



Obesity Prediction with EHR Data: A Deep Learning Approach with Interpretable Elements

MEHAK GUPTA, University of Delaware, USA

THAO-LY T. PHAN and H. TIMOTHY BUNNELL, Nemours Children's Health, USA

RAHMATOLLAH BEHESHTI, University of Delaware, USA

Childhood obesity is a major public health challenge. Early prediction and identification of the children at an elevated risk of developing childhood obesity may help in engaging earlier and more effective interventions to prevent and manage obesity. Most existing predictive tools for childhood obesity primarily rely on traditional regression-type methods using only a few hand-picked features and without exploiting longitudinal patterns of children's data. Deep learning methods allow the use of high-dimensional longitudinal datasets. In this article, we present a deep learning model designed for predicting future obesity patterns from generally available items on children's medical history. To do this, we use a large unaugmented electronic health records dataset from a large pediatric health system in the United States. We adopt a general LSTM network architecture and train our proposed model using both static and dynamic EHR data. To add interpretability, we have additionally included an attention layer to calculate the attention scores for the timestamps and rank features of each timestamp. Our model is used to predict obesity for ages between 3 and 20 years using the data from 1 to 3 years in advance. We compare the performance of our LSTM model with a series of existing studies in the literature and show it outperforms their performance in most age ranges.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health informatics**;

Additional Key Words and Phrases: Childhood obesity, electronic health records, temporal data, deep learning, long short-term memory, transfer learning

ACM Reference format:

Mehak Gupta, Thao-Ly T. Phan, H. Timothy Bunnell, and Rahmatollah Beheshti. 2022. Obesity Prediction with EHR Data: A Deep Learning Approach with Interpretable Elements. *ACM Trans. Comput. Healthcare* 3, 3, Article 32 (April 2022), 19 pages. <https://doi.org/10.1145/3506719>

1 INTRODUCTION

Childhood obesity is a major public health problem across the globe and in the United States. In 2019, the prevalence of obesity was 18.5%, affecting almost 13.7 million US children and adolescents aged 18 or less [18]. Childhood obesity can continue into adulthood as a major risk factor for chronic diseases such as diabetes, cancer, and cardiovascular diseases [13]. Preventing childhood obesity has been actively pursued in pediatric programs. However, decades of rigorous research have shown that prevention and management of obesity are not easy [6]. This is partly due to our limited understating of obesity and the complex interactions among myriad factors,

Authors' addresses: M. Gupta, University of Delaware, Newark, Delaware, USA, 19716; email: mehakg@udel.edu; T.-L. T. Phan and H. Timothy Bunnell, Nemours Children's Health, Wilmington, Delaware, USA, 19803; emails: ThaoLy.Phan@nemours.org, Tim.Bunnell@nemours.org; R. Beheshti, University of Delaware, Newark, Delaware, USA, 19716; email: rbi@udel.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

2637-8051/2022/04-ART32 \$15.00

<https://doi.org/10.1145/3506719>

ACM Transactions on Computing for Healthcare, Vol. 3, No. 3, Article 32. Publication date: April 2022.

including biological and environmental ones, which are known to contribute to obesity. The motivation for this work is to use routinely captured **electronic health record (EHR)** data in pediatric facilities to predict obesity without using data sources that are generally unavailable in pediatric EHR datasets, including maternity, pre-maternity, or lifestyle data. While flawed and incomplete, such EHR data is what providers are working with, and having a model embedded in the EHR data already available at a pediatric facility can help providers identify at-risk patients without collecting extra data. Many predictors that other models use, such as maternal and pre-gestational data in more carefully crafted data collection studies, are not available to the providers at pediatric facilities. Young children are particularly underdiagnosed, with only 31% of overweight preschoolers versus 76% of overweight adolescents so identified [48]. Pediatric providers only infrequently use recommended **Centers for Disease Control and Prevention's (CDC)** age- and gender-specific BMI charts [51], explicitly designed to help screen for unhealthy weight trajectories. Accordingly, an objective tool might help providers identify high-risk patients and also drive the sensitive conversation with parents who do not recognize their children as overweight and may even fail to recognize the health risks of increased weight [15, 57].

In this study, we present a set of predictive models of childhood obesity using a longitudinal dataset of children derived from the EHRs of a large pediatric healthcare system. EHR data comprises clinical data along with its related temporal information. It can consist of many features with sporadic recording patterns. This makes EHR data sparse (especially when used in a one-hot encoding format for machine learning tasks). Also, EHR data is irregularly sampled, as there are no regular time intervals between the patients' visits. The number of unique condition diagnoses, medications, procedures, and lab results collected in EHR datasets is generally very large. This leads to an exceptionally large feature (input) space for a prediction model, despite each visit having only a small subset of total unique conditions, medications, procedures, and measurements recorded. The models we present in this study try to address these issues effectively. The EHR data used in this work comprises the records of patients' diagnosed conditions, prescribed medications, performed procedures, and recorded laboratory results in any visit. Compared to existing obesity predictive models in this domain, our model uses a much larger dataset (44 million rows with 68,003 unique patients) for training and considers a larger set of confounders for predicting outcomes. Our model is based only on the standard EHR data already available in many hospitals.

Besides only using the existing data on standard pediatric EHRs, another key difference between our models and the comparable models in this domain is considering the temporal changes in the children's health patterns. The major limitation of existing obesity models is twofold. First, available obesity models focus on single (or only a few) future point predictions for overweight or obese prediction (such as Santorelli et al. [61] at age 2, Weng et al. [72] at age 3, Redsell et al. [56] at age 5, Levine et al. [32] at age 5 and stratified by sex, Robson et al. [58] at age 5, Hammond et al. [23] at age 5, Druet et al. [14] at ages 7 and 14, Steur et al. [65] at age 8, Manios et al. [37, 38] at ages 9 and 13, Pei et al. [49] at age 10, and Graversen et al. [19] during adolescent years). These single-point prediction models cannot be generalized to predict the future BMI trajectories starting from various points in early childhood and adolescence. Obesity is prevalent in all age groups in childhood and adolescence. This makes the application of these models limited, as they cannot assist in predicting obesity status in other ages. The second limitation about existing models relates to using aggregated patterns instead of longitudinal patterns for developing the models [79]. For obesity, this is a major limitation, since rigorous research has shown that longitudinal patterns of obesity-related measures (such as the body weight) have a strong correlation with the future obesity patterns [43]. Similarly, a large body of research has shown that childhood obesity patterns are sensitive to distinct patterns of weight gain such that a more acute and rapid weight gain predicts a different severity of obesity versus a more gradual weight gain [43]. Aggregating EHR datasets (e.g., by calculating the average values) loses valuable knowledge from this type of longitudinal dataset. The presented models in this work follow a **recurrent neural network (RNN)** architecture with **long short-term memory (LSTM)** cells and can learn a patient's representation from the temporal data collected over various visits of the patient. As one of the major drawbacks of deep learning models like RNNs is the lack of interpretability, we use embedding weights on the input layer and softmax activations on LSTM layers of our networks to calculate the importance of

the features and attention weights for each input timestamp. We then used the importance score for the features and attention weights for timestamps to visualize different rankings. Apart from the temporal components, EHR data also contains static elements that do not change with every visit, such as sex, race, ethnicity, and zip code for each patient. We used a separate feed-forward network for the static data and merged its output with the main architecture.

Our models can predict **body mass index (BMI)** values from ages 3 to 20. According to the CDC [45], children in the highest 5th BMI percentile among the children of the same age and sex are considered obese. Having the estimated BMI values, we specifically look at the problem of classifying children as obese and non-obese. Specifically, the main contributions of this article include presenting a predictive model of childhood obesity that (1) uses unaugmented EHR data, (2) can learn complex sequential patterns, and (3) is trained and evaluated on a very large sample ($n = 68,003$ patients). Our models can predict obesity status at three different future time points (next 1, 2, and 3 years). Furthermore, we propose a mechanism to add interpretability to this model at both the feature and timestamp level for the predictive task, which provides insights into the important clinical events at individual and population levels. We perform comparisons between several machine learning techniques that ignore temporality and our-RNN based models that capture temporality in the data and show that our models can achieve better results. While we focus on an obesity-related problem, our approach should be also usable for studying similar problems using EHR data.

2 RELATED WORK

Predictive models have recently shown a lot of promise in estimating future health outcomes in various biomedical applications, a trend similar to many non-biomedical domains [16, 30, 68, 69]. Clinical predictive models are becoming prevalent and largely adopted due to a rapid increase in the availability of large-scale medical data [4]. Until recently, most of the clinical predictive models were primarily developed based on regression and logistic regression or other types of statistical analysis [79]. Traditional methods (including the machine learning ones) are 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 [62].

Many deep learning studies in this domain use RNNs, which refer to a subset of deep neural architectures used on sequential datasets. Unlike basic feed-forward networks, RNNs can learn the long-term dependencies in temporal data by sharing parameters through deep computational graphs. However, capturing long-term dependencies using RNNs faces the vanishing gradient problem where gradient values become too small. Hochreiter and Schmidhuber [25] introduced the LSTM gating mechanism where the gradient can flow for long durations. LSTM gates can learn to keep important information and discard irrelevant ones from the previous time steps. This way, they can pass on the important information in the network for long durations. For additional details about RNN and LSTM architectures, we refer the readers to other references [5, 25].

