2018 17th IEEE International Conference on Machine Learning and Applications

Recurrent Neural Networks based Obesity Status Prediction Using Activity Data

Qinghan Xue*†, Xiaoran Wang*, Samuel Meehan*, Jilong Kuang*, Jun Alex Gao* and Mooi Choo Chuah†
*Samsung Research America. Mountain View, CA 94043
Email: {x.wang, s.meehan, jilong.kuang, alex.gao}@samsung.com
†Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015
Email: qix213@lehigh.edu; chuah@cse.lehigh.edu

***Abstract*—Obesity, a serious public health concern worldwide, increases the risk of many diseases, including hypertension, stroke, and type 2 diabetes. To tackle this problem, researchers collect diverse types of data, which includes biomedical, behavioral and activity, and utilize machine learning techniques to mine hidden patterns for obesity status improvement prediction. While existing machine learning methods such as Recurrent Neural Networks (RNNs) provide exceptional results, it is challenging to discover hidden patterns of the sequential data due to the irregular observation time instances. Meanwhile, the lack of understanding of why those learning models are effective also limits further improvements on their architectures. Thus, we develop a RNN based time-aware architecture to handle irregular observation times and identify relevant feature extractions from longitudinal patient records for obesity status improvement prediction. Evaluations of real-world data involving activity data collected from wearables and electronic health records demonstrate that our proposed method can capture the underlying structures in users' time sequences with irregularities, and achieve an accuracy of 77% in predicting the obesity status improvement.**

***Index Terms*—obesity surveillance, activity data, hidden patterns, sequential data, recurrent neural network**

I. Introduction

In recent years, the advent and rapid adoption of mobile health (mHealth) [1], [2] enabled by wearable technologies have made continuous monitoring of environment and lifestyle a concrete possibility. For example, tiny low cost sensors such as smart watches and wristbands track a variety of parameters that range from steps taken and hours slept, to heart rate variability, which enable intelligent healthcare applications to provide estimates of steps, calories, fitness assessment, rehabilitation, activity duration, and may even offer disease pre-diagnosis [3]–[5].

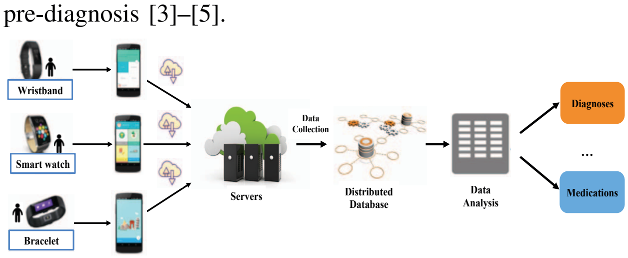


Fig. 1: End-to-end Data Flow

Fig 1 shows an example of a system where data streams are collected using various sensors/mobile devices and analyzed to understand users' physical activity levels, and possible disease forecasting. A particular disease of interest is obesity since it is a critical worldwide problem. Early detection and status monitoring of such patients allow better care.

While studies have shown that machine learning methods can provide good performance in various healthcare applications for personalized disease diagnosis, medication, and treatments, it is still challenging to learn efficient patterns from heterogeneous healthcare data. To overcome this, recently, deep learning techniques (e.g., RNN) have been adopted in medical patterns and patient representation learning. For example, Long-Short Term Memory (LSTM) [6], one of such popular variants which can handle long-term event dependencies by utilizing a gated architecture, has recently been applied to health informatics [7]–[9] with promising results.

Despite such successes, discovering hidden patterns of the sequential data is still an open challenge since it requires intelligent segmentation and clustering of the time series data. For example, the time lapse between successive elements in patient records can vary from days to months, which may lead to suboptimal performance for the traditional LSTM models. In addition, it is also difficult to interpret their impressive performance, especially when the data is high-dimensional, which in turn limits the ability to design better architectures. To address these challenges, three research questions have been raised as follows:

- (1) RQ1: "Can we predict diseases status based on individual records?"
- (2) RQ2: "How to build an appropriate learning model that can deal with the irregular data collection times, and learn hidden patterns from time series features?"
- (3) RQ3: "How does one interpret such a predictive model?"

