IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 25, 2019, accepted January 4, 2020, date of publication January 22, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968608

Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Technique

SAJIDA PERVEEN[1,2], MUHAMMAD SHAHBAZ[1,3], TANZILA SABA[4], (Senior Member, IEEE), KARIM KESHAVJEE[3,5], AMJAD REHMAN[4], (Senior Member, IEEE), AND AZIZ GUERGACHI[3,6,7]

[1]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore 39161, Pakistan
[2]Artificial Intelligence and Data Analytics (AIDA) Laboratory, Prince Sultan University, Riyadh 12435, Saudi Arabia
[3]Research Laboratory for Advanced System Modelling, Ryerson University, Toronto, ON M5B 2K3, Canada
[4]Artificial Intelligence and Data Analytics (AIDA) Laboratory, CCIS, Prince Sultan University, Riyadh 12435, Saudi Arabia
[5]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, ON M5S, Canada
[6]Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada
[7]Ted Rogers School of Information Technology Management, Ryerson University, Toronto, ON M5B 2K3, Canada

Corresponding author: Sajida Perveen (sajida.uaar@gmail.com)

This work was supported by the Artificial Intelligence and Data Analytics Laboratory (AIDA), Prince Sultan University, Riyadh, Saudi Arabia.

ABSTRACT Clinical researchers use prognostic modeling techniques to identify a-prior patient health status and characterize progression patterns. It is highly desirable to predict future health condition especially to implement preventive and intervention strategies in pre-diabetic individuals. Hidden Markov Model (HMM) and its variants are a class of models that provide predictions concerning future condition by exploiting sequences of clinical measurements obtained from a longitudinal sample of patients. Despite the advantages of using these models for prognostic modeling, it still face barriers and significant challenges, to effectively learn dynamic interactions, when using irregularly sampled longitudinal Electronic Medical Records (EMRs) data. Newton's divide difference method (NDDM) is a classical approach for handling irregular data in terms of divided difference. However, as it is polynomial approximation technique, it suffers with Runge Phenomenon. The problem can be even more severe when the interval is a bit extended. Therefore, to tackle this problem, we proposed a novel approximation method based on NDDM as a component with HMM in order to estimate the 8 years risk of developing Type 2 Diabetes Mellitus (T2DM) in a particular individual. The proposed method is evaluated on real world clinical data obtained from CPCSSN. The results demonstrated that our proposed technique has the ability to exploit the available irregularly sampled EMRs data for effective approximation and improved prediction accuracy.

INDEX TERMS Type 2 diabetes mellitus (T2DM), risk prediction, Newton's divided difference method (NDDM), irregular and sparsely sampled data handling, approximation technique, HMM, prognostic modeling, machine learning, risk scoring.

I. INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a significant public health problem that is approaching epidemic proportions globally [1]. It is a metabolic syndrome characterized by hyperglycemia [2]. T2DM can leads to lifelong dysfunction, failure and damage of different vital organs, particularly eyes, kidney, nerves, heart, and blood vessels [3], [4]. Diabetic patients are at increased risk for developing peripheral vascular, cardiovascular and cerebrovascular malady [4], and in its most severe forms, diabetes may lead to death.

Diabetes is one of the leading endocrine drivers to the global burden of disease [5]. The predominance of diabetes is consistently escalating at an alarming pace especially in developing countries [6]. According to World Health Organization (WHO) diabetes is at present the fifth most common

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

reason for death in the world [7]. According to the International Diabetes Foundation (IDF) after every six seconds one person dies of diabetes [8]. Beside these, in 2017, approximately 425 million adults were living with diabetes; the number is projected to ascend to 629 million by 2045 [9]. Furthermore, the subsequent economic and societal burdens due to healthcare expenditures of diabetes are of significant immensity [6], [45]. Globally, it was estimated that, in 2017, approximately $727 billion were spent on healthcare expenditure related to diabetes and its complications which accounts for 12% of global spending on healthcare [10]. Moreover, a huge loss in global economy is also witnessed due to reduced productivity caused by patients' health condition. Furthermore, diabetic individuals remain oblivious of their disease status, even the disease is persisting for years, as they often have no overt symptom at first, which makes the situation much worst [2], [6]. Therefore, early diagnosis of T2DM is a challenging problem. When considering the above narrated facts diabetes seems to have more than its fair share of challenges, when compared to other diseases.

These alarming figures, undoubtedly, require incredible considerations and implementation to optimally treat diabetic patients and to prevent or halt its development in vulnerable individuals. Early and intensive prevention strategies in high risk individuals not only improve the disease outcome but also reduce other complications associated with therein. To this end, application of Machine Learning techniques (ML) is presently more than ever before, indispensable and vital in efforts to transform intelligently large pool of data into valuable information [11], [46], [47], [48]. In particular, Electronic Medical records (EMRs), an inter-organizational, comprehensive, patient-centric longitudinal collection of health records, resulted from remarkable advances in biotechnology and health sciences play an integral part for prognostic prediction of T2DM risk [12]. Therefore, EMRs data is becoming the key driving force for the adaptation of data analytics techniques in healthcare domain, bringing the opportunities to foster the quality of healthcare services and support healthcare providers by providing comprehensive healthcare information of a particular patient [11], [13], [14].

Machine learning based techniques have the ability to provide preliminary judgment about disease progression that could serve as a reference for healthcare providers. Hidden Markov Model (HMM) has been widely used for modeling dynamic systems [15]–[17]. Although, it is a potentially powerful and conceptually simple technique to model disease surveillance data [18], [19], [44], it appears to has been rarely used in public health practice [43].

Furthermore, the nature of the clinical setting, together with the format of the EMRs data particularly presents versatile substantial challenges that confound classical HMM and grossly violate the model assumption. For example, consider the problem of learning from irregular and sparely sample data which is a common but complicated problem in almost every healthcare data, presents significant challenges in inference and learning process, and/or may significantly harm the performance of downstream applications [20], [21]. Hence, it is impractical to directly feed irregular and sparsely sampled data into HMM for prognostic prediction of T2DM risk in an individual. In addition, there can also be substantial uncertainty about the underlying temporal processes due to the sparsity of observations.

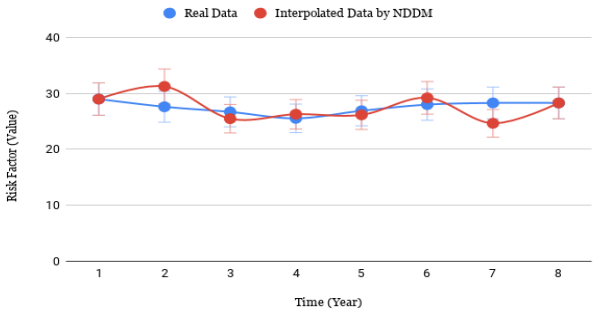

FIGURE 1. Example dataset derived from CPCSSN that is affected with Runge Phenomena.

