

# An Interpretable Prediction Model for Obesity Prediction using EHR Data\*

Mehak Gupta<sup>1,†</sup>, Thao-Ly T. Phan<sup>3,4</sup>, George Datto<sup>3,4</sup>, Timothy Bunnell<sup>1,2</sup>, Rahmatollah Beheshti<sup>1,5,†</sup>

<sup>1</sup> Computer and Information Sciences, University of Delaware, Newark, DE, USA

<sup>2</sup> Department of Biomedical Research, Nemours Alfred I. duPont Hospital for Children, Wilmington, DE, USA

<sup>3</sup> Department of Pediatrics, Nemours Alfred I. duPont Hospital for Children, Wilmington, DE, USA

<sup>4</sup> Department of Pediatrics, Thomas Jefferson University, Philadelphia, PA, USA

<sup>5</sup> Epidemiology Program, University of Delaware, Newark, DE, USA

† Email: [mehakg@udel.edu](mailto:mehakg@udel.edu) (MG), [rbi@udel.edu](mailto:rbi@udel.edu) (RB)

## ABSTRACT

Childhood obesity is a major public health challenge. Obesity in early childhood and adolescence can lead to obesity and other health risks in adulthood. Early prediction and identification of high-risk populations can help to prevent its development. With early identification, proper interventions can be used for its prevention. In this paper, we build prediction models to predict future BMI from baseline medical history data. We used unaugmented Nemours EHR (Electronic Health Record) data as represented in the PEDSnet (A pediatric Learning Health System) common data model. We trained variety of machine learning models to perform binary classification of obese, and non-obese for children in early childhood ages and during adolescence. We explored if deep learning techniques that can model the temporal nature of EHR data would improve the performance of predicting obesity as compared to other machine learning techniques that ignore temporality. We also added attention layer at top of rnn layer in our model to compute the attention scores of each hidden layer corresponding to each input timestep. The attention score for each timestep were computed as an average score given to all the features associated with the timestep. These attention scores added interpretability at both timestep level and the features associated with the timesteps.

## KEYWORDS

• Childhood obesity • Electronic health records • Temporal data

## 1 Introduction

Childhood obesity is a major public health problem across the globe as well as in the US. In 2019, the prevalence of obesity was 18.5% affecting almost 13.7 million US children and adolescents aged 18 or less [1]. Childhood obesity can continue into adulthood and is known to be a major risk factors for chronic diseases such as diabetes, cancer, and cardiovascular diseases [2]. Preventing childhood obesity has been actively pursued in pediatric programs. However, decades of rigorous research and experiments have shown that prevention is not easy [3]. This is partly due to our limited understating of the disease and the complex interactions of a myriad of various factors that are known to contribute to obesity. These factors include biological, social and environmental ones. Additionally, knowing the limited resources available to the healthcare systems, identifying children at the highest risk of developing obesity is another obstacle facing prevention programs. In such a complex domain, predictive models have been shown to be effective in informing decision makers and providers in designing and delivering prevention interventions.

In this paper, we present a predictive model of childhood obesity developed using a longitudinal dataset of children derived from the electronic health records (EHR) of a pediatric healthcare system. EHR data consists of clinical data along with its related temporal information. EHR datasets are generally very sparse and

complex due to the amount of information captured and irregular sampling. EHR data consists of records for each visit, which consists of conditions diagnosed, drugs prescribed, procedures performed, and laboratory results recorded for a visit. The number of unique condition diagnosis, drugs, procedures, and lab results collected in EHR data is huge. This leads to a very large feature space for our prediction model. However, each visit will only have very small subset of total unique conditions, drugs, procedures and measurements recorded. Due to the sparse feature space associated with each visit, we removed features which are absent or not recorded in more than 98% of the population. For all the data between 0 to 20 years of age, we used observation window of 2 years starting from 0 to 15 years and predicted obesity 1, 2 and 3 years in future. Models take data in observation window as input data from age  $x$  to  $x+2$  years and predict the output label at ages  $x+3$ ,  $x+4$  and  $x+5$  years, where  $x$  ranges from 0 to 15.

Both the model and the approach are unique. Our model uses a larger dataset (44 million rows with 68029 unique patients) for training and considers a larger set of confounders for predicting outcomes. Unlike other obesity prediction models which focus at single age point in childhood and adolescence[4], we created different prediction models based on the input baseline data and the age point in future at which the outcome label is to be predicted. Beside these, the main contribution of this study is presenting predictive models that consider the temporal changes of the children’s health patterns. A large body of research has shown that childhood obesity patterns are sensitive to different patterns of weight gain such that more acute and rapid weight gain predicts a different severity of obesity than more chronic and gradual weight gain [5]. Traditional statistical approaches rely on aggregated data, which ignores the temporality of data. We used a Recurrent Neural Network (RNN) architecture with Long Short-term Memory (LSTM) cells which learns the patient representation from the temporal data collected over various visits of the patient. This patient representation captures the temporality of input EHR data. Additionally, as the major drawback of deep learning models like RNNs is the lack of interpretability, we have used embedding weights on input layer and softmax activations on LSTM layers to calculate the importance of features and attention weights for each input timestep. The importance score for features and attention weights for timesteps were visualized to interpret important features and timesteps at individual and population level. This LSTM model is trained to predict future BMI value and then classify the BMI value as obese and non-obese. Apart from the time series data collected for visits of patients over time, EHR data also contains static data. The static data in EHR data does not change with every visit. This data consists of sex, race, ethnicity and zip code for each patient. We used separate feed-forward network for the static data and concatenated the output from this feed-forward network to the outputs obtained from LSTM cells.

Our models can predict the body mass index (BMI defined as height in kg over height squared in meter) in various ages. Having the estimated BMI values, we specifically look at the problem of classifying children as obese (above 95th percentile), and non-obese at the ages between 3 and 20 years according to the growth charts for children and teens provided by US Center for Disease Control (CDC) [6]. As the model predicts the BMI values fairly, it can be used for answering similar types of questions other than what is the focus of the current study. This specific problem (classifying obese versus non-obese) relates to identifying those at the highest risk.

The main contributions of this paper are (i) The prediction model which uses LSTM cell layers on multivariate irregularly spaced time series data to predict outcome at 3 different time points in future. We proposed a model architecture and mechanism to add interpretability to the lstm model for multivariate time-series data. We also added embedding layer and softmax layer to our proposed model which scores the importance of each feature and timestep for the predictive task. This mechanism adds interpretability at both feature and timestep level for the predictive task by computing embedding weights on input layer and timestep attention weights on the lstm layer. This provides insights into important clinical events at individual and population level. (ii) Performance comparison between machine learning techniques that ignore temporality with the machine learning models that capture temporality in the data in predicting childhood obesity.

## 2 Related Work

### 2.1 Recurrent Neural Networks (RNN)

Recurrent neural nets take advantage of the concept of parameter sharing across the model. Unlike the basic feedforward network where each input feature is learned separately and have separate parameters, recurrent neural nets share the parameters and generalize the model across different forms on input. This property of RNN is used for many NLP problems where the same information can be found in different locations depending on the formation of input data. For temporal data also RNN can be used to learn long-term dependencies by sharing parameters through the deep computational graph.

RNN works by updating the value of the hidden state at each time step based on both the current state and the value of the hidden state of the previous time stamp. This helps propagate the information from all previous hidden states at earlier time steps to hidden state at the current time step.

