	0.6463±0.0270	0.6842±0.0039
$\mathrm{DNN_{CE}}$	0.6609±0.0621	0.6166±0.0844	0.6524±0.0357	0.6898±0.0020
$\mathrm{DNN_{FL}}$	0.6654±0.0309	**0.6321**±0.0260	0.6590±0.0207	0.6923±0.0027
MoEDL	**0.7565**±0.0071	0.5839±0.0127	**0.7266**±0.0033	**0.7221**±0.0041

B. Comparison with GMO

In the original GMO, the model selection is only based on the non-dominated order μ_c , that is all models with $\mu_c = 1$ are selected as the optimal model set. To analyze the impact of our model selection strategy on GMO performance, we replaced the original model selection strategy of GMO (only based on non-dominated order) to our model selection strategy, while the rest part of this modified GMO (mGMO) remains the same as the original GMO. The number of optimal models and the number of base learners were set to 20 and 50, respectively. The comparison results are summarized in Table 2. The mGMO achieves lightly better AUC than the original GMO. Because the model selection of the original GMO only considers the score of a non-dominated sorting, some of the models selected have extremely imbalanced sensitivity and specificity and may negatively impact the final predictive results.

Table 2: The comparison results of GMO and modified GMO.

	Sensitivity	Specificity	Accuracy	AUC
GMO	0.7722±0.0102	0.5026±0.0207	0.7206±0.0058	0.6871±0.0102
mGMO	0.7464±0.0153	0.5585±0.0392	0.7104±0.0058	**0.6974**±0.0073

C. Comprehensive comparison with other models

We used the mean and standard deviation (Std) of each evaluation criterion to quantify the performance of the SVM, DNN_{FL}, mGOM and MoEDL. The results of the comparison of the four methods are reported in Figure 4. DNN_{FL} obtained better results than SVM, because the deep learning method is more proficient at dealing with heterogeneity, sparseness, random errors, and inconsistencies than traditional machine learning methods. Between the two ensemble methods that we compared, MoEDL obtained better results than mGMO. This is primarily because the base learner of mGMO is based on an SVM. MoEDL achieves the most stable results with the lowest Std values. This is because MoEDL selects from a variety of models so that more robust results can be achieved on the test dataset. The receiver operating characteristic curves and AUC values for the four methods compared are shown in Figure 5. These results demonstrate that the proposed MoEDL outperforms the other three methods. Meanwhile, all results show that deep learning-based methods perform better than the traditional machine learning methods.

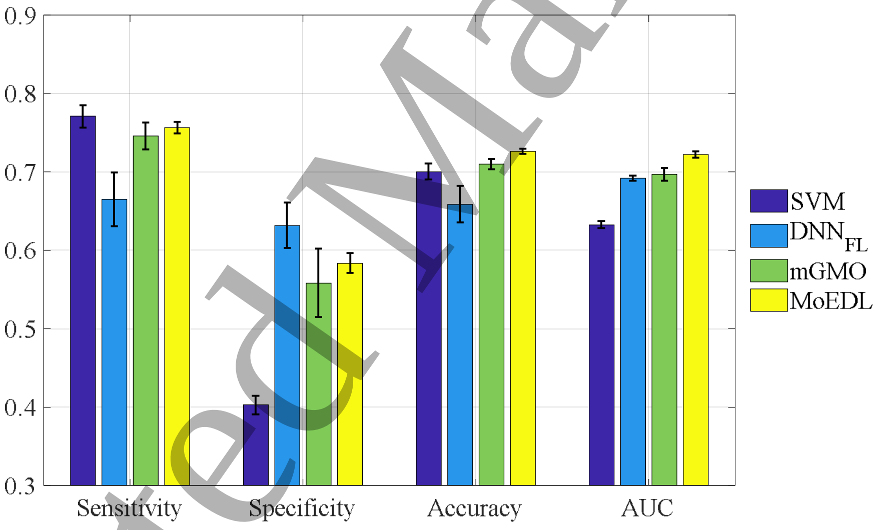


Figure 4. Comparison of four methods.