Many clinical predictive models have been developed using RNNs to predict health problems like heart failure [8, 10, 11], diabetes [52–54], high blood pressure [33], and hospital readmission [73]. To name a few examples, Choi et al. [8] have used a **gated recurrent unit (GRU)** architecture for the multi-label prediction that predicts the diagnosis and medication codes of the future visits. Choi et al. [11] have also used RNNs with GRU cells for heart failure prediction, and Pham et al. [52] have used RNNs with LSTM for predicting readmission for diabetic and mental health patients. Maragatham and Devi [39] used a LSTM model for heart failure prediction. Xu et al. [75] proposed a multi-task RNN for risk prediction of multiple **major adverse cardiovascular events (MACEs)** on EHR data. Mei et al. [41] used raw EHR data to construct a model with RNN for sequential data modeling. 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 the existing work relies on traditional machine learning methods. Example studies include using logistic regression or linear regressions [19, 56, 65, 72] and the random forest [23] where they used selected perinatal factors like BMI, birth weight,

maternal weight, and sociodemographic characteristics to predict obesity at certain age points in childhood and adolescence. A couple of recent studies have reviewed the machine learning models to predict childhood and adolescent obesity and compared these studies by also including an earlier version of our current work [12, 20–22, 55].

One of the major drawbacks of deep learning models is the lack of interpretability. The lack of interpretability reduces the value of predictive models, especially in the medical domain. If medical practitioners cannot understand how an outcome is predicted by a model, relying on the model's outcomes will not be practical [50]. Many attempts have been recently made to make sense of the outcome of these models. For instance, Bahdanau et al. [3] proposed the attention mechanism, which is a method originating from machine translation applications. This attention mechanism can improve interpretability at the time level, as it gives attention scores to the timestamps. However, for multivariate time series, we also need to consider feature importance at each timestamp. Zhang et al. [77] used a hierarchical attention mechanism by using a convolutional operation. Choi et al. [7] developed an interpretable model with two levels of attention weights learned from two reverse-time GRU models. Jin et al. [27] used two separate RNN networks to compute attention weights for analyzing EHR data. In our work, we follow the use of attention mechanisms to improve the interpretability of the RNN-based models for multivariate time-series data to get the importance score for the timestamps and for each feature in the timestamps.

3 DATA

3.1 Data Description

The EHR data used in this work was extracted from Nemours Children's Health, which is a large network of pediatric health in the United States, primarily spanning the states of Delaware, Florida, New Jersey, and Pennsylvania. The dataset is a portion of the larger PEDSnet dataset, containing EHR data from several major US Pediatric Health Systems [66]. Our data is extracted from over two million distinct patients from the Nemours EHR system with patient records dating from 2002 to 2019. Inclusion criteria for the patients in the cohort used in this study included having (1) at least 5 years of medical history, (2) no evidence of type 1 diabetes, and (3) no evidence of cancer, sickle cell disease, developmental delay, or other complex medical conditions. In addition to including patients with obesity in our cohort, an equal number of normal-weight patients were also selected by random sampling from the normal-weight population. The analysis dataset was further screened for inconsistencies, particularly those related to birth dates and measurement dates, dropping records with missing or implausible dates. The dataset was anonymized. All of the dates were skewed randomly per patient by ± 180 days. The data access and processing steps were approved by the Nemours institutional review board. Each record in our dataset relates to one visit and captures the visit start and end time and all the condition, procedure, medication, and measurement variables recorded for that visit. It also contains demographic data for each patient. The medical codes are standardized terminologies of SNOMED-CT, RxNorm, CPT, and LOINC [67] for both clinical and demographic facts.

3.2 Data Representation and Preprocessing

The EHR data extracted for this study consisted of 44 million records from 68,029 patients and 34,96,559 distinct visits. Medical codes in the data consisted of 20,300 conditions, 10,167 procedures, 6,163 medications, and 7,693 measurements (lab results) variables. On average, 22 conditions, 34 procedures, 15 medications, and 49 measurement variables were recorded per patient in the data. Since the average number of medical variables recorded per patient was far lower than the total number of unique variables in the data, the dataset had a very sparse representation (especially once we represent it using one-hot encoding for the training on machine learning models). We experimented with **principal component analysis (PCA)** and skip-gram methods for feature selection and dimensionality reduction [10]. Neither of these methods led to satisfactory results due to the very large size of our feature space. Finally, to reduce dimensionality, we followed a similar approach used in some other studies

Table 1. Characteristics of the Cohort Used in This Study

Name	Value
Total number of patients	68,003
Total number of visits	33,34,047
Avg. number of visits per patient	51
Number of females	31,003 (45%)
Number of males	37,000 (54%)
Age	0–20, Mean = 5
Race and Ethnicity	
White or Caucasian	33,244
Black or African American	25,329
Non-Hispanic	58,894
Other	17,834

[23]; we removed rare variables that did not have enough information and were absent in 98% of the population in the data. This resulted in 1,850 medical variables with 635 conditions, 407 procedures, 380 medications, and 428 measurement variables. Additionally, few measurement variables had inconsistent or incorrect values. More than 50% of entries in such variables contained text like “sample not collected properly” and “sample destroyed.” We removed 113 such measurement variables that had missing and corrupt values. For any visit, if the same medical variable was recorded more than once, we took the latest value. There were several correlated features in the dataset like weight, height, BMI, obesity, and BMI percentile. These variables had a different rate of missingness in the dataset. BMI, height, and weight were missing in 25%, 25%, and 7% of the visits in the final cohort, respectively. Obesity was a condition variable that was only recorded if a patient was screened for obesity in that visit. Considering the missing rates, and to ensure our dataset can capture as much obesity-related information, we did not do any further feature selection for these correlated variables. We also removed 26 patients from the data who did not have any BMI recorded throughout their medical history. The final dataset consisted of 68,003 patients, 33,34,047 visits, and 1,737 medical variables. Some additional details about the data are listed in Table 1.

We preprocessed the data to represent all the condition and procedure variables as binary variables (1 if present and 0 if not recorded for the visit). Few medication variables were present as continuous variables in original data, where the values contain information about the amount of medication prescribed to a patient during a visit. However, many medication variables were present as binary variables. Therefore, after the preprocessing, we represented all the medication variables as binary variables. Measurement variables in the cohort were present as continuous variables. These continuous variables were normalized to the 0–1 range during preprocessing for model training. We have used zero imputation for missing entries in the EHR data. In medical data, where missingness carries meaning, imputation needs to include sufficient information about the reason for missingness to avoid bias [36]. However, most imputation techniques assume data to be missing at random [36]. Zero imputation sets the missing values to 0, such that it did not contribute to the patient’s risk. It also has the potential to improve model performance in settings where missingness carries meaning [63]. Moreover, deep learning models like LSTMs have shown that they learn from missingness patterns represented by zero imputation [35]. With zero imputation it is also possible for the new patients to be assessed without the need for an end-user to impute the values.

As the EHR data consists of patient records as a sequence of visits with each visit containing various medical codes, we use an incremental representation, where we first represent medical codes using code-level representation. Then, we use a code-level representation of all medical codes for each visit record and represent visits using visit-level representation. Finally, we use visit-level representation for each visit and represent the pa-



Fig. 1. Visit-level representation of the EHR data. Each visit V_t is represented as a vector of condition, medication, procedure, and measurement variables. The set of all visits for a patient forms a patient's representation.

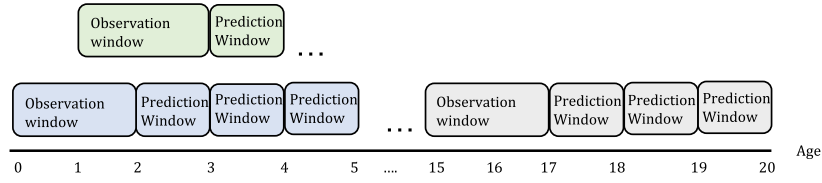


Fig. 2. Observation and prediction window design for each sub-cohort. For each observation window, there are three different prediction windows shown using the same color.

tients using a patient-level representation. In code-level representation, medical codes consist of all the unique condition, medication, 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 $|C|$, medication 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 $|M|$. For visit-level representation, we denote a visit at time t as V_t , which is the concatenation of the condition, medication, procedure, and measurement code vectors. The size of V_t is $|V| = |C| + |D| + |P| + |M|$. We represent condition, medication, 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 the visit V_t . All the measurement variables are represented by the corresponding continuous values $M_t \in \mathbb{R}^{|M|}$ for the visit V_t . Figure 1 shows the visit-level representation of our EHR data. EHR data for each patient is the sequence of visit-level vectors for that patient. Finally, patient-level representation is the sequence of the visit vectors for a patient. We denote the patients with $S: \{s_1, s_2, \dots, s_{|S|}\}$, where the i th patient s_i with T visits is represented with the matrix $s_i \in \mathbb{R}^{|T| \times |V|}$.

In addition to the medical data, which changes with each visit, the EHR data also consists of static demographic data, which does not change with every visit. We represented the demographic variables, i.e., sex, race, ethnicity, insurance type, and zip code (indicating the approximate location of the patient) as categorical variables. Table 1 shows the distribution of some of the demographic variables in our data.