In order to tackle those research questions, in this paper, we take the obesity disease as a use case and solve an obesity status improvement prediction task using activity data collected via a mobile phone application named FeatForward. We recruit 275 participants who are obese and collect activity data using this phone application for 6 months. The application collect useful information such as previous diagnoses, blood test results, and activity levels (e.g., step counts).

Then, based on these collected data, we develop a Recurrent Neural Network (RNN) based time-aware learning model that performs obesity status improvement prediction. In addition, we interpret the behavior of our learning model by analyzing the expected responses of the hidden state units given certain inputs. Finally, we evaluate the performance of our proposed

978-1-5386-6805-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMLA.2018.00139

model using a real-world dataset. The experimental results show that our proposed method can capture the underlying patterns in users' time series with irregular data collection time instances, and achieve an accuracy of 77% for the obesity status improvement prediction.

Our major contributions are:

- A mobile application FeatForward was developed to collect data from 275 participants for 6 months and a RNN-based time-aware model was designed to forecast individuals' obesity status.
- To handle the irregularity in data collection time instances, we also directly incorporate the day-week-month effect to improve the prediction performance.
- An analytical framework is developed to understand how the features contribute to our obesity status improvement predictive model.
- Extensive experimental results using real-world data show that our mechanism is accurate and generalizable.

The rest of the paper is organized as follows. Section II discusses related work. Section III provides brief descriptions of our data collection program. Section IV describes the proposed method in detail. Section V presents the evaluation results of our approach with real-world data. Section VI concludes the paper and highlights our future directions.

II. Related Work

In this section, we briefly review existing works in healthcare, which are closely related to our proposed method in this paper from two areas. The first one is recent works on clinical data mining and exploiting deep learning methods in the healthcare domain. The other is the existing explainable deep learning approaches that have utilized visualization to help understand machine learning models.

A. Clinical Data Mining

1) Conventional Machine Learning on Health Data: Clinical Data Mining (CDM) is the application of data mining techniques using clinical data [10], to extract relevant knowledge and make clinical decisions [11]. Marwa et al. in [12] have also presented an algorithm to generate prediction models based on the information gathered on patient's first physician visit. In addition, previous studies ([13], [14]) have used individual physical details including behavioral, sleeping, voice acoustic, and social patterns to estimate a person's mood. While their mechanisms are beneficial, they ignore the long-term dependencies among collected data records.

2) Deep Learning on Health Data: As deep learning has achieved great success recently, researchers have begun attempting to apply neural network based methods to clinical temporal health data [15]–[17]. For example, RETAIN and GRAM in [16], [17] are two state-of-the-art models utilizing RNNs for future disease predictions. While those models achieve good prediction performance, they do not exploit mobility or activity patterns for diseases monitoring and lack model interpretations.

B. Models Interpretation

In the field of visualization, most of the existing explainable deep learning approaches mainly focus on understanding and analyzing model predictions or the training process offline after the model training is completed. Recent work ([18]–[20]) has exhibited the effectiveness of visual analytics in understanding, diagnosing and presenting neural networks. While visualizations have achieved considerable success on deep learning models, they only provide an overall analysis without exploring hidden states in detail and are not scalable when hidden state dimensions have increased.

