

# PREDICTION OF DIABETES FROM ELECTRONIC HEALTH RECORDS

Philip de Melo

Department of Nursing and Allied Health , Norfolk State University , 700 Park Avenue  
Norfolk, VA 23504 , USA

## **ABSTRACT**

*Electronic Health Records (EHRs) encompass patients' diagnoses, hospitalizations, and medication histories, offering a wealth of data. Although EHR-based research, particularly in decision prediction, has made significant strides, challenges remain due to the inherently sparse and irregular nature of EHR data, which limits their direct application in time-series analysis.*

*Physicians treating individuals with chronic illnesses must anticipate the progression of their patients' conditions, as accurate forecasts enable more informed and timely treatment decisions. The strength of prediction lies in prevention—intervening early is often more effective than attempting to reverse damage later. In this study, we present a data-driven model designed to deliver accurate and efficient predictions of disease trajectories using electronic health records (EHRs) from Veterans Affairs hospitals.*

*Prediction of disease progression represents a fundamental challenge. EHRs contain vast volumes of frequently updated, high-dimensional, and irregularly spaced data in various formats, including numerical, textual, image, and video data. To address this complexity, we propose a new approach for predicting the progression of diabetes. This method has the potential to improve early intervention, prevent further health deterioration, and ultimately extend patients' lives.*

*The method is based on the PM GenAI, a novel approach that significantly improves classification and regression results. The method is compared to traditional techniques such as ARIMA, LTMS, and RF showing significant improvement in disease progression evaluation. The method is demonstrated on diabetes data.*

## **1. INTRODUCTION**

Electronic Health Records (EHRs) contain comprehensive documentation of patient status, making them a valuable source for tracking health information and enabling data-driven clinical decision-making. Unlike data collected from clinical trials or targeted biomedical studies, EHR data are not curated to address specific research hypotheses. Instead, they are primarily designed for patient monitoring, which introduces several complexities.

EHR data often exhibit challenging characteristics: they are typically uncurated (not deliberately organized or filtered), low in quality (rarely subjected to systematic audits), high-dimensional (containing thousands of distinct medical events), sparse (with many missing or zero values), heterogeneous (collected from diverse sources), temporal (recorded over time), incomplete, large-scale, and multimodal (capturing various data types such as images, notes, and lab results).

These complexities present obstacles for using raw EHR data directly in predictive modeling— an area of machine learning focused on building statistical models to forecast clinical outcomes. To

overcome this, a critical step involves transforming raw EHR data into meaningful, machine readable representations.

Representation learning, a key technique in machine learning, automates the extraction of useful features from raw data. In the context of EHRs, patient representation learning refers to developing structures, mathematical representations of patient data that can be effectively used by predictive models. These representations draw on multiple EHR modalities (e.g., clinical notes, lab results, medications) and must be structured in a way that facilitates accurate diagnosis, disease phenotyping, and outcome prediction.

The most relevant information in an Electronic Health Record (EHR) about a patient depends on the clinical context (e.g., routine visit, emergency, chronic disease management). However, the core essential elements typically include:

Patient Demographics Full name, date of birth, sex Contact information Insurance details Emergency contact	Active Medical Conditions Chronic diseases (e.g., diabetes, hypertension, asthma) Acute illnesses (e.g., infections, injuries) Status and history of each condition
Current Medications Drug names, dosages, routes, frequency Start/end dates All medications including over the counter or supplements	Allergies and Adverse Reactions Known drug, food, or environmental allergies Description of the reaction (e.g., rash, anaphylaxis)
Vital Signs and Measurements Blood pressure, heart rate, respiratory rate Temperature, oxygen saturation Weight, height, BMI	Lab Results and Diagnostic Tests Blood tests (e.g., glucose, cholesterol, hemoglobin A1C) Imaging reports (X-ray, CT, MRI) Pathology and microbiology results Trends over time
Past Medical and Surgical History Major illnesses, hospitalizations Surgeries for dates Family history of chronic or genetic conditions	Immunizations Vaccination status (e.g., flu, COVID-19, hepatitis B) Dates and types of vaccines
Clinical Notes Physician, nurse, and specialist documentation Assessment, diagnosis, and care plans	Lifestyle and Behavioral Information Smoking and alcohol use Physical activity level
Progress and discharge notes	Diet and nutrition Sleep habits
Encounter and Visit History Dates and reasons for past visits Diagnoses and treatments provided Emergency room or inpatient stays	Risk Factors and Social Determinants Occupational risks Housing or food insecurity Mental health indicators Substance use

Health systems in many developed countries are under mounting strain due to aging populations, the growing burden of chronic diseases, and rising per capita healthcare costs. To address these challenges, policymakers are shifting from a reactive care model—focused on treating illnesses as they occur—to a proactive approach that emphasizes early intervention to prevent negative health outcomes. Population Health Management (PHM) has emerged as a strategy to realize this shift,

aiming to achieve the “triple aim”: improving overall population health, enhancing the patient experience, and reducing healthcare costs (Berwick et al., 2008).

A core component of PHM is the effective use of data—particularly in identifying individuals at risk of future health complications, such as the onset of chronic disease (World Health Organization, 2023). Early identification allows healthcare systems to intervene and support at-risk patients, helping them maintain better health and reducing long-term healthcare utilization (Main et al., 2022). Advances in Deep Learning (DL) offer promising support for this effort, providing the ability to analyze vast healthcare datasets, uncover early indicators of disease, and predict future health trajectories within populations.

P. de Melo, (2025), demonstrated a new algorithm called PM GenAI (Principal Model Generative Artificial Intelligence) that combined with DL significantly increased the accuracy of diagnostics of diabetes reaching up to 97% of accuracy and sensitivity.

Deep learning (DL) holds significant promise in transforming the management of chronic conditions, particularly in populations affected by diseases such as Type 2 Diabetes Mellitus and Chronic Obstructive Pulmonary Disease (COPD). The incidence of chronic illnesses is steadily increasing across industrialized nations, with over one-third of EU citizens now reporting at least one chronic condition (Eurostat, 2023).