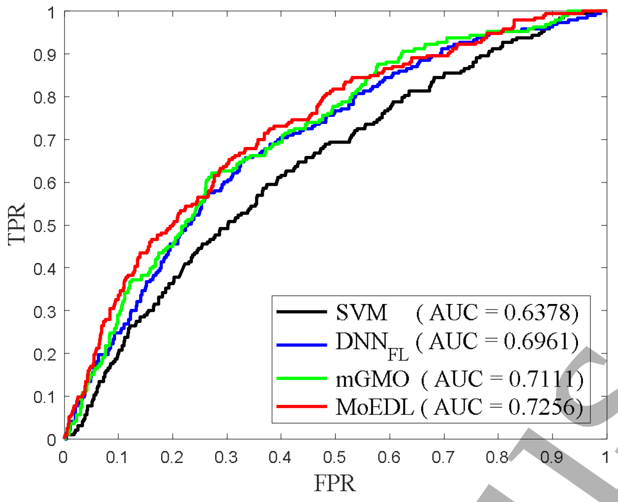


Figure 5: Receiver operating characteristic curves and AUC values of four methods.

The *p*-values of the paired *t*-test between MoEDL and each of other methods are shown in Table 3. These results show that there is a statistically significant difference between proposed MoEDL and other methods at a significant level of 0.05.

Table 3: *p*-values in paired *t*-test between MoEDL and other methods.

Method	SVM	DNN$_{MSE}$	DNN$_{CE}$	DNN$_{FL}$	GMO	mGMO
p-value	<0.0001	<0.0001	<0.0001	<0.0001	0.0014	0.0033

3.5 Importance of the individual feature in MoEDL

We adopted the method in [29] to investigate the importance of individual features, which is performed by artificially modifying their values in the acceptable range for the corresponding variable value, and evaluating changes in neural network predictions. Specifically, for all the test samples, we artificially changed each feature value to their corresponding maximal and minimal feature value, and then each modified test sample was fed into the trained MoEDL model. The importance of individual features was evaluated by the AUC change in MoEDL predictions. The results are given in Table 4. In order to visualize the change better, the magnitude of AUC change for each feature are shown in Figure 6. A larger change indicates the greater contribution of this features on predicted outcomes in MoEDL. The main contributing features are tumor size, regional dose, T staging, and N staging.

Table 4. The importance analysis of individual features in MoEDL. The first and second column of the AUC value corresponds to the result using the minimal or maximal value of the corresponding feature, respectively. For example, if we artificially changed the "Age" feature to its minimal value "32", the resulting AUC was 0.7055. Similarly, if we changed

the “Age” feature to its maximal value “92”, the resulting AUC was 0.7080.

Feature		Feature range		AUC	
Numerical feature	Age, years	32	92	0.7055	0.7080
	Tumor size, mm	5	430	0.7097	0.6845
	Regional Dose, cGy	1200	7740	0.6720	0.7019
	Boost Dose, cGy	280	6000	0.7105	0.6973
Ordinal feature	TNM T	T1	T4	0.6980	0.7033
	TNM N	N0	N3	0.6958	0.6714
	TNM M	M0	M1	0.7108	0.7017
	TNM Stage	Stage I	Stage IV	0.7040	0.7003
	Derived AJCC T	T1	T4	0.7005	0.7037
	Derived AJCC N	N0	N3	0.6927	0.6631
	Derived AJCC M	M0	M1	0.7108	0.7039
	Derived AJCC Stage	Stage I	Stage IV	0.7011	0.6928
Nominal feature	Sex	Male	Female	0.7125	0.7092
	Histologic	Small cell carcinoma	Other	0.7113	0.7086
	Primary Site	Main bronchus	other	0.7118	0.7071
	Regional Nodes Examined	Yes	No	0.7084	0.6984
	Surgical Margins	Yes	No	0.6898	0.7067
	Treatment Volume	Chest/lung	Other	0.7088	0.7000
	Surgery or not	Yes	No	0.7089	0.7035
	Chemo or not	Yes	No	0.6914	0.6981

