Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes

Yaohua Wang
Electrical and Computer Engineering
University of Iowa
USA

Lisanne Van Dijk
Radiation Oncology
UT M.D. Anderson Cancer Center
USA

Abdallah S. R. Mohamed
Radiation Oncology
UT M.D. Anderson Cancer Center
USA

Clifton David Fuller
Radiation Oncology
UT M.D. Anderson Cancer Center
USA

Xinhua Zhang
Computer Science
University of Illinois at Chicago
USA

G. Elisabeta Marai
Computer Science
University of Illinois at Chicago
USA

Guadalupe Canahuate
Electrical and Computer Engineering
University of Iowa
USA

ABSTRACT

Patient-Reported Outcome (PRO) surveys are used to monitor patients' symptoms during and after cancer treatment. Late symptoms refer to those experienced after treatment. While most patients experience severe symptoms during treatment, these usually subside in the late stage. However, for some patients, late toxicities persist negatively affecting the patient's quality of life (QoL). In the case of head and neck cancer patients, PRO surveys are recorded every week during the patient's visit to the clinic and at different follow-up times after the treatment has concluded. In this paper, we model the PRO data as a time-series and apply Long-Short Term Memory (LSTM) neural networks for predicting symptom severity in the late stage. The PRO data used in this project corresponds to MD Anderson Symptom Inventory (MDASI) questionnaires collected from head and neck cancer patients treated at the MD Anderson Cancer Center. We show that the LSTM model is effective in predicting symptom ratings under the RMSE and NRMSE metrics. Our experiments show that the LSTM model also outperforms other machine learning models and time-series prediction models for these data.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

Long Short-Term Memory (LSTM), Patient Reported Outcomes (PRO), Late Toxicity, Symptom Severity Prediction

1 INTRODUCTION

During head and neck cancer treatment, patients may experience different symptoms with different severity during and after treatment [4, 27, 28]. A commonly used way to monitor patients' symptoms is to record symptom severity or occurrence through a questionnaire survey, which is commonly known as Patient-Reported Outcomes (PRO) data. Much research is done over these PRO data in order to identify symptoms in an early stage and guide treatment decisions, as well as investigating the relationships among these symptoms [19, 24]. In this work, we use data collected from head and neck cancer patients treated at the M.D. Anderson Cancer Center using the M.D. Anderson Symptom Inventory questionnaire [5] and more specifically, the Head-Neck Module (MDASI-HN) [22]. The module is comprised of 28 questions, 13 referring to core symptoms related to cancer (systemic), 9 to head and neck symptoms (local), and the remaining 6 to symptom-burden interference with daily activities (life general). Patients rated the severity of their symptoms on a scale of 0 to 10 with 0 being mild or no existence and 10 being very severe (the worst imaginable). During treatment, symptoms are experienced with greater severity than after treatment. Ideally, we would like to see that all symptoms have receded in the late stage (e.g. a year after treatment), but in some cases, symptoms persist affecting the Quality-of-Life (QoL) of the patients in the long term.

Previous research over the MDASI-HN PRO data applies factor analysis and cluster analysis to cluster and investigate symptom progression [25]. These researches look at a particular snapshot in time to cluster either the patients, by their experienced symptoms, or the symptoms given the patient's ratings. In this work, we

approached the problem from a different perspective by modeling the PRO data as a time-series and applying the Long-Short Term Memory (LSTM) Neural Network model [12] to predict late symptom's rating 6 weeks and 12 months after treatment. Since PRO data is self-reported, patients may skip questions or entire questionnaires altogether, resulting in many missing values. To overcome this issue, we applied several methods for missing data imputation and evaluate the performance of the LSTM model for each method. We show that the LSTM model is effective in predicting symptom ratings under the RMSE and NRMSE metrics. In our experiments, the LSTM model also outperforms other machine learning models. Furthermore, we show that for this particular task as it has been observed in other domains, the more data used, the better the model obtained, even when the data needs to be imputed.

The main contributions of this paper can be summarized as follows. This is the first work that looks at predicting late toxicity for head and neck cancer patients using LSTM. We evaluate different imputation methods for completing the data, including applying LSTM recursively. We compare the LSTM performance against ARIMA and other machine learning models.