As our cohort is extracted such that each patient has at least 5 years of data, sub-cohorts for every 5-year age range were created for the predictive models. Picking a longer range (than 5 years) would have resulted in a small number of patients as there were fewer patients who had records of more than 5 years (due to various reasons like having no visit or a provider change). We divided the complete cohort into 5-year ranges starting from each age from 0 to 15 years. This procedure resulted in 16 age ranges for the models (models for the ages of 0 to 5, 1 to 6, ..., and 15 to 20). For every 5-year data, we then used a fixed observation window of the initial 2 years and predicted obesity for 1, 2, and 3 years in the future. This way, we ended up with 48 sub-cohorts by creating 3 sub-cohorts for each of the 16 five-year time periods. For each of the 48 sub-cohorts, we used only those patients with at least one visit in the observation and one visit in the prediction window. Figure 2 depicts how we created the sub-cohorts, as well as the observation and the prediction windows for each sub-cohort. The model is trained separately on data in the observation window of each sub-cohort to predict the future BMI values in the respective prediction window.

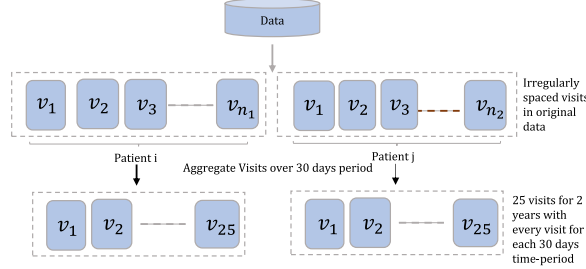


Fig. 3. Time sequences for the proposed LSTM model. Irregularly spaced visits for each patient (such as patient i and j) in a 2-year period are mapped to 25 equally distributed 30-day intervals.

4 METHOD

After obtaining all the sub-cohorts, we transformed the data so that it can be given as input to the LSTM model that we use as our predictive model. In general, visits have irregular time intervals and each patient has a different number of visits. To transform these irregularly spaced and unequal number of 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 of the 2-year observation window, resulting in $(2 * 365)/30 \approx 25$ equally spaced sequences for each patient. Figure 3 shows how new sequences are obtained from unequal and irregularly spaced input time sequences. Any condition, medication, and procedure variable observed at least once over the 30-day time period are denoted by “1” in these new sequences. Continuous variables are averaged over the 30-day time period. If there were no visits for a patient in any of the 30-day time periods, the corresponding vector for that period was set to contain all zeros. The zero vectors acted as padding to maintain an equal sequence length for the patients. In addition to the conditions, procedures, medications, and measurements, the time interval between each visit sequence (Δt , capturing the time intervals between the non-empty sequences) was also added to the end of each visit vector. Adding time interval values has been shown to enrich the time-series input in similar studies [11].

While deep learning models (such as LSTMs) show superior performances compared to traditional machine learning models, they are difficult to interpret due to their so-called “black box” nature [29]. Such property may reduce the practicality of deploying these methods in medical domains. To mitigate such concerns, we enhanced the LSTM model by adding extra layers to improve its interpretability. Because of the mixed nature of our datasets, we have considered two levels of interpretability: time-level and feature-level interpretability. The time-level interpretability refers to ranking visits, and the feature-level interpretability refers to ranking the features present in the visits according to their importance in predicting the final output. Figure 4 shows the complete architecture of the LSTM model with enhanced interpretability. Moreover, Figure S1 in the supplementary material shows the basic LSTM model architecture without components for interpretability.

In our architecture, $V = \{v_1, v_2, \dots, v_t\}$ is a multivariate time-series input consisting of the visits for a patient observed over timesteps $T = (1, 2, \dots, t)$. A visit v_i is a d -dimensional vector consisting of all medical features for the visit v_i . The first layer in our architecture is an embedding layer, which is used to reduce the input dimension space from the d -dimensional to the n -dimensional space ($n \ll d$). The output of the embedding layer is a multivariate time-series $X = \{x_1, x_2, \dots, x_t\} \in \mathbb{R}^{t \times n}$. We used $d = 1737$, and $n = 700$. This embedding layer is also used to add interpretability at the feature level, which we discuss in detail later. After the embedding layer, we used LSTM cells in two recurrent layers for training our model over the sequential visit-level (dynamic) data. The input to the LSTM layers is the multivariate time-series X . Each LSTM layer has a hidden layer dimension of size 700, which results in 3,922,800 trainable parameters. The output of the LSTM layers is another multivariate time series in the n -dimensional space $H = \{h_1, h_2, \dots, h_t\} \in \mathbb{R}^{t \times n}$. The softmax layer after the LSTM layers

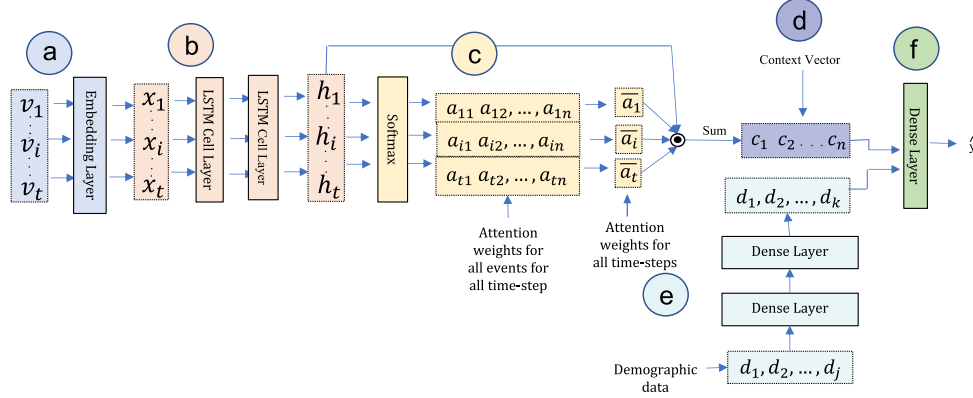


Fig. 4. Overview of proposed LSTM architecture including the elements to enhance its interpretability. (a) Multivariate time-series input vector V passes through the embedding layer. (b) The embedding layer's output X passes through two LSTM layers. (c) A softmax layer generates attention weights. (d) A context vector is the weighted sum of the hidden state vectors h_t and a_t . (e) A separate feed-forward network for the demographic data. (f) A dense (fully connected) layer to obtain the final output.

generates attention scores $\{a_{11}, a_{12}, \dots, a_{1n}; \dots; a_{t1}, a_{t2}, \dots, a_{tn}\} \in \mathbb{R}^{t \times n}$, where $a_{ij} \in (0, 1]$. The average of all values in each vector $\{a_{i1}, a_{i2}, \dots, a_{in}\}$ is then calculated to derive \bar{a}_i , where $i = \{1, 2, \dots, t\}$. The \bar{a}_i scores obtained from the softmax layer are used to add time-level interpretability, which we discuss later. This softmax layer has 25 neurons, each related to one of the 25 timestamps, and adds additional 650 trainable parameters to our model. The weighted sum of the hidden states (h_i) and attention scores (\bar{a}_i) is used to obtain the final context vector C . There is also a separate feed-forward network for the demographic (static) data. The demographic input is a j -dimensional vector $\{d_1, d_2, \dots, d_j\} \in \mathbb{R}^j$ that is passed through two dense layers with 10 and 15 neurons and adds 50 and 165 trainable parameters, respectively. The output of the LSTM layer is then concatenated with the output of this two-layer network. This concatenated output is then passed through one final dense layer consisting of 100 neurons with 71,600 trainable parameters for predicting the BMI values.

Time-level interpretability. As mentioned previously, we added a softmax layer on top of the LSTM layers to compute the attention score for each timestamp. In general, each LSTM unit generates a hidden vector output $H = \{h_1, h_2, \dots, h_i, \dots, h_t\} \in \mathbb{R}^{t \times n}$, where $i = \{1, 2, \dots, t\}$. The hidden state h_i is computed by applying the non-linear transformation on the input x_i at the time i and the hidden state of previous time step h_{i-1} :

$$h_i \leftarrow \text{LSTM}(x_i, h_{i-1}), \quad (1)$$

where h_i is an n -dimensional vector, which is the same as the dimension of the hidden layer of the LSTM network. We then calculated the attention score for each hidden state by using a softmax layer:

$$a_{i1}, a_{i2}, \dots, a_{ij}, a_{ij+1}, \dots, a_{in} = \text{softmax}(h_i), \quad (2)$$

where h_i is an n -dimensional vector. $\{a_{i1}, a_{i2}, \dots, a_{ij}, a_{ij+1}, \dots, a_{in}\}$ is also an n -dimensional vector, where $a_{ij} \in (0, 1]$. To obtain a scalar attention score value for each hidden state, we took the average of all the values in the vector $\{a_{i1}, a_{i2}, \dots, a_{ij}, a_{ij+1}, \dots, a_{in}\}$:

$$\bar{a}_i = \left(\sum_{j=1}^n a_{ij} \right) \div n, \quad (3)$$

The value of \bar{a}_i is calculated for each hidden state h_i . These scores show importance scores for each h_i and are used to visualize the visits that are given most importance by the LSTM layer. These scores are then used to compute the weighted sum of the hidden states:

$$C = \sum_{i=1}^T \bar{a}_i * h_i, \quad (4)$$

where \bar{a}_i is a scalar value as shown in Equation (3) and h_i is an n -dimensional vector. C is also an n -dimensional vector $\{c_1, c_2, \dots, c_j, c_{j+1}, \dots, c_n\}$. The vector C contains information of multivariate input time series weighted by the importance score of each timestep.