This growing prevalence is accompanied by escalating healthcare costs (Holman, 2020). For instance, chronic diseases account for up to 80% of healthcare expenditures in the EU and 86% in the USA, with projections indicating continued cost increases in the years ahead (Holman, 2020). Since these conditions are driven by both modifiable and non-modifiable risk factors, early prediction and identification can enable individuals and healthcare providers to implement preventive strategies, potentially delaying onset, improving clinical outcomes, and reducing the economic burden.

This study explores a new technology based on data augmentation and deep learning (DL) approaches in forecasting the future onset of long-term chronic conditions (LTCs) and other critical adverse health outcomes using Electronic Health Records (EHR). The widespread adoption of EHR systems, driven by digital transformation efforts in healthcare, has made EHR data increasingly available. However, these datasets present inherent challenges, namely sparsity and high dimensionality.

**Sparsity:** Many patients have data only for a small subset of possible features (e.g., not all lab tests are ordered for everyone). **High dimensionality:** EHRs include numerous possible codes for diagnoses, medications, procedures, and lab tests. **Implication:** This makes feature selection and model training more computationally challenging and prone to overfitting.

Historically, predictive models for disease diagnosis have relied heavily on domain expertise for manual feature engineering—a process that is resource-intensive and time-consuming. This dependence has made clinicians' time a bottleneck in model development, particularly considering growing shortages of qualified healthcare professionals in many regions.

In EHRs, health care is represented through structured vocabularies like ICD-10, SNOMED, NDC, and LOINC. prompting researchers to borrow techniques from Natural Language Processing (NLP). Notably, word embeddings and recurrent neural networks (RNNs) including Gated Recurrent Units (GRUs) and Long Short-Term Memory networks (LSTMs) have been adapted to model patient histories as sequences of events (Pham et al., 2016).

More recently, Med-BERT applied the popular BERT architecture which was evaluated on the prediction of diabetic heart failure (Rasmy et al.,2021). BEHRT used an adapted BERT model to create a multilabel classifier able to predict diagnoses in the next 6–12 months from previous diagnosis history (Li et al.,2020). Hi-BEHRT offered an updated version of BEHRT with an improved pre-training strategy capable of modeling longer patient histories, and increased the data scope to include medications, procedures, GP tests, drinking and smoking status as well as binned measurements for BMI and blood pressure (Li et al.,2022).

ExBEHRT also extended the feature scope of BEHRT, similarly including observation values for BMI and smoking status as well as procedures, laboratory types, age, race, and gender ( Rupp et al.,2023). Wornow et al. (2023) have pointed out however that more needs to be done to prove the practical utility of these foundation models for electronic medical records (FEMRs) to health systems. They emphasize the importance of articulating how such models could fit into clinical workflows, demonstrating their ability to improve predictive performance in contexts where less labeled data is available, and suggesting ways in which they could simplify model deployment.

This paper builds upon the previous research (de Melo, 2024) and proposes an effective methodology by which data can be taken from EHRs and used for disease prediction. It should be underscored that PHM practitioners are typically engaged with a broader range of determinants of health than other areas of clinical practice; for example, race, gender, economic deprivation, mental health, unmet social care needs and housing status (Buck et al.,2018). These determinants may be captured by many organizations including social and community care providers, local government and third sector organizations, but these may not be standardized in the same way as medical vocabularies and there is likely to be considerable local variation.

The main contributions of this research can be summarized as follows:

1. Developing an effective approach enables incorporating an augmented data set and novel features associated with wider determinants of health.
2. Demonstrating the effectiveness of pre-trained code embeddings to enhance predictive performance for key PHM outcomes where limited data is available, both within and across sub-populations.
3. Investigating whether the new algorithm named PREDMOD (Predictive Modeling with Augmented Data and Deep Learning) enables clinicians to invoke features from EHRs and conduct the predictive studies of disease progression with determining “critical” points of the progression.

Although the existing predictive model mainly focuses on the prediction of single diseases, rather than considering the complex mechanisms of patients from a holistic review it can be extended to comprehensive representations of patient EHR data. Advances in patient representation learning techniques will be essential for powering patient-level EHR analyses. Future work will still be devoted to leveraging the richness and potential of available EHR data through multivariate analysis as an extension of PREDMOD.

## **2. DATA EXTRACTION FROM EHR**

Data extraction from Electronic Health Records (EHRs) involves retrieving structured or unstructured information about a patient’s medical history, treatments, lab results, vital signs, medications, and other clinical events. Here's an overview of how data is typically extracted from EHRs, including common sources, formats, and methods:

EHRs store a wide range of information. Common categories include:

<b>Data Type</b>	<b>Examples</b>
<b>Demographics</b>	Age, sex, ethnicity, address
<b>Vitals</b>	Blood pressure, heart rate, glucose levels
<b>Diagnoses</b>	ICD-10 codes, diagnosis descriptions
<b>Medications</b>	Drug name, dose, frequency
<b>Labs</b>	Glucose, HbA1c, cholesterol levels
<b>Procedures</b>	CPT codes, surgical history
<b>Clinical Notes</b>	Free-text physician notes
<b>Allergies</b>	Allergy type, reaction severity
<b>Imaging Reports</b>	Radiology/CT/MRI results
<b>Encounters</b>	Admission/discharge, visit summaries

EHR data may be stored in:

- Relational databases (SQL) – e.g., Epic Clarity, Cerner Millennium
- FHIR APIs (Fast Healthcare Interoperability Resources) – standards-based access to EHRs
- HL7 feeds – older messaging standard for health data exchange
- CSV/Excel exports – flat files used for smaller-scale analysis
- Clinical Data Warehouses (CDWs) – consolidated data sources used by hospitals

There are the following tools to extract and process EHR data:

<b>Tool/Library</b>	<b>Purpose</b>
pandas	Load and analyze CSV/Excel data
sqlalchemy / pyodbc	Connect to SQL-based EHRs
requests	Interact with FHIR APIs
fhirclient (SMART on FHIR)	Access FHIR resources
spaCy / scispaCy	NLP on unstructured clinical notes

The data set was extracted using FHIR/API and represents diabetes recordings averaged over a week time.