III. Data Collection

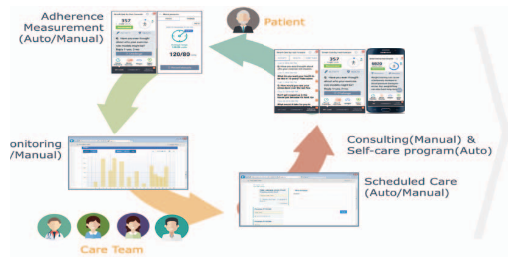


Fig. 2: FeatForward Study Overview

In our work, we use a smartphone application called FeatForward that was developed by our team to collect data from 275 participants for 6 months. 49.8% of the participants are over 50 years old and 50.2% are below 50. Such data allows us to infer three highly prevalent but often undiagnosed conditions including diabetes, hypertension, and obesity that participants may have. The application provides an intuitive interface (Fig 2) for users to record their data including gender, age, weight, blood pressure, etc. at three milestones (enrollment, midpoint, closeout). It also allows participants to voluntarily record their activity information (step counts) at different time slots (e.g., every 20 minutes or hourly) every day. The data can be divided into two 3-months periods, the 1^{st} one is from enrollment to midpoint while the 2^{nd} one is from midpoint to the end of the study. The application also shows the participants the trends of their health status after 3 months and encourages them to increase their activity levels if they do not achieve their health related goals (expressed in terms of clinically relevant threshold (CRT), a measure of the percentage reduction of their BMI values).

In our pilot study, we required all participants to wear the provided smartwatch during all hours except while they sleep for accurate and comprehensive data collection. The recorded information is shown in Table I.

IV. Methodology

In this section, we consider the obesity disease as a use case and utilize the collected data from the FeatForwad pilot study (described in section III) to generate clinical learning models to predict obesity status improvement and demonstrate the usefulness of using activity data collected via wearables for predicting health status.

A. Preliminaries

Different from other application domains (e.g., image and speech analysis), the problems in healthcare are more complicated. For example, the diseases are highly heterogeneous which make it hard for physicians to understand their causes and how they progress completely. Thus, in this sub-section, we first discuss the three research questions we raised in section I that helps to define our design requirements.

TABLE I: Collected Data Features

Classification	Detailed	Collection	Notes
Personal Data	gender, age, height, disease	manual input	occurs at registration
Biometric Data	step count	once per hour	obtained through S-Health
	blood pressure, blood sugar, weight, heart rate	manual input	through the S-Health GUI input window
Messaging History	activity/education	whenever event occurs	automatic program transmission
	communication	whenever event occurs	foctor-patient messaging
Environment	weather information		weather forecast

1) RQ1: Can we forecast obesity disease status based on individual records?: The question includes identifying specific measurements which can contribute towards improving prediction performance. In [21], Tryon has noted that step count information is a preferred metric for quantifying physical activity. Additional risk factors which impact prediction results that need to be considered include users' demographics and health histories e.g., his/her height, age, and fitness level.

2) RQ2: How to generate an appropriate model that can deal with irregular observation times?: In the pilot study, participants record their activity information at different time slots each day, and how their BMIs vary at the 3 milestones. Such raw data collected via the FeatForward application can be broken down into segments of time series data and mined for predicting obesity status improvement. For example, the obesity status prediction can be described as a (m, n) task, where a feature vector x_d^P can be extracted from data recorded in day d of a participant P. $x_d^P = \{x_{d_1}^P, x_{d_2}^P, \ldots, x_{d_m}^P\}$, where $x_{d_m}^P$ is a sequence of feature vectors of a participant P collected at time slot m of day d. The change of a participant's BMI value (a measure of obesity status improvement) will be predicted based on his/her histories in the previous n days $\{x_1^P, x_2^P, \ldots, x_n^P\}$. In this work, we proposed a RNN based time-aware model to learn long-term dependencies among time series features to perform such a prediction.

3) RQ3: How to interpret such a learning model?: While recent developments in the health informatics community have seen widespread adoption of modern machine learning methods, relatively little attention has been paid to understand the properties of its representations and predictions. Thus, in our work we try to interpret the changes of individual hidden states based on different inputs.