Feature-level interpretability. As mentioned earlier, we added an embedding layer after the input layer to add feature-level interpretability. We employ the simple method proposed in [9] to interpret medical code representations and enhance them using attention scores. We used the weights from the embedding layer to compute the importance score for the features in each timestep. The weight matrix of the embedding layer is of the form $W_i = w_1, w_2, \dots, w_d \in \mathbb{R}^{n \times d}$, where d is the dimension of the input time series v_i , and n is the reduced dimension of x_i at timestep i . To obtain the feature ranking for each input feature in v_i at the timestep i , we took the product of embedding of v_i (weight matrix W_i from the embedding layer) and the attention scores of v_i obtained from the softmax layer, $b_i = \{a_{i1}, a_{i2}, \dots, a_{ij}, \dots, a_{in}\} \in (0, 1]^n$. Equation (5) shows the importance score calculation for features at timestep i :

$$s_i = W_i^T b_i, \quad (5)$$

where s_i is a d -dimensional vector containing the importance score for all input features.

We have additionally used transfer learning to enhance the model performance by learning from a larger dataset. While dividing our input data into the 48 sub-cohorts could improve its performance on learning specific age range patterns, this also meant reducing the input size for each of the models. Table 2 shows the number of patients with and without obesity in each sub-cohort. The issue of having smaller sample sizes was especially more visible as the number of samples reduced gradually with the increasing age ranges. Reduction in the number of samples in pediatric datasets is common due to the higher rate of visits in earlier years of children's life. To improve the performance of the model, we used the complete dataset for all age ranges and created three initial models (instead of 48) for predicting obesity at 1 year, 2 years, and 3 years in the future, respectively. After this, each of the 3 general models has been used as the basis for the 16 separate predictive models related to a similar prediction window.

5 EXPERIMENTS

For training the models, we split the data into 60:20:20 as training, validation, and test data. Data split is performed such that the proportion of samples having and not having obesity is the same in the training and test data as in the original data. We used an Adadelta optimizer [76] with an initial learning rate of 0.05. We experimented with other learning rates (0.001, 0.01, 0.05, 0.1). We also tried different optimizer settings (RMSprop and Adam optimizer) the learning rates mentioned before. We found best train and validation error convergence with Adadelta and a learning rate of 0.05. Both L1 and L2 regularizations were used on the first LSTM layer. Models were trained on a Tesla V100-SXM2 32 GB GPU. We trained the models separately on different sub-cohorts using different observation and prediction age windows as explained in Section 3.2. All models were trained on data in the observation window (first column in Table 2) to predict the future BMI value in the respective prediction window (second column in Table 2). The predicted BMI values were then used to classify each sample into obesity and without-obesity classes.

To evaluate the performance of our proposed model, we created two baseline models that follow the traditional methods that are commonly used in the literature. Such methods aggregate the patterns on the dataset while ignoring the temporality of the EHR data. We used linear regression and random forest as the baseline models

Table 2. The Observation and Prediction Window Setup for the 48 Models, as Well as the Number of Patients with and without Obesity, Corresponding to Each of the Three Prediction Windows in the Second Column

Observation Window (Age Year)	Prediction Window (Age Year)	# of Samples with Obesity	# of Samples w/o Obesity
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

for comparison. To do this, we aggregated data over all the visits in the observation window for each patient (corresponding to any of the input sub-cohorts). Target labels were the labels at the prediction age. Aggregation for binary medical codes for the conditions, medications, and procedures is performed such that each medical code represents the frequency of its occurrence over the 2 years, and for the continuous variables, we took the average over the 2 years. For the BMI and body weight, we took the maximum and latest recorded values in the observation window.

We compared the performance of the baseline models and our proposed model. For the baseline models (linear regression and random forest), we used 10-fold cross-validation and report mean results over the complete data. For the LSTM models, we did not use cross-validations (due to heavy computing cost) and only report results on the test data. Figure 5 shows the accuracy and **area under the receiver operating characteristic curve (AUC)** for all models separately based on the prediction window size. It compares the performance of (1) the proposed LSTM model with the interpretable elements trained using transfer learning, (2) random forest regressor, and (3) linear regression over different prediction windows separately. Refer to Table S1 in the supplemental materials for an extended set of results of the proposed model. The additional results discussed in supplemental materials also include comparisons to another variation of the proposed model, which is the model without transfer learning. Figure 6 demonstrates the effect of the size of the prediction window in predicting obesity at different prediction ages. It compares the sensitivity, **positive predictive value (PPV)**, accuracy, and AUC of the proposed model.

Additionally, we have computed the feature importance at both the individual and population levels. In Figure 7, we ranked the features for the three most important visits for a sample patient to demonstrate the ability of our model to explain its predictions for any patient. To do this, we picked the top three visits with the highest attention weights and then ranked the features for those visits by calculating their importance score using Equation (5). Refer to Figure S3 in the supplemental materials for the visualization of the attention weights given to input timestamps. We also ranked feature importance at the population level by averaging feature importance for the top three visits for all the individual samples. Table 3 shows the top 20 most important features at the population level.

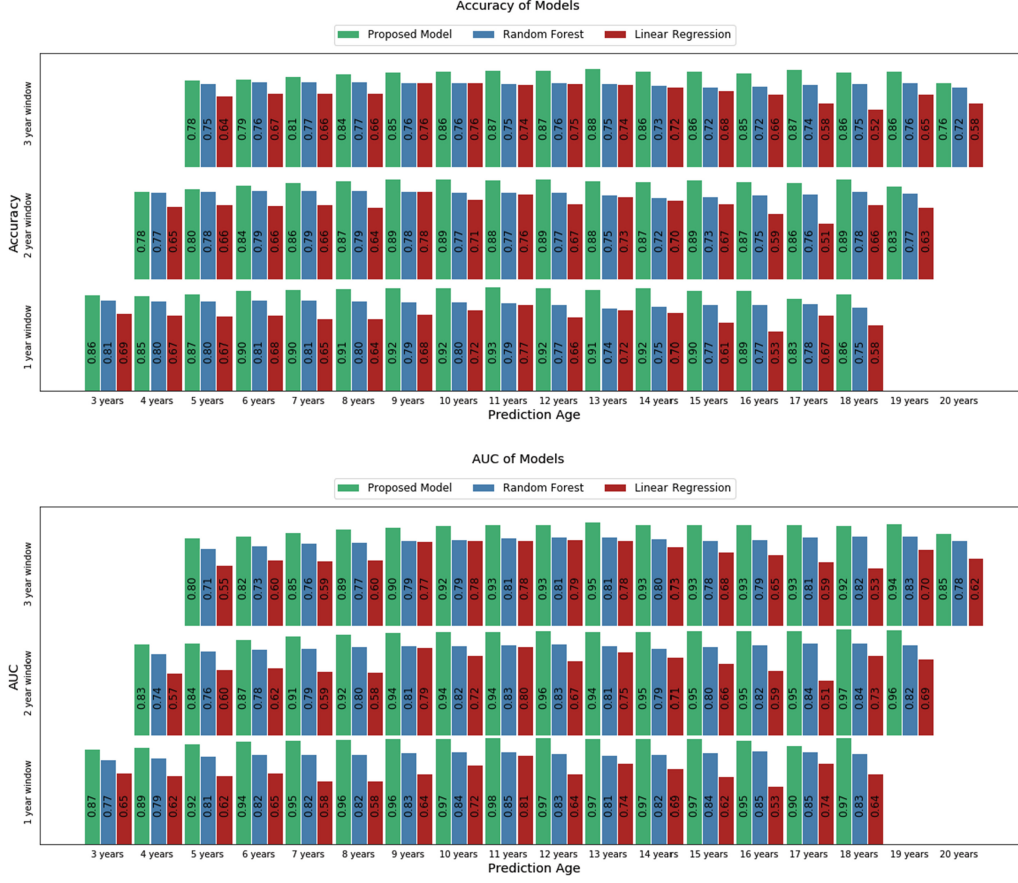


Fig. 5. Comparing accuracy (top panel) and AUC (bottom panel) results of (1) the proposed model, (2) random forest regressor, and (3) linear regression. Each of the three rows in the top and bottom panels refers to one of the three prediction windows at 1, 2, and 3 years in the future. See Figure 2 for the configuration used for defining the observation and prediction windows.

Finally, we compared the performance of our model against other existing childhood obesity prediction studies (Table 4). We compared the performance at the same prediction age with a 3-year window in our work as reported in existing works. If the existing work reported performance for a group of ages, then we took the average of performance at all the ages in the group at a 3-year window in our work.