### 3. DATA PROCESSING

#### 3.1. Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable  $y$  and one or more independent variables  $x$ . It assumes that this relationship is linear, meaning:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$y$ : Dependent variable (what we want to predict)  
 $x$ : Independent variable (predictor)

$\beta_0$ : Intercept (value of  $y$  when  $x=0$ )

$\beta_1$ : Slope (how much  $y$  changes for a one-unit increase in  $x$ )

$\varepsilon$ : Error term (captures noise or randomness)

Original data for glucose measured in 100 weeks and retrieved from the patient's EHR (Figure 1) while Figures 2 and 3 depict the results of linear regression applied to the original data. Linear regression shows a trend that was deduced from the data interval (0,100)

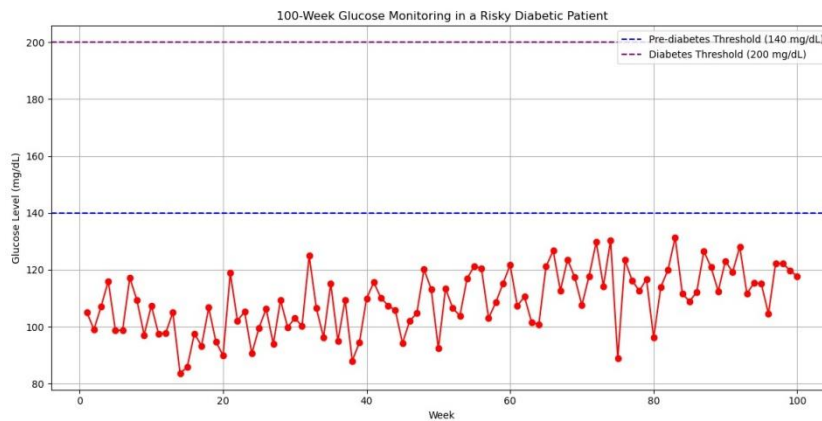


Figure 1: Original Glucose data set: The data were taken daily and averaged over the week.

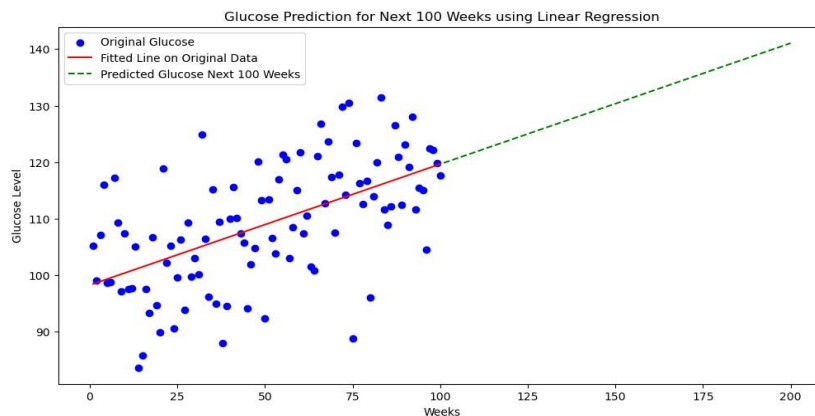


Figure 2: Linear regression algorithm applied to original data.

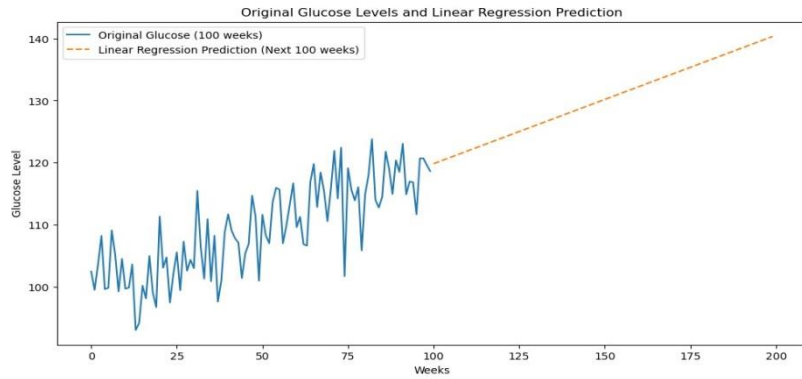


Figure 3: Linear regression algorithm applied to original data (Continuous function).

### 3.2. ARIMA- Based Prediction

ARIMA (Autoregressive Integrated Moving Average) is a powerful statistical model used for time series forecasting. It combines three key ideas:

The name ARIMA(p, d, q) refers to:

1. AR (Autoregressive, p): Uses the past values (lags) of the series to predict the future.

$$y_t = \varphi + y_{t-1} + \dots + \varphi_p y_{t-p} + \varepsilon_t$$

I (Integrated, d): Represents the number of differences needed to make the time series stationary (i.e., constant mean and variance over time). If d=1, it means we use the first difference:

$$y'_t = y_t - y_{t-1}$$

2. MA (Moving Average, q):

Uses past forecast errors in a regression-like model.

$$y_t = \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Putting it all together:

$$y'_t = \varphi_1 y'_{t-1} + \dots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where  $y'_t$  is the differenced series (depending on d).

Time series data with trend, seasonality, or noise, when forecasting future values, works best when data is stationary (non-stationary series are different)

Model Selection (How to choose p, d, q): ADF Test or KPSS Test: Check stationarity and determine d, ACF (Autocorrelation Function): Helps identify q, PACF (Partial ACF): Helps identify p, Use AIC/BIC for model comparison. Figure 4 is the ARIMA prediction. It takes the data at the final point (week 100) and extrapolates this value to week 200.