However, there is one mathematical problem with remembering long-term dependencies over time using recurrent neural nets. The problem of giving low weights to long-term interactions as compared to short-term interactions. This is the problem of vanishing gradient where gradient associated with error as we move to earlier time steps will decay exponentially. This happens due to the repeated operation of multiplying the gradient with matrix, which at time step  $t$  is the same as multiplying with  $t$  derivate of the matrix. Therefore, after a few time steps gradient becomes 0.

This problem of vanishing gradient is approached by a gating mechanism implemented in Long Short-term memory (LSTM) cells and Gated Recurrent unit cells. In this paper we are using LSTM cells.

#### 2.1.1 Long Short-Term Memory

Hochreiter et al. [7] introduced the gating mechanism where the gradient can flow for long durations. These gates learn to keep important information and throw irrelevant information from previous time steps. This way they pass on the important information in the network for long durations.

LSTM has forget gates that decide what information to keep or forget from previous hidden states and the input gate decides what information is important from the current state. Forget gates makes a decision using sigmoid function and input gates first use sigmoid layer to decide what information to update from current input and then use tanh layer to form an intermediate candidate vector from current input that can be added to get current cell state. The final output is generated by the dot product of output from forget gate and input gate. [8]

### 2.2 Time Series Analysis of Electronic Health Records

Clinical predictive models are becoming more and more prevalent [9]. Until recently most of the clinical predictive models were primarily developed based on regression and logistic regression or other types of statistical analysis [4]. Over the last decade, there has been an increase in the medical data collected in the form of EHR. To use the traditional methods, input features need to be selected by medical domain experts. Traditional methods (including the machine learning ones) are also not very effective in capturing the non-linear and temporal relationships in the complex EHR data. Recently, deep learning techniques have shown a lot of success in clinical predictive modeling [10]. Many clinical predictive models have been developed using deep learning techniques to predict various health problems like heart failure [11] [12] [13], diabetes[14], high blood pressure [15], and hospital readmission [16]. To the best of our knowledge, there are very few works that use deep learning for obesity prediction in early childhood and adolescence using EHR data from medical facilities in the United States.

However, and despite the urgent need, there is not a lot of work done in the field of obesity predictive modeling leveraging large scale datasets and advanced machine learning techniques. Most of existing work rely on traditional machine learning methods. Example studies include using logistic regression [17] [18] [19] [20], linear regression [21], and the random forest [22]. Our study used EHR data and applied the deep learning technique to capture the temporal nature of the data. Additionally, available predictive models of obesity focus on predicting BMI or overweight or obese label at a single age point in the future based on certain baseline data [23] [24]. These single-point prediction models are not generalized to predict the future

BMI trajectory starting from various points in early childhood and adolescence. Obesity prevalence is 13.9% among 2- to 5-year-olds, 18.4% among 6- to 11-year-olds, and 20.6% among 12- to 19-year-olds [1]. Since, obesity is prevalent in all age groups in childhood and adolescence, our prediction model predict obesity for all ages between 3 to 20 years instead of focusing on any single age point in childhood and adolescence. In our work, we have built different models using different input data, which can be used to keep track of future BMI trajectories from infancy to adolescence. We also analyzed results from all the models to see what age range is more accurate to predict BMI at a certain age in the future. Our work does not assume if any particular age range is best suited for predicting BMI at a certain age point in the future. We did not focus on any particular critical age range to predict BMI at a certain age in the future.

Also, our model is based only on the EHR data already available in the hospitals. Some work used questionnaires [25], and census data to predict obesity [22]. We did not use any other external data, and as the features that we use are commonly recorded in any standard EHR system, our models can be readily applicable to many healthcare systems. This also means that our models can be used with no additional cost in collecting any external data.

## 2.3 Prediction Model Interpretability

The major drawback of deep learning models is the lack of interpretability. The lack of interpretability reduces the value of prediction models especially in medical domain. If medical practitioners cannot understand how the outcome is predicted by a model, it is difficult to rely on the results. Many attempts have been made recently to make sense of the outcome of these models. The attention mechanism proposed by [26] is used in NLP for machine translation. This attention mechanism improves interpretability at time level i.e it gives attention scores to timesteps, but for multivariate time-series we also need to look at feature importance at each timestep. Choi et al. [27] develop an interpretable model with two levels of attention weights learned from two reverse-time GRU models, respectively. In our work we continue the use of attention mechanism to improve interpretability of the RNN based models for multivariate time-series to get importance score for timesteps and then get the importance score for each feature in the timesteps.

# 3 Data

## 3.1 Dataset description

The data was extracted from the Nemours Children Health System, which is a large network of pediatric health in the US primarily spanning the states of Delaware, Pennsylvania, and New Jersey. The dataset is a portion of larger dataset of the larger PEDSnet dataset [28] containing EHR data from over 10 major US Children’s Health Systems. Inclusion criteria for patients in our dataset included: (i) At least 5 years of medical history. (ii) No evidence of Type 1 diabetes. (iii) No evidence of Cancer, Sickle Cell Disease, Developmental Delay, or other complex medical conditions. (iv) An equal number of normal weight and overweight or obese patients were selected by random sampling from the normal weight population. The dataset was anonymized. Further details about the anonymization process are provided in Supplemental Materials. All of the dates were skewed randomly per patient by +/- 180 days. All the steps were approved by Nemours Institutional Review Board. The dataset consists of 44,401,791 records from 68029 distinct patients. Each record captures the timestamp for each visit start and end time and all the condition, procedure, drug, and measurement variables recorded for each visit. It also contains demographic data for each patient along with its date of birth. Some facts about the data are listed in Table 1.

**Table 1: EHR Data Statistics**

Total number of patients	68,029
Total number of visits	44,401,791

Avg. number of visits per patient	51	
Number of females	31,014 (45%)	
Number of Males	37,015 (54%)	
Avg age of a patient	5	
Race and Ethnicity	White or Caucasian	33244
	Black or African American	25329
	Non-Hispanic or Latino	58894
	Others	17834

## 3.2 Data Representation and Preprocessing

The EHR data extracted for this study consists of 20,300 condition diagnosis variables, 10,167 procedure variables, 6,163 drug variables, and 7,693 lab-results (measurement) variables. All the condition and procedure variables were recorded as binary variables in the original data. Binary variables are the variables which are recorded as 1 if present and 0 if not recorded for the visit. Few drug variables were recorded as continuous variables where the values contain information about the amount of drug prescribed to a patient in a visit. However, many drug variables were recorded as binary variables and did not have the amount of drug prescribed information in the cohort. Measurement variables in the cohort were recorded as continuous variables. These continuous variables were normalized for model training.

We represented EHR data using code-level, visit-level representation and patient-level representation. This means that each record in the data representation corresponds to one visit recorded in the EHR data. EHR data consists of patient records as sequence of visits with each visit containing various medical codes. The medical codes are standardized terminologies of SNOMED-CT, RxNorm, CPT, and LOINC for both clinical and demographic facts. All medical codes are represented as code-level representation and each visit is a visit-level record that is set of all medical codes and each patient is a sequence of visit records for that patient. More details about these representations are provided in sections below.

