A hybrid deep learning model for forecasting lymphocyte depletion during radiation therapy

Journal:	*IEEE Journal of Biomedical and Health Informatics*
Manuscript ID	JBHI-00573-2021
Manuscript Type:	Paper
Date Submitted by the Author:	12-Mar-2021
Complete List of Authors:	Ebrahimi, Saba; University of Houston Lim, Gino; University of Houston Hobbs, Brian; The University of Texas at Austin Mohan, Radhe; The University of Texas MD Anderson Cancer Center Lin, Steven; The University of Texas MD Anderson Cancer Center Cao, Wenhua; The University of Texas MD Anderson Cancer Center

A hybrid deep learning model for forecasting lymphocyte depletion during radiation therapy

Saba Ebrahimi, Gino Lim, Brian Hobbs, Steven H Lin, Radhe Mohan, and Wenhua Cao

***Abstract*—Recent studies have shown that severe depletion of the absolute lymphocyte count (ALC) induced by radiation therapy (RT) reduces the survival of patients with many solid tumors. In this paper, we aimed to predict radiation-induced lymphocyte depletion in esophageal cancer patients during the course of RT based on patient characteristics and dosimetric features. We proposed a hybrid model using a deep neural network and a long short-term memory network in a stacked structure to predict a trend toward ALC depletion before or in the early stages of RT treatment. Several important prediction metrics were employed to evaluate the performance of the proposed model compared to other commonly used prediction methods. The results showed that the proposed model outperformed off-the-shelf prediction methods with a reduction of least 30% in the mean square error (MSE) of weekly ALC predictions based on pretreatment data. Moreover, using an extended model based on first-week data reduced the MSE of predictions by 69.6% compared to the model based on the pretreatment data. In conclusion, our model performed well in predicting radiation-induced lymphocyte depletion for RT treatment planning. The ability to predict ALC will enable physicians to evaluate individual RT treatment plans for lymphopenia risk and to identify patients at high risk who would benefit from modified treatment approaches.**

***Index Terms*—Deep learning, LSTM, radiation therapy, lymphopenia**

I. Introduction

The application of artificial intelligence and machine learning methods to extract insights from data is becoming increasingly attractive in many fields, including healthcare. Although many healthcare applications have been developed, those that can predict disease progression [1], [2], treatment outcomes [3], or potential side effects [4], [5] play an important role in improving patients' care.

Radiation therapy (RT) is an effective treatment option for many cancer patients. An RT patient undergoes a series of treatment sessions over several weeks to deliver a prescribed dose of radiation to the tumor. The clinical goal of RT is to maximize the radiation-induced damage to the tumor, killing all cancerous cells, while minimizing toxic effects on surrounding healthy tissues [6].

Recent studies have shown that the absolute lymphocyte count (ALC) is very sensitive to radiation exposure; by killing the circulating lymphocytes in the radiation field, RT suppresses the immune system [7]– [8]. The resulting reduction in ALC causes radiation-induced lymphopenia (RIL), a common toxic effect of RT [9], [10]. Clinical studies have shown that severe lymphopenia can reduce the survival of patients with a number of solid tumors, including esophageal cancer [11], [12]. Therefore, the ability to reliably predict RIL on the basis of pretreatment factors (i.e., dosimetric factors, treatment factors, and patient-specific factors) would improve RT planning. Specifically, predicting the risk of lymphocyte depletion during early RT fractions could identify patients who are at high risk of grade 4 RIL and who may stand to benefit from RIL mitigation strategies [13], [14] and modified treatments that may ultimately improve their survival.

Several studies have shown strong associations between pretreatment factors and the risk of severe RIL in various cancers [10]– [15]. van Rossum et al. [13] developed a pretreatment clinical nomogram based on age, planning target volume, body mass index, radiation modality, and baseline ALC to determine the risk of grade 4 RIL for new patients. Zhu et al. [15] developed a hybrid deep learning model to classify patients with grade 4 RIL based on patient characteristics and dosimetric features. However, models that can forecast the kinetics of lymphocyte loss after fractionated radiation exposures in order to identify high-risk patients are lacking. It is critical to fill this gap to develop RIL mitigation strategies and improve the effectiveness of RT for cancer patients. Therefore, in this study, we aimed to predict radiation-induced lymphocyte depletion in esophageal cancer patients during the course of RT on the basis of pretreatment or early-treatment information.

Deep learning models have been developed to extract information from various kinds of data and for many tasks [16], [17]. Recurrent neural networks (RNNs) achieved significant results in extracting temporal information from sequential data such as text, audio, video, and time series [16]. The main advantage of RNNs is that they can maintain memory of recent events and update their current state based on both past states and current input data [18], [19]. Hochreiter and Schmidhuber [20] proposed the long short-term memory (LSTM) network as an improved variant of the RNN to handle the long-term dependency and vanishing gradient issues of RNNs. LSTM networks have been widely used for various kinds of tasks, including speech recognition [21], [22], image captioning [23], [24], trajectory prediction [25], [26], and text embedding [27], [28]. However, an LSTM network cannot be used alone for the current problem because the significant features that may predict RIL do not have uniform characteristics. A potential solution for this issue is to develop a stacked structure.

In a stacked structure, the algorithm nonlinearly integrates predictors in order to achieve higher prediction accuracy

Saba Ebrahimi, Gino Lim are with Department of Industrial Engineering, University of Houston, Texas, Houston

Brian Hobbs is with Department of Population Health, The University of Texas at Austin, Texas, Austin

Steven H Lin is with Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Texas, Houston

Radhe Mohan, Wenhua Cao are with Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Texas, Houston

and reduce generalization error. Deep-stacked models can outperform state-of-the-art deep learning and machine learning models such as tree-based ensemble models and extreme gradient boosting algorithms [29], [30]. Therefore, we propose a hybrid deep-stacked model that combines a deep neural network with an LSTM network for different groups of features in a stacked structure. The proposed structure consists of 4 channels to process 4 categories of features with different characteristics; LSTM is used to process the sequential features. We established 3 models to predict an ALC depletion trend on the basis of pretreatment, first-week, and second-week treatment information. To evaluate the performance of our proposed hybrid deep-stacked model, we calculated well-known prediction metrics and compared the results with other common prediction methods.

In summary, the contribution of this paper is:

- A hybrid deep-stacked structure based on pretreatment information is proposed to predict RIL for new esophageal cancer patients during the course of RT.
- The proposed hybrid deep-stacked structure can use information from different groups of features with different characteristics to predict weekly ALC without requiring a large amount of data, while reducing bias and the adverse effects of any noise in the data.
- The developed model is flexible and can be extended easily to account for early-treatment predictions (i.e., at the end of week 1 or 2), and a discriminative kernel layer was developed to evaluate the importance weight of each value in the input sequence.

The rest of this paper is organized as follows: section 2 covers the data description, data preprocessing, and our proposed hybrid deep-stacked model; section 3 presents the experimental results and discusses the key findings of our study; and section 4 concludes and provides directions for further research.

II. Materials and methods

A. Data description

This study was approved by the University of Texas MD Anderson Cancer Center institutional review board. All methods performed here were in accordance with the Health Insurance Portability and Accountability Act. Data from 860 patients who received concurrent chemoradiotherapy (with or without surgery) for biopsy-proven esophageal cancer between January 2004 and November 2017 at MD Anderson Cancer Center were used for this study. All patients were treated with proton or photon radiation modalities with a total radiation dose of 50.4 Gy over 5 weeks. All included patients also had available baseline ALC values and 3 or more documented weekly ALC values during treatment.

B. Variable selection

The variables of interest for the prediction were ALC after each week of treatment. Predictor variables were selected on the basis of their clinical relevance, their low level of missingness (<20%), and the results of the correlation analyses presented in previous studies on this dataset by Zhu et al. [15] and van Rossum et al. [13]. As a result, 52 features were selected as predictors and categorized into 4 main groups on the basis of their clinical and analytical characteristics: (1) dosimetric features contained 30 dose-volume metrics such as V_5, V_{10}, … ,V_{45} (V_x refers to the percentage of the volume that received at least x Gy radiation dose) and mean dose for 3 organs at risk: the lung, heart, and spleen; (2) other numerical treatment-related and patient-specific parameters were nondosimetric numerical features including age, body mass index, total blood volume, planning target volume, and blood component profiles at baseline (red blood cells, white blood cells, and others); (3) nondosimetric categorical features such as RT modality (proton or photon), race, sex, tumor location, tumor histologic characteristics, and use of induction chemotherapy; (4) sequential features included the sequence of 5 weekly ALC values.