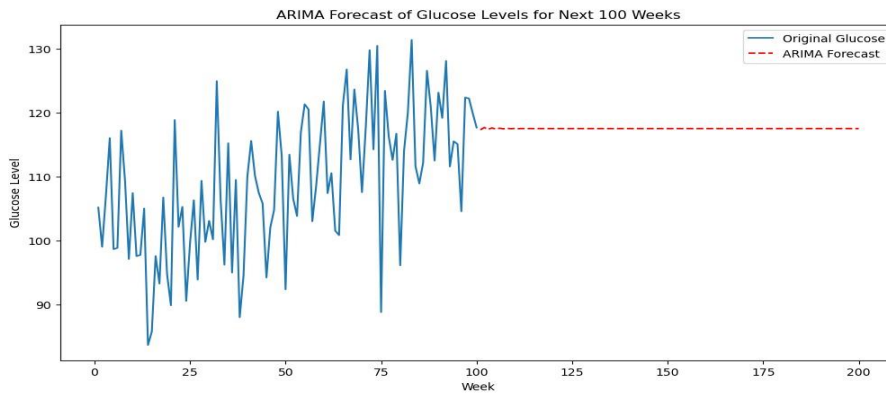


Figure 4: Result of ARIMA algorithm prediction.

### 3.3. Random Forest prediction

Random Forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to improve accuracy and robustness. It is widely used for both classification and regression tasks.

A decision tree is a model that splits data into branches to make predictions. It's prone to overfitting (too closely modeling the training data), especially when deep. Random Forest uses the idea of ensemble learning — combining multiple models to produce a better result.

It builds many decision trees and averages their predictions (for regression) or uses majority voting (for classification).

Random Forest trains each tree on a random sample (with replacement) of the training data. This reduces variance and prevents overfitting. At each split in a tree, only a random subset of features is considered. This introduces diversity among trees and makes the model less correlated. The outputs of all trees are combined: Classification and regression. Key parameters are number of trees in the forest, maximum depth of each tree, number of features to consider when looking for the best split, whether bootstrap samples are used. Figure 5 shows an important element of RF, bootstrapping. Figure 4 depicts the original data (in black) and 3 bootstrap samplings.

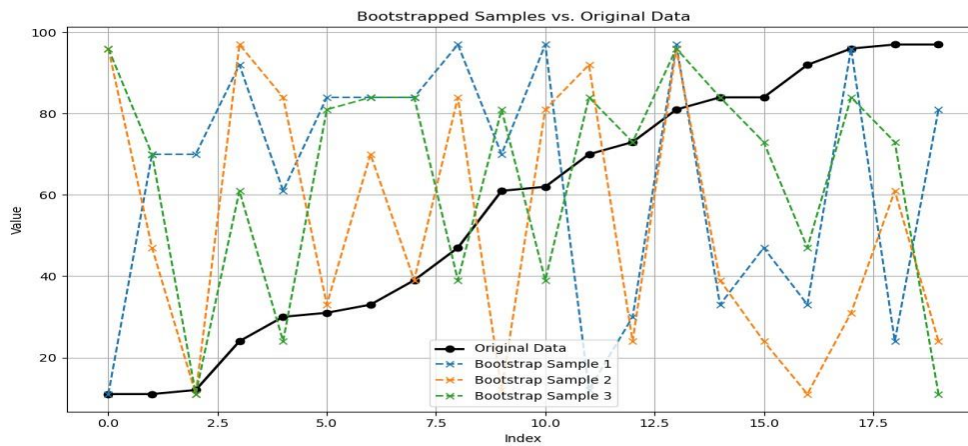


Figure 5: Original dataset and three bootstrap samples.

RF advantages include High accuracy, handles non-linear relationships, works well with highdimensional data, reducing overfitting through averaging, and handles missing values and categorical features. RF is used to predict disease progression from EHR data. Figure 6 depicts the result of RF prediction. It is similar to ARIMA (a horizontal line taken as the last sample of the original data).

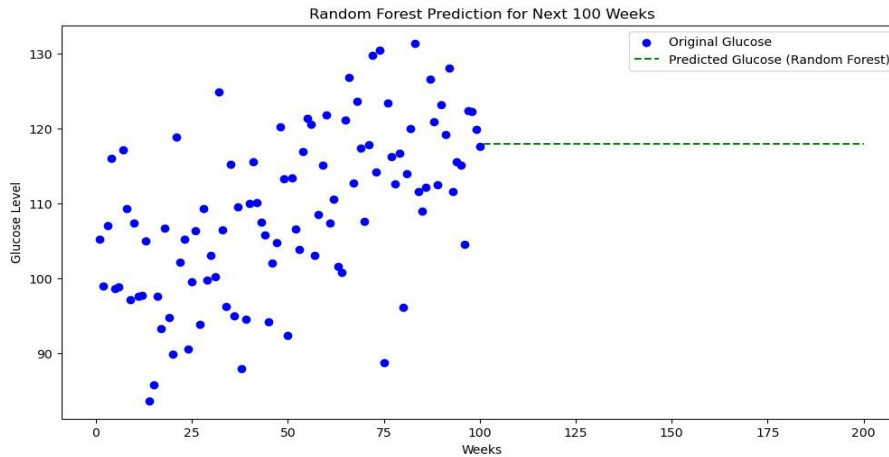


Figure 6: Result of RF algorithm prediction.

Random Forests are non-parametric and rely heavily on observed patterns. They don't extrapolate — they just replicate what they've seen. After week 100, no new behavior is learned, so the model "freezes" predictions at the nearest known values.

### 3.4. LSTM Prediction

Long Short-Term Memory networks (LSTM) are a specialized type of recurrent neural network (RNN) designed to retain information over long sequences, effectively allowing important data to persist throughout time steps. Unlike traditional RNNs, which struggle with learning longterm dependencies due to the vanishing gradient problem, LSTMs are specifically engineered to overcome this limitation.

Originally introduced by Hochreiter and Schmidhuber, LSTMs address the shortcomings of conventional RNNs and earlier machine learning models. They use a unique memory cell architecture that enables them to maintain information over extended sequences without losing relevance or stability.

To illustrate, consider watching a video or reading a book—you naturally retain previous scenes or chapters to understand the current context. RNNs mimic this behavior by using past inputs to influence current processing. However, standard RNNs tend to "forget" over time. LSTMs solve this by preserving long-term dependencies more effectively.

LSTM models can be implemented in Python using libraries such as Keras, making them accessible for a wide range of deep learning applications including time series forecasting, natural language processing, and speech recognition.