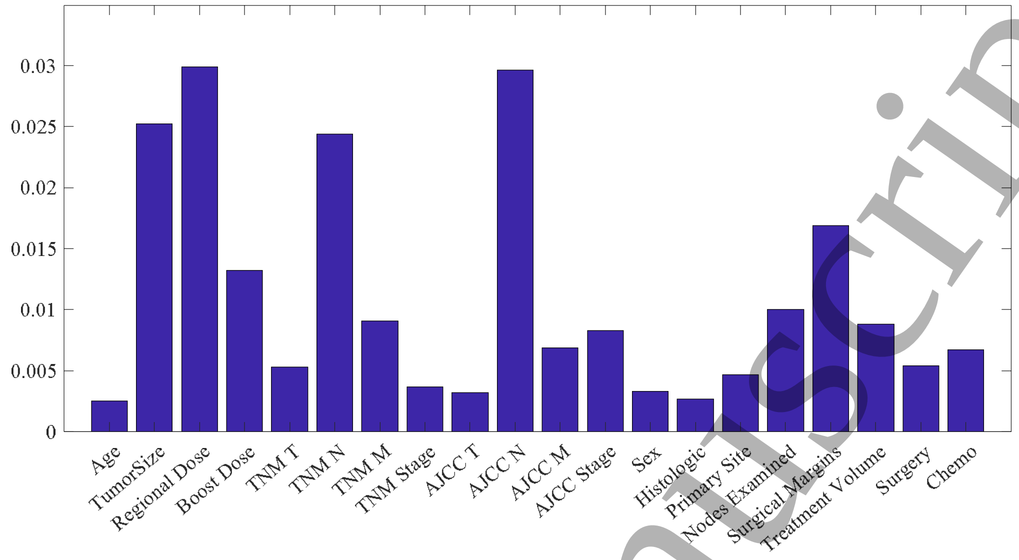


Figure 6. The magnitude of AUC changes for all features

4. Discussion and conclusion

Accurately predicting treatment outcomes can facilitate personalized treatment plans and follow-up schedules. In this work, we explored both ensemble deep learning and multi-objective optimization techniques that analyze electronic health records data to predict high risk of treatment failure after radiotherapy in patients with lung cancer. The method that we developed, MoEDL, is a reliable and dynamic method that makes the following contributions: 1) The dynamic ensemble deep learning strategy can randomly search and generate several different deep perceptron networks whose architecture and key parameters are dynamically adjusted to increase diversity and reduce the time spent tuning hyper-parameters; and 2) The adaptive model selection strategy, based on multi-objective optimization, can adjust the prediction model to focus more on certain evaluation criteria (sensitivity, specificity, or AUC) to meet different practical applications.

Although deep learning-based approaches are powerful for handling EHR data, most deep learning architectures are developed manually by human experts, so they require large amounts of time for hyper-parameter tuning. Because of this, there is a growing interest in automated neural architecture search (NAS) methods. NAS methods have been shown to outperform manually designed architectures for certain tasks [36]. NAS can potentially be incorporated into the proposed MoEDL framework to improve its performance. For example, NAS could provide a better method to design the search space, which includes specifying the size and the content of the search space; it could also enable us to develop a search strategy to quickly find well-performing architectures and avoid premature convergence. We will explore how to further utilize the NAS to improve our method's performance using EHRs in a future study.

In addition, missing data is a common phenomenon in EHR-based data mining. Missing values in EHRs often result from human error, such as lack of collection or lack of documentation [37]. There are many different approaches to address missing values, and these approaches are usually divided into two broad categories: deletion methods and imputation methods [38]. Deletion methods, also called "complete case analysis" or "listwise deletion," simply exclude cases with missing values. Although deletion methods can be used with any kind of data and without additional computation, they can cause removing a large amount of potentially usable information. Deletion methods are preferred in cases where the percentage of missing values is relatively low (<5%) [39]. Imputation methods usually replace each missing value with another value determined from a reasonable guess. In this work, we implemented a K-Nearest Neighbor (KNN) imputation method [40], which replaces each missing attribute value with the median value of that attribute from the K-nearest neighbors. One way that we could improve the MoEDL is by implementing more advanced imputation techniques, such as multiple imputation by chained equations (MICE) [41].