Zhu et al. [15] reported high collinearity between dosimetric features based on the high variance inflation factor of each dosimetric predictor. This is because of the sequential and highly intercorrelated nature of dose-volume histograms. We used the t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction method to reduce the effect of this severe collinearity among dosimetric features without losing information. t-SNE is a nonconvex and nonlinear method to reduce the dimensionality of data by considering the similarity between features that follow a conditional exponential probability distribution as

$$P(j|i) = \frac{\exp \dfrac{\left\|x_i - x_j\right\|^2}{2\sigma_i^2}}{\sum_{i\neq j} \exp \dfrac{\left\|x_i - x_j\right\|^2}{2\sigma_i^2}} \ , \tag{1}$$

where $p(j|i)$ is the similarity between features x_i and x_j considering the original data features as x_1, x_i, … ,x_N. The total similarity between these 2 variables is the mean value of 2 conditional probabilities divided by the number of features ($p_{ij} = \frac{p(j|i)+p(i|j)}{2N}$). The new d dimensional data y_1, y_i, … ,y_d must reflect the p_{ij} as much as possible. So, q_{ij} can be estimated as

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{i\neq j} \left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}} \ . \tag{2}$$

The values of new features can be calculated by minimizing the Kullback–Leibler (KL) divergence between the distributions of data before (P) and after (Q) dimensionality reduction.

$$\min KL(P|\,(Q) = \textstyle\sum_{i\neq j} p_{ij} log \dfrac{p_{ij}}{q_{ij}} \tag{3}$$

In our proposed model, t-SNE was implemented with 3 components, an optimal perplexity value of 60, a learning rate of 10, and principal component analysis (PCA) initialization in a maximum of 5000 iterations.

C. Data preprocessing

There were some missing values for some features in the dataset; these were considered missing at random. To avoid losing information by removing the missing values, they were imputed. Missing values in nondosimetric numerical features and dosimetric features were imputed by median value, and missing values in sequential features were handled with multiple imputation. Multiple imputation with Bayesian ridge regression was used to impute missing ALC values for weeks 4 and 5 in 2 steps: first, we imputed 9 missing ALC values for week 4; then, we imputed 124 missing ALC values for week 5. The imputation did not change the data distribution and variability.

Due to the uncertainty caused by ALC measurement error, there were some patients with odd weekly ALC trends, which we considered as outliers. Since removing these outliers from the dataset was not feasible because of the size of the dataset, we used Holt's double exponential smoothing (DES) method to remove noise from the data. DES is a popular smoothing method for time series with trends. It assigns exponentially decreasing weights to observations as the observations get older [31]. The smoothing method helps to remove or reduce volatility or other types of noise and allows important patterns to stand out. The following equations were used to determine the smoothed values with this method:

$$F_{t+m} = S_t + mT_t \ , \tag{4}$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \ , \tag{5}$$

$$S_t = \alpha y_t + (1-\alpha)(S_{t-1} + T_{t-1}) \ , \tag{6}$$

where y_t refers to the actual value at time t, F_{t+m}, T_t, and S_t are the forecast for the period $t + m$, the trend estimate, and the exponentially smoothed series, respectively. α denotes the process smoothing constant, and β refers to the trend smoothing constant ($-1 \leq \alpha,\ \beta \leq 1$). DES with a damped trend was done only for patients with unreasonable trends, which was defined by RT experts as having an increased ALC after weeks 2 or 3, or an increase in ALC value greater than 0.1 after weeks 4 or 5. Min-max normalization (i.e., $x = \frac{x - x_{min}}{x_{max} - x_{min}}$) was applied to all features, and the data were split into training and test sets in a ratio of 7:3 (602:258), respectively, using a stratified random sampling scheme.

D. Structure of the hybrid deep-stacked model

The general idea of developing our hybrid stacked model was to combine the knowledge from 4 channels of features and train a meta-model. This was expected to reduce bias and achieve a robust model that reduced the effect of any possible noise caused by the imputation or the randomness of the data. The hybrid deep-stacked model with 4 channels of input (based on the 4 categories of features) was developed to predict the weekly ALC depletion trend during RT using pretreatment information. The first 3 branches of the structure consisted of dense layers that predicted a sequence of 5 weekly ALC values from dosimetric, nondosimetric numerical, and nondosimetric categorical features. The least absolute shrinkage and selection operator (Lasso) regularization and dropout methods were added for nondosimetric categorical features to avoid the adverse effect of sparsity on the prediction and overfitting. The last branch in the structure considered the sequential features, for which we developed an encoder-decoder LSTM network structure to encode the sequential input (i.e., the encoder) and to predict a sequence of weekly ALC values (i.e., the decoder). Since we aimed to make our pretreatment predictions on the basis of pretreatment information only, we only included the baseline ALC value at the beginning of the treatment (i.e., week 0) as an input; thus, we had a one-to-many LSTM structure. Then, all predictions from each branch were concatenated and fed into combined dense layers to predict the weekly ALC values. Fig 1. represents the model structure schema and data preprocessing flow.

The inner connections of the LSTM cells for prediction based on sequential features were based on the following mathematical expressions:

$$f_t = \sigma\left(W_{fh}\ h_{t-1} + W_{fx}\ x_t + b_f\right), \tag{7}$$

$$i_t = \sigma\left(W_{ih}\ h_{t-1} + W_{ix}\ x_t + b_i\right), \tag{8}$$

$$\widetilde{c}_t = \tanh\left(W_{\widetilde{c}h}\ h_{t-1} + W_{\widetilde{c}x}\ x_t + b_{\widetilde{c}}\right), \tag{9}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{c}_t, \tag{10}$$

$$o_t = \sigma\left(W_{oh}\ h_{t-1} + W_{ox}\ x_t + b_o\right), \tag{11}$$

$$h_t = o_t \tanh\left(c_t\right). \tag{12}$$

Where h_{t-1}, x_t, c_{t-1} denote the current hidden state, the input of the cell, and the current cell state vectors of the LSTM, respectively. W and b are the weight matrices and bias vector parameters of each layer in the LSTM cells. f_t is the forget gate's activation vector, which decides what information will be discarded from the cell state. This decision was made by a sigmoid function that returned 1 when it completely kept the information or 0 when it completely discarded all the information. Moreover, $\widetilde{c}_t$, i_t, o_t are the cell input, the input/update gate, and output gate activation vectors, respectively. The input gate decides what information is to be updated (sigmoid function in i_t) and what new information (hyperbolic tangent function in $\widetilde{c}_t$) is to be added and stored in the cell state, and the output gate uses updated cell state and hidden state information to decide what information can be output.

It is feasible to change the treatment plan for RT patients to avoid or mitigate grade 4 RIL. Therefore, the ability to predict lymphopenia at the early stages of RT could help to validate pretreatment predictions or prompt modification