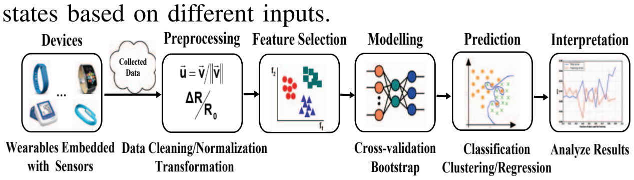


Fig. 3: Workflow of Our Method

B. Methodology Overview

The collected data is processed through several steps as shown in Fig 3. First, data cleaning and transformation is performed to remove missing information. Next, feature extraction and selection is carried out to produce learning models with high prediction accuracy. Finally, the learnt model need to be interpreted to uncover the impact of different features on the prediction results.

C. Detailed Design

In this sub-section, several key steps for processing data, namely, data cleaning, data transformation, feature extraction and data analytics/modeling are described in detail respectively.

TABLE II: Extracted Features

Features	Type	Description
Step Information	Time Slot	step count at every time slot
	Extracted Features	day: average steps, maximum/minimum steps
		week: maximum/minimum steps
		total steps
		how frequently does the participant walk
		# of days that has larger steps than the average steps (per day/week) of a specific participant
		# of days that has larger steps than the average steps (per day/week) of all participants
Demographic		gender, age, marital, adult in household, highest degree, hispanic or latino, race, occupation

1) Data Cleaning & Normalization: Collected data is noisy with some missing data. Typically, missing data may be caused by either participants forgetting to wear the sensor devices or failing to record their data through the mobile application. In our work, we replace any missing value with the average value obtained from non-missing entry values of that feature.

2) Data Transformation: To handle the issue of participants collecting data at irregular observation times, we introduce a day-week-month variable to organize participants' data into different non-overlapping time "windows" with each window being k days. Such transformation makes analysis more efficient but also has some side effects. For instance, unwise segmentation may result in the sparsity and missing data problems since there could be no observations for some features in some time windows. In addition, by dividing longitudinal data into "windows", the model may be less sensitive to capturing long-term feature patterns.

3) Feature Selection & Extraction: Since no one knows precisely which features (participants' characteristics or activity information) are more critical for obesity status prediction, we conduct feature selection operations to assess their usefulness in constructing the learning models. Cross-sectional studies [22] have shown that the daily step count is inversely related to body mass index (BMI), hypertension, and diabetes and hence will be considered.

Though daily average step feature is an important factor in estimating obesity risk, it does not describe the frequency, intensity or duration of a person's physical activity. Thus, we also include additional features: (i) intensity: maximum or minimum steps per day/week; (ii) frequency: number of days where the daily step counts exceed a threshold (set to twice the average lowest daily step counts); (iii) duration of physical activity: a number of days that the participants will walk more steps than the average step count when compared with other participants. In addition, we also consider a patient's demographics since the steps that a person takes vary based on his/her height, age, and fitness level. For example, frail, elderly individuals tend to take slower and fewer steps while younger individuals often take more steps. All the features extracted and used in our work are shown in Table II.

4) Model Construction & Interpretation: Next, we build learning models to predict obesity status improvement. To

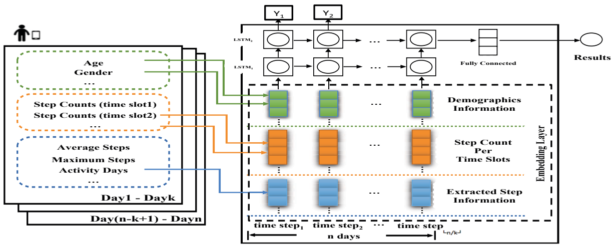


Fig. 4: Network Architecture of Our Method

compare the effectiveness of the neural network models, we first trained two traditional machine learning classifiers, namely Logistic Regression (LR) and Random Forest (RF).

(1) Baseline Classifiers

The reason why we select LR and RF as our baseline classifiers is because: (i) the L2-regularizer in LR classifier is more robust to limited training examples; (ii) the RF classifier is a robust and ensemble-based machine learning method. Both of these models assume all features are independent and not time correlated.