Therefore, we proposed a novel technique based on NDDM to accommodate irregular and sparsely sampled EMRs data with the objective to overcome the analytical challenges that arise primarily due to such data types. Basically, NDDM is a classical technique employed for interpolating polynomial in terms of divided differences. Divided differences are also independent of the order of arguments. Furthermore, NDDM works well as compared to other approximation techniques [22]–[25]. However, as it is polynomial approximation so it suffers with Runge Phenomenon (RP) as depicted in Figure 1. In general, RP is the divergence of the interpolant at edges of an interval. The problem can be even more severe when the interval is a bit extended. Hence, to cope up with the above mentioned problems, we devise a novel method that not only has the ability to deal with the RP but also effectively transform heterogeneous time series clinical data measured at irregular intervals into irregular data in order to develop a robust method to investigate the ongoing risk T2DM in an individual.

II. METHODOLOGY

A. HEALTHCARE DATA

The dataset being employed to evaluate our proposed method is acquired from CPCSSN. It is especially focusing on five chronic diseases including diabetes and three neurological diseases. It makes it realizable to collect health data from all participating networks across Canada into a centralized database (http://cpcssn.ca/). The data is further utilized for research purposes that may lead to improve healthcare management worldwide. CPCSSN contains the information of 172,168 unique patients and consist of a total of 812,007 records, spanning 13 years timeframe (from 2003 to 2015) and every record encompasses of distinct traits concerning vital signs, diagnosis and demographics. Each patient is described by the clinical measurements that have been collected within 13 years including information related to diastolic Blood Pressure (dBP),

systolic Blood Pressure (sBP), Fasting Blood Glucose (FBG), Triglycerides(TG), Body Mass Index (BMI), High Density Lipoprotein (HDL), Glycated Hemoglobin (HbA(1c) and Gender. The general time period defined for each patient to be included in the study sample is the same, i.e. 8 years regardless the disease status. To capture a representative cohort in the derived study sample, we included information related to only those individuals who have at least five clinical visits by June 2015 and have data for all the risk factors consider relevant in this study. Hence, a total of 170,250 patients do not satisfied minimum inclusion criteria and excluded from the data. Hence, a total of 1981 individuals were included in this prospective research study sample. The final dataset comprised of 775 (61.03%) female and 1143 (38.96%) male and among them 584 (23.49%) are diabetic patients, described by clinical measurements as mentioned above. An abstract overview of the CPCSSN can be found in [2]. CPCSSN acquired written consent form all the participating networks.

Our primary objective of this research is to evaluate the potential of EMRs in prognosticating ongoing risk of T2DM in a particular individual. In this research we specifically ask three major questions. (1) Can a machine learning based prognostic model be developed or derived to estimate 8 years risk of T2DM? (2) Can we effectively identify potentially relevant risk factors to estimate the T2DM risk? (3) Is it possible to propose a novel approximation technique that leverage irregularly and sparsely sample EMRs data represent it to regular space data to prognosticate 8 years risk of T2DM in a particular individual? This type of prognostic prediction provides evidence to make better decisions around patients' risk, individualized treatments and timely interventions [26], [27].

B. PROPOSED METHOD

Keeping in view the aim of this research and significant challenges inherent with the EMRs data the proposed methodology can be divided into two major components. (1) Handling sparse and irregularly sampled EMRs data and (2) prognostic modeling using approximated data prepared in the first phase in order to investigate the ongoing risk of T2DM.

1) HANDLING SPARSE AND IRREGULARLY SAMPLED EMRs DATA

Suppose a regularly sampled EMRs data that consists of n independent instances $D = \{S_1, S_2, \ldots\ldots, S_n\}$ recorded at uniformly distributed time interval. Each S_i associated with a list of time points $t_i = \{t_{i1}, t_{i2}, \ldots\ldots, t_{in}|S_i|\}^T$, and the related values, $y_i = \{y_{i1}, y_{i2}, \ldots, y_{im}|S_i|\}^T$. Whereas, irregularly sampled data is considered as a sparse matrix with features and a time dimension. NDDM, a classical technique, is used for interpolating polynomial in terms of divided differences. It can be defined as follows: given a set of pairs of numbers $(x_0, f_0), (x_1, f_1), \ldots\ldots\ldots, (x_n, f_n)$, where $x_1, x_2, \ldots\ldots, x_n$ are distinct and not necessarily distributed equally over time. While on the contrary, f_i can be obtained empirically from an observation or experiment or it can be the value of some mathematical function f(x). The interpolation problem is to find a polynomial $P_n(x)$ such that $P_n(x) = f_0, P_n(x_1) = f_1, \ldots\ldots\ldots, P_n(x_n) = fn$.

The polynomial $P_n(x)$ is used to estimate value for all x such that $P_n(x)$ is approximately f(x) or to get values for x_s at which no measurement was taken. Furthermore, following the approximation method, all D-dimensions of x_s are represented in term of output defined on the regularly spaced set of reference time points.

It can be written in the Newton form as follows [28]:

$$\begin{aligned} P_n(x) = {} & f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0) \\ & \times (x - x_1) + f[x_0, x_1, x_2, x_3](x - x_0) \\ & \times (x - x_1)(x - x_2) + \ldots\ldots \\ & + f[x_0, x_1, x_2, \ldots x_n](x - x_0) \\ & \times (x - x_1)(x - x_2) \ldots\ldots (x - x_n) \end{aligned}$$

where $f[x_0]$, $f[x_0, x_1]$ and $f[x_0, x_1, x_2]$ are the first, second, and third order finite divided differences respectively that can be defined as below:

$$\begin{aligned} f[x_0] &= f(x_0) \\ f[x_0, x_1] &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\ f[x_0, x_1, x_2] &= \frac{\frac{f(x_2) - f(x_1)}{x_1 - x_0} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \end{aligned}$$

Correspondingly, n^{th} Divided Difference can be calculated as below:

$$f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+n}] = \frac{f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+n}] - f[x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+n-1}]}{x_{i+n} - x_i}$$

a: NEWTON DIVIDED DIFFERENCE METHOD (NDDM) AND RUNGE PHENOMENON

NDDM is a polynomial interpolation technique that is utilized to approximate the unknown values by exploiting the previously known information from the related observations. Perveen et al. [49] also utilized NDDM for leveraging EMRs data to develop prognostic model for T2DM and this work is primarily based on this research. Let D is a EMRs data containing information related to n distinct individuals $\{S_1, S_2, \ldots\ldots, S_n\}$ recorded over time T. Each S_i is represented as an order set of measurements taken over a particular time interval $t_i = \{t_{i1}, t_{i2}, \ldots\ldots, t_{in}|S_i|\}^T$ with $t_i \epsilon R$, $\forall_i \epsilon [1, k]$ as well, typically correspond to a series of risk factor i.e. high density lipoprotein or/and body mass index of a patient over a particular time interval. Given fixed risk factors d_i, we can represent D-dimensional time series for data case k as a real valued tuple $S_{dk=}(x_{dk}, y_{dk})$ where $y_{dk} = x_{dk} = \{t_{i1}, t_{i2}, \ldots\ldots, t_{ik}|S_i|\}^T$, is a series of time interval at which observations are defined and $y_{dk} = \{y_{i1}, y_{i2}, \ldots\ldots, y_{ik}|S_i|\}^T$ is corresponding observed values, typically a series of risk factor. Each S_i is already assigned label based on relevant patient's most recent laboratory test results. Underlying D-dimensional, irregular and sparsely