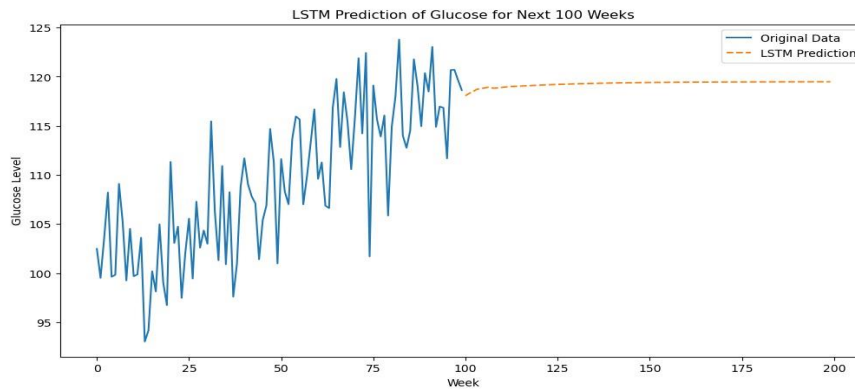


Figure 7: LSTM algorithm prediction

We can see that LSTM (Figure 7) produce predictions as almost horizontal lines. The models fail to generalize and simply repeat a mean value or flat linear trend for all future steps. This happens when there is no sufficient data samples.

## 4. DATA AUGMENTATION AND PREDICTION

### 4.1. Gaussian Augmentation and Prediction

One of the key advantages of data augmentation is that it relies solely on the existing training data, making it a cost-efficient method to expand both the size and diversity of the dataset. It is a powerful strategy to address overfitting, a common issue in deep learning where models perform well on training data but poorly on new, unseen data. By generating additional, varied samples, data augmentation helps models learn more generalizable patterns, thereby improving performance and robustness.

Another major benefit lies in addressing class imbalance, where some categories have significantly fewer examples than others, leading to biased predictions. Data augmentation helps mitigate this by creating synthetic instances for underrepresented classes, resulting in a more balanced dataset and better overall classification accuracy.

By exposing the model to a broader array of data variations, augmentation strengthens its ability to handle real-world input variability and reduces its sensitivity to noise or artifacts present in the original training set.

One popular technique for synthetic data generation is Gaussian augmentation, which is based on the Gaussian Mixture Model (GMM). This probabilistic model represents a complex data distribution as a combination of several Gaussian (normal) distributions, capturing multiple modes and patterns within the dataset.

Mathematically, a GMM is expressed as:

$$p(x) = \sum_{i=1}^k \pi_i N(x|\mu_i, \Sigma_i)$$

where:

$\pi_i$  are the mixture weights,  
 $\mu_i$  and  $\Sigma_i$  are the mean and covariance of the  $i^{\text{th}}$  Gaussian component,  
 $K$  is the number of components in the mixture.

This flexible model is widely used for generating realistic synthetic data that reflects the statistical properties of the original dataset. Figure 8 illustrates the original and predicted data.

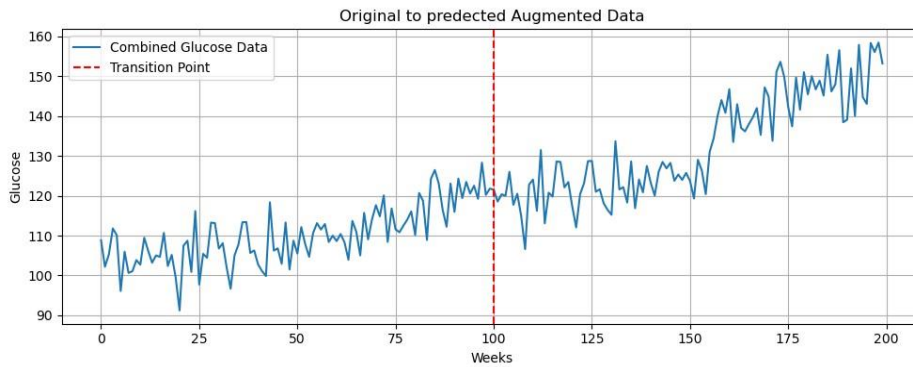


Figure 8: Predicted data after week 100 using augmented data

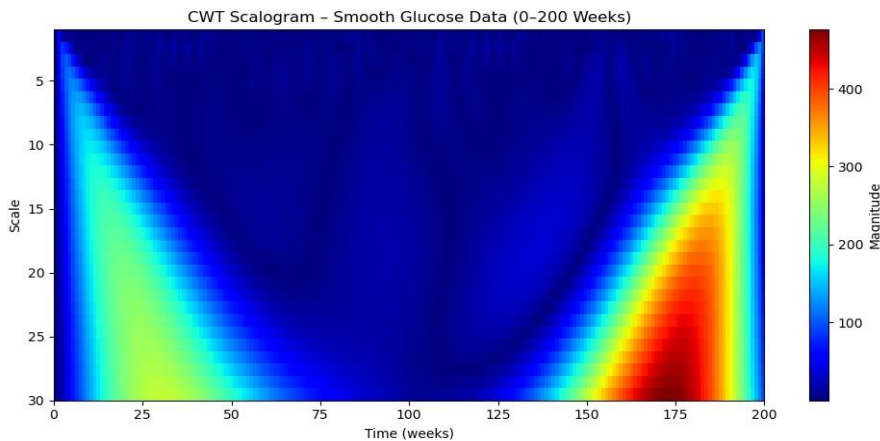


Figure 9: CWT scalogram depicts the critical zone (considerable rise of glucose after 150 weeks)

A Continuous Wavelet Transform (CWT) scalogram (Figure 9 and Figure 11) is a visual representation of how the frequency content of a signal changes over time. It is especially useful for non-stationary time series data like physiological signals (e.g., glucose levels), where frequency characteristics may vary across time.

What the CWT Scalogram Shows:

Time (X-axis): Represents the temporal evolution of the signal, shows when certain patterns or events occur.

Scale or Frequency (Y-axis): The "scale" is inversely related to frequency. For high scale  $\rightarrow$  low frequency (longer-term trends). Low scale  $\rightarrow$  high frequency (short-term fluctuations). This axis

tells at what scale or frequency components are present. Color Intensity (Z-axis, represented by color) represents the magnitude (or energy) of the wavelet coefficients at each time and scale.

