Improving Prediction of Late Symptoms using LSTM and Patient-reported Outcomes for Head and Neck Cancer Patients

Yaohua Wang
Electrical and Computer Engineering
The University of Iowa
Iowa City, United States
yaohua-wang@uiowa.edu

Lisanne Van Dijk
Radiation Oncology
UT M.D. Anderson Cancer Center
Houston, United States
dijkvansanne@gmail.com

Abdallah S. R. Mohamed
Radiation Oncology
UT M.D. Anderson Cancer Center
Houston, United States
asmohamed@mdanderson.org

Mohamed Naser
Radiation Oncology
UT M.D. Anderson Cancer Center
Houston, United States
manaser@mdanderson.org

Clifton David Fuller
Radiation Oncology
UT M.D. Anderson Cancer Center
Houston, United States
cdfuller@mdanderson.org

Xinhua Zhang
Computer Science
University of Illinois at Chicago
Chicago, United States
zhangx@uic.edu

G. Elisabeta Marai
Computer Science
University of Illinois at Chicago
Chicago, United States
gmarai@uic.edu

Guadalupe Canahuate
Electrical and Computer Engineering
The University of Iowa
Iowa City, United States
guadalupe-canahuate@uiowa.edu

***Abstract*—Patient-Reported Outcomes (PRO) are collected directly from the patients using symptom questionnaires. In the case of head and neck cancer patients, PRO surveys are recorded every week during treatment with each patient's visit to the clinic and at different follow-up times after the treatment has concluded. PRO surveys can be very informative regarding the patient's status and the effect of treatment on the patient's quality of life (QoL). Processing PRO data is challenging for several reasons. First, missing data is frequent as patients might skip a question or a questionnaire altogether. Second, PROs are patient-dependent, a rating of 5 for one patient might be a rating of 10 for another patient. Finally, most patients experience severe symptoms during treatment which usually subside over time. However, for some patients, late toxicities persist negatively affecting the patient's QoL. These long-term severe symptoms are hard to predict and are the focus of this study. In this work, we model PRO data collected from head and neck cancer patients treated at the MD Anderson Cancer Center using the MD Anderson Symptom Inventory (MDASI) questionnaire as time series. We impute missing values with a combination of K nearest neighbor (KNN) and Long Short-Term Memory (LSTM) neural networks, and finally, apply LSTM to predict late symptom severity 12 months after treatment. We compare performance against clinical and ARIMA models. We show that the LSTM model combined with KNN imputation is effective in predicting late-stage symptom ratings for occurrence and severity under the AUC and F1 score metrics.**

***Index Terms*—Long Short-Term Memory (LSTM), Patient Reported Outcomes (PRO), Late Toxicity, Symptom Severity Prediction, KNN Baseline Imputation**

I. Introduction

In clinical practice, Patient-Reported Outcomes (PRO) are considered an important complement to capture the patient's condition outside of the medical visit so that physicians can understand more about the patients and improve their QoL. PRO data have a major impact on clinical decision-making, and informing clinical practice [1]. In general, PRO data are collected directly from the patients as survey responses and often ask the patient to rate different symptoms in terms of severity over some time period. PRO data is routinely collected in clinical practice and specific questionnaires have been created and validated for different diseases and conditions. For head and neck cancer patients, the Head and Neck module MDASI-HN from the M.D. Anderson Symptom Inventory (MDASI) [2] questionnaire is a validated and widely used instrument. The module contains 28 symptom-related questions, in which, 13 questions are related to the systemic core symptoms of cancer, 9 are related to local head and neck symptoms, and the last 6 are related to life general symptoms associated with daily activities. During and after treatment, patients repeatedly rate their symptoms on a scale from 0 to 10 to indicate their symptom severity, where 0 stands for no experience of the symptom and 10 stands for greatest severity.

PRO surveys can be very informative regarding the patient's status and the effect of treatment on the patient's quality of life (QoL). However, processing PRO data is challenging for

2575-2634/23/$31.00 ©2023 IEEE
DOI 10.1109/ICHI57859.2023.00047

several reasons. First, missing data is frequent as patients might skip a question or a questionnaire altogether. Dropping the patients that are missing any or some of the questionnaire responses is not a feasible alternative because a significant portion of the patients would have missing data at one time or another. Another challenge is that PROs are patient-dependent and somewhat subjective measures, e.g. a rating of 5 for one patient might be a rating of 10 for another patient. Symptoms are also often correlated, co-occurring, or product of the same underlying cause. For head and neck cancer, most patients experience severe symptoms during treatment and ideally, these would subside over time. However, in some cases, patients continue to experience health and/or Quality-of-Life (QoL) debilitating symptoms even years after the end of treatment [3]–[5]. These long-term severe symptoms are hard to predict and are the focus of this study. We are interested in developing methods that can predict these moderate-to-severe long-term symptoms and identify patients at risk so that interventions can be implemented to minimize the risk and ultimately improve patients' QoL.

In this paper, we propose the use of Long Short-Term Memory (LSTM) neural networks to model the symptom rating trajectories for PRO data. We model PRO data collected using MDASI-HN questionnaires as a time series and focus on training an LSTM model to predict late toxicity at 12 months after treatment (M12). In our preliminary work [6], we show that recursively using an LSTM neural network to predict and impute missing PRO data at different time points is able to outperform linear interpolation and other imputation methods. The LSTM model recursively predicts the following time step using data from the prior time points starting from a baseline or time 0 at the start of treatment. This formulation makes the proposed model generalizable to other types of longitudinal PRO data collected for surveillance or monitoring.