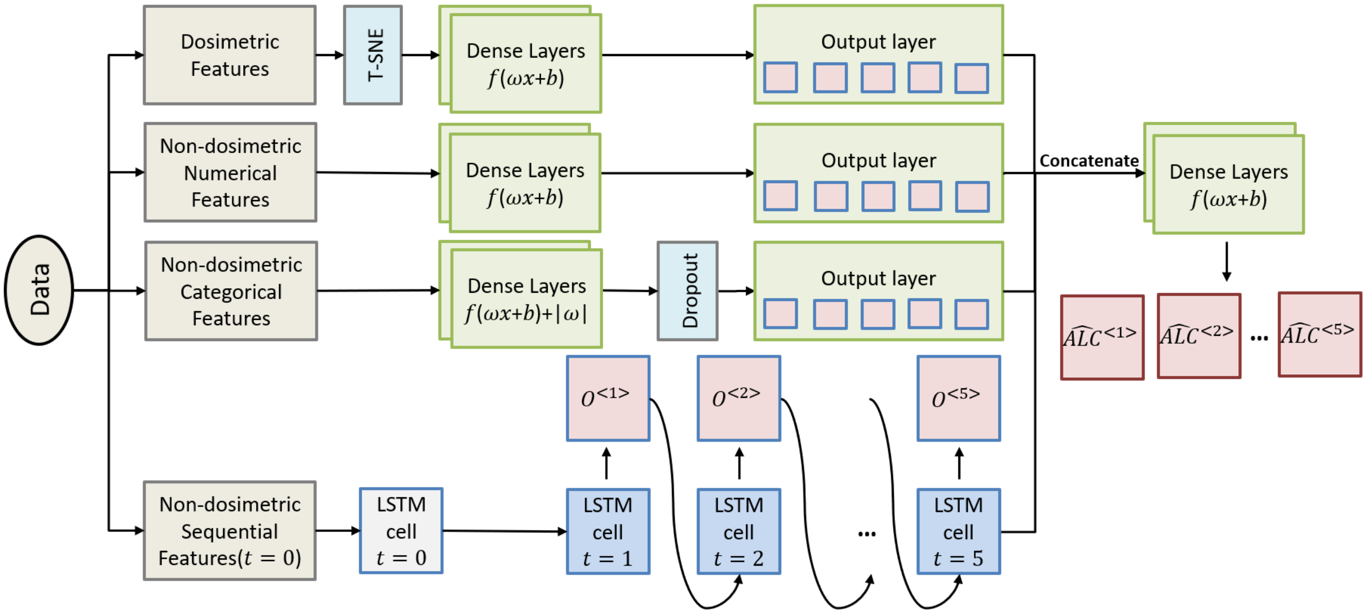


Fig. 1. The model structure for predictions based on pretreatment data

of the treatment plans for high-risk patients without losing much time. Thus, we extended our model to account for early-treatment predictions at the first and second weeks of RT to be used for validation or correction of pretreatment predictions. The model was extended to make predictions at the end of the first week of RT by adding the measured ALC value data for the first week to the LSTM encoder. So, the LSTM structure was modified to have a many-to-many structure. Also, the second week data was added to make predictions at the end of the second week of treatment. For these 2 extended structures, a discriminative layer was added to evaluate the importance weight of each value in the input sequence. The output of this layer was discriminative weights and weighted values of LSTM encoder output, which helped the model to focus on the most relevant part of the input sequence for each output. Fig. 2 and Fig. 3 show the modified structure for predictions after 1 and 2 weeks of treatment, respectively.

E. Training algorithm and model configuration

Algorithm 1 shows the training flow of the proposed models. Each deep learning model was trained with 75 epochs using a batch size of 16 and was implemented using Adam optimizer [32] with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. The L1 regularizer weight of 0.2 and dropout rate of 0.2 were considered for the dense layers of nondosimetric categorical features. The regularization of the hyperparameters of the training set was adjusted empirically until the loss functions of both the training and testing sets declined according to similar trends and without significant gaps between them.

Algorithm 1: Training process

```
Initialization: Initial network weights $\theta_0 \rightarrow \theta$and initial state, size of output sequence $N_{out}$, number of epochs $N_e$, number of batches $N_b$, $e = 1, b = 1$
For $e < N_e$ do
    For $b < N_b$ do
        Feed each feature group to the corresponding branch and obtain the output:
        $Y_{Dose} = f_{Dose}(X_{Dose})$
        $Y_{NonDose} = f_{NonDose}(X_{NonDose})$
        $Y_{NonDoseCat} = f_{NonDoseCat}(X_{NonDoseCat})$
        $X_{encode}, h_0, c_0 = f_{encode}(X_{Seq}, initial\ state)$
        Obtain each LSTM output sequence starting from $i = 1$:
        If $|X_{Seq}| > 1$ then
            $a_{context}, w_0 = f_{\mathrm{discr}}(X_{encode}, h_0)$
            For $i < N_{out}$ do
                $X_{decode}^i, h_i, c_i = f_{decode}(a_{context}, h_{i-1}, c_{i-1})$
                $a_{context}, w_i = f_{\mathrm{discr}}(X_{encode}, h_i)$
            End for
        Else:
            For $i < N_{out}$:
                $X_{decode}^i, h_i = f_{decode}(X_{encode}, h_{i-1})$
            End for
    End for
        $X_{decode} = [X_{decode}^1, X_{decode}^2, \dots, X_{decode}^{N_{out}}]$
        $Y_{out} = f_{\mathrm{final}}(Y_{Dose}, Y_{NonDose}, Y_{NonDoseCat}, X_{decode})$
        Optimize the loss using Adam optimizer [32]
        Calculate the batch total loss using MSE=$\|Y - \hat{Y}\|^2$
        Update network weights $\theta + \Delta\theta \rightarrow \theta$
End for
```

F. Evaluation metrics

In order to evaluate the performance of the proposed models, several important prediction metrics were calculated using predictions for the test data, including mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and explained variance (EV). These evaluation metrics were defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \tag{13}$$

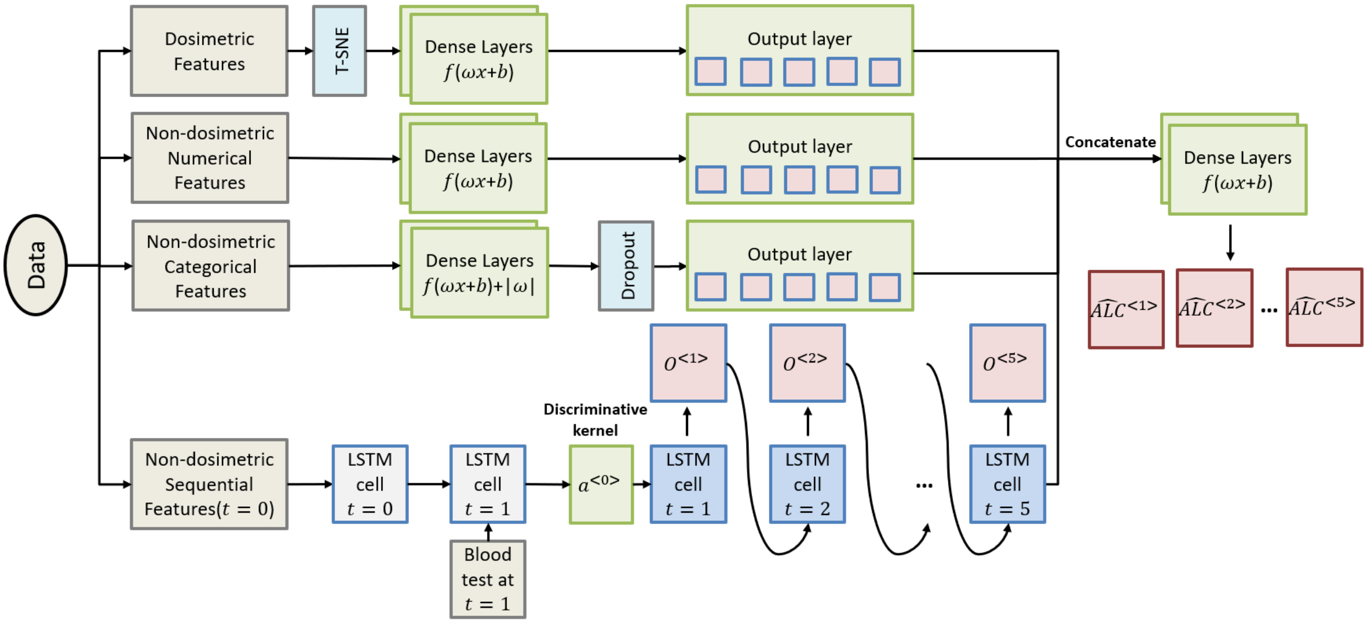


Fig. 2. The model structure for predictions after 1 week of treatment

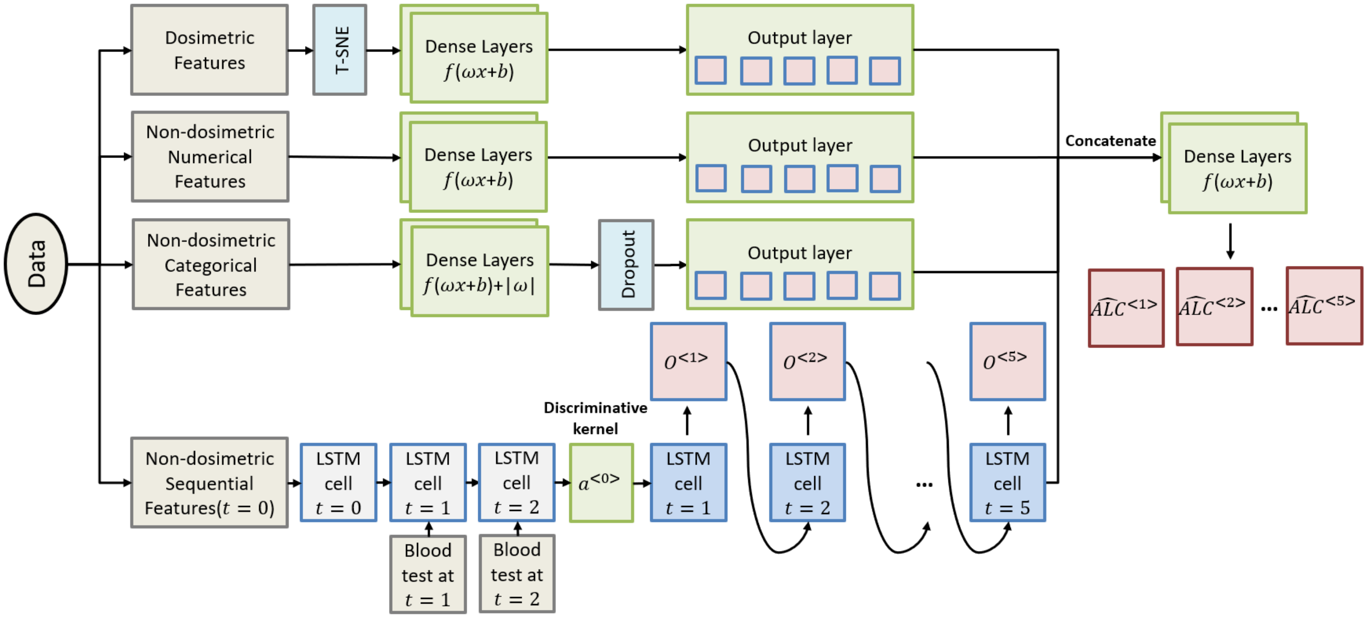


Fig. 3. The model structure for predictions after 2 weeks of treatment

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2}{n}} \qquad (14)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right| \qquad (15)$$

$$EV = 1 - \left(Var(Y_i - \widehat{Y}_i)/Var(Y)\right) \qquad (16)$$

Where Y_i and $\widehat{Y_i}$ are the true values and predicted values, respectively.