2 RELATED WORK

MDASI-HN PRO Data. PRO data has been widely collected physically and electronically in the clinical area since it has an important meaning of evaluating the treatment benefits [6]. The PRO data used in this project is an MDASI-HN [22] questionnaire with 28 symptoms to be rated on a scale from 0 to 10 with 0 being mild or no symptoms and 10 being very severe, during and after treatment. As shown in Table 1, the 28 symptoms can be divided into three types of toxicity. All patients are asked to fill MDASI-HN surveys before the start of treatment (baseline) and then weekly for the 6 weeks of the duration of the treatment. Patients are also asked at their follow-up visits 6 weeks, 6 months, and 12 months after treatment.

Using the MDASI-HN PRO data, several studies focus on identifying symptom clusters at a single timepoint [10, 14, 23]. Prior research mainly used two methods to find symptom clusters, one is factor analysis such as principal component analysis and the other one is cluster analysis such as hierarchical agglomerative clustering [2, 7, 11, 25]. These studies focus on a single time point analysis, whereas we model the PRO data as a time series.
Time Series Prediction and Imputation. Prediction of time series is typically done by looking at the previous values in the series and deciding the value at the current time step. Auto-regressive Integrated Moving Average (ARIMA) [1] is a commonly used method for the prediction of time series. The model combines the Auto-regressive (AR) and Moving Average (MA) models that are suitable for univariate time series modeling. In the AR model, the output depends on its lags while in the MA model, the output depends only on the lagged forecast errors. More recently, Long short term memory (LSTM) Recurrent Neural Networks have gained more popularity in time series prediction [9] and healthcare domain [15]. Specifically, LSTM networks were used to mimic the pathologist decision and other diagnostic applications [18, 30]; LSTM networks were used to recognize sleep patterns in multi-variate time-series

Toxicity	Symptoms
Systemic	fatigue, constipation, nausea, sleep, memory, appetite, drowsy, vomit, numb
Local	pain, mucus, swallow, choke, voice, skin, taste, mucositis, teeth, shortness of breath (SOB), dry mouth
Life general	general activity, mood, work, relations, walking, enjoy, distress, sad

Table 1: The 28 symptoms in the MDASI-HN questionnaire grouped into three types of toxicities.

clinical measurements [16]. However, to the best of our knowledge, it has not been previously applied to PRO data.

Data imputation methods such as Multiple Imputation by Chained Equations (MICE) [3, 21], linear regression, Kalman filtering [13], among others, can be used to impute time series data.

3 PROPOSED APPROACH

In this section, we first describe the methodological approach including data pre-processing and the methods used for data imputation.

3.1 Long Short Term Memory (LSTM)

Since PRO data with patients self-reporting on the severity of their symptoms is collected over time, we model it as a time series. If we can learn from the patients' answers over a period of time and predict their answers next week or in 6 months, we could proactively make recommendations to minimize the symptom's burden and therefore, improve the patients' quality of life. For example, we can record patients' responses to different symptoms from week 0 to week 5 during the treatment and predict what the ratings for those symptoms would be in week 6. Such prediction, if accurate, can be useful to make patients aware of the risks and prescribe exercises or medication that can help patients cope with the symptoms to avoid having to adjust treatment and improve the long-term quality of life of the patients.

Long short term memory (LSTM) neural networks are a type of recurrent neural network (RNN) proven effective in predicting time series [12]. Unlike a traditional neural network, LSTMs have a feedback structure to store the memory of the events happened in the past and use it as a parameter in prediction. The basic structure of an LSTM model takes 3 different pieces of information: the current input data, the short-term memory (hidden states) from the previous cell, and the long-term memory (cell state). This 3-dimensional data structure (number of samples, number of time steps, number of parallel time series on features) is pushed through the LSTM gates, which are used to regulate the information to be kept or discarded, i.e. selectively remove any irrelevant information. The LSTM model is able to memorize the time-series pattern of each patient's response and be able to predict late toxicity. Another advantage of the LSTM is the diversity of the inputs and outputs. LSTM can handle multiple predictions simultaneously. In a many-to-one mode, the LSTM would learn from many patients and predict one symptom. In a many-to-many mode, the LSTM would learn from many patients and predict all 28 symptoms for the test data.

Taking advantage of this, we are able to generate predictions for all 28 symptoms using one trained LSTM with many-to-many mode.