### 3.2.1 Code-level Representation:

In code-level representation, medical codes consist of all the unique condition, drug, procedure and measurement variables in the complete data. We denote condition codes with the vector  $C: \{c_1, c_2, \dots, c_{|C|}\}$  with a size of  $|C|$ , drug codes with the vector  $D: \{d_1, d_2, \dots, d_{|D|}\}$  with a size of  $|D|$ , procedure codes with the vector  $P: \{p_1, p_2, \dots, p_{|P|}\}$  with a size of  $|P|$ , and measurement codes  $M: \{m_1, m_2, \dots, m_{|M|}\}$  with a size of  $|M|$ .

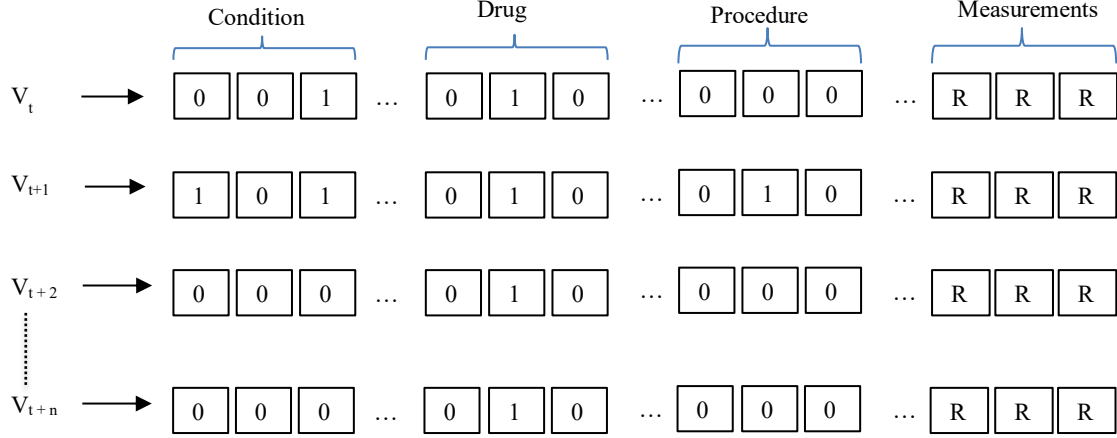
### 3.2.2 Visit-level Representation:

We denote visit at time  $t$  as  $V_t$ , which is the concatenation of condition, drug, procedure and measurement code vectors. The size of  $V_t$  is  $|V_t| = |C| + |D| + |P| + |M|$ . We represented condition, drug and procedure codes for visit  $V_t$  as binary vectors  $C_t \in \{0, 1\}^{|C|}$ ,  $D_t \in \{0, 1\}^{|D|}$ , and  $P_t \in \{0, 1\}^{|P|}$  respectively where “1” represents the presence of the corresponding code for a visit  $V_t$ . All the measurement variables were represented by the corresponding continuous values  $M_t \in \mathbb{R}^{|M|}$  for visit  $V_t$ . Figure 1 depicts the visit-level representation of our EHR data. EHR data for each patient is the sequence of visit-level vectors for that patient.

### 3.2.3 Patient-level Representation:

Patient-level representation is sequence of visit vectors for the patient. We denote patients  $S: \{s_1, s_2, \dots, s_{|N|}\}$  as  $S$ , where the  $i$ -th patient  $s_i$  with  $n$  visits is represented as matrix  $s_i \in R^{n \times |V|}$

Fig 1. Visit-level Representation of EHR data



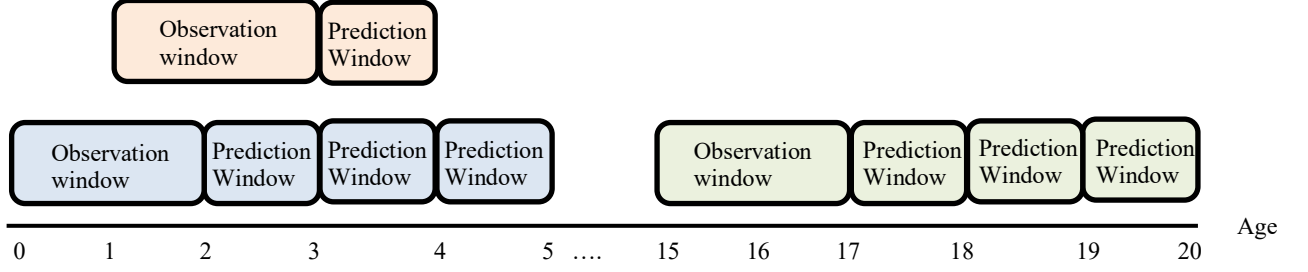
The EHR data also consist of demographic data. Demographic information consists of sex, race, ethnicity and zip code (indicating the approximate location of the patient). We represented demographic variables i.e., sex, race, ethnicity and zip code as category variables. Table 1 shows the distribution of race and ethnicity distributions in the data. Insurance information is also represented as category variable.

The visit-level representation of complete EHR data consists of large number of features including all unique condition, drug, procedure, and measurement variables. Many of these features are not present in most of the population. We removed the features which do not have enough information (the event occurred in at least 2% of the population in the cohort) to reduce the number of sparsity in the feature space. The feature space reduced to 3% when we only considered features that have enough information. Total number of features in final cohort were 1737.

We divided final data into sub-cohorts for different age ranges to predict obesity between 3 to 20 years of age. The data is extracted such that each patient has at least 5 years of data, and sub-cohorts for every 5 years of age range are created. Due to reasons like relocation, change in insurance families change hospitals, etc., any patient might not have data for all ages from 0 to 20 years. 5 years is feasible range where a patient has data for any consecutive 5 years at same medical facility. To get enough samples for training and testing of the prediction model we divided the complete cohort into 5 years of age range starting from 0 to 15 years of age which resulted in 16 age cohorts. If we pick more than 5 years of data then number of patients decrease as there are fewer patients who have records of more than 5 years at one facility.

For every 5 years data, we used a fixed observation window of 2 years and predicted obesity for 1, 2 and 3 years in future. Three sub-cohorts are created for each 5 years of age range such that each patient in the sub-cohort has at least one visit in the observation window and one visit at the age of prediction. For example, for creating the sub-cohort for age range of 0 to 5 years, 3 sub-cohorts are created such that patients included in the sub-cohort have at least one visit in the observation window of 0 to 2 years and at least one visit in the prediction window of 2 to 3 for predicting obesity at 3, at least one visit in the prediction window of 3 to 4 for predicting obesity at 4 and at least one visit in the prediction window of 4 to 5 for predicting obesity at 5 patient. Hence, there are 16 age cohort of 5 years time-period starting from 0 to 15. By creating 3 sub-cohorts for each 5 years time-period we derived 48 sub-cohorts. Fig. 2 depicts the way we created the sub cohorts and the observation window and prediction window for these cohorts. Prediction models are trained on data in the observation window to predict the future BMI value in the respective prediction window.

**Fig. 2 Sub cohort design and observation and prediction window for each sub-cohort. Three out of 16 example 5-year windows are shown.**



## 4 Method

Our proposed model is used to predict future BMI value. The predicted BMI value is used to classify patients as obese (more than 95th percentile) or non-obese (less than 95th percentile). The classification of BMI for different percentile is done according to the BMI-for-age charts provided by CDC [6]. This table provides label based on the age, gender and BMI for children from 24 months to 20 years of age. Children in the top 95 percentile are labeled as obese. For infants aged from 0 to 2 years classification is performed according to the Data Table of Infant Weight-for-age Charts provided by CDC [29].