Brighter or more intense colors indicate stronger presence of that frequency component at that point in time. For example, a red area at time 175 and scale 30 might suggest a significant high frequency event at that time (jump of the glucose value).

#### 4.2. Gibbs Augmentation and Prediction

Gibbs Sampling-based Data Augmentation for Time Series involves generating new synthetic data points that statistically follow the distribution of the original time series. While traditional Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method used in Bayesian inference, we can adapt its principles to time series augmentation by iteratively sampling each point conditioned on the previous one. A Gibbs distribution is defined as:

$$p(x, y) = \frac{1}{Z} \exp(-\beta E(\tilde{x}, y)) \quad (6)$$

$E(x,y)$ : Energy function (often related to loss or negative log-likelihood),  $\beta$ : controls sharpness,  $Z$ : Partition function (normalization constant). The augmented samples can be derived from the conditional probability (to ensure correspondence of augmented samples and labels  $y$ ). For unlabeled data such as time series, we use the following procedure: Let

$$x_1, \dots, x_M$$

If we want to sample from the joint distribution  $P(x_1, \dots, x_M)$ , Gibbs sampling updates each variable one at a time:

$$x_i^{(t+1)} \sim P \left( x_i | x_i^{(t+1)} \dots x_{i-1}^{(t+1)}, x_{i+1}^{(t+1)} \dots x_M^{(t)} \right)$$

In the time series context, each  $x_t$  can be sampled from a conditional distribution:

$$x_t \sim P(x_t | x_{k-1}, x_{t-2} \dots)$$

This captures temporal dependencies, like in an autoregressive model.

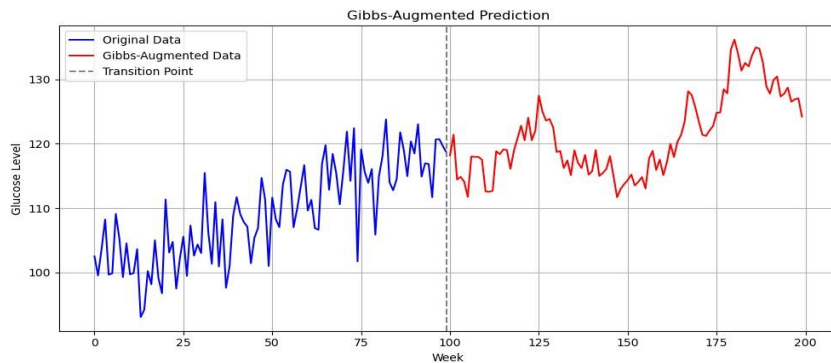


Figure 10: Gibbs prediction of the original data

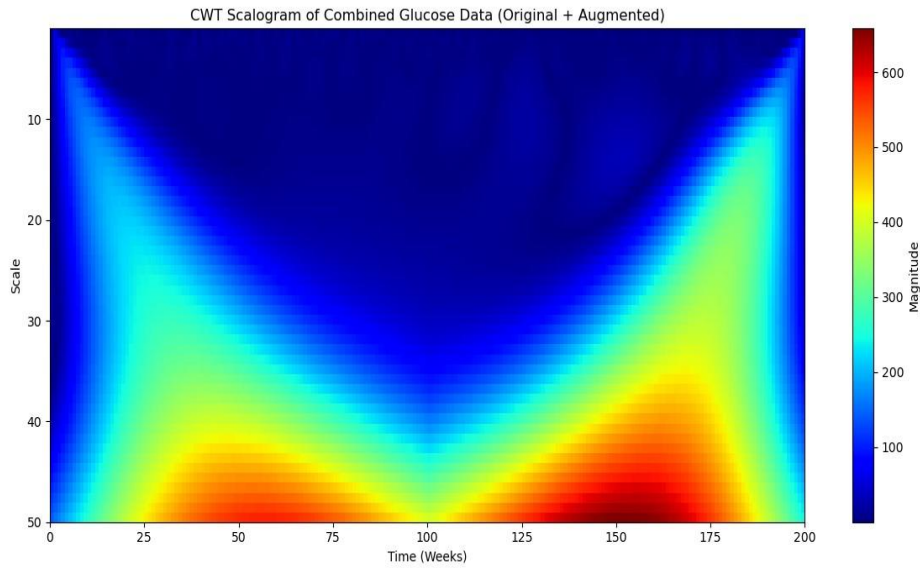


Figure 11: CWT scalogram depicts the critical zone (considerable rise of glucose at 150 week mark)

The algorithm divides the data into test and training. In the training data set it defines the trend and statistics: Mean vector ( $\mu$ ), the center of the distribution, covariance matrix ( $\Sigma$ ): the shape, spread, and orientation. The PDF of a Gaussian component is:

The covariance matrix is:

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$

In this expression:

$$N_k = \sum_{i=1}^N \gamma_{ik}$$

The trend is recovered from:

$$T_t = \frac{1}{k} \sum_{i=-\frac{k}{2}}^{\frac{k}{2}} Y_{t+i}$$

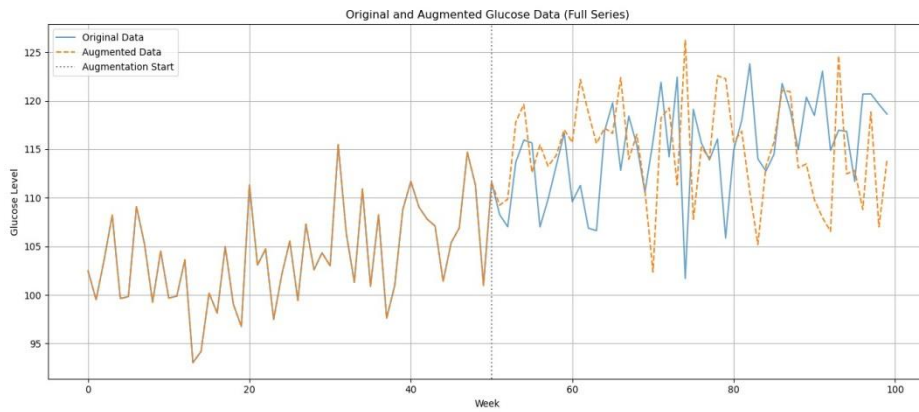


Figure 12: Original (red) and Gaussian augmented data (test data in the interval 50-100)