(2) Deep Learning Models

For deep learning models, we consider a Long Short-term Memory (LSTM) network since it can capture correlations across different behavioral sequences over time. The network architecture of our model is illustrated in Fig 4, where inputs grouped every k days are fed into the model and prediction results are generated as the output $Y_{\lfloor n/k \rfloor}$ at the final time step $\lfloor n/k \rfloor$.

First, we segment the collected time series data into different segments, each lasting k days and generate corresponding features from the collected records in each segment, as shown in Table II. For instance, $k = 7$ means the data collected every week is used to generate useful features, e.g., the average step counts per time slot per week, the average daily step counts, the maximum weekly step counts, etc.

An embedding layer is also introduced to convert those raw data with different types of features with distinct semantics into a dense vectorial representation. The outputs of the embedding layer will be fed into a 2-layer LSTM model, where the first layer is used to extract features from the vectorial representation and the 2^{nd} layer is used to learn high level abstraction representations. The output of the 2^{nd} layer is fed to a fully connected layer to perform prediction.

(3) Learning Models Interpretation

Inspired by the idea of using interpretable representations [20] to explain functions of network components, we use a similar method to interpret the learned representation of our proposed model.

At each time step t, the model takes an input X_t^i, and updates the hidden state using:

$$h_t^i = f(Wh_{t-1} + VX_t^i) \tag{1}$$

where W and V are weight matrices and f is a nonlinear activation function. After the updates, h_t^i is considered to capture the long-term memory and used to compute the intermediate or final output. $\triangle h_t^i = h_t^i - h_{t-1}$ is deterministic with regards to the input X_t^i once the previous history h_{t-1} is given, and can be used to reflect the degree a hidden state unit in the model is influenced by the input. Thus, based on $\triangle h_t^i$, the model's response to the input X_t can be computed as follows:

$$s(X_t) = E(\triangle h_t | X_t) = \frac{\sum_{i=1}^{|X_t|} |\triangle h_t^i|}{\sum_{i=1}^{|X_t|} 1} \tag{2}$$

where $s(X_t)$ is the average absolute expected responses of $|X_t|$ input sequences from all participants. For each input X_t^i, we can perturb the value of its various dimension corresponding to different features to see which one causes the largest change in $s(X_t)$ and hence determine which feature plays more important role in the prediction task.

V. Performance Evaluation

In this section, we train both the traditional classifiers and our RNN-based deep learning model to predict obesity status improvement suign the collected dataset. Then, we present our results in dealing with the research questions that are described in Section I.

A. Experimental Setup

To evaluate the performance of our scheme, we conduct experimental evaluations on real-world data, which is collected by the FeatForward application described in Section III. It tracks 275 participants for a 6-month period which can be divided into two 3-months periods, namely from enrollment to midpoint (1^{st} 3 months) and from midpoint to end of the study (2^{nd} 3 months). The measured metrics consist of an individual's time series of daily step counts (measured using 20 minutes time windows), his/her medical records and demographics (as shown in Table II). One can use statistics from different time windows to derive learning models. For example, we can conduct training using measurements collected with different day time windows ($k =$ one day, one week, one month) and daily time slots (i.e., $m = 4, 6, 12$). Since some participants dropped out after the midpoint or they do not record enough step counts information, so after data cleaning, we obtain dataset D_1, which contains 323 instances. Each instance contains historical measurements of a participant for 3 months. Next, we conducted 10-fold cross validation, where the cross validation splits the datasets into training (80%), validation (10%) and testing sets (10%). Our training model is used to predict whether a participant's BMI change exceeds the clinically relevant threshold (CRT) at the end of 3 months. For our training, we apply a dropout mechanism and conduct regularization to overcome any possible overfitting problem.