In this work, we make several important improvements to the LSTM prediction of PRO data. Since the LSTM recursive imputation needs a starting initial value for the time series and to prevent the exclusion of patients missing such baseline or initial symptom rating from the analysis, we evaluate two baseline imputation methods: mean imputation and KNN imputation. For KNN, we compute the similarity between patients using the available clinical variables including AJCC staging (T, N, and M), tumor location, and treatment. Furthermore, to account for patient variability, we subtract the baseline score from the subsequent symptom ratings, train the LSTM over the delta changes from the baseline, and show that this normalization further improves prediction. Finally, since symptom cluster research has identified groups of related symptoms [7], we focus on the symptom cluster that includes Dry mouth, Mucus, Swallowing, and Taste [8] given the prevalence of these symptoms after the end of treatment for head and neck cancer patients. We refer to this cluster as DMST for the initials of the symptoms involved. We compared the prediction outcome of the LSTM models against regression models using clinical data and PRO (ARIMA) and show that LSTM is an effective way of predicting late symptoms for head and neck cancer patients.

The rest of this paper is organized as follows. Section II presents related work. Section III describes the proposed approach. Section IV presents the experimental results using MDASI questionnaires. Finally, we conclude in Section V.

II. Related Work

MDASI-HN PRO Data. Patient Reported Outcomes (PRO) is data collected directly from the patient and it is widely used to evaluate treatment benefits and measure symptom burden for the patient [9]. The PRO data used in this project is the MD Anderson Symptom Inventory (MDASI) Head and Neck (HN) module. The MDASI-HN [10] is a 28-symptom questionnaire where patients rate symptoms on a scale from 0 to 10 with 0 being not present and 10 being the worse ever. As shown in Table I, the 28 symptoms can be divided into three broad groups: systemic (common to all cancers), local (specific to head and neck), and life interference. Patients are asked to fill out MDASI-HN surveys before the start of treatment (baseline), weekly during treatment, and at their follow-up visits 6 weeks, 6 months, and 12 months after treatment.

Using MDASI-HN PRO data, several studies focus on identifying symptom clusters at a single timepoint [8], [11]–[13]. From these preliminary researches, Dry mouth, Mucus, Swallowing, and Taste (DMST) are four of the most severe symptoms. Furthermore, cluster analysis using MDASI-HN data shows these symptoms have small relative distances and are highly related [8]. 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 [7], [14]–[16]. These studies focus on a single time point analysis, whereas we model the PRO data as a time series.

Time Series Prediction and Imputation. When comes to time series prediction, Auto-regressive Integrated Moving Average (ARIMA) [17] is a commonly used method. The model combines an Auto-regressive (AR) model which predicts the variable using a linear combination based on its previous values and a Moving Average (MA) model which uses the past prediction error rather than past value in prediction. In handling the missing values in the data, the ARIMA model can compute without struggling by skipping them in the update stage.

More recently, Long Short-Term Memory (LSTM) Recurrent Neural Networks have gained more popularity in time series prediction [18] and healthcare domain [19]. Specific applications of LSTM in healthcare include mimicking the pathologist's decision and other diagnostic applications [20], [21], classification of sleep patterns in multi-variate time-series clinical measurements [22], and predicting symptom severity in the acute and late stages after the treatment [6].

III. Proposed Approach

In this section, we describe the methodological approach including data pre-processing and training metrics used in this work. Figure 1 shows an overview of the proposed

TABLE I
THE 28 SYMPTOMS IN THE MDASI-HN QUESTIONNAIRE BY TOXICITY TYPE. SYSTEMIC REFERS TO COMMON CANCER SYMPTOMS, LOCAL TO MORE HEAD AND NECK SPECIFIC SYMTOMS, AND LIFE GENERAL TO THE INTERFERENCE RATINGS RELATED TO QUALITY OF LIFE.

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

methodology and each component is described in detail in the following subsections.

A. Data

The data contains several clinical features in addition to the sequential MDASI-HN PRO questionnaires. The clinical data is mainly categorical data including variables such as sex, tumor location, AJCC staging, T-stage, N-stage, M-stage, performance score, HPV status, and treatment. The three numerical variables correspond to age, height, and weight. Numerical variables were discretized in order to have a homogeneous set of features for distance computation as described in the next section. MDASI-HN PRO questionnaires were repeatedly administered to patients during treatment (at baseline, start, and end of treatment including weeks 2-6, 6 weeks, 6 months, and 12 months after treatment) for a total of 11 time points.

Since our focus is on late-stage symptoms, i.e. symptoms that affect patients (with moderate to high severity) long after the end of treatment, we evaluate rating prediction for symptoms at M12.

B. Preprocessing

We compute the BMI of the patients and grouped them into 4 categories (underweight, normal, obese, and morbidly obese). Age is discretized into four groups defined by 25%, 50%, and 75%-tile values. We treat the MDASI-HN-PRO-data as a time series and extract the ratings for four different symptoms: Dry mouth, Mucus, Swallowing, and Taste (DMST). Prior research has identified these symptoms as a cluster using MDASI-HN data [8]. We decide to focus on this symptom cluster because their average severity at the late stage, i.e. at M12, is moderate to severe, which means they have a long-term effect on patients. To account for patient variability, we subtract the baseline rating from the subsequent ratings for each symptom. The baseline is the rating provided by the patient before the start of the treatment. After subtraction, all the patients have their initial rating set at zero and all subsequent ratings as the delta change from their baseline. The MDASI-HN PRO data have lots of missing symptom ratings due to patients' skipping a question or failing to complete a questionnaire, loss of follow-up, or any other omissions during data collection. To prepare the PRO data for LSTM training, we transform the data 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 LSTM can be applied recursively to impute different time steps, but it needs to start from a known baseline value. Therefore, we distinguish between the baseline imputation which is needed before applying the LSTM and the imputation of subsequent time points by the LSTM model.

C. Baseline Imputation

In order to complete the missing data at baseline, we consider two different approaches: mean imputation and k-nearest neighbor (KNN) based imputation.