### 4.1 Baseline Model

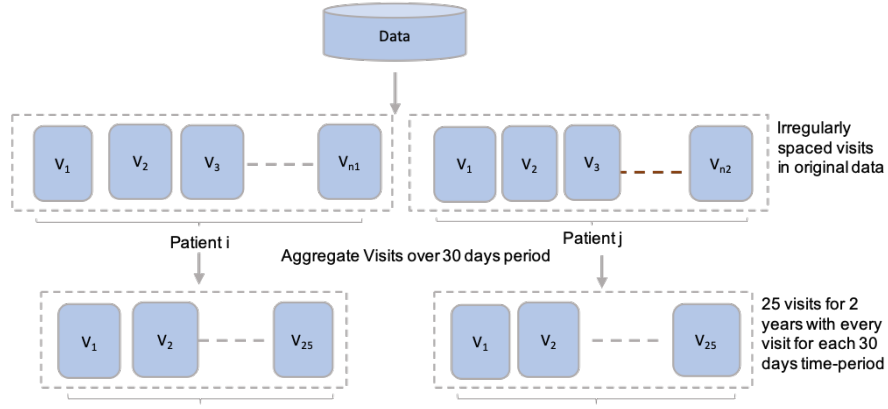
As briefly discussed in the Related Work Section, comparable predictive models of childhood obesity are not being used in clinical settings for screening or consulting (including at Nemours Health System where the data comes from). Therefore, to evaluate the performance of our proposed LSTM-based model, we created two baseline models that follow the traditional methods and aggregate the dataset while ignoring the temporality of the EHR data. We used linear regression and random forest regressor as the baseline models for comparison. To do this, we aggregated data over all the visits for each patient corresponding to any of the input sub-cohorts. All the visit records in the observation window are aggregated for each patient. Target labels will be the labels at the prediction age. Aggregation for binary medical codes of condition, drug and procedure type is performed such that each medical code represents the frequency of its occurrence over the 2 years, and for continuous variables we took average over the 2 years. For the BMI and body weight, we took the maximum BMI and bodyweight recorded in the observation window. We also took the last BMI and body weight recorded for the observation window. BMI is classified as non-obese (less than 85th percentile), and obese (more than 95th percentile).

### 4.2 LSTM model

After obtaining all the sub-cohorts as explained in Section 3.2, we transformed the data so that it can be given as input to the LSTM model. Clinical visits obtained in section 3.2 are represented by the medical codes associated with that visit. In general, (clinical) visits have irregular time intervals and each patient has a different number of visits. To transform these irregularly spaced and unequal number of clinical visits, we combined the visit data over a small fixed time window resulting in an equal number of time intervals. We combined visits over the 30-day time-periods for each observation window of the 2-year training windows, resulting in 25 equally spaced sequences for each patient. Fig. 3 shows how new sequences are obtained from unequal and irregularly spaced input time sequences. Any condition, drug and procedure variable observed at least once over 30-day time-period is denoted by 1 in new sequences. Continuous variables were averaged over the 30-day time-period. If there are no visits for a patient in any of the 30-day time-periods, the corresponding vector for that period contained all zeros. The zero vectors acted as padding to maintain equal

sequence length for all patients. Such equally spaced time intervals between input time series are preferred representation for RNN models. The width of the fixed time window is related to the stability of clinical events for the prediction task. We experimented with different window sizes of 6 months, 3 months, 30 days and 15 days. 30 days window size seemed to best capture the variations in clinical trajectories of patients for predicting obesity. In addition to conditions, procedure, drugs, and measurements the time intervals between each visit sequence were also added to the end of each visit's vectors. These time intervals capture the time intervals between the non-empty sequences. This procedure (adding time interval values) has been shown to enrich the time-series input in other similar studies [13].

**Fig. 3 Time sequences for LSTM model – Irregularly spaced visits for each patient (such as patient i and j) in each 2 year period is mapped to 25 distributed intervals.**



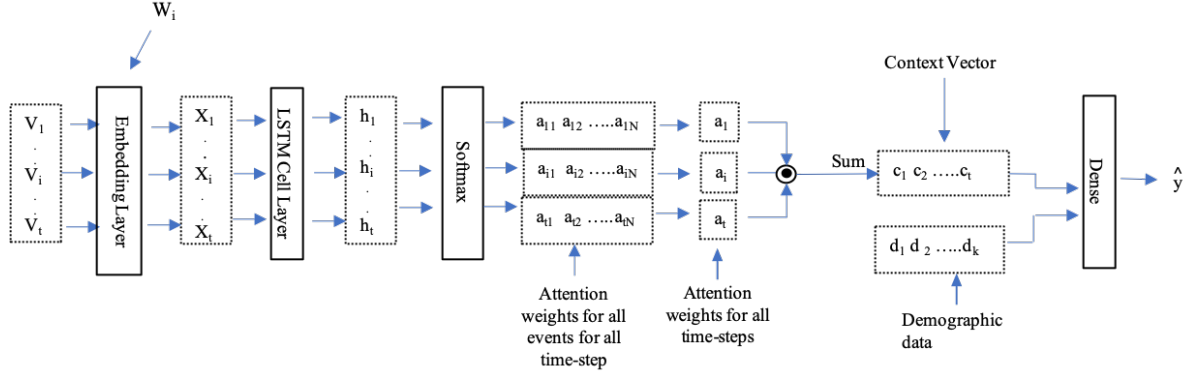
We used LSTM cells in the recurrent neural network for obesity prediction. The output of the LSTM layer is concatenated with the demographic input. This concatenated output is then passed through dense layers for predicting BMI value. The architecture of the complete model is shown in Fig. 4.

### 4.3 Interpretability

While deep learning models show superior performances compared to traditional machine learning models, they are difficult to interpret due to their so called “black box” architecture[30]. This may reduce the practicality of deploying them in medical domains. To mitigate such concerns, we enhanced our basic LSTM model to provide interpretability. Since our data consists of multivariate time-series we need interpretability such that we can score visits according to their importance in predicting output and also rank features present in the visit according to their importance in predicting output. We will refer to these two levels of interpretability as time-level and feature-level interpretability. Fig. 4 shows the enhanced model architecture to achieve interpretability. We will further elaborate on how we achieved the interpretability at both time-level and feature-level using the model architecture in Fig. 4.



**Fig. 4 LSTM Model Architecture with Interpretability**



#### 4.3.1 Interpretability at Time Level

To enhance the interpretability at a time level, we added a softmax layer on top of the LSTM layers to compute the “attention score” for each timestep. In general, each LSTM unit generates a hidden output at each time step. Hidden state  $h_j$  is computed by applying the non-linear transformation on input  $x_j$  to the LSTM unit at time  $j$  and hidden state of previous time step  $h_{j-1}$  (Eq. 1). Each hidden state  $h_j$  is of the length of the hidden layer size of the LSTM layer. Suppose  $h_j$  is of length  $|h_j| = N$ , then  $N$  softmax scores will be assigned to the  $h_j$ . We calculated attention score for each hidden state output by computing the softmax score for each value in all the hidden states and then taking an average of all the scores in each hidden state (Eq. 2, 3). These scores are then used to compute the weighted sum of hidden layers (Eq. 4). The vector  $c$  obtained is then used to predict future BMI. The scores computed using the softmax layer are used to visualize the visits that are given most importance by the LSTM layer.