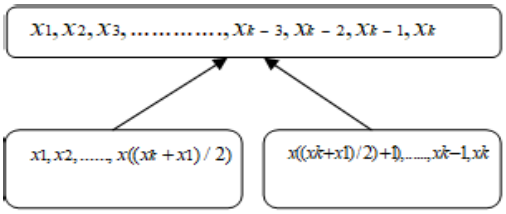

FIGURE 2. Windowing and the subintervals.

sampled time series data can have observations at different time interval as well as different total number of observations. Hence, for each irregular and sparsely sampled S_i we classify the given set of pairs of numbers $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots, (x_k, y_k)$ into upper and lower edges represented by x_1 and xk respectively. Then we divided $y_1, y_2, y_3, \ldots\ldots y_k$ into two subparts with respect to time interval i.e. $x1$ to $(xk + x1)/2$ in one interval and $((xk + x1)/2) + 1$ to xk in second interval as depicted in Figure 2. If value at point is required to approximate, values at point x_1, x_2 and x_4 are utilized. The same process continues to approximate a particular value till the value of point $x(x_k + x_1)/2$.

However, to find any missing value i.e x_n or $f(x_n)$ in interval two, the available values, x_{n+1} to x_k and x_{n-1} are utilized. For example, if missing value is at point Xk-1, only the value of point x_k and x_{k-2} will be utilized to find value $f(x_{k-1})$ for poin x_{k-1}, while all the previous values will be ignored. If value at point x_{k-2} is required, values at point x_k, x_{k-1} and x_{k-3} are utilized. The process continues until the value of point $x(x_k + x_1)/2$.

In specific scenario, if missing value is at point $x(x_k + x_1)/2$ then value at point $x((x_k + x_1)/2) + 1$ is utilized to find its value based on scheme being applied in interval one. On the other hand if missing value is at point, $x((x_k + x_1)/2) + 1$ then value at point $x(x_k + x_1)/2$ is utilized to find its value based on scheme being applied in interval two. Kingpin is selection of points to calculate the values for missing spots. Subsequently, on the bases of selected value points, divided differences will be calculated and further procedure of NDDM will be followed. According to the above mentioned description the pseudocode of the proposed INNDM is depicted in Figure 3. Under this setting, complete data set is not considered, only the main points or divided intervals are employed. As the interval for calculation is not lengthy, Runge Phenomenon will not exist anymore giving rise to improved performance for further processing.

2) PROGNOSTIC MODELLING

a: HIDDEN MARKOV MODEL

Following the application of our proposed approximation method based on NDDM as described above all D-dimensions of the input multivariate time series EMRs have been represented in term of regularly spaced set of reference time points $r_{1n}, r_{2n} \ldots\ldots r_{in} | S_i$. Thus we refer the complete set of approximation method output as $S_{yn=}(x_{dn}, y_{dn})$, which can be represented as a matrix of $nd \times T$.

To obtain missing values on points existing between interpolated data interval, consisting of N distinct no. $x_1, x_2, x_3, \ldots, x_n$ for function f:

Input : $x_1, x_2, x_3 .., x_n$; Values $f(x_1), f(x_2), f(x_3), \ldots, f(x_n)$ as

$F_{10}, F_{20}, \ldots, F_{n0}$, missing value point x_a ;

Output: Missing Value for point $x_a i.e. f(x_a)$

Step 1: Set $x_{check} = (x_n + x_1)/2$; Set $x_{LB} = 1$;, $x_{UB} = 1$;

/* x_{LB} and x_{UB} are used to save lower bound and upper bound points for interval respectively, used for calculation. Whereas x_{check} is used to determine interval in which calculations will be performed.*/

Step 2: $if(x_a \leq x_{check})$

$x_{LB} = x_1$;

$x_{UB} = x_a + 1$;

$elseif(x_a \succ x_{check})$

$x_{LB} = (x_a - 1)$;

$x_{UB} = x_n$;

Step 3:

$x_{count} = x_{LB}$;

$Count = 0$

$While(x_{count} \leq x_{UB})$

{

$if(F(x_{count}, 0)! = NULL)$

{

$Count = Count + 1$

}

$x_{count}++$;

}// Count is used to determine no of time loop will run

Step 4: For i=1,2,..... *Count*

For j=1,2,.....i

Set:

$$Fx_{LB}+i,\ x_{LB}+j = \frac{Fx_{LB}+i,\ x_{LB}+j-1 \ - \ Fx_{LB}+i-1,\ x_{LB}+j-1}{(x_{LB}+i)-(x_{LB}+i-j)}$$

Output: The value $F(x_a)$ for point x_a where

$$F(x_a) = i = \sum_{i=0}^{Count} F_{LB} + i, LB + i \prod_{j=1}^{i} (x_a - x_{LB} + j)$$

FIGURE 3. Pseudocode for Improved Newton divided difference Method (INDDM).

Subsequently, we also identified potentially relevant risk factors from the CPCSSN dataset, using logistic regression analysis. The identified potentially relevant risk factors of T2DM were also validated by an expert endocrinologist with more than 20 years' experience in medical practice. As the study sample data hold continuous type of data thus, to evaluate the proposed method, a systematic approach, named GaussianHMM, a variant of classical hidden makrov model has been applied, as the underlying inference model, for modeling prognostic prediction of T2DM. It has been shown to be a robust finite probability density distribution model, especially for dynamic systems [29], [30]. It allows us to predict temporal latent (hidden) states based on a set of observed variables by incorporating markov chain assumption, such

as $p(S_t|S_{t-1}) = p(S_t|S_{t-1}, S_{t-2}, S_{t-3}, \ldots\ldots\ldots S_0)$. The basic concept of HMM is that the observed variables in the underlying system have no one-to-one association with hidden states but are linked to states by the joint probability distribution. It is a doubly stochastic process that incorporates a Markov chain principal as the basic stochastic process [31]. Therefore, a HMM is a stochastic process of connected probabilistic states, where each state generates an observation i.e. $p(O_t|S_t) = p(O_t|S_t, S_{t-1}, S_{t-2}, S_{t-3}, \ldots\ldots\ldots S_0)$.

The primary idea related to HMM was published by Baum and Petrie [32].

b: EXPERIMENTAL SETTING

In this study, a series of experiments are designed to analyze and compare INDDM with standard NDDM for handling irregular and sparsely sampled EMRs data as input for prognostic prediction task. All experiments are conducted using the previously described data obtained from CPCSSN. The objective here is to identify ongoing risk of T2DM in a particular individual based on his/her clinical measurements.

In the first experiment, we evaluated the predictive performance of GaussianHMM using multivariate time series EMRs data related to only those patients who are enrolled in CPCSSN and have complete information related to each risk factor considered potentially significant in this study and do not have any differential loss to follow up. Hence, this derived study sample contained information related to 911 individuals of aged $\geq$18 years.