G. Comparison models

To evaluate the performance of our proposed hybrid deep-stacked model, we compared the results with other off-the-shelf prediction methods. Support vector machine (SVM), linear regression (LR), LR with lasso regularization (LR-Lasso), LR with ElasticNet (LR-ElasticNet), regression with stochastic gradient descent (SGD), decision tree, extra tree, and random forest models were developed for comparison with the proposed model. The same training and testing sets used for the hybrid deep-stacked model were used for all of these models. The hyperparameters for each model were selected based on a grid search to make the best possible predictions.

III. Results

A. Prediction based on pretreatment data

The proposed model was trained using data from 602 patients in the training set and another 258 patients in the test set. The MSE value of predictions using the baseline hybrid deep-stacked (HDS-t0) model for the 258 patients in the test set was 0.046. Fig. 4 shows the true ALC depletion trends versus the predicted curves for 15 randomly selected patients in the test set. As shown in the figure, the model provided accurate predictions with only small errors; these predictions are therefore suitable for use in pretreatment analysis to opt out patients at high risk of significant ALC reduction after RT.

Table I compares the performance metrics of 8 prediction models: SVM, LR, LR-Lasso, LR-ElasticNet, SGD, decision tree, extra tree, random forest, and the proposed hybrid model for pretreatment predictions (HDS-t0). As we can see from the table, LR achieved the best results among the 8 prediction methods, with the lowest MSE, 0.0657. The proposed HDS-t0 model outperformed the LR model, with a reduction of 30.6% in the MSE of the predicted values. Our model also improved upon several other metrics, including the normalized RMSE (NRMSE) (−16.8%), MAE (−5.82%), and EV (+21.2%), compared to the best off-the-shelf model (i.e., LR).

TABLE I
Comparison of prediction performance metrics (MSE, NRMSE, MAE, and EV) ofeight common prediction models and the proposed HDS-t0 model for predictions based on pretreatment information

	MSE	NRMSE	MAE	EV
HDS	0.0456	0.0332	0.1457	0.7260
SVM	0.0736	0.0422	0.1635	0.5552
Random Forest	0.0698	0.0411	0.1596	0.5745
LR-Elastic Net	0.0859	0.0456	0.1771	0.4764
Decision Tree	0.0776	0.0433	0.1716	0.5269
Extra Tree	0.0691	0.0409	0.1593	0.5787
SGD	0.0839	0.0450	0.1722	0.4891
LR-Lasso	0.0859	0.0456	0.1771	0.4764
Linear Regression	0.0657	0.0399	0.1547	0.5992

B. Predictions after 1 and 2 weeks of treatment

As discussed in section II.D, the proposed hybrid deep-stacked model was extended to use the ALC value obtained after 1 week of RT to forecast an ALC trend for the rest of the treatment. Fig. 5 shows the true ALC depletion trends versus the predicted curves for the same 15 patients in the test set. As shown in the figure, the extended model, HDS-t1, provided more accurate predictions than did the HDS-t0 model. The HDS-t1 model achieved an MSE value of 0.014 for the test set predictions, a reduction of 69.6% compared to the HDS-t0 model. These results suggest that data from early stages of RT can be used to estimate the final patient response to treatment with more confidence than the pretreatment data. Therefore, our model's predictions after the first week can be used to validate the pretreatment predictions or modify the treatment plan for patients at high risk of grade 4 RIL during the early stages of treatment.

To evaluate the effect of collecting more data during treatment on our model's ALC predictions, we also predicted the future ALC values after 2 weeks of treatment using the same test set with the HDS-t2 model. Fig. 6 shows the true ALC depletion trends and the predicted curves using the HDS-t2 model for the same 15 patients in the test set.

Prediction metrics were calculated for the predictions based on pretreatment data using the HDS-t0 model and data from after the first and second weeks of treatment using the HDS-t1 and HDS-t2 model structures, respectively. Table II summarizes the MSE, NRMSE, MAE, and EV of each model for the predicted weekly ALC values of patients in the test set. As shown in the table, using the first-week data reduced the MSE of predictions by 69.6% compared to the model based on pretreatment data. Moreover, adding second-week data improved the MSE value by 42.9% over the HDS-t1 model and 82.6% over the HDS-t0 model. Therefore, we can conclude that augmenting the model with the first-week treatment data can significantly improve the pretreatment predictions, although the magnitude of improvement is smaller when additional weekly data (i.e., week 2 data) are included. This suggests that the difference between the baseline ALC value and that measured after the first week can provide very useful information to predict the ALC trend during the rest of treatment. By updating the predictions with the measured ALC after the first week of treatment, we can reduce the risk of false-negative pretreatment risk predictions. Also, physicians could use this method to validate the pretreatment predictions, update treatment plans, or develop mitigation strategies. This model could be also used after delivery of 1 fraction of radiation instead of after 1 week to reduce time and cost.

TABLE II
Comparison of prediction performance metrics (MSE, NRMSE, MAE, and EV) for predictions based on HDS-t0, HDS-t1, and HDS-t2 models

	MSE	NRMSE	MAE	EV
HDS-t0	0.046	0.033	0.146	0.726
HDS-t1	0.014	0.018	0.069	0.917
HDS-t2	0.008	0.014	0.046	0.954

Fig. 7 and Fig. 8 show the scatter plots and distributions of weekly ALC values for real and predicted values based on each model. As shown in these figures, the distribution of predicted values using the HDS-t1 model was closer to the distribution of real values than were the pretreatment predictions using the HDS-t0 model. Similarly, the HDS-t2 model achieved more accurate predictions than the HDS-t1 model. Fig. 9 represents the boxplot of the residual values (i.e., $ALC_i - \widehat{ALC_i}$) normalized by the mean ALC value within each week. This figure shows that the median of the error (i.e., the normalized residual value) within each week was almost zero for all 3 models, which suggests that the models performed well to predict weekly ALC for more than 50% of the patients in the test set. Moreover, the range and interquartile range of the error in weeks 3 to 5 was the lowest for the HDS-t2 model, and the HDS-t1 model showed lower error values than the HDS-t0 model. This result is in agreement with our previous results comparing the models' performance.

The minimum ALC value during treatment, known as the ALC nadir, is an important factor in determining the occur-