Hidden state is calculated as follows,

$$h_j \leftarrow f(x_j, h_{j-1}), \text{ where } h_j \in R \quad (1)$$

If  $|h_j| = N$ , then attention score per hidden state will be calculated as,

$$b_1, b_2, \dots, b_j, b_{j+1}, \dots, b_t = \text{softmax}(h_1, h_2, \dots, h_j, h_{j+1}, \dots, h_t), \quad (2)$$

where  $b_j = a_{j1}, a_{j2}, \dots, a_{ji}, a_{ji+1}, \dots, a_{jN}$

$$a_j = \left( \sum_{i=1}^N a_{ji} \right) \div N \quad (3)$$

Attention scores for each timestep is then calculated as below,

$$c = \sum_{i=1}^n a_i * h_i \quad (4)$$

#### 4.3.2 Interpretability at Feature Level

We further enhanced the model to rank the input features in the multivariate time-series data. We added the embedding layer on top of the input layer. We used weights from the embedding layer to compute the importance score for features in each timestep. Softmax scores for the timestep are multiplied element wise with the embedding weight matrix for each input feature.

$$s_i = b_j \odot W_i \quad (5)$$

Eq. 6 shows the  $s_i$  importance score calculation for  $i^{\text{th}}$  feature, where  $b_j$  is the softmax score output after lstm layer and  $W_i$  is weight matrix for the  $i^{\text{th}}$  feature from the embedding layer.

Fig. 4 shows the complete architecture of the proposed LSTM model with interpretability mechanism.

#### 4.4 Transfer Learning

Transfer learning is used to enhance model performance by learning from a larger dataset. In our experiments, we created different sub-cohorts for different age ranges. As shown in Table 2 the number of samples reduced gradually with increasing age range. Due to the low number of samples model performance also decreases as will be seen in the next section. To improve the performance of the model we used the complete dataset for all age ranges. We initially created three models for predicting obesity at - 1 year in the future, 2 years in the future and 3 years in the future. After this, each of the three general models has been used as the basis for the 16 separate predictive models related to a similar prediction window.

### 5 Experiment and Result Evaluation

For training the LSTM models, we split data into 60:20:20 as training, validation and test data. Data split is performed such that the proportion of obese and non-obese samples is the same in train and test data as in original data. Table 2 shows the number of obese and non-obese samples in each sub-cohort are shown in Table 2.

We used two LSTM layers for all models and Adadelata optimizer with a learning rate of 0.01. Both L1 and L2 regularization were used on the first LSTM layer. Two fully connected layers were used for the feed-forward network for demographic data. We used the softmax layer and embedding weights to obtain the importance score at both time and feature level. We trained different models on different sub-cohorts based on different observation window and prediction age as explained in section 3.2. All models are trained on data in the observation window to predict the future BMI value in the respective prediction window. Then we classify the predicted BMI into obese and non-obese labels.

**Table 2. Number of Obese and Non-obese Samples**

Observation Window (Age years)	Prediction Age (Age year)	# of Obese Samples	# of Non-Obese Samples
0-2	3, 4, 5	5556, 7492, 8192	27842, 25906, 25206
1-3	4, 5, 6	7382, 8081, 8307	25464, 24765, 24539
2-4	5, 6, 7	6697, 6878, 7014	19977, 19796, 19660
3-5	6, 7, 8	5697, 5788, 6216	16227, 16136, 15708
4-6	7, 8, 9	4840, 5194, 5679	13492, 13138, 12653
5-7	8, 9, 10	4428, 4813, 5117	11182, 10797, 10493
6-8	9, 10, 11	4085, 4368, 4574	9032, 8749, 8543
7-9	10, 11, 12	3773, 3941, 4047	7542, 7374, 7268

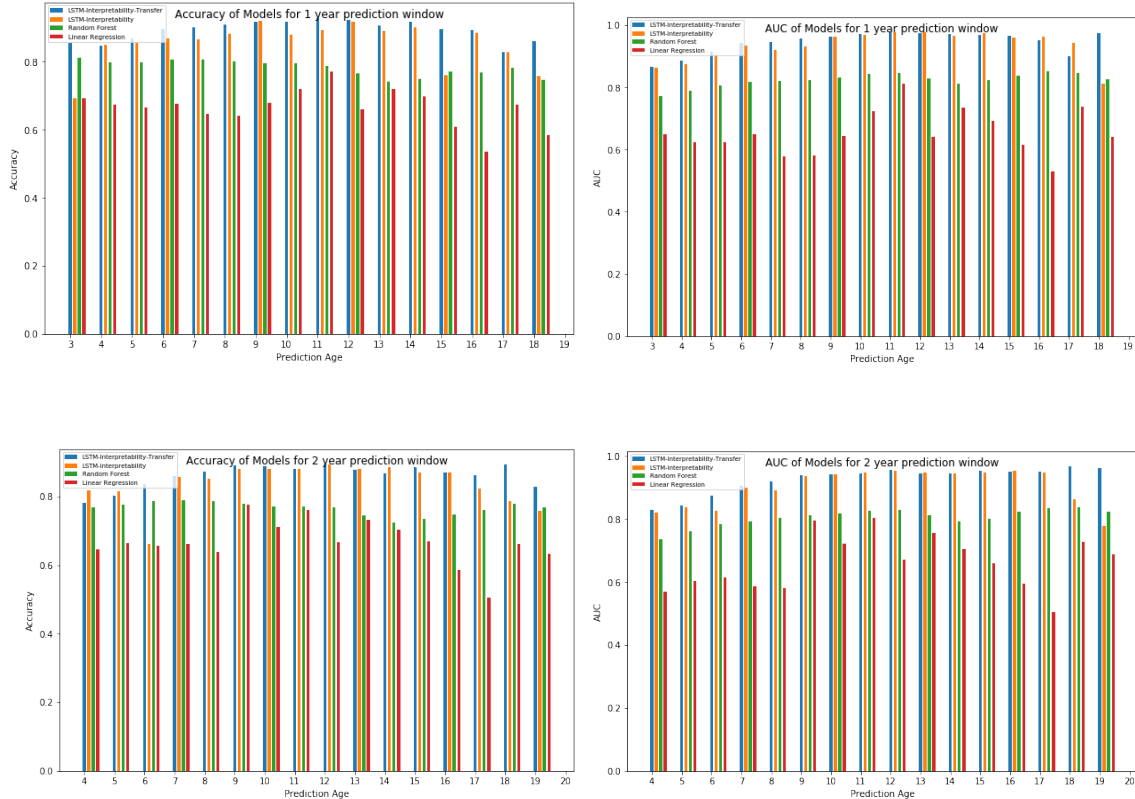
8-10	11, 12, 13	3273, 3354, 3387	6118, 6037, 6004
9-11	12, 13, 14	2671, 2729, 2692	4933, 4875, 4912
10-12	13, 14, 15	2116, 2078, 2078	3718, 3756, 3756
11-13	14, 15, 16	1507, 1502, 1530	2688, 2693, 2665
12-14	15, 16, 17	1052, 1059, 1087	1766, 1759, 1731
13-15	16, 17, 18	651, 665, 690	1044, 1030, 1005
14-16	17, 18, 19	250, 249, 260	358, 359, 348
15-17	18, 19, 20	55, 57, 55	87, 85, 87