In addition, since in the collected datasets, participants share their BMI measurements, we can use such measurements to label each participant as positive or negative instance depending on whether their BMI change percentage exceeds the CRT. In our case, we set CRT to 5% since past studies, e.g, [23] have shown that it is reasonable to achieve an average weight loss of 5-10% during 3-6 months. With such labeling method, we obtain 53% positive and 47% negative instances.

For our obesity status improvement learning model, we use 25 hidden units for our LSTM and apply a dropout of 0.5. We train our model for 150 epochs, where each epoch is defined as the process of feeding the whole training set to a model. All our experiments are conducted on Mac Pro with an Intel Core i7 processor running at 2.5GHz, 16GB memory and an external GTX 1080.

TABLE III: Impact of Different Selected Features

Selected Features	Time Slot	Feature Size	Accuracy (LR)	Accuracy (RF)
Time Slots	4	4	54.4%	53.1%
Time Slot+Demographic	4	12	55.9%	54.7%
All Features	4	23	59.0%	63.1%
Time Slots	6	6	54.2%	56.5%
Time Slot+Demographic	6	14	56.2%	57.8%
All Features	6	25	60.6%	63.7%
Time Slots	12	12	53.8%	53.2%
Time Slot+Demographic	12	20	55.1%	53.8%
All Features	12	31	56.9%	60.2%

B. Performance Evaluation

1) RQ1: Can we forecast obesity health status improvement? How to select essential features?: In this subsection, we conduct Exp1 to evaluate the forecasting performance of traditional models. Such performance results are used as baselines for comparison with deep-learning based model in subsequent subsection.

Exp1: We first use the traditional classifiers LR and RF (described in section IV) to generate the learning models using different combinations of extracted features (Table II) so that we can compare their performances: (i) we only use daily average step feature to build the models; (ii) instead of using the daily average step count feature, we use step counts measured using m different time slots; (iii) we use both time slot features and demographics to train the learning models. In this experiment, we use dataset D_1 with different m values ($m = 4, 6, 12$).

The results are shown in Table III. From the results, we can observe that each of the two classifiers (LR or RF) achieves significantly higher accuracy than random guessing (e.g., 0.5). It is better to use all features to generate learning models since the obesity status improvement depends not only on the number of step counts but also on the frequent, intensity, duration of a participant's physical activity bouts as well as his/her demographics. In addition, the performance of RF is better than LR (by 3% on average) when all features are used. It is expected since RF is less likely to overfit and it also learns better the correlations among different features. Moreover, we find that the performance of all models varies when m increases from 4 to 12, with the best performance being achieved when $m = 6$. This can be explained as follows: more noises in the training data will be observed with a larger number of time slots (e.g., 12) since more time slots will have zero measurements. With fewer number of time slots (e.g., 4), the observed step counts for different time slots may not differ much from one another and hence affects the prediction results.

2) RQ2: Which model performs best in obesity status improvement predictions?: We conduct two experiments (Exp2.1 & Exp2.2) to examine how the selection of window size k affects our results. For example, $k = 30$ means that the input of a predictive model is trained based on the data extracted from every month.

Exp2.1: Since those traditional classification approaches (Exp1) assume all features are independent and do not consider the time correlation of these features, we use a single layer LSTM architecture in this experiment so that we can see if such learning model can perform better than traditional methods. We also set $n = 90$, $k = 3, 7, 30$ and assign $\lfloor n/k \rfloor$ (i.e., 30, 12, 3) time steps for such LSTM model, where the input of each node is the related feature vector generated based on k days.