Mean Imputation. In this approach, we simply take the average of the existing ratings of the patients for each symptom at baseline or the first time point, i.e., week 0, and impute all missing values with the calculated average.

KNN Based Imputation. With the assumption that patients with similar demographics, disease stage, tumor location, and treatment would have similar baselines, we decided to compute the similarity between patients using these clinical variables. We apply KNN to the clinical data of each patient and compute the average baseline among the most similar patients to fill the missing baseline rating. The hope is that the imputed baseline will be more diverse than using the global average and this would translate into a better prediction of late symptom ratings.

While KNN is a non-parametric approach, we still need to define several important parameters: 1) which features to use for computing KNN imputation? 2) Which distance metric would be more suitable for these data?, and 3) What is a proper K for the K nearest neighbor?

As a result, we developed the following KNN imputation methodology, as it is shown in the KNN imputation pipeline depicted in Figure 1.

To answer the first question, we applied a feature selection algorithm (i.e. ridge regression) to identify the relevant features from the 13 clinical features available in the dataset. The distribution of clinical features is shown in Table II. As the outcome for the regression model, we created a binary outcome (0, 1) to indicate whether the patient experienced any of the selected DMST symptoms at baseline. Once the relevant features were identified, we then apply KNN over this reduced set. Since all the features are categorical, we consider Overlap and Goodall3 [23] as potential similarity metrics. Overlap similarity between two patients X and Y is the number of attributes (features) where the two patients fall into the same category and is defined by:

$$\mathtt{Overlap}(\mathtt{X},\mathtt{Y}) = \sum_{\mathtt{i}=1}^{\mathtt{d}} \begin{cases} 1 & \text{if } X_i = Y_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

d is the number of attributes used in the similarity.

Goodall3 similarity applies a penalty to the attributes where the two patients match using the sample probability of the attribute value and it is defined as:

$$\mathtt{Goodall3}(\mathtt{X},\mathtt{Y}) = \sum_{\mathtt{i}=1}^{\mathtt{d}} \begin{cases} 1 - p_i^2(X_i) & if\ X_i = Y_i \\ 0 & otherwise \end{cases} \quad (2)$$

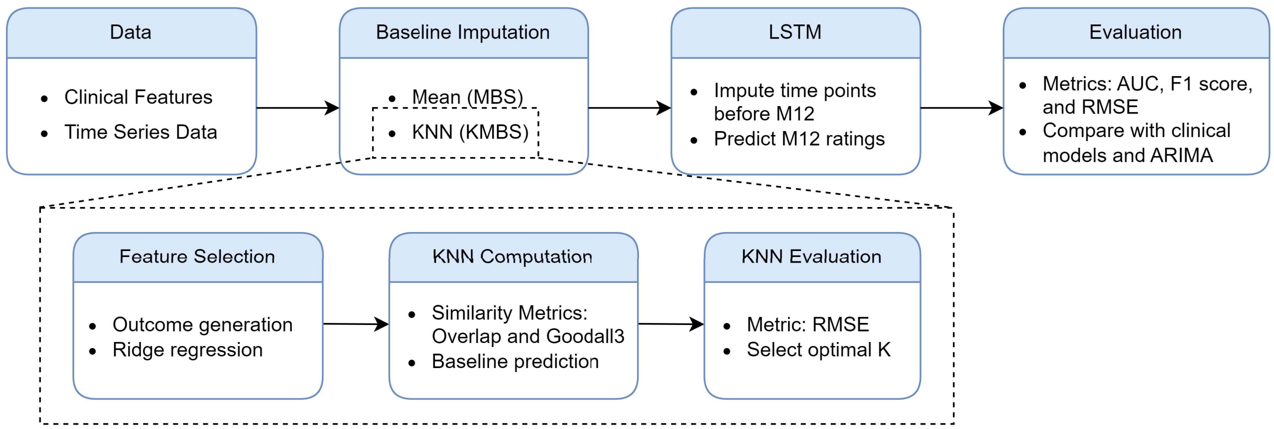


Fig. 1. Overview of the proposed methodology. The top row shows the overall processing pipeline and the bottom row further details the KNN imputation processing pipeline.

TABLE II
DATA DEMOGRAPHICS FOR THE CLINICAL FEATURES AVAILABLE PRIOR TO TREATMENT. A SET OF SELECTED FEATURES WERE USED FOR THE KNN BASELINE IMPUTATION.

Feature	Value(s)	Percentage
Sex	Male / Female	88.21% / 13.36%
AJCC	7th / 8th	56.74% / 43.26%
T_numeric*	0	6.68%
	1	28.92%
	2	39.85%
	3	13.12%
	4	11.42%
N_numeric*	0	12.64%
	1	36.94%
	2	48.24%
	3	2.19%
Location	BOT	43.38%
	Tonsil	42.89%
	Others	13.73%
Treatment*	Induction_RT	6.08%
	CC	71.69%
	no_chemo	5.59%
	Induction + CC	8.14%
BMI	Underweight	10.45%
	Normal	13.37%
	Obese	36.94%
	Morbid Obese	37.67%
Perform. Score*	0	66.71%
	1	20.17%
	2	1.94%
	3	0.61%
	4	0%
Enrollment_1*	Yes / No	0.85% / 99.15%
Enrollment_2*	Yes / No	1.58% / 98.42%
Enrollment_3	Yes / No	2.55% / 97.45%
HPV Status	Yes / No	72.42% / 27.58%
Age	Median, (25% - 75%)	60, (54 - 67)

* indicates the selected features before the KNN imputation.

where d is the number of attributes, $f_i(x)$ is the frequency of value x in attribute i, and $p_i^2(x) = \frac{f_i(x)(f_i(x)-1)}{N(N-1)}$, where N is the total number of patients.

When we compute both the Overlap and the Goodall3 similarities, patients with a similarity score less than 50% are dropped from the KNN set to minimize the effect of potential outliers.