To feed the data into the LSTM model, the original data was transformed into a 3-dimensional array where the first dimension corresponds to the patients, the second dimension to the time steps, and the third dimension to the symptoms. The number of patients corresponds to the number of samples in the training data. The number of time steps depends on what late toxicity we are evaluating. For the toxicity at 6 weeks after treatment, the number of time steps is 6, while for the toxicity at 12 months after treatment, the number of time steps is 11. The number of symptoms is always 28.

3.2 Data Imputation

Many of the symptom severity scores in the PRO data have NaN values. Some patients do not have enough follow-up time to collect later time points. Like many other machine learning models, LSTM requires complete data. As a proof of concept, we first consider three imputation methods: Linear interpolation, Kalman interpolation [13], and Multiple Imputation by Chained Equations (MICE) [3].

Both Linear and Kalman interpolations are uni-variable imputations. The first non-NA value was replicated to the start of the time series and the last-NA value was replicated to the end of the time series. Kalman smoothing requires at least three observations. When the time series contained less than 3 observations, the spline method was used to interpolate. Since the MDASI-HN ratings range from 0 to 10, imputed values that resulted in negative values were replaced by 0, and values larger than 10 were replaced by 10. Before applying the Kalman smoothing, the time series for the patient was scaled between 0 and 1, imputation applied, and then the data was scaled back to the original 0-10 scale.

MICE [3] is a multivariate imputation so the time series for all the patients are used simultaneously. That is, through an iterative series of predictive models, each specified variable in the data set is imputed using the other variables in the data set. The predictive model used to impute values can be various. In this work, we used predictive mean matching (pmm) where for each missing entry, the method forms a small set of candidates from all complete cases that have predicted values that is close to the predicted value for the missing entry and select a random candidate from the set to replace the missing value. Besides, we did 5 iterations for the imputation.

We also apply the LSTM model recursively to predict intermediate time points and then use the predicted data to train the next time step in the model.

All the data imputations described were done using only the PRO time-series data without considering any other clinical variables such as gender, age, cancer staging, or treatment.

Each imputation method produces a different complete version of the dataset. The complete data was then transformed into the correct input size of the LSTM model, and an LSTM model was trained on each of the imputed datasets. The predictions for all 28 symptoms were then compared using the Root Mean Squared Error (RMSE) and Normalized RMSE (NRMSE) metrics as defined below.

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)} \qquad (1)$$

where $\hat{\theta}$ is the vector of observed values of the variable being predicted and θ being the predicted values and E is the expected value.

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \qquad (2)$$

where y_{max} and y_{min} are the maximum and minimum of actual data.

4 EXPERIMENTAL RESULTS

4.1 Experimental Setup

MDASI questionnaires were collected from 823 patients weekly for 6-weeks during treatment and for 3-time points after treatment (6-weeks, 6 Months, and 12 Months). The original data was split into two series: from baseline to 6-weeks after treatment, and from baseline to 12-month after treatment. We then applied the three interpolation methods on the two versions of the data and generated 6 imputed data sets. The time point to be predicted was not imputed and patients with missing surveys for that time point were excluded from the model. For testing, we only considered patients with complete data, that is, patients that completed all the PRO surveys for each time point. We ended up with a total of 651 and 483 patients predicting 6-weeks after treatment and for 12-months after treatment, respectively. All data imputation was done in R using the imputeTS [20] and mice packages. For MICE, the method used was predictive mean matching (pmm) and only one imputation was used with 5 iterations.

The LSTM models are built based on the open-source PyTorch framework. The input and output dimensions were set to 28. We used Mean Square Error (MSE) as the loss function and Stochastic Gradient Descent (SGD) as the optimizer with the learning rate of 0.215. Using a grid search for parameter tuning, we set the number of hidden layers to 1 and the number of hidden dimensions to 8. During training, we used early stopping criteria to prevent over-fitting. The RMSE score is calculated by applying the square root function on the PyTorch MSE metric and the NRMSE score is calculated by dividing the RMSE with the range (max - min) of the actual data. All the networks were run on NVIDIA GeForce RTX 2070 GPU with 8GB of memory.

4.2 PRO Data Summary

The PRO data is summarized in Figure 1 using the average symptom severity for the different time points. As can be seen, patients experience severe symptoms during the treatment and over time most of the symptoms return to baseline. However, for some symptoms and for some patients the toxicity persists even after 12 months after treatment.