6 DISCUSSION

In this study, we have developed a new model for predicting childhood obesity patterns using EHR data. We employ an LSTM network and a separate feed-forward network to model dynamic (time-series) and static data in the EHR data. To transform the irregularly spaced and unequal number of clinical visits, we combined the visits over a 30-day time window and obtained a regularly spaced and an equal number of clinical visits for each patient. We experimented with different window sizes of 6 months, 3 months, 30 days, and 15 days, and found

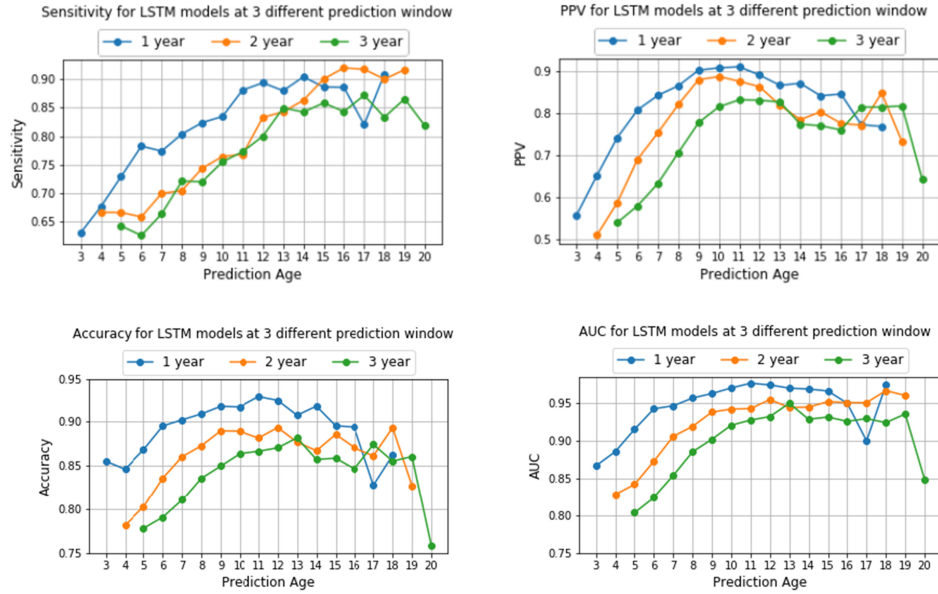


Fig. 6. Comparing prediction performances at different ages. Plots compare sensitivity, PPV, accuracy, and AUC of the proposed model across 1, 2, and 3 years in the future, shown by blue, orange, and green lines, respectively.

Table 3. Feature Rankings Were Obtained by Averaging the Importance Score of the Features from all the Samples in the Test Data

Feature Ranking	Feature Description	Feature Ranking	Feature Description
1	BMI [Percentile] per age and gender	11	Abnormal weight gain
2	Obese/non-obese label	12	Anomaly of chromosome pair 21
3	Allergic urticaria	13	Erythrocytes (#/volume) in body fluid
4	Childhood obesity	14	Obesity
5	Morbid obesity	15	Hyperactive behavior
6	Suspected clinical finding	16	Tachycardia
7	Achondroplasia	17	Requires respiratory syncytial virus
8	MCH (entitic mass) by automated count	18	Vaccination
9	Cholesterol in LDL/cholesterol in HDL	19	CO2 1712
10	Hearing loss	20	pH of blood

out that a 30-day window size shows the best performance in capturing the variations in clinical trajectories of patients for predicting obesity. Notably, we achieved an AUC of 0.80, 0.93, and 0.92 using a 3-year input window for predicting obesity at 5 years (preschool), 11 years (prepubertal), and 18 years (post-pubertal).

As shown in Figure 5, the performance of the proposed model was noticeably better than the two baseline models, i.e., linear regression and random forest. Our results show that the temporal patterns captured in EHR data encapsulate important information for future outcome prediction and should not be ignored. The existing body of research also shows that data temporality is important to capture weight gain trajectories and other

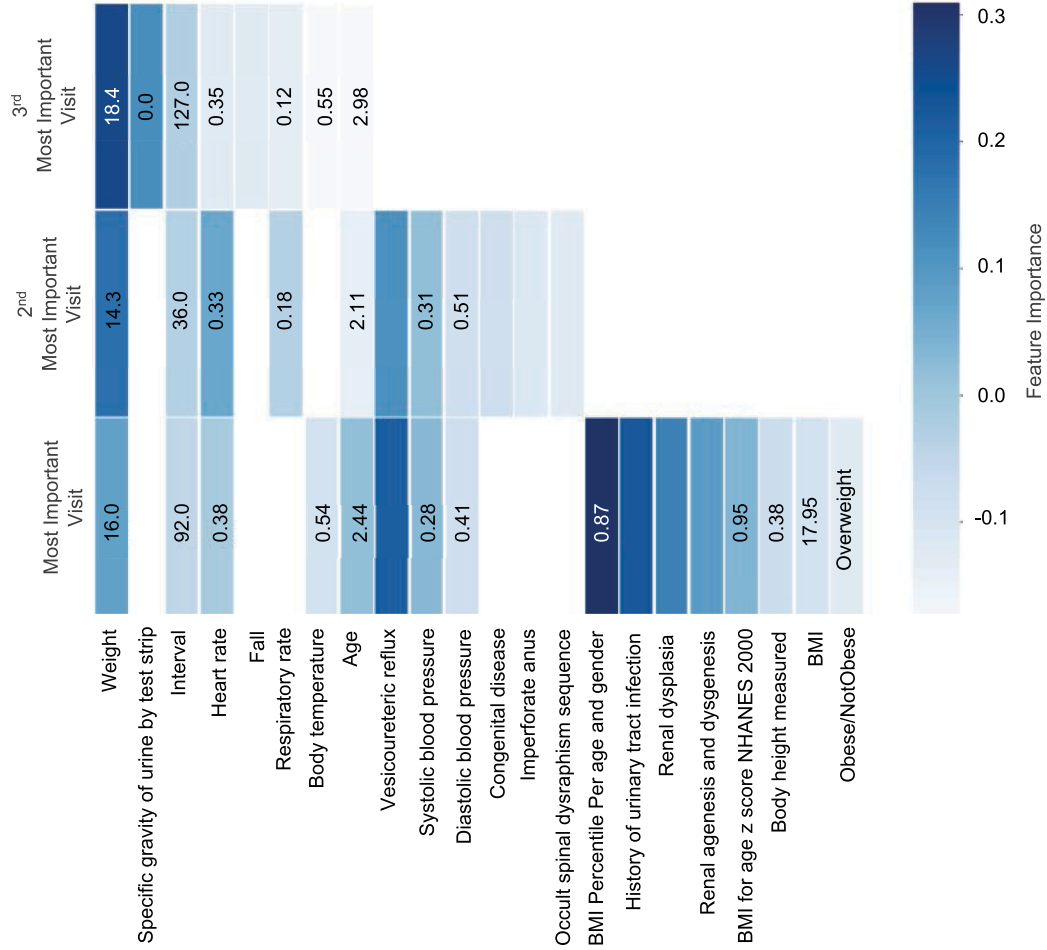


Fig. 7. Ranking of the features for the three most important visits (timestamps) for one patient (sample). The gradient bar on the right shows the importance score for each feature (darker is more important). For the cells related to the measurements, each cell also contains the actual measurement's numerical value (such as weight and height). Non-measurement cells do not have a value.

medical histories over time [43]. Therefore, the RNN architecture used in our model improves the prediction performance by taking into consideration the temporality of the data, a property that traditional methods do not have. We also used static features with a feed-forward network, which contributed to the overall performance of our model. Moreover, transfer learning helps to further improve the performance of the models for the sub-cohorts with a low number of samples by learning from the samples in other sub-cohorts. For instance, prediction at age 20 using a 3-year window achieves accuracy and AUC of 0.45 and 0.55, respectively, without using the transfer learning technique. However, when transfer learning is used, we achieved accuracy and AUC of 0.76 and 0.85, respectively. Refer to Figure S2 for a comparison of the results obtained from training the proposed model with and without transfer learning. Here, transfer learning helps to improve the performance of the models on each sub-cohort by learning from the samples of other sub-cohorts.

Table 4. Comparison of Our Work with Existing Work

Study Group	Prediction Age	Method Used	Their Results	Our Results
Morandi et al. [44]	7 and 16 years	Logistic Reg.	0.78, 0.75 (AUC)	0.85, 0.92 (AUC)
Druet et al. [14]	7–14 years	Logistic Reg.	0.77 (AUC)	0.83 (AUC)
Weng et al. [72]	3 years	Logistic Reg.	0.75 (AUC)	0.87 (AUC)
Redsell et al. [56]	5 years	Logistic Reg.	0.67 (AUC)	0.80 (AUC)
Manios et al. [37]	9–13 years	Logistic Reg.	0.64 (AUC)	0.92 (AUC)
Robson et al. [58]	5 years	Logistic Reg.	0.78 (AUC)	0.80 (AUC)
Steur et al. [65]	8 years	Logistic Reg.	0.75 (AUC)	0.88 (AUC)
Hammond et al. [23]	5 years	Lasso Reg.	0.81 and 0.76 (AUC) for girls and boys	0.80 (AUC) gender combined
Lingren et al. [34]	1–6 years	Naïve Bayes	0.90 (precision) and 0.44 (sensitivity)	0.54 (precision) and 0.63 (sensitivity)
Zheng and Ruggiero [78]	14–18 years	Neural network	0.84 (accuracy)	0.85 (accuracy)

Reg.: Regression.