To evaluate and compare the effectiveness of the two similarity metrics we consider different numbers of nearest neighbors (K) and calculate the RMSE between the actual patients' rating per symptom and the KNN predicted baseline per symptom. To minimize the bias of the baseline, we calculated the mean of the patients who have a baseline per symptom and subtracted it from the baseline. Therefore, the KNN is predicting the delta of the baseline, and the mean is later added back to the KNN prediction when compared to the actual patients' rating.

After deciding a suitable value of K and which similarity metric to use for the KNN model, we split the patients into two groups based on the absence of baseline symptom ratings, apply the KNN to each patient with a missing baseline, and take the average of the similar patients that the KNN identified. In this way, all the missing baselines were imputed.

D. Long Short-Term Memory (LSTM)

We picked LSTM neural network as the predictive model because it has proven effective in time series prediction [24]. Unlike traditional neural networks, LSTM is a type of recurrent neural network (RNN) that has a 3-gate feedback structure to memorize the important part of the input data and forget the unimportant, in other words, LSTM stores the memory of past events and use it to predict the future events. Furthermore, LSTM has the advantage of a diversity of inputs and outputs, which means, in our case, LSTM can take multiple patients' ratings over time on multiple symptoms as inputs and predict their responses to multiple symptoms in the late stage.

Since patients respond to the survey providing their symptom severity ratings periodically during and after treatment, we can learn from their responses over a past time period and predict the responses in a future time period so that recommendations can be made proactively to minimize patients' symptom burden thus improving their quality of life. We applied Long short-term memory (LSTM) neural networks in two ways during the training process. By recursively predicting the symptom ratings at each time point, we are able to impute the missing data. Using the complete data, we then predict the symptom burden at M12.

LSTM Imputation. After the baselines have been imputed, we then trained our LSTM model on the complete baseline data and let it predict the missing values at week 1. The predicted values are used to fill in the missing values for week 1. Subsequently, the same process is applied to predict the following weeks. This recursive process is repeated until all the missing ratings before M12 are fulfilled.

Late Symptom Prediction. We train the LSTM on all the time points before M12 to predict M12 symptom severity. All data prior to the prediction time point have been imputed. To account for patient variability in their ratings, we subtracted the baseline rating from all subsequent time steps including M12 ratings. The LSTM model predicts the delta of each patient and after prediction, the baseline is added back to the predicted rating.

Two approaches are evaluated for late symptom prediction depending on which method was used for baseline imputation. When the mean imputation is used, we refer to this approach as **Mean Baseline Subtracted (MBS)**. When KNN imputation is used, we refer to this approach as **KNN-based Mean Baseline Subtracted (KMBS)**.

E. Evaluation

We evaluate the 12-month symptom rating predictions (M12) using 5-cross validation and compute RMSE, AUC, and F1 scores between the predicted ratings and the actual ratings provided by the patients. We compare LSTM prediction performance with two models. The first one, referred to as the Clinical model, uses a logistic regression on all the clinical features and baseline symptom ratings to predict 12-month after-treatment symptom ratings. The second method is the auto-regressive integrated moving average (ARIMA) statistic model. ARIMA, as mentioned above, is able to handle the missing values and is used to predict 12-month after-treatment symptom ratings using all prior time points available.

IV. Experimental Results

In this section, we first describe the experimental setup and data statistics. Then we present the results for the KNN baseline imputation evaluation, compare the LSTM performance using both imputation methods for DMST symptoms, and finally compare the predictive performance of LSTM against ARIMA and a clinical regression model.

A. Experimental setup

For the KNN feature selection, a ridge regression model was trained using a 75/25 split and a grid search for the parameter $\alpha = 10^i$ with i $\in [-2, 6]$ with 0.5 increments. The minimum validation error was found for $\alpha = 1$, and features were selected with threshold ≥ 0.01. The selected features were T numeric, N numeric, Treatment, Performance score, Enrollment 1, and Enrollment 2 (highlighted in Table II).

For the LSTM imputation and prediction, we used Mean Square Error (MSE) as the loss function and Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.215. Using a grid search for parameter tuning considering hidden layers between 1-5 and the number of hidden dimensions between 1-10, we set the number of hidden layers to 1 and the number of hidden dimensions to 10. We also used early stopping criteria when training the LSTM model to prevent over-fitting. All the networks were run on NVIDIA GeForce RTX 2070 GPU with 8GB of memory. The LSTM model is built based on the open-source TensorFlow framework. Numpy, Pandas, and Scikit-Learn libraries were also used. As for the ARIMA model, we used the pre-built first-order autoregressive ARIMA in the statsmodels (ver. 0.13.0) library by setting the order to (1, 0, 0).

B. Data statistic

The MDASI-HN module contains a considerable number of patients with missing ratings at each time point. Figure 2 shows the percentage of patients with missing symptom ratings for each time point before month 12 ($<$ M12). The solid line is the percentage for all patients while the dashed line corresponds to patients with known M12 ratings, i.e. the subset of patients used in this work. As can be seen, both sets show similar distribution with missing data around 20% at baseline, raising to 42-56% during treatment. In fact, there are only 26 out of 823 patients who have all questionnaires fully completed for all time points. Therefore, simply dropping the patients with missing ratings would dramatically reduce the size of the sample. As a result, data imputation for the missing ratings is needed.