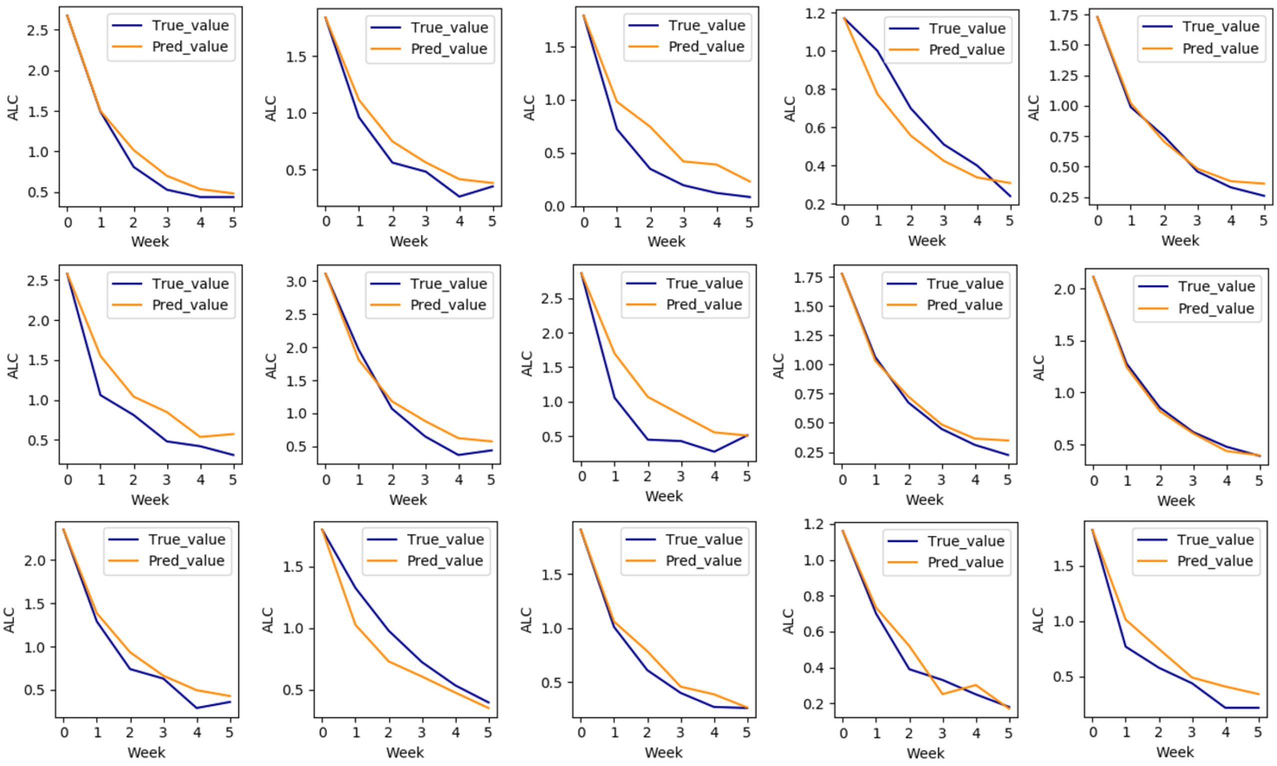


Fig. 4. Predicted ALC trends using the HDS-t0 model (orange) versus the real values (blue) for 15 randomly selected patients in the test set

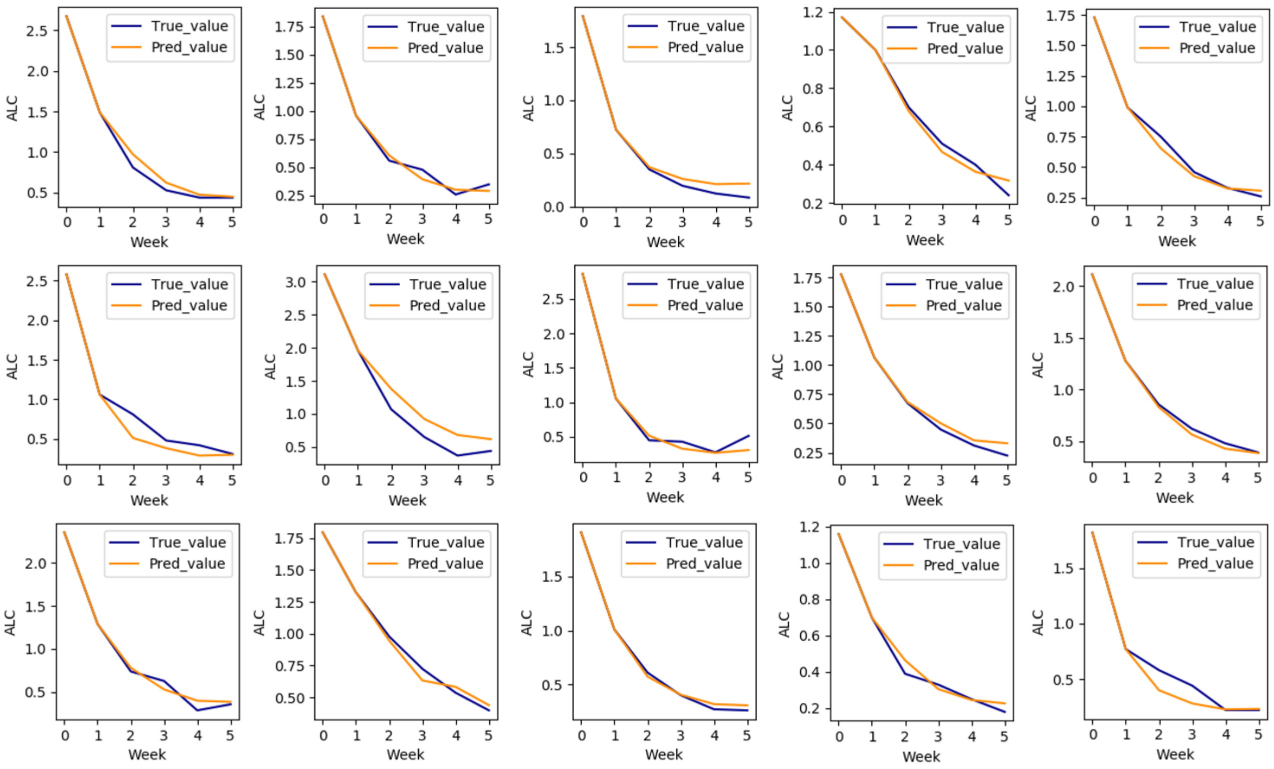


Fig. 5. Predicted ALC trends based on the HDS-t1 model (orange) versus real values (blue) for 15 randomly selected patients in the test set

rence of grade 4 RIL for RT patients. Thus, we determined the minimum ALC value during the 5 weeks of treatment for the real data and 3 predictions. Fig. 10 shows the histogram, boxplot, and kernel density estimation of the ALC nadir during 5 weeks of treatment based on the real values and the predicted values from the HDS-t0, HDS-t1, and HDS-t2 models. As

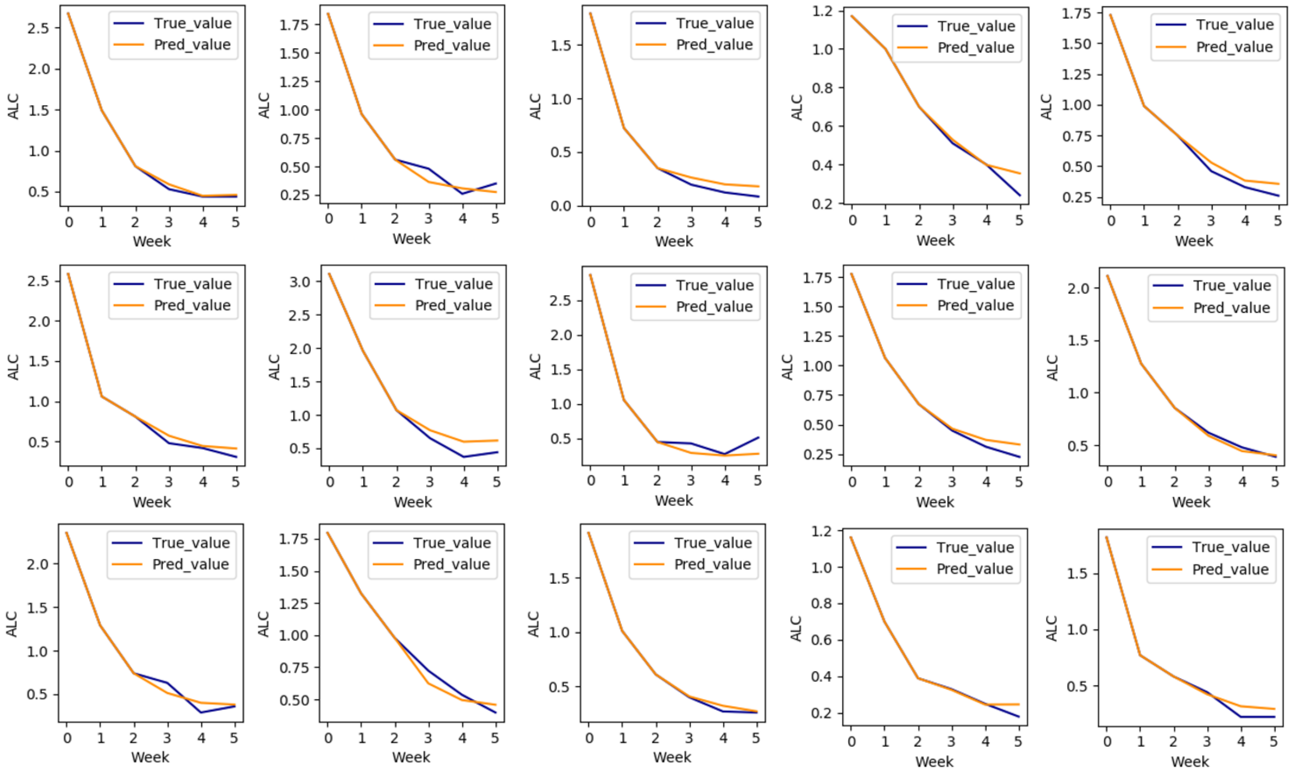


Fig. 6. Predicted ALC trends based on the HDS-t2 model (orange) versus real values (blue) for 15 randomly selected patients in the test set

shown in the figure, the HDS-t2 model achieved the best predictions and the most similar distribution to the real data in terms of the range (real: 1.722, HDS-t2: 1.915), mean (real: 0.298, HDS-t2: 0.315), median (real: 0.25, HDS-t2: 0.318), interquartile range (real: 0.189, HDS-t2: 0.142), and kernel density estimation.