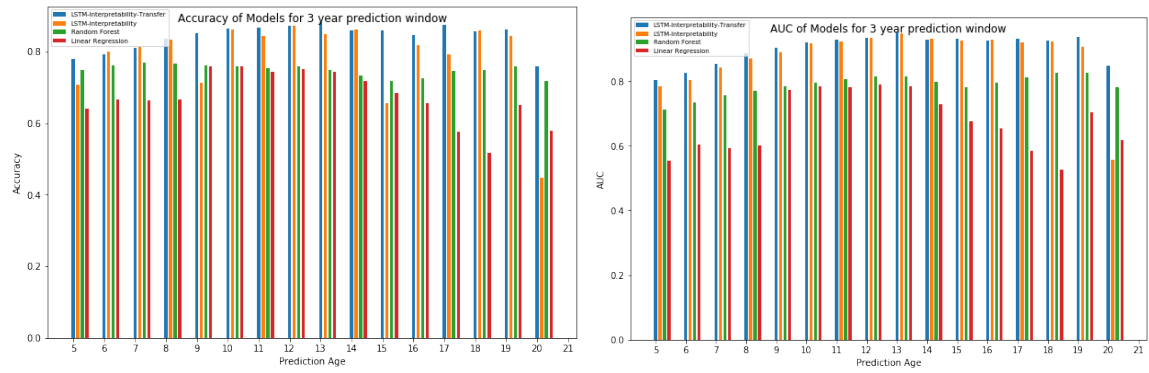
We evaluated and compared the performance of baseline models as explained in section 4.1, LSTM with interpretability as explained in section 4.3, and LSTM with Interpretability trained on the larger dataset using transfer learning as explained in section 4.4. We compared accuracy, and AUC scores from all the models. For baseline models of linear regression and random forest regressor, we did 10-fold cross-validation and reported mean results over complete data. For LSTM models we reported results on test data.

Fig. 5 shows the Accuracy and AUC for all models separately based on prediction window size.

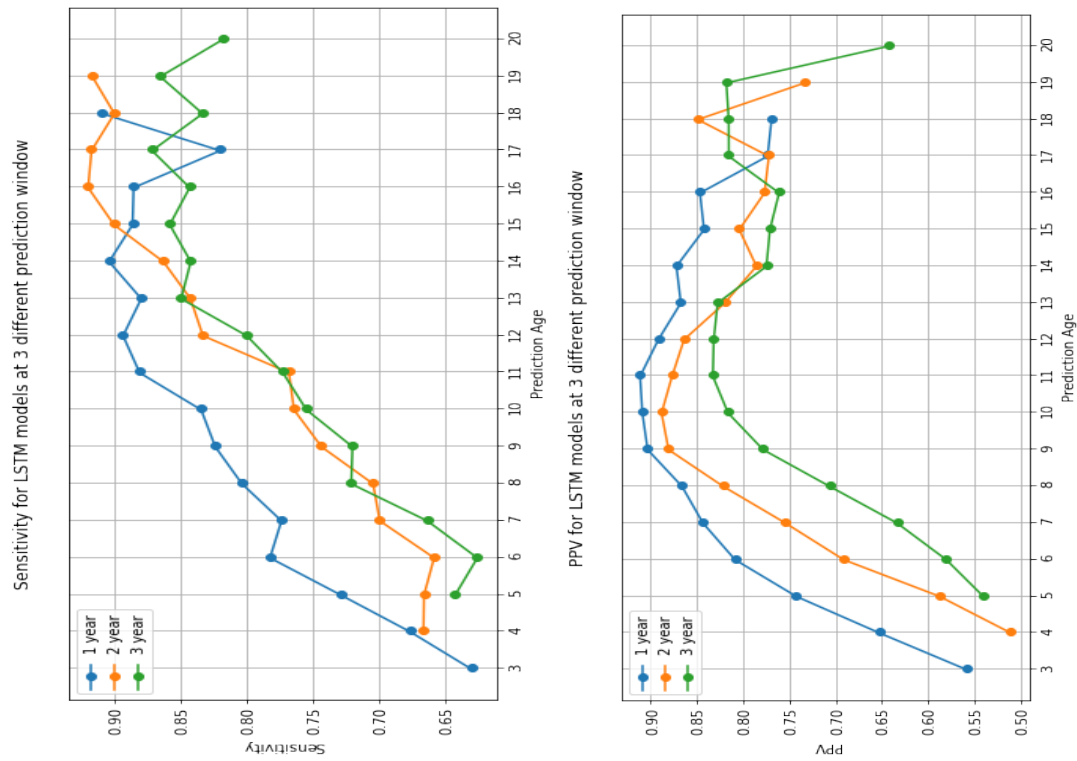
Fig. 6 further compares the results of the LSTM models based on prediction window size. It shows how LSTM models perform for different prediction age based on the size of the prediction window.

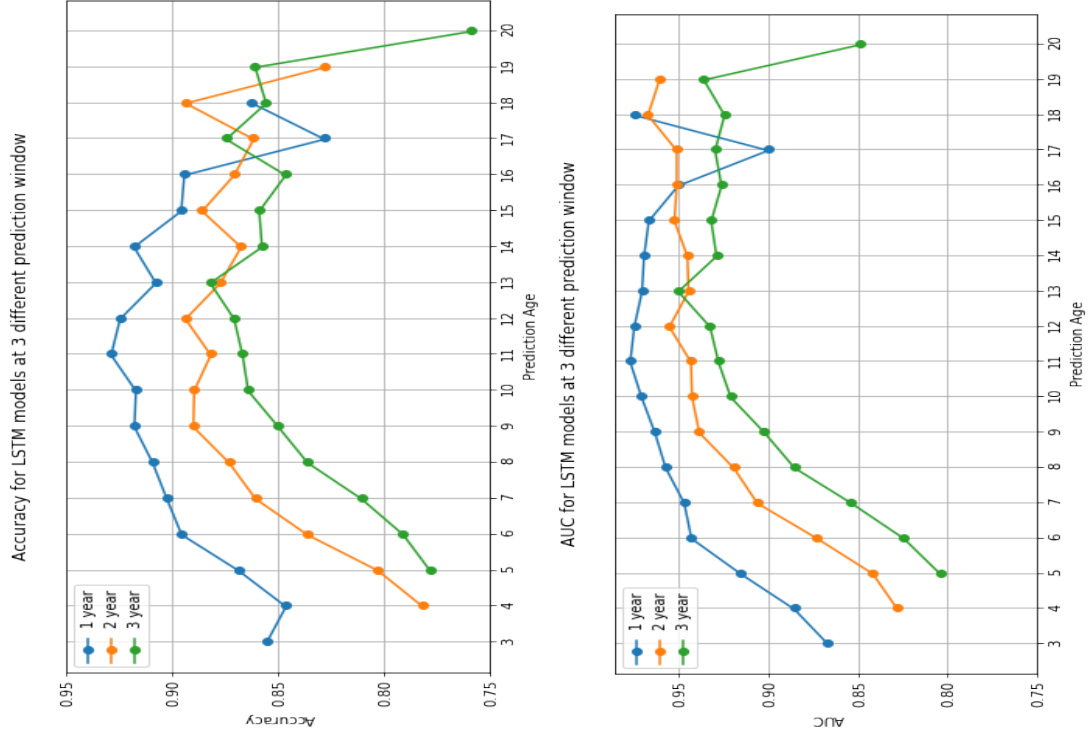
**Fig. 5 Comparing Accuracy and AUC Results of all Obesity Prediction Models**