In the second experiment, we incorporated multivariate irregular data obtained from CPCSSN directly as input, denoted by original features matrix, inclusion criteria for this dataset can be seen under healthcare data section. Whereas, in the third experiment the same dataset was preprocessed using NDDM in order to handle with sparsity and irregularity in data before developing and evaluating the model.

On the other hand, in the fourth experiment the same irregular and sparsely sampled dataset, used in second experiment, was again refurbished or approximated using our proposed novel technique(INDDM) primarily based on NDDM to get values for x_s at each reference time point $r_i|S_i$ for which no measurement was taken. The descriptive information related to the above mentioned subsets of CPCSSN data is depicted in Table 1.

HMM based applications primarily belong to the category of classification and it can be categorized into two sub phases: namely training and a decoding phase. In the training phase, a set of parameters, generally parameterized by a set of probabilities, is calculated based on training dataset of observed sequences. On the other hand, in the decoding phase we compute the most probable state path that could have generated the observed sequence. The objective of this phase is to classify input sequence from validation data.

In this study hold-out method was used in each experiment for the development and validation of our model.

TABLE 1. An abstract overview of the derived study samples.

Predictors	Findings	
Demographic (Gender, Age)	Regularly sampled data	Irregularly sampled data
Sample size without duplicate	911	1918
Female, sample size (%)	556, (61.03)	775 (61.03)
Male age mean ± SD,Years	58.97±11.83	63.19± 11.74
Female age mean ± SD,Years	53.03± 11.02	57.73± 11.92
Vital Signs/ clinical measures		
sBP, mean ± SD, mm Hg	127.611±15.86	128.611±15.86
Diabetes Mellitus frequency (%)	214 (23.49)	584 (23.49)
Lab Values		
Fasting blood glucose, mean±SD, mmol/L	5.573 ± 1.93	6.029± 1.51
Triglycerides, mean ± SD, mmol/L	1.705±1.027	1.72±1.02
High-Density Lipoprotein, mean ± SD, mmol/L	1.313±0.366	1.356±0.39
Light Density Lipoprotein, mean ± SD, mmol/L	2.47±0.97	2.442± 0.851
HbA(1c), mean ± SD, mmol/L	6.316±0.824	6.268±0.95
Cholesterol mean ± SD, mmol/L	4.938±1.178	5.409±0.59
Body Mass Index, mean ± SD, kg/m^2	28.76 ± 5.818	29.81 ± 6.362

SD, Standard Deviation; sBP, systolic Blood Pressure; HbA(1c), Glycated Hemoglobin.

Thus, in each experiment the derived EMRs dataset is further divided into two subsets namely raining and testing datasets (divided in the ratio of 80% and 20% respectively).

In all the above mentioned experiments training and test datasets contained information related to only potentially significant risk factors, those demonstrated positive association with T2DM when logistic regression analysis was performed as shown in Table 2.

HMM also has structural assumptions and assumed to be composed of the set of hidden states (corresponding to $S = \{s_1, s_2, s_3 \ldots\ldots s_m\}$ diabetic and non-diabetic in our scenario) along with a set of observation variables, V_k where $K = 1, 2, \ldots\ldots 8$.

As our data retained continuous type of variables thus, the observation probability assumes the Gaussian distribution hence, the standard GaussianHMM is specified by $\lambda = \{A, \mu, \sigma, \pi\}$ where A, μ, σ and π represent the transition probability matrix, mean and variance of the distribution corresponding to the state s_i and Baum-Welch algorithm [33] was used to draw these parameters from the training data as follows:

$$\theta = \left(\begin{array}{ll} \pi = \pi_i = \{q_1 = s_i\} & A = a_{i,j} = p(q_{t+1} = s_j | q_t = s_i) \\ \textit{Priorprobability}, & \textit{Transition propabilities matrix}, \\ & B = b_i(K) = b_i(O_t = V_k) = \mathcal{N}(V_k, \mu_i, \sigma_i) \\ & \textit{Emission probabilities matrix} \end{array} \right)$$

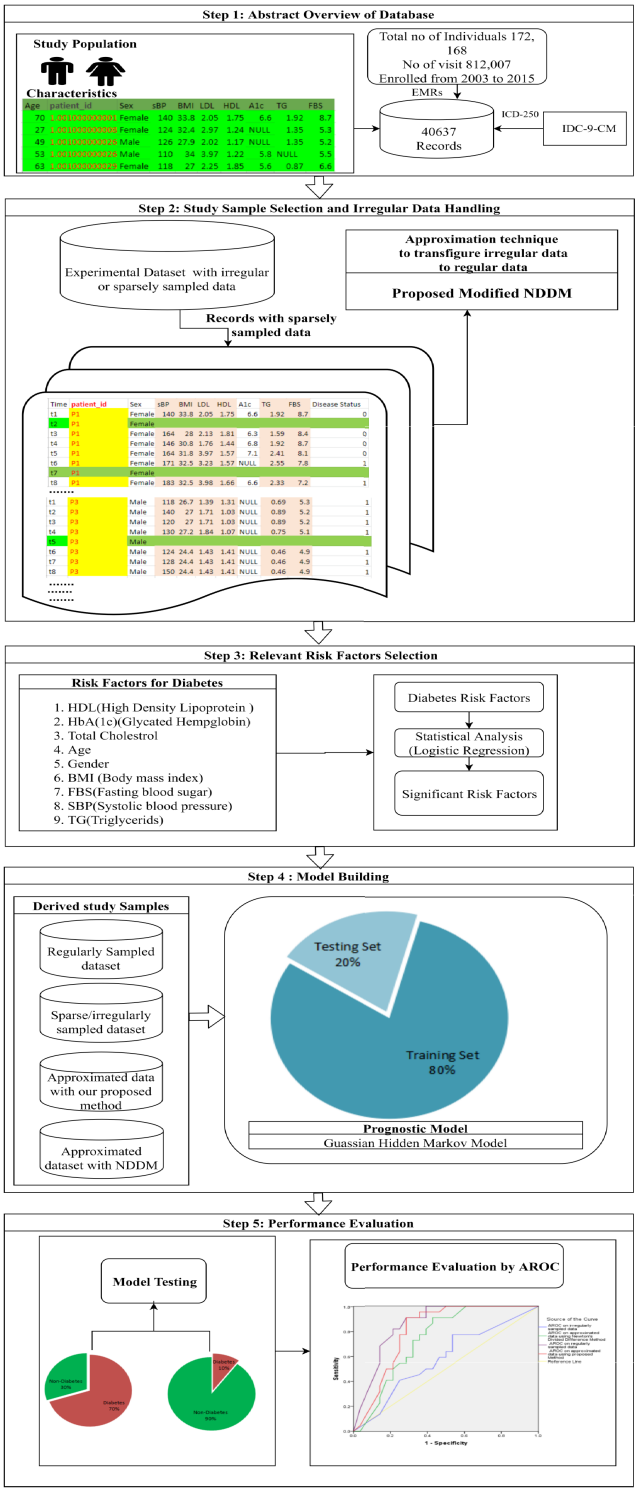

FIGURE 4. Framework for the prediction of T2DM using irregular and sparsely sampled data.