For the 2 extended models, HDS-t1 and HDS-t2, a discriminative layer was added to evaluate the importance weight of each value in the input sequence. The importance weights of each data value in the input sequence for each predicted value in the output sequence were obtained from this discriminative layer. Using the HDS-t1 model, the input sequence contained the ALC value at the baseline (i.e., week 0) and after the first week of treatment (i.e., week 1), and the output sequence was the predicted ALC values at the ends of weeks 2 to 5.

Fig. 11 (a) shows the importance weights for each input-output pair of the HDS-t1 model. As shown in the figure, the importance of baseline ALC was higher for the predicted ALC values of the last weeks of treatment (i.e., weeks 4 and 5) than the early ones (i.e., weeks 2 and 3). Likewise, Fig. 11 (b) presents the importance weights for each input-output pair based on the obtained results from the HDS-t2 model, which suggest the same conclusion as HDS-t1. These results suggest that the models could capture long-term as well as short-term dependencies.

IV. Conclusion

In this paper, a hybrid deep-stacked model is proposed to predict RT-induced lymphocyte depletion for esophageal cancer patients during the course of RT. The proposed stacked structure used 4 categories of features in 4 branches, which reduced the bias and adverse effects of any possible noise in the data. The model was extended to account for predictions made after the initial part of treatment (i.e., at the end of weeks 1 or 2), and a discriminative kernel layer was developed to evaluate the importance weight of each value in the input sequence. In order to evaluate the performance of the proposed models, important prediction metrics were compared with those from 8 off-the-shelf prediction methods. The results showed that the proposed model outperformed these off-the-shelf prediction methods in predicting weekly ALC values. Moreover, using the extended model based on the first-week data reduced the MSE of predictions compared to the model based on the pretreatment data. We conclude that augmenting the model with data from early stages of treatment (i.e., weeks 1 or 2) can significantly improve ALC predictions. Therefore, the HDS-t0 model using pretreatment data can be used in RT treatment planning to predict lymphocyte depletion during the course of RT. This prediction will help to select patients for RT and develop lymphopenia-mitigating strategies to ultimately improve patients' survival. After treatment is started, further predictions in the early stages of treatment can be used to validate the pretreatment predictions and, if necessary, modify the treatment plan for high-risk patients in order to preserve lymphocytes.

Our proposed deep learning framework is flexible and transferrable to other, related toxic effects of RT. It also can be used after delivering 1 fraction of radiation instead of after 1

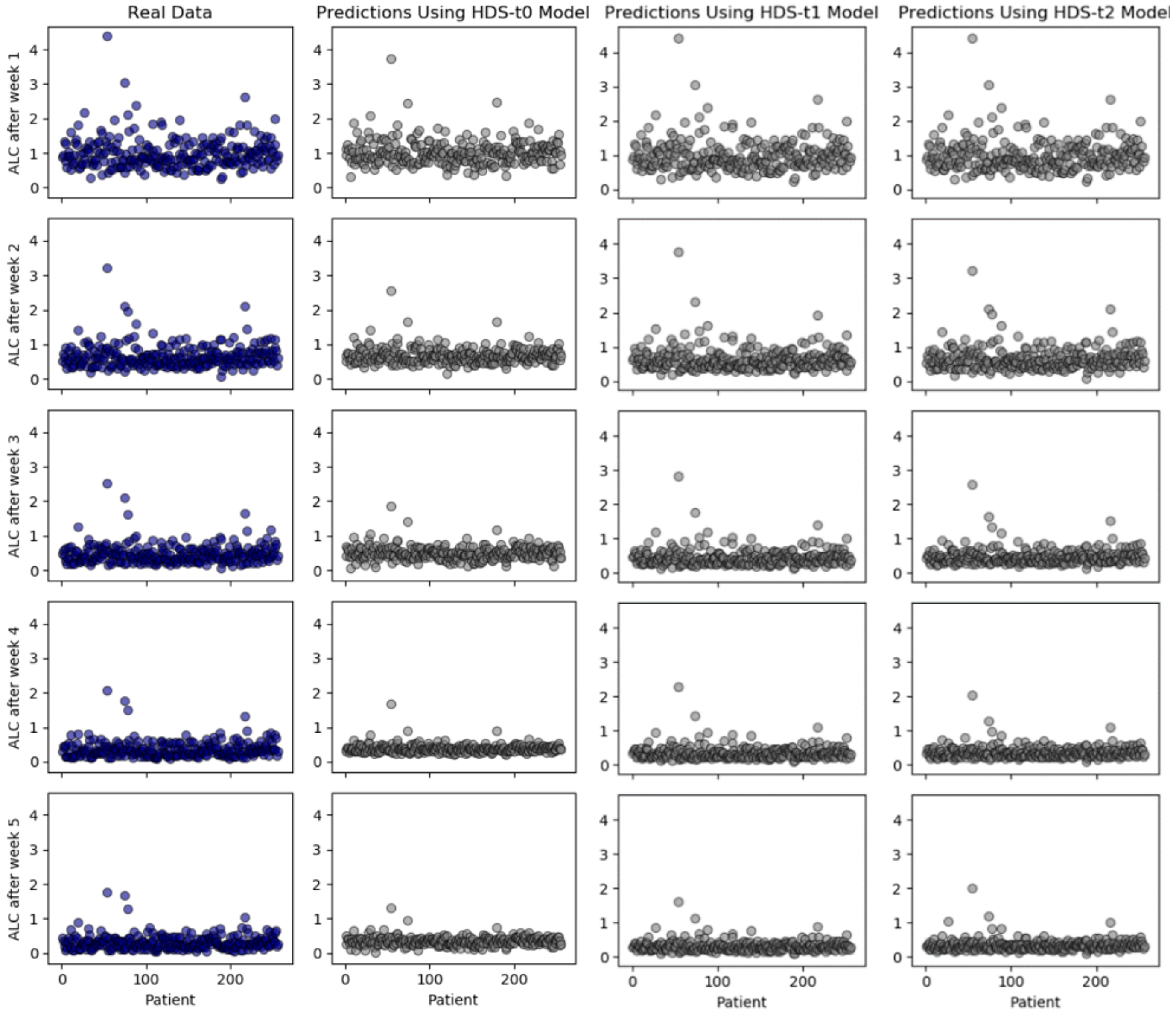


Fig. 7. Scatter plots of weekly ALC values for real and predicted values from each model

week by using fraction-based clinical data to reduce time and cost. For future work, further analysis and discussions will be made based on different patient profiles to investigate the impact of different clinical factors on RIL and to estimate the risk associated with each predicted ALC value.

References

[1] I. S. Stafford, M. Kellermann, E. Mossotto, R. M. Beattie, B. D. MacArthur, and S. Ennis, "A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases," *npj Digital Medicine*, vol. 3, no. 1, pp. 30–30, 2020. [Online]. Available: 10.1038/s41746-020-0229-3;https://dx.doi.org/10.1038/s41746-020-0229-3

[2] C. K. Fisher, A. M. Smith, and J. R. Walsh, "Machine learning for comprehensive forecasting of Alzheimer's Disease progression," *Sci. Rep*, vol. 9, no. 1, 2019.

[3] S. Yousefi, "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models," *Sci. Rep*, vol. 7, no. 1, 2017.

[4] B. C. Munsell, "Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data," *NeuroImage*, vol. 118, pp. 219–230, 2015.

[5] J. Weiss, F. Kuusisto, K. Boyd, J. Liu, and D. Page, "Machine Learning for Treatment Assignment: Improving Individualized Risk Attribution," *AMIA Annu. Symp. Proc. AMIA Symp*, vol. 2015, pp. 1306–1315, 2015.

[6] G. J. Lim, L. Kardar, S. Ebrahimi, and W. Cao, "A risk-based modeling approach for radiation therapy treatment planning under tumor shrinkage uncertainty," *European Journal of Operational Research*, vol. 280, no. 1, pp. 266–278, 2020. [Online]. Available: 10.1016/j.ejor.2019.06.041;https://dx.doi.org/10.1016/j.ejor.2019.06.041

[7] S. G. Ellsworth, "Field size effects on the risk and severity of treatment-induced lymphopenia in patients undergoing radiation therapy for solid tumors," *Advances in Radiation Oncology*, vol. 3, no. 4, pp. 512–519, 2018. [Online]. Available: 10.1016/j.adro.2018.08.014;https://dx.doi.org/10.1016/j.adro.2018.08.014

[8] S. A. Grossman, "Immunosuppression in Patients with High-Grade Gliomas Treated with Radiation and Temozolomide," *Clin. Cancer Res*, vol. 17, no. 16, pp. 5473–5480, 2011.