Our results in Figure 6 show that the closer the observation window to the prediction time, the better the performance of the model. This means that the prediction results obtained using a 1-year prediction window are better than the prediction results obtained on a 2-year window, which are in turn better than the prediction results obtained on a 3-year prediction window. This happens because the closer the observation window is to the prediction time, the shorter the prediction window, and therefore, the model has to predict for the nearer future (which is easier as compared to predicting a more distant future). For both accuracy and AUC, all of the plots show a semi-bell-shaped curve. In the beginning, the performance of the model increases, and then it starts to decrease after a certain age. When we compare our results to the number of samples in Table 2, we observe that with decreasing number of samples in our cohort (in older ages), the performance of the deep learning model increases (up to a certain point), possibly because of less noise and being a naturally simpler task (i.e., children maintaining a similar status). However, a very low number of samples in a cohort also does not help in learning a deep learning model effectively (as in the case of the older ages). As an example case, there is a sharp decrease in performance for prediction at the age of 20. For the prediction at the age of 20, our model uses an observation window of 15 to 17 years of age, with a very small number of samples. This small number of samples in the observation window of 15 to 17 years has more impact on the 3-year prediction window as compared to the 1-year and 2-year prediction window.

Besides superior predictive performance, we have also included interpretability capabilities in our proposed model. In Figure 7, we ranked the features for the three most important visits for a random sample to demonstrate the ability of our model to explain the prediction for any sample. In the case of one sample patient shown in Figure 7, we can see that his/her previous weight and BMI are the most important features, which are naturally expected for predicting obesity. Among other highly ranked features, we can see that vesicoureteral reflux is given a very high importance score. This condition is a type of kidney disease that is highly correlated to obesity in children [74]. We also calculated the feature ranking over the complete dataset (with the samples that are predicted as those with obesity) by averaging feature importance for the top three visits to obtain the population-level feature ranking. As shown in Table 3, the most important population-level features in predicting childhood obesity as determined by our model were BMI, previous obesity levels (recorded as obese/non-obese), childhood obesity, morbid obesity, and obesity, matching what was reported in similar studies [46, 64, 74]. Other predictors of obesity with a lower prevalence that seemed to fit the existing work were achondroplasia [24, 60] and the anomaly of chromosome 21 [17]. Hyperactive behavior has a known association with obesity, though the direction of causality is undetermined. Obesity is known to be among the main risk factors of high LDL (bad

cholesterol and low HDL (good) cholesterol [70]. **Mean Corpuscular Hemoglobin (MCH)** and Erythrocytes are also shown to be related to chronic inflammation of obesity [71], and CO₂ and pH levels of blood may be related to obesity hypoventilation syndrome [40, 42]. Another variable associated with obesity that is listed in Table 3 is hearing loss, which is positively correlated with obesity [31]. Our results collectively show that feature ranking obtained from the proposed LSTM model can generate results that coincide with existing studies in the field of obesity research [59].

Similar to other predictive models, while our models cannot show any causation relationship, identifying the top features that the model has been using for its predictions may improve the explainability of our model. To further analyze the potential of the feature-level interpretability of our proposed model in understanding the roots of childhood obesity, we separately ranked the features for the patients who moved from without-obesity to with-obesity status, and the patients with obesity in the observation window who also had obesity within the prediction window. Table S2 and Table S3 in the supplemental materials show the list of top 20 features for these sub-groups. By comparing these top-ranked features for the two sub-groups, we found that factors like prematurity (retinopathy of prematurity), acute tonsillitis, and chronic adenoiditis were among the strong predictors in those who moved from without-obesity to with-obesity status. Premature infants generally start with lower weight than their peers, but many gain weight rapidly [47]. Acute tonsillitis and chronic adenoiditis are strong predictors of obesity because of rapid weight gain after having a tonsillectomy and adenoidectomy (which are the procedures to remove tonsils and adenoids) [2]. Factors like severe obesity are seen in patients who remain obese in both the observation and prediction windows. This is consistent with existing findings showing that those who develop obesity early on tend to have a more persistent obesity status than their counterparts who started with a normal weight [28]. A higher ranking of the LDL cholesterol feature in patients who remain obese within the observation and prediction window shows that those who develop obesity early also develop comorbidities of their obesity.

Moreover, as shown in Table 4, our proposed deep learning model shows a better performance compared to traditional machine learning models. Unlike existing work that focuses on a single age point to predict obesity, our model can predict obesity (BMI) patterns across a wide age range, starting from early childhood to early adulthood. Ziauddeen et. al. [79] reported a list of existing studies on predicting childhood obesity, and all of these studies had used a logistic regression method on aggregated data. The data in these studies consisted of selected few variables such as maternal data and physical measurements of children. However, by using an LSTM-based architecture, our model was able to use temporal data and achieve a higher AUC compared to similar prediction age points. Other studies have also reviewed existing work on obesity prediction that uses logistic and other machine learning models [12, 55]. We list a set of comparable studies in Table 4. Some other studies related to obesity prediction have presented linear equations to estimate fat-free mass to access body fatness in children [1, 26]. These studies are not directly comparable to our work, as they only estimate the fat-free mass and not the total body weight. Additionally, most of such studies report single R^2 or RMSE scores over a wide range of ages such as 4 to 15 or 3 to 29, instead of at different ages (as done in our work). Another limitation of the linear models for obesity prediction is that they work on a selected few predictor variables. Deep learning models, as the one proposed in our work, use a large number of predictor variables and can thus be extended to study the effect of other medical conditions in predicting obesity similar to the results shown in Table 3.

7 LIMITATIONS AND FUTURE WORK

The current work is limited in several ways. Our model can predict a maximum of 3 years in the future. Long-term predictions require access to long-term medical histories. Since our data was collected in such a way that patients should have at least 5 years of data, we were only able to use a total 5-year window for performing supervised prediction tasks. In the future, we plan to explore other semi-supervised techniques like autoencoders to extend our prediction window with the complete data without dividing that into smaller datasets. The features in our EHR data contained only non-grouped features. Therefore, there was a large number of features and most of

them were not recorded for a major portion of the population in the data. In the future, we aim to explore the opportunities to group medical variables and reduce the number of features in a more standardized manner. This may also help in obtaining better results from the feature-level interpretability process introduced here. Reducing the number of input features will also help to reduce the number of model parameters and will reduce its computational complexity. Currently, the results of feature-level interpretability consist of correlated features like existing obesity level, BMI, and weight. Our work can be extended by employing other feature selection techniques that can focus more on non-correlated features and possibly find high-risk features that are difficult for pediatricians to detect for predicting obesity. Other ways for extending our models include adding another attention layer after the initial dense layers for processing static (demographic) data and expanding our transfer learning process by using data from other healthcare systems. Finally, we note that while in this article we have used our proposed model for childhood obesity prediction, it can also be applied to other healthcare conditions for future outcome prediction tasks using EHR data. Our model can be used in general for any outcome prediction using temporal and static features. Moreover, the layers added in the model for interpretability can help identify which timestamps and features are important for outcome prediction. This can help identify major risk factors for different health conditions and can help design interventions targeting those risk factors.

8 CONCLUSION

In this study, we presented a new deep neural network architecture based on the common LSTM architecture that combines both static and dynamic clinical features for predicting childhood obesity in the next 1, 2, and 3 years. Specifically, we have used a network with LSTM cells combined with a separate feed-forward network for training our model. We trained the models using a large pediatric EHR dataset. An additional transfer learning process was used to improve the performance of the models by using sub-cohorts of the complete dataset. We achieved an AUC of 0.80, 0.93, and 0.92 for a 3-year window at 5 years (preschool), 11 years (prepubertal), and 18 years (post-pubertal), respectively. We also compared the recurrent model with other machine learning models for the task of predicting childhood obesity and found that our LSTM-based model demonstrates better performance compared to traditional machine learning models that ignore the temporality of the data by aggregating the data. We also added interpretable elements to our proposed model by ranking features in the top three most important visits during the 2 years of the observation (training) window for each patient. We have also calculated feature rankings for all samples in the data with obesity status. Such a list of ranked features can provide further insights into the functionality of the models and demonstrate the features that contribute to childhood obesity most. Our models have the potential to be incorporated into existing EHR-based pediatric systems as a clinical decision support tool, helping both providers and families identify a child's risk for obesity earlier and make informed decisions about the need for earlier adoption of healthier lifestyles to prevent or treat childhood obesity.

ACKNOWLEDGMENTS

Our study was supported by NIH awards, P20GM103446 and P20GM113125, and RWJF award 76778.

REFERENCES

- [1] Hesham Saleh Al-Sallami, Ailsa Goulding, Andrea Grant, Rachael Taylor, Nicholas Holford, and Stephen Brent Duffull. 2015. Prediction of fat-free mass in children. *Clinical Pharmacokinetics* 54, 11 (2015), 1169–1178.
- [2] A. Amadeo, L. De Sanctis, J. P. Giordanella, P. J. Monteyrol, B. Fauroux, et al. 2016. Obesity and obstructive sleep apnea in children. *Archives de Pediatrie: Organe Officiel de la Societe Francaise de Pediatrie* 24 (2016), S34–S38.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [4] Giorgio Bedogni, A.B. Tsybakov, and S. Berlin. 2009. Clinical prediction models—a practical approach to development, validation and updating. *Development* 18, 500 (2009), 53–99.
- [5] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. 2017. *Deep Learning*, vol. 1. MIT press Massachusetts, USA.