where $\mathcal{N}$ is Gaussian probability density function that can be defined as below:

$$p\,(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma_0}\,\exp\left(\frac{-(x-\mu)^2}{2\sigma}\right)$$

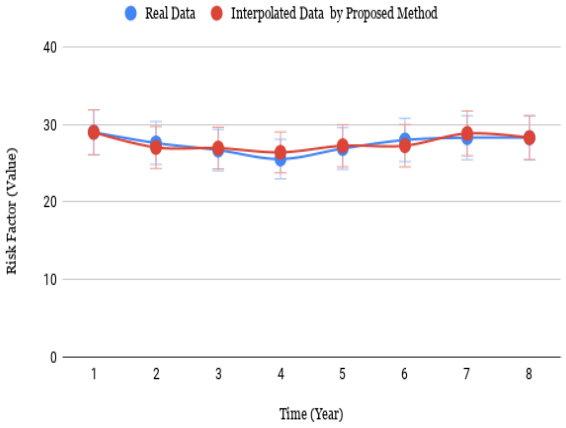

FIGURE 5. Approximation results using improved Newton's divided difference method.

The training phase of GaussianHMM, undoubtedly, is more complicated and tricky to settle down than the decoding phase. Nevertheless, the decoding procedure is the key to solve the classification problem. To solve the decoding problem Viterbi algorithm [34] is widely used [35]. In general, it is considered as the masximum a posteriori estimation of the most likely sequence of hidden states given the model the observed sequence $O = \{O_t^{(l)}, t = 1, 2, , \ldots\ldots\ldots\ldots T, l = 1, 2, 3, \ldots\ldots\ldots L\}$ where $O_t \epsilon R^D$ and l is the number of observation variables.

Subsequently, Viterbi decoding technique from HMM API (Hmmlearrn) was employed to carry out diagnostic and prognostic inferences in order to investigate the 8 year risk T2DM.

Subsequently, to analyze the relative differences among methods across different subset of data and to evaluate the generalizability of each model Area under the Receiver Operating Characteristic Curve (AROC) was incorporated as the evaluation measure. It demonstrates the performance of the model, without regard to class distribution or error costs.

The statistical analysis was performed using IBM SPSS Statistics (version 19) and Python (Version 2.7) was used as the development tool. An abstract overview of the process involved in the methodological design of this study is depicted in Figure 2.

III. RESULTS

As depicted above, in this research we proposed a novel approximation method based on NDDM to leverage irregularly and sparsely sample EMRs data to prognosticate 8 years risk of ongoing T2DM in a particular individual.

The experimental results demonstrated that the proposed INDDM performs better as compare to the NDDM as depicted in Figure 5 and 2 respectively. The proposed method effectively deals with the Runge phenomenon when approximated the value on edges. Furthermore, the produced results are also close to the real value.

As a secondary analysis, a logistic regression analysis is also conducted to assess the significant p-value of each risk factor with the incident of T2DM. To obtain the promising

TABLE 2. Analysis of the association between individual risk factor and incident of T2DM.

Explanatory variables	OR (95% C.I.)	P Value
Age	1.002 (.999 -1.006)	3.49E-22
sBP	.995 (.993-.997)	7.02E-07
Body Mass Index;	1.036 (1.030 -1 .052)	1.60E-58
Light Density Lipoprotein	.621 (.528 -.732)	1.56E-23
High Density Lipoprotein	.577 (.480 -.695)	5.63E-24
HbA(1c)	12.565 (10.902 -14.482)	7.94E-143
Triglycerides	1.183 (1.093 -1.281)	5.34E-07
Fasting blood glucose	5.907 (1.281- 5.967)	0.000
Total Cholesterol	.935 (.795-1.098)	0.411
Intercept		3.63E-145

Nagelkerke $R^2 = 0.546$

OR, Odds Ratio; C.I. Confidence Interval; sBP, systolic Blood Pressure; HbA(1c), Glycated Hemoglobin

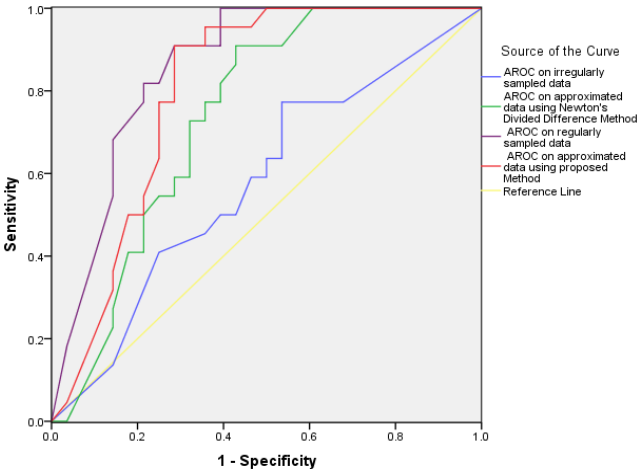

FIGURE 6. Comparative analysis of predictive performance of prognostic model in term of area under receiver operating characteristic curve over approximated and derived datesets from CPCSSN.

results, it is very crucial to effectively select the relevant features before the development of classification model. According to logistic regression analysis all the risk factors were positively associated with the incident of diabetes except total cholesterol as demonstrated in Table 2.

A comparative analysis of the logistic analysis result demonstrated that HbA(1c) is the most significant factor of T2DM as compare to other risk factors included in this analysis. It has the best predictive power to investigate the ongoing risk of T2DM with an odds ratios of ($p < 0.0005$, OR = 12.565 [95% CI, 10.902 -14.482]). It can also be deduced that HbA(1c) solely was the prime factor capable of prognosticate T2DM risk. While on the contrary, FBG is standing at second ($p < 0.0005$, OR = 5.907 [95% CI, 1.281- 5.967]); so these findings confirm the common sense and the clinical diagnosis basis.

According to results it can also be observed that elevated BMI, triglycerides and age are associated with progressively higher risk of developing T2DM ($p < 0.0005$, OR = 1.036 [95% CI, 1.030 −1 .052; $p < 0.0005$, OR = 1.183 [95% CI, 1.093 −1.281; $p < 0.0005$, OR = 1.002 [95% CI, 0.999 −1.006 respectively). It can be observed that not all overweight or obese patients have T2DM, and vice versa. But it is a well accepted and readily accessible concept in the existing literature that the more fatty tissue you have, the more resistant your cells become to insulin [42]. However, total cholesterol does not added into the model and does not show any significant relation with T2DM risk, thus, excluded from the risk factors list.

Subsequently, four prognostic models were built from four different subsets of CPCSS datasets, as mentioned above. These four subsets of dataset comprise of only positively correlated risk factors. Among these four subsets of dataset two were original datasets, however comprise of regular and irregularly sampled datasets and directly feed into GaussianHMM without any modification. While, the remaining two were approximated datasets using NDDM and improved NDDM proposed in this study. The derived dataset resulted in a total of 1918 unique individuals with 15,344 clinical visits, spanning 8 years timeframe. An abstract and descriptive information related to eligible cohorts is already depicted in Table 1.