C. KNN Baseline Imputation

We compare the performance of the baseline imputation for KNN using Overlap and Goodall3 distance metrics over the selected features for different values of K. Figure 3 shows the RMSE between the patients' original baselines and the patients' KNN predicted baselines. We also computed the normalized KNN predicted baselines (Z-score) and obtained the RMSE between the original baselines and the denormalized KNN predicted baselines. As it is shown in Figure 3, the KNN with the Overlap metric shows a lower RMSE than the KNN with the Goodall3 metric for all the five values of K evaluated between 1 and 20. The RMSE scores with the "Norm" baseline are all lower than the RMSE scores with the "Original" baseline. These results showed the use of Z-score normalization of patients' scores and computing the KNN prediction as the average of the scaled values improved

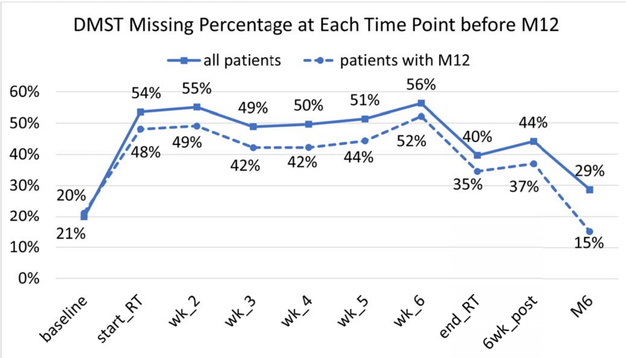


Fig. 2. Average percentage of missing data for DMST symptoms at each time point before M12. The solid line shows the average percentage for all the patients, whereas the dashed line shows the percentage of missing data only for patients with known M12 ratings.

the RMSE by smoothing the outliers. Moreover, the similarity between each patient and their KNN set was always above the 50% threshold. For the 158 patients with missing baseline rating for dry mouth, 85% had a perfectly matched set of nearest neighbors and from the remaining 15% (25 patients), only 1 had some neighbors at 50% similarity.

Since the RMSE lines shown in Figure 3 have a hinge at K = 5 and decrease slowly after, both K=5 and K=10 are good candidates for KNN baseline imputation. For the remaining experiments, the KNN imputation method reported is using the Overlap metric with K = 10 and Z-score normalization.

D. LSTM Evaluation

Table III shows the patient distribution for DMST symptoms and the three thresholds for symptom ratings used to evaluate the LSTM performance: occurrence (≥ 1), mild-severe (≥ 3), and moderate-severe (≥ 5). The distribution is shown for two patient groups. The M12 group corresponds to the set of patients with a known M12 symptom rating whereas the $M12_{IB}$ group corresponds to the patients with known M12 ratings that needed an imputed baseline. The $M12_{IB}$ group is a subset of the M12 group and corresponds to the patients affected by the baseline imputation method. This set is used to evaluate more clearly, the effect of baseline rating imputation. As can be seen in the Table, among DMST symptoms, Dry mouth is the most prevalent symptom with 82% of patients having the symptom and 28% feeling it moderately-severely at M12. The other symptoms (Mucus, Swallowing, and Taste), on the other hand, occur in the majority of the patients (55% - 70%) but only a smaller percentage (10% - 16%) of patients experience moderate-severe occurrences. While the $M12_{IB}$ group for the DMST symptoms is only around 100 patients, the distributions among the three thresholds are still similar to those for the entire cohort.

Symptom Rating Normalization. To evaluate the effect of baseline imputation and patient rating normalization, we first evaluate LSTM performance over the $M12_{IB}$ group. Figure 4 shows the performance of the LSTM models when the symptoms ratings are normalized by subtracting the baseline from all subsequent ratings (MBS and KMBS) or not (MB and KMB). The figure shows the F1 scores for symptom occurrence (rating ≥ 1). As can be seen, there is a performance improvement when the symptoms ratings are normalized (MB vs. MBS and KMB vs. KMBS). Subtracting the baseline ratings increased F1 score performance between $[3.85\%, 30.36\%]$ for DMST symptoms. This normalization makes all the patients' baselines the same, and the LSTM is effectively trained to predict delta from baseline. Given the better performance of this normalization for all symptoms, moving forward we only compare the MBS and KMBS LSTM approaches to other methods.

LSTM Prediction with Imputed Baseline. Figure 6 shows the performance comparison for M12 DMST symptom predictions in terms of RMSE for the $M12_{IB}$ patients. We compare the performance of the LSTM models (MBS and KMBS) against ARIMA and the clinical regression models. As can be seen, for Mucus, Swallowing, and Taste symptoms, the RMSEs of the clinical model are the highest, i.e. perform the worst, ranging between $[3.5, 4.3]$, and the ARIMA model achieves the highest RMSE of 2.8 for Dry mouth symptom. In contrast, LSTM models perform considerably better with RMSE being 7.5% - 59% lower than ARIMA and 22% - 165% lower than the clinical model predictions across DMST symptoms.

Figure 5 shows the performance comparison for M12 occurrence predictions over the $M12_{IB}$ group in terms of F1 scores. As can be seen, the MBS and KMBS-based LSTM achieve higher F1 scores than those of the clinical and the ARIMA model for all DMST symptoms. The performance is closest between the models for Dry mouth prediction, with the clinical model performing the lowest at almost 80% and the KMBS-LSTM model achieving the highest F1 score at 87.5%. The lowest performing symptom for the clinical and ARIMA models is Mucus, where the LSTM approach still achieves F1 score ≥ 0.7. Overall, the KMBS-LSTM approach achieves the highest F1 score for all the symptoms except for Swallowing where the MBS-LSTM approach shows the best relative performance between the compared models.

Figure 7 shows the average AUC performance for the DMST symptoms at two severity thresholds (≥ 1 and ≥ 3). As can be seen, the AUCs of the clinical model are significantly lower than the AUCs of the ARIMA model and LSTM models, ranging between $[0.39, 0.66]$. Both the ARIMA model and LSTM models show competitive performance with values ranging between $[0.73, 0.83]$. The LSTM models show a higher average AUC than the ARIMA models. At thresholds 1 and 3, the AUCs of the ARIMA model range between $[0.73, 0.79]$ and $[0.77, 0.8]$ whereas the AUCs of the LSTM models range between $[0.76, 0.83]$ and $[0.79, 0.83]$.