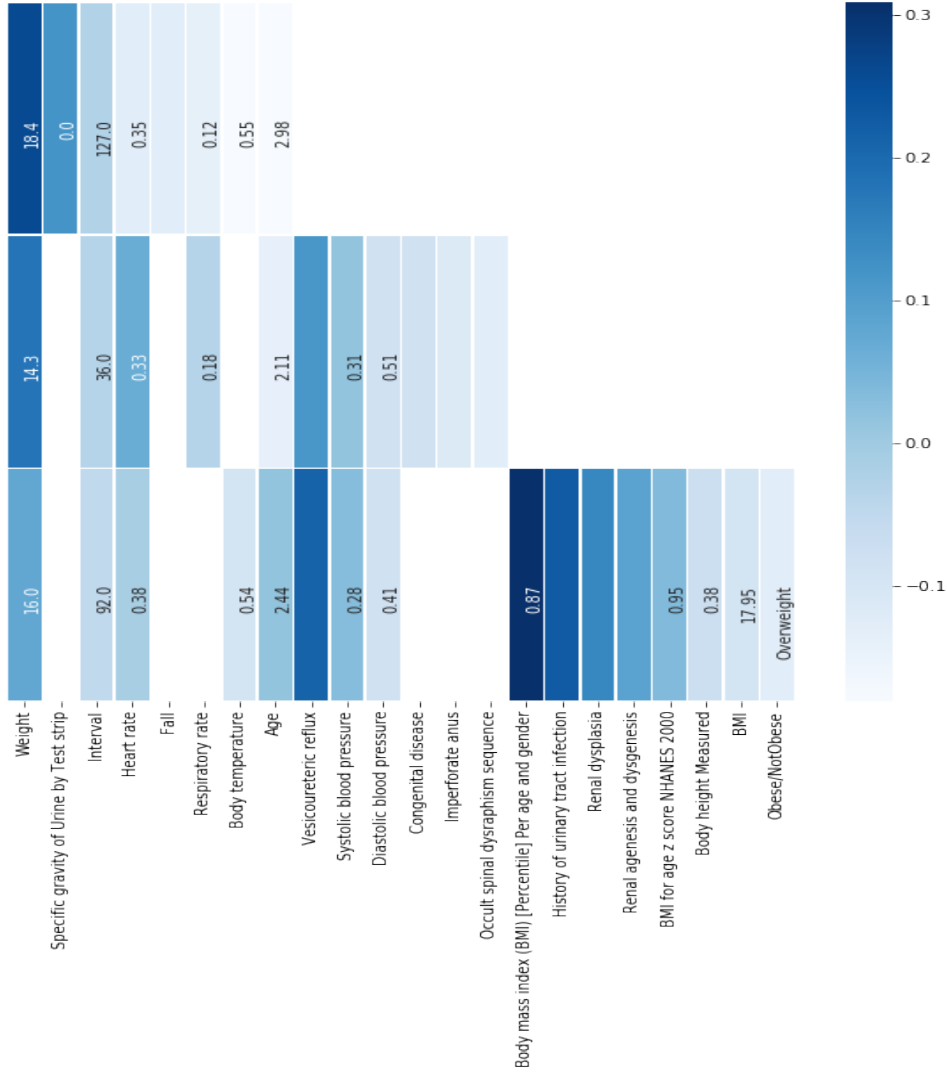
**Fig. 6 Results of LSTM with attention for different prediction window size**





We added the attention mechanism to the LSTM model by using softmax activation on the output of the LSTM network to interpret the model results. This allowed us to rank the timesteps and input features according to their importance in output prediction by using softmax activations on the LSTM layer and embedding weights on the input layer. Feature importance is computed at both the individual level and population level. Fig. 7 shows the ranking of the features in the top 3 important visits. This is feature importance for a sample individual patient. We also ranked feature importance at the population level by averaging feature importance for top 3 visits for all samples. The values in each cell of Fig. 7 is the measurement value for corresponding feature. Table 3 shows the top 20 most important features at the population level.

**Fig. 7 Feature Importance for Top 3 important timesteps**



**Table 3 Top 20 Feature Importance**

Top 20 Features
Body mass index (BMI) [Percentile] Per age and gender
Obese/Non-Obese Label
Allergic urticaria
Childhood obesity
Morbid obesity
Suspected clinical finding
Achondroplasia
MCH [Entitic mass] by Automated count
Cholesterol in LDL/Cholesterol in HDL [Mass Ratio] in Serum or Plasma
Hearing loss
Abnormal weight gain

Anomaly of chromosome pair 21
Erythrocytes [# /volume] in Body fluid
Obesity
Hyperactive behavior
Tachycardia
Requires respiratory syncytial virus vaccination
CO2 1712
pH of Blood
Hypoplastic left heart syndrome

## 6 Discussion

Our proposed model is applied to the obesity prediction in childhood and adolescence using EHR data. We employ the LSTM network and separate feed-forward network to model time series and static data in the EHR data. As shown in Fig. 5 the performance of the LSTM is better than the baseline model of Linear regression and random forest regressor. This shows that a recurrent neural network improves performance by taking into consideration the temporality of the data. This temporality is important to capture weight gain trajectories and other medical history overtime [5]. However random forest regressor shows higher performance for prediction at age of 20 using 3-year prediction window. This is due to the low number of samples in the corresponding sub-cohort and LSTM shows poor performance due to overfitting. But transfer learning helps improves performance for this sub-cohort by learning from samples of other sub-cohorts.

As shown in Fig. 5 the results obtained from LSTM model trained using transfer learning are higher as compared to LSTM trained on samples of specific sub-cohorts only. We can see that transfer learning helps improve prediction performance especially for cohorts with a low number of samples. We can see in Fig. 5 that the performance of the last sub-cohort has significantly improved over the model trained on corresponding sub-cohorts only. Data samples for last sub-cohort were very less and by using transfer learning it improves the performance of sub-cohorts with low number of samples.

In Fig. 6 we show the results of LSTM model with interpretability (shown in Fig. 4) trained using transfer learning. We can see from Fig. 6 that the closer the observation window is to the prediction time, the better the performance of the model. Also, the plots are all showing a bell curve. In the beginning performance increases and then it starts to decrease after a certain prediction age. The reason for decrease in performance for the second half is because of decrease in number of samples and visits for that sub-cohort. There is sharp decrease in accuracy for 3-year window as compared to 1-year and 2-year window. This is because the low number of samples in last cohort has more impact on larger prediction window as compared to smaller prediction window.

We have also ranked the features in each visit to provide insight into the prediction results. As shown in Fig. 7 we ranked the features for 3 most important visits. We pick the top 3 visits with highest attention weights and then ranked features for those visits. We ranked these features by calculating the importance score for features using Eq. 5. Here we can see that Weight and BMI are most important features which is expected for predicting obesity. Also, we can see that Vesicoureteral Reflux is given very high importance score. This condition is a type of kidney disease which is highly correlated to obesity in children. [31]

In Fig. 7 we see the feature ranking for one sample. We also calculated the feature ranking over the complete dataset (with samples that are predicted obese) to get population level feature ranking. This shows the most important features in predicting obesity in children. As expected, BMI and previous and existing obesity level has the highest impact. Cholesterol and abnormal weight gain are also correlated to obesity. Erythrocytes is related to kidney inflammation which could also be a sign for future obesity. Tachycardia and Heart rate are related to higher heart rate. Feature ranking also shows that Hyperactive behaviour is also an important factor in predicting future obesity. This coincides with the study in [32].