The superiority of our proposed technique is to improve the prognostic prediction results' precision. To evaluate the efficacy of our proposed technique in handling irregular and sparsely sample EMRs data and to investigate the ongoing risk of T2DM, we developed a comparative analysis from the experimental results. Figure 6 demonstrated a comparative analysis among the AROCs of the four prognostic models that were built from four different subsets of CPCSS datasets as mentioned above. As expected, the predictive performance of the GaussianHMM over approximated data generated using our proposed method was considerably better in general (AROC 80.4%, p-value < 0.0005, SE = 0.064 [95% CI, (0.679-0.930)]) than the performance over approximated dataset using NDDM (AROC 0.740 p-value < 0.0005, SE = 0.071 [95% CI, (0.602-0.879)]), as shown in Table 3.

It has a good balance of specificity and sensitivity with an AROC value of 80.4% which is statistically significant with p-value <0.0005 with the Confidence Interval (CI) of 95%. However, the AROC value of GaussianHMM over irregularly sampled EMRs data is 59.7% which is comparatively low. The results drawn from the regularly sampled dataset were also compared with our proposed method, as shown in Figure 6. It can be observed that our Improved NDDM technique for handling irregular and sparsely sampled data performed reasonably well and demonstrating a high reliability of discrimination for the GaussianHMM model over the approximated dataset when compared with the performance over regularly sampled data.

IV. DISCUSSION

Physiological information contained in EMRs is the crucial source for disease prognostic modeling. Nevertheless, the structure of the longitudinal clinical data along with the nature of the clinical settings present substantial hurdles that confound machine learning algorithms in inference and learning process, and/or may significantly influence the performance of downstream applications [20], [21].

This is particularly true about HMM based prognostic model [36]. Specifically, the dynamic scope of the time scale in EMRs is one of the possibly contributing elements for irregular and sparsely sampled clinical information which is a common but complicated problem in almost every healthcare data. As mentioned above in this study we also utilized HMM based method as a prognostic model to investigate the ongoing risk of T2DM. However, HMM presumes that the measurement data is gathered regularly or in such a way that the time interval between two consecutive observations is constant but this is not the case commonly.

Although, NDDM, a recursive division process, works well as compared to other interpolation techniques [22]–[25], [37] when handling irregular and sparse data, but as it is polynomial interpolation, its interpolants are vulnerable to the Runge Phenomenon [38], [39], as depicted in Figure 1.

In this work, we proposed a technique to handle irregular and sparsely sampled data along with overcoming the RP, particularly on a finite interval and showed experimentally that our proposed method can achieve much lower errors and have the ability to produce the approximated value near the real value as shown in Figure 4. Nevertheless, the error has not been converge to zero as shown in Figure but it can be observed that the Runge Phenomenon has reduced to a satisfactory level when we compared it to the results shown in Figure 1. The results depicted in Figure 4 are also sub-geometric rate of convergence is achieved. Furthermore, computational results revealed that the proposed method enumerated more near actual values as compared to NDDM. The average runtime complexity of the proposed method is O(n^2), same as NDDM where n is the number of points in polynomial involved in interpolation. However, the major difference between the two is the value of n involved in calculations. In NDDM, all values involved in polynomial are used for interpolation while in improved NDDM, whole polynomial is divided into 2 intervals and only half of the values of the polynomial are used for the procedure. This division not only decreases complexity of algorithm by reducing the value of n, but this division removes the major drawback of NDDM i.e. Runge Phenomenon. The decrease in complexity results in less computation time for a particular set of points as compared to NDDM.

Once the dataset is prepared, we also performed logistic regression analysis on the derived dataset. The objective of this statistical analysis was to explore the relation between the contributing risk factors and prevalence of diabetes mellitus. Thus this statistical analysis resulted in a set of potentially significant risk factors for developing T2DM as shown in Table 2. All the risk factors incorporated in this analysis were positively associated with T2DM incidents except total cholesterol. Therefore, total cholesterol was excluded from the further analysis. HbA(1c) and FBG were ranked as the most positively associated risk factors ($p < 0.0005$, OR $= 12.565$ [95% CI, 10.902 -14.482] and $p < 0.0005$, OR $=$ 5.907 [95% CI, 1.281- 5.967] respectively) with the ability to solely predict the diabetes incidence. Hence, these findings corroborate with the common sense and the clinical diagnosis basis. Furthermore, the result of the logistic regression analysis is also conforming to the existing studies [2]. In view of the objective of the proposed research four prognostic models based on GaussianHMM, a variant of classical HMM, were built from regular, irregular and approximated datasets derived from CPCSSN data. These subsets of CPCSS data comprise of only significant risk factors resulted from logistic regression analysis.

The first major contribution of this study is to propose a novel method based on NDDM for handling the irregular EMRs data for disease modeling. In this study we also utilized GaussianHMM, a variant of classical hidden makrov model, as the underlying inference model, for modeling prognostic prediction of T2DM to manage the irreversible and adverse outcomes. Therefore, four prognostic models were built from four different derived study sample, as mentioned above. According to the results predictive performance of the GaussianHMM over approximated data using our proposed method was considerably better (AROC 80.4%, p-value < 0.0005, SE $= 0.064$ [95% CI, (0.679-0.930)]) than the performance over approximated dataset using NDDM (AROC 0.740 p-value < 0.0005, SE $=$ 0.071 [95% CI, (0.602-0.879)]), as shown in Table 3 with their respective odd ratios. The graphical representation of the results of predictive models, related to four above mentioned experimental settings, are depicted in Figure 6.

The AROC value of GaussianHMM over irregularly sampled EMRs data is 59.7% which is comparatively low or it could be just considered as better than the random guess and do not provide significant information to take decisive steps. The proposed method along with the GaussianHMM yielded AROC 80.4% which is 20.7% higher than the baseline method. Thus, from experimental results it can be inferred that our proposed method for handling irregular and sparse data contributed to improve the predictive performance comparatively.

On the other hand, results drawn using the regularly sampled dataset demonstrated high performance (AROC 85.9 %) among all other remaining derived datasets, as shown in Figure 6. Furthermore, calibrated results obtained from the HMM on our baseline data set can be seen in Perveen *et al.* [45]. Thus, it can be concluded from the drawn results that our proposed method performed reasonably well and demonstrating a higher reliability of discrimination for the GaussianHMM model over the approximated dataset

TABLE 3. Results of the GaussianHMM over derived data samples.

	AROC	(SE) Std. Error [a]	Asymptotic Sig.[b]	Asymptotic 95% CI	
				Lower Bound	Upper Bound
Over the irregularly sampled data	0.597	0.082	0 .343	0.418	0.739
Over the regularly sampled data	0.859	0.053	0.000	0.754	0.963
Over the approximated data using our proposed method	0.804	0.064	0.000	0.679	0.930
Over the approximated data using Newton 's divided difference method	0.740	0.071	0.004	0.602	0.879

when compared with the performance over regularly sampled data.