The main contributing features are tumor size, regional dose, T staging, and N staging in MoEDL through feature importance analysis in Section 3.5. These results agree well with our clinical knowledge. For example, tumor size, T staging, and N staging are characteristics which define the extent of the disease. Limited, early stage disease (small tumor, low T stage, and no nodal involvement) typically responds well to treatment with excellent prognosis; whereas locally advanced (larger tumor, high T stage, and multiple lymph node involvement) portends a worse prognosis and a higher risk of recurrence. Simultaneously, the amount of dose received in the treatment volume is a direct correlation to outcome as predicted by MoEDL. Inadequate treatment could result in little response, no response, or response without durability. This finding is consistent with prescribing, achieving, and delivering adequate radiotherapy dosing to treatment volume as well as the importance clinicians that have placed on dose escalation/de-escalation studies.

In summary, we have developed MoEDL, a method that uses dynamic ensemble deep learning and adaptive model selection based on multi-objective optimization. This method predicts treatment outcomes by analyzing information extracted from EHRs better than other conventional methods. Based on accurate outcome predictions, we can stratify risk of treatment failure after radiotherapy in patients with lung cancer. This method will help in designing personalized treatment plans and follow-up schedules, which could increase survival rates of patients with lung cancer after radiation therapy.

Acknowledgment

This work was supported in part by National Institutes of Health (R01 EB020366). The authors thank Dr. Jonathan Feinberg for scientific editing.

Appendix A. Extracted features

The process of extracting features from the EHR dataset is mainly composed of the following two parts.

First, we use two items, “Cancer Status” and “Recurrence Type 1st”, to determine the sample label, which is defined as:

$$\text{label} = \begin{cases} 1 \quad \textit{if} \;\; \text{Recurrence Type } 1^{\text{st}} \neq 00 \;\; \& \;\; \text{Cancer Status} = 2 \\ 0 \quad \textit{if} \;\; \text{Recurrence Type } 1^{\text{st}} = 00 \;\; \& \;\; \text{Cancer Status} = 1 \end{cases} \qquad \text{(A.1)}$$

where label “0” denotes disease-free and “1” denotes never been disease-free; “Recurrence Type 1 st = 00” means the patient became disease-free after treatment and has not had a recurrence; “Cancer Status = 1” means no evidence of this tumor; and “Cancer Status = 2” means evidence of this tumor. The specific meaning of all the items mentioned in this paper can be found in the North American Association of Central Cancer Registries data dictionary [42].

Second, on the basis of characteristics of items that are closely related to radiation therapy, we extract 20 features, including demographic parameters, tumor characteristics, and treatment schemes. Due to the heterogeneous nature of EHR data, we divided all features into three categories of attributes—numerical, ordinal, and nominal attributes—and then adopted different extraction methods for each attribute feature. The details and extraction methods for all 20 features are introduced separately below.

A.1. Numerical features

A numerical or continuous feature is one for which the raw code is always a number and can be measured. There are 4 numerical features, as listed below. All numerical features are extracted by converting original code values to numerical ones except in “unknown” cases. The four numerical features and corresponding patient characteristics are listed in Table A1.

Table A1: Numerical feature patient characteristics.