4.3 Training LSTM using only complete data

Figure 2 shows the training and validation performance for the LSTM model trained for predicting symptom severity 6-weeks after treatment when only complete data (without imputation) is used in the analysis. As can be seen, the validation RMSE score is higher than the training RMSE score all the time, which can be caused by insufficient samples in the training set. The same trend was observed for the NRMSE metric as well.
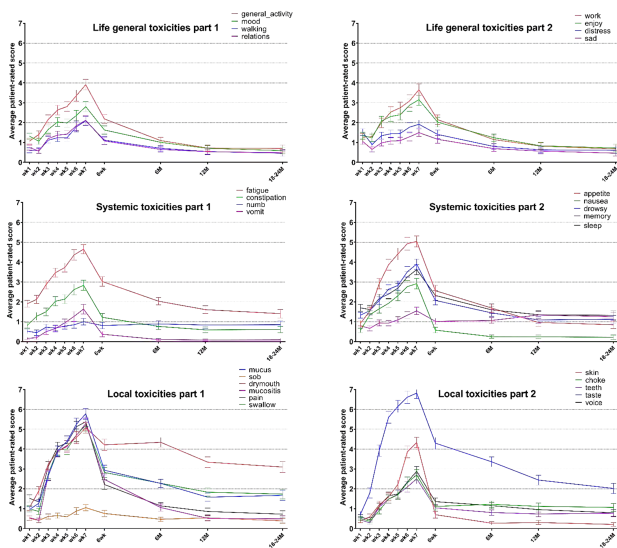
Figure 1: Average PRO scores trajectory and 95% confidence weekly during, 6 weeks, 3-6, 12 and 18-24 months after treatment for all 28 toxicities included in this study among all patients. Abbreviations: sob= shortness of breath; numb= numbness; wk=week; M=month
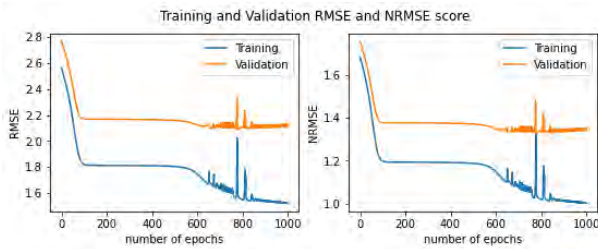

Figure 2: LSTM training and validation performance for predicting symptom severity 6-weeks after treatment using only complete data (original).

4.4 Data Imputation Evaluation

To evaluate the imputation methods, we compare the model performance in terms of RMSE and NRMSE when the LSTM was trained with complete data generated using the different imputation methods. Figure 3 shows the RMSE performance for training and validation of predicting symptom severity 6-weeks after treatment using complete datasets imputed with Linear interpolation, Kalman interpolation, MICE, and LSTM-recursive imputation methods. As can be seen, the model performance is better than when only complete data is used. Furthermore, the performance between the three imputation methods is similar and the models do not start to overfit until after 600 epochs for Linear interpolation and almost 1500 for the recursive LSTM. The final model uses the early stopping strategy to improve model performance.
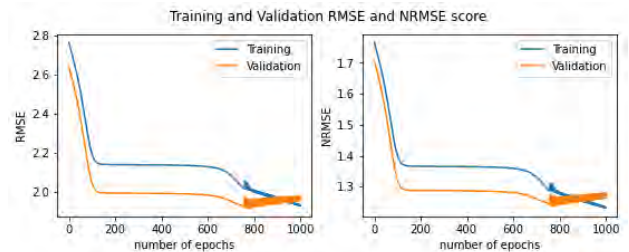
Figure 4 shows the validation RMSE metric for the final models trained over the complete data (original) and the four different
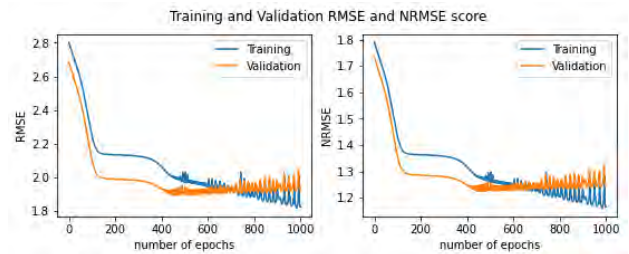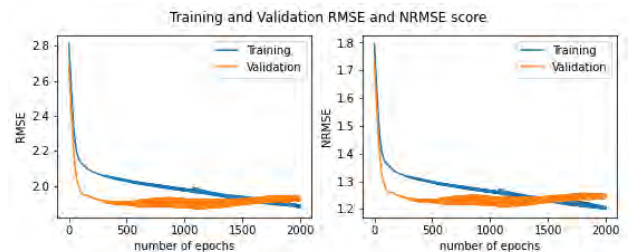