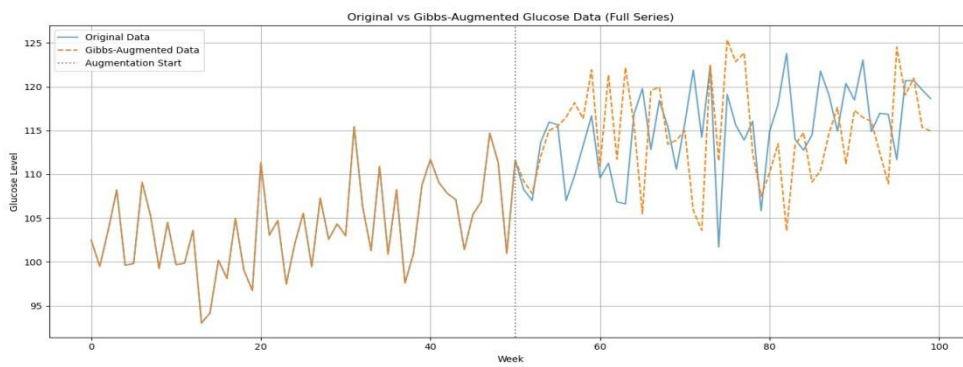


Figure 13: Original (red) and Gibbs augmented data (test data in the interval 50-100)

Type of augmentation	KS p-value
In Gaussian augmentation	KS p-value: 0.9667
In Gibbs augmentation,	KS p-value: 0.8693

Both approaches are acceptable, but Gaussian augmentation is slightly better.

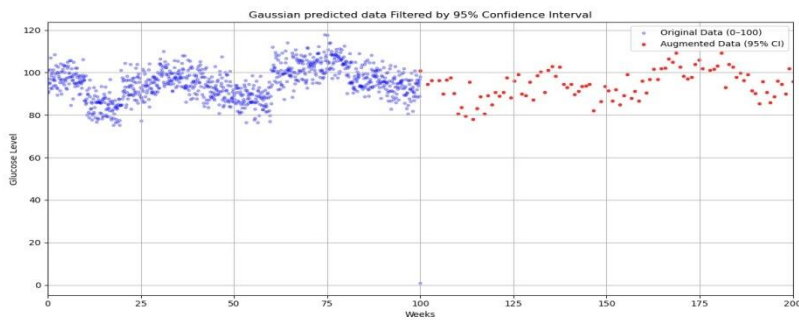


Figure 14: Original (blue) and predicted data (test data in the interval 100-200)

## 5. CONCLUSION

This paper presents a comprehensive overview of methodologies, applications, and implications of using Electronic Health Records (EHR) for detecting and modeling disease progression. With

the increasing adoption of EHR systems, the opportunity to utilize longitudinal patient data for early detection, monitoring, and prediction of disease trajectories has expanded significantly. We discuss computational approaches, data preprocessing, modeling techniques, and ethical considerations associated with this area of healthcare analytics.

Chronic diseases, such as diabetes require sustained monitoring and timely interventions. Traditional clinical models rely on episodic observations, limiting early and accurate detection of progression. EHR systems offer a rich, continuous data source encompassing clinical notes, laboratory results, medication records, imaging, and demographic information. Leveraging this data with machine learning and statistical modeling can yield valuable insights into disease progression pathways. The paper explores how EHR data can be used to identify early signs of disease progression, reviews computational methods for modeling disease trajectories, examines challenges in data quality, interoperability, and privacy.

EHR data includes structured (e.g., lab results, diagnosis codes) and unstructured (e.g., clinical notes) data collected over time. Standard sources include HL7, FHIR, and OMOP-based data warehouses. Preprocessing includes Standardizing units and terminologies (e.g., LOINC, SNOMED CT), structuring data into patient timelines, using imputation techniques or excluding incomplete records and ensuring compliance with HIPAA and GDPR regulations.

Modeling Approaches include: Random forests, linear and polynomial regression, Neural Networks, Deep Learning and Sequence Models: LSTM and Transformer-based models for time-series EHR data.

Applications focus on: Diabetes Progression including glucose readings to predict progression from pre-diabetes to type 2 diabetes, EHR data is used to model deterioration patterns, identifying early signs such as elevated BNP levels or declining ejection fraction, natural language processing (NLP) techniques applied to pathology reports and progress notes to detect recurrence risk of cancer recurrence.

Challenges and limitations in disease progressing include: Data quality ( EHRs often contain errors, redundancies, or incomplete records, interoperability (data silos and inconsistent formats hinder integration, bias and fairness (algorithms may reflect biases present in historical data), privacy (strict governance is necessary to protect patient data).

Detecting and modeling disease progression using EHR data represents a powerful advancement in precision medicine. When implemented responsibly, it can enhance clinical decision-making, improve patient outcomes, and reduce healthcare costs. Continued research, cross-disciplinary partnerships, and policy support are essential to fully realize its potential.

There is a common attitude towards the disease progression evaluation from EHRs is that these predictions are inaccurate and risky. They aim to estimate values beyond the observed range of a dataset by extending patterns found within the data. Disease progression evaluation based on modern data science is a powerful but inherently uncertain technique. Its effectiveness depends on the quality of the data, appropriateness of the model, and proximity of the extrapolated region to observed data. When used responsibly and with proper uncertainty quantification, it can be a valuable tool for informed forecasting and strategic planning. In this paper, we discuss a new approach based on the

## 6. RECOMMENDATIONS

- Invest in interoperable, standardized EHR systems.

- Foster collaborations between clinicians, data scientists, and ethicists.
- Encourage open datasets and model benchmarking initiatives.
- Establish regulatory frameworks for AI-assisted diagnosis.