Characteristics	Outcome	
	Never been disease-free (+)	Disease-free (−)

	no. (%)	Mean SD	Median (range)	no. (%)	Mean SD	Median (range)
Age, years						
Exact age	814 (100)	66±11	65 (32-92)	193 (100)	66±11	65 (39-89)
Tumor size, mm						
Exact size	704 (86.49)	45±31	37 (5-430)	178 (92.23)	32±19	28 (5-112)
Unknown or not applicable	110 (13.51)	-	-	15 (7.77)	-	-
Regional Dose, cGy						
Exact dose	732 (89.93)	5567±1116	6000 (1200-7740)	187 (96.89)	5673±797	5891 (3400-7400)
Unknown or not applicable	82 (10.07)	-	-	6 (3.11)	-	-
Boost Dose, cGy						
Boost was not administered	752 (92.39)	-	-	184 (95.34)	-	-
Exact dose	57 (7.00)	2152±1066	2000 (280-6000)	7 (4.40)	1820±564	1620 (1000-2600)
Unknown or not applicable	5 (0.61)	-	-	2 (0.80)	-	-

A.2. Ordinal features

An ordinal feature is one that can be ordered and ranked, but not measured. There are 8 ordinal features, as listed below. For all ordinal raw data, we first convert the original order to a numerical order; then, the same clinical phenotype is fused with different schema by averaging. For example, the four features TNM T, TNM N, TNM M, and TNM Stage are extracted from TNM clinical and TNM pathologic records. Meanwhile, Derived AJCC T, AJCC N, AJCC M, and AJCC Stage are extracted from AJCC6 and AJCC7 records. AJCC6 and AJCC7 are the sixth and seventh edition, respectively, of the Cancer Staging Manual by the American Joint Committee on Cancer. All ordinal features and corresponding patient characteristics are listed in Table A2.

Table A2: Ordinal feature patient characteristics.

Characteristics	Outcome	
	Never been disease-free (+)	Disease-free (−)
TNM T, no. (%)		
T1	208 (25.55)	89 (46.11)
T2	227 (27.89)	56 (29.02)
T3	168 (20.64)	21 (10.88)
T4	187 (22.97)	20 (10.36)
Unknown or not applicable	24 (2.95)	7 (3.63)
TNM N, no. (%)		
N0	262 (32.19)	110 (56.99)
N1	86 (10.57)	21 (10.88)
N2	315 (38.70)	55 (28.50)
N3	136 (16.71)	5 (2.59)
Unknown or not applicable	15 (1.84)	2 (1.04)
TNM M, no. (%)		
M0	796 (97.78)	190 (98.45)
M1	9 (1.11)	0 (0)
Unknown or not applicable	9 (1.11)	3 (1.55)
TNM Stage, no. (%)		

Stage I	173 (21.25)	91 (47.15)
Stage II	91 (11.18)	27 (13.99)
Stage III	525 (64.50)	71(36.79)
Stage IV	9 (1.11)	0 (0)
Unknown or not applicable	16 (1.97)	4 (2.07)
Derived AJCC T, no. (%)		
T1	50 (6.14)	26 (13.47)
T2	227 (27.89)	72 (37.31)
T3	277 (34.03)	59 (30.57)
T4	219 (26.90)	28 (14.51)
Unknown or not applicable	41 (5.04)	8 (4.15)
Derived AJCC N, no. (%)		
N0	267 (32.80)	112 (58.03)
N1	65 (7.99)	18 (9.33)
N2	338 (41.52)	58 (30.05)
N3	130 (15.97)	4 (2.07)
Unknown or not applicable	14 (1.72)	1 (0.52)
Derived AJCC M, no. (%)		
M0	784 (96.31)	192 (99.48)
M1	24 (2.95)	1 (0.52)
Unknown or not applicable	6 (0.74)	0 (0.00)
Derived AJCC Stage, no. (%)		
Stage I	176 (21.62)	92 (47.67)
Stage II	76 (9.34)	26 (13.47)
Stage III	540 (66.34)	72 (37.31)
Stage IV	1 (0.12)	0 (0)
Unknown or not applicable	21 (2.58)	3 (1.55)

A.3. Nominal features

A nominal feature is used for labeling variables without providing any quantitative value. All nominal scales correspond to different classes and are mutually exclusive (no overlap). We extract 8 nominal features, as listed below. Based on the specific medical meaning, each item's raw code is converted to a corresponding scale by sub-class merging and small-class merging grouping methods. All nominal features and corresponding patient characteristics are listed in Table A3.