(a) Linear
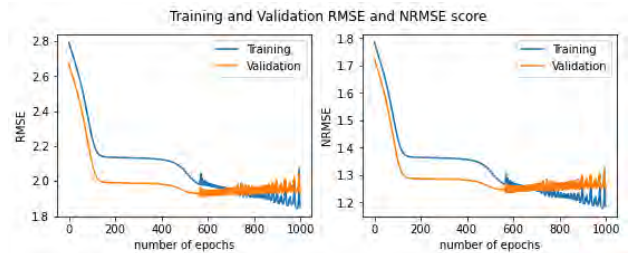
(b) Kalman

(c) MICE

(d) LSTM

Figure 3: LSTM RMSE performance for predicting symptom severity 6 weeks after treatment using Linear, Kalman, MICE, and LSTM imputed datasets.

imputed datasets. Linear imputation, in 6Wk after prediction, has the lowest validation RMSE score of and 1.9371 among the non-recurrent imputation methods whereas the original completed data has 2.1653 RMSE. The LSTM-recursive imputation shows the best overall performance for all metrics on late symptom predictions for 6Wk and 12Mo. The LSTM-recursive has an overall lowest RMSE of 1.9142 and 1.4231 for 6Wk and 12Mo, respectively. The LSTM-recursive method also achieved the best overall performance under the NRMSE metric (not shown). Worth noting is the fact that, regardless of what imputation method was used, the models trained
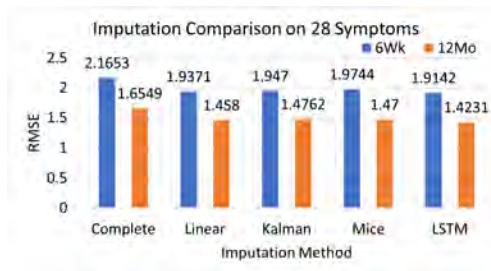
Figure 4: LSTM performance in terms of RMSE metrics for the validation set when the model is trained with only complete data (original) vs. complete data imputed with linear interpolation, Kalman interpolation, MICE, or LSTM methods.

with imputed data, and therefore more data, performed better than the models trained with only complete data.

4.5 LSTM performance on individual symptoms

Next, we want to evaluate the LSTM performance on individual symptoms. For these experiments, we used the LSTM-recursive interpolation dataset, as it had the best overall RMSE score for all 28 symptoms combined.

Figure 5, shows the (a) RMSE and (b) NRMSE score for each symptom predicted 6-weeks and 12-month after treatment. As can be seen, symptoms like taste and dry-mouth have higher RMSE scores for predictions at both time points while other symptoms like vomit and nausea have lower RMSE scores. However, when we look at the NRMSE metrics for the same symptoms, we see that the NRMSE scores are higher for some of those symptoms with lower RMSE. The reason is that NRMSE takes into account the range of the responses for the given time point. At 6-weeks after treatment, many symptoms still have a larger severity range than 12-months after treatment. Most of the symptoms have a lower RMSE score for 6-weeks after when compared to 12-months after treatment. The only exceptions are constipation and teeth, for which the RMSE score shows a slight increase for 12-months when compared to the 6-weeks prediction.

4.6 Comparison with other ML Models

For comparison to other machine learning models, we focus on three prevalent symptoms: pain, taste and general activity. The reason that we limit the number of symptoms is that the other models considered cannot handle multiple predictions simultaneously. We compare the LSTM performance with the performance of six other popular models: supported vector machine (SVM), K-nearest neighbour (KNN), random forest (RF), Gaussian naive Bayes (GauNB), multi-layer perceptron (MLP), and ARIMA.