- [6] Sara N. Bleich, Kelsey A. Vercammen, Laura Y. Zatz, Johannah M. Frelief, Cara B. Ebbeling, and Anna Peeters. 2018. Interventions to prevent global childhood overweight and obesity: A systematic review. *Lancet Diabetes & Endocrinology* 6, 4 (2018), 332–346.
- [7] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, 3512–3520.
- [8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. PMLR, 301–318.
- [9] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1495–1504.
- [10] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686* (2016).
- [11] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2017. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* 24, 2 (2017), 361–370.
- [12] Gonzalo Colmenarejo. 2020. Machine learning models to predict childhood and adolescent obesity: A review. *Nutrients* 12, 8 (2020), 2466.
- [13] William H. Dietz. 1998. Health consequences of obesity in youth: Childhood predictors of adult disease. *Pediatrics* 101, Supplement 2 (1998), 518–525.
- [14] Céline Druet, Nicolas Stettler, Stephen Sharp, Rebecca K. Simmons, Cyrus Cooper, George Davey Smith, Ulf Ekelund, Claire Lévy-Marchal, Marjo-Ritta Jarvelin, Diana Kuh, et al. 2012. Prediction of childhood obesity by infancy weight gain: An individual-level meta-analysis. *Paediatric and Perinatal Epidemiology* 26, 1 (2012), 19–26.
- [15] Kathryn C. Eckstein, Laura M. Mikhail, Adolfo J. Ariza, J. Scott Thomson, Scott C. Millard, Helen J. Binns, et al. 2006. Parents' perceptions of their child's weight and health. *Pediatrics* 117, 3 (2006), 681–690.
- [16] Damodar Reddy Edla, Diwakar Tripathi, Ramalingaswamy Cheruku, and Venkatanaresbhabu Kuppli. 2018. An efficient multi-layer ensemble framework with BPSOGSA-based feature selection for credit scoring data analysis. *Arabian Journal for Science and Engineering* 43, 12 (2018), 6909–6928.
- [17] Tycho Foerste, Matthew Sabin, S. Reid, and D. Reddihough. 2016. Understanding the causes of obesity in children with trisomy 21: Hyperphagia vs physical inactivity. *Journal of Intellectual Disability Research* 60, 9 (2016), 856–864.
- [18] Centers for Disease Control and Prevention. 2021. Childhood obesity facts. <https://www.cdc.gov/obesity/data/childhood.html>.
- [19] Lise Graversen, Thorkild I.A. Sørensen, Thomas A. Gerds, Liselotte Petersen, Ulla Sovio, Marika Kaakinen, Anneli Sandbaek, Jaana Laitinen, Anja Taanila, Anneli Pouta, et al. 2015. Prediction of adolescent and adult adiposity outcomes from early life anthropometrics. *Obesity* 23, 1 (2015), 162–169.
- [20] Mehak Gupta and Rahmatollah Beheshti. 2020. Time-series Imputation and Prediction with Bi-Directional Generative Adversarial Networks. *arXiv:2009.08900* [cs.LG]
- [21] Mehak Gupta, Thao-Ly T. Phan, H. Timothy Bunnell, and Rahmatollah Beheshti. 2021. Concurrent imputation and prediction on EHR data using bi-directional GANs: Bi-GANs for EHR imputation and prediction. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB'21)*. Association for Computing Machinery, New York, NY, Article 7, 9 pages. <https://doi.org/10.1145/3459930.3469512>
- [22] Mehak Gupta, Thao-Ly T. Phan, Timothy Bunnell, and Rahmatollah Beheshti. 2019. Obesity prediction with EHR data: A deep learning approach with interpretable elements. *arXiv preprint arXiv:1912.02655* (2019).
- [23] Robert Hammond, Rodoniki Athanasiadou, Silvia Curado, Yindalon Aphinyanaphongs, Courtney Abrams, Mary Jo Messito, Rachel Gross, Michelle Katzow, Melanie Jay, Narges Razavian, et al. 2019. Predicting childhood obesity using electronic health records and publicly available data. *PloS One* 14, 4 (2019), e0215571.
- [24] Jacqueline T. Hecht, Opal Jean Hood, Robert J. Schwartz, Jill C. Hennessey, Barbara A. Bernhardt, William A. Horton, John M. Opitz, and James F. Reynolds. 1988. Obesity in achondroplasia. *American Journal of Medical Genetics* 31, 3 (1988), 597–602.
- [25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [26] Mohammed T. Hudda, Mary S. Fewtrell, Dalia Haroun, Sooky Lum, Jane E. Williams, Jonathan C.K. Wells, Richard D. Riley, Christopher G. Owen, Derek G. Cook, Alicja R. Rudnicka, et al. 2019. Development and validation of a prediction model for fat mass in children and adolescents: Meta-analysis using individual participant data. *BMJ* 366 (2019).
- [27] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. Carepre: An intelligent clinical decision assistance system. *ACM Transactions on Computing for Healthcare* 1, 1 Article 6 (mar 2020), 20 pages. <https://doi.org/10.1145/3344258>
- [28] Aaron S. Kelly, Sarah E. Barlow, Goutham Rao, Thomas H. Inge, Laura L. Hayman, Julia Steinberger, Elaine M. Urbina, Linda J. Ewing, and Stephen R. Daniels. 2013. Severe obesity in children and adolescents: Identification, associated health risks, and treatment approaches: A scientific statement from the American Heart Association. *Circulation* 128, 15 (2013), 1689–1712.
- [29] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5686–5697.

- [30] Venkatanareshbabu Kuppli, Diwakar Tripathi, and Damodar Reddy Edla. 2020. Credit score classification using spiking extreme learning machine. *Computational Intelligence* 36, 2 (2020), 402–426.
- [31] Anil K. Lalwani, Karin Katz, Ying-Hua Liu, Sarah Kim, and Michael Weitzman. 2013. Obesity is associated with sensorineural hearing loss in adolescents. *The Laryngoscope* 123, 12 (2013), 3178–3184.
- [32] R. S. Levine, D. L. Dahly, and M. C. J. Rudolf. 2012. Identifying infants at risk of becoming obese: Can we and should we? *Public Health* 126, 2 (2012), 123–128.
- [33] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. 2014. Deep learning for healthcare decision making with EMRs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'14)*. IEEE, 556–559.
- [34] Todd Lingren, Vidhu Thaker, Cassandra Brady, Bahram Namjou, Stephanie Kennebeck, Jonathan Bickel, Nandan Patibandla, Yizhao Ni, Sara L. Van Driest, Lixin Chen, et al. 2016. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Applied Clinical Informatics* 7, 3 (2016), 693–706.
- [35] Zachary C. Lipton, David Kale, and Randall Wetzel. 2016. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Machine Learning for Healthcare Conference*. PMLR, 253–270.
- [36] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.
- [37] Yannis Manios, Manolis Biribilis, George Moschonis, George Biribilis, Vassilis Mougios, Christos Lionis, George P. Chrousos, et al. 2013. Childhood obesity risk evaluation based on perinatal factors and family sociodemographic characteristics: CORE index. *European Journal of Pediatrics* 172, 4 (2013), 551–555.
- [38] Yannis Manios, Elpis Vlachopapadopoulou, George Moschonis, Feneli Karachaliou, Theodora Psaltopoulou, Dimitra Koutsouki, Gregory Bogdanis, Vilelmine Carayanni, Angelos Hatzakis, and Stefanos Michalacos. 2016. Utility and applicability of the “Childhood Obesity Risk Evaluation” (CORE)-index in predicting obesity in childhood and adolescence in Greece from early life: The “National Action Plan for Public Health.” *European Journal of Pediatrics* 175, 12 (2016), 1989–1996.
- [39] G. Maragatham and Shobana Devi. 2019. LSTM model for prediction of heart failure in big data. *Journal of Medical Systems* 43, 5 (2019), 1–13.
- [40] Juan F. Masa, Jean-Louis Pépin, Jean-Christian Borel, Babak Mokhlesi, Patrick B. Murphy, and Maria Ángeles Sánchez-Quiroga. 2019. Obesity hypoventilation syndrome. *European Respiratory Review* 28, 151 (2019).
- [41] Jing Mei, Shiwan Zhao, Feng Jin, Lingxiao Zhang, Haifeng Liu, Xiang Li, Guotong Xie, and M. Xu. 2017. Deep diabetologist: Learning to prescribe hypoglycemic medications with recurrent neural networks. *Studies in Health Technology and Informatics* 245 (2017), 1277–1277.
- [42] Babak Mokhlesi. 2010. Obesity hypoventilation syndrome: A state-of-the-art review. *Respiratory Care* 55, 10 (2010), 1347–1365.
- [43] Paulo Orlando Alves Monteiro and Cesar G. Victora. 2005. Rapid growth in infancy and childhood and obesity in later life—A systematic review. *Obesity Reviews* 6, 2 (2005), 143–154.
- [44] Anita Morandi, David Meyre, Stéphane Lobbens, Ken Kleinman, Marika Kaakinen, Sheryl L. Rifas-Shiman, Vincent Vatin, Stefan Gaget, Anneli Pouta, Anna-Liisa Hartikainen, et al. 2012. Estimation of newborn risk for child or adolescent obesity: Lessons from longitudinal birth cohorts. *PLoS One* 7, 11 (2012), e49919.
- [45] Centers for Disease Control National Center for Health Statistics and Prevention. 2001. Data Table of BMI-for-age Charts. https://www.cdc.gov/growthcharts/html/_charts/bmiagerev.htm.
- [46] Ken K. Ong and Ruth J. F. Loos. 2006. Rapid infancy weight gain and subsequent obesity: Systematic reviews and hopeful suggestions. *Acta Paediatrica* 95, 8 (2006), 904–908.
- [47] Mei-Chen Ou-Yang, Yao Sun, Melissa Liebowitz, Chih-Cheng Chen, Min-Lin Fang, Weiwei Dai, Tang-Wei Chuang, and Jyu-Lin Chen. 2020. Accelerated weight gain, prematurity, and the risk of childhood obesity: A meta-analysis and systematic review. *PLoS One* 15, 5 (2020), e0232238.
- [48] Sarah Harvey O'Brien, Richard Holubkov, and Evelyn Cohen Reis. 2004. Identification, evaluation, and management of obesity in an academic primary care center. *Pediatrics* 114, 2 (2004), e154–e159.
- [49] Z. Pei, Claudia Flexeder, E. Fuertes, Elisabeth Thiering, Berthold Koletzko, C. Cramer, Dietrich Berdel, Irina Lehmann, Carl-Peter Bauer, and Joachim Heinrich. 2013. Early life risk factors of being overweight at 10 years of age: Results of the German birth cohorts GINIplus and LISAplus. *European Journal of Clinical Nutrition* 67, 8 (2013), 855–862.
- [50] Mor Peleg, Samson Tu, Jonathan Bury, Paolo Ciccarese, John Fox, Robert A. Greenes, Richard Hall, Peter D. Johnson, Neill Jones, Anand Kumar, et al. 2003. Comparing computer-interpretable guideline models: A case-study approach. *Journal of the American Medical Informatics Association* 10, 1 (2003), 52–68.
- [51] Eliana Miller Perrin, Kori B. Flower, and Alice S. Ammerman. 2004. Body mass index charts: Useful yet underused. *Journal of Pediatrics* 144, 4 (2004), 455–460.
- [52] Trang Pham, Truyen Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.
- [53] Ramin Ramazi, Christine Perndorfer, Emily Soriano, Jean-Philippe Laurenceau, and Rahmatollah Beheshti. 2019. Multi-modal predictive models of diabetes progression. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 253–258.