This shows that feature ranking obtained using the lstm model in Fig. 4 and calculated using Eq. 5 gives results that coincide with existing medical studies [33].

In future we can add attention to static demographic data as well. In this work we fixed the observation to 2 years of data, in future we can employ the proposed model with larger observation window size. We can also expand transfer learning training by using data from other medical facilities.

## 7 Conclusion

We applied lstm model to predict obesity. We can see that lstm model shows better performance as compared to traditional machine learning models. The performance obtained from our proposed model is comparable to performance in other cohort studies for obesity prediction. [22] [34] [17] [4] We also added interpretability to the prediction model. Interpretability is achieved by ranking features in top 3 most important timesteps. We also calculated feature ranking for all samples in the data that were predicted obese in future. This gave us the list of features ranked according to their importance in predicting future obesity. We also used transfer learning to train model on all age sub-cohorts which helped us improve performance for sub-cohorts that have very low number of samples.

## References

- [1] Centres for Disease Control and Prevention, “Childhood obesity facts,” 24-Jun-2019. [Online]. Available: <https://www.cdc.gov/obesity/data/childhood.html>.
- [2] W. H. Dietz, “Health Consequences of Obesity in Youth: Childhood Predictors of Adult Disease,” *Pediatrics*, vol. 101, no. Supplement 2, pp. 518–525, Mar. 1998.
- [3] S. Bleich, K. Vercammen, L. Zatz, J. Frelie, C. Ebbeling, and A. Peeters, “Interventions to prevent global childhood overweight and obesity: a systematic review,” *Lancet Diabetes Endocrinol.*, vol. 6, January 10.
- [4] N. Ziauddeen, P. J. Roderick, N. S. Macklon, and N. A. Alwan, “Predicting childhood overweight and obesity using maternal and early life risk factors: a systematic review,” *Obes. Rev.*, vol. 19, no. 3, pp. 302–312, 2018.
- [5] P. O. A. Monteiro and C. G. Victora, “Rapid growth in infancy and childhood and obesity in later life—a systematic review,” *Obes. Rev. Off. J. Int. Assoc. Study Obes.*, vol. 6, no. 2, pp. 143–154, May 2005.
- [6] N. C. for H. S. Centres for Disease Control and Prevention, “Data Table of BMI-for-age Charts,” 23-Aug-2001. [Online]. Available: [https://www.cdc.gov/growthcharts/html\\_charts/bmiagerev.htm](https://www.cdc.gov/growthcharts/html_charts/bmiagerev.htm).
- [7] J. Schmidhuber and S. Hochreiter, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning. vol. 1,” 2016.
- [9] G. Bedogni, A. B. Tsybakov, and S. Berlin, “Clinical prediction models—a practical approach to development, validation and updating,” *development*, vol. 18, no. 500, pp. 53–99, 2009.
- [10] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, 2017.
- [11] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, “Doctor ai: Predicting clinical events via recurrent neural networks,” 2016, pp. 301–318.
- [12] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, “Medical concept representation learning from electronic health records and its application on heart failure prediction,” *ArXiv Prepr. ArXiv160203686*, 2016.
- [13] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, “Using recurrent neural network models for early detection of heart failure onset,” *J. Am. Med. Inform. Assoc.*, vol. 24, no. 2, pp. 361–370, 2016.
- [14] T. Pham, T. Tran, D. Phung, and S. Venkatesh, “Deepcare: A deep dynamic memory model for predictive medicine,” 2016, pp. 30–41.
- [15] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, “Deep learning for healthcare decision making with EMRs,” 2014, pp. 556–559.
- [16] N. Wickramasinghe, “Deepr: a convolutional net for medical records,” 2017.
- [17] A. Morandi *et al.*, “Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts,” *PloS One*, vol. 7, no. 11, p. e49919, 2012.



- [18] M. Steur *et al.*, “Predicting the risk of newborn children to become overweight later in childhood: the PIAMA birth cohort study,” *Int. J. Pediatr. Obes.*, vol. 6, no. sup3, pp. e170–178, 2011.
- [19] Y. Manios *et al.*, “Childhood Obesity Risk Evaluation based on perinatal factors and family sociodemographic characteristics: CORE index,” *Eur. J. Pediatr.*, vol. 172, no. 4, pp. 551–555, 2013.
- [20] C. Druet *et al.*, “Prediction of childhood obesity by infancy weight gain: an individual-level meta-analysis,” *Paediatr. Perinat. Epidemiol.*, vol. 26, no. 1, pp. 19–26, 2012.
- [21] Z. Pei *et al.*, “Early life risk factors of being overweight at 10 years of age: results of the German birth cohorts GINIplus and LISAplus,” *Eur. J. Clin. Nutr.*, vol. 67, no. 8, p. 855, 2013.
- [22] R. Hammond *et al.*, “Correction: Predicting childhood obesity using electronic health records and publicly available data,” *PloS One*, vol. 14, no. 10, pp. e0223796–e0223796, 2019.
- [23] M. Adnan, W. Husain, and N. Rashid, “Parameter identification and selection for childhood obesity prediction using data mining,” 2012, vol. 35, pp. 75–80.
- [24] T. M. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs, “Machine learning techniques for prediction of early childhood obesity,” *Appl. Clin. Inform.*, vol. 6, no. 03, pp. 506–520, 2015.
- [25] J. O. Robson, S. G. Verstraete, S. Shiboski, M. B. Heyman, and J. M. Wojcicki, “A risk score for childhood obesity in an urban Latino cohort,” *J. Pediatr.*, vol. 172, pp. 29–34, 2016.
- [26] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *ArXiv Prepr. ArXiv14090473*, 2014.
- [27] 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,” 2016, pp. 3504–3512.
- [28] “Home,” *PEDSnet*. [Online]. Available: <http://pedsnet.org>. [Accessed: 01-Dec-2019].
- [29] N. C. for H. S. Centres for Disease Control and Prevention, “Data Table of Infant Weight-for-age Charts,” 23-Aug-2001. [Online]. Available: [https://www.cdc.gov/growthcharts/html\\_charts/wtageinf.htm](https://www.cdc.gov/growthcharts/html_charts/wtageinf.htm).
- [30] J. Krause, A. Perer, and K. Ng, *Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models*. 2016.
- [31] “Vesicoureteral Reflux (VUR) | NIDDK,” *National Institute of Diabetes and Digestive and Kidney Diseases*. [Online]. Available: <https://www.niddk.nih.gov/health-information/urologic-diseases/hydronephrosis-newborns/vesicoureteral-reflux>. [Accessed: 01-Dec-2019].
- [32] E. A. Fliers *et al.*, “ADHD is a risk factor for overweight and obesity in children,” *J. Dev. Behav. Pediatr. JDBP*, vol. 34, no. 8, pp. 566–574, 2013.
- [33] K. Sahoo, B. Sahoo, A. K. Choudhury, N. Y. Sofi, R. Kumar, and A. S. Bhadoria, “Childhood obesity: causes and consequences,” *J. Fam. Med. Prim. Care*, vol. 4, no. 2, p. 187, Apr. 2015.
- [34] S. F. Weng, S. A. Redsell, D. Nathan, J. A. Swift, M. Yang, and C. Glazebrook, “Estimating Overweight Risk in Childhood From Predictors During Infancy,” *Pediatrics*, vol. 132, no. 2, pp. e414–e421, Aug. 2013.