Figure 6 shows the RMSE comparison for the 6-weeks after treatment prediction of pain, taste, and general activity symptoms for the 6 ML models and two LSTM: the one trained with the linear interpolated data (LSTM_L) and the recursive LSTM (LSTM*2). As can be seen, the LSTM model yields the lowest RMSE scores for

all three symptoms. For pain, LSTM has an RMSE of 1.7794 which is over 20% lower than the MLP prediction, which has the second-lowest RMSE score of 2.2646. LSTM also outperformed all other models for the 12-month prediction of these symptoms but results are omitted for brevity. Interestingly, for taste and general activity, the LSTM trained over the data imputed using linear interpolation shows a slightly lower error than the recursive LSTM imputed data. A plausible explanation is that interpolation uses the before and after values in the series to impute the missing values and intuitively symptom severity follow a linear increase/decay. In contrast, LSTM only uses past information to forecast symptoms' severity. In the future, it could be worth exploring alternatives that could combine both methods.

5 CONCLUSION

In this work, we used the PRO data from the MDASI-HN module and applied the LSTM model to predict late toxicity from head and neck cancer treatment. An accurate prediction can help identify personalized symptom risk profiles and proactively prescribe exercises or medication that can help patients cope with symptoms to avoid having to adjust treatment and improve the long-term quality of life of the patients.

To deal with the missing data, we applied three interpolation methods, linear, Kalman, and MICE. In addition, we also applied LSTM recursively to complete the data. We compared the performance of the LSTM model in terms of RMSE and NRMSE. The results show that using linear interpolation as the imputation method, though it is the simplest of the methods used, yielded better performance than Kalman and MICE imputations. While the LSTM imputation produces lower overall error measures than using linear interpolation, linear interpolation performed better than LSTM imputation for some individual symptoms. In all cases, the use of imputed data produced a better model than using only complete data. Furthermore, the LSTM model outperforms other machine learning models in the prediction of individual symptoms including pain, taste, and general activity symptoms.

As future work, we would like to evaluate whether the inclusion of clinical data [8, 17, 26, 29] into the analysis would further improve the predictive power of the LSTM models. There are different ways in which these clinical variables, which are mostly categorical, can be leveraged into the LSTM model to further improve model performance and symptom prediction.

REFERENCES

[1] Ratnadip Adhikari and R. K. Agrawal. 2013. An Introductory Study on Time Series Modeling and Forecasting. *CoRR* abs/1302.6613 (2013). arXiv:1302.6613 http://arxiv.org/abs/1302.6613
[2] A. Aktas, D. Walsh, and L. Rybicki. 2010. Symptom clusters: myth or reality? *Palliative medicine* 24(4) (2010), 373−−385. https://doi.org/10.1177/0269216310367842
[3] Stef Buuren and Catharina Groothuis-Oudshoorn. 2011. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45 (12 2011). https://doi.org/10.18637/jss.v045.i03
[4] Kaitlin M. Christopherson, Alokananda Ghosh, Abdallah Sherif Radwan Mohamed, et al. 2019. Chronic radiation-associated dysphagia in oropharyngeal cancer survivors: Towards age-adjusted dose constraints for deglutitive muscles. *Clin.*
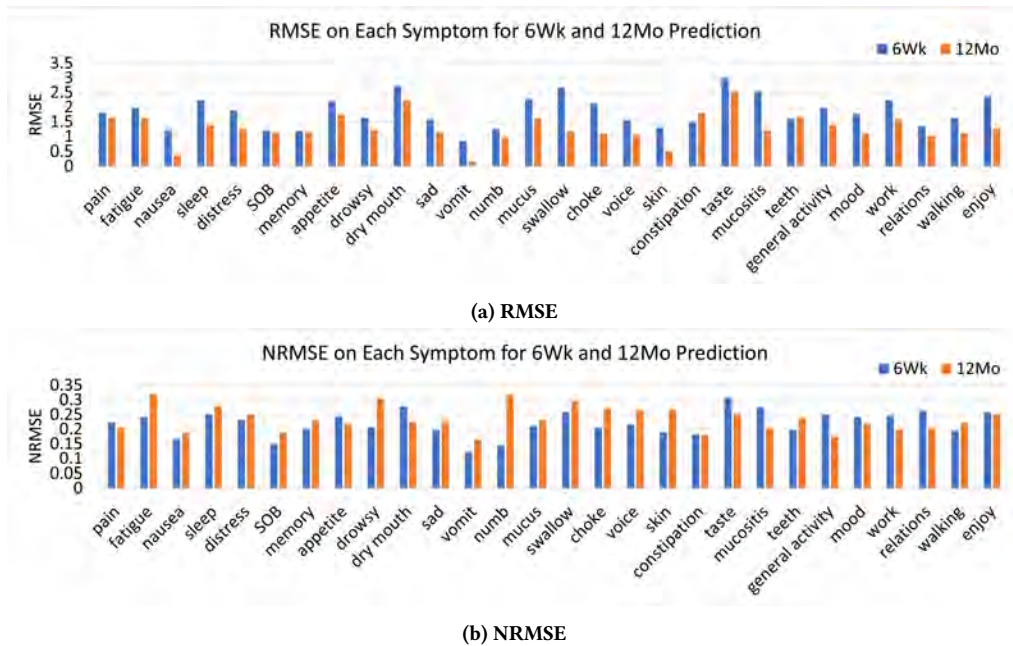
(a) RMSE