## REFERENCES

- [1] Berwick D. M., Nolan T. W., Whittington J. (2008). The triple aim: care, health, and cost. *Health Aff.* 27, 759, 769. 10.1377/hlthaff.27.3.759
- [2] Buck D., Baylis A., Dougall D., Robertson R. (2018). *A Vision for Population Health: Towards a Healthier Future*. London: Kings Fund.
- [3] deMelo P., (2025). Accurate Classification of Diabetes using Optimized Deep Learning Algorithm, *Journal Diabetes Meletus*, (accepted for publication)
- [4] de Melo (2024) *Public Health Informatics and Technology*. AAAS press, ISBN ISBN 13 9798893729535
- [5] Eurostat (2023). Self-Perceived Health Statistics Available online at: [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Selfperceived\\_health\\_statistics&oldid=509628](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Selfperceived_health_statistics&oldid=509628)
- [6] Holman H. R. (2020). The relation of the chronic disease epidemic to the health care crisis. *ACR Open Rheumatol.* 2, 167, 173. 10.1002/acr2.11114
- [7] Li Y., Mamouei M., Salimi-Khorshidi G., Rao S., Hassaine A., Canoy D., et al. (2022). Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inf.* 27, 1106, 1117. 10.1109/JBHI.2022.3224727
- [8] Main C., Haig M., Kanavos P. (2022). *The Promise of Population Health Management in England: From Theory to Implementation*. London: The London School of Economics and Political Science.
- [9] Pham T., Tran T., Phung D., Venkatesh S. (2016). “Deepcare: a deep dynamic memory model for predictive medicine,” in *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20* (Auckland: Springer; ), 30, 41.
- [10] Rasmy L., Xiang Y., Xie Z., Tao C., Zhi D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Dig. Med.* 4, 86. 10.1038/s41746-021-00455-y
- [11] Rupp M., Peter O., Pattipaka T. (2023). Exbehrt: Extended transformer for electronic health records to predict disease subtypes & progressions. *arXiv [preprint]*. 10.1007/978-3-031-39539-0\_7
- [12] World Health Organization (2023). *Population Health Management in Primary Health Care: a Proactive Approach to Improve Health and Well-Being: Primary Health Care Policy Paper Series*. No. WHO/EURO: 2023-7497-47264-69316. World Health Organization. Regional Office for Europe.
- [13] Wornow M., Xu Y., Thapa R., Patel B., Steinberg E., Fleming S., et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Dig. Med.* 6, 135. 10.1038/s41746-023-00879-8
- [14] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts, “Deep representation learning of patient data from electronic health records (EHR): a systematic review,” *J Biomed Inform*, 2020.
- [15] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sunl, “RAIM: recurrent attentive and intensive model of multimodal patient monitoring data,” *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2018*, pp. 2565–2573.
- [16] S. Yang, X. W. Zheng, C. Ji and X. C. Chen, “Multi-layer Representation Learning and Its Application to Electronic Health Records,” *Neural Process Lett.* 2021 ; 53 ( 2 ): 1417–1433.
- [17] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, “Attend and diagnose: clinical time series analysis using attention models,” *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2018*, pp. 4091–4098.
- [18] Y. Si, and K. Roberts, “Deep patient representation of clinical notes via multi-task learning for mortality prediction,” *AMIA Jt Summits TranslSciProc*, 2019, pp. 779–788.

- [19] L. Liu, H. Li, Z. Hu, H. Shi, Z. Wang, J. Tang, and M. Zhang, "Learning hierarchical representations of electronic health records for clinical outcome prediction," *AMIA AnnuSympProc*, 2019, pp. 597–606.
- [20] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm, "Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk," *Scientific Reports*, 2020, 10 ( 1 ): 1111.
- [21] F. Yuan, S. Chen, K. Liang and L. Xu, "Research on the coordination mechanism of traditional chinese medicine medical record data standardization and characteristic protection under big data environment," Shandong People's Publishing House.
- [22] X. M. Yu and H. Wang, "Intelligent data mining-frequent Patterns for uncertain data," *tsinghua university press*, 2018, 06.
- [23] X. W. Zheng, X. M. Yu, Y. Q. Yin, T. T. Li and X. Y. Yan, "Three-dimensional feature maps and convolutional neural network-based emotion recognition," *International Journal of Intelligent Systems* 36 ( 2021 ): 6312–6336.
- [24] X. W. Zheng, M. Zhang, T. Li, C. Ji and B. Hu, "A novel consciousness emotion recognition method using ERP components and MMSE," *J Neural Eng.* 2021 Apr 18; 18 ( 4 ).
- [25] Y. Q. Yin, X. W. Zheng, B. Hu, Y. Zhang and X. C. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.* 100 ( 2021 ): 106954.
- [26] Y. Jiang, Y. Zheng, S. Hou, Y. Chang, and J. C. Gee, "Multimodal image alignment via linear mapping between feature modalities, *J HealthcEng.*" 2017, pp. 1–6.
- [27] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2Vec: deep representation learning for clinical temporal data." *Neurocomputing*, 2019, pp. 31–42.
- [28] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: predicting clinical events via recurrent neural networks," *JMLR Workshop ConfProc*, 2016, 56 : 301–318.
- [29] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches," *Computer Science*, 2014, pp. 103–111.
- [30] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997, 9 ( 8 ): 1735–1780.
- [31] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar. "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Aff*, 2014, 33 ( 7 ): 1123–1131.
- [32] R. Miotto, W. Fei, and W. Shuang, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, 2017, 19 ( 6 ).
- [33] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1903–1911.
- [34] C. Chao, X. Cao, L. Jian, J. Bo, and W. Fei, "An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease," *Proceedings of the 2017 SIAM International Conference on Data Mining*, 2017, pp. 198–206.
- [35] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. P. Wei, "Predicting the risk of heart failure with EHR sequential data modeling," *IEEE Access*, 2018, pp. 9256–9261.
- [36] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2Vec: a personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, 2018, pp. 65333–65346.
- [37] S. Rendle, "Factorization machines," *The 10th IEEE International Conference on Data Mining*, Sydney, Australia, 14–17 December 2010.
- [38] T. Shen, J. Jia, T. S. Chua, W. Hall, and B. Chen, "PEIA: personality and emotion integrated attentive model for music recommendation on social media platforms," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34 ( 01 ): 206–213.