[9] J. L. Campian, X. Ye, M. Brock, and S. A. Grossman, "Treatment-related Lymphopenia in Patients With Stage III Non-Small-Cell Lung Cancer," *Cancer Investigation*, vol. 31, no. 3, pp. 183–

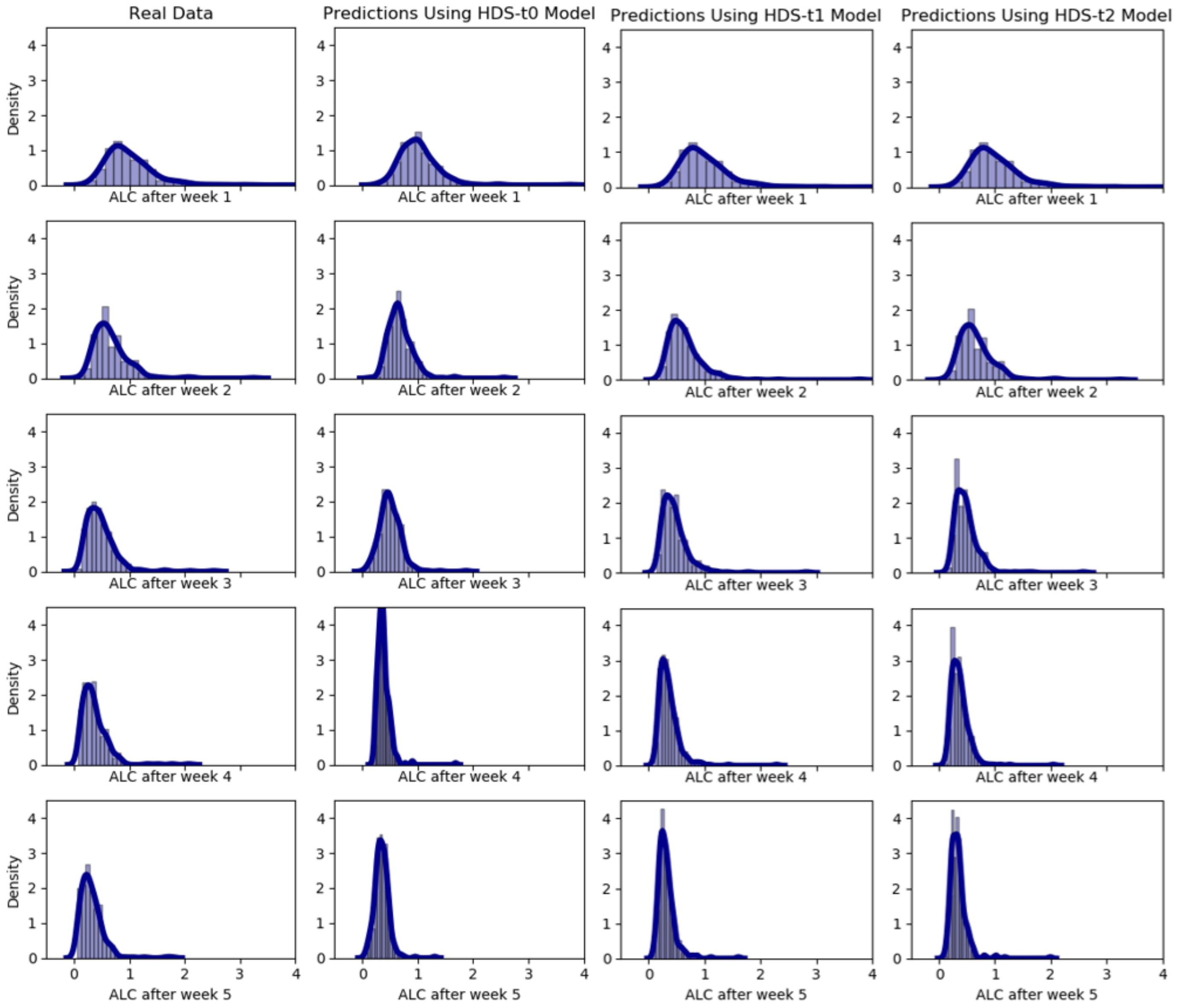


Fig. 8. The distribution of weekly ALC values for real and predicted values based on each model

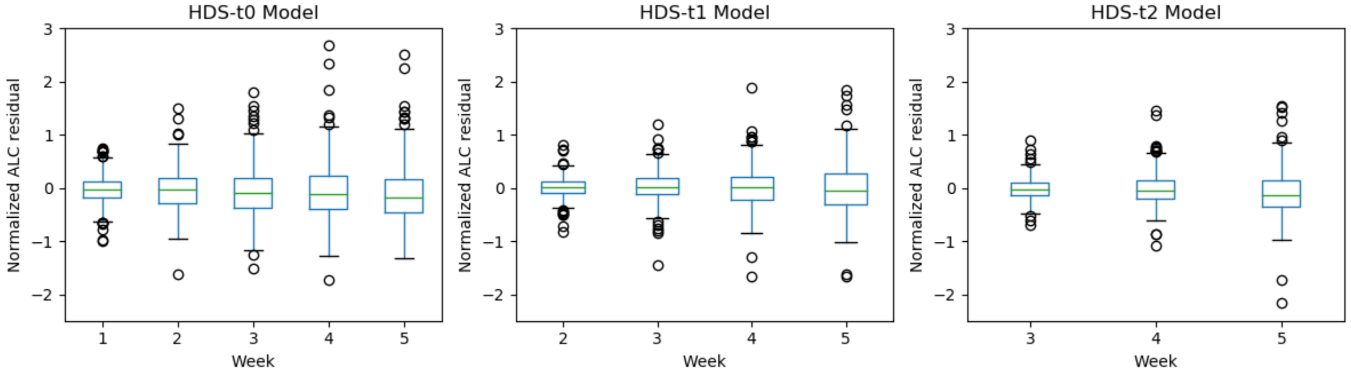
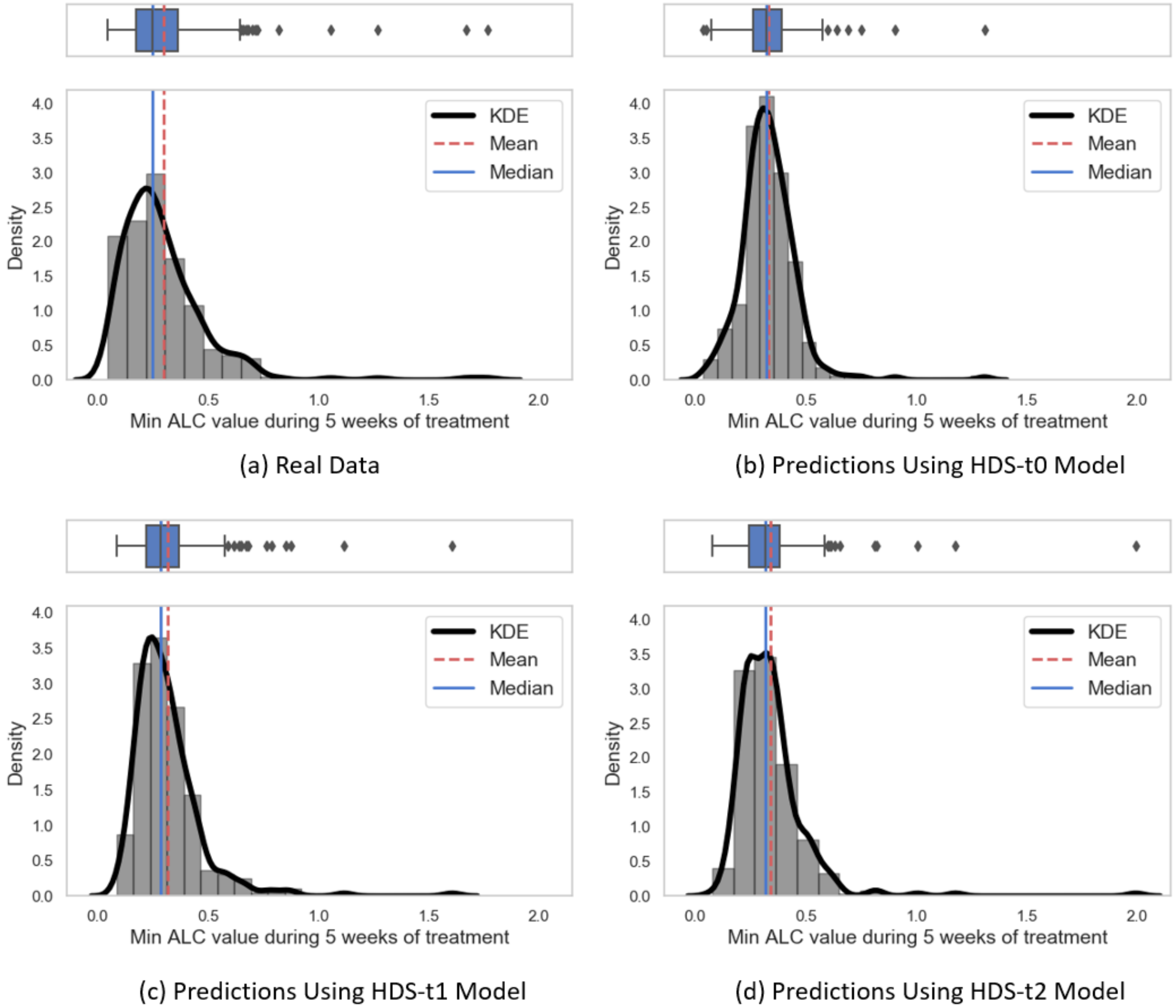