LSTM Prediction for the Entire Cohort. Figure 8 shows the performance comparison for the M12 rating prediction of the DMST symptoms for the entire cohort (M12 group). The figure shows F1 scores for the Clinical, ARIMA, and LSTM models (MBS and KMBS) at each of the three different

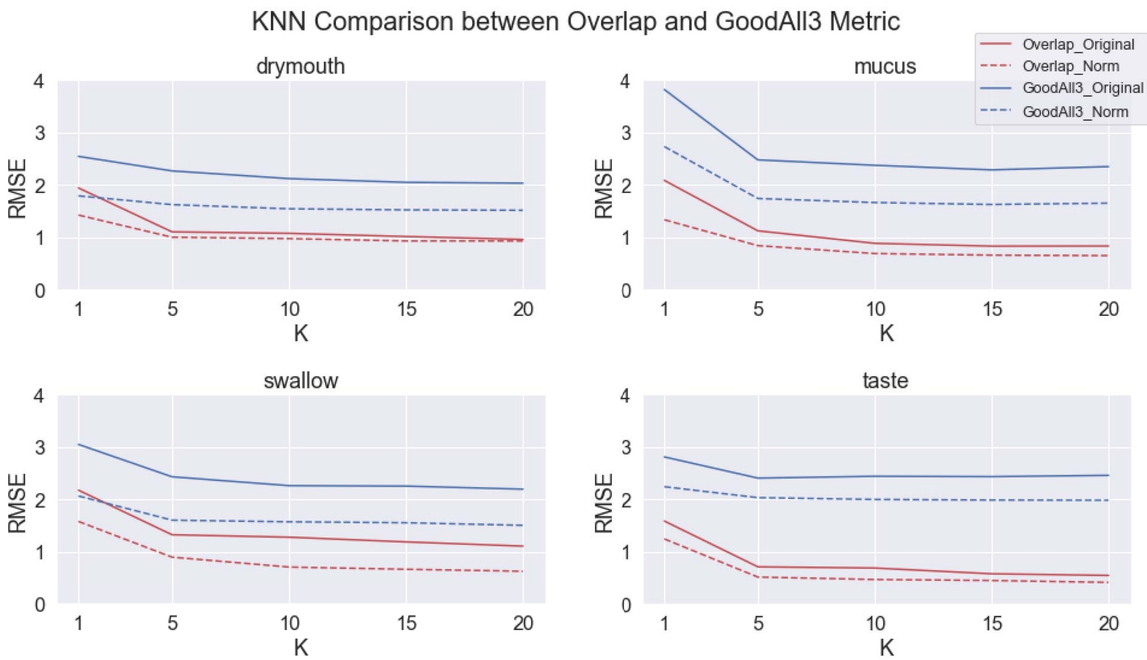


Fig. 3. KNN Baseline Imputation Comparison between Overlap metric and Goodall3 Metric as a function of K. _Original means the RMSE was computed between original baselines and KNN-predicted baselines; _Norm means the RMSE was computed between the original baselines and the normalized (and later denormalized) KNN-predicted baselines. Overall, the _Norm KNN-predicted baselines have lower RMSE than the _Original KNN-predicted baselines.

TABLE III
PATIENTS' RATING DISTRIBUTION FOR DMST SYMPTOMS 12 MONTHS AFTER TREATMENT (M12) AND FOR PATIENTS WITH KNOWN M12 BUT MISSING BASELINE RATING ($M12_{IB}$). $M12_{IB}$ IS THE SET OF PATIENTS FOR WHICH BASELINE RATINGS NEED TO BE IMPUTED.

	Dry mouth				Mucus				Swallowing				Taste			
	M12		$M12_{IB}$		M12		$M12_{IB}$		M12		$M12_{IB}$		M12		$M12_{IB}$	
# Patients	464		96		463		98		461		100		459		97	
Symptom Threshold	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**	**Cnt**	**%**
Occurrence (Rating ≥ 1)	380	82	78	81	248	54	55	56	300	65	66	66	316	69	69	71
Mild-Severe (Rating ≥ 3)	223	48	52	54	115	25	23	23	191	41	30	30	172	37	39	40
Moderate-Severe (Rating ≥ 5)	129	28	23	24	46	10	7	7	46	10	9	9	75	16	20	21

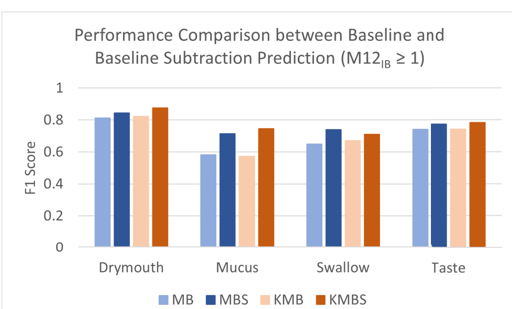


Fig. 4. Evaluating the impact of baseline subtraction on the prediction of DMST symptom occurrence. The first two bars in each column (blue pair) correspond to the mean imputation and the last two bars (orange pair) correspond to the KNN imputation. The baseline subtraction has a positive effect on the performance metric (F1 score) for both mean and KNN-imputation approaches for all symptoms.

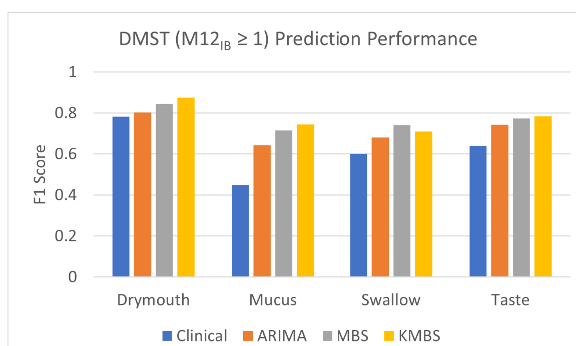


Fig. 5. Performance comparison for the clinical model, ARIMA, and LSTM methods on the DMST symptoms occurrence at M12 for the baseline imputed patients ($M12_{IB}$). MBS stands for the mean baseline subtracted approach, and KMBS stands for the KNN-based mean baseline subtracted approach.