Diabetes is a costly ailment. According to ADA in 2017, approximate cost of diagnosed diabetes was $327 that includes $237 spent for medical costs and $90 in terms of reduced productivity [40]. In November, 2018 National Health Service (NHS) bill for blood glucose-lowering medicines for the first time excelled £1 billion. An ongoing survey in sub-Saharan Africa estimated that the expense of diabetes in that area is equal to 1.2% of combined total national output [41]. In the absence of proper and early interventions, this burden will continue to increase and will leave our healthcare system in unsustainable future.

Therefore, the proposed technique is an effective tool to increase awareness in pre-diabetics, prevent and/or mange the risk of developing diabetes. Utilization of such risk scoring methods as the first step for screening high risk individuals is more pragmatic than conducting FBG tests because the latter one is invasive as well as time consuming and costly. It is also crucial to provide guidance and awareness to vulnerable individuals, providers and communities to manage the adverse health outcomes. Furthermore, the proposed method is also aligned with the goals of world health Organization.

In spite of the promising results, our proposed technique has various limitations. First one is, the requirement for further processing, the proposed method required further processing entailed for the selection of values vital for interpolation. This additional processing may increase computation time for a single value. Another hitch in suggested procedure is that, due to decrease in past and future values for interpolation, may be the system lead to averaged result rather than taking into account the sense of future and past values. Moreover, for points other than in the mid of polynomial, improved NDDM proved to work satisfactory, and exhibit better performance as compared to NDDM. Nonetheless, when required point is in the mid of polynomial, sometimes NDDM manifest slightly preferable staging.

V. CONCLUSION

The results indicate the proposed method bears a significant potential to sift through and meaningfully put together the sparsely and irregularly sampled data, thereby extracting useful information with enhanced efficiency as well as effectiveness. The very ability of this method enables to make a better use of EMRs in carrying out the predictive measures for a particular disease. The proposed technique also uncovers the underlying state with likely future projection of the disease. Moreover, it also uncovers the inbuilt temporal dependencies present in the longitudinal EMRs data which are imperative for arriving at decisive steps for disease management, together with the effective utilization of healthcare resources. Therefore, it may prove as a promising tool for the identification of the individuals prone to high risk; and a mean to prompt the healthcare entities to adopt innovative preventive measures in respect of diabetes. o judge the cost effectiveness of the proposed model, further study remains warranted.

ACKNOWLEDGMENT

The authors thank their colleagues at CPCSSN for their help with the datasets.

CONFLICT OF INTERESTS AND FINANCIAL DISCLOSURES

The authors have no conflicts of interest to declare.

REFERENCES

[1] S. A. Tabish, "Is diabetes becoming the biggest epidemic of the twenty-first century?" *Int. J. Health Sci.*, vol. 1, no. 2, p. 5, 2007.
[2] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques," *IEEE Access*, vol. 7, pp. 1365–1375, 2019.
[3] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 27, no. 1, p. S5, 2004.
[4] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation," *Diabetic Med.*, vol. 15, no. 7, pp. 539–553, 1998.
[5] J. Bhutani and S. Bhutani, "Worldwide burden of diabetes," *Indian J. Endocrinol. Metabolism*, vol. 18, no. 6, p. 868, 2014.
[6] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Comput. Sci.*, vol. 82, pp. 115–121, Jan. 2016.
[7] P. Balakumar, K. Maung-U, and G. Jagadeesh, "Prevalence and prevention of cardiovascular disease and diabetes mellitus," *Pharmacol. Res.*, vol. 113, pp. 600–609, Nov. 2016.
[8] S. Pugazhenthi, L. Qin, and P. H. Reddy, "Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease," *Biochim. Biophys. Acta-Mol. Basis Disease*, vol. 1863, no. 5, pp. 1037–1045, May 2017.
[9] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. Ohlrogge, and B. Malanda, "IDF diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018.