(b) NRMSE

Figure 5: Validation RMSE (a) and NRMSE (b) scores for the prediction of individual symptoms 6Wk and 12Mo after treatment.

(a) Pain

(b) Taste
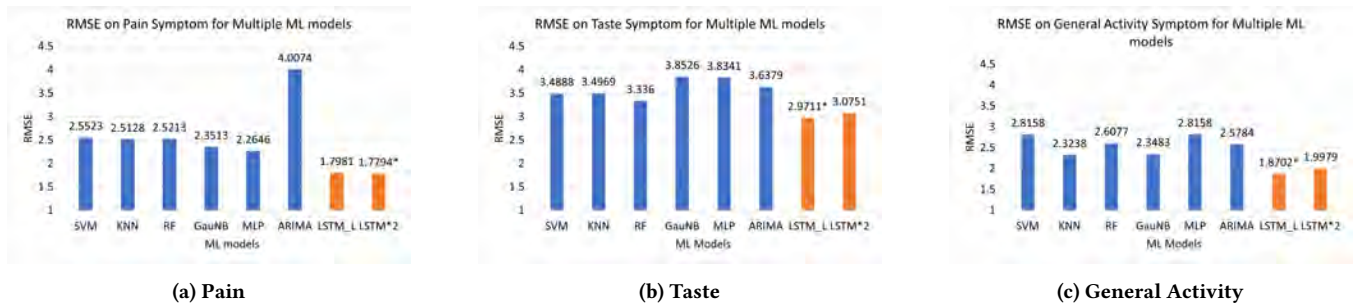

(c) General Activity

Figure 6: 6-week prediction RMSE on pain, taste, and general activity symptoms comparison for several machine learning models. Abbreviations: SVM = supported vector machine, KNN = K-nearest neighbour, RF = random forest, GauNB = Gaussian naive-bayes, MLP = multi-layer perceptron, ARIMA = Auto Regressive Integrated Moving Average, LSTM_L = LSTM on Linear Imputed data, LSTM*2 = LSTM on LSTM-recursive imputed data. * indicates the lowest RMSE score.

Transl. Rad. Onc. 18 (Sept. 2019), 16–22. https://doi.org/10.1016/j.ctro.2019.06.005

[5] C. S. Cleeland, T. R. Mendoza, X. S. Wang, et al. 2000. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer* 89(7) (2000), 1634−−1646. https://doi.org/10.1002/1097-0142(20001001)89:7<1634::aid-cncr29>3.0.co;2-v

[6] S. J. Coons, S. Eremenco, J. J. Lundy, et al. 2015. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. *The patient* 8(4) (2015). https://doi.org/10.1007/s40271-014-0090-z

[7] S. T. Dong, D. S. Costa, P. N. Butow, et al. 2016. Symptom Clusters in Advanced Cancer Patients: An Empirical Comparison of Statistical Methods and the Impact on Quality of Life. *Journal of pain and symptom management* 51(1) (2016), 88−−98. https://doi.org/10.1016/j.jpainsymman.2015.07.013

[8] Hesham Elhalawani, Abdallah SR Mohamed, Aubrey L White, et al. 2017. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Scientific data* 4 (2017), 170077.

[9] Steven Elsworth and Stefan Güttel. 2020. Time Series Forecasting Using LSTM Networks: A Symbolic Approach. arXiv:2003.05672 [cs.LG]

[10] S. A. Eraj, M. K. Jomaa, C. D. Rock, et al. 2017. Long-term patient reported outcomes following radiation therapy for oropharyngeal cancer: cross-sectional

assessment of a prospective symptom survey in patients ≥65 years old. *Rad. onc.* 12(1) (2017). https://doi.org/10.1186/s13014-017-0878-9

[11] G. Fan, L. Filipczak, and E. Chow. 2007. Symptom clusters in cancer patients: a review of the literature. *Current oncology (Toronto, Ont.)* 14(5) (2007), 173−−179. https://doi.org/10.3747/co.2007.145

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Rudolph Emil Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering* 82, Series D (1960), 35–45.