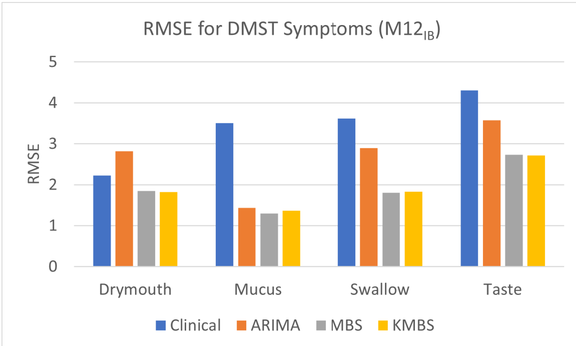


Fig. 6. RMSE for the prediction of DMST symptoms in the $M12_{IB}$ group. Overall, clinical and ARIMA approaches have higher RMSE than LSTM-based (MBS and KMBS) approaches.

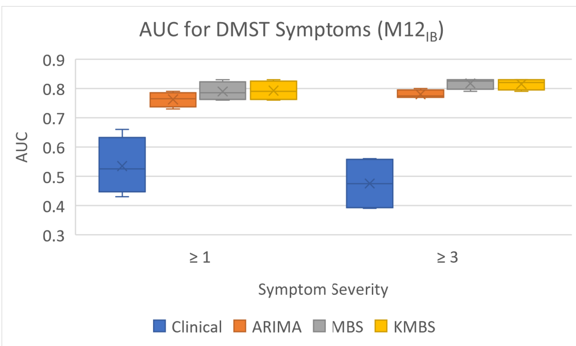
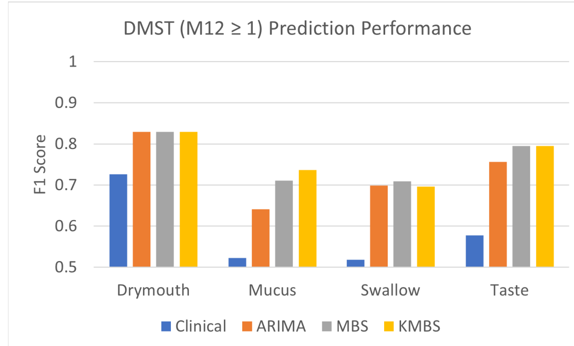
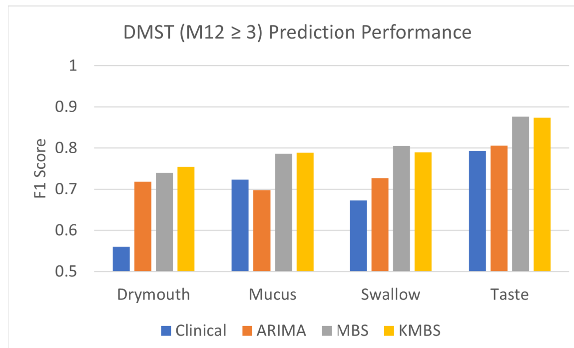


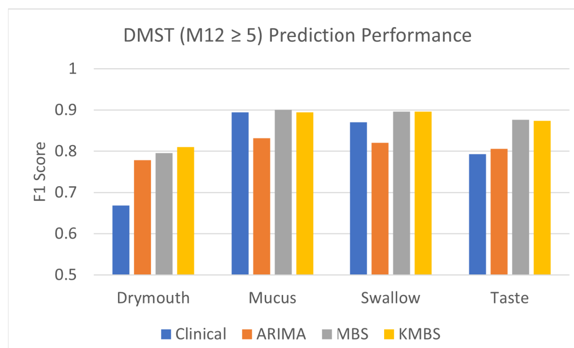
Fig. 7. AUC for the prediction of DMST symptoms in the $M12_{IB}$ group. The clinical approach has overall lower AUC performance than ARIMA, MBS, and KMBS approaches.



(a) Symptom Occurrence (Rating ≥ 1)



(b) Symptom Occurrence (Rating ≥ 3)



(c) Symptom Occurrence (Rating ≥ 5)

Fig. 8. Prediction performance on the DMST symptoms at M12 for the three severity thresholds (M12 data). Overall, MBS and KMBS approaches show better performance than Clinical and ARIMA approaches.

symptom severity thresholds. As can be seen, for all the symptoms, at each threshold, the LSTM prediction has either a similar or better performance than the clinical and ARIMA models. The performance difference between mean-baseline imputation (MBS) and KNN imputation (KMBS) imputation is negligible between the two. The main reason is that no imputation is needed for the large majority of patients diluting the performance impact for the baseline imputation. In any case, baseline imputation does not seem to degrade the performance of the LSTM. The AUC scores over the entire cohort had similar distributions as the F1 scores and are omitted for brevity. It is worth noting that the LSTM still showed better performance in terms of AUC when compared to the other models, ranging between $[0.75, 0.89]$. The clinical models had AUC around 0.5 and ARIMA between $[0.72, 0.84]$.

V. Conclusion

In this work, we used the MDASI-HN PRO data and applied LSTM neural network for symptom rating imputation and late-stage (M12) symptom rating prediction. A significant challenge for these patient-reported outcome (PRO) data is the subjectivity or patient variability when responding to their symptom severity. Some patients might be more sensitive to some symptoms while others might dismiss them or not even notice them. To account for this subjectivity we subtract the baseline symptom rating from subsequent ratings effectively making all patients have the same baseline and showing performance improvements over raw rating prediction.