[10] P. Zhang, X. Zhang, J. Brown, D. Vistisen, R. Sicree, J. Shaw, and G. Nichols, "Global healthcare expenditure on diabetes for 2010 and 2030," *Diabetes Res. Clin. Pract.*, vol. 87, no. 3, pp. 293–301, Mar. 2010.
[11] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression," *Sci. Rep.*, vol. 8, no. 1, p. 2112, 2018.
[12] R. D. Cebul, T. E. Love, A. K. Jain, and C. J. Hebert, "Electronic health records and quality of diabetes care," *New England J. Med.*, vol. 365, no. 9, pp. 825–833, Sep. 2011.
[13] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annu. Rev. Public Health*, vol. 36, no. 1, pp. 345–359, Mar. 2015.
[14] D. Blumenthal and M. Tavenner, "'The meaningful use' regulation for Electron. Health records," *New England J. Med.*, vol. 363, no. 6, pp. 501–504, 2010.
[15] D. Chen, Z. Runtong, S. Xiaopu, W. V. Li, and H. Zhao, "Predicting the interaction between treatment processes and disease progression by using hidden Markov model," *Symmetry*, 2018.
[16] Y. Li, S. Swift, and A. Tucker, "Modelling and analysing the dynamics of disease progression from cross-sectional studies," *J. Biomed. Inform.*, vol. 46, no. 2, pp. 266–274, Apr. 2013.
[17] R. D. Saraçoglu, "Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction," *Eng. Appl. Artif. Intell.*, vol. 25, no. 7, pp. 1523–1528, Oct. 2012.
[18] P. Srikanth, "Using Markov chains to predict the natural progression of diabetic retinopathy," *Int. J. Ophthalmol.*, vol. 8, no. 1, p. 132, 2015.
[19] M. El Nahas, S. Kassim, and N. Shikoun, "Profile hidden Markov model for detection and prediction of hepatitis C virus mutation," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, p. 251, 2012.
[20] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proc. 2nd ACM SIGHIT Symp. Int. Health Informat. (IHI)*, 2012, pp. 389–398.
[21] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2018, pp. 1174–1182.
[22] R. B. Srivastava and P. K. Srivastava, "Comparison of Lagrange's and Newton's interpolating polynomials," *J. Exp. Sci.*, to be published.
[23] A. A. Al-Hussainan, B. M. Al-Eideh, and Y. S. H. Al-Zalzalah, "The adjusted empirical Lorenz curve using a Newton's divided difference formula," *Int. J. Appl. Math.*, vol. 7, no. 3, pp. 325–332, 2001.
[24] A. Al-Ayyoub, "Pipelined algorithm for Newton's divided difference interpolation," *Comput. Struct.*, vol. 58, no. 4, pp. 689–701, Feb. 1996.
[25] H. E. Salzer, "A multi-point generalization of Newton's divided difference formula," *Proc. Amer. Math. Soc.*, vol. 13, no. 2, pp. 210–212, Apr. 1962.
[26] M. Mashayekhi, F. Prescod, B. Shah, L. Dong, K. Keshavjee, and A. Guergachi, "Evaluating the performance of the framingham diabetes risk scoring model in canadian electronic medical records," *Can. J. Diabetes*, vol. 39, no. 2, pp. 152–156, Apr. 2015.
[27] J. Yoon, A. Alaa, S. Hu, and M. Schaar, "ForecastICU: A prognostic decision support system for timely prediction of intensive care unit admission," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1680–1689.
[28] E. E. Kalu, *Numerical Methods With Applications: Abridged*. Abu Dhabi, United Arab Emirates: Lulu. 2009.
[29] P. Kenny, M. Lennig, and P. Mermelstein, "Speaker adaptation in a large-vocabulary Gaussian HMM recognizer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 9, pp. 917–920, Sep. 1990.
[30] T. Artières, J. M. Marchand, P. Gallinari, and B. Dorizzi, "Stroke level modeling of on line handwriting through multi-modal segmental models," in *Proc. Int. Workshop Frontiers Handwriting Recognit. (IWFHR)*, 2000, pp. 1–11.
[31] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[32] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966.
[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B (Methodol.)*, vol. 39, no. 1, pp. 1–22, 1977.
[34] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.
[35] Y. Lifshits, S. Mozes, O. Weimann, and M. Ziv-Ukelson, "Speeding up HMM decoding and training by exploiting sequence repetitions," *Algorithmica*, vol. 54, no. 3, pp. 379–399, 2009.
[36] Y. Y. Liu, S. Li, F. Li, L. Song, and J. M. Rehg, "Efficient learning of continuous-time hidden Markov models for disease progression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3600–3608.
[37] J. H. Mathews, *Numerical Methods for Computer Science, Engineering, and Mathematics*. Upper Saddle River, NJ, USA: Prentice-Hall, 1986.
[38] J. P. Boyd, "Six strategies for defeating the Runge phenomenon in Gaussian radial basis functions on a finite interval," *Comput. Math. Appl.*, vol. 60, no. 12, pp. 3108–3122, Dec. 2010.
[39] B. Fornberg and J. Zuev, "The Runge phenomenon and spatially variable shape parameters in RBF interpolation," *Comput. Math. Appl.*, vol. 54, no. 3, pp. 379–398, Aug. 2007.
[40] American Diabetes Association, "Economic costs of diabetes in the US in 2017," *Diabetes Care*, vol. 41, no. 5, pp. 917–928, 2018.
[41] C. Moucheraud, C. Lenz, M. Latkovic, and V. J. Wirtz, "The costs of diabetes treatment in low- and middle-income countries: A systematic review," *BMJ Global Health*, vol. 4, no. 1, Feb. 2019, Art. no. e001258.
[42] J. Edqvist, A. Rawshani, M. Adiels, L. Björck, M. Lind, A.-M. Svensson, S. Gudbjörnsdottir, N. Sattar, and A. Rosengren, "BMI and mortality in patients with new-onset type 2 diabetes: A comparison with age- and sex-matched control subjects from the general population," *Diabetes Care*, vol. 41, no. 3, pp. 485–493, Mar. 2018.
[43] R. E. Watkins, S. Eagleson, B. Veenendaal, G. Wright, and A. J. Plant, "Disease surveillance using a hidden Markov model," *BMC Med. Inform. Decis. Making*, vol. 9, no. 1, p. 39, 2009.
[44] D. Madigan, "Bayesian data mining for health surveillance," in *Spatial and Syndromic Surveillance for Public Health*. A. B. Lawson and K. Klienman, Eds. Chichester, U.K.: Wiley, 2005, pp. 203–221.
[45] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Prognostic modeling and prevention of diabetes using machine learning technique," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.
[46] M. Yamin and A. A. A. Sen, "Improving privacy and security of user data in location based services," *Int. J. Ambient Comput. Intell.*, vol. 9, no. 1, pp. 19–42, Jan. 2018.
[47] N. A. Mhetre, A. V. Deshpande, and P. N. Mahalle, "Trust management model based on fuzzy approach for ubiquitous computing," *Int. J. Ambient Comput. Intell.*, vol. 7, no. 2, pp. 33–46, Jul. 2016.
[48] H. Kort and J. Van Hoof, "Telehomecare in The Netherlands: Barriers to implementation," *Int. J. Ambient Comput. Intell.*, vol. 4, no. 2, pp. 64–73, Apr. 2012.
[49] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "A hybrid approach for modeling type 2 diabetes mellitus progression," *Frontiers Genet.*, vol. 10, p. 1076, Jan. 2019.

SAJIDA PERVEEN is currently pursuing the Ph.D. degree in healthcare informatics with the University of Engineering and Technology, Lahore, Pakistan, under the supervision of Prof. M. Shahbaz. Her research interest includes mining interesting and implicit patterns from the structured and unstructured data.

MUHAMMAD SHAHBAZ received the Ph.D. degree from Loughborough University, U.K. He is currently a Full Professor with the Department of Computer Science and Engineering, University of Engineering and Technology. He has delivered several talks in the industry at National and International levels and in various conferences around the world. He has a wide experience in the field of data science and has published more than 60 articles in the same domain. His research interests include healthcare informatics, fog computing, data science, and artificial intelligence.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently serving as an Associate Professor and Associate Chair of the Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, KSA. She has over one hundred publications that have around 1800 citations with H-index 28. Her most publications are in the biomedical research published in ISI/SCIE indexed. Her primary research focus, in recent years, is medical imaging, pattern recognition, data mining, MRI analysis, and soft-computing. She received the Best Student Award from the Faculty of Computing, UTM, in 2012. Due to her excellent research achievement, she is included in Marquis Who's Who (S & T) 2012. She is the Leader of the Artificial Intelligence and Data Analytics Research Laboratory, PSU, and an active professional member of ACM, AIS, and IAENG organizations. She is the PSU WiDS (Women in Data Science) Ambassador at Stanford University. She is currently an Editor and reviewer of reputed journals and on the panel of TPC of International conferences. On the accreditation side, she is a skilled lady with ABET and NCAAA quality assurance.

KARIM KESHAVJEE is currently an Adjunct Professor with the University of Toronto and a Visiting Scientist with Ryerson University. He trained as a Family Physician and is a Health Informatics Consultant. He designs and implements large scale disease focused research projects that are embedded in clinical practice. His goal is to implement knowledge translation into clinical practice for the improvement of patient experiences and outcomes, and for the improvement of health provider productivity.

AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia, in 2010, with specialization in forensic documents analysis and security. He is currently a Senior Researcher with the AIDA Laboratory, College of Computer and Information Sciences, Prince Sultan University, Riyadh, KSA. He is author of more than 100 indexed journal articles. His keen interests are in data mining, health informatics, and pattern recognition. He is a reviewer in many international impact-factor journals, and also a technical committee program member in a number of international conferences.

AZIZ GUERGACHI received the Ph.D. degree from the University of Ottawa, Canada. He is currently a Full Professor with the Ted Rogers School of Management, Ryerson University, and an Adjunct Professor with the Department of Mathematics and Statistics, York University, Canada. His research interests include system modeling, explainable AI, and data analytics.