- [54] Ramin Ramazi, Christine Perndorfer, Emily C. Soriano, Jean-Philippe Laurenceau, and Rahmatollah Beheshti. 2021. Predicting progression patterns of type 2 diabetes using multi-sensor measurements. *Smart Health* 21 (2021), 100206. <https://doi.org/10.1016/j.smhl.2021.100206>
- [55] Ilkka Rautiainen and Sami Äyrämö. 2022. Predicting overweight and obesity in later life from childhood data: A review of predictive modeling approaches. *Computational Sciences and Artificial Intelligence in Industry* (2022), 203–220.
- [56] Sarah A. Redsell, Stephen Weng, Judy A. Swift, Dilip Nathan, and Cris Glazebrook. 2016. Validation, optimal threshold determination, and clinical utility of the infant risk of overweight checklist for early prevention of child overweight. *Childhood Obesity* 12, 3 (2016), 202–209.
- [57] Elizabeth Reifsnider, Alma R. Flores-Vela, Diana Beckman-Mendez, Hoang Nguyen, Colleen Keller, and Shannon Dowdall-Smith. 2006. Perceptions of children’s body sizes among mothers living on the Texas-Mexico border (La Frontera). *Public Health Nursing* 23, 6 (2006), 488–495.
- [58] Jacob O. Robson, Sofia G. Verstraete, Stephen Shiboski, Melvin B. Heyman, and Janet M. Wojcicki. 2016. A risk score for childhood obesity in an urban latino cohort. *Journal of Pediatrics* 172 (2016), 29–34.
- [59] Krushnapriya Sahoo, Bishnupriya Sahoo, Ashok Kumar Choudhury, Nighat Yasin Sofi, Raman Kumar, and Ajeet Singh Bhadoria. 2015. Childhood obesity: Causes and consequences. *Journal of Family Medicine and Primary Care* 4, 2 (2015), 187.
- [60] Celine Saint-Laurent, Laura Garde-Etayo, and Elvire Gouze. 2019. Obesity in achondroplasia patients: from evidence to medical monitoring. *Orphanet Journal of Rare Diseases* 14, 1 (2019), 1–9.
- [61] Gillian Santorelli, Emily S. Petherick, John Wright, Brad Wilson, Haider Samiei, Noël Cameron, and William Johnson. 2013. Developing prediction equations and a mobile phone application to identify infants at risk of obesity. *PLoS One* 8, 8 (2013), e71183.
- [62] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2017), 1589–1604.
- [63] Andrew J. Steele, Spiros C. Denaxas, Anoop D. Shah, Harry Hemingway, and Nicholas M. Luscombe. 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* 13, 8 (2018), e0202344.
- [64] Nicolas Stettler, Babette S. Zemel, Shiriki Kumanyika, and Virginia A. Stallings. 2002. Infant weight gain and childhood overweight status in a multicenter, cohort study. *Pediatrics* 109, 2 (2002), 194–199.
- [65] Marinka Steur, Henriette A. Smit, C. Maarten A. Schipper, Salome Scholtens, Marjan Kerkhof, Johan C. De Jongste, Annemien Haveman-Nies, Bert Brunekreef, and Alet H. Wijga. 2011. Predicting the risk of newborn children to become overweight later in childhood: The PIAMA birth cohort study. *International Journal of Pediatric Obesity* 6, 2 Part 2 (2011), e170–e178.
- [66] A. Pediatric Learning Health System. 2021. PEDSnet: A Pediatric Learning Health System. <https://pedsnet.org>.
- [67] A. Pediatric Learning Health System. 2021. PEDSnet Common Data Model. <https://pedsnet.org/data/common-data-model>.
- [68] Diwakar Tripathi, Damodar Reddy Edla, and Ramalingaswamy Cheruku. 2018. Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification. *Journal of Intelligent & Fuzzy Systems* 34, 3 (2018), 1543–1549.
- [69] Diwakar Tripathi, Damodar Reddy Edla, Ramalingaswamy Cheruku, and Venkatanaresbhabu Kuppli. 2019. A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. *Computational Intelligence* 35, 2 (2019), 371–394.
- [70] Jelena Vekic, Aleksandra Zeljkovic, Aleksandra Stefanovic, Zorana Jelic-Ivanovic, and Vesna Spasojevic-Kalimanovska. 2019. Obesity and dyslipidemia. *Metabolism* 92 (2019), 71–81.
- [71] Jennifer Vuong, Yuelin Qiu, Myanh La, Gwen Clarke, Dorine W. Swinkels, and George Cembrowski. 2014. Reference intervals of complete blood count constituents are highly correlated to waist circumference: Should obese patients have their own “normal values”? *American Journal of Hematology* 89, 7 (2014), 671–677.
- [72] Stephen F. Weng, Sarah A. Redsell, Dilip Nathan, Judy A. Swift, Min Yang, and Cris Glazebrook. 2013. Estimating overweight risk in childhood from predictors during infancy. *Pediatrics* 132, 2 (2013), e414–e421.
- [73] Nilmini Wickramasinghe. 2017. Deepr: A convolutional net for medical records. (2017).
- [74] Gabrielle Williams, Jeffery T. Fletcher, Stephen I. Alexander, and Jonathan C. Craig. 2008. Vesicoureteral reflux. *Journal of the American Society of Nephrology* 19, 5 (2008), 847–862.
- [75] Enliang Xu, Shiwan Zhao, Jing Mei, Eryu Xia, Yiqin Yu, and Songfang Huang. 2019. Multiple MACE risk prediction using multi-task recurrent neural network with attention. In *2019 IEEE International Conference on Healthcare Informatics (ICHI ’19)*. IEEE, 1–2.
- [76] Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *ArXiv abs/1212.5701* (2012).
- [77] Jinghe Zhang, Kamran Kowsari, James H. Harrison, Jennifer M. Lobo, and Laura E. Barnes. 2018. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6 (2018), 65333–65346.
- [78] Zeyu Zheng and Karen Ruggiero. 2017. Using machine learning to predict obesity in high school students. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM’17)*. IEEE, 2132–2138.
- [79] Nida Ziauddeen, Paul J. Roderick, Nicholas S. Macklon, and Nisreen A. Alwan. 2018. Predicting childhood overweight and obesity using maternal and early life risk factors: A systematic review. *Obesity Reviews* 19, 3 (2018), 302–312.

Received January 2020; revised November 2021; accepted December 2021