TABLE IV: Impact of Different Time Window Sizes

Model	Time Slot	Feature Size	Window Size	Step Nodes	Accuracy (Validation)	Accuracy (Testing)
LSTM (1 layer)	6	25	3 days	30	61.3%	60.9%
LSTM (1 layer)	6	25	1 week	12	69.4%	68.7%
LSTM (1 layer)	6	25	1 month	3	64.8%	64.3%

TABLE V: Impact of Different Learning Models

Model	Time Slot	Feature Size	Window Size	Step Nodes	Accuracy (Validation)	Accuracy (Testing)
LR	6	25			61.1%	60.6%
RF	6	25			64.3%	63.7%
CNN	6	25			65.9%	65.6%
LSTM (1 layer)	6	25	1 week	12	69.4%	68.7%
LSTM (2 layers)	6	25	1 week	12	77.6%	77.2%
LSTM (3 layers)	6	25	1 week	12	73.1%	72.5%

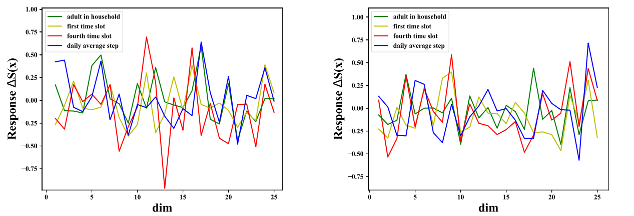
Exp2.2: Instead of only using a single layer LSTM as in Exp2.1, we use other learning models, which include a 2-layer CNN model and multi-layer LSTMs.

The results are shown in Table IV and V. From the results, we can discover that there is a rapid accuracy increase from $k = 3$ to $k = 7$, but the improvement saturates as k increases beyond 7 and in fact drops when $k = 30$. The results reveal that the step counts captured in a week contains sufficient information to allow us to forecast obesity status improvement at the end of the 3 months period. This finding coincides with a commonly used activity assessment method called the International Physical Activity Questionnaire (IPAQ) [24], where the central question it has is "The Time People Spent Being Physically Active in the Last 7 Days".

The CNN and LSTM-based deep learning models perform better than the traditional learning models with the LSTM-based model performs 15% better. However, the CNN model performs poorer than LSTM models since it lacks the capability to learn long-term correlations among time series features. In addition, the results also show that among all learning models, the 2-layer LSTM model has the highest accuracy. In this 2-layer LSTM model, the first layer acts as a feature extractor and the second layer learns the correlations among long-term dominant features. The prediction performance drops when more than 2 layers are used for the LSTM-based model since overfitting begins to appear in this complicated model.

In addition, we have also separated the participants into 2 age-group (above/below 50 years old) and trained a 2-layer LSTM model for each age group. We obtain an accuracy of 80.7% for those above 50 years old and an accuracy of 74.6% for those below. This can be explained as follows: the weight loss of younger adults is also greatly influenced by their diets which is not captured in the application. Older adults typically have better diets and hence their weight loss can be more easily predicted using step counts.

3) RQ3: How to interpret the learning model: In order to better understand our learning scheme, we conduct Exp3 to explain the inner working of our 2-layer LSTM model and rationale behind its predictions.



(a) Interpretation of the 10^{th} Time Step (b) Interpretation of the 12^{th} Time Step

Fig. 5: Two-layer LSTM Model Interpretation

Exp3: Inspired by the idea of using interpretable representations to explain functions of network components, we propose a method to interpret the hidden state units using different inputs. Since our model typically has many-to-many relationships between hidden state units and inputs, we perform the following steps to study how different variables affect the various hidden state units: (i) we first choose four features (shown in Fig 5), (ii) for each selected feature, we compute $S(X_t)$ values with all input features as well as with modified features where that selected feature is zeroed out and record the differences in these two $S(X_t)$ values, (iii) based on the computed difference values, we determine which time steps has the largest impact on the prediction results. In this experiment, we set $m = 6$ (each time slot equals 4 hours).

After checking the prediction accuracy of each time step, we found two crucial time steps (e.g., 10^{th} & 12^{th}) and further analyze their hidden units' responses. The hidden units' responses of these two time steps are shown in Fig 5 (a)(b). From the results, we can see that the hidden state units in the left and right ends are more responsive/sensitive to the important feature "daily average step" than the less important feature "adult in household". We also find that different time slot features yield different impacts on the expected responses of hidden units, with the "fourth time slot" having the most effect. This coincides with the observation that the five highest hourly step counts reported by participants fall under the 4^{th} time slot and the first time slot has the least activity. These results confirm that our model learns relevant information and allow us to infer the most relevant variables that can be used to help participants improve their BMI losses e.g., be more active during the "fourth time slot".