For patients with missing baseline ratings, we evaluated two different baseline imputation approaches. The first approach, MBS, simply computed the average based on the baseline-existing patients, while the second approach, KMBS, is a KNN imputation method that incorporates clinical information into the baseline imputation by computing each patient's baseline from the average of K nearest neighbors with known baseline and the most similar clinical features. When we compare the predictive performance for only patients with imputed baseline, the KMBS method shows better performance than MBS. However, when we consider the entire set of patients, the model performance is comparable for both methods as the majority of patients have baseline ratings.

For evaluating LSTM M12 prediction, we consider three rating thresholds for occurrence (≥ 1), mild-severe (≥ 3), and moderate-severe (≥ 5) symptoms. We evaluated performance using RMSE, AUC, and F1 score metrics and show that LSTM predictions outperform other models, with ARIMA showing comparable performance in some cases and with the clinical regression models underperforming in most cases. The better performance of LSTM and ARIMA models indicates that the use of longitudinal PRO is highly predictive of long-term symptoms.

In conclusion, we have shown that LSTM can accurately predict late symptoms for oropharyngeal patients. Our ultimate goal is to embed symptom prediction into a clinical decision support tool that can be used to quantify the potential risks for late symptoms and allow physicians to preemptively prescribe exercises or medication to help patients cope with the symptoms or avoid them together improving the long-term QoL of patients. In future work, we would like to extend this work to predict time-to-event and account for the gap difference between collected time points.

ACKNOWLEDGEMENTS

This work was partially supported by NIH award NCI-R01-CA258827.

REFERENCES

[1] Kyte D.G. Aiyegbusi O.L. et al. Rivera, S.C. The impact of patient-reported outcome (pro) data from clinical trials: a systematic review and critical analysis. *Health Qual Life Outcomes*, 17(156), 2019.

[2] C. S. Cleeland, T. R. Mendoza, X. S. Wang, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer*, 89(7):1634—1646, 2000.

[3] Kaitlin M. Christopherson, Alokananda Ghosh, Abdallah Sherif Radwan Mohamed, et al. Chronic radiation-associated dysphagia in oropharyngeal cancer survivors: Towards age-adjusted dose constraints for deglutitive muscles. *Clin. Transl. Rad. Onc.*, 18:16–22, September 2019.

[4] A. Wentzel, P. Hanula, T. Luciani, et al. Cohort-based T-SSIM Visual Computing for Radiation Therapy Prediction and Exploration. *IEEE Trans. Vis. and Comp. Graphics*, 26(1):949–959, January 2020.

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

[6] Yaohua Wang, Guadalupe M Canahuate, Lisanne V Van Dijk, Abdallah S. R. Mohamed, Clifton David Fuller, Xinhua Zhang, and Georgeta-Elisabeta Marai. Predicting late symptoms of head and neck cancer treatment using lstm and patient reported outcomes. In *Proceedings of the 25th International Database Engineering amp; Applications Symposium*, IDEAS '21, page 273–279, New York, NY, USA, 2021. Association for Computing Machinery.

[7] H. M. Skerman, P. M. Yates, and D. Battistutta. Multivariate methods to identify cancer-related symptom clusters. *Res. Nursing & Health*, 32(3):345—360, 2009.

[8] David I. Rosenthal, Tito R. Mendoza, Clifton D. Fuller, Katherine A. Hutcheson, X. Shelley Wang, Ehab Y. Hanna, Charles Lu, Adam S. Garden, William H. Morrison, Charles S. Cleeland, and G. Brandon Gunn. Patterns of symptom burden during radiotherapy or concurrent chemoradiotherapy for head and neck cancer: A prospective analysis using the university of texas md anderson cancer center symptom inventory-head and neck module. *Cancer*, 120(13):1975–1984, 2014.

[9] S. J. Coons, S. Eremenco, J. J. Lundy, et al. Capturing Patient-Reported Outcome (PRO) Data Electronically: The Past, Present, and Promise of ePRO Measurement in Clinical Trials. *The patient*, 8(4), 2015.

[10] D. I. Rosenthal, T. R. Mendoza, M. S. Chambers, et al. 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):923—931, 2007.

[11] S. A. Eraj, M. K. Jomaa, C. D. Rock, et al. 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.

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

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

[14] G. Fan, L. Filipczak, and E. Chow. Symptom clusters in cancer patients: a review of the literature. *Current oncology (Toronto, Ont.)*, 14(5):173–179, 2007.

[15] A. Aktas, D. Walsh, and L. Rybicki. Symptom clusters: myth or reality? *Palliative medicine*, 24(4):373—385, 2010.

[16] S. T. Dong, D. S. Costa, P. N. Butow, et al. 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):88—-98, 2016.

[17] Ratnadip Adhikari and R. K. Agrawal. An Introductory Study on Time Series Modeling and Forecasting. *CoRR*, abs/1302.6613, 2013.

[18] Steven Elsworth and Stefan Güttel. Time Series Forecasting Using LSTM Networks: A Symbolic Approach, 2020.

[19] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, et al. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Frontiers in Big Data*, 3:4, 2020.

[20] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, et al. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. *CoRR*, abs/1707.02485, 2017.

[21] G. Maragatham and S Devi. LSTM Model for Prediction of Heart Failure in Big Data. *J Med Syst*, 111, 2019.

[22] Zachary C. Lipton, David C. Kale, Charles Elkan, et al. Learning to Diagnose with LSTM Recurrent Neural Networks, 2017.

[23] Shyam Boriah, Varun Chandola, and Vipin Kumar. *Similarity Measures for Categorical Data: A Comparative Evaluation*, pages 243–254.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, 9:1735–80, 12 1997.