Table A3: Nominal feature patient characteristics.

Characteristics	Outcome	
	Never been disease-free (+)	Disease-free (−)
Sex, no. (%)		
Male	459 (56.39)	100 (51.81)
Female	355 (43.61)	93 (48.19)
Histologic, no. (%)		
Small cell carcinoma	87 (10.69)	11 (5.70)
Non-small cell carcinoma	99 (12.16)	22 (11.40)
Other	628 (77.15)	160 (82.90)
Primary Site, no. (%)		
Main bronchus	45 (5.53)	3 (1.55)

Upper lobe, lung	478 (58.72)	123 (63.73)
Middle lobe, lung	31 (3.81)	7 (3.63)
Lower lobe, lung	196 (24.08)	49 (25.39)
Other	64 (7.86)	11 (5.70)
Regional Nodes Examined, no. (%)		
No nodes were examined	623 (76.54)	146 (75.65)
Nodes were examined	30 (3.69)	11 (5.70)
No regional nodes were removed	141 (17.32)	35 (18.13)
Other	20 (2.46)	1 (0.52)
Surgical Margins, no. (%)		
No residual tumor	38 (4.67)	25 (12.95)
No primary site surgery	745 (91.52)	164 (84.97)
Residual tumor	28 (3.44)	4 (2.07)
Unknown or not applicable	3 (0.37)	0 (0)
Treatment Volume, no. (%)		
Chest/lung	632 (77.64)	123(63.73)
Lung (limited)	142(17.44)	62 (32.12)
Chest wall	9 (1.11)	4 (2.07)
Other	31 (3.81)	4 (2.07)
Surgery or not, no. (%)		
With surgery	250 (30.71)	63 (32.64)
Without surgery	564 (69.29)	130 (67.36)
Chemo or not, no. (%)		
With chemo	525 (64.50)	96 (49.74)
Without chemo	289 (35.50)	97 (50.26)

References

1. *World Health Organization.* https://www.who.int/news-room/fact-sheets/detail/cancer.

2. Birkhead, G.S., M. Klompas, and N.R. Shah, *Uses of electronic health records for public health surveillance to advance public health.* Annual review of public health, 2015. **36**: p. 345-359.

3. Shickel, B., et al., *Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis.* IEEE journal of biomedical and health informatics, 2017. **22**(5): p. 1589-1604.

4. Zhou, Z., et al., *Predicting distant failure in early stage NSCLC treated with SBRT using clinical parameters.* Radiotherapy and Oncology, 2016. **119**(3): p. 501-504.

5. Zhou, Z., et al., *Multi-objective radiomics model for predicting distant failure in lung SBRT.* Physics in Medicine & Biology, 2017. **62**(11): p. 4460.

6. Nickerson, P., et al. *Deep neural network architectures for forecasting analgesic response*. in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016. IEEE.

7. Tran, T., et al., *Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM).* Journal of biomedical informatics, 2015. **54**: p. 96-105.

8. Li, H., et al., *Identifying informative risk factors and predicting bone disease progression via deep belief networks.* Methods, 2014. **69**(3): p. 257-265.

9. Esteban, C., et al. *Predicting clinical events by combining static and dynamic information using recurrent neural networks*. in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. 2016. Ieee.

10. Miotto, R., et al., *Deep patient: an unsupervised representation to predict the future of patients from the electronic health records.* Scientific reports, 2016. **6**: p. 26094.

11. Nguyen, P., et al., *$\mathtt {Deepr} $: a convolutional net for medical records.* IEEE journal of biomedical and health informatics, 2016. **21**(1): p. 22-30.

12. Bengio, Y., I.J. Goodfellow, and A. Courville, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.

13. Lv, X., et al., *Clinical relation extraction with deep learning.* IJHIT, 2016. **9**(7): p. 237-248.

14. Choi, Y., C.Y.-I. Chiu, and D. Sontag, *Learning low-dimensional representations of medical concepts.* AMIA Summits on Translational Science Proceedings, 2016. **2016**: p. 41.