[14] M. Kamal, M. P. Barrow, J. S. Lewin, et al. 2019. Modeling symptom drivers of oral intake in long-term head and neck cancer survivors. *Supportive care in cancer* 27(4) (2019), 1405–1415. https://doi.org/10.1007/s00520-018-4434-4

[15] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, et al. 2020. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Frontiers in Big Data* 3 (2020), 4. https://doi.org/10.3389/fdata.2020.00004

[16] Zachary C. Lipton, David C. Kale, Charles Elkan, et al. 2017. Learning to Diagnose with LSTM Recurrent Neural Networks. arXiv:1511.03677 [cs.LG]

[17] Timothy Luciani, Andrew Wentzel, Baher Elgohari, et al. 2020. A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine. *J. Biomed. Informatics* 5 (2020), 100067.

[18] G. Maragatham and S Devi. 2019. LSTM Model for Prediction of Heart Failure in Big Data. *J Med Syst* 111 (2019). https://doi.org/10.1007/s10916-019-1243-3

[19] G. E. Marai, C. Ma, A. T. Burks, et al. 2019. Precision Risk Analysis of Cancer Therapy with Interactive Nomograms and Survival Plots. *IEEE Trans. Vis. Comp. Graphics* 25(4) (2019), 1732−−1745. https://doi.org/10.1109/TVCG.2018.2817557

[20] Steffen Moritz and Thomas Bartz-Beielstein. 2017. ImputeTS: Time Series Missing Value Imputation in R. *The R Journal* 9, 1 (2017), 207–218. https://doi.org/10.32614/RJ-2017-009

[21] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, et al. 2015. Comparison of different Methods for Univariate Time Series Imputation in R. arXiv:1510.03924 [stat.AP]

[22] D. I. Rosenthal, T. R. Mendoza, M. S. Chambers, et al. 2007. Measuring head and neck cancer symptom burden: the development and validation of the M. D. Anderson symptom inventory, head and neck module. *Head & neck* 29(10) (2007), 923−−931. https://doi.org/10.1002/hed.20602

[23] D. I. Rosenthal, T. R. Mendoza, C. D. Fuller, et al. 2014. Patterns of symptom burden during radiotherapy or concurrent chemoradiotherapy for head and neck. *Cancer* 120(13) (2014), 1975–1984. https://doi.org/10.1002/cncr.28672

[24] T. Sheu, D. Vock, A. Mohamed, et al. 2017. Conditional Survival Analysis of Patients With Locally Advanced Laryngeal Cancer: Construction of a Dynamic Risk Model and Clinical Nomogram. *Scientific Reports* (2017). https://doi.org/10.

1038/srep43928

[25] H. M. Skerman, P. M. Yates, and D. Battistutta. 2009. Multivariate methods to identify cancer-related symptom clusters. *Res. Nursing & Health* 32(3) (2009), 345−−360. https://doi.org/10.1002/nur.20323

[26] J. Tosado, L. Zdilar, H. Elhalawani, et al. 2020. Clustering of Largely Right-Censored Oropharyngeal Head and Neck Cancer Patients for Discriminative Groupings to Improve Outcome Prediction. *Scientific reports* 10(1) (2020). https://doi.org/10.1038/s41598-020-60140-0

[27] A. Wentzel, P. Hanula, T. Luciani, et al. 2020. Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration. *IEEE Trans. Vis. and Comp. Graphics* 26, 1 (Jan. 2020), 949–959. https://doi.org/10.1109/TVCG.2019.2934546

[28] Andrew Wentzel, Peter Hanula, Lisanne V van Dijk, et al. 2020. Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy. *Radiotherapy and Oncology* 148 (2020), 245–251.

[29] Luka Zdilar, David M Vock, G Elisabeta Marai, et al. 2018. Evaluating the effect of right-censored end point transformation for radiomic feature selection of data from patients with oropharyngeal cancer. *JCO clinical cancer informatics* 2 (2018), 1–19.

[30] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, et al. 2017. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. *CoRR* abs/1707.02485 (2017). arXiv:1707.02485 http://arxiv.org/abs/1707.02485