VI. Conclusions and Future Works

In recent years, new healthcare applications utilizing emerging smart devices with embedded sensors for improving users' health have become popular. In this paper, we have proposed a RNN based time-aware architecture to predict obesity status improvement using participants' data collected via wearables, e.g., blood pressures, step counts and their demographics. Our experimental results confirm that our framework can decently forecast obesity status improvement using users' activity and health measurement data. Furthermore, we have also provided some interpretations on how different variables affect our model. Thus, we believe that our effort is a good step in understanding how collected activity data and physical health measurements can be utilized to predict users' health status improvement. As for the future work, we intend to apply our predictive model to infer improvements in hypertension or diabetes health statuses. In addition, further prediction accuracy improvement can be made if we can integrate additional information such as participants' diets.

References

[1] A. J. Bandodkar, I. Jeerapan, and J. Wang, "Wearable chemical sensors: Present challenges and future prospects," *Acs Sensors*, vol. 1, no. 5, pp. 464–482, 2016.

[2] T. Quisel, L. Foschini, A. Signorini, and D. C. Kale, "Collecting and analyzing millions of mhealth data streams," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1971–1980.

[3] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," *Computer networks*, vol. 54, no. 15, pp. 2688–2710, 2010.

[4] R. D. Caytiles and S. Park, "A study of the design of wireless medical sensor network based u-healthcare system," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 3, pp. 91–96, 2014.

[5] L. Filipe, F. Fdez-Riverola, N. Costa, and A. Pereira, "Wireless body area networks for healthcare applications: Protocol stack review," *International Journal of Distributed Sensor Networks*, vol. 11, no. 10, p. 213705, 2015.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 507–516.

[8] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 198–206.

[9] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[10] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, "Clinical data mining: a review," *Yearbook of medical informatics*, vol. 18, no. 01, pp. 121–133, 2009.

[11] I. Epstein, *Clinical data-mining: Integrating practice and research*. Oxford University Press, 2009.

[12] M. Elamin, P. Bede, A. Montuschi, N. Pender, A. Chio, and O. Hardiman, "Predicting prognosis in amyotrophic lateral sclerosis: a simple algorithm," *Journal of neurology*, vol. 262, no. 6, pp. 1447–1454, 2015.

[13] P. Chow, H. Xiong, K. Fua, W. Bonelli, B. A. Teachman, and L. E. Barnes, "Sad: Social anxiety and depression monitoring system for college students," 2016.

[14] Y. Huang, H. Xiong, K. Leach, Y. Zhang, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes, "Assessing social anxiety using gps trajectories and point-of-interest data," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 898–903.

[15] N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," *arXiv preprint arXiv:1511.07938*, 2015.

[16] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[17] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 787–795.

[18] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[19] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 233–240.

[20] Y. Ming, S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu, "Understanding hidden memories of recurrent neural networks," *arXiv preprint arXiv:1710.10777*, 2017.

[21] W. W. Tryon, *Activity measurement in psychology and medicine*. Springer Science & Business Media, 2013.

[22] D. R. Bassett, L. P. Toth, S. R. LaMunion, and S. E. Crouter, "Step counting: a review of measurement considerations and health-related applications," *Sports Medicine*, vol. 47, no. 7, pp. 1303–1315, 2017.

[23] "NHI TEXT BOOK," https://www.nhlbi.nih.gov/health-pro/guidelines/current/obesity-guidelines/e_textbook/txgd/4311.htm.

[24] "International Physical Activity Questionnaire," https://snaped.fns.usda.gov/materials/international-physical-activity-questionnaire-opaq.