15. Choi, E., et al., *Medical concept representation learning from electronic health records and its application on heart failure prediction.* arXiv preprint arXiv:1602.03686, 2016.

16. Che, Z., et al. *Deep computational phenotyping*. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015. ACM.

17. Quanming, Y., et al., *Taking human out of learning applications: A survey on automated machine learning.* arXiv preprint arXiv:1810.13306, 2018.

18. Feurer, M., et al. *Efficient and robust automated machine learning*. in *Advances in neural information processing systems*. 2015.

19. Kotthoff, L., et al., *Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA.* The Journal of Machine Learning Research, 2017. **18**(1): p. 826-830.

20. Zoph, B. and Q.V. Le, *Neural architecture search with reinforcement learning.* arXiv preprint arXiv:1611.01578, 2016.

21. Liu, C., et al. *Progressive neural architecture search*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

22. Katz, G., E.C.R. Shin, and D. Song. *Explorekit: Automatic feature generation and selection*. in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016. IEEE.

23. Kanter, J.M. and K. Veeramachaneni. *Deep feature synthesis: Towards automating data science endeavors*. in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2015. IEEE.

24. Zhou, Z.-H., *Ensemble learning.* Encyclopedia of biometrics, 2015: p. 411-416.

25. Huang, G., et al., *Snapshot ensembles: Train 1, get m for free.* arXiv preprint arXiv:1704.00109, 2017.

26. Zhou, Z., et al., *Constructing multi-modality and multi-classifier radiomics predictive models through reliable classifier fusion.* arXiv preprint arXiv:1710.01614, 2017.

27. Kim, P., *Matlab deep learning*, in *With Machine Learning, Neural Networks and Artificial Intelligence*. 2017, Springer.

28. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM transactions on

intelligent systems and technology (TIST), 2011. **2**(3): p. 27.

29. Ibragimov, B., et al., *Neural networks for deep radiotherapy dose analysis and prediction of liver SBRT outcomes.* IEEE journal of biomedical and health informatics, 2019.

30. Bishop, C.M., *Pattern recognition and machine learning*. 2006: springer.

31. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.

32. Loshchilov, I. and F. Hutter, *Sgdr: Stochastic gradient descent with warm restarts.* arXiv preprint arXiv:1608.03983, 2016.

33. Lobo, J.M., A. Jiménez-Valverde, and R. Real, *AUC: a misleading measure of the performance of predictive distribution models.* Global ecology and Biogeography, 2008. **17**(2): p. 145-151.

34. Deb, K., et al., *A fast and elitist multiobjective genetic algorithm: NSGA-II.* IEEE transactions on evolutionary computation, 2002. **6**(2): p. 182-197.

35. Lin, T.-Y., et al. *Focal loss for dense object detection*. in *Proceedings of the IEEE international conference on computer vision*. 2017.

36. Elsken, T., J.H. Metzen, and F. Hutter, *Neural architecture search: A survey.* arXiv preprint arXiv:1808.05377, 2018.

37. Wells, B.J., et al., *Strategies for handling missing data in electronic health record derived data.* Egems, 2013. **1**(3).

38. Soley-Bori, M., *Dealing with missing data: Key assumptions and methods for applied analysis.* Boston University, 2013.

39. Schafer, J.L., *Analysis of incomplete multivariate data*. 1997: Chapman and Hall/CRC.

40. Faisal, S. and G. Tutz, *Nearest neighbor imputation for categorical data by weighting of attributes.* arXiv preprint arXiv:1710.01011, 2017.

41. Azur, M.J., et al., *Multiple imputation by chained equations: what is it and how does it work?* International journal of methods in psychiatric research, 2011. **20**(1): p. 40-49.

42. *Data Dictionary – NAACCR.* http://datadictionary.naaccr.org/.