Fig. 9. Boxplots of predicted ALC values normalized to the mean ALC value for each week based on each model

188, 2013. [Online]. Available: 10.3109/07357907.2013.767342;https: //dx.doi.org/10.3109/07357907.2013.767342

[10] B. P. Venkatesulu, S. Mallick, S. H. Lin, and S. Krishnan, “A systematic review of the influence of radiation-induced lymphopenia on survival



(a) Real Data

(b) Predictions Using HDS-t0 Model

(c) Predictions Using HDS-t1 Model

(d) Predictions Using HDS-t2 Model

Fig. 10. Histograms, boxplots, and kernel density estimations (KDE) of minimum ALC values during 5 weeks of treatment based on the real and predicted values using HDS-t0, HDS-t1, and HDS-t2 models

outcomes in solid tumors," *Critical Reviews in Oncology/Hematology*, vol. 123, pp. 42–51, 2018. [Online]. Available: 10.1016/j.critrevonc.2018.01.003;https://dx.doi.org/10.1016/j.critrevonc.2018.01.003

[11] R. Davuluri, "Lymphocyte Nadir and Esophageal Cancer Survival Outcomes After Chemoradiation Therapy," *Int. J. Radiat. Oncol. Biol. Phys*, vol. 99, no. 1, pp. 2017–2017.

[12] Y. Shiraishi, "Severe lymphopenia during neoadjuvant chemoradiation for esophageal cancer: A propensity matched analysis of the relative risk of proton versus photon-based radiation therapy," *Radiother. Oncol*, vol. 128, no. 1, pp. 154–160, 2018.

[13] P. S. N. V. Rossum, "Prediction of Severe Lymphopenia During Chemoradiation Therapy for Esophageal Cancer: Development and Validation of a Pretreatment Nomogram," *Pract. Radiat. Oncol*, vol. 10, no. 1, pp. 16–26, 2020.

[14] S. G. Ellsworth, A. Yalamanchali, H. Zhang, S. A. Grossman, R. Hobbs, and J.-Y. Jin, "Comprehensive Analysis of the Kinetics of Radiation-Induced Lymphocyte Loss in Patients Treated with External Beam Radiation Therapy," *Radiation Research*, vol. 193, no. 1, pp. 73–73, 2019. [Online]. Available: 10.1667/rr15367.1;https://dx.doi.org/10.1667/rr15367.1

[15] C. Zhu, "A novel deep learning model using dosimetric and clinical information for grade 4 radiotherapy-induced lymphopenia prediction," *Phys. Med. Biol*, vol. 65, no. 3, pp. 35 014–35 014, 2020.

[16] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Comput*, vol. 31, no. 7, pp. 1235–1270, 2019.

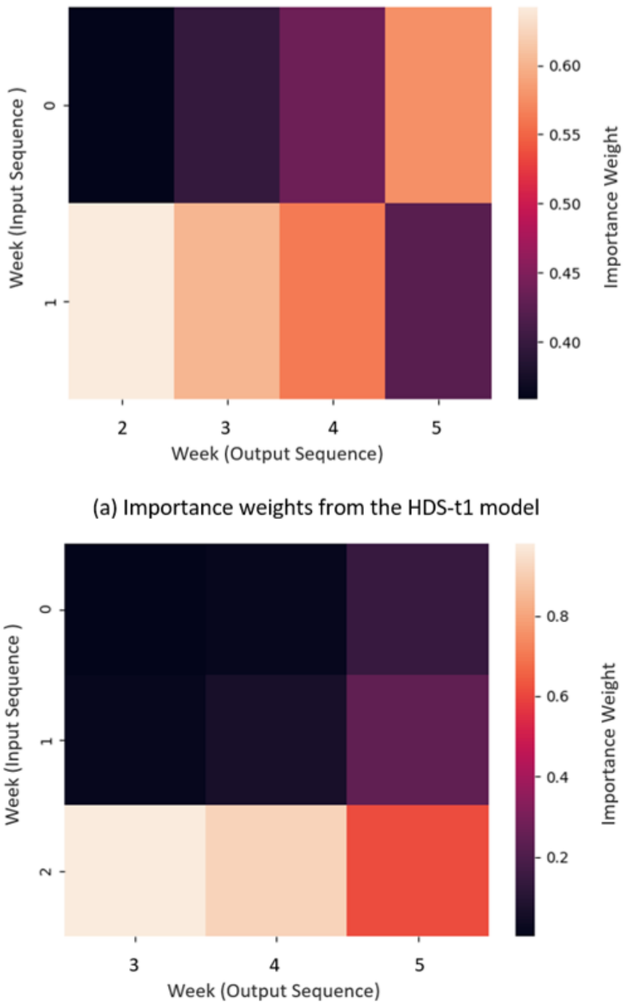
[17] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241–265, 2018. [Online]. Available: 10.1016/j.ymssp.2017.11.024;https://dx.doi.org/10.1016/j.ymssp.2017.11.024

[18] J. L. Elman, "Finding Structure in Time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990. [Online]. Available: 10.1207/s15516709cog1402_1;https://dx.doi.org/10.1207/s15516709cog1402_1

[19] B. Tung, V.-W. Chen, and Soo, "A comparative study of recurrent neural network architectures on learning temporal sequences," *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 4, pp. 1945–1950, 1996.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: 10.1162/neco.1997.9.8.1735;https://dx.doi.org/10.1162/neco.1997.9.8.1735

[21] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing*



(a) Importance weights from the HDS-t1 model

(b) Importance weights from the HDS-t2 model

Fig. 11. Heatmaps of importance weights for each input-output pair based on the obtained results from the HDS-t1 (a) and HDS-t2 (b) models

Research, vol. 9, no. 4, pp. 235–245, 2019. [Online]. Available: 10.2478/jaiscr-2019-0006;https://dx.doi.org/10.2478/jaiscr-2019-0006

[22] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[23] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 988–997.

[24] X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, and X. Niu, "Image captioning with triple-attention and stack parallel LSTM," *Neurocomputing*, vol. 319, pp. 55–65, 2018.

[25] S. Dai, L. Li, and Z. Li, "Modeling Vehicle Interactions via Modified LSTM Models for Trajectory Prediction," *IEEE Access*, vol. 7, pp. 38 287–38 296, 2019.

[26] F. Altché and A. D. L. Fortelle, "An LSTM network for highway trajectory prediction," *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 353–359, 2017.

[27] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 117–125.

[28] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 694–707, 2016. [Online]. Available: 10.1109/taslp.2016.2520371;https://dx.doi.org/10.1109/taslp.2016.2520371

[29] G. Zhong, K. Zhang, H. Wei, Y. Zheng, and J. Dong, "Marginal Deep Architecture: Stacking Feature Learning Modules to Build Deep Learning Models," *IEEE Access*, vol. 7, pp. 30 220–30 233, 2019. [Online]. Available: 10.1109/access.2019.2902631;https://dx.doi.org/10.1109/access.2019.2902631

[30] M. Jiang, J. Liu, L. Zhang, and C. Liu, "An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms," *Phys. Stat. Mech. Its Appl*, vol. 541, pp. 122 272–122 272, 2020.

[31] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, 2004. [Online]. Available: 10.1016/j.ijforecast.2003.09.015;https://dx.doi.org/10.1016/j.ijforecast.2003.09.015

[32] D. P. Kingma and J. Ba, 2017. [Online]. Available: http://arxiv.org/abs/1412.6980